MULTIPLE CHANGE POINTS DETECTION IN VECTOR AUTOREGRESSIVE MODELS

By

PEILIANG BAI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2021

I dedicate this to
my mother, Ruihua Shi
and my beloved girlfriend, Yue Bai

ACKNOWLEDGEMENTS

There are many people to whom I would like to express my deepest appreciation. First, I would like to thank my advisor, Professor George Michailidis, for the continuous support of my PhD study and related research. He has been providing guidance for every interesting and novel research projects, as well as my career to grow in to a mature researcher. I would like to thank him for all his encouragement, criticism, patience and broad knowledge.

My special thanks go to Professor Abolfazl Safikhani. Thanks for his excellent mentorship and generous help with my research, I will never forget the stimulating discussion with him, and his innovative research led to our productive collaborations over the years. Thanks Professor Hani Doss for his meticulous guidance for my Master's thesis and his inspiring courses. I thank Professor Arunava Banerjee for his interesting class in computer science and serving as my thesis committee member. I should also thank Professor James Hobert for his serving as my thesis committee.

I also want to thank all the staff at the Department of Statistics and the Institution of Informatics at the University of Florida. From the course enrollment to my PhD defense, Christine Miron and Alethea Geiger were always there to help me out.

Most of all, my deepest gratitude goes to my mother Ruihua, my girl friend Yue, and my happy doggy, who are always my strongest supporters. Having them is the luckiest thing in my life.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Figures page

LIST OF ALGORITHMS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

MULTIPLE CHANGE POINTS DETECTION IN VECTOR AUTOREGRESSIVE MODELS

By

Peiliang Bai

May 2021

Chair: George Michailidis
Major: Statistics

Change points detection is a very popular topic in statistics, and the study of change point detection has a wide range of applications. Recent techniques for detecting change points in high dimensional time series data have a restriction on the sparse transition matrix, however, in many applications the autoregressive dynamics of the time series exhibits also a low dimensional structure. An investigation on change point detection for a generalized structured autoregressive time series model is needed.

In the first part of the dissertation, we study the change point detection problem on a high dimensional vector autoregressive model, whose transition matrices is further decomposed into a fixed low rank component plus a time varying sparse component, and we propose a blocked fused lasso algorithm to detect the change points. Next, we extend this structure to a general one: both low rank and sparse components are time varying. For the general model, we first address the problem of detecting a single change point using an exhaustive search algorithm and establish a finite sample error bound for its accuracy. Then, we consider the results to the case of multiple change points that can grow as a function of the sample size. We propose a two-step algorithm, wherein the first step, an exhaustive search for a candidate change point is employed for overlapping rolling windows, and subsequently a backwards elimination procedure is used to screen out redundant candidates. The two-step strategy yields consistent estimates of the number of change points and the locations of the change points. To reduce the computational complexity of

the two-step algorithm, we also establish a surrogate VAR model with a weakly sparse transition matrix.

In the second part of dissertation, we work on establishing a fast detection algorithm that accurately detects the number of change points together with their location and subsequently estimates the model parameters in each stationary segment. We also show the computational effectiveness on both synthetic and real data sets.

CHAPTER 1
INTRODUCTION

1.1 Overview of Change Point Detection

Change point detection considers identifying the differences in the given data. It tries to estimate the number of changes as well as locate the changes. The study of change point detection has a long history in statistics (Bai 1997), signal processing (Basseville & Nikiforov 1993), economics and finance (Frisén 2008), quality control (Qiu 2013), risk analysis (McGlohon et al. 2009), surveillance and environmental monitoring (Nobre & Stroup 1994), and neuroscience (Koepcke et al. 2016). A change point represents a discontinuity in the parameters of the data generating process.

Most of change point detection problem can be considered either in an (1) *offline* setting, or (2) an *online* one. In the first case, a sequence of observations is fixed and questions of interest include: (i) whether there exist change points and (ii) if there exist change points, identify their locations, as well as estimate the model parameters. The offline version has been investigated extensively for various models, including univariate/multivariate time series models (Bai 1997, Killick et al. 2012), graphical models (Roy et al. 2017), and factor models (Barigozzi et al. 2018). In the online setting, one continuously obtains new observations and the main interest is in *quickest detection* of the change point, and the accuracy of the estimated location of change points. The online version problem has been considered in limited literature, including using a Bayesian framework (Adams & MacKay 2007), unsupervised approach (Zameni et al. 2020), and CUSUM based methodology (Yu et al. 2020). In this dissertation, we focus on offline change point detection.

Another thrust of the literature has focused on developing scalable and efficient algorithms to detect the change points. In (Killick et al. 2012), the author proposed a dynamic programming algorithm to detect multiple change points in an univariate time series in linear time. Safikhani & Shojaie (2020) established a fused lasso algorithm to identify multiple change points in a sparse structured VAR model, and Bai et al. (2020) has extended to blocked fused lasso and applied to a VAR model with a fixed low rank plus

a fluctuating sparse structure. In (Wang et al. 2021), the author established an optimal change point detection algorithm in sparse dynamic networks by applying Universal Singular Value Thresholding (USVT) and local refinement, and extended to stochastic block matrices.

1.2 Change Point Detection in High Dimensional Time Series Models

In recent years, the study of high dimensional time series has become increasingly important in diverse domains, including macroeconomics (Kilian & Lütkepohl 2017, Stock & Watson 2016), financial economics modeling (Billio et al. 2012, Lin & Michailidis 2017), molecular biology (Michailidis & d'Alché Buc 2013) and neuroscience (Friston et al. 2014, Schröder & Ombao 2019). There are two main modeling paradigms for capturing the cross-correlated and auto-correlated features in the high dimensional time series data: (i) dynamic factor and latent models (Bai & Ng 2008, Stock & Watson 2002, 2016, Lam et al. 2011, Li et al. 2014); and (ii) vector autoregressive (VAR) models (Lütkepohl 2013, Kilian & Lütkepohl 2017). The precondition of models in (i) is that the common dynamics of a large number of time series are driven by a relatively small number of latent factors. While VAR models aim to capture the self and cross auto-correlation structure in the time series, but the number of parameters need to be estimated grow quadratically in the number of time series under consideration. Besides these two modeling paradigms, various structural assumptions have been proposed and investigated in the literature, for example, the most popular one with *sparsity* (Basu & Michailidis 2015) structure. In many applications, the dynamics of the time series exhibits also low dimensional structure, which inspired to the introduction of reduced rank auto-regressive models (Velu et al. 1986, Wang & Bessler 2004).

In many application domains including those mentioned above, non-stationary time series are considered. One of the interpretable time series model is *piecewise-stationary*. Under this assumption, the time series data are modeled as approximately stationary between neighboring change points. There are a number of literature on change point

detection for piecewise stationary time series models. While most of those literature consider the sparse structure. (Bardsley et al. 2017) developed tests for the change point detection in functional factor models, while (Barigozzi et al. 2018) applied the binary segmentation procedure for detecting multiple change points in factor models. Wang et al. (2019) and Safikhani & Shojaie (2020) considered change points detection in *sparse* VAR models. Bai et al. (2020) investigated change points detection in *sparse* plus *low rank* VAR models.

1.3 Contributions of This Work

The key contributions of this work include proposing a generalized change point detection framework in low rank plus sparse structured piece-wise stationary VAR models, establishing a novel *information ratio* to leverage the strength of signal from the low rank and the sparse components, and introducing proper conditions to ensure identifiability of change points together with estimation of transition matrices of VAR models; developing efficient algorithms to estimate and locate the change points, as well as to estimate the model parameters within each stationary segment; constructing theoretical properties for its accuracy; illustrating the effectiveness of the proposed algorithms and strategies on both synthetic and real data sets.

1.4 Preliminary and Notations

Throughout this dissertation, we denote with a superscript "$\star$" the true value of the corresponding model parameters. Further, for any $p \times p$ matrix A, we use A' to denote its transpose, use $A^\dagger$ to represent the conjugate transpose of A, and we use $\|A\|_1$, $\|A\|_2$, $\|A\|_F$, and $\|A\|_*$ to represent the vectorzied ℓ_1-norm, spectral norm, Frobenius norm, and nuclear norm of the matrix, respectively. We also use $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues of the matrix A, respectively. As for any set $\mathcal{S}$, we denote its cardinality by $|\mathcal{S}|$, and its complement by $\mathcal{S}^c$. We also define the Hausdorff distance between two countable sets on the real line as: $d_H(A, B) = \max_{b\in B} \min_{a\in A} |b - a|$.

CHAPTER 2
MULTIPLE CHANGE POINT DETECTION IN LOW RANK AND SPARSE HIGH DIMENSIONAL VAR MODELS

In this chapter, we investigate the change point detection problem in low rank plus sparse high dimensional vector autoregressive (VAR) model. More precisely, the transition matrix of VAR model comprises of two components: a *fixed* low rank component, and a *time-varying* sparse component, respectively. The essential technical contributions include:

- Introduction of proper conditions to identify the low rank and the sparse components in piece-wise stationary VAR time series.
- Development of an efficient three-step algorithm based on fused lasso to estimate and locate the change points as well as to estimate the model parameters within each interval.
- Establish a blocked fused lasso regression model to accelerate the change points detection procedure.
- Construction of a cross-validation type method to select the key tuning parameters involved in the proposed algorithm.

2.1 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 Section 2.5. Specifically, suppose we have $T+1$ time points and there exist m_0 change points:

$$0 = t_0^\star < t_1^\star < \cdots < t_{m_0}^\star < t_{m_0+1}^\star = T,$$

such that for $t_{j-1}^\star \leq t < t_j^\star$, $j = 1, \ldots, m_0+1$, the structured VAR(1) process is given by:

$$X_t = B_j' X_{t-1} + \epsilon_t \quad \text{and} \quad B_j = L^\star + S_j^\star \tag{2-1}$$

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 throughout this chapter, Σ_ϵ is set to be fixed over segments. We

then 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 $t_j^\star$, 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 Chandrasekaran et al. (2011), 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 Agarwal et al. (2012), 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 2.2.1, Assumption H2), which reflects similar constraints imposed on the radius of non-identifiability in Agarwal et al. (2012), to constrain the signal strength originating from the low-rank component that will render changes in sparse components *detectable*.

2.2 The Change Point Detection Procedure and Theoretical Properties

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

2.2.1 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_T 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 $(\lceil \frac{T}{b_T} \rceil + 1)p^2$ 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_T to be too large as explained below.

Define a sequence of time points $1 = r_0 < r_1 < \cdots < r_{k_T+1} = T$ corresponding to the end points of the blocks (i.e. $r_{i+1} - r_i = b_T$ and $k_T = \lceil \frac{T}{b_T} \rceil$). Subsequently, by using the same notation as in the model (2-1), 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

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

block respectively, translated in matrix form notation as follows: let

$\mathcal{X} = [\mathbf{X}_{r_1}, \ldots, \mathbf{X}_{r_{k_T+1}}]' \in \mathbb{R}^{n\times p}$, $\mathcal{Y} = [\mathbf{Y}_{r_1}, \ldots, \mathbf{Y}_{r_{k_T+1}}]' \in \mathbb{R}^{n\times p}$, $\mathcal{E} = [\boldsymbol{\epsilon}_{r_1}, \ldots, \boldsymbol{\epsilon}_{r_{k_T}}]' \in \mathbb{R}^{n\times p}$

and

$$\mathcal{Z} = \begin{bmatrix} \mathbf{X}'_{r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{X}'_{r_2} & \mathbf{X}'_{r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}'_{r_{k_T+1}} & \mathbf{X}'_{r_{k_T+1}} & \cdots & \mathbf{X}'_{r_{k_T+1}} \end{bmatrix} \in \mathbb{R}^{T\times pk_T}.$$

We then formulate the model (2-1) into the following linear regression problem

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

wherein $\Theta = [\theta'_1, \ldots, \theta'_{k_T}]' \in \mathbb{R}^{pk_T\times p}$. We set $\theta_1 = S^\star_1$; for $i = 2, 3, \ldots, k_T$, and for the subsequent ones we set

$$\theta_i = \begin{cases} S^\star_{j+1} - S^\star_j, & \text{when } i = t_j \text{ for some } j, \\ 0, & \text{otherwise.} \end{cases} \tag{2-3}$$

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 (2-2) implies that the model coefficients Θ and L can be estimated through the following restricted penalized least squares problem:

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

In the objective function above, $\Omega \overset{\text{def}}{=} \{L \in \mathbb{R}^{p\times p} : \|L\|_\infty \leq \frac{\alpha}{p}\}$ 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; $\lambda_{1,T}$, $\lambda_{2,T}$ and $\lambda_{3,T}$ 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 Rinaldo et al. (2009). Notice that the space of feasible solutions for optimization problem (2-4) consists of all pairs (C, D) such that the square p-dim matrix C is low-rank and belongs to the space Ω, while the matrix $D \in \mathbb{R}^{pk_T\times p}$ is sparse. Based on Assumption A3, the number of blocks k_T is much larger than m_0. This expansion on the space of model parameters is a crucial development in Step 1.

Remark 2-1. *The computational complexity of estimating the sparse components in* (2-4) *is of order* $\mathcal{O}(k_T p^2)$ *Bleakley & Vert (2011). If the size of the blocks is set to 1 (i.e.* $b_T = 1$*) the method would revert to a standard fused lasso penalty Rinaldo et al. (2009). However, to speed up computations, we allow* b_T *to increase as a function of the sample size. On the other hand, larger values of* b_T *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 (2-4) 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 Basu & Michailidis (2015). 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:

(A1) 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

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

(A2) Define the information ratio

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

Then, with fixed γ, we obtain that $\|L^\star\|_\infty \le \gamma^{-1} M_S$ by A1. In this model, we recommend choosing γ in the range $\gamma \ge 1$.

(A3) 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 $\Delta_T = \min_{1 \le j \le m_0} |t_{j+1}^\star - t_j^\star|$ and $d_T^\star = \sum_{j=1}^{m_0+1} d_j^\star$, there exists a vanishing positive sequence γ_T such that, as $n \to +\infty$,

$$\frac{\Delta_T}{T\gamma_T} \to +\infty, \ \limsup \frac{b_T}{T\gamma_T} \le C < \frac{1}{12}, \frac{d_T^\star \log p}{T\gamma_T} \to 0 \text{ and } \frac{r^\star p}{T\gamma_T} \to 0.$$

Assumption A1 is standard in the high-dimensional liner regression literature, while Assumption A2 ensures identifiablity of the model parameters, as discussed previously in this Section. Assumption A3 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 Safikhani & Shojaie (2020),

Assumptions A2 and A3 in Harchaoui & Lévy-Leduc (2010), and Assumptions H2 and H3 in Chan et al. (2014).

Theorem 2-1. *Suppose Assumptions A1-A3 hold. Choose the tuning parameters as* $\lambda_{1,T} = C_1 \frac{b_T}{T}\sqrt{\frac{p}{T\gamma_T}}$, $\lambda_{2,T} = 2C_2\sqrt{\frac{\log T + 2\log p}{T}}$ *and* $\lambda_{3,T} = C_3 \frac{b_T}{T}\sqrt{\frac{\log p}{T\gamma_T}}$ *for some large constants* $C_1, C_2, C_3 > 0$. *Then, as* $T \to +\infty$,

$$\mathbb{P}(|\widehat{\mathcal{A}}_T| \geq m_0) \to 1 \quad \textit{and} \quad \mathbb{P}(d_H(\widehat{\mathcal{A}}_T, \mathcal{A}_T^\star) \leq T\gamma_T) \to 1.$$

Remark 2-2. *In Theorem 2-1, We express the tuning parameters* $\lambda_{1,T}$ *and* $\lambda_{3,T}$ *in different forms. Note that under the setting in Theorem 2-1, the quantities* $\frac{b_T}{T}\sqrt{\frac{p}{T\gamma_T}}$ *and* $Td_T^\star\lambda_{3,T}$ *equal to* $\frac{b_T}{T\gamma_T}\sqrt{\gamma_T}\sqrt{\frac{p}{T}}$ *and* $C_3 b_T\sqrt{d_T^\star}\sqrt{\frac{d_T^\star \log p}{T\gamma_T}}$, *respectively. Further, we have a positive vanishing sequence* $\{\gamma_T\}$ *satisfying* $\limsup \frac{b_T}{T\gamma_T} \leq C < \frac{1}{12}$ *and* $\frac{d_T^\star \log p}{T\gamma_T} \to 0$ *in Assumption H3, which yields to* $\lambda_{1,T} \propto C_1\sqrt{\frac{p}{T}}$ *and* $\lambda_{3,T} = o((Td_T^\star)^{-1})$.

2.2.2 Step 2: Screening

Since the set of estimated break points $\widehat{\mathcal{A}}_T$ is a superset of $\mathcal{A}_T$, 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: $1 = s_0 < s_1 < \cdots < s_m < s_{m+1} = T$. 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 $\mathcal{Z}_{s_1,\ldots,s_m} = \mathrm{diag}(\mathbf{X}_{s_1}, \ldots, \mathbf{X}_{s_{m+1}}) \in \mathbb{R}^{n \times (m+1)p}$, 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

$$\mathcal{Y} = \mathcal{Z}_{s_1,\ldots,s_m}\Theta_{s_1,\ldots,s_m} + \mathcal{X}L + \Xi, \tag{2-5}$$

where $\Xi \stackrel{\text{def}}{=} [\xi_1, \xi_2, \ldots, \xi_T]' \in \mathbb{R}^{T\times p}$ 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 $i = 1, 2, \ldots, m+1$.

$$(\widehat{L}, \widehat{\Theta}_{s_1,\ldots,s_m}) = \underset{L,\Theta_{s_1,\ldots,s_m}}{\arg\min} \sum_{i=1}^{m+1} \frac{1}{s_i - s_{i-1}} \big\|\mathbf{Y}_{s_i} - \mathbf{X}_{s_i}(\theta_{(s_{i-1},s_i)} + L)\big\|_2^2 + \eta_{(s_{i-1},s_i)}\|\theta_{(s_{i-1},s_i)}\|_1 + \eta_L\|L\|_*. \tag{2-6}$$

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

$$L_T(\mathbf{s}; \eta_T) = \Big\|\mathcal{Y} - \mathcal{Z}_{s_1,\ldots,s_m}\widehat{\Theta}_{s_1,\ldots,s_m} - \mathcal{X}\widehat{L}\Big\|_F^2 + \sum_{i=1}^{m+1} \eta_{(s_{i-1},s_i)}\|\widehat{\theta}_{(s_{i-1},s_i)}\|_1 + \eta_L\|\widehat{L}\|_*, \tag{2-7}$$

where $\mathbf{s} = (s_1, \ldots, s_m)$. Then, for a penalty sequence ω_T (which can be selected in accordance to Assumption A4 below), we consider the following *information criterion*

$$\mathrm{IC}(\mathbf{s}; \eta_T) = L_T(\mathbf{s}; \eta_T) + m\omega_T. \tag{2-8}$$

The second step of our strategy selects a subset of initial $\widehat{m}$ change points derived from (2-4) by solving

$$(\widetilde{m}, \widetilde{t}_j; j = 1, 2, \ldots, \widetilde{m}) = \underset{0\le m\le\widehat{m}, \mathbf{s}\in\widehat{\mathcal{A}}_T}{\arg\min} \mathrm{IC}(\mathbf{s}; \eta_T). \tag{2-9}$$

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

(A4) Assume that

$$\frac{m_0 T\gamma_T(d_T^{\star 2} + r^{\star 2})}{\omega_T} \to 0 \text{ and } \frac{\Delta_T}{m_0\omega_T} \to +\infty.$$

(A5) There exists a large positive constant $c > 0$ such that (a) if $|s_i - s_{i-1}| \le T\gamma_T$, then $\eta_{(s_{i-1},s_i)} = c\sqrt{T\gamma_T \log p}$ and $\eta_L = c\sqrt{T\gamma_T p}$; (b) if there exists $t_j^\star$ and $t_{j+1}^\star$ such that $|s_{i-1} - t_j^\star| \le T\gamma_T$ and $|s_i - t_{j+1}^\star| \le T\gamma_T$, then, $\eta_{(s_{i-1},s_i)} = 2\big(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_T^\star \frac{T\gamma_T}{s_i - s_{i-1}}\big)$ and $\eta_L = 2c\sqrt{\frac{p}{T\gamma_T}}$; (c) otherwise, $\eta_{(s_{i-1},s_i)} = 2\big(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_T^\star\big)$ and $\eta_L = 2c\sqrt{\frac{p}{T\gamma_T}}$.

Assumption A4 connects the penalty term ω_T defined in the information criterion to the minimum spacing allowed between break points. Assumption A5 specifies the magnitude

(rate) of the tuning parameters used in the least squares problem given in (2-6). 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. Zhang & Huang (2008)). 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 in Chan et al. (2014), Safikhani & Shojaie (2020), Chan et al. (2017).

Theorem 2-2. *Suppose Assumptions A1-A5 hold. Then, as $T \to +\infty$, the minimizer $(\widetilde{m}, \widetilde{t}_j; j = 1, 2, \ldots, \widetilde{m})$ of (2-9) satisfies*

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

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

$$\mathbb{P}\big(\max_{1 \le j \le m_0} |\widetilde{t}_j - t_j^\star| \le B m_0 T \gamma_T (d_T^{\star 2} + r^{\star 2}) \big) \to 1.$$

Remark 2-3. *For the case of finite m_0, the sequence γ_T can be chosen as $\gamma_T = (rp + d_T^\star \log p)^{1+v/2}/T$ for some small $v > 0$. Assuming that the low-rank component and total degree of sparsity satisfy $d_T^{\star 2} + r^{\star 2} = o\left((rp + d_T^\star \log p)^{v/2}\right)$, then the consistency rate for identifying the relative location of true break points -$t_j^\star/T$- is of the order $(rp + d_T^\star \log p)^{1+v}/T$ in Theorem 2-2. Finally, in this setting, ω_T can be chosen as $(rp + d_T^\star \log p)^{1+2v}$ and the minimum spacing allowed between consecutive break points -Δ_T- must be at least of order $(rp + d_T^\star \log p)^{1+3v}$. Comparing the consistency rates with those in Theorem 3 in Safikhani & Shojaie (2020), 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 2-4. *If $r = 0$ (no low-rank component present in the model), the consistency results are similar to those in Safikhani & Shojaie (2020). Specifically, Theorem 2-2 could be seen as an extension of Theorem 3 in Safikhani & Shojaie (2020). Further, whenever*

$r = 0$, the total number of time series components could be of order $o\left(e^T\right)$, while for $r \geq 1$, we must have $p = o(T)$ since the low-rank component in each transition matrix is potentially dense. This is similar to the stationary (no break points) case discussed in Basu et al. (2019).

2.2.3 Step 3: Consistent Parameter Estimation

The main idea to consistently estimate the model parameters is that Theorem 2-1 and Theorem 2-2 indicate that removing the estimated change points together with an adequate R_T-radius neighborhood around them will also remove the true change points. Hence, the remainder time segments would be stationary. Theorem 2-1 points out that the radius R_T can be as small as $T\gamma_T$, while Theorem 2-2 establishes that this radius should be at least $Bm_0T\gamma_T({d_T^\star}^2 + r^{\star 2})$ for some large constant $B > 0$, in order to drop out redundant change points.

Given the results in Theorem 2-2, 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 2-2, we have

$$\mathbb{P}\big(\max_{1\leq j\leq m_0} |\widetilde{t}_j - t_j^\star| \leq R_T\big) \to 1,$$

as $T \to +\infty$. 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 $r_{j1} = \widetilde{t}_j - R_T - 1$ and $r_{j2} = \widetilde{t}_j + R_T + 1$ for $j = 1, 2, \ldots, m_0$ and let $r_{02} = 1$ and $r_{(m_0+1)1} = T$. 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.

Next, 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

$$\mathcal{Y}_j = \mathcal{X}_j(S_j + L) + \epsilon_j, \tag{2-10}$$

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

$$(\widehat{L}, \widehat{S}_j) = \underset{L \in \Omega, S_j}{\arg\min} \frac{1}{N_j} \|\mathcal{Y}_j - \mathcal{X}_j(S_j + L)\|_F^2 + \rho_j \|S_j\|_1 + \rho_L \|L\|_*. \tag{2-11}$$

Then, the error bound for each estimated segment is:

Theorem 2-3. *Suppose Assumptions A1-A5 hold, m_0 is unknown and $R_T = Bm_0 n \gamma_T (d_T^{\star 2} + r^{\star 2})$. 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 Negahban et al. (2012) . Then, as $n \to +\infty$, the optima $(\widehat{L}, \widehat{S}_j)$ of (2-11) satisfies:*

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

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 $\Delta_T > \delta T$ for some positive constant δ in order to ensure that all N_j's are of the same order T. Then, when considering all estimated segments simultaneously, (2-10) 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 $\mathcal{X}_r = \text{diag}(\mathcal{X}_1, \ldots, \mathcal{X}_{m_0+1})$, 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

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

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

Theorem 2-4. *Suppose Assumptions A1-A5 hold, m_0 is unknown and define $R_T = Bm_0T\gamma_n({d_T^\star}^2 + r^{\star 2})$. Assume that $\Delta_T > \delta T$ for some large positive constant δ, and $\rho_T = C_1\sqrt{\frac{\log N + 2\log p}{N}} + C_2\frac{\tau}{p\gamma}$, $\rho_L = C_1'\sqrt{\frac{p}{N}}$ for some large enough constants $C_1, C_2, C_1' > 0$ and curvature parameter $\tau > 0$ in the restricted strong convexity assumption Negahban et al. (2012) . Then, as $T \to +\infty$, the optimal $(\widehat{L}, \widehat{\mathbf{S}})$ satisfies*

$$\|\widehat{\mathbf{S}} - \mathbf{S}^\star\|_F^2 + (m_0 + 1)\|\widehat{L} - L^\star\|_F^2 = \mathcal{O}\big(\frac{r^\star pm_0 + d_T^\star \log p}{N} + \frac{d_T^\star}{p^2\gamma^2}\big).$$

Remark 2-5. *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 Agarwal et al. (2012), or for stationary data in Basu et al. (2019). Therefore, as seen in the statement of Theorems 2-3 and 2-4, 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 $d_T^\star$, and the information ratio γ and does not vanish even in the presence of infinite samples.*

Remark 2-6. *All optimization problems introduced in our methodology including (2-4), (2-6) and (2-11) are convex and can be solved by proximal gradient methods by combining algorithms developed in Basu et al. (2019) and Safikhani & Shojaie (2020). 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).*

Now, we summarize the main steps for the proposed three-step BFL algorithm in the

following algorithm diagram.

Algorithm 1 Three-step change point detection by Block Fused Lasso (BFL).

1. **Input:** Time series data $\{X_t\}$, $t = 0, 1, \ldots, T$, tuning parameters $\lambda_{1,T}$, $\lambda_{2,T}$, block size b_T, convergent tolerances δ;

2. **Initialize:** Low rank component estimator $\hat{L}^{(0)}$, sparse components estimators $\hat{S}_j^{(0)} = \mathbf{0}_{p\times p}$.

3. **while** $\|\hat{L}^{(k)} - \hat{L}^{(k-1)}\|_F^2 \geq \delta$ **do:**

   - Remove the estimated low rank component $\hat{L}^{(k-1)}$ from the time series data;
   - Partition the time axis into blocks of size b_n and fix the VAR parameters within each block. Then, obtain candidate change points together with estimates of the sparse components for each block: $\hat{S}_1^{(k)}, \ldots, \hat{S}_{\hat{m}+1}^{(k)}$ by solving (2-4);
   - Re-estimate low rank component $\hat{L}^{(k)}$ by (2-4) with fixed sparse components for each block $\hat{S}_j^{(k)}$, for $j = 1, \ldots, \hat{m} + 1$.

4. **Screening:** Remove the low rank estimate $\hat{L}^{(\infty)}$ from the time series data; Use backward elimination algorithm (BEA), similar to Algorithm 3.

5. **Estimation step:** Let $f_j(L, S_j)$ be the objective function formulated in (2-11), then for jth estimated interval I_j:
   while $|f_j(\hat{L}^{(k-1)}, \hat{S}_j^{(k-1)}) - f_j(\hat{L}^{(k)}, \hat{S}_j^{(k)})| \geq \epsilon$ **do:**

   - Remove $\hat{L}^{(k-1)}$ from the time series data;
   - Estimate $\hat{S}_j^{(k)}$ in the interval I_j using modified time series data by previous step;
   - Remove $\hat{S}_j^{(k)}$ from each interval I_j, for $j = 1, 2, \ldots, \hat{m} + 1$;
   - Re-estimate low rank component and obtain $\hat{L}^{(k)}$ by using the time series data without sparse component effects.

6. **Output:** Estimated change points: $s_1, \ldots, s_{\hat{m}}$; estimated model parameters low rank $\hat{L}^{(\infty)}$ and sparse $\hat{S}_j^{(\infty)}$.

2.3 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 (3rd) 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. Figure 2-1 illustrates the 1-off diagonal structure for transition matrices in scenario A-D and F with values $-\gamma\|L^\star\|_\infty$, $\gamma\|L^\star\|_\infty$ and $-\gamma\|L^\star\|_\infty$, respectively, and a randomly sparse structure in scenario E. We set $\gamma = 4$ and the locations of two change points at $t_1^\star = \lfloor n/3 \rfloor$ and $t_2^\star = \lfloor 2n/3 \rfloor$.

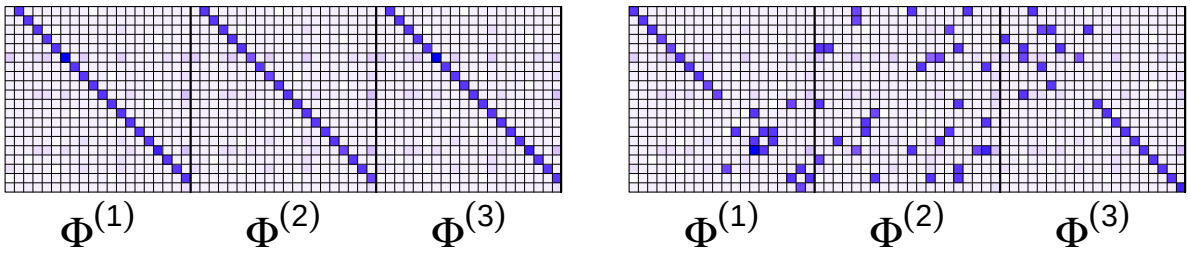


Figure 2-1. True structure of transition matrices: 1-off diagonal and randomly sparse.

2.3.1 Scenarios Examined

In this section, we present different numerical simulation scenarios in the following list, where we examine different sample sizes, dimensions of transition matrix, information ratios, and the limits of the power of change point detection of proposed methodology.

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)* $t_1^\star = \lfloor T/6 \rfloor$, $t_2^\star = \lfloor T/3 \rfloor$ and $t_3^\star = \lfloor 2T/3 \rfloor$; *(b)* $T = 600$ with $t_1^\star = \lfloor T/6 \rfloor$, $t_2^\star = \lfloor T/4 \rfloor$, $t_3^\star = \lfloor T/3 \rfloor$, $t_4^\star = \lfloor 2T/3 \rfloor$ and $t_5^\star = \lfloor 5T/6 \rfloor$. It should be noted that we adopt smaller block sizes $b_T = \sqrt{T}/2$ or $b_T = \sqrt{T}/5$ 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 Figure 2-1 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 $t_1^\star = \lfloor T/3 \rfloor$ and $t_2^\star = \lfloor 2T/3 \rfloor$.

Table 2-1 summarizes all model parameters in the various scenarios discussed above.

Table 2-1. Model parameters under different scenario settings.

Case	p	T	$t_j^\star/T$	$r^\star$	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

2.3.2 Tuning Parameter Selection

There are a number of tuning parameters in the developed three-step strategy: $\lambda_{1,T}$, $\lambda_{2,T}$, $\lambda_{3,T}$, $\boldsymbol{\eta}_T$, 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.

$\lambda_{1,T}$: We use fixed $\lambda_{1,T}$ in accordance to the nature of the application. In most cases, we manually choose $\lambda_{1,T}$ in the range $[\sqrt{\frac{p}{T}}, 10\sqrt{\frac{p}{T}}]$.

$\lambda_{2,T}$: We can select $\lambda_{2,T}$ 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 $\lambda_{2,T}$. The parameters estimated in the first step are used to predict the time series at time points in $\mathcal{T}$. The value of $\lambda_{2,T}$ which minimizes the mean squared prediction error over $\mathcal{T}$ is the cross-validated choice of $\lambda_{2,T}$.

$\lambda_{3,T}$: As previously discussed, the rate for $\lambda_{3,T}$ vanishes fast as T increases. Thus for simplicity, we suggest setting $\lambda_{3,T}$ 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 $\lambda_{1,T}$.

ρ_L: This parameter is suggested to be kept fixed; in practice, it was set in the range $[\sqrt{\frac{p}{T}}, 10\sqrt{\frac{p}{T}}]$.

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

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

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 Safikhani & Shojaie (2020).

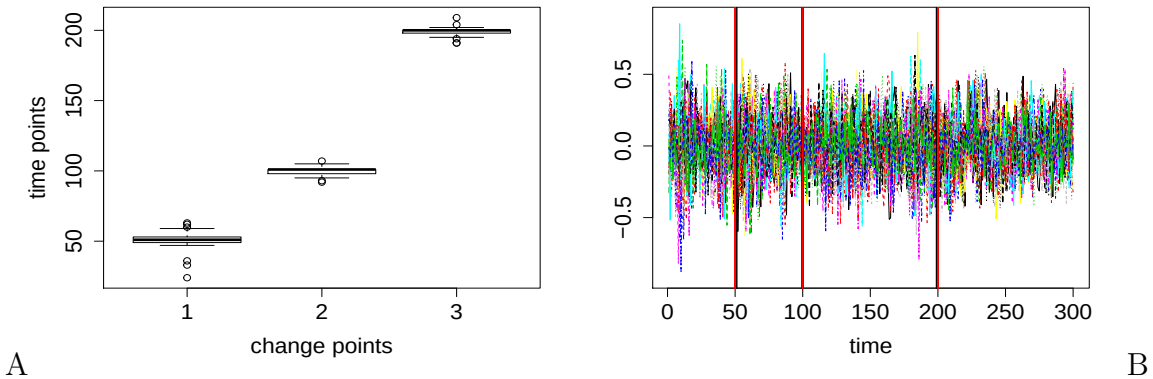
2.3.3 Simulation Results

We evaluate the empirical performance of our algorithm by considering the mean and standard deviation of the estimated change point locations relative to the sample size, i.e. $\widetilde{t}_j/T$, 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*: $[t^\star_{j-1} + \frac{t^\star_j - t^\star_{j-1}}{5}, t^\star_j + \frac{t^\star_{j+1} - t^\star_j}{5}]$. 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.

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

As an illustration of the variability of the estimates, Figure 2-2 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.



A B

Figure 2-2. Box-plot for estimated change points in scenario C.1 and mean results plot for estimated change points (black lines) and true change points (red lines).

The results in Table 2-2 illustrate the performance of change point detection for each of the settings considered in Table 2-1. 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 2-2 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.

Table 2-2. Results for change point selection under parameters settings in Table 3-1.

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

Tables 2-3 and 2-4 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 2-4. 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.

Table 2-3. Performance evaluation of low-rank component under different model settings.

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

Table 2-4. Performance evaluation of sparse components under different model settings.

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

2.4 Real Data Applications

2.4.1 Surveillance Video Data

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

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

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*.

Figure 2-3 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: $\widehat{t}_1 = 115$ 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; $\widehat{t}_{11} = 695$ empty lobby.

In Figure 2-4, 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.

Figure 2-3. The detected change points corresponding to the following times and events.

Figure 2-4. 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.

2.4.2 Stock Data

Next, we employ the proposed detection strategy to identify change points in *weekly* financial stock price data, covering 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 Meucci (2009), or constructed from a large scale diverse portfolio Fama & French (1992, 2015)). 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 Barigozzi et al. (2018). Table 2-5 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 Barigozzi et al. (2018). Figure 2-5 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).

Table 2-5. 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	

The overall (normalized) density of the time varying sparse component based on a 3-factor model is plotted in Figure 2-6. 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

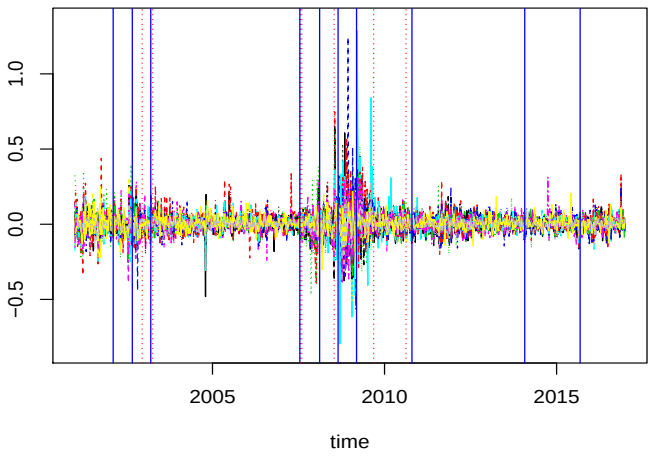


Figure 2-5. Detected change points in the log-returns data during the 2001-2016 period. Red dashed lines are change points selected by Factor Analysis Model Barigozzi et al. (2018); blue solid lines indicate the change points selected by our model.

corresponding values are 1.569 and 1.582, respectively. This result indicates that a rank 3 component is preferable. More specifically, we have the following detailed time periods and

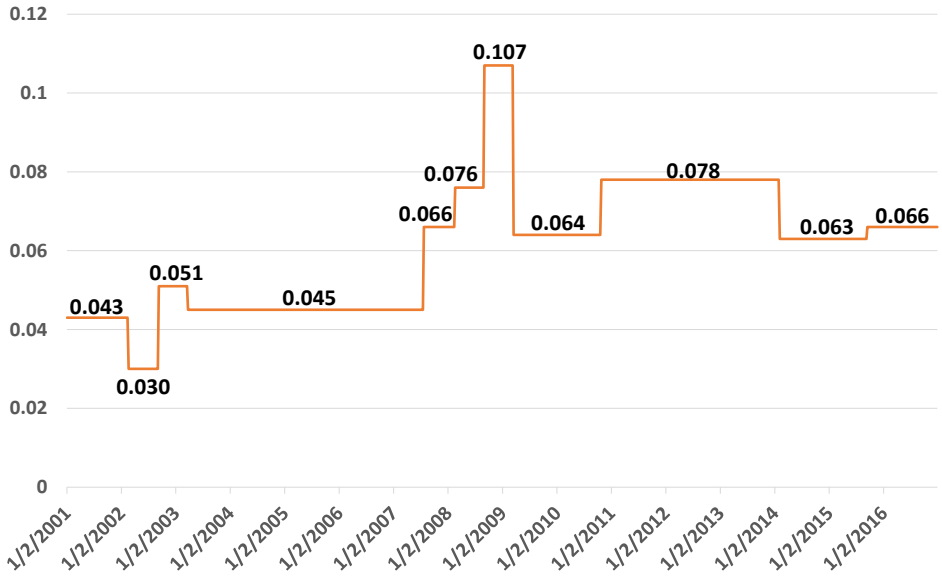


Figure 2-6. Connectivity for each estimated sparse components in different selected time periods.

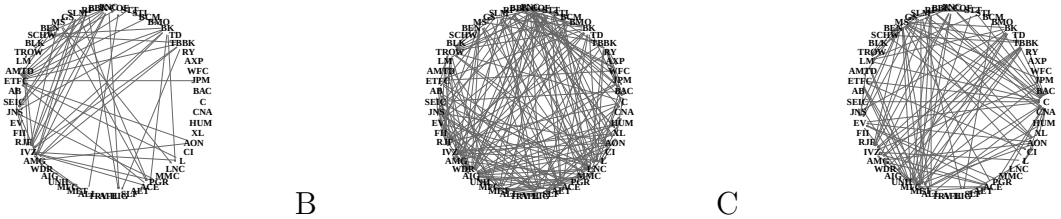
their corresponding connectivity sparsity levels: 1/2/01-2/12/02 period: 4.3%; 2/12/02 -

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%.

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.

Figure 2-7 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.



A B C

Figure 2-7. Estimated connectivity based on selected time periods: structure of the connections in the pre-crisis period with 62 edges (A); structure of the connections among selected companies during the crisis period with 228 edges (B); structure of the connections in the post-crisis period with 114 edges in (C).

2.5 Concluding Remarks

In this chapter, 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 2.2. Note that the developed methodology can be extended to VAR(d) with $d > 1$ in a similar way as discussed in Basu et al. (2019). 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.

2.6 Auxiliary Lemmas and Their Proofs for Chapter 2

Lemma 2-1. *Under the assumptions of Theorem 2-2, for $m < m_0$, there exist constants $c_1, c_2 > 0$ such that:*

$$\mathbb{P}\left(\min_{(s_1,\ldots,s_m)\subset\{1,\ldots,T\}} L_T(s_1,\ldots,s_m;\eta_T) > \sum_{t=1}^{T}\|\epsilon_t\|_2^2 + c_1\Delta_T - c_2 mT\gamma_T({d_T^\star}^2 + r^2)\right) \to 1,$$

where $\Delta_T = \min_{1\le j\le m_0+1} |t_j^\star - t_{j-1}^\star|$.

Proof of Lemma 2-1. Since $m < m_0$, there exists a point t_j such that $|s_i - t_j^\star| > \Delta_T/4$. In order to find a lower bound on the sum of the least squares, we consider three different cases: (a) $|s_i - s_{i-1}| \le T\gamma_T$; (b) there exist two true break points $t_j^\star$, $t_{j+1}^\star$ such that $|s_{i-1} - t_j^\star| \le T\gamma_T$ and $|s_i - t_{j+1}^\star| \le T\gamma_T$; and (c) otherwise. Here we only consider one candidate for each case. Denote the estimated parameter in each of the estimated segments below by $\hat\theta$ and $\hat L$ and we use the similar notation as the proof of Theorem 2, let $\hat\Delta_\theta = \hat\theta - S_{j+1}^\star$ and $\hat\Delta_L = \hat L - L^\star$.

For case (a), consider the case where the interval (s_{i-1}, s_i) is inside a true segment. In other words, suppose there exists j such that $t_j^\star \le s_{i-1} < s_i \le t_{j+1}^\star$, then we have

$$\begin{aligned}\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\hat\theta + \hat L)' X_{t-1}\|_2^2 &= \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 + 2\sum_{t=s_{i-1}}^{s_i-1} X_{t-1}'(\hat\Delta_\theta + \hat\Delta_L)\epsilon_t + \sum_{t=s_{i-1}}^{s_i-1} \|X_{t-1}'(\hat\Delta_\theta + \hat\Delta_L)\|_2^2 \\ &\ge \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 - 2\left|\sum_{t=s_{i-1}}^{s_i-1} X_{t-1}'(\hat\Delta_\theta + \hat\Delta_L)\epsilon_t\right| \\ &\ge \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 - c_1\sqrt{T\gamma_T \log p}\|\hat\Delta_\theta\|_1 - c_2\sqrt{T\gamma_T p}\|\hat\Delta_L\|_*,\end{aligned} \tag{2-12}$$

therefore, given the tuning parameter based on the Assumptions A4 and A5, we have

$$\begin{aligned}&\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\hat\theta + \hat L)' X_{t-1}\|_2^2 + \eta_{(s_{i-1},s_i)}\|\hat\theta\|_1 + \eta_L\|\hat L\|_* \\ &\qquad\ge \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 - c\sqrt{T\gamma_T \log p}\|S_{j+1}^\star\|_1 - c\sqrt{T\gamma_T p}\|L^\star\|_*.\end{aligned} \tag{2-13}$$

Then, we can derive the result by combining all intervals together.

For case (b), we consider the case where $s_{i-1} < t_j^\star$ and $s_i < t_{j+1}^\star$. Similarly, according

to the definition of estimated values, considering the interval $(t_j^\star, s_i)$ we have

$$\begin{aligned}
&\frac{1}{s_i - t_j^\star}\sum_{t=t_j^\star}^{s_i-1}\|X_t - (\hat{\theta}+\hat{L})'X_{t-1}\|_2^2 + \eta_{(s_{i-1},s_i)}\|\hat{\theta}\|_1 + \eta_L\|\hat{L}\|_* \\
&\leq \frac{1}{s_i - t_j^\star}\sum_{t=t_j^\star}^{s_i-1}\|X_t - (S_{j+1}^\star + L^\star)'X_{t-1}\|_2^2 + \eta_{(s_{i-1},s_i)}\|S_{j+1}^\star\|_1 + \eta_L\|L^\star\|_*.
\end{aligned} \tag{2-14}$$

After some algebraic rearrangements and by the optimality of $(\hat{L}, \hat{\Theta}_{s_1,\ldots,s_m})$, we have

$$\begin{aligned}
0 &\leq \frac{1}{s_i - t_j^\star}\sum_{t=t_j^\star}^{s_i-1}\|X_{t-1}'(\hat{\Delta}_\theta + \hat{\Delta}_L)\|_2^2 \\
&\leq \frac{2}{s_i - t_j^\star}\sum_{t=t_j^\star}^{s_i-1}X_{t-1}'(\hat{\Delta}_\theta + \hat{\Delta}_L)\epsilon_t + \eta_{(s_{i-1},s_i)}(\|S_{j+1}^\star\|_1 - \|\hat{\theta}\|_1) + \eta_L(\|L^\star\|_* - \|\hat{L}\|_*) \\
&\leq \left(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_T^\star \frac{T\gamma_T}{s_i - s_{i-1}}\right)\|\hat{\Delta}_\theta\|_1 + \eta_{(s_{i-1},s_i)}(\|S_{j+1}^\star\|_1 - \|\hat{\theta}\|_1) \\
&+ c\sqrt{\frac{p}{T\gamma_T}}\|\hat{\Delta}_L\|_\star + \eta_L(\|L^\star\|_* - \|\hat{L}\|_*) \\
&\leq \frac{1}{2}\eta_{(s_{i-1},s_i)}\|\hat{\Delta}_\theta\|_1 + \eta_{(s_{i-1},s_i)}(\|S_{j+1}^\star\|_1 - \|\hat{\theta}\|_1) + \frac{1}{2}\eta_L\|\hat{\Delta}_L\|_\star + \eta_L(\|L^\star\|_* - \|\hat{L}\|_*).
\end{aligned} \tag{2-15}$$

For estimating $\theta_{(s_{i-1},s_i)}$ and L, we define $\mathcal{Q}$ to be the weighted regularizer $\mathcal{Q}(\theta, L) := \|L\|_* + \frac{\eta_{(s_{i-1},s_i)}}{\eta_L}\|\theta\|_1$. Using the above definition and the decomposition of the nuclear norm and ℓ_1 norm respectively, we can derive that is further bounded by

$$\frac{3}{2}\eta_L\mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}}, \hat{\Delta}_L^A) - \frac{1}{2}\eta_L\mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}^c}, \hat{\Delta}_L^B), \tag{2-16}$$

where the subspaces pairs (A, B) and $(\mathcal{I}, \mathcal{I}^c)$ are detailed introduced in the next section. By the triangle inequality $\mathcal{Q}(\hat{\Delta}_\theta, \hat{\Delta}_L) \leq \mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}}, \hat{\Delta}_L^A) + \mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}^c}, \hat{\Delta}_L^B)$, and substituting (2-16) into (2-15), we obtain

$$\mathcal{Q}(\hat{\Delta}_\theta, \hat{\Delta}_L) \leq 4\mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}}, \hat{\Delta}_L^A).$$

We directly get the following based on Lemma 1 in Agarwal et al. (2012)

$$\text{rank}(\hat{\Delta}_L^A) \leq 2r \quad \text{and} \quad \langle\hat{\Delta}_L^A, \hat{\Delta}_L^B\rangle = 0,$$

which implies $\|\hat{\Delta}_L^A\|_* \leq \sqrt{2r}\|\hat{\Delta}_L^A\|_2 \leq \sqrt{2r}\|\hat{\Delta}_L\|_2$ and $\|\hat{\Delta}_\theta|_{\mathcal{I}}\|_1 \leq \sqrt{d_T^\star}\|\hat{\Delta}_\theta|_{\mathcal{I}}\|_2 \leq \sqrt{d_T^\star}\|\hat{\Delta}_\theta\|_2$. Therefore,

$$\eta_L \mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}}, \hat{\Delta}_L^A) \leq \sqrt{(\eta_{(s_{i-1},s_i)}\sqrt{d_T^\star})^2 + (\eta_L\sqrt{2r})^2}\sqrt{\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2}. \tag{2-17}$$

Moreover, by the RSC condition, it implies that

$$\frac{\tau}{2}(\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2) - \frac{\eta_L}{2}\mathcal{Q}(\hat{\Delta}_\theta, \hat{\Delta}_L) \leq \frac{3}{2}\eta_L \mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}}, \hat{\Delta}_L^A) - \frac{1}{2}\eta_L \mathcal{Q}(\hat{\Delta}_\theta|_{\mathcal{I}^c}, \hat{\Delta}_L^B),$$

where $\tau > 0$ is the RSC constant. After some algebraic rearrangements, we have

$$\frac{\tau}{2}(\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2) \leq 8\sqrt{(\eta_{(s_{i-1},s_i)}\sqrt{d_T^\star})^2 + (\eta_L\sqrt{2r})^2}\sqrt{\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2}.$$

Therefore, we have

$$\sqrt{\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2} \leq \frac{16}{\tau}\sqrt{d_T^\star \eta_{(s_{i-1},s_i)}^2 + 2r\eta_L^2}. \tag{2-18}$$

Similar to case (a), we have

$$\sum_{t=t_j^\star}^{s_i-1} \|X_t - (\hat{\theta}+\hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c'|s_i - t_j^\star| \left(\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2 - \frac{3\eta_L}{2}\mathcal{Q}(\hat{\Delta}_\theta, \hat{\Delta}_L) \right). \tag{2-19}$$

Therefore, according to the upper bound of $\mathcal{Q}(\hat{\Delta}_\theta, \hat{\Delta}_L)$ from (2-18) and (2-19), we have

$$\sum_{t=t_j^\star}^{s_i-1} \|X_t - (\hat{\theta}+\hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 - c''(d_T^\star \log p + r^\star p), \tag{2-20}$$

where the last inequality is derived from assumption H5(b) and c'' is a large enough positive constant.

Now, let's consider the interval $(s_{i-1}, t_j^\star)$. By the conditions in the statement of Lemma, we have $|s_{i-1} - t_j^\star| \leq T\gamma_T$ and the following lower bound holds

$$\sum_{t=s_{i-1}}^{t_j^\star-1} \|X_t - (\hat{\theta}+\hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{t_j^\star-1} \|\epsilon_t\|_2^2 - c_1' {d_T^\star}^2 \sqrt{T\gamma_T \log p} - c_2' {r^\star}^2 \sqrt{T\gamma_T p}. \tag{2-21}$$

Combining the results of (2-20) and (2-21) gives

$$\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 - c_1' d_T^{\star 2} \sqrt{T\gamma_T \log p} - c_2' r^{\star 2} \sqrt{T\gamma_T p}. \tag{2-22}$$

For case (c), consider the case where $s_{i-1} < t_j^\star < s_i$, with $|s_{i-1} - t_j^\star| > \Delta_T/4$ and $|s_i - t_j^\star| > \Delta_T/4$. Note that the RSC condition is not satisfied in this case. Therefore, the convergence of the $\hat{\theta}$ and $\hat{L}$ cannot be verified. Now, similar to case (b), on both intervals $(s_{i-1}, t_j^\star)$ and $(t_j^\star, s_i)$, we have

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{t_j^\star-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \\
&\geq \sum_{t=s_{i-1}}^{t_j^\star-1} \|\epsilon_t\|_2^2 + c|t_j^\star - s_{i-1}| \|(\hat{\theta} - S_j^\star) + \hat{\Delta}_L\|_2^2 - c_1' \sqrt{|t_j^\star - s_{i-1}| \log p} \|\hat{\theta} - S_j^\star\|_1 - c_2' \sqrt{|t_j^\star - s_{i-1}| p} \|\hat{\Delta}_L\|_* \\
&\geq \sum_{t=s_{i-1}}^{t_j^\star-1} \|\epsilon_t\|_2^2 + c|t_j^\star - s_{i-1}| \|(\hat{\theta} - S_j^\star) + \hat{\Delta}_L\|_2^2 - c'|t_j^\star - s_{i-1}| \widetilde{\eta}_L \widetilde{\mathcal{Q}}(\hat{\theta} - S_j^\star, \hat{\Delta}_L),
\end{aligned}$$

where $\widetilde{\eta}_L = \sqrt{\frac{p}{|t_j^\star - s_{i-1}|}}$ and $\widetilde{\eta}_{(s_{i-1}, s_i)} = \sqrt{\frac{\log p}{|t_j^\star - s_{i-1}|}}$ and $\widetilde{\mathcal{Q}}$ is the weighted regularizer with respect to $\widetilde{\eta}_L$ and $\widetilde{\eta}_{(s_{i-1}, s_i)}$. Since we know the upper bound of $\widetilde{\mathcal{Q}}(\hat{\theta} - S_j^\star, \hat{\Delta}_L)$, we obtain a lower bound of $\|(\hat{\theta} - S_j^\star) + \hat{\Delta}_L\|_2^2$ is given by

$$\left(\|\hat{\theta} - S_j^\star\|_2^2 + \|\hat{\Delta}_L\|_2^2 \right) - \frac{\widetilde{\eta}_L}{2} \widetilde{\mathcal{Q}}(\hat{\theta} - S_j^\star, \hat{\Delta}_L).$$

After some substitutions of the corresponding terms, we obtain

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{t_j^\star-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \\
&\geq \sum_{t=s_{i-1}}^{t_j^\star-1} \|\epsilon_t\|_2^2 \\
&+ c|t_j^\star - s_{i-1}| \sqrt{\|\hat{\theta} - S_j^\star\|_2^2 + \|\hat{\Delta}_L\|_2^2} \left(\sqrt{\|\hat{\theta} - S_j^\star\|_2^2 + \|\hat{\Delta}_L\|_2^2} - 2(1 + \frac{c'}{c}) \sqrt{2r^\star \widetilde{\eta}_L^2 + d_T^\star \widetilde{\eta}_{(s_{i-1}, s_i)}^2} \right).
\end{aligned}$$

Similarly, consider the interval $(t_j^\star, s_i)$. We have

$$\begin{aligned}
&\sum_{t=t_j^\star}^{s_i-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \\
&\geq \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 \\
&+ c|s_i - t_j^\star| \sqrt{\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2} \left(\sqrt{\|\hat{\Delta}_\theta\|_2^2 + \|\hat{\Delta}_L\|_2^2} - 2(1 + \frac{c'}{c}) \sqrt{2r^\star \widetilde{\eta}_L^2 + d_T^\star \widetilde{\eta}_{(s_{i-1}, s_i)}^2} \right).
\end{aligned}$$

From Assumptions A3 and A5(c), we know that $\|S_{j+1}^\star - S_j^\star\|_2 \geq v > 0$, either $\|S_{j+1}^\star - \hat{\theta}\|_2 \geq v/4$ or $\|S_j^\star - \hat{\theta}\|_2 \geq v/4$. Assume that we have $\|S_j^\star - \hat{\theta}\|_2 \geq v/4$, then in the interval $(s_{i-1}, t_j^\star)$, for some large positive constants C_1 and C_1', we have

$$\sum_{t=s_{i-1}}^{t_j^\star - 1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{t_j^\star - 1} \|\epsilon_t\|_2^2 + C_1 \Delta_T, \tag{2-23}$$

and for the interval $(t_j^\star, s_i)$, we have

$$\sum_{t=t_j}^{s_i-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=t_j}^{s_i-1} \|\epsilon_t\|_2^2 - C_1'(d_T^\star \log p + r^\star p). \tag{2-24}$$

Aggregating (2-23) and (2-24) implies that

$$\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 + C_1 \Delta_T - C_1'(d_T^\star \log p + r^\star p). \tag{2-25}$$

The other situation in which we have $|s_{i-1} - t_j^\star| > T\gamma_T$ and $|t_j^\star - s_i| > T\gamma_T$, using similar augments as above, we have the following lower bound

$$\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\hat{\theta} + \hat{L})' X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 - C_1' T\gamma_T \left({d_T^\star}^2 + {r^\star}^2 \right).$$

Combining all of these cases together leads to the result. □

2.7 Technical Proofs for Main Theorems in Chapter 2

Proof of Theorem 2-1. First, we focus on the second part. Suppose for some $j = 1, 2, \ldots, m_0$, $|\widehat{t}_j - t_j^\star| \geq T\gamma_T$. Then, there exists a true break point $t_{j_0}^\star$ which is isolated from all the estimated points, i.e., $\min_{1 \leq j \leq m_0} |\widehat{t}_j - t_{j_0}^\star| > T\gamma_T$. The idea is to show the

estimated AR parameter $\widehat{S}_j$ in the interval $[t^\star_{j_0-1} \vee \widehat{t}_j, t^\star_{j_0+1} \wedge \widehat{t}_{j+1}]$ converges in ℓ_2 to both $S^\star_j$ and $S^\star_{j+1}$ which contradicts assumption H3.

Due to the definition of $(\widehat{L}, \widehat{\Theta})$ in (2-4), the value of the function defined in (2-4) is minimized exactly at $(\widehat{L}, \widehat{\Theta})$. Denote the closest r_i to the right side of $t^\star_{j_0-1}$ by s_{j_0-1} and the closest r_i to the left side of $t^\star_{j_0}$ 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 $\psi_{\widehat{t}_j} = S^\star_{j_0} - \widehat{S}_j$ and $\psi_{s_{j_0}} = \widehat{S}_{j+1} - S^\star_{j_0}$ 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, $\widehat{\theta}_{s_{j_0-1} \vee \widehat{t}_j} = \widehat{S}_{j+1} - \widehat{S}_j$. Denoting $\Psi = [\psi'_1, \psi'_2, \ldots, \psi'_{k_n}]' \in \mathbb{R}^{pk_T \times p}$, we obtain

$$\frac{1}{n}\|\mathcal{Y} - \mathcal{X}\widehat{L} - \mathcal{Z}\widehat{\Theta}\|_2^2 + \lambda_{1,n}\|\widehat{L}\|_* + \lambda_{2,n}\|\widehat{\Theta}\|_1 + \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\widehat{\theta}_j\right\|_1 \\ \leq \frac{1}{n}\|\mathcal{Y} - \mathcal{X}L^\star - \mathcal{Z}\Psi\|_2^2 + \lambda_{1,n}\|L^\star\|_* + \lambda_{2,n}\|\Psi\|_1 + \lambda_{3,n}\sum_{l=1}^{k_n}\left\|\sum_{j=1}^{l}\psi_j\right\|_1. \tag{2-26}$$

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 (Basu et al. 2019). 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 (2-26) leads to

$$\frac{1}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\|\widetilde{\mathcal{X}}(\widehat{\Delta}_L + \widehat{\Delta}_S)\|_2^2 \\ \leq \frac{2\langle \widehat{\Delta}_L + \widehat{\Delta}_S, \widetilde{\mathcal{X}}'\widetilde{\mathcal{E}}\rangle}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} + \frac{T\lambda_{1,T}\big(\|L^\star\|_* - \|\widehat{L}\|_*\big)}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \\ + \frac{T\lambda_{2,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\Big(\|S^\star_{j_0} - \widehat{S}_{j+1}\|_1 + \|S^\star_{j_0} - \widehat{S}_j\|_1 - \|\widehat{S}_{j+1} - \widehat{S}_j\|_1\Big) + \frac{n\lambda_{3,T}}{b_T}\Big(\|S^\star_{j_0}\|_1 - \|\widehat{S}_{j+1}\|_1\Big)$$

$$\leq \frac{2}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \langle \widehat{\Delta}_L + \widehat{\Delta}_S, \widetilde{\mathcal{X}}' \widetilde{\mathcal{E}} \rangle + \frac{n\lambda_{1,n}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \big(\|\widehat{\Delta}_L^A\|_* - \|\widehat{\Delta}_L^B\|_* \big) \tag{2-27}$$
$$+ \frac{2T\lambda_{2,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \|\widehat{\Delta}_S\|_1 + \frac{T\lambda_{3,T}}{b_T} \big(\|\widehat{\Delta}_S\|_{1,\mathcal{I}} - \|\widehat{\Delta}_S\|_{1,\mathcal{I}^c} \big) + \frac{2T\lambda_{1,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \sum_{j=r+1}^{p} \sigma_j(L^\star),$$

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 Agarwal et al. (2012) 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 (2-27) implies the following inequality

$$\langle \widehat{\Delta}_L + \widehat{\Delta}_S, \widetilde{\mathcal{X}}' \widetilde{\mathcal{E}} \rangle \leq \|\widetilde{\mathcal{X}}' \widetilde{\mathcal{E}}\|_{\text{op}} \big(\|\widehat{\Delta}_L^A\|_* + \|\widehat{\Delta}_L^B\|_* \big) + \|\widetilde{\mathcal{X}}' \widetilde{\mathcal{E}}\|_\infty \big(\|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \|\widehat{\Delta}_S\|_{1,\mathcal{I}^c} \big). \tag{2-28}$$

Substituting (2-28) into (2-27) and considering the conditions for $\lambda_{1,T}$, $\lambda_{2,T}$ and $\lambda_{3,T}$, we have

$$\begin{aligned}
&\frac{1}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \|\widetilde{\mathcal{X}}(\widehat{\Delta}_L + \widehat{\Delta}_S)\|_2^2 \\
&\leq \frac{3T\lambda_{1,T}}{2b_T} \|\widehat{\Delta}_L^A\|_* + \frac{3T\lambda_{3,T}}{2b_T} \|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{2T\lambda_{3,T}}{b_T} \|S_{j_0}^\star\|_{1,\mathcal{I}^c} \\
&+ \Big(\frac{2T\lambda_{2,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} + C\sqrt{\frac{\log p}{T\gamma_T}} \Big) \|\widehat{\Delta}_S\|_1 + \frac{2T\lambda_{1,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \sum_{j=r+1}^{p} \sigma_j(L^\star) \\
&\leq \frac{3T\lambda_{1,T}}{2b_T} \|\widehat{\Delta}_L^A\|_* + \frac{3T\lambda_{3,T}}{2b_T} \|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{2T\lambda_{3,T}}{b_T} \|S_{j_0}^\star\|_{1,\mathcal{I}^c} \\
&+ \frac{n\lambda_{3,T}}{2b_T} \|\widehat{\Delta}_S\|_1 + \frac{2T\lambda_{1,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \sum_{j=r+1}^{p} \sigma_j(L^\star) \\
&= \frac{3T\lambda_{1,T}}{2b_T} \|\widehat{\Delta}_L^A\|_* + \frac{3T\lambda_{3,T}}{2b_T} \|\widehat{\Delta}_S\|_{1,\mathcal{I}} + \frac{T\lambda_{3,T}}{2b_T} \|\widehat{\Delta}_S\|_1.
\end{aligned} \tag{2-29}$$

The first inequality holds with high probability converging to 1 due to part (a) in Lemma 2 and the fact that $s_{j_0} - s_{j_0-1} \vee \widehat{t}_j \geq \frac{1}{2} T\gamma_T$ and $b_T \leq \frac{1}{4} T\gamma_T$ by assumption H3. The second inequality is based on triangle inequality and the selection for $\lambda_{2,T}$ and $\lambda_{3,T}$. 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 Basu et al. (2019),

there exists a constant $\tau > 0$ such that

$$\frac{1}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} \|\widetilde{\mathcal{X}}(\widehat{\Delta}_L + \widehat{\Delta}_S)\|_2^2 \geq \frac{\tau}{2}\big(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2\big) - \frac{T\lambda_{3,T}}{2b_T}\|\widehat{\Delta}_S\|_1. \tag{2-30}$$

Inserting the inequality (2-30) into (2-29), we have

$$\begin{aligned}\frac{\tau}{2}\big(\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2\big) &\leq \frac{3T\lambda_{1,T}}{2b_T}\|\widehat{\Delta}_L^A\|_* + \frac{5T\lambda_{3,T}}{2b_T}\|\widehat{\Delta}_S\|_1 \\ &\leq \sqrt{\big(\frac{3T\lambda_{1,T}}{2b_T}\sqrt{2r^\star}\big)^2 + \big(\frac{5T\lambda_{3,T}}{2b_T}\sqrt{d_T^\star}\big)^2}\sqrt{\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2}.\end{aligned}$$

Further, combining with our tuning parameters assumption, we obtain

$$\|\widehat{\Delta}_L\|_2^2 + \|\widehat{\Delta}_S\|_2^2 \leq \frac{4}{\tau^2}\left(\frac{9C_1^2}{2}\frac{r^\star p}{T\gamma_T} + \frac{25C_3^2}{4}\frac{d_T^\star \log p}{T\gamma_T}\right), \tag{2-31}$$

it implies that

$$\|\widehat{L} - L^\star\|_2^2 + \|S_{j_0}^\star - \widehat{S}_{j+1}\|_2^2 = o_p\left(\frac{r^\star p + d_T^\star \log p}{T\gamma_T}\right), \tag{2-32}$$

which indicates that $\|\widehat{L} - L^\star\|_2^2 + \|S_{j_0}^\star - \widehat{S}_{j+1}\|_2^2$ 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 $\|\widehat{L} - L^\star\|_2^2 + \|S_{j_0+1}^\star - \widehat{S}_{j+1}\|_2^2$ converges to zeros as well, which leads to that $\|S_{j_0}^\star - \widehat{S}_{j+1}\|_2^2 - \|S_{j_0}^\star - \widehat{S}_j\|_2^2$ converges to zero as well, and this implies to a contradiction to the first part of Assumption A3. Therefore, we proved the second part of the theorem.

The first part can be proved as follows. We assume that $|\widehat{\mathcal{A}}_T| < m_0$, 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 $\|S_{j+1}^\star - S_j^\star\|_2$ converges to zero and therefore contradicts with Assumption A3. □

Proof of Theorem 2-2. 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 2-1 that there exist points $\widehat{t}_j \in \widehat{\mathcal{A}}_T$

satisfying that $\max_{1\le j\le m_0}|\widehat{t}_j - t_j^\star| \le T\gamma_T$. According to the arguments in Lemma 2-1, we get that there exists a constant $K > 0$ such that

$$L_T(\widehat{t}_1,\ldots,\widehat{t}_{m_0};\eta_T) \le \sum_{t=1}^{T}\|\epsilon_t\|_2^2 + Km_0T\gamma_T({d_T^\star}^2 + r^{\star 2}). \tag{2-33}$$

To prove (2-33), we only need to consider one of the estimated segments. Suppose $s_{i-1} < t_j^\star < s_i$ with $|t_j^\star - s_{i-1}| \le T\gamma_T$. 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 $\widehat{\Delta}_\theta = \widehat{\theta} - S_{j+1}^\star$. Then, similar to the proof of Lemma 3 case (b), we have

$$\begin{aligned}\sum_{t=t_j^\star}^{s_i-1}\|X_t - (\widehat{\theta}+\widehat{L})'X_{t-1}\|_2^2 &\le \sum_{t=t_j^\star}^{s_i-1}\|\epsilon_t\|_2^2 + c_3|s_i - t_j^\star|\|\widehat{\Delta}_\theta + \widehat{\Delta}_L\|_2^2 \\ &\quad + c'\left(\sqrt{|s_i - t_j^\star|\log p}\|\widehat{\Delta}_\theta\|_1 + \sqrt{|s_i - t_j^\star|p}\|\widehat{\Delta}_L\|_*\right) \\ &\equiv \sum_{t=t_j^\star}^{s_i-1}\|\epsilon_t\|_2^2 + J_1 + J_2.\end{aligned} \tag{2-34}$$

Now, according to the convergence rate of the error in Lemma 2-1 case (b), we obtain

$$J_1 \le c_3|s_i - t_j^\star|\big(\|\widehat{\Delta}_\theta\|_2^2 + \|\widehat{\Delta}_L\|_2^2\big) \le \mathcal{O}_p\big(T\gamma_T({d_T^\star}^2 + r^{\star 2})\big), \tag{2-35}$$

and

$$J_2 = c'|s_i - t_j^\star|\left(\sqrt{\frac{\log p}{s_i - t_j^\star}}\|\widehat{\Delta}_\theta\|_1 + \sqrt{\frac{p}{s_i - t_j^\star}}\|\widehat{\Delta}_L\|_*\right) \le \mathcal{O}_p\big(T\gamma_T({d_T^\star}^2 + r^{\star 2})\big). \tag{2-36}$$

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

$$\begin{aligned}&\sum_{t=s_{i-1}}^{t_j^\star-1}\|X_t - (\widehat{\theta}+\widehat{L})'X_{t-1}\|_2^2 \\ &\le \sum_{t=s_{i-1}}^{t_j^\star-1}\|\epsilon_t\|_2^2 + c_3|t_j - s_{i-1}|\|(\widehat{\theta} - S_j^\star) + \widehat{\Delta}_L\|_2^2 + c'\left(\sqrt{|t_j^\star - s_{i-1}|\log p}\|\widehat{\theta} - S_j^\star\|_1 + \sqrt{|t_j^\star - s_{i-1}|p}\|\widehat{\Delta}_L\|_*\right) \\ &\le \sum_{t=s_{i-1}}^{t_j^\star-1}\|\epsilon_t\|_2^2 + 2c_3|t_j^\star - s_{i-1}|\big(\|\widehat{\Delta}_\theta + \widehat{\Delta}_L\|_2^2 + \|(S_{j+1}^\star - S_j^\star) + \widehat{\Delta}_L\|_2^2\big)\end{aligned} \tag{2-37}$$

$$+ c'\big(\sqrt{|t_j^\star - s_{i-1}|\log p}\big(\|\widehat{\Delta}_\theta\|_1 + \|S_{j+1}^\star - S_j^\star\|_1\big) + \sqrt{|t_j^\star - s_{i-1}|p}\|\widehat{\Delta}_L\|_*\big)$$
$$\leq \sum_{t=s_{i-1}}^{t_j^\star - 1} \|\epsilon_t\|_2^2 + \mathcal{O}_p\left(T\gamma_T(d_T^{\star\,2} + r^{\star 2})\right)$$

Therefore, combining (2-34) to (2-37) yields

$$\sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\widehat{\theta}+\widehat{L})' X_{t-1}\|_2^2 + \eta_{(s_{i-1},s_i)}\|\widehat{\theta}\|_1 + \frac{s_i - s_{i-1}}{T}\eta_L\|\widehat{L}\|_* = \sum_{t=s_{i-1}}^{s_i-1} \|\epsilon_t\|_2^2 + \mathcal{O}_p\left(T\gamma_T(d_T^{\star\,2} + r^{\star 2})\right). \tag{2-38}$$

Taking the union of all $m_0 + 1$ estimated intervals leads to the result (2-33).

Applying Lemma 2-1 and noting that under the conditions specified in Assumption A4, we obtain

$$\begin{aligned}
\mathrm{IC}(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}) &= L_T(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}; \eta_T) + \widetilde{m}\omega_T \\
&\geq L_T(\widehat{t}_1, \ldots, \widehat{t}_{m_0}; \eta_T) + m_0\omega_n + c_1\Delta_T - c_2 m_0 T\gamma_n(d_T^{\star\,2} + r^{\star 2}) - (m_0 - \widetilde{m})\omega_T \\
&\geq L_T(\widehat{t}_1, \ldots, \widehat{t}_{m_0}; \eta_T) + m_0\omega_T,
\end{aligned} \tag{2-39}$$

which leads to the proof of case (a).

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

$$L_T(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}; \eta_T) \geq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 - c_2\widetilde{m}T\gamma_T(d_T^{\star\,2} + r^{\star 2}). \tag{2-40}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T} \|\epsilon_t\|_2^2 - c_2\widetilde{m}T\gamma_T(d_T^{\star\,2} + r^{\star 2}) + \widetilde{m}\omega_T \\
&\leq \mathrm{IC}(\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}) \leq \mathrm{IC}(\widehat{t}_1, \ldots, \widehat{t}_{m_0}) \leq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + Km_0 T\gamma_T(d_T^{\star\,2} + r^{\star 2}) + m_0\omega_T,
\end{aligned} \tag{2-41}$$

which implies that

$$(\widetilde{m} - m_0)\omega_T \leq (Km_0 + c_2\widetilde{m})T\gamma_T(d_T^{\star\,2} + r^{\star 2}),$$

which contradicts Assumption A4. Now we proved the first part of Theorem 2-2.

For the second part, we let $B = 2K/c$, if there exists a true change point $t_j^\star$ such that

$\min_{1\le j\le m_0}|\widetilde{t}_j - t_j^\star| \ge Bm_0T\gamma_T({d_T^\star}^2 + r^{\star 2})$, then by similar arguments as in Lemma 2-1, we have

$$\sum_{t=1}^{T}\|\epsilon_t\|_2^2 + cBm_0T\gamma_T({d_T^\star}^2 + r^{\star 2}) \le \sum_{t=1}^{T}\|\epsilon_t\|_2^2 + Km_0T\gamma_T({d_T^\star}^2 + r^{\star 2}),$$

which contradicts to $B = 2K/c$. Therefore, we complete the proof. □

Proof of Theorem 2-3. It follows along the lines of the proof of Proposition 4 in Basu et al. (2019). 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 $\left\|\mathcal{X}_j'\mathcal{E}_j/N_j\right\|_\infty$. These two conditions can be verified by Lemma 6 directly. Therefore, we can derive the following result

$$\begin{aligned}\frac{1}{N_j}\|\mathcal{Y}_j - \mathcal{X}_j(\widehat{L}+\widehat{S}_j)\|_F^2 + \rho_j\|\widehat{S}_j\|_1 + \rho_L\|\widehat{L}\|_* \\ \le \frac{1}{N_j}\|\mathcal{Y}_j - \mathcal{X}_j(L^\star + S_j^\star)\|_F^2 + \rho_j\|S_j^\star\|_1 + \rho_L\|L^\star\|_*,\end{aligned} \tag{2-42}$$

we define the same weighted regularizer as in Lemma 2-1 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

$$\frac{1}{N_j}\|\mathcal{X}_j(\widehat{\Delta}_L + \widehat{\Delta}_{S_j})\|_F^2 \le \frac{3}{2}\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}|_{\mathcal{I}_j}, \widehat{\Delta}_L^A) - \frac{1}{2}\rho_L\mathcal{Q}(\widehat{\Delta}_{S_j}|_{\mathcal{I}_j^c}, \widehat{\Delta}_L^B). \tag{2-43}$$

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

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

substituting the inequality above in (2-43), then we have

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

Therefore, we obtain

$$\|\widehat{\Delta}_L\|_F^2 + \|\widehat{\Delta}_{S_j}\|_F^2 \le \frac{16}{\tau^2}(2r^\star\rho_L^2 + d_j^\star\rho_j^2). \tag{2-44}$$

Combining the choices for the tuning parameters specified in Theorem 2-3 and (2-44), we

can obtain the posited result. □

CHAPTER 3
MULTIPLE CHANGE POINT DETECTION IN HIGH DIMENSIONAL REDUCED RANK VAR MODELS

In this chapter, we generalize the structure of high dimensional VAR models to both time-varying low rank and sparse components. We start by investigating the single change point scenario, and formulate the model with an objective function, together with providing a detection procedure based on exhaustive search, and establish theoretical properties for the change point and the model parameter estimates. Then, we discuss the case of multiple change points, and we introduce a two-step algorithm based on rolling window mechanism and establish consistency of the obtained estimates for the change points and model parameters. To reduce computational complexity, we also provide a weakly sparse surrogate model and establish that under certain regularity conditions on the structure of the transition matrices of the reduced rank model, the estimated change points from the surrogate model are consistent ones of data generated by the former. At last, we present a number of numerical experiments to evaluate the performance of the single and multiple change points detection procedures. Two real data sets are analyzed using the proposed detection procedures as well.

3.1 Single Change Point Detection

3.1.1 Model Formulation

We start by introducing a piece-wise stationary structured VAR(1) time series model with a single change point. Suppose that there is a p-dimensional time series $\{X_t\}$ which is observed at $T+1$ points: $t = 0, 1, \ldots, T+1$. Further, there exists a single change point at $0 < \tau^\star < T$, such that the observed time series can be modeled in accordance to the following two models in the intervals $[0, \tau^\star)$ and $[\tau^\star, T)$, respectively:

$$\begin{aligned} X_t &= A_1^\star X_{t-1} + \epsilon_t^1, \quad t = 1, 2, \ldots, \tau^\star, \\ X_t &= A_2^\star X_{t-1} + \epsilon_t^2, \quad t = \tau^\star + 1, \ldots, T, \end{aligned} \tag{3-1}$$

where $X_t \in \mathbb{R}^p$ is a vector of observed time series at time t, $A_1^\star$, $A_2^\star$ are $p \times p$ transition matrices for the corresponding models in the two time intervals respectively, and ϵ_t^1, ϵ_t^2 are p-dimensional independent and identical error processes drawn from Gaussian distribution

with mean zero and covariance matrix $\sigma^2 I$ for some fixed constant $\sigma > 0$. It's further assumed that the transition matrices comprises of two time-varying components, a low rank component and a sparse component:

$$A_1^\star = L_1^\star + S_1^\star \quad \text{and} \quad A_2^\star = L_2^\star + S_2^\star, \tag{3-2}$$

where the rank of low rank components and the sparsity level of sparse components are denoted by $\text{rank}(L_j^\star) = r_j^\star$, $d_j^\star = \|S_j^\star\|_0$ for $j = 1, 2$, respectively, and satisfy $r_j^\star \ll p$ and $d_j^\star \ll p^2$.

3.1.2 Detection Procedure

Consider $\{X_t\}$ is a sequence of observations generated from a VAR(1) model proposed in (3-1) with the structured transition matrices posited in (3-2). Then for any fixed time point $\tau \in \{1, 2, \ldots, T\}$, we construct objective functions for estimating the model parameters in the intervals $[1, \tau)$ and $[\tau, T)$, respectively:

$$\ell(L_1, S_1 | \mathbf{X}^{[1,\tau)}) \stackrel{\text{def}}{=} \frac{1}{\tau - 1} \sum_{t=1}^{\tau-1} \|X_t - (L_1 + S_1)X_{t-1}\|_2^2 + \lambda_1 \|S_1\|_1 + \mu_1 \|L_1\|_*, \tag{3-3}$$

and

$$\ell(L_2, S_2; \mathbf{X}^{[\tau:T)}) \stackrel{\text{def}}{=} \frac{1}{T - \tau} \sum_{t=\tau}^{T-1} \|X_t - (L_2 + S_2)X_{t-1}\|_2^2 + \lambda_2 \|S_2\|_1 + \mu_2 \|L_2\|_*, \tag{3-4}$$

where $\mathbf{X}^{[b:e)}$ denotes the data $\{X_t\}$ from time points b to e, and the non-negative tuning parameters λ_1, λ_2, μ_1, and μ_2 control the regularization of the sparse and the low-rank components in the corresponding transition matrices.

Next, we introduce the objective function with respect to the time point τ:

$$\ell(\tau | L_1, S_1, L_2, S_2) \stackrel{\text{def}}{=} \frac{1}{T - 1} \left(\sum_{t=1}^{\tau-1} \|X_t - (L_1 + S_1)X_{t-1}\|_2^2 + \sum_{t=\tau}^{T-1} \|X_t - (L_2 + S_2)X_{t-1}\|_2^2 \right),$$

then the estimator $\widehat{\tau}$ of the change point $\tau^\star$ is derived by the following optimization problem:

$$\widehat{\tau} \stackrel{\text{def}}{=} \arg\min_{\tau \in \mathcal{T}} \ell(\tau; \widehat{L}_{1,\tau}, \widehat{L}_{2,\tau}, \widehat{S}_{1,\tau}, \widehat{S}_{2,\tau}), \tag{3-5}$$

where the search domain $\mathcal{T} \subset \{1, 2, \ldots, T\}$, for each $\tau \in \mathcal{T}$, the estimators $\widehat{L}_{1,\tau}$, $\widehat{L}_{2,\tau}$, $\widehat{S}_{1,\tau}$, $\widehat{S}_{2,\tau}$ are derived by minimizing the objective functions in (3-3) and (3-4) with tuning parameters $\mu_{1,\tau}$, $\mu_{2,\tau}$, $\lambda_{1,\tau}$, and $\lambda_{2,\tau}$, respectively.

We describe the key steps in estimating the change point $\tau^\star$ as well as the model parameters in the following Algorithm 2.

Algorithm 2 Single Change Point Detection via Exhaustive Search

1. **Input:** Time series data $\{X_t\}$, $t = 0, 1, \ldots, n$; search domain $\mathcal{T} \subset \{1, 2, \ldots, T\}$;
2. **while** $\tau \in \mathcal{T}$ **do:**
   Estimate the low rank and sparse components on the sub-intervals $[1, \tau)$ and $[\tau, T)$, respectively:

$$(\widehat{L}_{1,\tau}, \widehat{S}_{1,\tau}) := \underset{\substack{L_1 \in \Omega \\ L_1, S_1 \in \mathbb{R}^{p\times p}}}{\arg\min} \left\{ \frac{1}{\tau - 1} \sum_{t=1}^{\tau-1} \|X_t - (L_1 + S_1)X_{t-1}\|_2^2 + \lambda_{1,\tau}\|S_1\|_1 + \mu_{1,\tau}\|L_1\|_* \right\},$$

$$(\widehat{L}_{2,\tau}, \widehat{S}_{2,\tau}) := \underset{\substack{L_2 \in \Omega \\ L_2, S_2 \in \mathbb{R}^{p\times p}}}{\arg\min} \left\{ \frac{1}{T - \tau} \sum_{t=\tau}^{T-1} \|X_t - (L_2 + S_2)X_{t-1}\|_2^2 + \lambda_{2,\tau}\|S_2\|_1 + \mu_{2,\tau}\|L_2\|_* \right\},$$

   Estimate the change point $\widetilde{\tau}$:

$$\widehat{\tau} := \underset{\tau \in \mathcal{T}}{\arg\min} \left\{ \frac{1}{T - 1} \left(\sum_{t=1}^{\tau-1} \|X_t - (\widehat{S}_{1,\tau} + \widehat{L}_{1,\tau})X_{t-1}\|_2^2 + \sum_{t=\tau}^{T-1} \|X_t - (\widehat{S}_{2,\tau} + \widehat{L}_{2,\tau})X_{t-1}\|_2^2 \right) \right\}.$$

   Updating the time point τ by $\tau + 1$
3. **Output:** The estimated change point $\widehat{\tau}$ and model parameters $\widehat{L}_{j,\widehat{\tau}}$, $\widehat{S}_{j,\widehat{\tau}}$, $j = 1, 2$.

3.1.3 Theoretical Properties

In this section, we aim to address the theoretical properties for the single change point detection. Due to the decomposition of the transition matrices posited in (3-2), the estimation suffers an identifiability issue. Specifically, when the low rank component $L_j^\star$ is sparse and the corresponding sparse component $S_j^\star$ is hided into the low rank component. In such a setting, estimating the model parameter as well as the change point becomes impossible. To resolve this issue, the key idea is to restrict the *spikiness* of the low rank component, so that it can be disinguishable from the sparse component. In Agarwal et al.

(2012), the author proposed a constraint space Ω defined as follows:

$$\Omega = \left\{ L_j^\star \in \mathbb{R}^{p\times p} : \|L_j^\star\|_\infty \leq \frac{\alpha_L}{p} \right\}, \quad j = 1, 2,$$

where the universal parameter α_L defines the *radius of identifiability* that controls the degree of separating the sparse component from the low rank one. Note that a larger α_L allows the low rank components absorb most of the signal, thus making it harder to identify the sparse component, and vice versa.

Now, we introduce an important quantity for future developments, the *information ratio* which measures the relative strength of the maximum signal in the transition matrices $A_j^\star$'s generated by these two components, defined as:

$$\gamma_j \stackrel{\text{def}}{=} \frac{\|L_j^\star\|_\infty}{\|S_j^\star\|_\infty}, \quad j = 1, 2.$$

Remark 3-1. *Based on the definition of the information ratio, some algebra provides guidance on the identifiability conditions that need to be imposed on the transition matrices $A_j^\star$ and their constituent parts. Specifically, for the low rank component we obtain:*

$$\|A_2^\star - A_1^\star\|_2 \geq v_L - \frac{\alpha_L(\gamma_1 + \gamma_2)}{\gamma_1\gamma_2}.$$

Analogous derivations for the sparse component yield:

$$\|A_2^\star - A_1^\star\|_2 \geq v_S - 2\alpha_L/p,$$

where $v_L \equiv \|L_2^\star - L_1^\star\|_2 \geq 0$, $v_S \equiv \|S_2^\star - S_1^\star\|_2 \geq 0$ are norm differences for the low-rank and the sparse components, respectively.

Remark 3-2. *Suppose the low-rank components are dominant (i.e. $\gamma_1, \gamma_2 \geq 1$), but their ℓ_2 norm difference change is small; i.e. $\|L_2^\star - L_1^\star\|_2 \leq \epsilon$, with $\epsilon > 0$ being a small enough constant). Then, we have:*

$$\|A_2^\star - A_1^\star\|_2 \geq \frac{1}{\gamma_2}\left(\|L_2^\star\|_\infty - \frac{\alpha_L}{p}\frac{\gamma_2}{\gamma_1}\right) - \epsilon.$$

Note that since the low rank components are constrained to be in the Ω space -$\|L_2^\star\|_\infty \leq \alpha_L/p$- it implies that the transition matrices are identifiable, only if $\gamma_2 < \gamma_1$ and $\|S_2^\star\|_\infty > \|S_1^\star\|_\infty$. The roles of $L_2^\star$ and $L_1^\star$ can be swapped to obtain that only if $\gamma_2 \neq \gamma_1$ and $\|S_2^\star\|_\infty \neq \|S_1^\star\|_\infty$, is the change in the full transition matrices $A_j^\star$ identifiable, which is intuitive.

By using the defined constraint space Ω and information ratio γ_j above, we provide the necessary assumptions to establish the theoretical results.

(H1) There exists a positive constant $C_0 > 0$ such that

$$\Delta_T(v_S^2 + v_L^2) \geq C_0 \left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right),$$

where Δ_T is the spacing between the change point $\tau^\star$ and the boundary, and v_S, v_L are the jump sizes, defined as

$$\Delta_T = \min\{\tau^\star - 1, T - \tau^\star\}, \quad v_S = \|S_2^\star - S_1^\star\|_2, \quad v_L = \|L_2^\star - L_1^\star\|_2.$$

Further, at least one of v_S, v_L is strictly positive.

(H2) (Identifiability conditions) Consider low rank matrices $L_1^\star$, $L_2^\star$, and their corresponding Singular Value Decompositions (SVD): $L_j^\star = U_j^\star D_j^\star V_j^{\star'}$, where $D_j^\star = \mathrm{diag}(\sigma_1^j, \ldots, \sigma_{r_j}^j, 0, \ldots, 0)$, for $j = 1, 2$ and $U_j^\star, V_j^\star$ are orthonormal. Then,

(1) there exists a universal positive constant $M_S > 0$, such that for the sparse matrices $S_j^\star$, we have: $\|S_j^\star\|_\infty \leq M_S < +\infty$, $j = 1, 2$;

(2) there exists a large enough constant $c > 0$, such that the diagonal matrices $D_j^\star$ satisfy: $\max_{j=1,2} \|D_j^\star\|_\infty \leq c < +\infty$; further the orthonormal matrices $U_j^\star$ and $V_j^\star$ satisfy: $\max_{j=1,2} \left\{\|U_j^\star\|_\infty, \|V_j^\star\|_\infty\right\} = \mathcal{O}\left(\sqrt{\frac{\alpha_L}{r_{\max} p}}\right)$, where $r_{\max} = \max\{r_1^\star, r_2^\star\}$. In addition, we assume that $\alpha_L = \mathcal{O}\left(p\sqrt{\frac{\log(pT)}{T}}\right)$.

(3) the maximal sparsity level $d_{\max}^\star = \max\{d_1^\star, d_2^\star\}$ satisfies: $d_{\max}^\star \leq \frac{1}{C_{\max}} \sqrt{\frac{T}{\log(pT)}}$, for a large enough positive constant $C_{\max} > 0$.

(H3) (Restrictions on the search domain $\mathcal{T}$) The change point $\tau^\star$ belongs to the search domain by $\mathcal{T} \subset \{1, 2, \ldots, T-1\}$ and denote the search domain $\mathcal{T} \stackrel{\text{def}}{=} [a, b]$. Assume that, $a = \lfloor (d_{\max}^\star + \sqrt{r_{\max}^\star})^{1+\eta} \rfloor$ and $b = \lfloor T - (d_{\max}^\star + \sqrt{r_{\max}^\star})^{1+\eta} \rfloor$, and denote $|\mathcal{T}|$ as the length of the search domain, then:

$$\frac{|\mathcal{T}|}{d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)} \to +\infty,$$

where $\eta > 0$ is an arbitrarily small positive constant, $d^\star_{\max} = \max\{d^\star_1, d^\star_2\}$, and $r^\star_{\max} = \max\{r^\star_1, r^\star_2\}$.

Note that for any fixed time point τ in the search domain $\mathcal{T}$, let $(\lambda_{1,\tau}, \mu_{1,\tau})$ be the tuning parameters on $[1, \tau)$, and $(\lambda_{2,\tau}, \mu_{2,\tau})$ the tuning parameters on $[\tau, T)$, respectively. Then, the tuning parameters of the regularization terms are selected as follows: for constants $c_0, c_0' > 0$,

$$\begin{aligned}(\lambda_{1,\tau}, \mu_{1,\tau}) &= \left(4c_0\sqrt{\frac{\log p + \log(\tau - 1)}{\tau - 1}},\ 4c_0'\sqrt{\frac{p + \log(\tau - 1)}{\tau - 1}}\right), \\ (\lambda_{2,\tau}, \mu_{2,\tau}) &= \left(4c_0\sqrt{\frac{\log p + \log(T - \tau)}{T - \tau}},\ 4c_0'\sqrt{\frac{p + \log(T - \tau)}{T - \tau}}\right),\end{aligned} \tag{3-6}$$

Theorem 3-1. *Suppose Assumptions H1-H3 hold, and select the tuning parameters according to (3-6). Then, as $T \to +\infty$, there exists a large enough constant $K_0 > 0$ such that*

$$\mathbb{P}\left(|\widehat{\tau} - \tau^\star| \le K_0 \frac{d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)}{v_S^2 + v_L^2}\right) \to 1.$$

Note that the Theorem 3-1 provides the upper bound for the single change point. After we detect the change point, we establish a consistent estimation of model parameter as follows. First, given the estimated change point $\widehat{\tau}$, we remove it together with its R-radius neighborhood $\mathcal{U}(\widehat{\tau}, R)$ to ensure that the remaining time points consist two stationary segments. Based on the result in Theorem 3-1, the radius R can be of the order $\mathcal{O}\left(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\right)$.

Denote N_j be the length of the j-th segment after removing the R-radius neighborhoods; then we select another pair of tuning parameters:

$$(\lambda_j, \mu_j) = \left(4c_1\sqrt{\frac{\log p}{N_j}} + \frac{4c_1\alpha_L}{p},\ 4c_1'\sqrt{\frac{p}{N_j}}\right), \quad j = 1, 2, \tag{3-7}$$

for constants c_1, c_1' that can selected using cross-validation.

In accordance to the tuning parameters in (3-7), we establish the following theorem about the optimal estimation rate:

Theorem 3-2. *Suppose Assumptions H1-H3 hold, and select the tuning parameters according to (3-7). Then, as $T \to +\infty$, there exist universal positive constants $C_1, C_2 > 0$, so that the optimal solution of (3-5) satisfies*

$$\|\widehat{L}_j - L_j^\star\|_F^2 + \|\widehat{S}_j - S_j^\star\|_F^2 \leq C_1 \left(\frac{d_j^\star \log p + r_j^\star p}{N_j} \right) + C_2 \frac{d_j^\star \alpha_L^2}{p^2}, \quad j = 1, 2.$$

3.2 Multiple Change Point Detection

3.2.1 Model Formulation

Now, we consider a general multiple change points scenario. Similarly, we start by formulating a p-dimensional piece-wise VAR time series model with m_0 change points: $0 = \tau_0^\star < \tau_1^\star < \cdots < \tau_{m_0}^\star < \tau_{m_0+1}^\star = T$; then the model is written as:

$$X_t = \sum_{j=1}^{m_0+1} (A_j^\star X_{t-1} + \epsilon_t^j) \mathbf{I}(\tau_{j-1}^\star \leq t < \tau_j^\star), \quad t = 1, 2, \ldots, T, \tag{3-8}$$

where $A_j^\star$ is further decomposed into a low rank component $L_j^\star$ and a sparse component $S_j^\star$, and $\mathbf{I}(\tau_{j-1}^\star \leq t < \tau_j^\star)$ denotes the indicator function for the j-th stationary segment. Analogously to the single change point case, we define the sparsity level $d_j^\star = \|S_j^\star\|_0$ and rank $r_j^\star = \text{rank}(L_j^\star)$ for the components in each segment, wherein $d_j^\star \ll p^2$ and $r_j^\star \ll p$, (i.e., $d_j^\star = o(p^2)$ and $r_j^\star = o(p)$). Finally, ϵ_t^j's are independent and independently distributed zero mean Gaussian noise processes with covariance matrices $\sigma^2 I$, $j = 1, \ldots, m_0 + 1$.

For detecting the change points and estimating the model parameters consistently, the following minor modifications to the assumptions H1-H3 are required:

(H1') There exists a positive constant $C_0 > 0$ such that

$$\Delta_T \min_{1 \leq j \leq m_0} \{v_{j,S}^2 + v_{j,L}^2\} \geq C_0 (d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)),$$

where Δ_T is the minimum spacing defined as $\Delta_T \stackrel{\text{def}}{=} \min_{1 \leq j \leq m_0} |\tau_{j+1}^\star - \tau_j^\star|$, and the minimum norm differences (jump sizes) between two consecutive segments are defined as: $v_{j,S} \stackrel{\text{def}}{=} \|S_{j+1}^\star - S_j^\star\|_2$, and $v_{j,L} \stackrel{\text{def}}{=} \|L_{j+1}^\star - L_j^\star\|_2$.

(H2') Consider low rank matrices $L_j^\star$, and their corresponding Singular Value Decompositions: $L_j^\star = U_j^\star D_j^\star V_j^{\star\prime}$, where $D_j^\star = \text{diag}(\sigma_1^j, \ldots, \sigma_{r_j}^j, 0, \ldots, 0)$, for $j = 1, 2, \ldots, m_0 + 1$. Then,

(1) there exists a universal positive constant $M_S > 0$, such that for the sparse matrices $S_j^\star$, we have: $\|S_j^\star\|_\infty \leq M_S < +\infty$, $j = 1, \ldots, m_0 + 1$;

(2) there exists a large enough constant $c > 0$, such that the diagonal matrices $D_j^\star$ satisfy: $\max_{j=1,2} \|D_j^\star\|_\infty \leq c < +\infty$, and the orthonormal matrices $U_j^\star$ and $V_j^\star$ such that: $\max_{1\leq j\leq m_0+1} \left\{\|U_j^\star\|_\infty, \|V_j^\star\|_\infty\right\} = \mathcal{O}\left(\sqrt{\frac{\alpha_L}{r_{\max} p}}\right)$, where $r_{\max} = \max_{1\leq j\leq m_0+1} r_j^\star$. In addition, we assume that $\alpha_L = \mathcal{O}\left(p\sqrt{\frac{\log(pT)}{T}}\right)$.

(3) the maximal sparsity level $d_{\max}^\star = \max_{1\leq j\leq m_0+1} d_j^\star$ satisfies: $d_{\max}^\star \leq \frac{1}{C_{\max}}\sqrt{\frac{T}{\log(pT)}}$, for a large enough positive constant $C_{\max} > 0$.

(H3') There exists a vanishing positive sequence $\{\xi_T\}$ such that, as $T \to +\infty$,

$$\frac{\Delta_T}{T\xi_T(d_{\max}^{\star 3} + r_{\max}^{\star 2})} \to +\infty, \quad d_{\max}^{\star 2}\sqrt{\frac{\log p}{T\xi_T}} \to 0, \quad r_{\max}^{\star \frac{3}{2}}\sqrt{\frac{p}{T\xi_T}} \to 0,$$
$$\frac{\Delta_T(d_{\max}^\star \log p + r_{\max}^\star p)}{(T\xi_T)^2(d_{\max}^{\star 3} + r_{\max}^{\star 2})} \to C \geq 1,$$

for a positive constant $C > 0$.

3.2.2 A Two-step Detection Algorithm

Since the use of fused type penalties is not applicable to the low rank component and hence we develop the following two-step algorithm to detect multiple change points.

Step 1: It is based on Algorithm 1 provided in Appendix B that detects a single change point, additionally equipped with a *rolling window* mechanism to select *candidate* change points. We start by selecting an interval $[b_1, e_1) \subset \{1, 2, \ldots, T\}, b_1 = 1$, of length h and employ on it the exhaustive search Algorithm 1 to obtain a candidate change point $\widetilde{\tau}_1$. Next, we shift the interval to the right by l time points and obtain a new interval $[b_2, e_2)$, wherein $b_2 = b_1 + l$ and $e_2 = e_1 + l$. The application of Algorithm 1 to $[b_2, e_2)$ yields another candidate change point $\widetilde{\tau}_2$. This procedure continues until the last interval that can be formed, namely $[b_{\widetilde{m}}, e_{\widetilde{m}})$, where $e_{\widetilde{m}} = T$ and $\widetilde{m}$ denotes the number of windows of size h that can be formed. The following Figure 3-1 depicts this rolling-window mechanism. The blue lines represent the boundaries of each window, awhile the green dashed lines represent the candidate change point in each window. Note that the basic assumption for

Algorithm 1 is that there exists a single change point in the given time series. However, it can easily be seen in Figure 3-1 that **not** every window includes a single change point.

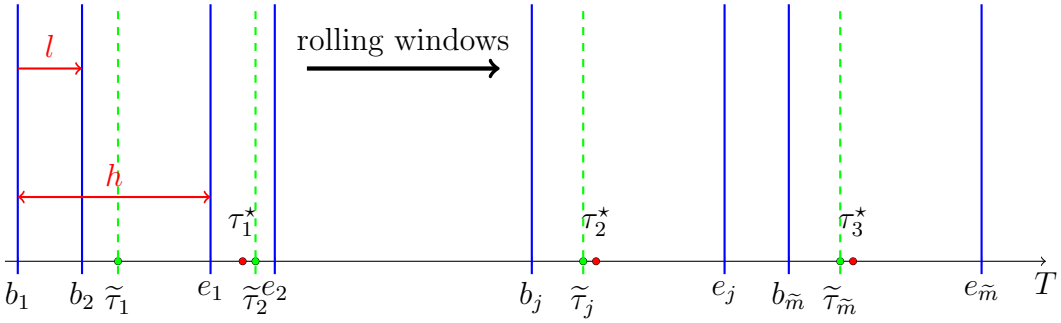
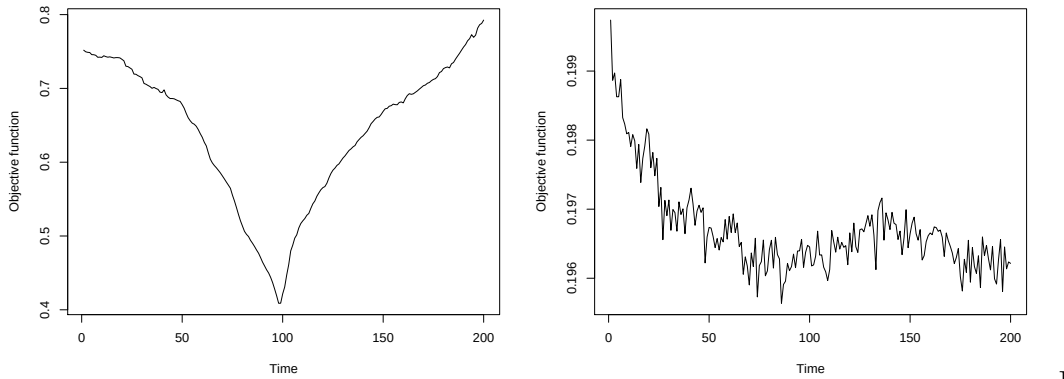


Figure 3-1. Depiction of the rolling windows strategy. There are three true change points: $\tau_1^\star$, $\tau_2^\star$, and $\tau_3^\star$ (red dots); the boundaries of the rolling-window are represented in blue lines; the estimated change points in each window are plotted in green dashed lines, where the subscript indicates the index of the window used to obtain it.

To showcase the last point, we compare the behavior of Algorithm 1 on an interval with and without a change point based on data generated from a low-rank plus sparse VAR process $\{X_t\}$ with $p = 20$. We select two windows of length $h = 200$, one containing a change point at $t = 100$ and another not containing a change point. Plots of the objective function (3-3) used in Algorithm 1 for these two windows are depicted in the panels A and B of Figure 3-2, respectively. It can be seen that in the presence of a change point, a



A B

Figure 3-2. Plots of the objective functions obtained by an application of Algorithm 1.

clearly identified minimum close to the true change point exists. Contrary, in the absence

of a change point, the objective function is mostly flat without a clearly identified minimum. Next, we introduce an assumption on the size of the window h used in the detection procedure:

(H4) Let h denote the length of the window in the rolling window algorithm. Further, the minimum spacing Δ_T and the vanishing sequence $\{\xi_T\}$ are defined as in Assumption H3', and let l denote the length by which the window is shifted to the right; it is assumed that:

$$0 < l \leq \max\{\frac{h}{2}, 1\}, \ \limsup_{T\to+\infty} \frac{h}{\Delta_T} < 1, \ \text{and} \ \liminf_{T\to+\infty} \frac{h}{T\xi_T} \geq 2.$$

Assumption H4 restricts h, so that asymptotically can not include more than a single true change point and also is not too small, so that the deviation bound and restricted eigenvalue conditions used for establishing theoretical properties of the estimates of the model parameters hold for each time segment. Further, this assumption places an upper bound on the shift l, to ensure that no true break point close to the boundary of windows would be missed by the proposed algorithm. The shift size can vary in $[1, h/2]$; a small l helps reduce the finite sample estimation error for locating the break points, while a large l speeds up the detection procedure, by considering fewer rolling windows.

Next, we establish theoretical guarantees for Step 1 of the proposed detection procedure. Denote by $\widetilde{\mathcal{S}}$ the set of candidate change points and by $\mathcal{S}^\star$ the set of true change points. Specifically, $\widetilde{\mathcal{S}}$ is defined as:

$$\widetilde{\mathcal{S}} \stackrel{\text{def}}{=} \left\{ \widetilde{t}_i \in [b_i, e_i) : \widetilde{t}_i = \underset{\tau\in[b_i,e_i)}{\arg\min}\, \ell(\tau; \widehat{L}_{1,\tau}, \widehat{L}_{2,\tau}, \widehat{S}_{1,\tau}, \widehat{S}_{2,\tau}), \quad i = 1, 2, \ldots, \widetilde{m} \right\},$$

where $[b_i, e_i)$ is the i-th rolling-window.

Next, we extend Theorem 3-1 to the multiple change points scenario:

Proposition 3-1. *Suppose Assumptions H1'-H3' and H4 hold, and select the tuning parameters for each rolling window according to (3-6). Then, as $T \to +\infty$, there exists a*

large enough constant $K > 0$ *such that*

$$\mathbb{P}\left(d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star) \leq K \frac{d^\star_{\max} \log(p \vee h) + r^\star_{\max}(p \vee \log h)}{\min_{1 \leq j \leq m_0}\{v^2_{j,S} + v^2_{j,L}\}} \right) \to 1.$$

Proposition 3-1 shows that the number of candidate change points identified in Step 1 of the algorithm is an overestimate of the true number of change points. Hence, a second *screening step* is required to remove the redundant ones.

Step 2: Let the candidate change points from Step 1 be denoted by $\{s_j\}$, $j = 1, 2, \cdots, \widetilde{m}$. Then, model (3-8) can be rewritten in the following form:

$$X_t = \sum_{i=1}^{\widetilde{m}+1} \left((L_{(s_{i-1},s_i)} + S_{(s_{i-1},s_i)}) X_{t-1} + \epsilon_t^i \right) \mathbf{I}(s_{i-1} \leq t < s_i), \quad t = 1, 2, \ldots, T,$$

where $L_{(s_{i-1},s_i)}$ and $S_{(s_{i-1},s_i)}$ denote for the low-rank and sparse components of the transition matrix in the interval $[s_{i-1}, s_i)$. We define
$0 = s_0 < s_1 < s_2 < \cdots < s_{\widetilde{m}} < s_{\widetilde{m}+1} = T$ and for ease of presentation use L_i and S_i instead of $L_{(s_{i-1},s_i)}$ and $S_{(s_{i-1},s_i)}$ for $i = 1, 2, \ldots, m+1$. We also define matrices
$\mathbf{L} \stackrel{\text{def}}{=} [L_1', L_2', \ldots, L_{\widetilde{m}+1}']'$ and $\mathbf{S} \stackrel{\text{def}}{=} [S_1', S_2', \ldots, S_{\widetilde{m}+1}']'$. Estimates for $\mathbf{L}$ and $\mathbf{S}$ are obtained as the solution to the following regularized regression problem:

$$(\widehat{\mathbf{L}}, \widehat{\mathbf{S}}) = \underset{L_i, S_i, 1 \leq i \leq \widetilde{m}+1}{\arg\min} \sum_{i=1}^{\widetilde{m}+1} \left\{ \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i - 1} \|X_t - (L_i + S_i) X_{t-1}\|_2^2 + \lambda_i \|S_i\|_1 + \mu_i \|L_i\|_* \right\},$$

with tuning parameters $(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \{(\lambda_i, \mu_i)\}_{i=1}^{\widetilde{m}+1}$. Next, we define the objective function with respect to $(s_1, s_2, \ldots, s_m)$:

$$\mathcal{L}_T(s_1, s_2, \ldots, s_m; \boldsymbol{\lambda}, \boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{i=1}^{\widetilde{m}+1} \left\{ \sum_{t=s_{i-1}}^{s_i - 1} \|X_t - (\widehat{L}_i + \widehat{S}_i) X_{t-1}\|_2^2 + \lambda_i \|\widehat{S}_i\|_1 + \mu_i \|\widehat{L}_i\|_* \right\}. \quad (3\text{-}9)$$

Then, for a suitably selected penalty sequence ω_T, specified in the upcoming Assumption H5, we consider the following *information criterion* defined as:

$$\mathrm{IC}(s_1, s_2, \ldots, s_m; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T) \stackrel{\text{def}}{=} \mathcal{L}_T(s_1, \ldots, s_m; \boldsymbol{\lambda}, \boldsymbol{\mu}) + m\omega_T. \quad (3\text{-}10)$$

The second step selects a subset of initial $\widetilde{m}$ change points from the first step by solving:

$$(\widehat{m}, \widehat{\tau}_i, i = 1, 2, \ldots, \widehat{m}) = \underset{0 \le m \le \widetilde{m}, (s_1, \ldots, s_m)}{\arg\min} \mathrm{IC}(s_1, \ldots, s_m; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T).$$

The following two additional assumptions on the minimum spacing Δ_T and the selection of tuning parameters are required to establish the Theorem 3-3.

(H5) Assume that $m_0 T \xi_T (d^{\star 2}_{\max} + r^{\star \frac{3}{2}}_{\max})/\omega_T \to 0$ and $m_0 \omega_T / \Delta_T \to 0$ as $n \to +\infty$.

(H6) Suppose $(s_1, \ldots, s_m)$ are a set of change points obtained from the Step 1, we consider the following scenarios: (a) if $|s_i - s_{i-1}| \le T\xi_T$, select $\lambda_i = c\sqrt{T\xi_T \log p}$ and $\mu_i = c\sqrt{T\xi_T p}$, for $i = 1, 2, \ldots, m$; (b) if there exist two true change points $\tau^\star_j$ and $\tau^\star_{j+1}$ such that $|s_{i-1} - \tau^\star_j| \le T\xi_T$ and $|s_i - \tau^\star_{j+1}| \le T\xi_T$, select $\lambda_i = 4\left(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d^\star_{\max} \frac{T\xi_T}{s_i - s_{i-1}}\right)$ and $\mu_i = 4\left(c\sqrt{\frac{p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r^\star_{\max}} \frac{T\xi_T}{s_i - s_{i-1}}\right)$; (c) otherwise, select $\lambda_i = 4c\sqrt{\frac{\log p + \log(s_i - s_{i-1})}{s_i - s_{i-1}}}$ and $\mu_i = 4c\sqrt{\frac{p + \log(s_i - s_{i-1})}{s_i - s_{i-1}}}$, for some large constant c.

The screening step for removing the redundant candidate change points are summarized in the following algorithm diagram.

Theorem 3-3. *Suppose Assumptions H1'–H3', and H4–H6 hold. As $T \to +\infty$, the minimizer $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{m}})$ of (3-10) satisfies: $\mathbb{P}(\widehat{m} = m_0) \to 1$. Further, there exists a large enough positive constant $B > 0$ so that*

$$\mathbb{P}\left(\max_{1 \le j \le m_0} |\widehat{\tau}_j - \tau^\star_j| \le B m_0 T \xi_T \frac{d^{\star 2}_{\max} + r^{\star \frac{3}{2}}_{\max}}{\min_{1 \le j \le m_0} \{v^2_{j,S} + v^2_{j,L}\}}\right) \to 1.$$

Remark 3-3. *For a finite number of change points m_0, the sequence $\{\xi_T\}$ can be selected as $(d^\star_{\max} \log(p \vee T) + r^\star_{\max}(p \vee \log T))^{1+\frac{\rho}{2}}/T$ for some small $\rho > 0$. Assuming that the maximum rank among all the low-rank components and the maximum sparsity level among all the sparse components satisfy $d^{\star 2}_{\max} + r^{\star \frac{3}{2}}_{\max} = o\left((d^\star_{\max} \log(p \vee T) + r^\star_{\max}(p \vee \log T))^{\frac{\rho}{2}}\right)$, then the order of detecting the relative location -$\tau^\star_j/T$- becomes $(d^\star_{\max} \log(p \vee T) + r^\star_{\max}(p \vee \log T))^{1+\rho}/T$ in Theorem 3-3. Finally, one can choose the penalty tuning parameter ω_T to be of order $(d^\star_{\max} \log(p \vee T) + r^\star_{\max}(p \vee \log T))^{1+2\rho}$ in this setting, and the minimum spacing Δ_T to be at least of order*

Algorithm 3 Screening via a Backwards Elimination Algorithm

1. **Input:** Time series data $\{X_t\}$, $t = 1, 2, \ldots, n$; candidate change points $\{\widetilde{t}_j\}$ for $j = 1, 2, \ldots, \widetilde{m}$.

2. **Initialize:** Define the interval partition of time axis based on candidate change points: $\mathcal{P} \stackrel{\text{def}}{=} \big\{\{1, \ldots, \widetilde{t}_1\}, \{\widetilde{t}_1 + 1, \ldots, \widetilde{t}_2\}, \ldots, \{\widetilde{t}_{\widetilde{m}} + 1, \ldots, n\}\big\}$. Set the initial value of information criterion is $W_0 = 0$ and the number of final selected change points $m = \widetilde{m}$.

3. **While** $W_{m-1} \leq W_m$ and $m \neq 1$ **do:**

   - Let $\widetilde{\mathbf{t}} \stackrel{\text{def}}{=} \{\widetilde{t}_1, \ldots, \widetilde{t}_m\}$ be the screened change points and define $W_m^* = \mathrm{IC}(\widetilde{\mathbf{t}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_n)$;
   - For each $j = 1, 2, \ldots, m$, we calculate $W_{m,-j} = \mathrm{IC}(\widetilde{t}/\{\widetilde{t}_j\}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_n)$, and define $W_{m-1} = \min_j W_{m,-j}$;
   - There are three cases:
     (a) If $W_{m-1} > W_m$, then no further step is needed. Return the current partition $\widehat{\mathcal{P}}$;
     (b) If $W_{m-1} \leq W_m$ and $m > 1$, set $j^* = \arg\min_j W_{m,-j}$, then we update candidate change points vector $\widetilde{\mathbf{t}} \leftarrow \widetilde{\mathbf{t}}/\widetilde{t}_{j^*}$ and $m \leftarrow m - 1$;
     (c) If $W_{m-1} \leq W_m$ and $m = 1$, return an empty set.

4. **Output:** The final set of screened change points $\{\widehat{t}_j\}$, for $j = 1, 2, \ldots, \widehat{m}$.

$(d^\star_{\max} \log(p \vee T) + r^\star_{\max}(p \vee \log T))^{2+\rho}$ *in accordance to Assumption H3'. Comparing the consistency rates provided in Theorem 3-3 with those in Safikhani & Shojaie (2020), the additional term* $r^\star_{\max}(p \vee \log T)$ *reflects the complexity of estimating the low-rank components in the model.*

Remark 3-4. *For the proposed p-dimensional VAR model with* T *observations and window size* $h = \mathcal{O}(T^\delta)$*, where* $\delta \in (0, 1]$*, the computational complexity of the first step is of order* $\mathcal{O}(Tp^3)$*, and the second screening step is of order* $\mathcal{O}(T^{1-\delta}p^3)$*, since an SVD is required to estimate the low-rank components for every search. Hence, the overall complexity is linear in* T*.*

Next, we establish a consistent estimator for the model parameters similar to the Theorem 3-2.

Theorem 3-4. *Given the estimated change points:* $1 = \widehat{\tau}_0 < \widehat{\tau}_1 < \cdots < \widehat{\tau}_{\widehat{m}} < \widehat{\tau}_{\widehat{m}+1} = T$*, let Assumptions H1'-H3' and H4 hold and remove the R-radius neighborhoods for each* $\widehat{\tau}_j$ *for* $j = 1, 2, \ldots, \widehat{m}+1$*. Further, by using the following tuning parameters:* $(\lambda_j, \mu_j) = \left(4c_1\sqrt{\frac{\log p}{N_j}} + \frac{4c_1\alpha_L}{p},\ 4c_1'\sqrt{\frac{p}{N_j}}\right)$*, where* c_1, c_1' *are positive constants. For* $T \to +\infty$*, there exist universal positive constants* $C_1', C_2' > 0$ *such that for each selected segment, the estimated low-rank and the sparse components satisfy*

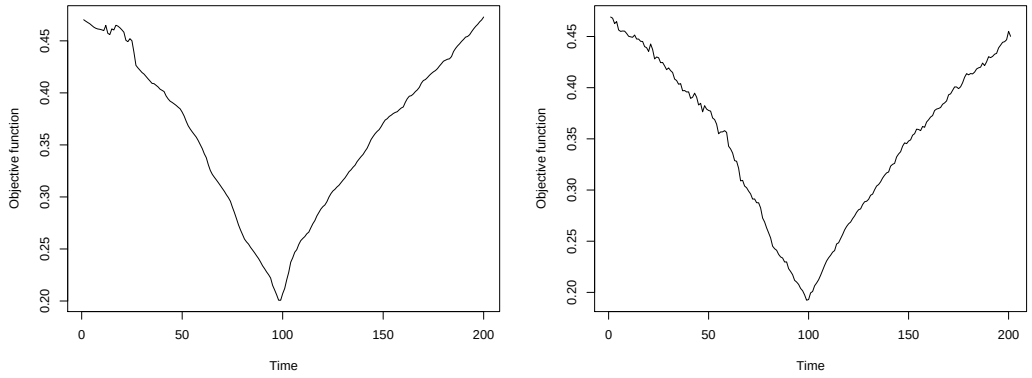
$$\|\widehat{L}_j - L_j^\star\|_F^2 + \|\widehat{S}_j - S_j^\star\|_F^2 \leq C_1'\left(\frac{d_j^\star \log p + r_j^\star p}{N_j}\right) + C_2'\frac{d_j^\star \alpha_L^2}{p^2}.$$

3.3 A Fast Procedure Based on Surrogate Model

Remark 3-4 illustrates that identifying multiple change points in a low rank and sparse VAR model is computationally expensive, due to the presence of the nuclear norm and the need for selecting the tuning parameters through a 2-dimensional grid search.

The question addressed next is whether there are settings wherein the nature of the signal in the norm difference $||A_j^\star - A_{j+1}^\star||_2$ is such that it can be *adequately captured* by a less computationally demanding surrogate model. For example, if the norm difference is primarily due to a large enough change in the sparse component, it is reasonable to expect that a *surrogate* VAR model with a *sparse* transition matrix may prove adequate under certain regularity conditions. However, if the norm difference is due to a change in the low-rank component, which by construction is dense, a pure sparse VAR model will not be adequate; however, a *weakly sparse* model may be sufficient. Indeed, some numerical evidence suggests that this is the case. Figure 3-3 presents plots of the objective functions of the original (panel A) and the surrogate weakly sparse (panel B) model under the same experimental setting for a low-rank plus sparse VAR process $\{X_t\}$ with $p = 20$, $T = 200$, and a single change point at $\tau^\star = 100$ with *changes in both the low-rank and sparse* components.

As can be seen, the plot for the surrogate weakly sparse model shares a similar pattern to that of the true model. However, in practice, we can not a priori guarantee a



A B

Figure 3-3. The curve of the objective function of the full low-rank plus sparse and surrogate weakly sparse models.

change both in the low-rank and the sparse component, simultaneously. Therefore, an extra assumption is required to ensure the detectability of the change points. Before we state it, we first introduce formally the surrogate piece-wise weakly sparse VAR model.

3.3.1 Formulation of the Surrogate Weakly Sparse VAR Model

A $p \times p$ real matrix A is weakly sparse, if it satisfies

$$\mathbb{B}_q(R_q) := \left\{ A \in \mathbb{R}^{p \times p} : \sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|^q \leq R_q \right\}, \tag{3-11}$$

for some $q \in (0, 1)$; namely, its entries are restricted in an ℓ_q ball of radius R_q (Negahban et al. 2012). Note that when $q \to 0^+$, this set converges to an exact sparse model, that is, $A \in \mathbb{B}_0(R_0)$, if and only if A has at most R_0 nonzero elements. When $q \in (0, 1)$, the set $\mathbb{B}_q(R_q)$ enforces a certain rate of decay on the ordered absolute values of A.

We focus the discussion on detecting a single change point and establish under what conditions the change point can be estimated consistently based on the weakly sparse surrogate model. Subsequently, we extend the result to the case of multiple change points using the proposed rolling window strategy.

Since the focus is on the weakly sparse VAR model, the detection procedure provided in Section 3.1 requires some modification.

We assume that $(A_1^\star, A_2^\star) \in \mathbb{B}_q(R_q)$, for some $q \in (0, 1)$ and $R_q > 0$. We also introduce

a modification on the Assumptions made in Sections 3.1.3. Based on Remark 3-3 and using the same notation as in the results in Sections 3.1, the counterpart of Assumption H1 becomes:

(W1) The weakly sparse assumption on the $A_j^\star$'s singles out *spiky* entries. Hence, one of the following needs to hold:

(1) If $\gamma_1, \gamma_2 \geq p$, then we require the minimum spacing Δ_T and the jump size $v_A = \|A_2^\star - A_1^\star\|_2$ satisfy:

$$\Delta_T v_A^2 \geq C_0^w \left(T^{\frac{q}{2}} R_q (\log(p \vee T))^{1-\frac{q}{2}} \right);$$

(2) Otherwise, the change point is identifiable as long as:

$$\Delta_T v_S^2 \geq C_0^w \left(T^{\frac{q}{2}} R_q (\log(p \vee T))^{1-\frac{q}{2}} \right).$$

Remark 3-5. *Assumption W1 is based on Remark 3-3. Note that if the low-rank components dominate the signal, then an adequate change in them is required to identify the change point; otherwise, we need different information ratios together with distinct spiky entries in the sparse components. The latter sufficient condition indicates that the changes in the spiky entries play an important role in identifying the change points. For the second case, if the low-rank components are not dominant in both segments, then an adequately large change in the sparse components is sufficient to determine the change point.*

3.3.2 Theoretical Properties

The following proposition provides a lower bound for the radius R_q, so that the true transition matrices $(A_1^\star, A_2^\star)$ that admit a low-rank plus sparse decomposition do belong to the above defined ℓ_q ball. We only discuss the case $0 < \gamma_1, \gamma_2 \leq p$. Analogous results for the other cases can be derived in a similar manner.

Proposition 3-2. *Let $q \in (0, 1)$ be fixed and $R_q > 0$ be the radius of $\mathbb{B}_q(R_q)$ defined in (3-11). Further, the transition matrices for the data generating model satisfy the following decomposition: $A_1^\star = L_1^\star + S_1^\star$ and $A_2^\star = L_2^\star + S_2^\star$, where $L_1^\star$, $L_2^\star$, $S_1^\star$, and $S_2^\star$ are the corresponding low-rank and sparse components. Then, $A_1^\star, A_2^\star$ belong to $\mathbb{B}_q(R_q)$ if R_q*

satisfies:

$$R_q \geq d^\star_{\max}\left(\left(\frac{\alpha_L}{p}\right)^q + M_S^q\right) + (p^2 - d^\star_{\max})|\sigma_{\max}|^q,$$

where $\sigma_{\max} = \max\{\|L_1^\star\|_2, \|L_2^\star\|_2\}$ *and* $d^\star_{\max} = \max\{d_1^\star, d_2^\star\}$.

Before we extend Theorem 3-1 to the surrogate weakly sparse model, a modification to the selection of tuning parameters is required. Recall that (3-6) identifies the tuning parameters for the low-rank plus sparse model, while for the surrogate weakly sparse model, the only parameter is the transition matrix $A_j^\star$ for $j = 1, 2$. Along with the notation defined in (3-6), the tuning parameters are given by:

$$\lambda^w_{1,\tau} = 4c_0^w\sqrt{\frac{\log p + \log(\tau - 1)}{\tau - 1}}, \quad \lambda^w_{2,\tau} = 4c_0^{w'}\sqrt{\frac{\log p + \log(T - \tau)}{T - \tau}}, \tag{3-12}$$

where $c_0^w, c_0^{w'} > 0$ are some positive constants selected by the similar method as c_0 and c_0' in (3-6), the selection procedure is provided in the next section. Since we employ the same exhaustive search algorithm in Algorithm 1, a similar assumption as H3 on the search domain $\mathcal{T}^w$ is required.

(W2) Using similar definitions to Assumption H3, denote the search domain by $\mathcal{T}^w \stackrel{\text{def}}{=} [a^w, b^w]$, and let $|\mathcal{T}^w|$ to be the length of $\mathcal{T}^w$. Then, we assume that,

$$a^w = \left\lfloor R_q\left(\frac{\log(p \vee T)}{T}\right)^{-\frac{q}{2}}\right\rfloor, \; b^w = \left\lfloor T - R_q\left(\frac{\log(p \vee T)}{T}\right)^{-\frac{q}{2}}\right\rfloor, \; \frac{|\mathcal{T}^w|}{T^{\frac{q}{2}}R_q(\log(p \vee T))^{1-\frac{q}{2}}} \to +\infty.$$

We are now in a position to extend the result in Theorem 3-1 in the following proposition.

Proposition 3-3. *Suppose Assumptions W1 and W2 hold and the transition matrices* $A_1^\star$ *and* $A_2^\star$ *in* (3-1) *belong to the set* $\mathbb{B}_q(R_q)$ *for some fixed constant* $q \in (0, 1)$ *and radius* $R_q > 0$, *such that* $c_1\sqrt{R_q}\left(\frac{\log p + \log T}{T}\right)^{\frac{1}{2}-\frac{q}{4}} \leq 1$ *for some constant* $c_1 > 0$. *Then, by employing Algorithm 1 and using the tuning parameters as in* (3-12), *there exists a large enough constant* $K_0^w > 0$ *such that, with respect to the jump size* $v_A = \|A_2^\star - A_1^\star\|_2$, *as* $T \to +\infty$

$$\mathbb{P}\left(|\widehat{\tau} - \tau^\star| \leq K_0^w\frac{T^{\frac{q}{2}}R_q\left(\log(p \vee T)\right)^{1-\frac{q}{2}}}{v_A^2}\right) \to 1.$$

The following Proposition extends the above result to the case of multiple change points based on the rolling window strategy previously described. The window size h can be selected by substituting the vanishing sequence $\{\xi_T\}$ in Assumption H4 by the vanishing sequence $\{\xi_T^w\}$ defined in Assumption W3 below, for the weakly sparse model.

Proposition 3-4. *Suppose Assumptions W1 and W2 hold and the transition matrices $A_j^\star$, $j = 1, \ldots, m_0 + 1$ belong to the set $\mathbb{B}_q(R_q)$ for some fixed constant $q \in (0, 1)$ and the ℓ_q-ball radius $R_q > 0$ satisfies that $\sqrt{R_q}\left(\frac{\log p + \log h}{h}\right)^{\frac{1}{2}-\frac{q}{4}} \leq 1$. Then, by employing the rolling window strategy, we obtain the candidate change points set $\widetilde{\mathcal{S}}_w = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{m}}\}$. Then, as $T \to +\infty$, there exists a large enough constant $K_1^w > 0$ such that,*

$$\mathbb{P}\left(d_H(\widetilde{\mathcal{S}}_w, \mathcal{S}^\star) \leq K_1^w \frac{h^{\frac{q}{2}} R_q \left(\log(p \vee h)\right)^{1-\frac{q}{2}}}{\min_{1\leq j\leq m_0} v_{j,A}^2}\right) \to 1,$$

where $v_{j,A} = \|A_{j+1}^\star - A_j^\star\|_2$.

Recall that the rolling-window mechanism will result in a number of *redundant* candidate change points. By using the surrogate weakly sparse model, we obtain a few redundant candidate change points as well. Therefore, we need to remove those redundant change points by using a similar screening step as introduced in the two-step algorithm in Section 3.2.2. Similarly, we also extend Assumptions H3', H5 and H6 to the weakly sparse scenario -Assumptions W3 and W4 given in Appendix C- in order to formally introduce the theoretical results for the surrogate model.

Employing the selected tuning parameters as detailed in Assumptions W3 and W4, we can establish consistent estimation of the change points.

Proposition 3-5. *Suppose Assumptions W1–W4 hold and denote $(\widehat{\tau}_1^w, \ldots, \widehat{\tau}_{\widehat{m}^w}^w)$. Then, as $T \to +\infty$, there exists a large enough positive constant $B^w > 0$ such that*

$$\mathbb{P}\left(\max_{1\leq j\leq m_0} |\widehat{\tau}_j^w - \tau_j^\star| \leq B^w m_0 T \xi_T^w \frac{R_q^2 \left(\log(p \vee T)/T\right)^{-q}}{\min_{1\leq j\leq m_0} v_{j,A}^2}\right) \to 1.$$

Remark 3-6. *Proposition 3-5 provides the consistency rate of the final estimated change points obtained by the surrogate weakly sparse model. In the case of m_0 being finite, we*

select the vanishing sequence $\{\xi_T^w\}$ *to be of order* $R_q^2 \left(\log(p \vee T)\right)^{(1+\rho+q)}/T$ *for some arbitrarily small constant* $\rho > 0$. *Therefore, the consistency rate in Proposition 3-5 becomes* $B' m_0 T^q R_q^4 \left(\log(p \vee T)\right)^{(1+\rho)}$. *According to Assumption W3, the penalty term* ω_T^w *can be selected to be of the order* $T^{1+q}\xi_T^w R_q^2 \left(\log(p \vee T)\right)^{\rho-q}$ *and the minimum spacing in the weakly sparse model* Δ_T *must be at least* $T^{1+q}\xi_T^w R_q^2 \left(\log(p \vee T)\right)^{2\rho-q}$.

In order to formulate the theoretical properties of the surrogate model, we present several useful definitions and results. Specifically, for a chosen threshold $\eta_j > 0$, we firstly define the thresholded subset:

$$\mathcal{J}(\eta_j) \stackrel{\text{def}}{=} \left\{(k,l) \in \{1,2,\ldots,p\}^2 : |A_j^\star(k,l)| > \eta_j\right\}.$$

Recalling the ℓ_1 decomposition with respect to $\mathcal{J}(\eta_j)$, then we derive the upper bound of the cardinality of $\mathcal{J}(\eta_j)$ in terms of the threshold η_j and ℓ_q-ball radius R_q. Note that we have:

$$R_q \geq \sum_{k,l} |A_j^\star(k,l)|^q \geq \sum_{(k,l)\in\mathcal{J}(\eta_j)} |A_j^\star(k,l)|^q \geq \eta_j^q |\mathcal{J}(\eta_j)|,$$

hence, $|\mathcal{J}(\eta_j)| \leq R_q \eta_j^{-q}$ for any $\eta_j > 0$. Here, we set $\eta_j \propto \lambda_{j,\tau}^w$, for $j = 1, 2$, and denote $\widetilde{\Delta}_{1,\tau} \stackrel{\text{def}}{=} \widetilde{A}_{1,\tau}^w - A_1^\star$, $\widetilde{\Delta}_{2,\tau} \stackrel{\text{def}}{=} \widetilde{A}_{2,\tau}^w - A_2^\star$, and $\widetilde{\Delta}_{1/2,\tau} \stackrel{\text{def}}{=} \widetilde{A}_{1,\tau}^w - A_2^\star$, where $\widetilde{A}_{j,\tau}^w$ represent the estimated transition matrices by the weakly sparse model with respect to time point τ. Based on the decomposition of the ℓ_1 norm and the discussion in Section 4.3 in Negahban et al. (2012), we obtain that:

$$\begin{aligned}
\|\widetilde{\Delta}_{1,\tau}\|_1 &\leq 4\sqrt{R_q}\eta_1^{-\frac{q}{2}}\|\widetilde{\Delta}_{1,\tau}\|_2 + 4R_q\eta_1^{1-q}, \\
\|\widetilde{\Delta}_{2,\tau}\|_1 &\leq 4\sqrt{R_q}\eta_2^{-\frac{q}{2}}\|\widetilde{\Delta}_{2,\tau}\|_2 + 4R_q\eta_2^{1-q}, \\
\|\widetilde{\Delta}_{1/2,\tau}\|_1 &\leq 4\sqrt{R_q}\eta_2^{-\frac{q}{2}}\|\widetilde{\Delta}_{1/2,\tau}\|_2 + 4R_q\eta_2^{1-q}.
\end{aligned}$$

Based on step 2 of in Section 3.2.2, denote by $s_1, s_2, \ldots, s_{\widetilde{m}}$ the candidate change points obtained from the rolling-window step. We analogously formulate the model as in

(3-9). Then, we estimate $A_{(s_{i-1},s_i)}$ by solving the following regularized problem:

$$\widehat{A}^w_{(s_{i-1},s_i)} = \underset{A\in\mathbb{B}_q(R_q)}{\arg\min} \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} \|X_t - AX_{t-1}\|_2^2.$$

Further, define the tuning parameter vector $\boldsymbol{\lambda}^w \overset{\text{def}}{=} (\lambda_1^w, \ldots, \lambda_{\widetilde{m}}^w)$ to obtain

$$\mathcal{L}_T^w(s_1, s_2, \ldots, s_m; \boldsymbol{\lambda}^w) \overset{\text{def}}{=} \sum_{i=1}^{\widetilde{m}+1} \left\{ \sum_{t=s_{i-1}}^{s_i-1} \|X_t - \widehat{A}_i^w X_{t-1}\|_2^2 + \lambda_i^w \|\widehat{A}_i^w\|_1 \right\}.$$

Then, we define the *information criterion* for the weakly sparse model as follows:

$$\text{IC}^w(s_1, s_2, \ldots, s_m; \boldsymbol{\lambda}^w, \omega_T^w) \overset{\text{def}}{=} \mathcal{L}_T^w(s_1, \ldots, s_m; \boldsymbol{\lambda}^w) + m\omega_T^w. \tag{3-13}$$

The final selected change points are given by:

$$(\widehat{m}^w, \widehat{\tau}_i^w, i = 1, 2, \ldots, \widehat{m}^w) = \underset{0\le m\le\widetilde{m}, (s_1,\ldots,s_m)}{\arg\min} \text{IC}^w(s_1, \ldots, s_m; \boldsymbol{\lambda}^w, \omega_T^w).$$

Then, we can use the exact same backward elimination algorithm as proposed in Algorithm 2 to screen the redundant candidate change points by substituting the information criterion function with the newly defined IC^w in (3-13).

3.4 Numerical Experiments

We start by investigating the performance of the exhaustive search algorithm for a single change point detection for the low-rank plus sparse VAR model and its surrogate counterpart and the two-step algorithm for detecting multiple change points for these models.

- **Data generation:** (1) We generate the time series data $\{X_t\}$ with a *single* change point at $\tau^\star = \lfloor T/2 \rfloor$ from model (3-1). We set the true ranks $r_1^\star = \lfloor p/15 \rfloor$, $r_2^\star = \lfloor p/15 \rfloor + 1$, and the information ratio $\gamma_1 = \gamma_2$ for most of the cases considered, unless otherwise specified. The low-rank components $L_1^\star$ and $L_2^\star$ are designed by randomly generating an orthonormal matrix U and singular values $\sigma_1, \ldots, \sigma_p$ to obtain $L_1^\star = \sum_{l=1}^{r_1^\star} \sigma_l \mathbf{u}_l \mathbf{u}_l'$, and $L_2^\star = \sum_{l=1}^{r_2^\star} \sigma_l \mathbf{u}_l \mathbf{u}_l'$, where $\mathbf{u}_l$ represents the l-th column of matrix U. Then, the sparse components share the same 1-off diagonal structure with values $-\|L_1^\star\|_\infty/\gamma_1$ and $\|L_2^\star\|_\infty/\gamma_2$, respectively. The error term $\{\epsilon_t\}$ is normally distributed from $\mathcal{N}_p(\mathbf{0}, 0.01\mathbf{I}_p)$. (2) In the *multiple* change points case, we create the time series data $\{X_t\}$ from model (3-8) with m_0 change points, the true ranks $r_j^\star$ are

randomly chosen from: $\lfloor p/10 \rfloor - 1, \lfloor p/10 \rfloor, \lfloor p/10 \rfloor + 1$ unless otherwise specified, and the information ratios are fixed to $\gamma_j = 0.25$. The low-rank components are designed in a similar way as the single change point case, and the j-th sparse components are generated by $(-1)^j \|L_j^\star\|_\infty / \gamma_j$.

- **Tuning parameter selection:** To select the tuning parameters related to optimization problem (3-3), we can use the theoretical values of λ_j and μ_j provided in (3-6) and (3-7), and select the constants c_0 and c_0' by using a grid search as follows:

  (1) Choose an equally spaced sequence within $[0.001, 10]$ as the range for constants c_0 and c_0' to construct the grid $\mathcal{G}(\lambda, \mu)$;

  (2) Next, extract a time point every k time points (we set $k = 5$ in all numerical settings) to construct the testing set $\mathcal{T}_{\text{test}}$, and use the remaining time points as the training set $\mathcal{T}_{\text{train}}$, and denote the corresponding estimated transition matrix $\widehat{A}_{(\lambda,\mu)}$ with respect to the tuning parameters (λ, μ);

  (3) Select the tuning parameters $(\widehat{\lambda}, \widehat{\mu})$ satisfying:

$$(\widehat{\lambda}, \widehat{\mu}) = \underset{(\lambda,\mu)\in\mathcal{G}(\lambda,\mu)}{\arg\min} \left\{ \frac{1}{|\mathcal{T}_{\text{test}}|} \sum_{t\in\mathcal{T}_{\text{test}}} \|X_{t+1} - \widehat{A}_{(\lambda,\mu)} X_t\|_2^2 \right\}.$$

- **Window size selection:** The width of the rolling window plays an important role in the multiple change points scenario. In practice, we can manually select a suitable window-size, or we may use the following strategy. In Assumption H4, we provided conditions on the window size h and rolling step size l. Next, we discuss an iterative procedure for determining these two parameters in practice.
(1) Start with $h = cT^\delta$, and $l = h/4$, where δ is selected from 1 to 0.5 (equally spaced) and $0 < c < 1$ is a constant; (2) For a given δ, apply Algorithm 2 and obtain the final set of change points $\{\widehat{\tau}_1, \ldots, \widehat{\tau}_m\}$; (3) Repeat (2) until the number of the final set of change points does not change. Return the corresponding window size $\widehat{h}$.

- **Model evaluation:** We evaluate the performance of our algorithm by using the mean and standard deviation of the estimated change point locations relative to the number of observations as well as the boxplots for the estimated change point for each case. We use estimated rank, sensitivity (SEN), specificity (SPC), and relative error (RE) for the whole transition matrices and the low-rank and the sparse components as additional metrics to evaluate the performance of model.

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

For multiple change points settings, we also measure the *selection rate*. Specifically, a detected change point $\widehat{t}_j$ is counted as a *success* for the true change point $t_j^\star$, if and only if $\widehat{t}_j \in [t_j^\star - \frac{1}{10}(t_j^\star - t_{j-1}^\star), t_j^\star + \frac{1}{10}(t_{j+1}^\star - t_j^\star)]$. Then, the selection rate is defined by calculating the percentage of simulation replications with successes.

3.4.1 Guidelines for Applying the Methods to Data and Tuning Parameters Selection

Recall that the information ratio plays a key role in the identifiability of the change points under both the full and the surrogate model. Since the information ratio is unknown in practice, it becomes unclear whether the surrogate model is capable of detecting the underlying change points, even when the full model clearly can. However, the computational savings of the former make it an attractive candidate for a first pass at obtaining candidate change points. To that end, we outline below a strategy for deciding on the question of applicability of the surrogate model.

Step 1: Use the surrogate model and obtain candidate change points (after applying the screening step): $\widetilde{\tau}_1, \widetilde{\tau}_2, \ldots, \widetilde{\tau}_m$.

Step 2: Let $\widetilde{\mathcal{I}}_j \stackrel{\text{def}}{=} |\widetilde{\tau}_{j+1} - \widetilde{\tau}_j|$, for $j = 0, 1, \ldots, m$, where $\widetilde{\tau}_0 = 1$ and $\widetilde{\tau}_{m+1} = T$. Then, apply the full model on each selected time segment $\widetilde{\Delta}_j$. Suppose in the j-th segment $\widetilde{\mathcal{I}}_j$, we have estimated change points $\widehat{\tau}_1^{(j)}, \ldots, \widehat{\tau}_{T_j}^{(j)}$ obtained by the full L+S model, where $T_j = |\widetilde{\tau}_{j+1} - \widetilde{\tau}_j|$. Then the final estimated change points set is given by:

$$\left(\bigcup_{j=0}^{m} \{\widehat{\tau}_1^{(j)}, \ldots, \widehat{\tau}_{T_j}^{(j)}\} \right) \cup \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_m\}.$$

Next, we discuss how the following tuning parameters are selected.

- For the full low-rank plus sparse model, we need to select the tuning parameters λ_j and μ_j;
- For the surrogate weakly sparse model, the tuning parameter η_j needs to be selected;
- For the selection step for candidate change points, a proper window size is the key factor impacting the accuracy and speed of the algorithm. Further, the screening step requires the penalization parameter ω_n to be specified.

Our recommendations are summarized next:

(λ_j, μ_j) : We use the same selection procedure as discussed in Section 5 in the main manuscript; specifically, we use the theoretical values provided and select the constants c_0, c_0' instead. By using a grid search, we simultaneously pick these two tuning parameters for each specified segment.

η_j : According to Negahban et al. (2012), we adopt the theoretical assumption on η_j: $\eta_j \propto \xi_j$, where ξ_j is the corresponding lasso penalization parameter selected by `glmnet` and `sparsevar`. We typically choose $\eta_j \in [0.01\xi_j, 0.1\xi_j]$.

α_L : In practice, we choose α_L based on the goal of the application. We empirically choose α_L based on the theoretical value $cp\sqrt{\frac{\log(pT)}{T}}$, and choose the constant $c \in [0.1, 1]$ in order to obtain a satisfactory estimation of the sparse components.

h : For window-size h, a feasible selection method is introduced in Section 5.

ω_n : The idea is to first finish the backward elimination algorithm (BEA) until no break points are left. Then, we cluster the *jumps* in the objective function $\mathcal{L}_T$ into two subgroups, small and large. Intuitively, if removing a break point leads to a small jump in $\mathcal{L}_T$, then the break point is likely redundant. In contrast, larger jumps correspond to true break points. The smallest jump in the second group is thus a reasonable candidate for ω_n. The proposed algorithm is summarized as follow:

(i) Apply the BEA algorithm to the set $\widetilde{\mathcal{S}}$ until no break points are left. Denote the ordered deleted break points as $\widetilde{t}_{i_1}, \widetilde{t}_{i_2}, \ldots, \widetilde{t}_{i_{\widetilde{m}}}$.

(ii) For each $k = 1, 2, \ldots, \widetilde{m}$, set $v_k = \left|\mathcal{L}_T(\widetilde{t}_{i_k}, \ldots, \widetilde{t}_{i_{\widetilde{m}}}; \boldsymbol{\lambda}, \boldsymbol{\mu}) - \mathcal{L}_T(\widetilde{t}_{i_{k-1}}, \ldots, \widetilde{t}_{i_{\widetilde{m}}}; \boldsymbol{\lambda}, \boldsymbol{\mu})\right|$. Define $V = \{v_1, v_2, \ldots, v_{\widetilde{m}}\}$.

(iii) Apply k-means clustering algorithm (Hartigan & Wong (1979)) to the set V with two centers. Denote the subset with smaller center as the small subgroup, V_S, and the other subset as the large subgroup, V_L.

(iv) (a) If (between-group SS/total SS) in (iii) is high, set $\omega_n = \min V_L$.
(b) If (between-group SS/total SS) in (iii) is low, set $\omega_n = \max V$.

3.4.2 Performance for Detecting Single Change Point

We investigate the following factors: the dimension of the model p, the sample size T, the differences in the ℓ_2 norm, v_L and v_S of the two low-rank and sparse components, respectively and the information ratio γ. The following parameters settings are considered in our investigation.

(A) In the first setting, we consider the case that the low-rank component exhibits a very small change while the sparse one a large change. Further, the "total signal" in the transition matrix comes mostly from the sparse component and therefore, $\gamma_j < 1, j = 1, 2$.

(B) This setting is similar in structure to A: the low-rank components exhibit very small change, while the sparse components change by a significant amount, but the "total signal" in the transition matrix comes mostly from the former; i.e., $\gamma_j \geq 1$ for $j = 1, 2$.

(C) The structure of this setting is as in B, but different values of γ_j are considered.

(D) This setting is the reverse of B, wherein the low-rank components exhibit a large change, while the sparse ones a very small ones, and further $\gamma_j \geq 1, j = 1, 2$.

(E) This setting is similar in structure to C, but the information ratio $\gamma_j < 1, j = 1, 2$.

(F) The setting is similar to E, but an increasing $|\gamma_1 - \gamma_2|$ is considered.

The following Table 3-1 fully summarizes all the parameter settings for all scenarios introduced previously.

Table 3-1. Model parameters for different settings considered.

Case	p	T	$\tau^\star/T$	$(r_1^\star, r_2^\star)$	v_L	v_S	(γ_1, γ_2)
A.1	20	300	0.500	$(1, 3)$	0.10	1.5	$(0.25, 0.25)$
A.2	20	300	0.500	$(1, 3)$	0.25	1.5	$(0.25, 0.25)$
A.3	20	300	0.500	$(1, 3)$	0.50	1.5	$(0.25, 0.25)$
B.1	20	300	0.500	$(1, 2)$	0.25	2.0	$(2.0, 2.0)$
B.2	20	300	0.500	$(1, 2)$	0.50	2.0	$(2.0, 2.0)$
B.3	20	300	0.500	$(1, 2)$	0.75	2.0	$(2.0, 2.0)$
C.1	20	300	0.500	$(1, 2)$	0.25	2.0	$(1.75, 2.0)$
C.2	20	300	0.500	$(1, 2)$	0.25	2.0	$(1.25, 2.0)$
C.3	20	300	0.500	$(1, 2)$	0.25	2.0	$(1.0, 2.0)$
C.4	20	300	0.500	$(1, 2)$	0.25	2.0	$(0.5, 2.0)$
D.1	20	300	0.500	$(1, 2)$	3.0	0.75	$(1.5, 1.5)$
D.2	20	300	0.500	$(1, 2)$	3.5	0.75	$(1.5, 1.5)$
D.3	20	300	0.500	$(1, 2)$	4.0	0.75	$(1.5, 1.5)$
E.1	20	300	0.500	$(1, 3)$	2.5	0.15	$(0.25, 0.25)$
E.2	20	300	0.500	$(1, 3)$	3.0	0.15	$(0.25, 0.25)$
E.3	20	300	0.500	$(1, 3)$	4.5	0.15	$(0.25, 0.25)$
F.1	20	300	0.500	$(1, 2)$	2.5	0.25	$(0.5, 0.45)$
F.2	20	300	0.500	$(1, 2)$	2.5	0.25	$(0.5, 0.75)$
F.3	20	300	0.500	$(1, 2)$	2.5	0.25	$(0.5, 0.95)$

The results for these settings over 50 replications are given in Table 3-2. The first two columns record the mean and standard deviation of the estimated change point location, the third and fourth columns are the estimated ranks for the low-rank components, the fifth and sixth columns give the sensitivity and specificity of the estimated sparse components, and finally the last column shows the relative norm error of the estimated transition matrix $\widehat{A}$ to the truth $A^\star$, and we also provide the relative error of the estimated sparse components (low rank components) $\widehat{S}$ (or $\widehat{L}$) to the truth $S^\star$ (or $L^\star$).

For settings A and D, where the dominant components change significantly, the algorithm identifies the change point extremely accurately, as evidenced by the mean

Table 3-2. Performance of the L+S model under different simulation settings.

	mean	sd	$\widehat{r}_1$	$\widehat{r}_2$	SEN	SPC	Total RE/ Sparse RE / Low-rank RE
A.1	0.498	0.002	1.020	2.900	(1.000, 1.000)	(0.909, 0.976)	(0.186, 0.237)/(0.172, 0.220)/(0.582, 0.648)
A.2	0.499	0.002	1.020	2.820	(1.000, 1.000)	(0.910, 0.974)	(0.186, 0.241)/(0.172, 0.217)/(0.582, 0.759)
A.3	0.499	0.002	1.020	2.960	(1.000, 1.000)	(0.909, 0.979)	(0.186, 0.249)/(0.172, 0.225)/(0.582, 0.749)
B.1	0.530	0.090	1.000	1.340	(0.166, 0.108)	(0.947, 0.980)	(0.590, 0.579)/(1.140, 1.006)/(0.482, 0.413)
B.2	0.532	0.089	1.000	1.340	(0.166, 0.109)	(0.947, 0.979)	(0.590, 0.580)/(1.139, 1.006)/(0.482, 0.414)
B.3	0.534	0.089	1.000	1.330	(0.165, 0.109)	(0.947, 0.980)	(0.591, 0.580)/(1.140, 1.006)/(0.482, 0.413)
C.1	0.522	0.056	1.000	1.350	(0.237, 0.103)	(0.944, 0.978)	(0.592, 0.569)/(1.070, 1.015)/(0.459, 0.384)
C.2	0.497	0.005	1.000	1.300	(0.400, 0.120)	(0.948, 0.979)	(0.645, 0.575)/(0.953, 1.006)/(0.482, 0.397)
C.3	0.502	0.031	1.000	1.320	(0.629, 0.109)	(0.947, 0.978)	(0.646, 0.570)/(0.858, 1.007)/(0.499, 0.389)
C.4	0.497	0.005	1.000	1.300	(1.000, 0.132)	(0.927, 0.977)	(0.357, 0.559)/(0.381, 1.002)/(0.499, 0.381)
D.1	0.494	0.011	1.000	1.500	(0.301, 0.207)	(0.948, 0.978)	(0.654, 0.581)/(1.036, 0.969)/(0.543, 0.455)
D.2	0.494	0.008	1.000	1.920	(0.305, 0.325)	(0.948, 0.975)	(0.654, 0.639)/(1.037, 0.934)/(0.544, 0.478)
D.3	0.495	0.007	1.000	2.080	(0.307, 0.485)	(0.948, 0.972)	(0.653, 0.558)/(1.031, 0.878)/(0.544, 0.444)
E.1	0.477	0.048	1.200	3.060	(1.000, 1.000)	(0.727, 0.739)	(0.171, 0.193)/(0.160, 0.176)/(0.563, 0.674)
E.2	0.478	0.026	1.000	3.040	(1.000, 1.000)	(0.836, 0.932)	(0.185, 0.216)/(0.168, 0.191)/(0.673, 0.633)
E.3	0.496	0.015	1.000	3.000	(1.000, 1.000)	(0.917, 0.729)	(0.204, 0.254)/(0.180, 0.250)/(0.674, 0.776)
F.1	0.495	0.053	1.000	2.880	(1.000, 1.000)	(0.924, 0.958)	(0.405, 0.330)/(0.429, 0.330)/(0.603, 0.482)
F.2	0.487	0.039	1.000	3.520	(1.000, 0.996)	(0.925, 0.964)	(0.411, 0.415)/(0.437, 0.486)/(0.602, 0.429)
F.3	0.495	0.023	1.000	2.640	(1.000, 0.895)	(0.924, 0.970)	(0.405, 0.539)/(0.429, 0.688)/(0.602, 0.484)

estimate over 50 replicates and the very small standard deviation recorded. Further, the ranks of L_j are accurately estimated under setting A, and the specificity and sensitivity of S_j is close to 1. Under setting D, there is deterioration in the estimation of the rank of L_2, as well as in the sensitivity of both S_1 and S_2. In settings B and E, where there is a small change in the dominant component, the estimates of the change point deteriorate and also exhibit larger variability (especially in setting B). Under setting B, estimation of the rank of L_2 is also off, as is the sensitivity for the sparse components. Note that all estimated model parameters under setting E are very accurate, with a small deterioration in the specificity of the S_j's. In settings C and F, we examine how the behavior of the information ratio influences the accuracy of the change point detection. As the difference between γ_1 and γ_2 increases, the estimation accuracy improves of the change point improves markedly. The same happens for the model parameters under setting F. Note that the results for settings C and F are in accordance with Remark 3-1 that discusses how the detectability of the full transition matrix is controlled by the information ratio. We provide the performance of single change point detection based on the surrogate model in Table 4 in Appendix F.1.

Figure 3-4 depicts boxplots based on 50 replicates of the distance between the

location of the true change point and its estimate, i.e., $|\widehat{\tau} - \tau^\star|$. The yellow bars correspond to the full low-rank plus sparse model, while the orange ones to the surrogate model. In accordance to previous findings, under settings A and C, the results are comparable, as well as certain cases for setting E. On the other hand, under settings B, D and F, the full model clearly outperforms the surrogate one, even though in settings F2 and F3 the differences become smaller as the corresponding differences in the information ratios increase.

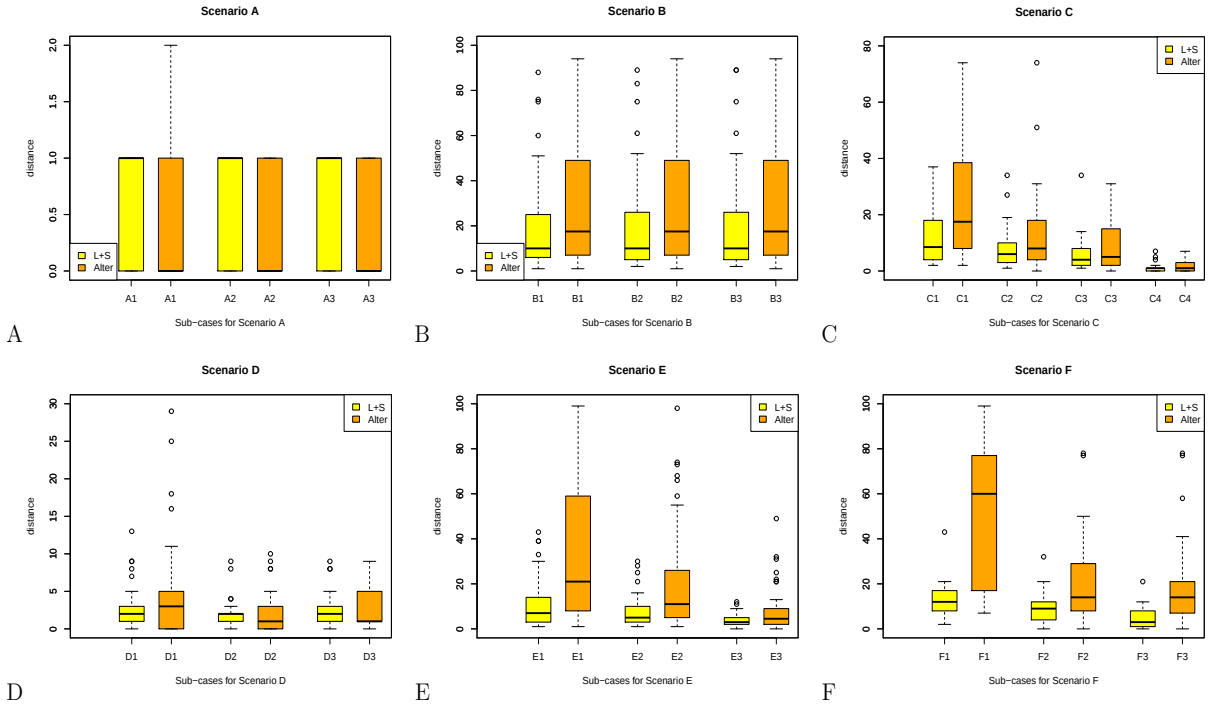


Figure 3-4. Boxplots for $|\widehat{\tau} - \tau^\star|$ under settings A–F with the full model and the surrogate weakly sparse model.

Next, Table 3-3 presents the extra settings for scenario G, which considers high-dimensional cases.

Table 3-3. Model parameters for different settings considered.

Case	p	T	$\tau^\star/T$	$(r_1^\star, r_2^\star)$	v_L	v_S	(γ_1, γ_2)
G.3	80	200	0.200	(1,3)	0.20	0.75	(0.25, 0.25)
G.4	80	200	0.800	(1,3)	0.20	0.75	(0.25, 0.25)
G.5	80	200	0.500	(3,1)	0.20	0.75	(0.25, 0.25)
G.6	80	200	0.500	(3,3)	0.20	0.75	(0.25, 0.25)
G.7	80	200	0.500	(5,3)	0.20	0.75	(0.25, 0.25)
G.8	50	200	0.500	(1,3)	0.45	0.40	(0.75, 0.75)

The results of those settings are presented in Table 3-4. We can easily observe that the accuracy of the estimated change points, as well as the transition matrices are satisfactory under the high dimensional setting. The estimated model parameters are also highly satisfactory based on the sensitivity and specificity metrics.

Table 3-4. Performance for simulation setting G.

Case	mean	sd	$\widehat{r}_1$	$\widehat{r}_2$	SEN	SPC	Total RE/ Sparse RE / Low-rank RE
G.3	0.203	0.016	1.051	2.773	$(0.887, 0.945)$	$(0.985, 0.957)$	$(0.782, 0.635)/(0.692, 0.503)/(0.897, 0.704)$
G.4	0.822	0.035	1.000	1.885	$(0.967, 0.918)$	$(0.955, 0.932)$	$(0.603, 0.745)/(0.531, 0.691)/(0.688, 0.880)$
G.5	0.515	0.056	2.750	1.050	$(0.965, 0.955)$	$(0.928, 0.987)$	$(0.606, 0.775)/(0.608, 0.742)/(0.776, 0.763)$
G.6	0.514	0.056	3.000	3.250	$(0.963, 0.977)$	$(0.926, 0.979)$	$(0.607, 0.853)/(0.608, 0.771)/(0.776, 0.820)$
G.7	0.502	0.006	5.900	4.025	$(0.991, 0.987)$	$(0.928, 0.939)$	$(0.575, 0.850)/(0.567, 0.669)/(0.826, 0.969)$
G.8	0.539	0.031	0.855	2.335	$(0.925, 0.902)$	$(0.865, 0.899)$	$(1.002, 0.975)/(1.200, 1.004)/(0.827, 0.927)$

3.4.3 Performance of the Surrogate Model for Single Change Point Detection

Table 3-5 summarizes the results of the surrogate weakly sparse model. Analogously to the results for the low-rank plus sparse model, under settings A and D the estimates of the change point are highly accurate. In settings B and E, the surrogate model performs worse than the full model, since the difference in the norm of the transitions matrices is rather small. Specifically, under setting B, the low-rank components contribute most of the "signal", even though their changes before and after the change point are rather small, thus effectively not satisfying Assumption W1. A similar reasoning justifies the rather poor performance of the surrogate model under setting E. In settings C and F, we investigate the case of different information ratios, covered in the second part of assumption W1. It can be seen that performance gradually improves by enlarging the differences between the information ratios. Estimation of the transition matrices is analogous to that under the full model; when the sparse component contributes most of the "signal" as in settings A, E and F, the relative error of is good and comparable to that of the full model. On the other hand, the relative error becomes worse than that obtained by the full model.

3.4.4 Performance for Detecting Multiple Change Points

We consider the same settings for each change point, as in case A in Section 3.4.2 with modified T and p, respectively. The specific scenarios under consideration are as follows:

Table 3-5. Performance of the surrogate model under different simulation settings.

Case	mean	sd	RE	Case	mean	sd	RE
A.1	0.498	0.002	(0.188, 0.201)	D.1	0.502	0.025	(0.766, 0.829)
A.2	0.498	0.002	(0.190, 0.200)	D.2	0.498	0.012	(0.763, 0.743)
A.3	0.498	0.002	(0.190, 0.206)	D.3	0.498	0.011	(0.762, 0.691)
B.1	0.538	0.125	(0.788, 0.814)	E.1	0.525	0.154	(0.199, 0.234)
B.2	0.538	0.125	(0.787, 0.815)	E.2	0.510	0.104	(0.200, 0.248)
B.3	0.539	0.124	(0.787, 0.814)	E.3	0.518	0.060	(0.198, 0.285)
C.1	0.550	0.112	(0.765, 0.814)	F.1	0.456	0.200	(0.431, 0.352)
C.2	0.515	0.076	(0.737, 0.798)	F.2	0.470	0.098	(0.411, 0.458)
C.3	0.494	0.041	(0.682, 0.783)	F.3	0.475	0.095	(0.415, 0.571)
C.4	0.501	0.008	(0.370, 0.775)				

(L) In the first case, we consider settings with different number of change points. Specifically, we investigate the following three cases: (1) $T = 1200$ with $\tau_1^\star = \lfloor T/6 \rfloor$, $\tau_2^\star = \lfloor T/3 \rfloor$, $\tau_3^\star = \lfloor T/2 \rfloor$, $\tau_4^\star = \lfloor 2T/3 \rfloor$, and $\tau_5^\star = \lfloor 5T/6 \rfloor$; (2) $T = 1800$ with $\tau_1^\star = \lfloor T/10 \rfloor$, $\tau_2^\star = \lfloor 3T/10 \rfloor$, $\tau_3^\star = \lfloor T/2 \rfloor$, $\tau_4^\star = \lfloor 7T/10 \rfloor$, and $\tau_9^\star = \lfloor 9T/10 \rfloor$; (3) $T = 2400$ with $\tau_1^\star = \lfloor T/10 \rfloor$, $\tau_2^\star = \lfloor T/4 \rfloor$, $\tau_3^\star = \lfloor 2T/5 \rfloor$, $\tau_4^\star = \lfloor 3T/5 \rfloor$, and $\tau_5^\star = \lfloor 4T/5 \rfloor$.

(M) In the second case, we consider p large enough to satisfy $p^2 > T$ with two change points: $\tau_1^\star = \lfloor T/3 \rfloor$ and $\tau_2^\star = \lfloor 2T/3 \rfloor$.

(N) In the last scenario, the change in sparsity patterns is considered. We consider a different sparsity pattern rather than the 1-off diagonal structure in the sparse components.

The detailed model parameters are listed in the Table 5 in the Appendix F.2. Table 3-6 presents the mean and standard deviation of the estimated locations of the change points, relative to the sample size T, together with the selection rate, as defined at the beginning of the current section. For all cases under settings L and M, the two-step algorithm obtains very accurate results, also exhibiting little variability. The complex random sparse pattern considered in setting N leads to a small deterioration in the selection rate. The locations of the estimated change points together with box plots of $|\widehat{\tau}_j - \tau_j^\star|$ for scenario N over 50 replicates are depicted in the 3-5.

3.4.5 A Comparison of Run Times between the Full and the Surrogate Model

We undertake such a comparison for settings A, C and D presented in Section 3.4.2. The results averaged over 50 replicates indicate that Algorithm 1 for the full model takes

Table 3-6. Results for multiple change point selection by full L+S model.

Case	points	truth	mean	sd	selection rate	Case	points	truth	mean	sd	selection rate
	1	0.1667	0.1667	0.0004	1.00	M.1	1	0.3333	0.3331	0.0005	1.00
	2	0.3333	0.3333	0.0003	1.00		2	0.6667	0.6665	0.0004	1.00
L.1	3	0.5000	0.4999	0.0003	1.00	M.2	1	0.3333	0.3329	0.0003	1.00
	4	0.6667	0.6665	0.0004	1.00		2	0.6667	0.6667	0.0006	1.00
	5	0.8333	0.8335	0.0004	1.00	N.1	1	0.3333	0.3311	0.0125	0.94
	1	0.1000	0.0999	0.0002	1.00		2	0.6667	0.6656	0.0056	0.98
	2	0.2500	0.2500	0.0000	1.00	N.2	1	0.1667	0.1683	0.0115	0.92
L.2	3	0.4000	0.3999	0.0002	1.00		2	0.8333	0.8267	0.0181	0.94
	4	0.6000	0.6000	0.0000	1.00	N.3	1	0.3333	0.3302	0.0121	0.98
	5	0.8000	0.7999	0.0001	1.00		2	0.6667	0.6655	0.0119	0.98
	1	0.1000	0.1000	0.0000	1.00						
	2	0.3000	0.3000	0.0000	1.00						
L.3	3	0.5000	0.5000	0.0000	1.00						
	4	0.7000	0.6999	0.0002	1.00						
	5	0.9000	0.8998	0.0002	1.00						

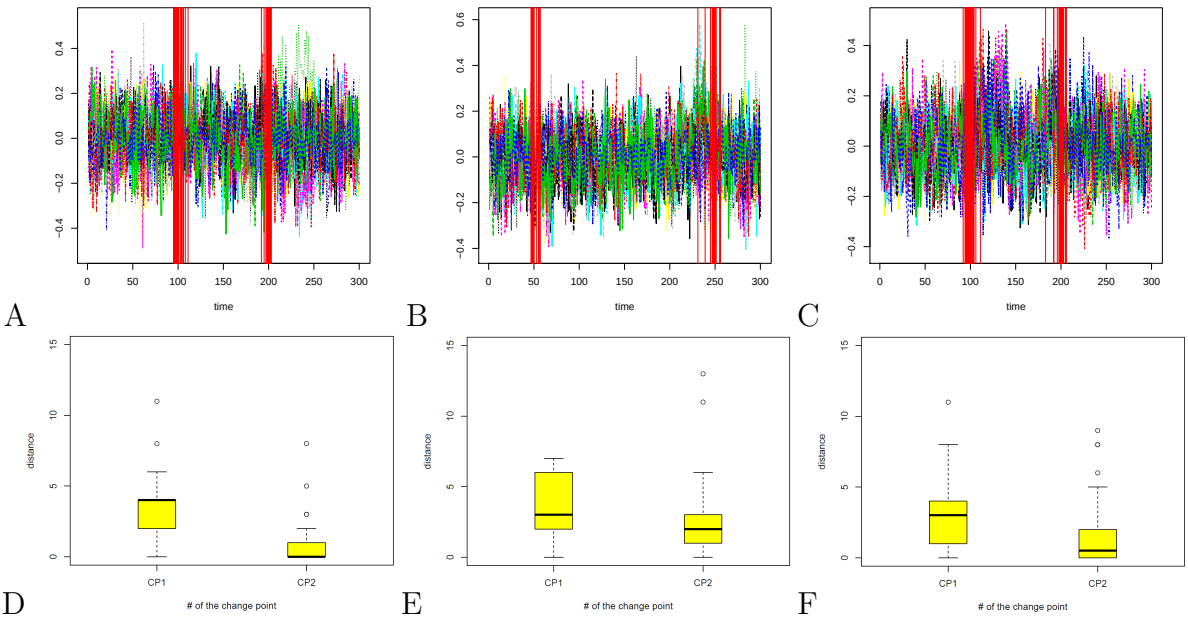


Figure 3-5. Final selected change points (red lines) by using two-step algorithm and boxplots for $|\widehat{t} - t^{\star}|$ under different scenario N settings.

approximately 900 secs per replicate, while the surrogate model less than 200 secs. For the two-step Algorithm 2, the average run time for the full model takes approximately 3.5 hours per replicate, while that for the surrogate model approximately 20 minutes per replicate. The results are plotted in the following Figure 3-6. The high computational cost of the exhaustive search procedure for the full model is apparent and is due to performing

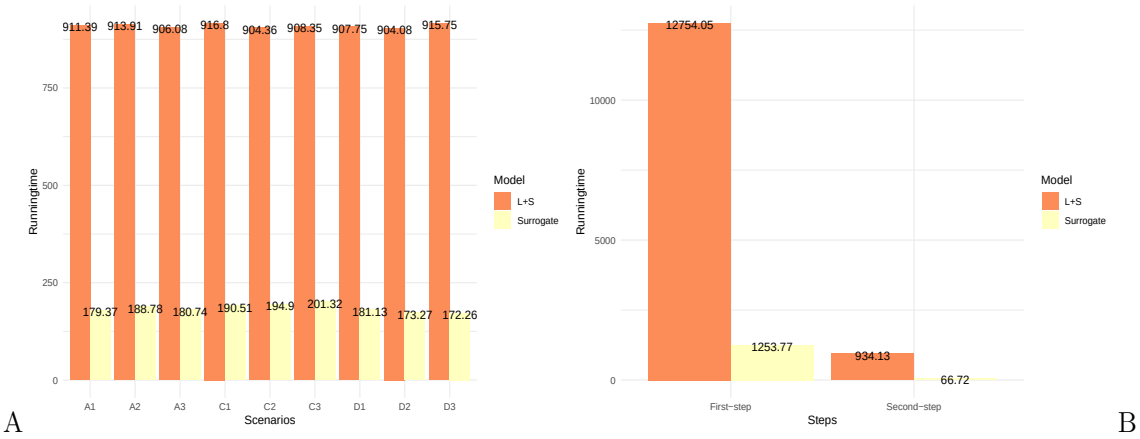


Figure 3-6. Comparison the run times for the full low-rank plus sparse and alternative weakly sparse models.

multiple SVDs, while the surrogate model provides significant computational savings and hence justify its use when suitable, based the theoretical developments and guarantees presented in Section 3.3.2.

3.4.6 Comparisons between the LR+Sparse VAR Model and Factor Model

3.4.6.1 A comparison with the factor-based model under scenarios L.1 and L.2

In Barigozzi et al. (2018), the authors investigate detection of multiple change points in a static factor model. Note that a factor model assumes that the data exhibit low rank structure in the contemporaneous dependence (correlation) structure, as opposed to their lead-lag (autocorrelation) structure as in a VAR model. Next, we provide results for scenarios L.1 and L.2 in Table 3-7, respectively.

Table 3-7. Results for change point selection by low rank plus sparse VAR model and a factor-based model.

Case	Model	points	truth	mean	sd	selection rate	Model	points	truth	mean	sd	selection rate
		1	0.1667	0.1667	0.0004	1.00		1	0.1667	0.1657	0.0061	0.78
		2	0.3333	0.3333	0.0003	1.00		2	0.3333	0.3340	0.0057	0.44
L.1	L+S VAR model	3	0.5000	0.4999	0.0003	1.00	Factor model	3	0.5000	0.5005	0.0065	0.56
		4	0.6667	0.6665	0.0004	1.00		4	0.6667	0.6706	0.0065	0.46
		5	0.8333	0.8335	0.0004	1.00		5	0.8333	0.8361	0.0073	0.70
		1	0.1000	0.0999	0.0002	1.00		1	0.1000	0.1000	0.0046	0.94
		2	0.2500	0.2500	0.0000	1.00		2	0.2500	0.2528	0.0041	0.84
L.2	L+S VAR model	3	0.4000	0.3999	0.0002	1.00	Factor model	3	0.4000	0.4011	0.0040	0.94
		4	0.6000	0.6000	0.0000	1.00		4	0.6000	0.6038	0.0028	0.70
		5	0.8000	0.7999	0.0001	1.00		5	0.8000	0.8020	0.0029	0.94

It can be seen that the accuracy of the location of the detected change points by the

factor-based model is high; nevertheless, the selection rate (# of times that it correctly identifies the right number of change points) is significantly lower than that of the VAR model. To illustrate further the latter point, the mean/median Hausdorff distance between the estimated change points set $\widetilde{\mathcal{S}}$ and the true change points set $\mathcal{S}^\star$ is tabulated in Table 3-8. The result is not particularly surprising, since the true data generating mechanism is according to the posited low rank plus sparse VAR model.

Table 3-8. Hausdorff distance $d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star)$ comparison with factor change point model.

Case	Model	mean$(d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star))$	std$(d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star))$	median$(d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star))$
L.1	Our model	1.46	2.6358	0.00
	Factor model	27.26	18.6808	19.00
L.2	Our model	1.45	2.3774	0.00
	Factor model	33.34	23.4082	30.00

3.4.6.2 A comparison with the factor-based model under a dynamical factor model (DFM) generating mechanism

For a further comparison between the detection strategy based on the factor model (Barigozzi et al. 2018) and the two step rolling window strategy based on the posited VAR model, we employed a *dynamical factor model* to generate the data. Hence, the data are generated according to:

$$X_t = \Lambda_j F_t + e_t, \quad F_t = \Psi F_{t-1} + \epsilon_t,$$

where $\Lambda_j, j = 1, 2, \ldots, m_0 + 1$ are factor loadings and Ψ is a diagonal transition matrix of the latent process $\{F_t\}$ with independent and identically distributed standard normal entries. Both error terms e_t and ϵ_t are independent and identically normally distributed with mean zero and variance $0.01\mathbf{I}$. The dimension of the time series is set to $p = 20$, the sample size to $T = 300$, the locations of the change points at $t_1^\star = 100$ and $t_2^\star = 200$, and the dimension of the latent factor process to $r = 5$. The loadings matrices Λ_j are generated at random, with varying magnitudes across different stationary segments.

Table 3-9 tabulates the mean and standard deviation of the relative location of the detected change points, as well as the selection rate for the two-step procedure for the VAR model and the procedure based on the factor model over 50 replications. It can be seen

Table 3-9. Comparison of the two-step strategy for the VAR model and the strategy based on factor model under a DFM data generating mechanism.

Method	CP	Truth	Mean	Sd	Selection rate
Two-step strategy for a L+S VAR model	1	0.333	0.342	0.050	0.80
	2	0.667	0.647	0.041	0.66
Factor model based strategy (Barigozzi et al. (2018))	1	0.333	0.232	0.107	0.06
	2	0.667	0.801	0.118	0.10

that the two-step strategy based on the posited low-rank plus sparse VAR model exhibits a significantly higher selection rate than the strategy based on the static factor model. Further, the former provides much more accurate estimates of the locations of the underlying change points.

Note that both models *misspecify* the true data generating mechanism. The factor model assumes a static factor structure (no autoregressive dynamics in the latent factor), whereas the VAR model assumes autoregressive dynamics on the observed data. The inferior performance of the strategy based on the factor model may be due to the detection mechanism used in Barigozzi et al. (2018), which first extracts principal components of the data across the whole observation interval and then leverages a binary segmentation algorithm to identify the change points.

3.4.7 A Comparison of the Two-step Strategy with a Dynamic Programming Algorithm

Next, prompted by a comment from a reviewer, we investigate a popular and generally applicable strategy for detecting multiple change points, namely one based on a dynamic program (see Wang et al. (2019) for a similar algorithm for detecting multiple change points in a *pure sparse* VAR model). The algorithmic details are given in Algorithm 4. As is well-known, the time complexity of a dynamic programming algorithm is $\mathcal{O}(T^2)$ namely quadratic in the number of time points. Contrary, as mentioned earlier the time complexity of the two step rolling window strategy is *linear* in the number of time points.

The setting under consideration is as follows: the data are generated according to the posited low rank plus sparse VAR model with $p = 20$, $T = 240$ and two change points located at $t_1^\star = 80$ and $t_2^\star = 160$. The ranks for each stationary segment remain the same

$r_j \equiv 1$ and the jump sizes are set to $v_L = 0.1$ and $v_S = 1.5$. The information ratio $\gamma_j = 0.25$ for all three stationary segments.

Algorithm 4 Penalized Dynamic Programming Algorithm.

1. **Input:** Time series data $\{X_t\}$, $t = 1, 2, \ldots, T$, the tuning parameter γ, the mark for change point detection *Flag* is set as false, an empty set to store estimated change points $\mathcal{C}$.
2. **While** $e < T - 1$ **do:**
   - $s \leftarrow s + 1$
   - **While** $t < T$ and *Flag* is false **do:**
     * $t \leftarrow t + 1$
     * **If** $\min_{s+1 \le l \le t-1} \{\mathcal{L}(s, l) + \mathcal{L}(l + 1, t) + \gamma\} < \mathcal{L}(s, t)$, **then:**
       · $s \leftarrow \min \arg\min_{s+1 \le l \le t-1} \{\mathcal{L}(s, l) + \mathcal{L}(l + 1, t) + \gamma\}$
       · $\mathcal{C} \leftarrow \mathcal{C} \cup \{s\}$
       · Set *Flag* as true
   - Set *Flag* as false
3. **Output:** The set of estimated change points $\mathcal{C}$.

The results are presented in Table 3-10 based on the following evaluation metrics: (1) the *running time* for the low rank plus sparse model and the surrogate weakly sparse model using the two step rolling window strategy, and the low rank plus sparse model using the DP algorithm; (2) the *number of estimated change points*; and (3) the mean and standard deviation of the *estimated change points*.

Table 3-10. Comparison of Proposed Two-step Algorithm with DP Algorithm.

	Model	Case (DP)
Running time (sec)	Two-step L+S VAR	942.11
	Two-step Surrogate	193.20
	DP L+S VAR	1471.72
No. of estimated change points	Two-step L+S VAR	2
	Two-step Surrogate	2
	DP L+S VAR	2
Estimated change points: $\widehat{t}_j/T$	Two-step L+S VAR	$(0.3334_{0.0005}, 0.6665_{0.0003})$
	Two Step Surrogate	$(0.3294_{0.0025}, 0.6708_{0.0103})$
	DP L+S VAR	$(0.3332_{0.0001}, 0.6665_{0.0000})$

3.4.8 A Comparison between the Two-step Algorithm and the Dynamic Programming Algorithm for the Low Rank plus Sparse VAR model.

Based on the procedure in dynamical programming algorithm, we implement the algorithm and set up an extra simulation setting O as follows:

(O) In this case, we set $p = 20$, $T = 300$ with two change points $t_1^\star = 100$ and $t_2^\star = 200$. The rank in each segment remains the same, while the jump sizes are $v_L = 0.1$ and $v_S = 1.5$. The information ratio $\gamma_j = 0.25$ for the $j = 1, 2, 3$ segments.

The results are presented in Table 3-11. It can be seen that the running time for the DP algorithm is 1.5 times longer than the two-step rolling window algorithm. Further, all three algorithms identify correctly both change points and their locations. However, the proposed two-step algorithm achieves this task faster. Further, the proposed two-step strategy using the surrogate model is much faster (about 10% of the computation time of the two-step approach with the full model and 6.5% of the computation time of the DP algorithm), but as explained in Section 4 of the main manuscript, it requires more stringent conditions.

Table 3-11. Comparison of Proposed Two-step Algorithm with Dynamical Programming Algorithm.

	Model	Case (DP)
Running time (sec)	Two-step	942.11
	Surrogate	193.20
	DP	1471.72
No. of estimated change points	Two-step	2
	Surrogate	2
	DP	2
Estimated change points: $\widehat{t}_j/T$	Two-step	$(0.3334_{0.0005}, 0.6665_{0.0003})$
	Surrogate	$(0.3294_{0.0025}, 0.6708_{0.0103})$
	DP	$(0.3332_{0.0001}, 0.6665_{0.0000})$

3.5 Applications

3.5.1 Change Point Detection in EEG Signals Data

There has been work in the literature on analyzing EEG data using low-rank models for task related signals, since the latter exhibit low-rank structure (Liu et al. 2018, Jao

et al. 2018). Next, we employ the full low-rank plus sparse model to detect change points in data from Trujillo et al. (2017). This data set recorded 72 channels of continuous EEG signals by using active electrodes. The sampling frequency is 256Hz and the total number of time points per EEG electrode is 122880 over 480 seconds. The stimulus procedure is that after a resting state (eliminated from the data set) lasting 8 mins, the subject alternates between a 1-min period with eyes open followed by a 1-min period with eyes closed, repeated four times. Hence, we expect that the employed model captures the low-rank structure associated with the task at hand (open/closed eyes), while the sparse component can capture idiosyncratic behavior across repetitions of the task.

To illustrate the proposed methodology, two subjects are selected; differences in the EEG signals over time are visible for the first subject, but not for the second one. The data are de-trended, by calculating the moving average of each EEG signal and removing it. Specifically, the period average, which is an unbiased estimator of trend, is given by $\hat{m}_l = \frac{1}{d}\sum_{t=1}^{d} X_{l+t}$; we select $d = 256$ in accordance to the frequency of the data, and we obtain the de-trended time series by removing the period average. In this work, we use 21 selected EEG channels and $T = 67952$ time points in the middle of the whole time series. According to the experiments described in Trujillo et al. (2017), there are five open/closed eyes segments in the selected time period with four change points approximately at locations: $\tau_1^\star \cong 11650$, $\tau_2^\star \cong 27750$, $\tau_3^\star \cong 44000$, and $\tau_4^\star \cong 60000$. The data are plotted in Figure 2 in Appendix G.2. Selection of the tuning parameters is based on the guidelines given in Appendix G.1. Note that to separate adequately the sparse component from the low-rank one, we set α_L based on its theoretical values provided in Assumption H2.

The change points estimated by the two-step algorithm are $\widehat{\tau}_1 = 9633$, $\widehat{\tau}_2 = 28529$, $\widehat{\tau}_3 = 43361$ and $\widehat{\tau}_4 = 60209$. The estimated change points are close to those identified based on the designed experiment. In order to quantify the differences among the estimated components across segments, we use the Hamming distance for both sparse and low-rank ones. The results are shown in Figure 3-7 in the form of a heat map that confirms the high

degree of similarity between all "eyes closed" segments (1, 3, 5) and all "eyes open" segments (2, 4), thus further confirming the accuracy of the methodology.

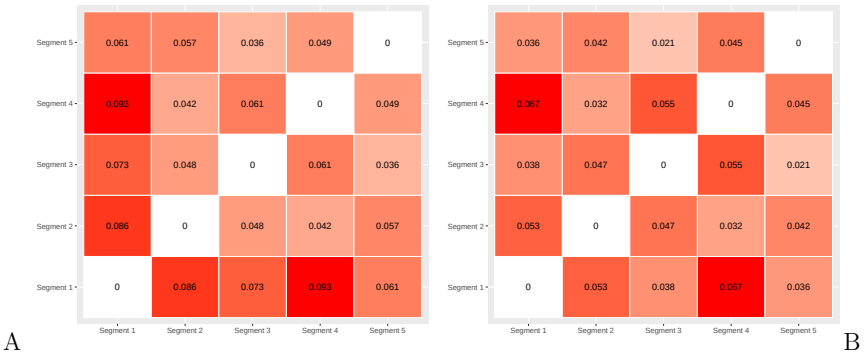


Figure 3-7. Hamming distance heat map among the estimated low-rank components (A) and sparse components (B).

3.5.2 An Application to Macroeconomics Data

We consider the macroeconomics data obtained from the FRED database McCracken & Ng (2016). This data set comprises of 19 key macroeconomic variables, corresponding to the "Medium" model analyzed in Bańbura et al. (2010) and covering the 1959–2019 period (723 observations). The original time series data are non-stationary and we de-trend them by taking first differences.

To select the tuning parameters (λ, μ), we employ a 2-dimensional grid search procedure. In our analysis, we set α_L based on its theoretical value in Assumption H2 to ensure identifiability of the sparse component from the low-rank one. The modified macroeconomics data together with all 6 estimated change points is provided in Figure 3-8, as well as the estimated change points are listed in Table 3-12.

Table 3-12. Estimated Change Points and Candidate Related Events.

Date (mm/dd/yyyy)	Candidate Related Events
02/01/1975	Aftermath of 1973 oil crisis
04/01/1977	Rapid build-up of inflation expectations
12/01/1980	Rapid increase of interest rates by the Volcker Fed
01/01/1994	Multiple events - see Appendix G.3
09/01/2008	Recession following collapse of Lehman Brothers
05/01/2010	Recovery from the Great Financial crisis of 2008

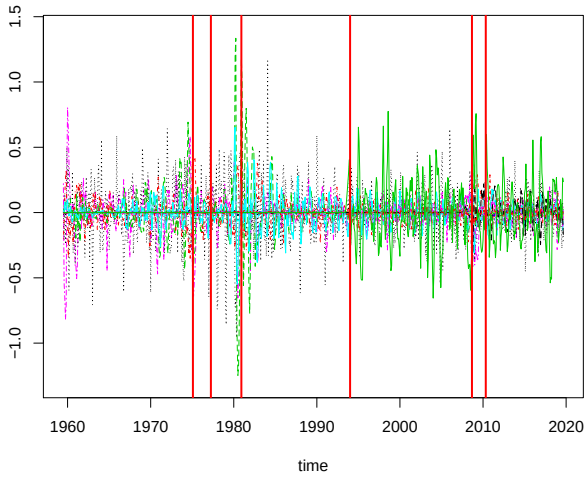


Figure 3-8. Macroeconomic indicators for the 1959-2019 period with all 6 estimated change points (red lines).

The first change point corresponds to the aftermath of the first oil crisis in 1973 and the collapse of the post-war Bretton Woods system of monetary management of commercial and financial relations among the leading western economics (Bordo & Eichengreen 2007), that led to low growth and sustained inflation. The second change point identified, marks the rapid build-up of inflation expectations (Kareken 1978) that led the Federal Reserve Board under Chairman Volcker to pursue a contractionary monetary policy through doubling the federal funds rate to 20% to fight-off persisting inflation expectations (Orphanides 2004). The next change point is associated with multiple events, including the Republican Party controlling the US House of Representatives for the first time since 1952 with a business and markets friendly agenda, and the ratification of the North American Free Trade Agreement. The last two change points are associated with the onset and exit of the Great Financial Crisis of 2008 that led to a deep recession, collapse and/or bailouts of various financial institutions, liquidity crunches and a debt crisis in peripheral countries in the Eurozone that exhibited a negative feedback to the US economy (Eichengreen 2014).

While the sparsity levels and ranks for each segment are plotted in Figure 3-9. The selected change points are presented in Figure 3-8.

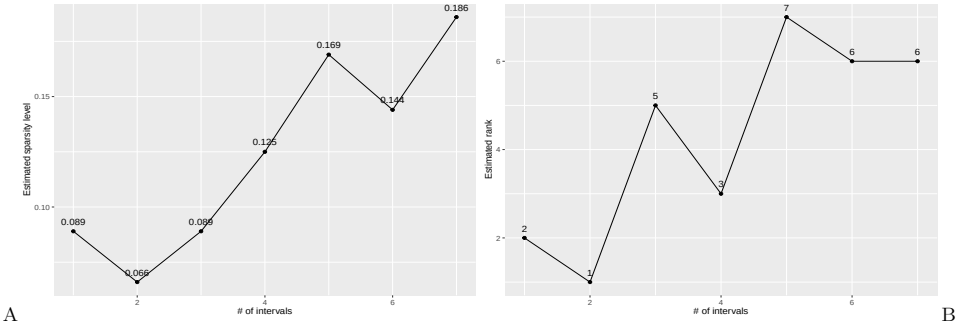


Figure 3-9. Estimated sparsity level (A) and estimated ranks (B) for each selected interval.

We also compare the results using the detection strategy based on the static factor model in Barigozzi et al. (2018). According to Fama & French (1996), we set the maximum number of factors to three and the estimated change points are listed in Table 3-13.

Table 3-13. Estimated Change Points by the Detection Strategy based on a Factor Model.

Date (mm/dd/yyyy)	Candidate Related Events
12/1/1979	Rapid increase of interest rates by the Volcker Fed
1/1/1985	Multiple events
11/1/1993	Multiple events
4/1/2008	Prequel to the Great Financial Crisis

The factor model misses important events, including the economic recovery following the Financial Crisis of 2008 and the recession following the first oil crisis of 1973. Further, it identifies a change point in early April of 2008, even though most of the macroeconomic (as opposed to financial market) indices started deteriorating in the summer of 2008 and tumbled in the 3rd quarter, following the collapse of Lehman Brothers in mid-September.

3.6 Concluding Remarks

The paper addressed the problem of multiple change point detection in reduced rank VAR models. The key innovation is the development of a two-step strategy that obtains consistent estimates of the change points and the model parameters. Other strategies for detecting multiple change points in high-dimensional models, such as fused penalties or binary segmentation type of procedures, either require very stringent conditions or are not directly applicable. Further, dynamic programming entails a quadratic computational cost in the number of time points compared to a linear cost for the proposed strategy. To

enhance computational efficiency, we introduced a surrogate weakly sparse model and identified sufficient conditions under which the aforementioned 2-step strategy detects change points in low-rank and sparse VAR models as accurately as using the correctly specified model, but at significant computational gains.

In the algorithmic and technical results presented, similar to the case of a sparse VAR model with change points (Wang et al. (2019)), we assume a simple structure on the error terms, i.e., in segment j, $\epsilon_t^j \sim \mathcal{N}(0, \sigma^2 I)$, where σ is a fixed constant independent of j. Such a simple structure on the covariance matrices of error terms ensures the identifiability of change points, since a change in the transition matrices would imply that the second order structure (the auto-correlation function) of the stochastic process before and after the change points have changed, thus the definition of change points becomes meaningful. It is of interest to investigate in future work a general covariance matrix Σ_E, or even segment specific ones Σ_E^j, including conditions that lead to changes in the segment specific auto-correlation function of the process.

Further, the proposed strategy is directly applicable to other forms of structured sparsity in the transition matrix of the VAR model, including low-rank plus structured sparse, or structured sparse plus sparse, as discussed for stationary models in Basu et al. (2019).

Finally, the presentation focused on a VAR model with a single lag, but both the modeling framework and the developed 2-step detection strategy can be extended to VAR(d) processes with $d > 1$ in a similar manner, as presented in Basu et al. (2019).

3.7 Auxiliary Lemmas for Chapter 3

Lemma 3-1. *Given a VAR(1) series $\{X_t\}$ and a time point s, for any true change point $\tau_j^\star$, if $|s - \tau_j^\star| \geq T\xi_T$, and $\tau_{j-1}^\star < s < \tau_j^\star$, there exist constants $c_i > 0$ such that with probability at least $1 - c_1 \exp(-c_2 \log p)$:*

$$\sup_{1 \leq j \leq m_0, |s-\tau_j^\star| \geq T\xi_T} \left\| (t_j^\star - s)^{-1} \left(\sum_{t=s}^{\tau_j^\star - 1} X_{t-1} X_{t-1}' - \Gamma_j(0) \right) \right\|_\infty \leq c_0 \sqrt{\frac{\log p}{\tau_j^\star - s}},$$

$$\sup_{1\leq j\leq m_0, |s-\tau_j^\star|\geq T\xi_T} \left\| (\tau_j^\star - s)^{-1} \sum_{t=s}^{\tau_j^\star - 1} X'_{t-1}\epsilon_t \right\|_\infty \leq c_0 \sqrt{\frac{\log p}{\tau_j^\star - s}}.$$

Similarly, there exist constants $c_i' > 0$, such that with probability at least $1 - c_1\exp(-c_2 p)$:

$$\sup_{1\leq j\leq m_0, |s-\tau_j^\star|\geq T\xi_T} \left\| (\tau_j^\star - s)^{-1} \left(\sum_{t=s}^{\tau_j^\star - 1} X_{t-1}X'_{t-1} - \Gamma_j(0) \right) \right\|_{op} \leq c_0 \sqrt{\frac{p}{\tau_j^\star - s}},$$

$$\sup_{1\leq j\leq m_0, |s-\tau_j^\star|\geq T\xi_T} \left\| (\tau_j^\star - s)^{-1} \sum_{t=s}^{\tau_j^\star - 1} X'_{t-1}\epsilon_t \right\|_{op} \leq c_0 \sqrt{\frac{p}{\tau_j^\star - s}}.$$

Proof of Lemma 3-1. The proof of the lemma follows along similar lines as that of Proposition 2.4 in Basu & Michailidis (2015), Proposition 3 in Basu et al. (2019) and Lemma 3 in Safikhani & Shojaie (2020) and thus is omitted. □

Consider the following two sets of subspaces $\{\mathcal{I}, \mathcal{I}^c\}$ and $\{A, B\}$ associated with some generic matrix $\Theta \in \mathbb{R}^{p\times p}$, in which the ℓ_1 norm and the nuclear norm are decomposable, respectively (Negahban et al. 2012). Specifically, let the singular value decomposition of Θ be $\Theta = U\Sigma V'$ with U and V being orthonormal matrices. Let $r = \text{rank}(\Theta)$, and U^r and V^r denote the first r columns of U and V associated with the first r singular values of Θ, respectively. Define:

$$A \stackrel{\text{def}}{=} \left\{ \Psi \in \mathbb{R}^{p\times p} : \text{row}(\Psi) \subseteq V^r \text{ and } \text{col}(\Psi) \subseteq U^r \right\},$$
$$B \stackrel{\text{def}}{=} \left\{ \Psi \in \mathbb{R}^{p\times p} : \text{row}(\Psi) \perp V^r \text{ and } \text{col}(\Psi) \perp U^r \right\}.$$

Let $\mathcal{J}$ be the set of indices in which Θ is nonzero. Analogously, we define

$$\mathcal{I} \stackrel{\text{def}}{=} \left\{ \Psi \in \mathbb{R}^{p\times p} : \Psi_{ij} = 0 \text{ for } (i,j) \notin \mathcal{J} \right\},$$
$$\mathcal{I}^c \stackrel{\text{def}}{=} \left\{ \Psi \in \mathbb{R}^{p\times p} : \Psi_{ij} = 0 \text{ for } (i,j) \in \mathcal{J} \right\}.$$

Therefore, we have $\|\Theta\|_* = \|\Theta\|_{*,A} + \|\Theta\|_{*,B}$ and $\|\Theta\|_1 = \|\Theta\|_{1,\mathcal{I}} + \|\Theta\|_{1,\mathcal{I}^c}$.

Lemma 3-2. *Define the error matrices $\widehat{\Delta}^L = \widehat{L} - L^\star$ and $\widehat{\Delta}^S = \widehat{S} - S^\star$ associated with any*

positive parameters λ, μ, and let the weighted regularizer $\mathcal{Q}$ be defined as:

$$\mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \stackrel{def}{=} \|\widehat{\Delta}^L\|_* + \frac{\lambda}{\mu}\|\widehat{\Delta}^S\|_1,$$

for the previously defined subspaces. Then, the following inequality holds:

$$\mathcal{Q}(L^\star, S^\star) - \mathcal{Q}(\widehat{L}, \widehat{S}) \leq \mathcal{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^L_{\mathcal{I}}) - \mathcal{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^L_{\mathcal{I}^c}).$$

Proof of Lemma 3-2. Based on the definition of the subspaces, we immediately get that $L^\star_B = 0$ and $S^\star_{\mathcal{I}^c} = 0$. Then, we get:

$$\begin{aligned}
\mathcal{Q}(\widehat{L}, \widehat{S}) = \mathcal{Q}(L^\star + \widehat{\Delta}^L, S^\star + \widehat{\Delta}^S) &= \|L^\star_A + L^\star_B + \widehat{\Delta}^L_A + \widehat{\Delta}^L_B\|_* + \frac{\lambda}{\mu}\|S^\star_{\mathcal{I}} + S^\star_{\mathcal{I}^c} + \widehat{\Delta}^S_{\mathcal{I}} + \widehat{\Delta}^S_{\mathcal{I}^c}\|_1 \\
&\geq \|L^\star_A + \widehat{\Delta}^L_B\|_* - \|\widehat{\Delta}^L_A\|_* + \frac{\lambda}{\mu}\left(\|S^\star_{\mathcal{I}} + \widehat{\Delta}^S_{\mathcal{I}}\|_1 - \|\widehat{\Delta}^S_{\mathcal{I}^c}\|_1\right) \\
&\geq \|L^\star_A\|_* + \|\widehat{\Delta}^L_B\|_* - \|\hat{\Delta}^L_A\|_* + \frac{\lambda}{\mu}\left(\|S^\star_{\mathcal{I}}\|_1 + \|\widehat{\Delta}^S_{\mathcal{I}}\|_1 - \|\widehat{\Delta}^S_{\mathcal{I}^c}\|_1\right).
\end{aligned}$$

Therefore, it follows that,

$$\begin{aligned}
\mathcal{Q}(L^\star, S^\star) - \mathcal{Q}(\widehat{L}, \widehat{S}) &= \left(\|L^\star_A\|_* + \frac{\lambda}{\mu}\|S^\star_{\mathcal{I}}\|_1\right) - \mathcal{Q}(\widehat{L}, \widehat{S}) \\
&\leq \|\widehat{\Delta}^L_B\|_* + \frac{\lambda}{\mu}\|\widehat{\Delta}^S_{\mathcal{I}^c}\|_1 - \left(\|\widehat{\Delta}^L_A\|_* + \frac{\lambda}{\mu}\|\widehat{\Delta}^S_{\mathcal{I}}\|_1\right) = \mathcal{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{\mathcal{I}^c}) - \mathcal{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_{\mathcal{I}}).
\end{aligned}$$

□

Recall that the exhaustive search algorithm requires examining every time point in the search domain $\mathcal{T}$. It can then be seen that for the fixed true change point $\tau^\star$ solving optimization problems (4) in the main paper on $[1, \tau)$ and $[\tau, T)$ includes a portion of time where the underlying model is *misspecified*. For example, assuming that $\tau > \tau^\star$, then solving (4) in the main paper cannot reach the optimal estimation error rate. Therefore, we require the following lemma to select the tuning parameters for intervals involving misspecified models.

Lemma 3-3. *Under the condition of Theorem 1 with $\tau > \tau^\star$, consider the interval* $[1, \tau)$

where the model is misspecified and further select tuning parameters

$$\lambda_{1,\tau} = 4c\sqrt{\frac{\log p + \log(\tau-1)}{\tau-1}}, \quad \mu_{1,\tau} = 4c\sqrt{\frac{p + \log(\tau-1)}{\tau-1}}.$$

Suppose that the search domain $\mathcal{T}$ satisfies Assumption H3; then, the following hold:

(1) for $T \gtrsim \log p$, with probability at least $1 - c_1 p^{-1}$:

$$\left\| \frac{1}{\tau-1} \sum_{t=1}^{\tau-1} X_{t-1}(X_t - (L_1^\star + S_1^\star)X_{t-1})' \right\|_\infty \leq \frac{\lambda_{1,\tau}}{2} + c_0 \frac{(\tau - \tau^\star)_+}{\tau-1}(M_S \vee \alpha_L)(d^\star_{\max} + \sqrt{r^\star_{\max}}), \tag{3-14}$$

(2) for $T \gtrsim p$, with probability at least $1 - c_1' \exp(-c_2' p) T^{1-c_3'}$:

$$\left\| \frac{1}{\tau-1} \sum_{t=1}^{\tau-1} X_{t-1}(X_t - (L_1^\star + S_1^\star)X_{t-1})' \right\|_{op} \leq \frac{\mu_{1,\tau}}{2} + c_0 \frac{(\tau - \tau^\star)_+}{\tau-1}(M_S \vee \alpha_L), \tag{3-15}$$

where $c_0, c_1, c_1', c_2', c_3'$ are some generic large enough positive constants. Symmetrically, we can obtain the following deviation bounds for the other side of interval $[\tau, T)$:

(1) with probability at least $1 - c_1 p^{-1}$:

$$\left\| \frac{1}{T-\tau} \sum_{t=\tau}^{T-1} X_{t-1}(X_t - (L_2^\star + S_2^\star)X_{t-1})' \right\|_\infty \leq \frac{\lambda_{2,\tau}}{2} + c_0 \frac{(\tau^\star - \tau)_+}{T-\tau}(M_S \vee \alpha_L)(d^\star_{\max} + \sqrt{r^\star_{\max}}), \tag{3'}$$

(2) with probability at least $1 - c_1' \exp(-c_2' p) T^{1-c_3'}$:

$$\left\| \frac{1}{T-\tau} \sum_{t=\tau}^{T-1} X_{t-1}(X_t - (L_2^\star + S_2^\star)X_{t-1})' \right\|_{op} \leq \frac{\mu_{2,\tau}}{2} + c_0 \frac{(\tau^\star - \tau)_+}{T-\tau}(M_S \vee \alpha_L), \tag{4'}$$

where $c_0, c_1, c_1', c_2', c_3'$ are some large enough positive constants, and the tuning parameters $\lambda_{2,\tau}$ and $\mu_{2,\tau}$ are given by:

$$\lambda_{2,\tau} = 4c\sqrt{\frac{\log p + \log(T-\tau)}{T-\tau}}, \quad \mu_{2,\tau} = 4c\sqrt{\frac{p + \log(T-\tau)}{T-\tau}}.$$

Proof of Lemma 3-3. First, we present the details for (3-14). Fix $t \in \mathcal{T}$, and define $Y_t \overset{\text{def}}{=} X_t - (L_1^\star + S_1^\star)X_{t-1}$ for $t = 2, \ldots, \tau$. Therefore, we obtain that $\mathbb{E}(Y_t) = 0$, while $\text{cov}(X_t, Y_t) \neq 0$. Then, by setting $\lambda_{1,\tau} = \frac{A}{\sqrt{\tau-1}}$ and $\mu_{1,\tau} = \frac{A'}{\sqrt{\tau-1}}$, with

$A \overset{\text{def}}{=} 4c\sqrt{\log p + \log(\tau-1)}$ and $A' \overset{\text{def}}{=} 4c\sqrt{p + \log(\tau-1)}$, and for some large enough constant $c > 0$, we obtain using a union bound the following:

$$\mathbb{P}\left(\max_{\tau\in\mathcal{T}} \lambda_{1,\tau}^{-1} \left\| \frac{1}{\tau-1}\sum_{t=1}^{\tau-1} X_{t-1}Y_t' \right\|_\infty - c_0\lambda_{1,\tau}^{-1}\frac{(\tau-\tau^\star)_+}{\tau-1}(M_S \vee \alpha_L)(d_{\max}^\star + \sqrt{r_{\max}^\star}) > \frac{1}{2}\right)$$
$$\leq \sum_{\tau\in\mathcal{T}} \mathbb{P}\left(\left\| \frac{1}{\tau-1}\sum_{t=1}^{\tau-1} X_{t-1}\epsilon_t' \right\|_\infty > \frac{A}{2\sqrt{\tau-1}}\right) \overset{\text{(i)}}{\leq} 6\sum_{\tau\in\mathcal{T}} \exp\left(-\frac{c_0' A^2}{4}\right) \leq \frac{6c_0''}{p} \to 0,$$

where c_0' and c'' are some large constants, and inequality (i) holds based on the results of Proposition 3 in Basu et al. (2019).

Next, to see (3-15), we use the same defined random process $Y_t = X_t - (L_1^\star + S_1^\star)X_{t-1}$ and the notations A and A' above. Then, we obtain:

$$\mathbb{P}\left(\max_{\tau\in\mathcal{T}} \mu_{1,\tau}^{-1} \left\| \frac{1}{\tau-1}\sum_{t=1}^{\tau-1} X_{t-1}Y_t' \right\|_{\text{op}} - \mu_{1,\tau}^{-1} c_0\frac{(\tau-\tau^\star)_+}{\tau-1}(M_S \vee \alpha_L) > \frac{1}{2}\right)$$
$$\leq \sum_{\tau\in\mathcal{T}} \mathbb{P}\left(\left\| \frac{1}{\tau-1}\sum_{t=1}^{\tau-1} X_{t-1}\epsilon_t' \right\|_{\text{op}} > \frac{A'}{2\sqrt{\tau-1}}\right) \overset{\text{(i)}}{\leq} 6\sum_{\tau\in\mathcal{T}} \exp\left(-\frac{c_1 A'^2}{4}\right) \leq \frac{6}{e^{c_1 p} T^{c_1-1}} \to 0,$$

for some large enough constant $c_1 > 0$. (i) is a direct application of the result of Proposition 3 in Basu et al. (2019) to this inequality with the choice of $\eta = \frac{A'}{2\sqrt{\tau-1}}$. Note that $T \gtrsim p$ ensures that $\eta \leq 4c'\frac{\log(\tau-1)}{\tau-1} < 1$, and then we can derive the anticipated results in (3-14) and (3-15). By using a similar procedure, (3') and (4') also follow. □

Next, we provide a proof for the *uniqueness* of the low rank and sparse decomposition. The main idea follows the *rank-sparsity incoherence* condition introduced in Chandrasekaran et al. (2011) followed by certain refinements in Hsu et al. (2011) to characterize a decomposition of a matrix including a low rank component L and a sparse component S. Before proving the following lemma, we require the following essential quantities:

1. *Maximum number of non-zero entries in any row or column of* S:

$$\alpha(\rho) \overset{\text{def}}{=} \max\left\{\rho\|\text{sign}(S)\|_{1\to 1}, \rho^{-1}\|\text{sign}(S)\|_{\infty\to\infty}\right\},$$

2. *Sparseness of the singular vectors of L*: let $L = UDV$, U and V are matrices of left and right orthonormal singular vectors corresponding to the non-zero singular values of L, and the rank of L is r. Define

$$\beta(\rho) \stackrel{\text{def}}{=} \rho^{-1}\|UU'\|_\infty + \rho\|VV'\|_\infty + \|U\|_{2\to\infty}\|V\|_{2\to\infty},$$

where

$$\text{sign}(M)_{i,j} = \begin{cases} -1, & \text{if } M_{i,j} < 0 \\ 0, & \text{if } M_{i,j} = 0 \\ +1, & \text{if } M_{i,j} > 0 \end{cases},$$

and further define the induced norm $\|M\|_{p\to q} \stackrel{\text{def}}{=} \max\{\|Mv\|_q : v \in \mathbb{R}^n, \|v\|_p \le 1\}$.

Additionally, we define two subspaces:

$$\Omega = \Omega(S) \stackrel{\text{def}}{=} \left\{X \in \mathbb{R}^{p\times p} : \text{supp}(X) \subset \text{supp}(S)\right\},$$

be the space of matrices whose supports are subsets of the support of S, and let

$$T = T(L) \stackrel{\text{def}}{=} \left\{X_1 + X_2 \in \mathbb{R}^{p\times p} : \text{range}(X_1) \subset \text{range}(L),\ \text{range}(X_2') \subset \text{range}(L')\right\}$$

Lemma 3-4. *Suppose Assumption H2 in the case of a single change point or Assumption H2' in the case of multiple change points is satisfied. Then, the low rank plus sparse decomposition of all transition matrices A_j^*'s for $j = 1, \ldots, m_0 + 1$ are unique and further the restricted space condition proposed in Agarwal et al. (2012) is satisfied.*

Proof of Lemma 3-4. Without loss of generality, we consider the multiple change points case (i.e., we investigate Assumption H2'). Assuming that at the j-th stationary segment, $A_j^\star = S_j^\star + L_j^\star$, and then an application of the singular value decomposition (SVD) on $L_j^\star$ yields: $L_j^\star = U_j D_j V_j'$, where $D_j = \text{diag}(\sigma_1^j, \ldots, \sigma_{r_j}^j, 0, \ldots, 0)$, σ_i^j is the i-th singular value for $L_j^\star$, for $j = 1, 2, \cdots, m_0 + 1$. Next, we consider the Assumption H2'-(1)-(3).

First, based on Assumptions H2'-(1)-(2), we get that $\|L_j^\star\|_\infty \le \frac{\alpha_L}{p}$ for $j = 1, \cdots, m_0 + 1$, which coincides with the constrained space condition proposed in Agarwal et al. (2012) and Basu et al. (2019).

Then, according to the definition of functions $\alpha(\rho)$ and $\beta(\rho)$, we derive that

$$\alpha(\rho) = \max\left\{\rho, \frac{d^{\star}_{\max}}{\rho}\right\}, \ \beta(\rho) = \frac{\alpha_L}{p}(1+\rho+\frac{1}{\rho}). \tag{3-16}$$

Thus, by using (3-16) together with Assumption H2'-(3), we get for $\rho = 1$:

$$\alpha(1)\beta(1) = 3d^{\star}_{\max}\frac{\alpha_L}{p} = \mathcal{O}\left(d^{\star}_{\max}\sqrt{\frac{\log(pT)}{T}}\right).$$

Hence, with the newly proposed Assumption H2'-(3), we obtain that $\alpha(1)\beta(1) < 1$, which satisfies the sufficient condition of uniqueness of decomposition in Theorem 1 in Hsu et al. (2011). □

Lemma 3-5. *Suppose that the Assumptions of Theorem 1 hold, and use the weighted regularizer $\mathcal{Q}$. Further, tor a fixed $\tau \in \mathcal{T}$, define $\widehat{\Delta}^L_{1,\tau} = \widehat{L}_{1,\tau} - L^{\star}_1$, $\widehat{\Delta}^S_{1,\tau} = \widehat{S}_{1,\tau} - S^{\star}_1$, $\widehat{\Delta}^L_{2,\tau} = \widehat{L}_{2,\tau} - L^{\star}_2$, and $\widehat{\Delta}^S_{2,\tau} = \widehat{S}_{2,\tau} - S^{\star}_2$ for two intervals $[1,\tau)$ and $[\tau, T)$, and the misspecified error terms $\widetilde{\Delta}^L_{1/2,\tau} = \widehat{L}_1 - L^{\star}_2$, $\widetilde{\Delta}^S_{1/2,\tau} = \widehat{S}_1 - S^{\star}_2$ for the interval $[\tau^{\star}, \tau)$, respectively. Then, for the tuning parameters $(\lambda_{1,\tau}, \mu_{1,\tau})$ and $(\lambda_{2,\tau}, \mu_{2,\tau})$ proposed in (10) of Theorem 1, we obtain that:*

$$\mathcal{Q}\left(\widehat{\Delta}^L_{1,\tau}, \widehat{\Delta}^S_{1,\tau}\right) \le 4\mathcal{Q}\left(\widehat{\Delta}^L_{1,\tau}|_A, \widehat{\Delta}^S_{1,\tau}|_{\mathcal{I}}\right), \quad \mathcal{Q}\left(\widehat{\Delta}^L_{2,\tau}, \widehat{\Delta}^S_{2,\tau}\right) \le 4\mathcal{Q}\left(\widehat{\Delta}^L_{2,\tau}|_A, \widehat{\Delta}^S_{2,\tau}|_{\mathcal{I}}\right),$$

and

$$\mathcal{Q}\left(\widetilde{\Delta}^L_{1/2,\tau}, \widetilde{\Delta}^S_{1/2,\tau}\right) \le 4\mathcal{Q}\left(\widetilde{\Delta}^L_{1/2,\tau}|_A, \widetilde{\Delta}^S_{1/2,\tau}|_{\mathcal{I}}\right).$$

Proof of Lemma 3-5. Assuming that $\tau > \tau^{\star}$, we investigate the behavior of the misspecified model in the interval $[1,\tau)$ and the non-misspecified model in the interval $[\tau, T)$, separately. Therefore, for the interval $[\tau, T)$ and $(\widehat{L}_{2,\tau}, \widehat{S}_{2,\tau})$, according to the defined objective functions $\ell(L_1, S_1; \mathbf{X}^{[1:\tau)})$ and $\ell(L_2, S_2; \mathbf{X}^{[\tau:T)})$ in the main text, we obtain that for

minimizing $\ell(L_2, S_2; \mathbf{X}^{[\tau:T)})$, we derive:

$$\begin{aligned}
&\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\|X_t - (\widehat{L}_2 + \widehat{S}_2)X_{t-1}\|_2^2 + \lambda_{2,\tau}\|\widehat{S}_2\|_1 + \mu_{2,\tau}\|\widehat{L}_2\|_* \\
\leq &\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\|X_t - (L_2^\star + S_2^\star)X_{t-1}\|_2^2 + \lambda_{2,\tau}\|S_2^\star\|_1 + \mu_{2,\tau}\|L_2^\star\|_*.
\end{aligned}$$

After some algebraic rearrangements, and due to the nature of the decomposition spaces (A, B) for the low-rank components and the corresponding decomposable support sets $(\mathcal{I}, \mathcal{I}^c)$ for the sparse components, we get:

$$\begin{aligned}
0 &\leq \frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\|X_{t-1}(\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S)\|_2^2 \\
&\leq \frac{2}{T-\tau}\sum_{t=\tau}^{T-1} X_{t-1}'(\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S)'\epsilon_t + \lambda_{2,\tau}\left(\|S_2^\star\|_1 - \|\widehat{S}_{2,\tau}\|_1\right) + \mu_{2,\tau}\left(\|L_2^\star\|_* - \|\widehat{L}_{2,\tau}\|_*\right) \\
&\overset{\text{(i)}}{\leq} 2c_0\sqrt{\frac{\log p + \log(T-\tau)}{T-\tau}}\|\widehat{\Delta}_{2,\tau}^S\|_1 + 2c_0\sqrt{\frac{p + \log(T-\tau)}{T-\tau}}\|\widehat{\Delta}_{2,\tau}^L\|_* \\
&+ \lambda_{2,\tau}\left(\|S_2^\star\|_1 - \|\widehat{S}_{2,\tau}\|_1\right) + \mu_{2,\tau}\left(\|L_2^\star\|_* - \|\widehat{L}_{2,\tau}\|_*\right) \\
&\leq \frac{\lambda_{2,\tau}}{2}\|\widehat{\Delta}_{2,\tau}^S\|_1 + \frac{\mu_{2,\tau}}{2}\|\widehat{\Delta}_{2,\tau}^L\|_* + \lambda_{2,\tau}\left(\|S_2^\star\|_1 - \|\widehat{S}_{2,\tau}\|_1\right) + \mu_{2,\tau}\left(\|L_2^\star\|_* - \|\widehat{L}_{2,\tau}\|_*\right) \\
&\leq \frac{3}{2}\mu_{2,\tau}\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L|_A, \widehat{\Delta}_{2,\tau}^S|_{\mathcal{I}}) - \frac{1}{2}\mu_{2,\tau}\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L|_B, \widehat{\Delta}_{2,\tau}^S|_{\mathcal{I}^c}),
\end{aligned}$$

where inequality (i) holds because of the deviation bound derived in Lemma 3-3. Therefore, we can further derive that:

$$\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L, \widehat{\Delta}_{2,\tau}^S) \leq 4\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L\|_A, \widehat{\Delta}_{2,\tau}^S\|_{\mathcal{I}}).$$

On the other hand, for the misspecified model in the interval $[1, \tau)$, by minimizing the objective function $\ell(L_1, S_1; \mathbf{X}^{[1:\tau)})$ to the intervals $[1, \tau^\star)$ and $[\tau^\star, \tau)$ separately, we obtain:

$$\begin{aligned}
&\frac{1}{\tau^\star - 1}\sum_{t=1}^{\tau^\star-1}\|X_t - (\widehat{L}_1 + \widehat{S}_1)X_{t-1}\|_2^2 + \lambda_{1,\tau}\|\widehat{S}_1\|_1 + \mu_{1,\tau}\|\widehat{L}_1\|_* \\
\leq &\frac{1}{\tau^\star - 1}\sum_{t=1}^{\tau^\star-1}\|X_t - (L_1^\star + S_1^\star)X_{t-1}\|_2^2 + \lambda_{1,\tau}\|S_1^\star\|_1 + \mu_{1,\tau}\|L_1^\star\|_*,
\end{aligned}$$

and

$$\frac{1}{\tau-\tau^\star}\sum_{t=\tau^\star}^{\tau-1}\|X_t-(\widehat{L}_1+\widehat{S}_1)X_{t-1}\|_2^2+\lambda_{1,\tau}\|\widehat{S}_1\|_1+\mu_{1,\tau}\|\widehat{L}_1\|_*$$
$$\leq\frac{1}{\tau-\tau^\star}\sum_{t=\tau^\star}^{\tau-1}\|X_t-(L_2^\star+S_2^\star)X_{t-1}\|_2^2+\lambda_{1,\tau}\|S_2^\star\|_1+\mu_{1,\tau}\|L_2^\star\|_*.$$

Similarly for the first inequality, after some algebraic rearrangements, we derive that

$$\begin{aligned}
0&\leq\frac{1}{\tau^\star-1}\sum_{t=1}^{\tau^\star-1}\|X_{t-1}(\widehat{\Delta}_{1,\tau}^L+\widehat{\Delta}_{1,\tau}^S)\|_2^2\\
&\leq\frac{2}{\tau^\star-1}\sum_{t=1}^{\tau^\star-1}X_{t-1}'(\widehat{\Delta}_{1,\tau}^L+\widehat{\Delta}_{1,\tau}^S)'\epsilon_t+\lambda_{1,\tau}\left(\|S_1^\star\|_1-\|\widehat{S}_{1,\tau}\|_1\right)+\mu_{1,\tau}\left(\|L_1^\star\|_*-\|\widehat{L}_{1,\tau}\|_*\right)\\
&\leq 2c_0\sqrt{\frac{\log p+\log(\tau-1)}{\tau-1}}\|\widehat{\Delta}_{1,\tau}^S\|_1+2c_0\sqrt{\frac{p+\log(\tau-1)}{\tau-1}}\|\widehat{\Delta}_{1,\tau}^L\|_*\\
&+\lambda_{1,\tau}\left(\|S_1^\star\|_1-\|\widehat{S}_{1,\tau}\|_1\right)+\mu_{1,\tau}\left(\|L_1^\star\|_*-\|\widehat{L}_{1,\tau}\|_*\right)\\
&\leq\frac{\lambda_{1,\tau}}{2}\|\widehat{\Delta}_{1,\tau}^S\|_1+\frac{\mu_{1,\tau}}{2}\|\widehat{\Delta}_{1,\tau}^L\|_*+\lambda_{1,\tau}\left(\|S_1^\star\|_1-\|\widehat{S}_{1,\tau}\|_1\right)+\mu_{1,\tau}\left(\|L_1^\star\|_*-\|\widehat{L}_{1,\tau}\|_*\right)\\
&\leq\frac{3}{2}\mu_{1,\tau}\mathcal{Q}(\widehat{\Delta}_{1,\tau}^L|_A,\widehat{\Delta}_{1,\tau}^S|_{\mathcal{I}})-\frac{1}{2}\mu_{1,\tau}\mathcal{Q}(\widehat{\Delta}_{1,\tau}^L|_B,\widehat{\Delta}_{1,\tau}^S|_{\mathcal{I}^c}),
\end{aligned}$$

since the second inequality can be derived by the same procedure. Therefore, we conclude that

$$\mathcal{Q}(\widehat{\Delta}_{1,\tau}^L,\widehat{\Delta}_{1,\tau}^S)\leq 4\mathcal{Q}(\widehat{\Delta}_{1,\tau}^L\|_A,\widehat{\Delta}_{1,\tau}^S\|_{\mathcal{I}})\text{ and }\mathcal{Q}(\widetilde{\Delta}_{1/2,\tau}^L,\widetilde{\Delta}_{1/2,\tau}^S)\leq 4\mathcal{Q}(\widetilde{\Delta}_{1/2,\tau}^L\|_A,\widetilde{\Delta}_{1/2,\tau}^S\|_{\mathcal{I}}).$$

□

Lemma 3-6. *Under Assumptions H1'-H5', for a set of estimated change points* $(s_1,s_2,\cdots,s_m)$ *with* $m<m_0$*, there exist universal positive constants* $c_1,c_2>0$ *such that:*

$$\mathbb{P}\left(\min_{(s_1,\ldots,s_m)}\mathcal{L}_n(s_1,\ldots,s_m;\boldsymbol{\lambda},\boldsymbol{\mu})>\sum_{t=1}^{T}\|\epsilon_t\|_2^2+c_1\widetilde{v}\Delta_T-c_2mT\xi_T\left(d_{\max}^{\star 2}+r_{\max}^{\star\frac{3}{2}}\right)\right)\to 1,$$

where $\widetilde{v}\overset{def}{=}\min_{1\leq j\leq m_0}\{v_{j,S}^2+v_{j,L}^2\}$.

Proof of Lemma 3-6. Suppose we obtain a set of candidate change points $(s_1,s_2,\cdots,s_m)$ obtained by the rolling-window strategy. Since $m<m_0$, there exists a true change point $t_j^\star$ satisfying $|s_i-t_j^\star|>\Delta_T/4$. In order to find a lower bound for $\mathcal{L}(s_1,\cdots,s_m;\boldsymbol{\lambda},\boldsymbol{\mu})$, based on

the vanishing sequence $\{\xi_T\}$ specified in Assumption H3', there are three different cases to consider: (a) $|s_i - s_{i-1}| \leq T\xi_T$, which implies that there is a negligibly small interval between two consecutive estimated change points s_{i-1} and s_i; (b) there exist two true change points $\tau_j^\star$ and $\tau_{j+1}^\star$ such that $|s_{i-1} - \tau_j^\star| \leq T\xi_T$ and $|s_i - \tau_{j+1}^\star| \leq T\xi_T$; and (c) otherwise.

Next, we introduce some additional notation used in the sequel. Let $\widehat{\Delta}^L$ and $\widehat{\Delta}^S$ denote the difference between the true expression and its estimate; i.e., $\widehat{\Delta}^L = L_{j+1}^\star - \widehat{L}_i$ and $\widehat{\Delta}^S = S_{j+1}^\star - \widehat{S}_i$, respectively. We denote by $\widetilde{\Delta}^L$ and $\widetilde{\Delta}^S$ the difference between the true expression and its estimate in the misspecified time segments; i.e., $\widetilde{\Delta}^L = L_j^\star - \widehat{L}_i$ and $\widetilde{\Delta}^S = S_j^\star - \widehat{S}_i$.

For case (a), without loss of generality, we assume that $\tau_j^\star < s_{i-1} < s_i < \tau_{j+1}^\star$ to obtain:

$$\begin{aligned}
\sum_{t=s_{i-1}}^{s_i} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 &= \sum_{t=s_{i-1}}^{s_i} \|\epsilon_t\|_2^2 + \sum_{t=s_{i-1}}^{s_i} \|X_{t-1}(\widehat{\Delta}^L + \widehat{\Delta}^S)\|_2^2 + 2\sum_{t=s_{i-1}}^{s_i} X_{t-1}'(\widehat{\Delta}^L + \widehat{\Delta}^S)\epsilon_t \\
&\geq \sum_{t=s_{i-1}}^{s_i} \|\epsilon_t\|_2^2 - 2\left|\sum_{t=s_{i-1}}^{s_i} \langle X_{t-1}'\epsilon_t, \widehat{\Delta}^L\rangle\right| - 2\left|\sum_{t=s_{i-1}}^{s_i} \langle X_{t-1}'\epsilon_t, \widehat{\Delta}^S\rangle\right| \\
&\geq \sum_{t=s_{i-1}}^{s_i} \|\epsilon_t\|_2^2 - c\sqrt{T\xi_T p}\|\widehat{\Delta}^L\|_* - c'\sqrt{T\xi_T \log p}\|\widehat{\Delta}^S\|_1.
\end{aligned}$$

Based on Assumption H6 on the selection of the tuning parameters, we conclude that:

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{s_i} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 + \lambda_i\|\widehat{S}_i\|_1 + \mu_i\|\widehat{L}_i\|_* \\
\geq &\sum_{t=s_{i-1}}^{s_i} \|\epsilon_t\|_2^2 - c\sqrt{T\xi_T p}\|L_{j+1}^\star\|_* - c'\sqrt{T\xi_T \log p}\|S_{j+1}^\star\|_1. \qquad (3\text{-}17)
\end{aligned}$$

For case (b), we assume that $s_{i-1} < \tau_j^\star$, $s_i < \tau_{j+1}^\star$, $|s_{i-1} - \tau_j^\star| \leq T\xi_T$, and $|s_i - \tau_{j+1}^\star| \leq T\xi_T$. Since the estimates $\widehat{L}_i$ and $\widehat{S}_i$ are the minimizers to the objective

function as (9) in the main paper, then we obtain:

$$\begin{aligned}
&\frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 + \lambda_i \|\widehat{S}_i\|_1 + \mu_i \|\widehat{L}_i\|_* \\
&\leq \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} \|X_t - (L^\star_{j+1} + S^\star_{j+1})X_{t-1}\|_2^2 + \lambda_i \|S^\star_{j+1}\|_1 + \mu_i \|L^\star_{j+1}\|_*.
\end{aligned} \tag{3-18}$$

Some algebraic rearrangements and based on Assumption H6 on the selection of the tuning parameters, we obtain that:

$$\begin{aligned}
0 &\leq \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i} \|X_{t-1}(\widehat{\Delta}^L + \widehat{\Delta}^S)\|_2^2 \\
&\leq \frac{2}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i} \langle X'_{t-1}\epsilon_t, \widehat{\Delta}^L + \widehat{\Delta}^S \rangle + \frac{2}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{\tau^\star_j - 1} \langle X'_{t-1}\epsilon_t, L^\star_j - L^\star_{j+1} + S^\star_j - S^\star_{j+1} \rangle \\
&+ \lambda_i(\|S^\star_{j+1}\|_1 - \|\widehat{S}_i\|_1) + \mu_i(\|L^\star_{j+1}\|_* - \|\widehat{L}_i\|_*) \\
&\leq 2\left(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d^\star_{\max} \frac{T\xi_T}{s_i - s_{i-1}} \right) \|\widehat{\Delta}^S\|_1 + \lambda_i(\|S^\star_{j+1}\|_1 - \|\widehat{S}_i\|_1) \\
&+ 2\left(c\sqrt{\frac{p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r^\star_{\max}} \frac{T\xi_T}{s_i - s_{i-1}} \right) \|\widehat{\Delta}^L\|_* + \mu_i(\|L^\star_{j+1}\|_* - \|\widehat{L}_i\|_*) \\
&\leq \frac{\lambda_i}{2}\|\widehat{\Delta}^S\|_1 + \lambda_i(\|S^\star_{j+1}\|_1 - \|\widehat{S}_i\|_1) + \frac{\mu_i}{2}\|\widehat{\Delta}^L\|_* + \mu_i(\|L^\star_{j+1}\|_* - \|\widehat{L}_i\|_*) \\
&= \frac{3}{2}\lambda_i \|\widehat{\Delta}^S\|_{1,\mathcal{I}} - \frac{1}{2}\lambda_i \|\widehat{\Delta}^S\|_{1,\mathcal{I}^c} + \frac{3}{2}\mu_i \|\widehat{\Delta}^L\|_{*,A} - \frac{1}{2}\mu_i \|\widehat{\Delta}^L\|_{1,B} \\
&= \frac{3}{2}\mu_i \mathcal{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_{\mathcal{I}}) - \frac{1}{2}\mu_i \mathcal{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{\mathcal{I}^c}).
\end{aligned} \tag{3-19}$$

The properties of the weighted regularizer $\mathcal{Q}$ have been discussed in Lemma 3-3. Therefore, in accordance to equation (3-19), we obtain that:

$$\mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \leq 4\mathcal{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_{\mathcal{I}}).$$

Moreover, an application of the Cauchy-Schwarz inequality leads to the following upper bound for the weighted regularizer $\mathcal{Q}$ with respect to the support sets $(A, \mathcal{I})$ defined before

Lemma 3-2:

$$\mu_i \mathcal{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_{\mathcal{I}}^S) \leq \mu_i \sqrt{r^\star_{\max}} \|\widehat{\Delta}^L\|_F + \lambda_i \sqrt{d^\star_{\max}} \|\widehat{\Delta}^S\|_F \leq \sqrt{\lambda_i^2 d^\star_{\max} + \mu_i^2 r^\star_{\max}} \sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2}. \tag{3-20}$$

Next, examining the first inequality in (3-19), we see that there exists a positive constant $c' > 0$ such that:

$$\frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i} \|X_{t-1}(\widehat{\Delta}^L + \widehat{\Delta}^S)\|_2^2 \geq c' \|\widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 \geq \nu(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) - \frac{1}{2}\mu_i \mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S), \tag{3-21}$$

where $\nu > 0$ is the curvature constant appearing in the RSC condition and the last inequality holds due to Lemma 3-3. By substituting (3-21) into (3-19), we obtain:

$$\begin{aligned} & \nu(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) - \frac{1}{2}\mu_i \mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \leq \frac{3}{2}\mu_i \mathcal{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_{\mathcal{I}}^S) - \frac{1}{2}\mu_i \mathcal{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{\mathcal{I}^c}^S) \\ \Longrightarrow\ & \nu(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) \leq 2\mu_i \mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \leq 8\sqrt{\lambda_i^2 d^\star_{\max} + \mu_i^2 r^\star_{\max}} \sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} \\ \Longrightarrow\ & \|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 \leq \frac{64}{\nu^2}(\lambda_i^2 d^\star_{\max} + \mu_i^2 r^\star_{\max}). \end{aligned} \tag{3-22}$$

Following analogous derivations to (3-23), one can similarly conclude that the error bound of the estimates can still be verified in the interval $[s_{i-1}, s_i)$. Note that there is a misspecified model in the interval $[s_{i-1}, \tau_j^\star)$, which is discussed separately.

First, consider the interval $[\tau_j^\star, s_i)$, for which we have:

$$\begin{aligned} & \sum_{t=t_j^\star}^{s_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \\ \geq & \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \|\widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 - c'\sqrt{|s_i - t_j^\star| \log p} \|\widehat{\Delta}^S\|_1 - c'\sqrt{|s_i - t_j^\star| p} \|\widehat{\Delta}^L\|_* \\ \geq & \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \left(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 - \frac{1}{2}\mu_i \mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \right) \\ & - c'\sqrt{|s_i - \tau_j^\star| \log p} \|\widehat{\Delta}^S\|_1 - c'\sqrt{|s_i - t_j^\star| p} \|\widehat{\Delta}^L\|_* \\ \overset{\text{(i)}}{\geq} & \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \left(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 - (\frac{1}{2} + \frac{1}{c})\mu_i \mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \right) \end{aligned} \tag{3-23}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\geq} \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \left(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 - (\frac{1}{2} + \frac{1}{c}) \sqrt{\lambda_i^2 d_{j+1}^\star + \mu_i^2 r_{j+1}^\star} \sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} \right) \\
&\geq \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} \left(\sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} - (\frac{1}{2} + \frac{1}{c}) \sqrt{\lambda_i^2 d_{j+1}^\star + \mu_i^2 r_{j+1}^\star} \right) \\
&\overset{\text{(iii)}}{\geq} \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 - c''|s_i - \tau_j^\star| \left\{ \frac{d_{j+1}^\star \log p + r_{j+1}^\star p}{s_i - s_{i-1}} + (M_S^2 d_{\max}^{\star 2} d_{j+1}^\star + \alpha_L^2 r_{\max}^\star r_{j+1}^\star) \left(\frac{T\xi_T}{s_i - s_{i-1}} \right)^2 \right\} \\
&\quad -2c''|s_i - \tau_j^\star| \left(M_S d_{\max}^\star d_{j+1}^\star \sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r_{\max}^\star} r_{j+1}^\star \sqrt{\frac{p}{s_i - s_{i-1}}} \right) \frac{T\xi_T}{s_i - s_{i-1}},
\end{aligned}
$$

where c, c', c'' are large enough positive constants which can be determined based on the tuning parameter rates. In (3-23), (i) holds because of the selection of tuning parameters; (ii) holds by using the result from (3-20); (iii) holds due to the verified error bound in (3-22) and the selection of the tuning parameters. Based on the results for $|s_i - \tau_{j+1}^\star| \leq T\xi_T$ and $|s_{i-1} - \tau_j^\star| \leq T\xi_T$, then we get that $T\xi_T/(s_i - s_{i-1}) \to 0$ as $T \to +\infty$; the latter together with Assumption H3' imply

$$
\frac{d_{\max}^\star \log p + r_{\max}^\star p}{s_i - s_{i-1}} \geq \left(M_S^2 d_{\max}^{\star 3} + \alpha_L^2 r_{\max}^{\star 2} \right) \left(\frac{T\xi_T}{s_i - s_{i-1}} \right)^2.
$$

It can be verified by some algebraic rearrangements that

$$
\begin{aligned}
&\frac{d_{\max}^\star \log p + r_{\max}^\star p}{s_i - s_{i-1}} \geq \left(M_S^2 d_{\max}^{\star 3} + \alpha_L^2 r_{\max}^{\star 2} \right) \left(\frac{T\xi_T}{s_i - s_{i-1}} \right)^2 \\
\iff &\left(\frac{s_i - s_{i-1}}{T\xi_T} \right) \frac{d_{\max}^\star \log p + r_{\max}^\star p}{T\xi_T} \geq M_S^2 d_{\max}^{\star 3} + \alpha_L^2 r_{\max}^{\star 2},
\end{aligned}
$$

which can be directly derived from Assumption H3'. Similarly, we can prove the following fact:

$$
\begin{aligned}
&\frac{d_{\max}^\star \log p + r_{\max}^\star p}{s_i - s_{i-1}} \geq \left(M_S d_{\max}^\star d_{j+1}^\star \sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r_{\max}^\star} r_{j+1}^\star \sqrt{\frac{p}{s_i - s_{i-1}}} \right) \frac{T\xi_T}{s_i - s_{i-1}} \\
\iff &\left(\frac{d_{\max}^\star \log p + r_{\max}^\star p}{T\xi_T} \right)^2 \geq \left(M_S d_{\max}^\star d_{j+1}^\star \sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r_{\max}^\star} r_{j+1}^\star \sqrt{\frac{p}{s_i - s_{i-1}}} \right)^2,
\end{aligned}
$$

the right-hand side being upper bounded by:

$$\left(M_S d^\star_{\max} d^\star_{j+1}\sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r^\star_{\max}} r^\star_{j+1}\sqrt{\frac{p}{s_i - s_{i-1}}}\right)^2$$
$$\leq \left(M_S^2 d^{\star^3}_{\max} + \alpha_L^2 r^{\star^2}_{\max}\right)\left(\frac{d^\star_{\max}\log p + r^\star_{\max} p}{s_i - s_{i-1}}\right).$$

Substituting the upper bound into the inequality above, the fact can be verified. Therefore, (3-23) can be further lower bounded by:

$$\sum_{t=t_j^\star}^{s_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \geq \sum_{t=t_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 - c''\left(d^\star_{j+1}\log p + r^\star_{j+1} p\right). \tag{3-24}$$

Next, we consider the misspecified model in the interval $[s_{i-1}, \tau_j^\star)$, which satisfies the condition $|\tau_j^\star - s_{i-1}| \leq T\xi_T$, by using the notation $\widetilde{\Delta}^L$ and $\widetilde{\Delta}^S$ as previously defined. Then,

$$\sum_{t=s_{i-1}}^{\tau_j^\star-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \geq \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon\|_2^2 - c'\left(\sqrt{T\xi_T\log p}\|\widetilde{\Delta}^S\|_1 + \sqrt{T\xi_T p}\|\widetilde{\Delta}^L\|_*\right)$$
$$\geq \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon\|_2^2 - c'\left\{\sqrt{T\xi_T\log p}\left(\|\widehat{\Delta}^S\|_1 + \|S^\star_{j+1} - S^\star_j\|_1\right) + \sqrt{T\xi_T p}\left(\|\widehat{\Delta}^L\|_* + \|L^\star_{j+1} - L^\star_j\|_*\right)\right\}$$
$$\geq \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 - c'\sqrt{T\xi_T\log p}\|\widehat{\Delta}^S\|_1 - c'\sqrt{T\xi_T p}\|\widehat{\Delta}^L\|_* - c_1' d^\star_{\max}\sqrt{T\xi_T\log p} - c_2'\sqrt{r^\star_{\max}}\sqrt{T\xi_T p}$$
$$\overset{\text{(i)}}{\geq} \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 - \sqrt{T\xi_T(s_i - s_{i-1})}\mu_i\mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) - c_1' d^\star_{\max}\sqrt{T\xi_T\log p} - c_2'\sqrt{r^\star_{\max}}\sqrt{T\xi_T p} \tag{3-25}$$
$$\overset{\text{(ii)}}{\geq} \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 - c_0'\sqrt{T\xi_T(s_i - s_{i-1})}\left(\lambda_i^2 d^\star_{j+1} + \mu_i^2 r^\star_{j+1}\right) - c_1' d^\star_{\max}\sqrt{T\xi_T\log p} - c_2'\sqrt{r^\star_{\max}}\sqrt{T\xi_T p}$$
$$\overset{\text{(iii)}}{\geq} \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon\|_2^2 - c_1' d^\star_{\max}\sqrt{T\xi_T\log p} - c_2'\sqrt{r^\star_{\max}}\sqrt{T\xi_T p} - 16c_0'\left(d^\star_{j+1}\log p + r^\star_{j+1} p\right)\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}}$$
$$-c_0''\left(\frac{(T\xi_T)^{\frac{3}{2}}}{\sqrt{s_i - s_{i-1}}}\right)\left((M_S^2 d^{\star^3}_{\max} + \alpha_L^2 r^{\star^2}_{\max})\left(\frac{T\xi_T}{s_i - s_{i-1}}\right) + 2\left(M_S d^{\star^2}_{\max}\sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L r^{\star\frac{3}{2}}_{\max}\sqrt{\frac{p}{s_i - s_{i-1}}}\right)\right),$$

where c', c_0', c_1' and c_2' are large enough positive constants. Note that (i) holds because of the selection of the tuning parameters and the relationships between the ℓ_1, ℓ_2 and nuclear

norms on the true transition matrices; (ii) holds because of the upper bound of the weighted penalty term $\mathcal{Q}$ derived in (3-20); (iii) holds because of the selection of the tuning parameters. Similar to (3-23), we need to find a further lower bound for (3-25). Let us first establish the following two facts:

$$16c_0'\left(d_{j+1}^\star \log p + r_{j+1}^\star p\right)\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}} \geq c_0''(M_S^2 d_{\max}^{\star 3} + \alpha_L^2 r_{\max}^{\star 2})\frac{(T\xi_T)^{\frac{5}{2}}}{(s_i - s_{i-1})^{\frac{3}{2}}},$$

and

$$\begin{aligned}
&16c_0'\left(d_{j+1}^\star \log p + r_{j+1}^\star p\right)\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}} \\
&\geq 2c_0''\left(\frac{(T\xi_T)^{\frac{3}{2}}}{\sqrt{s_i - s_{i-1}}}\right)\left(M_S d_{\max}^{\star 2}\sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L r_{\max}^{\frac{3}{2}}\sqrt{\frac{p}{s_i - s_{i-1}}}\right).
\end{aligned}$$

The first inequality can be rearranged as:

$$16c_0'\left(\frac{d_{j+1}^\star \log p + r_{j+1}^\star p}{T\xi_T}\right)\left(\frac{s_i - s_{i-1}}{T\xi_T}\right) \geq c_0''(M_S^2 d_{\max}^{\star 3} + \alpha_L^2 r_{\max}^{\star 2}),$$

which can be directly verified by using Assumption H3'. The right-hand side of the second inequality is upper bounded by the Cauchy-Schwarz inequality:

$$M_S d_{\max}^{\star 2}\sqrt{\frac{\log p}{s_i - s_{i-1}}} + \alpha_L r_{\max}^{\frac{3}{2}}\sqrt{\frac{p}{s_i - s_{i-1}}} \leq \left(M_S^2 d_{\max}^{\star 3} + \alpha_L r_{\max}^{\star 2}\right)^{\frac{1}{2}}\left(\frac{d_{\max}^\star \log p + r_{\max}^\star p}{s_i - s_{i-1}}\right)^{\frac{1}{2}},$$

and after substituting back into the second inequality, we need to establish:

$$\begin{aligned}
&16c_0'\left(d_{\max}^\star \log p + r_{\max}^\star p\right)\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}} \\
&\geq 2c_0''\left(\frac{(T\xi_T)^{\frac{3}{2}}}{\sqrt{s_i - s_{i-1}}}\right)\left(M_S^2 d_{\max}^{\star 3} + \alpha_L r_{\max}^{\star 2}\right)^{\frac{1}{2}}\left(\frac{d_{\max}^\star \log p + r_{\max}^\star p}{s_i - s_{i-1}}\right)^{\frac{1}{2}} \\
\iff &64c_0'\left(\frac{d_{\max}^\star \log p + r_{\max}^\star p}{T\xi_T}\right)\left(\frac{s_i - s_{i-1}}{T\xi_T}\right) \geq c_0''\left(M_S^2 d_{\max}^{\star 3} + \alpha_L r_{\max}^{\star 2}\right),
\end{aligned}$$

which has already been proven. The following facts are consequences of Assumption H3':

$$d_{\max}^\star\sqrt{T\xi_T \log p} \geq d_{\max}^\star \log p\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}} \text{ and } \sqrt{r_{\max}^\star}\sqrt{T\xi_T p} \geq r_{\max}^\star p\sqrt{\frac{T\xi_T}{s_i - s_{i-1}}}.$$

Therefore, we can derive a further lower bound for (3-25):

$$\sum_{t=s_{i-1}}^{\tau_j^\star-1}\|X_t-(\widehat{L}_i+\widehat{S}_i)X_{t-1}\|_2^2\geq\sum_{t=s_{i-1}}^{t_j^\star-1}\|\epsilon_t\|_2^2-c_1'd_{\max}^\star\sqrt{T\xi_T\log p}-c_2'\sqrt{r_{\max}^\star}\sqrt{T\xi_T p}. \quad (3\text{-}26)$$

Combining the results from (3-24) and (3-26), we get:

$$\sum_{t=s_{i-1}}^{s_i-1}\|X_t-(\widehat{L}_i+\widehat{S}_i)X_{t-1}\|_2^2\geq\sum_{t=s_{i-1}}^{s_i-1}\|\epsilon_t\|_2^2-c_2\left(d_{\max}^\star\sqrt{T\xi_T\log p}+\sqrt{r_{\max}^\star}\sqrt{T\xi_T p}\right). \quad (3\text{-}27)$$

For case (c), we firstly assume that $s_{i-1}<\tau_j^\star<s_i$, $|s_{i-1}-\tau_j^\star|>\Delta_T/4$ and $|s_i-\tau_j^\star|>\Delta_T/4$, respectively. Therefore, the interval $[s_{i-1},\tau_j^\star)$where the model is misspecified is not negligible as compared to the other interval $[\tau_j^\star,s_i)$; hence, we can not obtain the convergence rate of $\widehat{L}_i$ and $\widehat{S}_i$ on the whole interval $[s_{i-1},s_i)$. By using a similar procedure as in (3-19), similar results can be derived as long as we choose the tuning parameters as proposed in case (c) of Assumption H6:

$$\mathcal{Q}\left(\widehat{\Delta}^L,\widehat{\Delta}^S\right)\leq 4\mathcal{Q}\left(\widehat{\Delta}^L|_A,\widehat{\Delta}^S|_{\mathcal{I}}\right),\quad \mathcal{Q}\left(\widetilde{\Delta}^L,\widetilde{\Delta}^S\right)\leq 4\mathcal{Q}\left(\widetilde{\Delta}^L|_A,\widetilde{\Delta}^S|_{\mathcal{I}}\right).$$

Next, by following the same procedure as in the proof of case (b), we separately consider two intervals: $[s_{i-1},\tau_j^\star)$ and $[\tau_j^\star,s_i)$ as follows:

For the interval $[s_{i-1},\tau_j^\star)$, we adopt the same notation as before and obtain:

$$\begin{aligned}
&\sum_{t=s_{i-1}}^{\tau_j^\star-1}\|X_t-(\widehat{L}_i+\widehat{S}_i)X_{t-1}\|_2^2\\
&\geq\sum_{t=s_{i-1}}^{\tau_j^\star-1}\|\epsilon_t\|_2^2+c|\tau_j^\star-s_{i-1}|\|\widetilde{\Delta}^L+\widetilde{\Delta}^S\|_F^2-c'\sqrt{|\tau_j^\star-s_{i-1}|\log p}\|\widetilde{\Delta}^S\|_1-c'\sqrt{|\tau_j^\star-s_{i-1}|p}\|\widetilde{\Delta}^L\|_*\\
&\overset{\text{(i)}}{\geq}\sum_{t=s_{i-1}}^{\tau_j^\star-1}\|\epsilon_t\|_2^2+c|\tau_j^\star-s_{i-1}|\left(\|\widetilde{\Delta}^L\|_F^2+\|\widetilde{\Delta}^S\|_F^2-\frac{3\mu_i}{2}\mathcal{Q}(\widetilde{\Delta}^L,\widetilde{\Delta}^S)-\frac{(d_{\max}^\star+\sqrt{r_{\max}^\star})\|\widetilde{\Delta}^S\|_1+\|\widetilde{\Delta}^L\|_*}{s_i-s_{i-1}}\right)\\
&\geq\sum_{t=s_{i-1}}^{\tau_j^\star-1}\|\epsilon_t\|_2^2+c|\tau_j^\star-s_{i-1}|\sqrt{\|\widetilde{\Delta}^L\|_F^2+\|\widetilde{\Delta}^S\|_F^2}\left(\sqrt{\|\widetilde{\Delta}^L\|_F^2+\|\widetilde{\Delta}^S\|_F^2}-\sqrt{\lambda_i^2d_{\max}^\star+\mu_i^2r_{\max}^\star}\right),
\end{aligned} \quad (3\text{-}28)$$

where (i) can be verified by using a similar procedure as in (3-36) and (3-37) in the proof of Theorem 1; the last inequality is a direct consequence of the Cauchy-Schwarz inequality for the upper bound of $\mathcal{Q}(\widetilde{\Delta}^L, \widetilde{\Delta}^S)$ and Assumption H3' on the minimum spacing, which leads to the vanishing of the last term.

On the other hand, for the interval $[\tau_j^\star, s_i)$, we employ a similar procedure as in (3-29) to derive the following result:

$$\begin{aligned}
&\sum_{t=\tau_j^\star}^{s_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \\
&\geq \sum_{t=\tau_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 + c|s_i - \tau_j^\star| \sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} \left(\sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} - \sqrt{\lambda_i^2 d_{\max}^\star + \mu_i^2 r_{\max}^\star} \right).
\end{aligned} \tag{3-29}$$

According to Assumption H1', either $\|S_{j+1}^\star - S_j^\star\|_2 \geq v_S > 0$ holds or $\|L_{j+1}^\star - L_j^\star\|_2 \geq v_L > 0$ holds. By defining $\widetilde{v} = \min_{1 \leq j \leq m_0} \{v_{j,S}^2 + v_{j,L}^2\}$, it is not difficult to see that either $\|\widehat{\Delta}^S\|_F^2 \geq \widetilde{v}/4$ or $\|\widetilde{\Delta}^S\|_F^2 \geq \widetilde{v}/4$, or $\|\widehat{\Delta}^L\|_F^2 \geq \widetilde{v}/4$ or $\|\widetilde{\Delta}^L\|_F^2 \geq \widetilde{v}/4$. Without loss of generality, we assume that $\|\widetilde{\Delta}^L\|_F \geq \widetilde{v}/4$ and $\|\widetilde{\Delta}^S\|_F \geq \widetilde{v}/4$, respectively. We can then obtain a further lower bound for (3-28) as follows:

$$\begin{aligned}
\sum_{t=s_{i-1}}^{\tau_j^\star-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 &\geq \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 \\
&\quad + \frac{\sqrt{2}}{4} c|\tau_j^\star - s_{i-1}| \sqrt{\widetilde{v}} \left(\frac{\sqrt{2}}{4}\sqrt{\widetilde{v}} - \left(2 + \frac{4c'}{c}\right) \sqrt{\lambda_i^2 d_{\max}^\star + \mu_i^2 r_{\max}^\star} \right) \\
&\geq \sum_{t=s_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 + c_1 \widetilde{v} \Delta_T.
\end{aligned} \tag{3-30}$$

For the second interval $[\tau_j^\star, s_i)$, we obtain a lower bound for (3-29):

$$\begin{aligned}
\sum_{t=\tau_j^\star}^{s_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 &\geq \sum_{t=\tau_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 - c_2|s_i - \tau_j^\star| \left(\lambda_i^2 d_{\max}^\star + \mu_i^2 r_{\max}^\star\right) \\
&\geq \sum_{t=\tau_j^\star}^{s_i-1} \|\epsilon_t\|_2^2 - c_2 \left(d_{\max}^\star \log(p \vee \Delta_T) + r_{\max}^\star (p \vee \log \Delta_T) \right).
\end{aligned} \tag{3-31}$$

Next, by combining the lower bounds (3-30) and (3-31), we obtain:

$$\sum_{t=s_{i-1}}^{s_i-1}\|X_t-(\widehat{L}_i+\widehat{S}_i)\|_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1}\|\epsilon_t\|_2^2+c_1\widetilde{v}\Delta_T-c_2\left(d^\star_{\max}\log(p\vee\Delta_T)+r^\star_{\max}(p\vee\log\Delta_T)\right). \quad (3\text{-}32)$$

Notice that another scenario might arise in the (c) case: namely, $s_{i-1}<\tau_j^\star<s_i$, $T\xi_T<|s_i-\tau_j^\star|\ll\Delta_T$ and $T\xi_T<|\tau_j^\star-s_{i-1}|\ll\Delta_T$. By using analogous calculations as in case (b), we finally establish:

$$\sum_{t=s_{i-1}}^{s_i-1}\|X_t-(\widehat{L}_i+\widehat{S}_i)\|_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1}\|\epsilon_t\|_2^2-c_2'T\xi_T\left(d^{\star 2}_{\max}+r^{\star\frac{3}{2}}_{\max}\right). \quad (3\text{-}33)$$

Combining (3-17), (3-27), (3-32), and (3-33) and summing up all $m+1$ intervals leads to the final result. □

To verify the theoretical properties for the surrogate weakly sparse model, the following lemmas are required.

Lemma 3-7. *Consider the single change point scenario in Proposition 2 and also assume $\tau>\tau^\star$. For the misspecified model in the interval $[1,\tau)$ with tuning parameters provided in (12), and given that the search domain $\mathcal{T}^w$ satisfies Assumption W2, as $T\gtrsim\log p$, we obtain:*

(1) with probability at least $1-c_1p^{-1}$:

$$\left\|\frac{1}{\tau-1}\sum_{t=1}^{\tau-1}X_{t-1}(X_t-A_1^\star X_{t-1})'\right\|_\infty \leq \frac{\lambda^w_{1,\tau}}{2}+c_0M_S\frac{(\tau-\tau^\star)_+}{\tau-1}R_q\eta_{\min}^{-q},$$

(2) with probability at least $1-c_2p^{-1}$:

$$\left\|\frac{1}{T-\tau}\sum_{t=\tau}^{T-\tau}X_{t-1}(X_t-A_2^\star X_{t-1})'\right\|_\infty \leq \frac{\lambda^w_{2,\tau}}{2}+c_0M_S\frac{(\tau^\star-\tau)_+}{T-\tau}R_q\eta_{\min}^{-q},$$

where c_0,c_1,c_2 are some large positive constants.

Proof of Lemma 3-7. This proof is similar to the proof of Lemma 3-3. The key step is to measure the deviations for the misspecified model in the posited interval. In this case, since we assume that $\tau>\tau^\star$, the deviation on the interval $[\tau^\star,\tau)$ is upper bounded by

$c_0 M_S \frac{(\tau-\tau^\star)_+}{\tau-1}|\mathcal{J}(\eta_j)|$ for some large constant $c_0 > 0$. Then, substituting the upper bound of $|\mathcal{J}(\eta_j)|$ by $R_q \eta_j^{-q}$ implies the final result. □

Lemma 3-8. *Under Assumptions W1-W3, for a set of estimated change points* $(s_1, s_2, \cdots, s_m)$ *with* $m < m_0$ *and for the minimum spacing* Δ_T*, and jump size* $v_A = \min_{1\le j\le m_0} \|A^\star_{j+1} - A^\star_j\|_2$*, there exist universal positive constants* $c_1, c_2 > 0$ *such that:*

$$\mathbb{P}\left(\min_{(s_1,\ldots,s_m)} \mathcal{L}_T(s_1,\ldots,s_m;\boldsymbol{\lambda}^w) > \sum_{t=1}^{T}\|\epsilon_t\|_2^2 + c_1 v_A \Delta_T - c_2 m T \xi_T R_q^2 \left(\frac{\log(p\vee T)}{T}\right)^{-q}\right) \to 1.$$

Proof of Lemma 3-8. This lemma is proved in a similar manner as Lemma 4 in Safikhani & Shojaie (2020). □

3.8 Technical Proofs for Main Theorems in Chapter 3

In this section, we provide all technical proofs for main theorems, corollaries, and properties established in the main context.

Proof of Theorem 3-1. Let $\widehat{\tau}$ be the estimated change point obtained by solving optimization problem (4) addressed in the main text. Based on Algorithm xxx, we use the following objective function $\mathcal{L}(\tau)$ similar to (3) in the main text, which can be written as:

$$\mathcal{L}(\tau) = \sum_{t=1}^{\tau-1}\|X_t - (\widehat{L}_{1,\tau} + \widehat{S}_{1,\tau})X_{t-1}\|_2^2 + \sum_{t=\tau}^{T-1}\|X_t - (\widehat{L}_{2,\tau} + \widehat{S}_{2,\tau})X_{t-1}\|_2^2 \stackrel{\text{def}}{=} I_1 + I_2, \quad (3\text{-}34)$$

wherein $\widehat{L}_{j,\tau}$ and $\widehat{S}_{j,\tau}$ for $j = 1, 2$ are the optimizers of the convex programs (5) in the main. Note that the estimated low rank and sparse components are functions of τ. Without loss of generality, we assume that $\tau > \tau^\star$, and the length of misspecified interval $(\tau - \tau^\star)$ is large enough.

Denote by $(L_1^\star, S_1^\star)$ and $(L_2^\star, S_2^\star)$ the true components in the interval $[1, \tau^\star)$ and $[\tau^\star, T)$, respectively. Further, we use the notation $\widehat{\Delta}^L_{j,\tau} = \widehat{L}_{j,\tau} - L_j^\star$, $\widehat{\Delta}^S_{j,\tau} = \widehat{S}_{j,\tau} - S_j^\star$ for $j = 1, 2$. Then, by using the tuning parameters defined in (6) in the main text, we obtain

the corresponding lower bound of I_2:

$$
\begin{aligned}
I_2 &= \sum_{t=\tau}^{T-1} \|X_t - (\widehat{L}_{2,\tau} + \widehat{S}_{2,\tau})X_{t-1}\|_2^2 \\
&\geq \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 + \sum_{t=\tau}^{e-1} \|X_{t-1}(\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S)\|_2^2 - 2\left|\sum_{t=\tau}^{T-1} X_{t-1}'(\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S)'\epsilon_t\right| \\
&\overset{\text{(i)}}{\geq} \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 + c'|T-\tau|\|\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S\|_F^2 \\
&\quad - c''|T-\tau|\left(\sqrt{\frac{\log p + \log(T-\tau)}{T-\tau}}\|\widehat{\Delta}_{2,\tau}^S\|_1 + \sqrt{\frac{p+\log(T-\tau)}{T-\tau}}\|\widehat{\Delta}_{2,\tau}^L\|_*\right) \\
&\overset{\text{(ii)}}{\geq} \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 + c'|T-\tau|\left(\|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2 - \frac{\mu_{2,\tau}}{2}\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L, \widehat{\Delta}_{2,\tau}^S)\right) - c''|T-\tau|\mu_{2,\tau}\mathcal{Q}(\widehat{\Delta}_{2,\tau}^L, \widehat{\Delta}_{2,\tau}^S) \\
&\overset{\text{(iii)}}{\geq} \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 + c'|T-\tau|\sqrt{\|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2}\left(\sqrt{\|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2}\right. \\
&\quad \left. - \left(\frac{1}{2} + \frac{c''}{c'}\right)\sqrt{\lambda_{2,\tau}^2 d_{\max}^\star + \mu_{2,\tau}^2 r_{\max}^\star}\right) \\
&\geq \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 - c'|T-\tau|\left(\lambda_{2,\tau}^2 d_{\max}^\star + \mu_{2,\tau}^2 r_{\max}^\star\right) \\
&\geq \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 - c'\left(d_{\max}^\star \log p + r_{\max}^\star p + (d_{\max}^\star + r_{\max}^\star)\log(T-\tau)\right), \qquad (3\text{-}35)
\end{aligned}
$$

where (i) holds based on Lemma 3-3; (ii) is derived based on the following result together with Assumption H3 on the size of the search domain $\mathcal{T}$:

$$
\begin{aligned}
\|\widehat{\Delta}_{2,\tau}^L + \widehat{\Delta}_{2,\tau}^S\|_F^2 &\geq \|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2 - 2\left|\langle \widehat{\Delta}_{2,\tau}^L, \widehat{\Delta}_{2,\tau}^S\rangle\right| \\
&\geq \|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2 - 2\|\widehat{\Delta}_{2,\tau}^L\|_\infty\|\widehat{\Delta}_{2,\tau}^S\|_1 \\
&\geq \|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2 - \frac{2\alpha_L}{p}\|\widehat{\Delta}_{2,\tau}^S\|_1 \\
&\geq \|\widehat{\Delta}_{2,\tau}^L\|_F^2 + \|\widehat{\Delta}_{2,\tau}^S\|_F^2 - \lambda_{2,\tau}\|\widehat{\Delta}_{2,\tau}^S\|_1 - \mu_{2,\tau}\|\widehat{\Delta}_{2,\tau}^L\|_*; \qquad (3\text{-}36)
\end{aligned}
$$

(iii) holds because of an application of the Cauchy-Schwarz inequality to the result of Lemma 3-5.

Next, we derive a lower bound for I_1. Before stating the results, we first define the

misspecified error terms $\widetilde{\Delta}^L_{1/2,\tau} = \widehat{L}_{1,\tau} - L^\star_2$ and $\widetilde{\Delta}^S_{1/2,\tau} = \widehat{S}_{1,\tau} - S^\star_2$, then we similarly obtain that:

$$
\begin{aligned}
I_1 &= \sum_{t=1}^{\tau-1} \|X_t - (\widehat{L}_{1,\tau} + \widehat{S}_{1,\tau})X_{t-1}\|_2^2 \\
&\geq \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + \sum_{t=1}^{\tau^\star-1} \|X_{t-1}(\widehat{\Delta}^L_{1,\tau} + \widehat{\Delta}^S_{1,\tau})\|_2^2 + \sum_{t=\tau^\star}^{\tau-1} \|X_{t-1}(\widetilde{\Delta}^L_{1/2,\tau} + \widetilde{\Delta}^S_{1/2,\tau})\|_2^2 \\
&- 2\left|\sum_{t=1}^{\tau^\star-1} X'_{t-1}(\widehat{\Delta}^L_{1,\tau} + \widehat{\Delta}^S_{1,\tau})'\epsilon_t\right| - 2\left|\sum_{t=\tau^\star}^{\tau-1} X'_{t-1}(\widetilde{\Delta}^L_{1/2,\tau} + \widetilde{\Delta}^S_{1/2,\tau})'\epsilon_t\right| \\
&\overset{(i)}{\geq} \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + c|\tau^\star - 1|\|\widehat{\Delta}^L_{1,\tau} + \widehat{\Delta}^S_{1,\tau}\|_F^2 + c'|\tau - \tau^\star|\|\widetilde{\Delta}^L_{1/2,\tau} + \widetilde{\Delta}^S_{1/2,\tau}\|_F^2 \qquad (3\text{-}37) \\
&- c_1|\tau^\star - 1|\left(\sqrt{\frac{\log p + \log(\tau-1)}{\tau-1}}\|\widehat{\Delta}^S_{1,\tau}\|_1 + \sqrt{\frac{p + \log(\tau-1)}{\tau-1}}\|\widehat{\Delta}^L_{1,\tau}\|_*\right) \\
&- c'_1|\tau - \tau^\star|\left(\left(\sqrt{\frac{\log p + \log(\tau-1)}{\tau-1}} + \frac{M_S \vee \alpha_L}{\tau-1}(d^\star_{\max} + \sqrt{r^\star_{\max}})\right)\|\widetilde{\Delta}^S_{1/2,\tau}\|_1\right. \\
&\left. + \left(\sqrt{\frac{p + \log(\tau-1)}{\tau-1}} + \frac{M_S \vee \alpha_L}{\tau-1}\right)\|\widetilde{\Delta}^L_{1/2,\tau}\|_*\right) \\
&\overset{(ii)}{\geq} \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + c|\tau^\star - 1|\left(\|\widehat{\Delta}^L_{1,\tau}\|_F^2 + \|\widehat{\Delta}^S_{1,\tau}\|_F^2 - \frac{3\mu_{1,\tau}}{2}\mathcal{Q}(\widehat{\Delta}^L_{1,\tau}, \widehat{\Delta}^S_{1,\tau})\right) \\
&+ c'|\tau - \tau^\star|\left(\|\widetilde{\Delta}^L_{1/2,\tau}\|_F^2 + \|\widetilde{\Delta}^S_{1/2,\tau}\|_F^2 - \frac{3\mu_{1,\tau}}{2}\mathcal{Q}(\widetilde{\Delta}^L_{1/2,\tau}, \widetilde{\Delta}^S_{1/2,\tau})\right) \\
&- c''|\tau - \tau^\star|\frac{(d^\star_{\max} + \sqrt{r^\star_{\max}})\|\widetilde{\Delta}^S_{1/2,\tau}\|_1 + \|\widetilde{\Delta}^L_{1/2,\tau}\|_*}{\tau-1} \\
&\overset{(iii)}{\geq} \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + c'_1|\tau - \tau^\star|\left(v_S^2 + v_L^2\right) - c'_2\left(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee T)\right),
\end{aligned}
$$

where $C_0 \geq (M_S \vee \alpha_L)$ is some large constant. Inequality (i) is derived by the deviation bound in Lemma 3-3; (ii) holds due to the definition of the weighted regularizer $\mathcal{Q}$ and (3-36); (iii) is derived by substituting the differences of sparse components $\|\widehat{\Delta}^S_{1,\tau}\|_2$ and $\|\widetilde{\Delta}^S_{1/2,\tau}\|_2$ by v_S and the differences of low rank components $\|\widehat{\Delta}^L_{1,\tau}\|_2$ and $\|\widetilde{\Delta}^L_{1/2,\tau}\|_2$ by v_L, respectively.

Since we can not verify the RE condition on the misspecified interval $[1, \tau)$ and due to $\|S^\star_2 - S^\star_1\|_2 \geq v_S > 0$ and $\|L^\star_2 - L^\star_1\|_2 \geq v_L > 0$, then we have either $\|\widehat{\Delta}^S_{1,\tau}\|_2 \geq v_S/4$ or

$\|\widetilde{\Delta}^S_{1/2,\tau}\|_2 \geq v_S/4$, and either $\|\widehat{\Delta}^L_{1,\tau}\|_2 \geq v_L/4$ or $\|\widetilde{\Delta}^L_{1/2,\tau}\|_2 \geq v_L/4$. Assume that $\|\widetilde{\Delta}^S_{1/2,\tau}\|_2 \geq v_S/4$ and $\|\widetilde{\Delta}^L_{1/2,\tau}\|_2 \geq v_L/4$, then based on Assumptions H2 and H3, it implies that

$$\frac{(d^\star_{\max} + \sqrt{r^\star_{\max}})\|\widetilde{\Delta}^S_{1/2,\tau}\|_1 + \|\widetilde{\Delta}^L_{1/2,\tau}\|_*}{\tau - 1} \to 0,$$

hence, for some constants c_1, $c_2 > 0$ we get:

$$I_1 \geq \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + c_1|\tau - \tau^\star| - c_2\bigg(d^\star_{\max}\log p + r^\star_{\max}p + (d^\star_{\max} + r^\star_{\max})\log(\tau - 1)\bigg). \tag{3-38}$$

Combining (3-35) and (3-38) establishes that the objective function $\mathcal{L}(\tau)$ satisfies:

$$\mathcal{L}(\tau) \geq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + K_1\left(v_S^2 + v_L^2\right)|\tau - \tau^\star| - K_2\bigg(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\bigg). \tag{3-39}$$

Next, we prove the upper bound of $\mathcal{L}(\tau^\star)$. For some constant $K > 0$,

$$\mathcal{L}(\tau^\star) \leq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + K\bigg(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\bigg). \tag{3-40}$$

To see this result, by using a similar procedure, we have:

$$\mathcal{L}(\tau^\star) = \sum_{t=1}^{\tau^\star-1} \|X_t - (\widehat{L}_{1,\tau} + \widehat{S}_{1,\tau})X_{t-1}\|_2^2 + \sum_{t=\tau^\star}^{T-1} \|X_t - (\widehat{L}_{2,\tau} + \widehat{S}_{2,\tau})X_{t-1}\|_2^2 \stackrel{\text{def}}{=} J_1 + J_2.$$

Then, we obtain:

$$\begin{aligned} J_1 &\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + 2c|\tau^\star - 1|\Bigg(\|\widehat{\Delta}^L_{1,\tau^\star}\|_F^2 + \|\widehat{\Delta}^S_{1,\tau^\star}\|_F^2 + c'\sqrt{\frac{\log p + \log(\tau^\star - 1)}{\tau^\star - 1}}\|\widehat{\Delta}^S_{1,\tau^\star}\|_1 \\ &\quad + c'\sqrt{\frac{p + \log(\tau^\star - 1)}{\tau^\star - 1}}\|\widehat{\Delta}^L_{1,\tau^\star}\|_*\Bigg) \\ &\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + K_1\bigg(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\bigg), \end{aligned} \tag{3-41}$$

and similarly we have:

$$J_2 \leq \sum_{t=\tau^\star}^{T-1} \|\epsilon_t\|_2^2 + K_2\left(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\right). \tag{3-42}$$

Hence, combining inequalities (3-41) and (3-42) leads to the fact (3-39).

Based on (3-39) and (3-40), and using the fact that $\widehat{\tau}$ is the minimizer of optimization program (4) in the main text, we get that with high probability:

$$\sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + K_1(v_S^2 + v_L^2)|\widehat{\tau} - \tau^\star| - K_2\Big(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\Big) \leq \mathcal{L}(\widehat{\tau})$$
$$\leq \mathcal{L}(\tau^\star) \leq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + K\Big(d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)\Big). \tag{3-43}$$

Therefore, with high probability, for some large enough constant $K_0 > 0$, the following holds

$$|\widehat{\tau} - \tau^\star| \leq K_0 \frac{d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)}{v_S^2 + v_L^2}, \tag{3-44}$$

which concludes the proof of the Theorem. □

Proof of Theorem 3-2. This proof is similar to the proof of Proposition 4.1 in Basu & Michailidis (2015). The key steps in the proof that require verification are (a) the restricted strong convexity condition and (b) the deviation bound condition (see Appendix A) for the intervals $[1, \tau^\star - R)$ and $[\tau^\star + R, T)$, respectively, for the radius R. For (a), analogous arguments as in the proof of Theorem 4 in Safikhani & Shojaie (2020) establish the result. Further (b) follows from the result established in Lemma 3-1. □

Proof of Theorem 3-3. We first establish the following fact: suppose $(m_0, \widehat{\tau}_i, i = 1, 2, \ldots, m_0)$ is a subset of the candidate set $\widetilde{\mathcal{S}}$, which satisfies $\max_{1 \leq i \leq m_0} |\widehat{\tau}_i - \tau_i^\star| \leq T\xi_T$. Then, we can obtain the upper bound for $\mathcal{L}_n(\widehat{\tau}_1, \cdots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}, \boldsymbol{\mu})$:

$$\mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + Km_0T\xi_T(d^{\star^2}_{\max} + r^{\star\frac{3}{2}}_{\max}), \tag{3-45}$$

where $K > 0$ is a large enough constant.

To prove (3-45), we can focus on the estimated interval $(\widehat{\tau}_{i-1}, \widehat{\tau}_i)$ and corresponding estimates: $\widehat{L}_i$ and $\widehat{S}_i$. Suppose there is a true change point $\tau_j^\star$ such that: $\widehat{\tau}_{i-1} < \tau_j^\star < \widehat{\tau}_i$ with $|\tau_j^\star - \widehat{\tau}_{i-1}| \leq T\xi_T$. Similar to the proof in case (b) in Lemma 3-6, for the interval

$[\tau_j^\star, \widehat{\tau}_i)$, by choosing the same tuning parameters as in case (b) of Assumption H6, we have:

$$
\begin{aligned}
&\sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \\
&\leq \sum_{t=t_j^\star}^{\widehat{t}_i-1} \|\epsilon_t\|_2^2 + c_3|\widehat{\tau}_i - \tau_j^\star|\|\widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 + c_3'\left(\sqrt{|\widehat{\tau}_i - \tau_j^\star|p}\|\widehat{\Delta}^L\|_* + \sqrt{|\widehat{\tau}_i - \tau_j^\star|\log p}\|\widehat{\Delta}^S\|_1\right) \\
&\leq \sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|\epsilon_t\|_2^2 + 2c_3|\widehat{\tau}_i - \tau_j^\star|(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) + c_3'|\widehat{\tau}_i - \tau_j^\star|\left(\sqrt{\frac{p}{|\widehat{\tau}_i - \tau_j^\star|}}\|\widehat{\Delta}^L\|_* + \sqrt{\frac{\log p}{|\widehat{\tau}_i - \tau_j^\star|}}\|\widehat{\Delta}^S\|_1\right) \\
&\overset{\text{(i)}}{\leq} \sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|\epsilon_t\|_2^2 + 2c_3|\widehat{\tau}_i - \tau_j^\star|(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) + c_3'|\widehat{\tau}_i - \tau_j^\star|\frac{1}{2}\mu_i\mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \\
&= \sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|\epsilon_t\|_2^2 + c_3|\widehat{\tau}_i - \tau_j^\star|\left(2(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2) + \frac{c_3'}{2c_3}\mu_i\mathcal{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S)\right) \\
&\overset{\text{(ii)}}{\leq} \sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|\epsilon_t\|_2^2 + c_3|\widehat{\tau}_i - \tau_j^\star|\sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2}\left(2\sqrt{\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2} + \frac{c_3'}{2c_3}\sqrt{{\lambda_i}^2 d_{\max}^\star + {\mu_i}^2 r_{\max}^\star}\right) \\
&\overset{\text{(iii)}}{\leq} \sum_{t=\tau_j^\star}^{\widehat{\tau}_i-1} \|\epsilon_t\|_2^2 + \mathcal{O}_p\left(T\xi_T\left(d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}}\right)\right), \qquad (3\text{-}46)
\end{aligned}
$$

where (i) holds because of the selection of the tuning parameters (see Lemma 3-6 case (b)); (ii) holds since we can derive the error bound together with the upper bound of the weighted regularizer $\mathcal{Q}$; (iii) holds because of Assumptions H3' and H6.

For the other interval $[\widehat{\tau}_{i-1}, \tau_j^\star)$ we also get:

$$
\begin{aligned}
&\sum_{t=\widehat{\tau}_{i-1}}^{\tau_j^\star-1} \|X_t - (\widehat{L}_i + \widehat{S}_i)X_{t-1}\|_2^2 \\
&\leq \sum_{t=\widehat{\tau}_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 + c_3|\tau_j^\star - \widehat{\tau}_{i-1}|\|\widetilde{\Delta}^L + \widetilde{\Delta}^S\|_F^2 + c_3'\left(\sqrt{|\tau_j^\star - \widehat{\tau}_{i-1}|p}\|\widetilde{\Delta}^L\|_* + \sqrt{|\tau_j^\star - \widehat{\tau}_{i-1}|\log p}\|\widetilde{\Delta}^S\|_1\right) \\
&\leq \sum_{t=\widehat{\tau}_{i-1}}^{\tau_j^\star-1} \|\epsilon_t\|_2^2 + 2c_3|\tau_j^\star - \widehat{\tau}_{i-1}|\left(\|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 + \|L_{j+1}^\star - L_j^\star\|_F^2 + \|S_{j+1}^\star - S_j^\star\|_F^2\right) \\
&+c_3'\left(\sqrt{|\tau_j^\star - \widehat{\tau}_{i-1}|p}(\|\widehat{\Delta}^L\|_* + \|L_{j+1}^\star - L_j^\star\|_*) + \sqrt{|\tau_j^\star - \widehat{\tau}_{i-1}|\log p}(\|\widehat{\Delta}^S\|_1 + \|S_{j+1}^\star - S_j^\star\|_1)\right)
\end{aligned}
$$

$$\leq \sum_{t=\widehat{\tau}_{i-1}}^{\tau_j^\star - 1} \|\epsilon_t\|_2^2 + \mathcal{O}_p\left(T\xi_T\left(d_{\max}^{\star^2} + r_{\max}^{\star\frac{3}{2}})\right)\right). \tag{3-47}$$

Combining (3-46) and (3-47) and adding all $m_0 + 1$ intervals lead to (3-45).

Next, in order to prove the consistency of the number of estimated change points, we need to prove that: (a) $\mathbb{P}(\widehat{m} < m_0) \to 0$; and (b) $\mathbb{P}(\widehat{m} > m_0) \to 0$, respectively. To prove (a), we apply the result from Lemma 3-6, which leads to:

$$\begin{aligned}
&\mathrm{IC}(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{m}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T) = \mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{m}}; \boldsymbol{\lambda}, \boldsymbol{\mu}) + \widehat{m}\omega_T \\
&\overset{\text{(i)}}{>} \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + c_1\widetilde{v}\Delta_T - c_2\widehat{m}T\xi_T(d_{\max}^{\star^2} + r_{\max}^{\star\frac{3}{2}}) + \widehat{m}\omega_T \\
&\geq \mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}, \boldsymbol{\mu}) + m_0\omega_T + c_1\widetilde{v}\Delta_T - c_2 m_0 T\xi_T(d_{\max}^{\star^2} + r_{\max}^{\star\frac{3}{2}}) - (m_0 - \widehat{m})\omega_T \\
&\overset{\text{(ii)}}{\geq} \mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}, \boldsymbol{\mu}) + m_0\omega_T,
\end{aligned} \tag{3-48}$$

where (i) holds because of Lemma 3-6; (ii) holds because of Assumption H5. The result in (3-48) shows that $(\widehat{\tau}_1, \cdots, \widehat{\tau}_{\widehat{m}})$ is not the optimal solution to minimize IC function defined in (10) in the main; hence, we conclude that $\mathbb{P}(\widehat{m} < m_0) \to 0$. To prove (b), we assume that $(\widehat{\tau}_1, \cdots, \widehat{\tau}_{\widehat{m}})$ are the estimated change points with $\widehat{m} > m_0$. Then, similarly we get:

$$\mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{m}}; \boldsymbol{\lambda}, \boldsymbol{\mu}) \geq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 - c_2'\widehat{m}T\xi_T(d_{\max}^{\star^2} + r_{\max}^{\star\frac{3}{2}}). \tag{3-49}$$

Next, choose a subset $\{\widehat{\tau}_{i_1}, \cdots, \widehat{\tau}_{i_{m_0}}\}$ from $\{\widehat{\tau}_1, \cdots, \widehat{\tau}_{\widehat{m}}\}$ such that $\max_{1 \leq j \leq m_0} |\widehat{\tau}_{i_j} - \tau_j^\star| \leq T\xi_T$. Then, based on the definitions for $\mathrm{IC}(\widehat{\tau}_1, \cdots, \widehat{\tau}_{\widehat{m}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T)$ and $\mathrm{IC}(\widehat{\tau}_{i_1}, \cdots, \widehat{\tau}_{i_{m_0}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T)$ and using (3-49) we obtain:

$$\begin{aligned}
\sum_{t=1}^{T} \|\epsilon_t\|_2^2 - c_2'\widehat{m}T\xi_T(d_{\max}^{\star^2} + r_{\max}^{\star^2}) + \widehat{m}\omega_T &\leq \mathrm{IC}(\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{m}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T) \\
&\leq \mathrm{IC}(\widehat{\tau}_{i_1}, \ldots, \widehat{\tau}_{i_{m_0}}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \omega_T) \\
&\leq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + Km_0T\xi_T(d_{\max}^{\star^2} + r_{\max}^{\star\frac{3}{2}}) + m_0\omega_T,
\end{aligned} \tag{3-50}$$

which leads to:

$$(\widehat{m} - m_0)\omega_T \leq (Km_0 + c_2'\widehat{m})T\xi_T(d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}}). \tag{3-51}$$

Assumption $m_0 T\xi_T(d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}})/\omega_T \to 0$ implies that $m_0 < \widehat{m} \leq m_0$, which is a *contradiction*. Thus, we have established case (b) that $\mathbb{P}(\widehat{m} > m_0) \to 0$. Hence, we successfully prove that $\mathbb{P}(\widehat{m} = m_0) \to 1$.

The second part of Theorem 3 follows directly from the first part. By using similar arguments as in the proof of Theorem 1, it shows that for any estimated change point $\widehat{\tau}_j$, and corresponding true change point $\tau_j^\star$ such that:

$$\sum_{t=1}^{T} \|\epsilon_t\|_2^2 + c_1\widetilde{v}|\widehat{\tau}_j - \tau_j^\star| - c_2 m_0 T\xi_T\left(d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}}\right) \leq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + Km_0 T\xi_T\left(d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}}\right),$$

which implies that

$$\max_{1\leq j\leq m_0} |\widehat{\tau}_j - \tau_j^\star| \leq Bm_0 T\xi_T \frac{d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}}}{\min_{1\leq j\leq m_0}\{v_{j,S}^2 + v_{j,L}^2\}},$$

where $B > 0$ is a large enough constant. □

Proof of Corollary 1. This proof is similar to the proof of Theorem 4 in Safikhani & Shojaie (2020). We first remove the R-radius neighborhoods for each estimated change points $\widehat{\tau}_i$, we thus obtain the stationary segments $I_i \stackrel{\text{def}}{=} [\widehat{\tau}_i - R, \widehat{\tau}_i + R]$ for $i = 1, 2, \ldots, m_0$. Then, let N_i be the length of the i-th segment, the two key aspects that need to be verified are (a) the restricted strong convexity condition; (b) the deviation bound condition.

For each estimated segment I_i, the result of Theorem 3 suggests that $N_i = \mathcal{O}(\Delta_T)$; therefore, sufficiently large sample sizes are available to verify the RSC condition and the deviation bounds in each segment. The verification is similar to Proposition 4.1 in Basu & Michailidis (2015).

Therefore, by using the tuning parameters selected and the result in Proposition 1(a) in Basu et al. (2019), the final result follows. □

Proof of Corollary 2. This proof is similar to the proof of Theorem 3 and Theorem 1. By

using the conclusion in Theorem 3, we have $\mathbb{P}(\widehat{m} = m_0) \to 1$. Since we are using the similar procedure as singel change point detection proposed in Theorem 1, the estimated change points $\widetilde{\tau}_j$ satisfy the similar results as the proof of Theorem 1. Hence, for the jth refined change point:

$$|\widetilde{\tau}_j - \tau_j^\star| \leq K_0 \frac{d_j^\star \log(p \vee h) + r_j^\star (p \vee \log h)}{v_{j,S}^2 + v_{j,L}^2},$$

then combining all $\widehat{m}$ refined change points leads to the final result. □

Proof of Corollary 3. This proof is similar to the proof of Corollary 3 in Negahban et al. (2012). The main idea is to find an upper bound for the pseudo-sparsity level and an upper bound for the ℓ_1 norm of the true model parameter for the complementary sparse support set $\mathcal{J}(\eta_j)$, which have been already derived in the proof of Lemma 3-7.

The RSC condition can be verified as well for each estimated segment by using the same procedure as in the proof of Lemma 3-2. Applying Theorem 1 in Negahban et al. (2012) to the specific segment leads to the result.

Specifically, according to Theorem 1 in Negahban et al. (2012), with suitable selected tuning parameters, the error bound for the estimated model parameters is given by:

$$\|\widehat{A}_j^w - A^\star\|_F^2 \leq c_1 \lambda_j^{w^2} |\mathcal{J}(\eta_j)| + c_2 \lambda_j^w \left(c_3 \frac{\log p}{N_j} \|A^\star\|_{1,\mathcal{J}^c(\eta_j)}^2 + 4 \|A^\star\|_{1,\mathcal{J}(\eta_j)^c} \right);$$

therefore, by substituting the results of (20), we obtain

$$\|\widehat{A}_j^w - A^\star\|_F^2 \leq c_1 \lambda_j^{w^{2-q}} R_q + c_2 \left(\lambda_j^{w^{2-q}} R_q \right)^2 \frac{\log p}{\lambda_j^w N_j} \leq C_0 R_q \left(\frac{\log p}{N_j} \right)^{1-\frac{q}{2}},$$

where c_1, c_2, c_3, and C_0 are universal positive constants. □

Proof of Proposition 1. The result can be directly established by using the definition of the Hausdorff distance and the rolling-window mechanism provided in Algorithm xxx. Based on Assumption H4, the number of candidate change points $\widetilde{m}$ obtained by the rolling-window strategy satisfies $\widetilde{m} > T/\Delta_T > m_0$. Therefore, we get that $\mathbb{P}(\widetilde{m} \geq m_0) = 1$.

Based on the result of Theorem 1, for any true change point $\tau_j^\star$, once the window

includes $\tau_j^\star$, there exists an estimated change point $\widehat{\tau}_i$ satisfying with high probability:

$$|\widehat{\tau}_i - \tau_j^\star| \leq K \frac{d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)}{v_{j,S}^2 + v_{j,L}^2}$$

for some large enough positive constant K. Combining all m_0 change points, we obtain the final result. □

Proof of Proposition 2. Suppose that $A = L + S$ is one of the transition matrices in model (1). Further, suppose A is in the given ℓ_q-ball and the support set of the sparse component S is denoted by $\mathcal{I}$, and $|\mathcal{I}| = d^\star$. We can then get:

$$\begin{aligned} \sum_{i=1}^{p}\sum_{j=1}^{p} |A_{ij}|^q &= \sum_{(i,j)\in\mathcal{I}} |L_{ij} + S_{ij}|^q + \sum_{(i,j)\in\mathcal{I}^c} |L_{ij} + S_{ij}|^q \\ &= \sum_{(i,j)\in\mathcal{I}} |L_{ij} + S_{ij}|^q + \sum_{(i,j)\in\mathcal{I}^c} |L_{ij}|^q \stackrel{\text{def}}{=} J_1 + J_2. \end{aligned} \tag{3-52}$$

First, a Singular Value Decomposition of matrix L yields: $L = UDV'$, where $U = [u_1, \ldots, u_p] \in \mathbb{R}^{p\times p}$, $V = [v_1, \ldots, v_p] \in \mathbb{R}^{p\times p}$ are orthonormal matrices (i.e., for any u_i or v_j, $\|u_i\| = \|v_j\| = 1$), and $D = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, where σ_k is the k-th largest singular value of L, and r is the rank of L. We can then obtain:

$$\begin{aligned} J_2 &\leq \sum_{(i,j)\in\mathcal{I}^c} \left| \sum_{k=1}^{r} \sigma_k u_{ik} v_{jk} \right|^q \leq \sum_{(i,j)\in\mathcal{I}^c} \left| \left(\sum_{k=1}^{r} \sigma_k u_{ik}^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{r} \sigma_k v_{jk}^2 \right)^{\frac{1}{2}} \right|^q \\ &\leq \sum_{(i,j)\in\mathcal{I}^c} |\sigma_1|^q = |\sigma_1|^q (p^2 - d^\star). \end{aligned} \tag{3-53}$$

Next, due to the fact that $|L_{ij} + S_{ij}|^q \leq |L_{ij}|^q + |S_{ij}|^q$, we can obtain the following result

$$J_1 \leq \sum_{(i,j)\in\mathcal{I}} |L_{ij}|^q + \sum_{(i,j)\in\mathcal{I}} |S_{ij}|^q \leq d^\star \left\{ \left(\frac{\alpha_L}{p} \right)^q + M_S^q \right\}. \tag{3-54}$$

Combining the results (3-53) and (3-54) leads to the final result:

$$\sum_{i=1}^{p}\sum_{j=1}^{p} |A_{ij}|^q \leq d^\star \left(\left(\frac{\alpha_L}{p} \right)^q + M_S^q \right) + (p^2 - d^\star)|\sigma_1|^q. \tag{3-55}$$

□

Proof of Proposition 3. Let the transition matrices $A_1^\star$ and $A_2^\star \in \mathbb{B}_q(R_q)$, with a fixed $q \in (0,1]$, and R_q satisfying the condition proposed in Proposition 1. Also, assume that the associated true change point satisfies $\tau^\star \in [1, T)$. To establish the result, we follow a similar strategy as in the proof of Theorem 1. First, we establish:

$$\ell(\tau^\star) \leq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + c_0 T^{\frac{q}{2}} R_q \left(\log p\right)^{1-\frac{q}{2}}. \tag{3-56}$$

Split the objective function $\ell(t)$ as follows:

$$\ell(\tau^\star) = \sum_{t=1}^{\tau^\star-1} \|X_t - \widehat{A}_{1,\tau^\star} X_{t-1}\|_2^2 + \sum_{t=\tau^\star}^{T-1} \|X_t - \widehat{A}_{2,\tau^\star} X_{t-1}\|_2^2 \equiv I_1 + I_2.$$

Then, based on the definition ℓ_q norm, we are able to obtain that:

$$\begin{aligned}
I_1 &= \sum_{t=1}^{\tau^\star-1} \|X_t - \widehat{A}_{1,\tau^\star} X_{t-1}\|_2^2 \\
&\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + c_1|\tau^\star - 1|\|\widehat{A}_{1,\tau^\star} - A_1^\star\|_2^2 + c_1'\sqrt{|\tau^\star - 1|(\log p + \log(\tau^\star - 1))}\|\widehat{A}_{1,\tau^\star} - A_1^\star\|_1 \\
&\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + c_1|\tau^\star - 1|\|\widehat{A}_{1,\tau^\star} - A_1^\star\|_2 \left(\|\widehat{A}_{1,\tau^\star} - A_1^\star\|_2 + \frac{c_1'}{c_1}\sqrt{R_q}\left(\frac{\log p + \log(\tau^\star - 1)}{\tau^\star - 1}\right)^{\frac{1}{2}(1-\frac{q}{2})}\right) \\
&\quad + c_1' R_q \left(\frac{\log p + \log(\tau^\star - 1)}{\tau^\star - 1}\right)^{1-\frac{q}{2}} \\
&\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + c_1|\tau^\star - 1|\|\widehat{A}_{1,\tau^\star} - A_1^\star\|_2^2 + c_1' R_q \left(\frac{\log p + \log(\tau^\star - 1)}{\tau^\star - 1}\right)^{1-\frac{q}{2}} \\
&\leq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + c_1 T^{\frac{q}{2}} R_q \left(\log p + \log T\right)^{1-\frac{q}{2}}.
\end{aligned} \tag{3-57}$$

Analogously, we can get for I_2:

$$I_2 \leq \sum_{t=\tau^\star}^{T-1} \|\epsilon_t\|_2^2 + c_2 T^{\frac{q}{2}} R_q \left(\log p + \log T\right)^{1-\frac{q}{2}}. \tag{3-58}$$

Combining (3-57) and (3-58) leads to the result in (3-55). Next, we prove that for any fixed time point $\tau \in \mathcal{T}$, there exists some large enough constants $c_1, c_2 > 0$, together with jump

size $v_A \stackrel{\text{def}}{=} \|A_2^\star - A_1^\star\|_2$ such that the lower bound for $\ell(\tau)$ is given by:

$$\ell(\tau) \geq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 - c_1 T^{\frac{q}{2}} R_q (\log p + \log T)^{1-\frac{q}{2}} + c_2 v_A^2 |\tau - \tau^\star|. \tag{3-59}$$

Consider the interval $[1, \tau)$ and $[\tau, T)$ separately. Notice that, in this situation, we might have a misspecified model in that interval. Specifically, let us assume $\tau > \tau^\star$; then, the interval with a misspecified model corresponds to $[\tau^\star, \tau)$. We then have:

$$\ell(\tau) = \sum_{t=1}^{\tau-1} \|X_t - \widehat{A}_{1,\tau} X_{t-1}\|_2^2 + \sum_{t=\tau}^{T-1} \|X_t - \widehat{A}_{2,\tau} X_{t-1}\|_2^2 \equiv I_1 + I_2,$$

and for I_1:

$$\begin{aligned}
I_1 &= \sum_{t=1}^{\tau^\star-1} \|X_t - \widehat{A}_{1,\tau} X_{t-1}\|_2^2 + \sum_{t=\tau^\star}^{\tau-1} \|X_t - \widehat{A}_{1,\tau} X_{t-1}\|_2^2 \\
&\geq \sum_{t=1}^{\tau^\star-1} \|\epsilon_t\|_2^2 + c|\tau^\star - 1| \|\widehat{A}_{1,\tau} - A_1^\star\|_2^2 - c' \sqrt{|\tau^\star - 1|(\log p + \log(\tau-1))} \|\widehat{A}_{1,\tau} - A_1^\star\|_1 \\
&+ \sum_{t=\tau^\star}^{\tau-1} \|\epsilon_t\|_2^2 + \widetilde{c}|\tau - \tau^\star| \|\widehat{A}_{1,\tau} - A_2^\star\|_2^2 - \widetilde{c}' \sqrt{|\tau - \tau^\star|(\log p + \log(\tau-1))} \|\widehat{A}_{1,\tau} - A_2^\star\|_1 \\
&- \widetilde{c}''|\tau - \tau^\star| \frac{M_S R_q}{\tau - 1} \left(\frac{\log p + \log(\tau-1)}{\tau - 1} \right)^{-\frac{q}{2}} \|\widehat{A}_1 - A_2^\star\|_1 \\
&\stackrel{\text{(i)}}{\geq} \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 + c|\tau^\star - 1| \|\widehat{A}_{1,\tau} - A_1^\star\|_2 \left(\|\widehat{A}_{1,\tau} - A_1^\star\|_2 - \frac{c'}{c} \sqrt{R_q} \left(\frac{\log p + \log(\tau-1)}{\tau - 1} \right)^{\frac{1}{2}(1-\frac{q}{2})} \right) \\
&+ \widetilde{c}|\tau - \tau^\star| \|\widehat{A}_{1,\tau} - A_2^\star\|_2 \left(\|\widehat{A}_{1,\tau} - A_2^\star\|_2 - \frac{\widetilde{c}'}{\widetilde{c}} \sqrt{R_q} \left(\frac{\log p + \log(\tau-1)}{\tau - 1} \right)^{\frac{1}{2}(1-\frac{q}{2})} \right) \\
&- 4cR_q \left(\frac{\log p + \log(\tau-1)}{\tau - 1} \right)^{1-\frac{q}{2}} - 4\widetilde{c} R_q \left(\frac{\log p + \log(\tau-1)}{\tau - 1} \right)^{1-\frac{q}{2}} \\
&\stackrel{\text{(ii)}}{\geq} \sum_{t=1}^{\tau-1} \|\epsilon_t\|_2^2 - c_1 T^{\frac{q}{2}} R_q (\log p + \log T)^{1-\frac{q}{2}} + c_2 v_A^2 |\tau - \tau^\star|.
\end{aligned} \tag{3-60}$$

(i) holds due to Assumption W2 on the search domain $\mathcal{T}^w$; (ii) holds due to assuming that $\|\widehat{A}_1 - A_2^\star\|_2 \geq v_A/4 > 0$.

Analogously, we can derive a lower bound for I_2:

$$I_2 \geq \sum_{t=\tau}^{T-1} \|\epsilon_t\|_2^2 - c_1 T^{\frac{q}{2}} R_q (\log p + \log T)^{1-\frac{q}{2}}. \tag{3-61}$$

Hence, we proved the conclusion in (3-59). Next, by using (3-60) and (3-61), we obtain that with high probability the following holds:

$$\begin{aligned} &\sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 - c_1 h^{\frac{q}{2}} R_q (\log p + \log T)^{1-\frac{q}{2}} + c_2 v_A^2 |\widehat{\tau} - \tau^\star| \\ &\leq \ell(\widehat{\tau}) \leq \ell(t_j^\star) \leq \sum_{t=1}^{T-1} \|\epsilon_t\|_2^2 + c_0 T^{\frac{q}{2}} R_q \left(\log p + \log T\right)^{1-\frac{q}{2}}. \end{aligned} \tag{3-62}$$

Thus, the error bound for $|\widehat{\tau} - \tau^\star|$ is given by

$$|\widehat{\tau} - \tau^\star| \leq \frac{c_0 + c_1}{c_2} \frac{T^{\frac{q}{2}} R_q \left(\log(p \vee T)\right)^{1-\frac{q}{2}}}{v_A^2}, \tag{3-63}$$

for some constants $c_0, c_1, c_2 > 0$ and $v_A = \|A_2^\star - A_1^\star\|_2$. □

Proof of Proposition 4. The proof is analogous to that of Proposition 1. Based on the rolling-window mechanism and the result of Proposition 2, we can verify the result. In this case, we just need to replace the sample size T by the window-size h. □

Proof of Proposition 5. This proof follows in a similar manner to that of Theorem 3. By following the arguments in Lemma 3-8, we firstly verify the upper bound of $\mathcal{L}_T^w(\widehat{\tau}_1, \cdots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}^w)$ with respect to the set of estimated change points $(\widehat{\tau}_1, \cdots, \widehat{\tau}_{m_0})$. The latter satisfy $\max_{1 \leq i \leq m_0} |\widehat{\tau}_i - \tau_i^\star| \leq T\xi_T^w$. Similar to the proof of Theorem 2, we obtain that

$$\mathcal{L}_T(\widehat{\tau}_1, \ldots, \widehat{\tau}_{m_0}; \boldsymbol{\lambda}^w) \leq \sum_{t=1}^{T} \|\epsilon_t\|_2^2 + K' m_0 T \xi_T^w R_q^2 \left(\frac{\log(p \vee T)}{T}\right)^{-q},$$

where $K' > 0$ is a large enough constant and $\Delta_T = \min_{1 \leq i \leq m_0 - 1} |\tau_i^\star - \tau_{i+1}^\star|$.

Next, we establish: (a) $\mathbb{P}(\widehat{m} < m_0) \to 0$; (b) $\mathbb{P}(\widehat{m} > m_0) \to 0$, respectively. For (a),

we have that: denote $\widetilde{v}_A = \min_{1\le j\le m_0} v_{j,A}$, where $v_{j,A} \stackrel{\text{def}}{=} \|A^\star_{j+1} - A^\star_j\|_2$, then

$$\begin{aligned}
&\mathrm{IC}^w(\widehat{\tau}_1,\ldots,\widehat{\tau}_{\widehat{m}};\boldsymbol{\lambda}^w,\omega_T^w) = \mathcal{L}_T^w(\widehat{\tau}_1,\ldots,\widehat{\tau}_{\widehat{m}};\boldsymbol{\lambda}^w) + \widehat{m}\omega_T^w \\
>&\sum_{t=1}^{T}\|\epsilon_t\|_2^2 + c_1\widetilde{v}_A^2\Delta_T - c_2\widehat{m}T\xi_T^w R_q^2\left(\frac{\log(p\vee T)}{T}\right)^{-q} + \widehat{m}\omega_T^w \\
\ge&\mathcal{L}_T(\widehat{\tau}_1,\ldots,\widehat{\tau}_{m_0};\boldsymbol{\lambda}^w) + m_0\omega_T^w + c_1\widetilde{v}_A^2\Delta_T - c_2 m_0 T\xi_T^w R_q^2\left(\frac{\log(p\vee T)}{T}\right)^{-q} - (m_0-\widehat{m})\omega_T^w \\
\ge&\mathcal{L}_T(\widehat{\tau}_1,\ldots,\widehat{\tau}_{m_0};\boldsymbol{\lambda}^w) + m_0\omega_T^w,
\end{aligned}$$

which implies that the set of estimated change points $(\widehat{\tau}_1,\cdots,\widehat{\tau}_{\widehat{m}})$ are not the optimal solution for minimizing IC^w. Hence, we conclude that $\mathbb{P}(\widehat{m} < m_0) \to 0$. To prove (b), suppose the set of estimated change points $(\widehat{\tau}_1,\cdots,\widehat{\tau}_{\widehat{m}})$ satisfy $\widehat{m} > m_0$; hence, we similarly obtain the following result:

$$\mathcal{L}_T^w(\widehat{\tau}_1,\ldots,\widehat{\tau}_{\widehat{m}};\boldsymbol{\lambda}^w) \ge \sum_{t=1}^{T}\|\epsilon_t\|_2^2 - c_2'\widehat{m}T\xi_T^w R_q^2\left(\frac{\log(p\vee T)}{T}\right)^{-q}.$$

Choose a subset of $(\widehat{\tau}_1,\cdots,\widehat{\tau}_{\widehat{m}})$ with m_0 elements, such that $\max_{1\le i\le m_0}|\widehat{\tau}_{k_i} - \tau_i^\star| \le T\xi_T^w$. We then have:

$$\begin{aligned}
&\sum_{t=1}^{T}\|\epsilon_t\|_2^2 - c_2'\widehat{m}T\xi_T^w R_q^2\left(\frac{\log(p\vee T)}{T}\right)^{-q} + \widehat{m}\omega_T^w \\
\le&\mathrm{IC}^w(\widehat{\tau}_1,\ldots,\widehat{\tau}_{\widehat{m}};\boldsymbol{\lambda}^w,\omega_T^w) \le \mathrm{IC}^w(\widehat{\tau}_{k_1},\ldots,\widehat{\tau}_{k_{m_0}};\boldsymbol{\lambda}^w,\omega_T^w) \\
\le&\sum_{t=1}^{T}\|\epsilon_t\|_2^2 + K' m_0 T\xi_T^w R_q^2\left(\frac{\log(p\vee T)}{T}\right)^{-q},
\end{aligned}$$

which implies that $m_0 < \widehat{m} \le m_0$, which is a contradiction. Therefore, we have $\mathbb{P}(\widehat{m} = m_0) \to 1$. The error bound is established by similar arguments as in Lemma 3-6. □

Under the additional Assumptions (W5a)–(W5c), Proposition 6 can be verified by the following proof.

Proof of Proposition 6. According to Proposition 4 and Proposition 1, we separately obtain that:

$$d_H(\widetilde{\mathcal{S}}_w,\mathcal{S}^\star) = K'\frac{h^{\frac{q}{2}}R_q\left(\log(p\vee h)\right)^{1-\frac{q}{2}}}{\min_{1\le j\le m_0} v_{j,A}^2},$$

and

$$d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star) = K \frac{d^\star_{\max} \log(p \vee h) + r^\star_{\max}(p \vee \log h)}{\min_{1 \le j \le m_0}\{v_{j,S}^2 + v_{j,L}^2\}},$$

for some constants $K, K' > 0$. Since we have

$v_{j,A} = \|A^\star_{j+1} - A^\star_j\|_2 = \|(S^\star_{j+1} - S^\star_j) + (L^\star_{j+1} - L^\star_j)\|_2$ for $j = 1, 2, \ldots, m_0$, hence, we have $2v_{j,A}^2 \ge (v_{j,S}^2 + v_{j,L}^2)$. Then, suppose $p \gtrsim h$ we establish the left-hand side:

$$\begin{aligned}
\frac{d_H(\widetilde{\mathcal{S}}_w, \mathcal{S}^\star)}{d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star)} &\ge \frac{K'}{K} \frac{\min_j\{v_{j,S}^2 + v_{j,L}^2\}}{\min_j v_{j,A}^2} \frac{h^{\frac{q}{2}} R_q (\log(p \vee h))^{1-\frac{q}{2}}}{d^\star_{\max} \log(p \vee h) + r^\star_{\max}(p \vee \log h)} \\
&\ge \frac{K'}{2K} \frac{c_0^{\frac{q}{2}} (\log T)^{\frac{q}{2}} R_q (\log p)^{1-\frac{q}{2}}}{(d^\star_{\max} \log p + r^\star_{\max} p)^{1-\frac{q}{2}}} \\
&\ge \frac{K'}{2K} \frac{c_0' (\log T)^{\frac{q}{2}} \left\{ d^\star_{\max} \left(\frac{\alpha_L^q}{p^q} + M_S^q \right) + (p^2 - d^\star_{\max}) |\sigma_1|^q \right\} (\log p)^{1-\frac{q}{2}}}{(d^\star_{\max} \log p + r^\star_{\max} p)^{1-\frac{q}{2}}} \\
&\ge \frac{c_0' K}{2K'} \frac{d^\star_{\max} \left(\frac{\alpha_L^q}{p^q} + M_S^q \right) + (p^2 - d^\star_{\max}) |\sigma_1|^q}{\left(d^\star_{\max} + r^\star_{\max} \frac{p}{\log p} \right)^{1-q}} \\
&\ge \frac{c_0' K}{2K'} \frac{d^\star_{\max} \frac{\alpha_L^q}{p^q} + (p^2 - d^\star_{\max}) (\frac{\alpha_L}{p})^q}{\left(d^\star_{\max} + r^\star_{\max} \frac{p}{\log p} \right)^{1-q}} \ge \frac{c_0'' K}{2K'} \frac{p^{2-q}}{\left(d^\star_{\max} + r^\star_{\max} \frac{p}{\log p} \right)^{1-q}} \ge 1
\end{aligned} \tag{3-64}$$

Next, we determine an upper bound for the ratio of estimation errors.

$$\begin{aligned}
\frac{d_H(\widetilde{\mathcal{S}}_w, \mathcal{S}^\star)}{d_H(\widetilde{\mathcal{S}}, \mathcal{S}^\star)} &= \frac{K}{K'} \frac{\min_j\{v_{j,S}^2 + v_{j,L}^2\}}{\min_j v_{j,A}^2} \frac{h^{\frac{q}{2}} R_q (\log p)^{1-\frac{q}{2}}}{d^\star_{\max} \log p + r^\star_{\max} p} \\
&= \frac{K}{K'} \frac{\min_j\{v_{j,S}^2 + v_{j,L}^2\}}{\min_j v_{j,A}^2} \frac{c_0^{\frac{q}{2}} (\log T)^{\frac{q}{2}} R_q (\log p)^{1-\frac{q}{2}}}{(d^\star_{\max} \log p + r^\star_{\max} p)^{1-\frac{q}{2}}} \\
&\overset{\text{(i)}}{\le} \frac{K}{K'} \frac{\min_j\{v_{j,S}^2 + v_{j,L}^2\}}{\min_j v_{j,A}^2} \frac{c_0' (d^\star_{\max} + r^\star_{\max})^{1-\frac{q}{2}} p^{2-q} \max\{\alpha_L, M_S\}^q}{(d^\star_{\max} + r^\star_{\max})^{1-\frac{q}{2}}} \\
&= \frac{\min_j\{v_{j,S}^2 + v_{j,L}^2\}}{\min_j v_{j,A}^2} c_0'' p^{2-q} (\log T)^{\frac{q}{2}},
\end{aligned} \tag{3-65}$$

where c_0' and c_0'' are some large enough universal constants, and (i) holds due to Assumption (W5c). Since we only consider the case that the information ratios $0 < \gamma_j < p$, which indicates that the sparse components are dominating as well as the jump size of $A_j^\star$'s are lower bounded by a small enough constant, then we have the last equation in (3-65) is

bounded by $c_0'' p^{2-q} (\log T)^{\frac{q}{2}}$. Therefore, combining the results in (3-64) and (3-65) leads to the desired outcome. □

CHAPTER 4
A FAST DETECTION METHOD OF CHANGE POINTS IN FUNCTIONAL CONNECTIVITY NETWORKS

As we discussed in Chapter 3, the computational complexity of proposed two-step algorithm is $\mathcal{O}(Tp^3)$ due to an SVD is required to estimate the low rank components for every search. Note that a surrogate weakly sparse model is developed to achieve computational efficient.

In this chapter, we devote to establish a fast detection strategy to detect multiple change points in exceedingly long time series data based on vector autorgressive (VAR) models, followed by estimation of the Granger causal networks in the stationary segments identified. The proposed strategy is supported by theoretical arguments for its detection capability.

4.1 Model Formulation

Recall the model we discussed in Chapter 2 and Chapter 3 without the low rank components, a *purely* sparse VAR model with m_0 structural change points is defined in (4-1). Suppose we have m_0 change points $0 = t_0^\star < t_1^\star < \cdots < t_{m_0}^\star < t_{m_0+1}^\star = T + 1$, and for any time point t such that $t_j^\star \leq t < t_{j+1}^\star$, where $j = 1, 2, \ldots, m_0 + 1$, we have:

$$y_t = \Phi^{(1,j)} y_{t-1} + \cdots + \Phi^{(q,j)} y_{t-q} + \Sigma^{1/2} \epsilon_t, \tag{4-1}$$

where y_t is a p-vector of observed time series at time t; $\Phi^{(l,j)} \in \mathbb{R}^{p \times p}$ is the coefficient matrix corresponding to the l-th lag ($l \in \{1, \cdots, q\}$) for the j-th stationary segment. Note that in the sequel, we assume that the coefficient matrices $\Phi^{(l,j)}$ are *sparse*. The elements of $\Phi^{(l,j)}$ correspond to Granger causal effects Basu et al. (2015) and their collection is referred to as a Granger causal network Friston et al. (2013). Finally, ϵ_t is a Gaussian noise process with Σ denoting the covariance matrix which is assumed to be fixed. For ease of presenting the change point detection procedure, we fix $\Sigma = \sigma^2 \mathbf{I}$.

In each segment $[t_{j-1}, t_j)$, all model parameters are assumed to be fixed. However, the elements of the autoregressive matrices can potentially change with respect to both the structure and magnitudes of its entries between segments. The key objective is then to

detect the number and the location of change points t_j, in a computationally efficient manner that is also scalable for very large values of T. Of interest is also to estimate accurately the VAR parameters $\Phi^{(l,j)}$, under a high dimensional scaling ($p^2 \gg T$).

4.2 A Thresholded Block Segmentation Scheme (TBSS) Algorithm

The main idea of the proposed algorithm is to partition the time axis into blocks of size b_T and fix the VAR parameters within each block. Obviously, the maximum block size b_T that can be afforded, and bounded by the minimum distance D_T allowed between break points; namely $D_T = \min_{j\in\{0,\cdots,m_0\}} |t_j^\star - t_{j+1}^\star|$. Further, as discussed in the sequel, there is a trade-off between the computational cost of TBSS and the block size b_T.

To formally summarize the key steps of the TBSS algorithm, we define a sequence of time points $q = r_0 < r_1 < ... < r_{k_T+1} = T$, which plays the role of end points for the blocks; i.e. $r_{i+1} - r_i = b_T$ for $i = 0, ..., T-1$, and $k_T = \lceil \frac{T}{b_T} \rceil$ is the total number of blocks. In the following context, we assume that $T = k_T b_T$, and we define $\theta_1 = \Phi^{(\cdot,1)}$ and for $i = 2, 3, \ldots, k_T$, we define the difference variables θ_i as follows:

$$\theta_i = \begin{cases} \Phi^{(\cdot,j+1)} - \Phi^{(\cdot,j)}, & \text{if } t_j \in [r_{i-1}, r_i) \text{ for some } j, \\ 0, & \text{otherwise.} \end{cases} \tag{4-2}$$

Next, we form the following linear regression with the newly defined θ_i in (4-2) as the regression parameters, we define the design matrix as:

$$\mathcal{Z} = \begin{pmatrix} \mathbf{Y}_0' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{Y}_1' & \mathbf{Y}_1' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Y}_{k_T-1}' & \mathbf{Y}_{k_T-1}' & \cdots & \mathbf{Y}_{k_T-1}' \end{pmatrix},$$

where the block matrices $\mathbf{Y}_j \in \mathbb{R}^{b_T \times pq}$ is defined as: $\mathbf{Y}_j = (Y_{r_j}, Y_{r_j+1}, \ldots, Y_{r_{j+1}-1})$, for $j = 0, 1, \ldots, k_T - 1$. To be explicit, we define the pq-dimensional vector $Y_l = (y_l', \cdots, y_{l-q+1}')'$, where for any time points $l = q-1, q, \ldots, T$. Then, the design matrix $\mathcal{Z} \in \mathbb{R}^{T \times k_T pq}$. Similarly, the response matrix is given by the following:

$\mathcal{Y} = (y_q, \cdots, y_{r_1}, \cdots, y_{r_{k_T-1}}, \cdots, y_T)'$, where the model parameters can be defined in the compact form as: $\Theta = (\theta_1, \theta_2, \ldots, \theta_{k_T})' \in \mathbb{R}^{k_T pq \times p}$, and the corresponding error term can be similarly given by $E = (\epsilon_q, \cdots, \epsilon_{r_1}, \epsilon_{r_1+1}, \cdots, \epsilon_{r_2}, \cdots, \epsilon_{r_{k_T-1}}, \cdots, \epsilon_T)'$. Note that in this parameterization, $\theta_i \neq 0$ for $i \geq 2$ implies a change in the VAR coefficients $\Phi^{(.,j)}$ in the block $[r_{i-1}, r_i)$. Therefore, for $j = 1, \cdots, m_0$, the structural break points t_j would correspond to time points $r_i \geq 1$, for which $\theta_i \neq 0$.

The regression model can be succinctly written as $\mathcal{Y} = \mathcal{Z}\Theta + E$ and compactly expressed as:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Theta} + \mathbf{E}, \tag{4-3}$$

where $\mathbf{Y} = \text{vec}(\mathcal{Y})$, $\mathbf{Z} = I_p \otimes \mathcal{Z}$, $\boldsymbol{\Theta} = \text{vec}(\Theta)$, and the noise term $\mathbf{E} = \text{vec}(E)$, with $\otimes$ denoting the Kronecker product of two matrices, and $\pi_b = k_T p^2 q$. Then, the dimensions of the dependent variable, design matix and parameter vector are $\mathbf{Y} \in \mathbb{R}^{Tp \times 1}$, $\mathbf{Z} \in \mathbb{R}^{Tp \times \pi_b}$, $\boldsymbol{\Theta} \in \mathbb{R}^{\pi_b \times 1}$, and $\mathbf{E} \in \mathbb{R}^{Tp \times 1}$.

Step 1. Identification of candidate break points

Based on the linear regression representation in (4-3), the model parameters $\boldsymbol{\Theta}$ will be estimated via a regularized least squares objective function. Regularization is necessary to handle the growing number of parameters due to the presence of break points, as well as the number of time series p. Therefore, an initial estimate of the $\boldsymbol{\Theta}$ parameters is obtained by solving the following ℓ_1-penalized least squares regression:

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{T-q+1} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\Theta}\|_2^2 + \lambda_{1,T} \|\boldsymbol{\Theta}\|_1 + \lambda_{2,T} \sum_{i=1}^{k_T} \left\| \sum_{j=1}^{i} \theta_j \right\|_1 \right\}. \tag{4-4}$$

There are two penalty terms, the first term controlling, through a fused lasso penalty, the number of break points and the second term inducing sparsity in the parameters in accordance with the sparse nature of the posited model. This optimization problem is convex and hence can be solved efficiently with available algorithms in Tibshirani et al.

(2005). Denote the set of estimated break points obtained from solving (4-4) by

$$\widehat{\mathcal{A}}_T = \left\{ i \geq 2 : \widehat{\theta}_i \neq 0 \right\}.$$

The cardinality of this set corresponds to the estimated number of break points; i.e., $\widehat{m} = |\widehat{\mathcal{A}}_T|$. Further, let $\widehat{t}_j$, $j = 1, \ldots, \widehat{m}$ denote their estimated locations. Then, the relationship between $\widehat{\theta}_j$ and $\widehat{\Phi}^{(.,j)}$ in each of the estimated segments is given by:

$$\widehat{\Phi}^{(.,1)} = \widehat{\theta}_1 \text{ and } \widehat{\Phi}^{(.,j)} = \sum_{i=1}^{\widehat{t}_j} \widehat{\theta}_i, \quad j = 1, 2, \ldots, \widehat{m}. \tag{4-5}$$

Note that the block size b_T acts as a tuning parameter that *regulates* the number of model parameters to be estimated, given by $\pi_b = \lceil \frac{T}{b_T} \rceil p^2 q$. In the extreme case with $b_T = 1$, TBSS reverts to an exhaustive search of all time points to locate the structural breaks. Nevertheless, b_T can not also be too large. Based on results in the literature (see Assumption A3 in Safikhani & Shojaie (2020), Assumptions A2 and A3 in Harchaoui & Lévy-Leduc (2010) and Assumptions H2 and H3 in Chan et al. (2014)) and our calculations adapting arguments in Safikhani & Shojaie (2020), b_T can range in $\{1, \cdots, \lfloor \sqrt{T} \rfloor\}$. To see this, note that large values of the block size b_T result in reduced computational cost in Step 1, while increasing the computation time in Step 2, as explained later on. To keep the balance between the computation times of both steps and to minimize their total computation time, it is appropriate to select the block size b_T within the range $\{1, \cdots, \lfloor \sqrt{T} \rfloor\}$. On the other hand, clearly b_T can not be larger than the minimum spacing D_T between consecutive true break points since it that case, there may be more than one true break point in a certain block, while the strategy in (4-4) can at most detect a single break point within each block. Therefore, any selection of block size will implicitly impose restrictions on the minimum spacing D_T between consecutive break points or equivalently, it puts an upper bound on the total number of break points allowed in the model. Specifically, D_T should be much large than the block size b_T (asymptotically, we should have $b_T/D_T \to 0$ as $T \to +\infty$). For example, if $b_T = \lfloor \sqrt{T} \rfloor$, then D_T should be of order at

least $T^{0.5+\nu}$ for some small positive ν which means the total number of true break points m_0 must be at most of order $T^{0.5-\nu}$. Any block size larger than $\lfloor\sqrt{T}\rfloor$ restricts even further the total number of true break points allowed in the model, while not helping in reducing computation time. Thus, the range $\{1, \cdots, \lfloor\sqrt{T}\rfloor\}$ can be seen as the feasible set of possible block sizes. Note that restrictions on the minimum spacing D_T between consecutive break points are natural/typical in break point detection literature Chan et al. (2014). Even in the case of $b_T = 1$, as mentioned in Safikhani & Shojaie (2020), D_T has a lower bound of order $(\log p)^{1+\nu}$ for some positive ν. A practical procedure to determine a good value for b_T, in the absence of good prior information on the spacing D_T between consecutive break points, is outlined in Section 4.4.1.

The following theorem illustrates that under the first step, there always exist selected change points close to the true change points. Before stating the theorem, we need some additional assumptions.

A1 For any fixed j, denote the covariance matrices $\Gamma_j(h) = \text{cov}\left(y_t^{(j)}, y_{t+h}^{(j)}\right)$ for $t, h \in \mathbb{Z}$. Also, we assume that for $\kappa \in [-\pi, \pi]$, the spectral density matrices $f_j(\kappa) = (2\pi)^{-1}\sum_{l\in\mathbb{Z}}\Gamma_j(l)\exp(-i\kappa l)$ exist. Further, we assume that the largest and smallest eigenvalues of the spectral density matrices are bounded.

A2 The matrices $\Phi^{(\cdot,j)}$ are sparse and there exists a positive constant $M > 0$ such that

$$\max_{1\le j\le m_0+1} \|\Phi^{(\cdot,j)}\|_\infty \le M.$$

A3 There exists a positive constant v such that

$$\min_{1\le j\le m_0} \|\Phi^{(\cdot,j+1)} - \Phi^{(\cdot,j)}\|_2 \ge v > 0.$$

Moreover, there exists a vanishing positive sequence γ_T such that, as $T \to +\infty$,

$$\min_{1\le j\le m_0+1} \frac{D_T}{T\gamma_T} \to +\infty \text{ and } d^\star_{\max}\sqrt{\frac{\log p}{T\gamma_T}} \to 0,$$

where $d^\star_{\max} = \max_{1\le j\le m_0+1} d^\star_j$.

Remark 4-1. *Assumption A1 allows us to derive probability bounds in high dimension models. Assumption A2 relates to the maximum magnitude for each sparse transition*

matrices of the model. Assumption A3 combines the detection rate to the tuning parameter chosen in the optimization problems (4-4). Moreover, this assumption provides a minimum spacing requirement on the transition matrices in different segments. This can be regarded as the extension of Assumption H2 and H3 in Chan et al. (2014), and Assumptions A2 and A3 in Harchaoui & Lévy-Leduc (2010). Also, the proposed vanishing sequence $\{\gamma_T\}$ is directly related to the detection rate of the break points.

Then by using Hausdorff distance introduced in Section 1.4, we obtain the following theorem.

Theorem 4-1. *Suppose the assumptions A1-A3 proposed in Safikhani & Shojaie (2020) hold, then as $T \to +\infty$,*

$$\mathbb{P}(d_H(\widetilde{\mathcal{A}}_T, \mathcal{A}_T) \leq T\gamma_T) \to 1,$$

where $\{\gamma_T\}$ is a vanishing positive sequence satisfying the conditions in the assumption A3 in Safikhani & Shojaie (2020).

The proof of Theorem 4-1 is similar to the proof of Theorem 3 in Safikhani & Shojaie (2020), hence, we omit the details in this dissertation.

Remark 4-2. *Theorem 4-1 provides the rate of consistency for break point detection after step 1 of TBSS. It shows that for each true break point t_j, there exists a candidate break point $\widehat{t}_j$ obtained by step 1 satisfying that $|\widehat{t}_j - t_j| \leq T\gamma_T$. Moreover, the sequence γ_T also depends on the minimum spacing D_T as well as the model dimension p. One can choose $\gamma_T = (\log p \log T)/T$ or $\gamma_T = (\log\log T \log p)/T$. This means that the convergence rate for estimating the relative locations of the break points, i.e. -$\widehat{t}_j/T$-, could be as small as $(\log\log T \log p)/T$.*

Step 2. Screening redundant candidate change points

The magnitude of the required threshold η is selected by a completely data-driven method. The main idea is to combine the K-means clustering algorithm with the BIC

criterion to cluster the changes in the parameter matrix into two subgroups. The main steps are given by:

A. (Initialization): Define the jumps for each partitioned block by $v_k = \|\widehat{\theta}_k\|_2^2$, for $k = 2, 3, \ldots, k_T$ and let $v_1 = 0$. Define the set V as $V = (v_1, v_2, \ldots, v_{k_T})$. Denote the set of selected blocks with large jumps as J and set it to be the empty set. Also, set the value of the BIC function by $\text{BIC}^{\text{old}} = \infty$.

B. (Recursion state): Apply K-means clustering to the obtained jumps V with $K = 2$ clusters. Cluster V_S contains small jump values, while cluster V_L includes those with large jump sizes. Define $\eta = (\min V_L + \max V_S)/2$ as the threshold. Then we add the corresponding partitioned blocks in the large jump size into J, and compute the BIC by using the estimated parameters $\widehat{\Theta}$ with $\widehat{\theta}_i = \mathbf{0}$ for blocks $i \notin J$ and denoted as BIC^{new}. Compute the difference $\text{BIC}^{\text{diff}} = \text{BIC}^{\text{new}} - \text{BIC}^{\text{old}}$. Stop if $\text{BIC}^{\text{diff}} \geq 0$; otherwise, set $\text{BIC}^{\text{old}} = \text{BIC}^{\text{new}}$ and $V = V \backslash J$.

Step 3. Block clustering

In this step, we use the Gap statistic Friedman et al. (2001) to determine the number of clusters of the candidate break points selected in Step 2. The output of this step corresponds to clusters $C_i, i = 1, \cdots, \widetilde{m}$ of candidate break points.

Step 4. Exhaustive search for identifying a single break point for each cluster

For each selected cluster of break points $C_i, i = 1, 2, \cdots, \widetilde{m}$ from Step 3, we define the *search interval* (l_i, u_i) with the lower and upper bounds as follows:

$$l_i = \begin{cases} c_i - b_T, & \text{if } |C_i| = 1, \\ \min\{C_i\}, & \text{otherwise,} \end{cases} \quad \text{and} \quad u_i = \begin{cases} c_i + b_T, & \text{if } |C_i| = 1, \\ \max\{C_i\}, & \text{otherwise,} \end{cases}$$

where c_i is the unique element in C_i when $|C_i| = 1$. Denote the subset of corresponding block indices in the interval (l_i, u_i) by J_i with $J_0 = \{1\}$ and $J_{\widetilde{m}+1} = \{k_T\}$. Further denote the closest block end to $(\max J_{i-1} + \min J_i)/2$ as w_i. Now, the *local* parameter estimators are given by:

$$\widetilde{\Phi}^{(\cdot,i)} = \sum_{k=1}^{w_i} \widehat{\theta}_k, \tag{4-6}$$

for $i = 1, 2, \cdots, \widetilde{m} + 1$, where $\widehat{\theta}_k$ is derived by block fused lasso step from (4-4), for $k = 1, 2, \cdots, k_T$. Finally, for each $i = 1, 2, \cdots, \widetilde{m}$, the final estimated break points is

defined as:

$$\widetilde{t}_i = \underset{s\in(l_i,u_i)}{\arg\min} \left\{ \sum_{t=l_i}^{s-1} \|y_{t+1} - \widetilde{\Phi}^{(.,i)} Y_t\|_2^2 + \sum_{t=s}^{u_i-1} \|y_{t+1} - \widetilde{\Phi}^{(.,i+1)} Y_t\|_2^2 \right\}, \tag{4-7}$$

where $\widetilde{\Phi}^{(.,i)}$ are the local parameter estimates obtained by (4-6). The final set of estimated break points obtained by solving (4-7) are denoted by $\widetilde{\mathcal{A}}_T = \{\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}\}$.

Now, we are in the position to establish the main result of estimating consistently the locations of the final detected break points $\widetilde{t}_j$'s. The proof of Theorem 4-2 is provided in the Supplement.

Theorem 4-2. *Suppose the tuning parameters* $\lambda_{1,T} = \mathcal{O}\left(\sqrt{\frac{\log(p\vee T)}{T}}\right)$, *and* $\lambda_{2,T} = \frac{C}{T}\sqrt{\frac{\log p}{T\gamma_T}}$ *for some large enough constant* $C > 0$, *and a vanishing sequence* $\{\gamma_T\}$ *introduced in the assumption A3. Then, as* $T \to +\infty$, *we have*

$$\mathbb{P}\left(\max_{1\le j\le m_0} |\widetilde{t}_j - t_j| \le K d^\star_{\max} \log p\right) \to 1.$$

The sketch of the proof of Theorem 4-2 is provided in the Section 4.7.

Step 5. Model parameter estimation

Once the final set of break points have been identified from Step 4, we can estimated the transition matrices (and thus the Granger causal networks) by using the algorithms developed in Lin & Michailidis (2017) for stationary data. To ensure that the data in the time segments between break points are strictly stationary, we remove all time points in a R_T-radius neighborhood of the break points obtained in Step 4. The length of R_T needs to be at least b_T. Specifically, denote by $s_{j1} = \widetilde{t}_j - R_T - 1$, $s_{j2} = \widetilde{t}_j + R_T + 1$ for $j = 1, \cdots, \widetilde{m}$, and set $s_{02} = q$ and $s_{(\widetilde{m}+1)1} = T$. Next, define the intervals $I_{j+1} = [s_{j2}, s_{(j+1)1}]$ for $j = 0, \cdots, \widetilde{m}$. The idea is to form a linear regression on $\cup_{j=0}^{\widetilde{m}} I_{j+1}$ and estimate the auto-regressive parameters by minimizing an ℓ_1-regularized least squares criterion. Specifically, we form the following linear regression similar to (4-3):

$$\mathcal{Y}_s = \mathcal{X}_s B + E_s,$$

where $\mathcal{Y}_s = (y_q, \cdots, y_{s_{11}}, \cdots, y_{s_{\widetilde{m}2}}, \cdots, y_T)'$, $B = (\beta_1, \beta_2, \cdots, \beta_{\widetilde{m}+1})$, the corresponding error term $E_s = (\zeta_q, \cdots, \zeta_{s_{11}}, \ldots, \zeta_{s_{\widetilde{m}2}}, \cdots, \zeta_T)'$, and the design matrix is given by:

$$\mathcal{X}_s = \begin{pmatrix} \widetilde{\mathbf{Y}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{Y}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widetilde{\mathbf{Y}}_{\widetilde{m}}, \end{pmatrix}.$$

where the diagonal elements are given by: $\widetilde{\mathbf{Y}}_j' = (Y_{s_{j2}-1}, \cdots, Y_{s_{(j+1)1}-1})$, and $j = 0, 1, \ldots, \widetilde{m} - 1$. Then, we can rewrite it in compact form as:

$$\mathbf{Y_s} = \mathbf{Z_s B} + \mathbf{E_s},$$

where $\mathbf{Y_s} = \text{vec}(\mathcal{Y}_\mathbf{s})$, $\mathbf{Z_s} = I_p \otimes \mathcal{X}_\mathbf{s}$, $\mathbf{B} = \text{vec}(B)$, $\mathbf{E_s} = \text{vec}(E_\mathbf{s})$, and $\mathbf{s}$ is the collection of all s_{j1} and s_{j2} for $j = 0, \ldots, m_0 + 1$. Let $\widetilde{\pi} = (\widetilde{m} + 1)p^2q$, $N_j = s_{(j+1)1} - s_{j2}$ be the length of the interval I_{j+1} for $j = 0, \cdots, \widetilde{m}$ and $N = \sum_{j=1}^{\widetilde{m}} N_j$. Then, $\mathbf{Y_s} \in \mathbb{R}^{Np\times 1}$, $\mathbf{Z_s} \in \mathbb{R}^{Np\times\widetilde{\pi}}$, $\mathbf{B} \in \mathbb{R}^{\tilde{\pi}\times 1}$, and $\mathbf{E_s} \in \mathbb{R}^{Np\times 1}$. Therefore, we estimate the VAR parameters by solving the following ℓ_1 regularized optimization problem:

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left\{ \frac{1}{N} \left\| \mathbf{Y_s} - \mathbf{Z_s B} \right\|_2^2 + \rho_T \left\| \mathbf{B} \right\|_1 \right\}. \tag{4-8}$$

Similar to the previous two chapters, here we are going to provide consistent estimators for model parameters as well. The following theorem indicates that the minimizer $\widehat{\mathbf{B}}$ of (4-8) is consistent estimator for model parameter vec(Φ).

Theorem 4-3. *The minimizer* $\widehat{\mathbf{B}}$ *of* (4-8) *satisfies*

$$\|\widehat{\mathbf{B}} - vec(\Phi)\|_2 = \mathcal{O}_p\left(\sqrt{\frac{d^\star_{\max}\log p}{T}}\right) \quad \textit{and} \quad \|\widehat{\mathbf{B}} - vec(\Phi)\|_1 = \mathcal{O}_p\left(d^\star_{\max}\sqrt{\frac{\log p}{T}}\right),$$

where $d^\star_{\max} = \max_{j=1,2,\ldots,m_0+1} d^\star_j$.

Remark 4-3. *Theorem 4-3 provides the estimation error bound for the model parameters for each stationary segments partitioned by detected break points* $\widetilde{t}_j$. *The* ℓ_2 *rate*

$\sqrt{d^{\star}_{\max} \log p/T}$ *is of the similar order as the rates for regression i.i.d. samples (see Loh & Wainwright (2012), Basu & Michailidis (2015)), the* ℓ_1 *rate is derived by applying inequality between* ℓ_1 *and* ℓ_2 *norms on the sparse matrix Bühlmann & Van De Geer (2011).*

Now, we present the first four change point detection steps in the following algorithm diagram.

Algorithm 5 Threshold Block Segmentation Scheme (TBSS) Algorithm

1. **Input:** Time series data $\{X_t\}$, $t = 1, 2, \ldots, T$; regularization parameters $\lambda_{1,T}$ and $\lambda_{2,T}$; block size b_T.

2. **Block Fused Lasso:** Partition time series into blocks of size b_T and fix the coefficient parameters within each block. Then we estimate model parameters by solving (4-4). Then, obtain candidate break points:

$$\widehat{\mathcal{A}}_T = \left\{ i \geq 2 : \widehat{\theta}_i \neq 0 \right\}.$$

3. **Hard-thresholding:** We firstly denote the set of selected blocks with large jumps as J and initialize it as $J = \emptyset$, and initialize the value of the BIC function by $\text{BIC}^{\text{old}} = \infty$. **while** $\text{BIC}^{\text{diff}} < 0$ **do** :
   - Apply k-means clustering to the obtained jumps set V with $k = 2$ clusters. Use V_S to denote the set with small jump values, and V_L to denote the set with large jump values. Define $\eta = (\min V_L + \max V_S)/2$ as the threshold.
   - Add the corresponding blocks in V_L into J and compute the BIC by using the estimated parameter $\widehat{\Theta}$ with $\widehat{\theta}_i = 0$ for blocks $i \notin J$ and denoted as BIC^{new}.
   - Compute the difference between the BIC values: $\text{BIC}^{\text{diff}} = \text{BIC}^{\text{new}} - \text{BIC}^{\text{old}}$. Set $\text{BIC}^{\text{old}} = \text{BIC}^{\text{new}}$ and $V = V \backslash J$.

4. **Block Clustering:** In this step, we use the Gap statistics to determine the number of clusters of the candidate break points screened by Step 2. The output is $C = \{C_1, C_2, \ldots, C_{\widetilde{m}}\}$.

5. **Exhaustive Search:** For each selected cluster of break points C_i, for $i = 1, 2, \ldots, \widetilde{m}$ from Step 3. We define the search interval (l_i, u_i) as discussed in Section 4.2.
   For $i = 1, 2, \ldots, \widetilde{m} + 1$ **do:**
   - *Local* parameter estimators are defined by (4-6).
   - The final selected break point $\widetilde{t}_i$'s are defined in (4-7).

6. **Output:** The final estimated break points $\widetilde{\mathcal{A}}_T = \{\widetilde{t}_1, \ldots, \widetilde{t}_{\widetilde{m}}\}$.

In order to explicitly illustrate the main steps of TBSS algorithm, we present the first four steps (detection steps) of TBSS in Figure 4-1, where we set $p = 20$ and $T = 600$ with two true change points at $\tau_1^\star = 200, \tau_2^\star = 400$, respectively, and we use the default block size $b_T = \lfloor \sqrt{T} \rfloor \approx 25$. Here, TBSS algorithm accurately estimates the change points: $\widehat{\tau}_1 = 200$ and $\widehat{\tau}_2 = 400$.

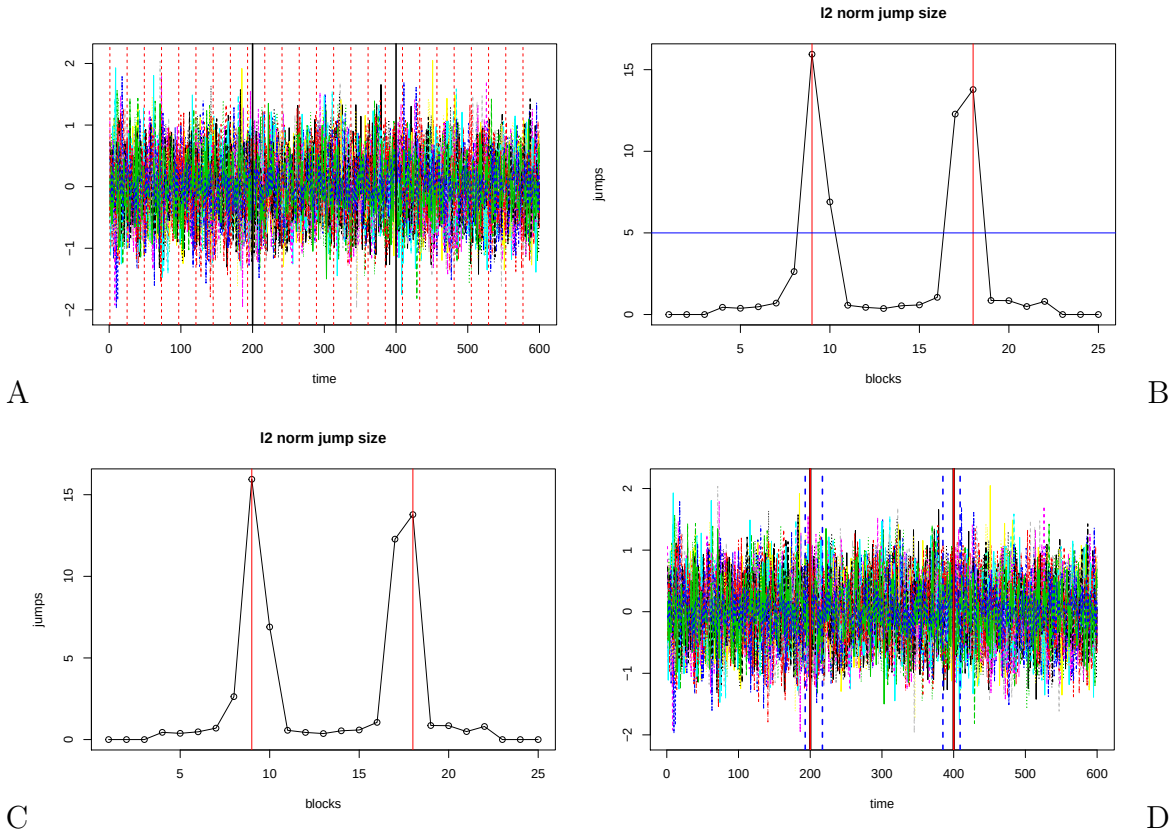


Figure 4-1. Illustration of main steps of TBSS algorithm.

In subplot A, we illustrate the data together with the blocks used in Step 1 of TBSS, those red dashed vertical lines indicate the end-points of blocks; then for Step 2 in subplot B, we provide the line plot of the ℓ_2 norm of the estimated parameters in each block, with the red lines indicating the blocks achieving locally maximum values, the horizontal blue line indicates the threshold η we found according to the jump sizes v_k; next, in subplot C, we present the Step 3 of TBSS, where two clusters of blocks are obtained comprising of large jump sizes; lastly, in Step 4 of subplot D, an exhaustive search of all the time points

contained in the selected clusters of blocks from Step 3 (indicated by the vertical blue dashed lines) is employed to identify the location of the final set of candidate break points, and gives the final estimated change points: $\widehat{\tau}_1 = 200$, and $\widehat{\tau}_2 = 400$ (red vertical lines).

4.3 Computational Complexity of TBSS

Note that the first four steps of TBSS correspond to estimating the number of true break points and their locations. To calculate the combined computational complexity of these steps, we assume that m_0 is finite. The effective sample size in Step 1 is k_T, which implies that the cost of solving the penalized regression in (4-3) is $\mathcal{O}\left(k_T p^2 q\right)$ Bleakley & Vert (2011). In Step 2, the initialization step in which model parameters in all blocks are estimated, requires the computational cost $\mathcal{O}\left(k_T p^2 q\right)$, while the recursively K-means clustering step needs the computational cost of order $\mathcal{O}\left(k_T\right)$ Friedman et al. (2001). Thus, the computational complexity of Step 2 is $\mathcal{O}\left(k_T p^2 q + k_T\right)$. The computational cost of Step 3 is similar to Step 2, since we cluster all candidates into $\widetilde{m}$ subsets, and the fact that $\widetilde{m} \leq k_T$ indicates that the order is upper bounded by $\mathcal{O}\left(k_T\right)$. The complexity of Step 4 is also similar to Step 2. According to the definition of intervals (l_i, u_i) in (4-7), the order for Step 4 is $\mathcal{O}\left(b_T p^2 q\right)$. Therefore, the total computational complexity of TBSS is $\mathcal{O}\left((k_T + b_T) p^2 q + k_T\right)$. Note that the number of blocks k_T is in fact T/b_T (rounded to the closest integer). Thus, the total computational complexity of TBSS can be written as a function of b_T only, i.e., $\mathcal{O}\left(\left(\frac{T}{b_T} + b_T\right) p^2 q + \frac{T}{b_T}\right)$. Selecting the block size $b_T = \mathcal{O}(\sqrt{T})$ yields the optimal computational cost $\mathcal{O}\left(\sqrt{T}(p^2 q + 1)\right)$.

To illustrate of computational complexity, we consider the following setting for a sparse lag 1 VAR model, whose transition matrix has non-zero elements in the first off diagonal, analogous to the pattern explored in Scenario A in the simulation studies (see Section 4.4). The number of time series under consideration is $p = 10, 20, 30$ with the number of time points $T = 20000$. Further, there are four equally-spaced break points located at $t_1 = 4000$, $t_2 = 8000$, $t_3 = 12000$, and $t_4 = 16000$.

We investigate the running time of the first four steps of TBSS for the following seven

block sizes b_T : 50, 75, 100, 120, 150, 200, and 300, respectively. The corresponding mean running times averaged over 50 replicates are depicted in Figure 4-2. It can be seen that initially the running time rapidly decreases as the block size increases from 50 to 120, but once it becomes larger than 150 it keeps increasing. Note that for a very small block size $b_T = 5$, the average running time for a exceeds 2.5 hours (approximate to 10,000 secs) for $p = 20$ and $p = 30$ scenarios, while a method employing mean shift models and binary segmentation Cho & Fryzlewicz (2015) requires over 5 hours.

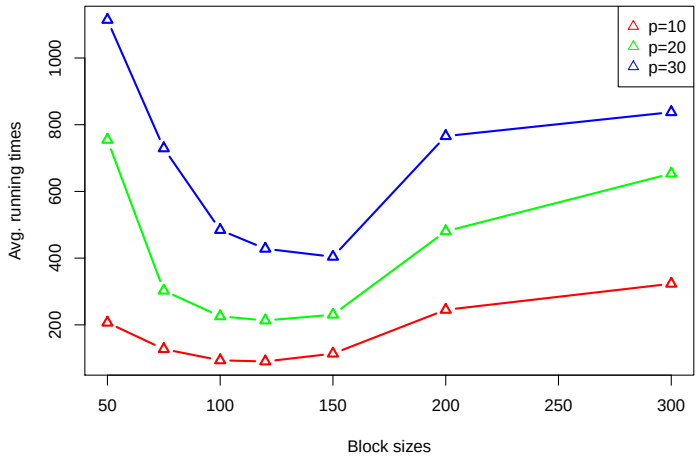


Figure 4-2. Averaged running time (in second) for different block sizes b_T, for VAR(1) models with 4 break points.

These numerical results are consistent with the complexity analysis and discussions above, wherein it was shown that the computational complexity of the detection phase (first four steps) of TBSS is $\mathcal{O}\left(\left(\frac{T}{b_T} + b_T\right) p^2 q + \frac{T}{b_T}\right)$, ignoring any (unknown) constants involved in the calculations for Steps 1-4 that may also differ across Steps. For this setting, the empirical optimal step size is 120 (for $p = 10, 20$) or 150 (for $p = 30$), which are close to $\lfloor \sqrt{T} \rfloor = 141$. Hence, it is recommended to set $b_T = \sqrt{T}$ in applications, provided that the underlying break points are not located too close to each other -see previous discussion on the interplay between the minimum spacing between consecutive break points D_T and the block size b_T.

4.4 Numerical Experiments

In this section, we assessed the performance of the proposed TBSS algorithm based on an extensive set of experiments on synthetic data. We start with an extensive discussion of how to select the tuning parameters in the algorithm, followed by a presentation of the simulation settings considered and the results obtained.

4.4.1 Tuning Parameter Selection

The TBSS algorithm is governed by a few tuning parameters. In Section 4.2, we provided a range for some of them based on some theoretical calculations. Next, we discuss in detail how these were selected in the simulation studies below and in the real data application.

$\lambda_{1,T}$: This parameter is selected 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 (block ends) 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 $\lambda_{1,T}$. The parameters estimated in Step 1 are used to predict the time series data at time points in $\mathcal{T}$. The value of $\lambda_{1,T}$ which minimizes the mean squared prediction error over $\mathcal{T}$ is the cross-validated choice of $\lambda_{1,T}$. The candidate sequence for $\lambda_{1,T}$ is selected as follows (see also Friedman et al. (2010)): pick a sequence of K_1 values for decreasing from $\lambda_{1,\max}$ to $\lambda_{1,\min}$ on the log-scale, where the maximum value $\lambda_{1,\max}$ is the smallest value that yields $\widehat{\theta}_i = 0$, for all $i = 1, 2, \cdots, k_T$, and the minimum value is set to be $\lambda_{1,\min} = \epsilon\lambda_{1,\max}$. We selected $\epsilon = 10^{-3}$ if the blocks size $b_T \leq 2p$ and $\epsilon = 10^{-4}$ otherwise. Finally, we selected $K_1 = 10$.

$\lambda_{2,T}$: This tuning parameter is also selected through cross-validation. Specifically, we set $\lambda_{2,T} = c\sqrt{\log p/T}$, where c is selected by similar procedure as $\lambda_{1,T}$: pick a decreasing sequence of K_2 values from $c_{\max}$ to $c_{\min}$, we set $c_{\max} = 0.1$ and $c_{\min} = \epsilon c_{\max}$, where $K_2 = 5$ and $\epsilon = 10^{-3}$.

ρ_T: This parameter controls the regularization term in optimization problem (4-8), and is selected by using cross-validation or BIC. In the TBSS algorithm, the ℓ_1 regularized optimization problems in Steps 1 and 5 are solved by using the `glmnet` package in `R`.

b_T: The default selection of block size is given by $\lfloor\sqrt{T}\rfloor$. In practice, b_T can be selected by grid search.

R_T: As we mentioned in 4.2, we set R_T equal to b_T in the last step of TBSS.

4.4.2 Simulation Studies

Next, we evaluate the performance of TBSS with respect to its accuracy of detecting both the number and location of break points, as well as the quality of the Granger causal network estimates ($\Phi^{(\cdot,j)}$'s).

To evaluate the performance of the first four steps of TBSS, we consider the mean and the standard deviation of the estimated break point locations relative to the sample size, i.e. $\widetilde{t}_j/T$, and the percentage of simulation runs where the correct number of break points are correctly detected. A detected break point is counted as a *success* for the j-th true break point, if it falls into the *selection interval*: $[t^\star_{j-1} + \frac{t^\star_j - t^\star_{j-1}}{5}, t^\star_j + \frac{t^\star_{j+1} - t^\star_j}{5}]$.

To measure the accuracy of the estimation of the sparse transition matrices ($\Phi^{(\cdot,j)}$'s), we use sensitivity (SEN), specificity (SPC), Matthew's Correlation Coefficient (MCC), and relative error in Frobenius norm (RE) as the evaluation criteria, defined next:

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}},\ \text{SPC} = \frac{\text{TN}}{\text{FP} + \text{TN}},\ \text{RE} = \frac{\|\text{Est.} - \text{Truth}\|_F}{\|\text{Truth}\|_F},$$
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}.$$

4.4.2.1 Numerical Scenarios and Model Parameter Settings

There are a number of factors influencing the performance of proposed algorithm, including the dimension of the model p, the number of observations T, the number of break points m_0, the spacing between them D_T, the lag of the auto-regressive model time series q, the sparsity level d_{kj} and the sparsity structure of transition matrices, and the block size b_T for selecting the candidate break points.

In this section, we first examine the impact of the block size of b_T in detection accuracy and computational speed based on scenario A discussed below. For all subsequent scenarios B-F, the block size is fixed to $b_T = \lfloor \sqrt{T} \rfloor$. The experimental scenarios considered are described next:

A. In this scenario, we investigate the impact of different block sizes b_T. For other parameters, we set $p = 20$, $T = 6000$ with four break points:

$t_1^\star = 1200, t_2^\star = 2400, t_3^\star = 3600$, and $t_4^\star = 4800$, the sparse transition matrices have all non-zero entries at 1-off diagonal, which is illustrated in the top panel of Figure 4-3.

B. In this scenario, we examine the influence of the number of break points m_0, for the following case: $p = 20$, $T = 21000$, with 6 equally spacing break points: $t_i^\star = 3000i$ for $i = 1, 2, \cdots, 6$. The sparsity pattern is as in scenario A.

C. In this scenario, we consider high-dimensional scaling. Specifically, we set $p = 40$, $p = 60$, and $p = 200$ respectively, with $T = 1000$ and single break point: $t_1^\star = 500$ and we set $p = 100$, $T = 1000$ with two change points: $t_1^\star = 333$, $t_2^\star = 666$. The sparsity pattern is as in scenario A.

D. In this scenario, we consider the effects of the time lag q in the presence of few break points; specifically, we examine a VAR(2) model in case D.1 with $p = 20$, $T = 3000$ and two break points at $t_1^\star = 1000$ and $t_2^\star = 2000$, and we assume that only the lag 1 autoregressive coefficients exhibit changes while the lag 2 ones remain the same. In D.2, we consider the setting $p = 20$ and $T = 2000$ with only one break point $t_1^\star = 1000$ with the assumption that both the lag 1 and lag 2 coefficients exhibit a change. The sparsity pattern is as in scenario A.

E. This scenario considers a random sparsity pattern in the transition matrices $\Phi^{(\cdot,j)}$, illustrated in the middle panel of Figure 4-3. The other model parameters are set to $p = 20$ and $T = 1000$ with two break points at $t_1^\star = 333$ and $t_2^\star = 667$.

F. The last scenario investigates the case that the transition matrices have similar patterns as real brain connectivity networks, which is presented in Figure 4-3 (bottom panel). In this scenario, we set $p = 21$, $T = 6000$ with four break points: $t_1^\star = 1200$, $t_2^\star = 2400$, $t_3^\star = 3600$, and $t_4^\star = 4800$.

Table 4-1 provides a summary of the specific settings and the exact location of the break points. The tuning parameters are selected according to the guidelines in Section 4.4.1. All numerical experiments are run in `R` version 3.6.3, on a PC equipped with 4 CPU cores at 3.6GHz frequency and 16GB DDR3 memory.

Next, we describe the data generation mechanism. In all scenarios, the error terms are independent and identically distributed from a Gaussian distribution with zero mean and variance 0.1, i.e. $\epsilon_t \overset{iid}{\sim} \mathcal{N}(0, 0.1\mathbf{I}_p)$. In order to ensure the stationarity of the VAR processes in each segment, the spectral radius of the transition matrices $\Phi^{(\cdot,j)}$ is equal to 0.8. In scenarios A, B, and C, the values of the non-zero elements of the transition matrices are set to -0.8 and 0.8 in an alternating manner. In scenario D, we consider a higher time lag $q = 2$, hence, the non-zero elements are set as follows: (a) in scenario D.1, we only

Table 4-1. Parameters settings aiming to large-scale time series data break point detection.

Case	b_T	p	T	$t_j^\star/T$	q
A.1	60				
A.2	80	20	6000	(0.200, 0.400, 0.600, 0.800)	1
A.3	100				
B.1	144	20	21000	$i/7$ for $i = 1, 2, \ldots, 6$	1
C.1		40			
C.2	31	60	1000	0.500	1
C.3		200			
C.4		100		$(0.333, 0.667)$	1
D.1	55	20	3000	$(0.333, 0.667)$	2
D.2	45	20	2000	0.500	2
E.1	31	20	1000	$(0.333, 0.667)$	1
E.2					
F.1	80	21	6000	$(0.200, 0.400, 0.600, 0.800)$	1

allow the changes in the first lag, specifically, the values of lag 1 and lag 2 for all three segments equal to $(0.6, 0.3)$, $(-0.6, 0.3)$ and $(0.6, 0.3)$, respectively; (b) in scenario D.2, we set the magnitudes of lag 1 and lag 2 for all segments changes: $(0.5, 0.35)$ and $(-0.6, -0.3)$, respectively. For scenario E, we consider two different random sparsity patterns shown in the panels B and C of Figure 4-3 with $p = 20$, $T = 1000$ and two change points. In the last scenario F, we construct the transition matrices similar to the real patterns in the functional connectivity networks of brain (here we estimate the connectivity networks of brain based on the real EEG data provided in Section 4.5), and generate VAR(1) time series data of $p = 21$, $T = 6000$ with four equally-spaced change points, which is similar to the real EEG data set in the next section.

4.4.2.2 Simulation Results

Table 4-2 presents the performance of the TBSS algorithm for Scenarios A-F. Overall, the detection performance of TBSS is satisfactory with having more than 80% selection rate in all simulation scenarios while in around half of them the selection rate is 100%. In scenario A, the detection and estimation performance seem to be robust with respect to changes in the block size. Most of the selection rates are above 90% in this scenario while the selection rate for the first change point in scenario A.3 drops to around 80%. This may

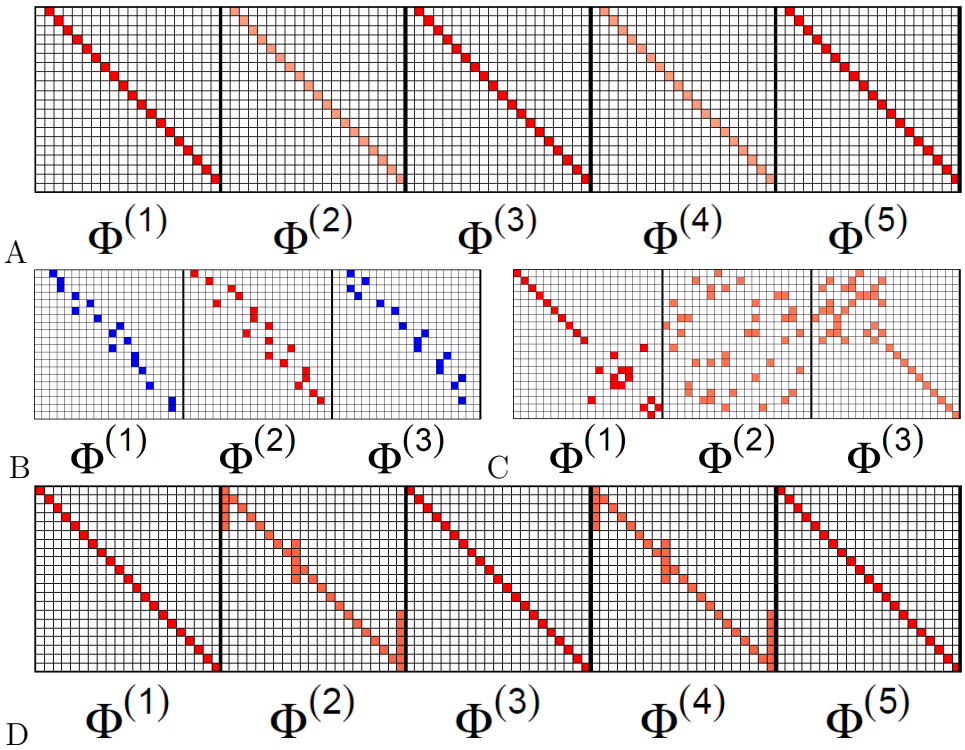


Figure 4-3. True transition matrices for different simulation scenarios.

be due to the fact that the block size in simulation A.3 is large compared to scenarios A.1 and A.2. Further, the estimated location of break points seem to align with the location of true break points in all scenarios as can be seen in the "mean" and "sd" columns of Table 4-2. The *mean* is computed as the average of relative location of estimated break points (i.e. $\widetilde{t}_j/T$) over all replicates while the *sd* is simply the standard deviation of the relative location of estimated change points. Moreover, TBSS yields perfect (100%) sensitivity, almost perfect ($\sim$ 100%) specificity and good ($\sim$ 94%) Matthew's correlation coefficient and fairly small relative error for the estimated transition matrices $\Phi^{(\cdot,j)}$. Of interest in simulation scenario D in which the time lag is 2 while the parameter estimation (close to 90% sensitivity, almost 100% specificity, and almost 90% Matthew's correlation coefficient) and anomaly detection rates (100%) are still very satisfactory. It is worth noting that in simulation C, the computation time is high compared to other simulations with similar sample size T. This is mainly due to the fact of having large number of time series components $p = 40, 60, 100$, and $p = 200$ in simulation C. Note that the number of

model parameters scales quadratically with respect p, i.e. p^2, and such large number of parameters slows down the detection/estimation procedure. Finally, TBSS performs very well in simulation scenarios E and F in which the sparsity pattern of transition matrices are selected randomly and mimics the brain connectivity patterns, respectively. This shows the robustness of the proposed algorithm with respect to changes in the sparsity pattern.

Table 4-2. Simulation results for scenarios A-F: change point selection rate and model parameter estimation.

Case	CP	mean	sd	rate	time	SEN	SPC	MCC	RE	Case	CP	mean	sd	rate	time	SEN	SPC	MCC	RE
A.1	1	0.200	0.000	0.92	21.61	1.00	0.99	0.94	0.048	C.1	1	0.500	0.000	1.00	12.75	1.00	0.99	0.99	0.075
	2	0.400	0.000	0.92						C.2	1	0.500	0.001	1.00	34.48	1.00	1.00	0.99	0.079
	3	0.601	0.002	1.00						C.3	1	0.527	0.012	1.00	630.18	1.00	0.99	0.97	0.11
	4	0.800	0.001	1.00						C.4	1	0.326	0.026	1.00	145.30	1.00	1.00	0.98	0.13
A.2	1	0.200	0.000	1.00	15.08	1.00	0.99	0.93	0.048		2	0.667	0.022	1.00					
	2	0.400	0.000	1.00						D.1	1	0.331	0.005	1.00	13.67	0.87	0.99	0.86	0.42
	3	0.600	0.000	1.00							2	0.666	0.001	1.00					
	4	0.800	0.000	1.00						E.1	1	0.333	0.000	1.00	4.23	1.00	0.96	0.73	0.11
A.3	1	0.200	0.000	0.80	9.16	1.00	0.99	0.94	0.047		2	0.666	0.001	1.00					
	2	0.400	0.000	1.00						E.2	1	0.333	0.001	0.92	4.59	0.98	0.93	0.72	0.30
	3	0.600	0.003	1.00							2	0.664	0.006	0.96					
	4	0.800	0.000	1.00							1	0.200	0.000	0.82					
B.1	1	0.143	0.003	0.92	72.71	1.00	1.00	0.99	0.029	F.1	2	0.400	0.000	1.00	14.19	1.00	0.97	0.82	0.16
	2	0.286	0.000	0.84							3	0.600	0.000	1.00					
	3	0.429	0.000	0.84							4	0.800	0.000	1.00					
	4	0.571	0.000	0.84															
	5	0.714	0.000	0.84															
	6	0.857	0.001	0.92															

4.5 Application to EEG Data for a Visual Task

We use the EEG data[1] analyzed in Trujillo et al. (2017). In this database, EEG signals from active electrodes for 72 channels are recorded at a sampling frequency of 256Hz, for a total of 480 seconds ($T = 122880$). The stimulus procedure tested on 22 subjects (Due to a technical recording error, one subject only received 240 seconds of recording time, and we exclude it from our data set) comprised of eight 1-min duration interleaved sessions with eyes open and closed. The preceding resting state data were removed and not considered in the analysis. The subjects were undergraduates at Texas State University (11 female, 11 male, mean age = 21.1, age range = 18 – 26) and participated in this study for course credit or monetary payment.

Note that there has been work in the literature on related experimental setups, wherein the subjects were asked to keep their eyes open or closed Agcaoglu et al. (2019),

[1]The raw data can be downloaded from: https://dataverse.tdl.org/dataverse/rsed2017.

Weng et al. (2020), Wang et al. (2015), Liu et al. (2020)). Some studies recruited adolescents (e.g. Agcaoglu et al. (2019), Weng et al. (2020)), while others college students (e.g. Wang et al. (2015)). Further, the design of the experiment differed, since the subjects did a session with eyes open and a session with eyes closed, as opposed to switching between them in a single session. Nevertheless, the focus of these studies was to identify both differences in functional connectivity brain networks in these two states, and also track fluctuations in them during the recorded sessions. Further, preprocessing issues were addressed, including discarding initial time points, motion correction and data smoothing.

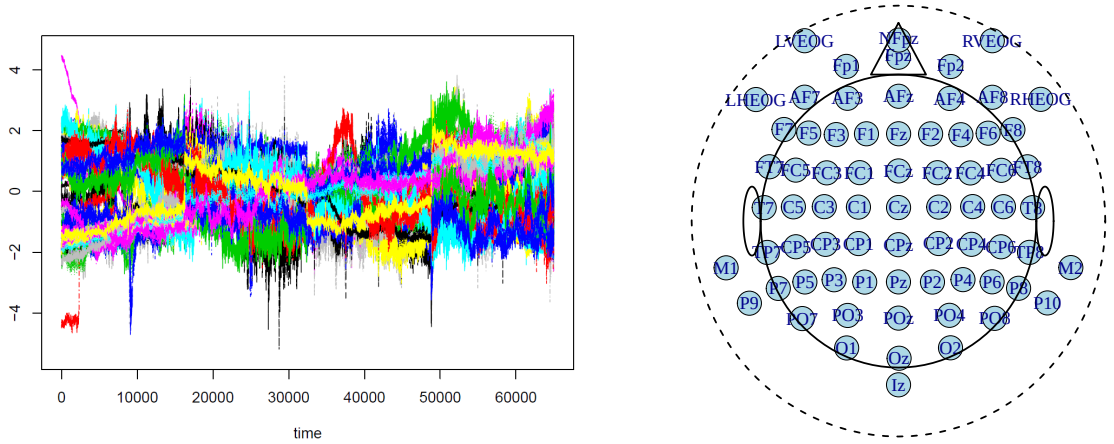
On the other hand, the objective of our analysis is listed as follows:

4.5.1 Data Pre-processing

In our analysis, we collect 65000 observations in the middle of the total recorded period of around 254 seconds (with 3 status switches). According to the experimental setup described in Trujillo et al. (2017), there are three applications of the stimulus (break points) during the selected time period, and the switches between the stimulus states (closed vs open eyes) occur at: $t_1^\star = 16280$, $t_2^\star = 32600$, and $t_3^\star = 48920$, respectively. These are considered as the locations of the *true* break points.

In practice, EEG signals require to be processed by detrending and re-filtering to obtain specific frequency band. The raw EEG signals data are depicted in the right panel in Figure 4-4, and exhibit obvious trend patterns due to known recording artifacts (de Cheveigné & Arzounian 2018). We applied *robust detrending* method proposed in de Cheveigné & Arzounian (2018) to remove the trend patterns in the raw data. Subsequently, we filtered the data in both alpha-band (8-13Hz) and beta-band (14-30Hz), respectively. We examine the selected 71 EEG channels (exclude NAS channel due to the erroneous recording of the signal), and we presented their locations in the right panel of Figure 4-4.

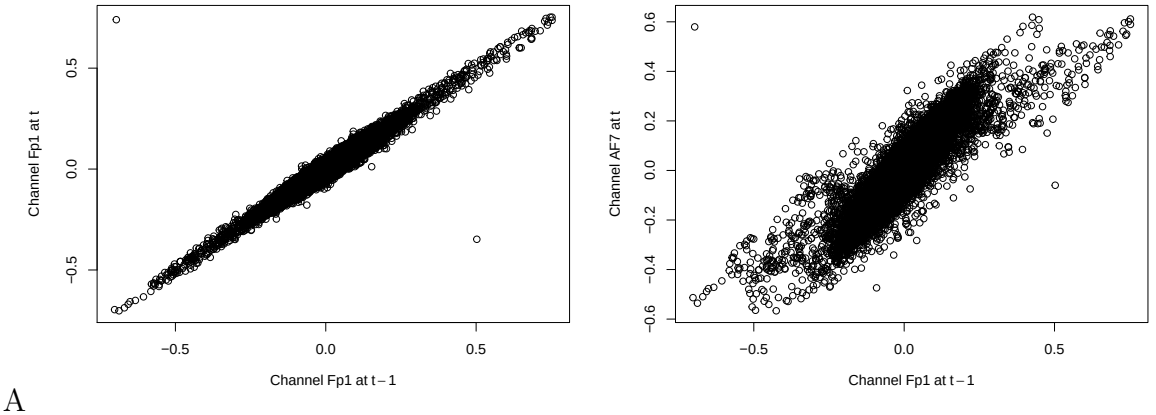
Next, we need to verify that the VAR(1) model exhibits a good fit for the EEG signals data. We firstly consider the following scatter plots as an example: (i) for the value



A B

Figure 4-4. Visualization for EEG channels data set and 71 EEG electrodes locations.

of the i-th channel at time point $t-1$: $x_i(t-1)$, we compare it with the i-th channel at time point t; (ii) for the i-th channel at time point $t-1$: $x_i(t-1)$, we compare it with the j-th channel at time point t, where $j \neq i$. In Figure 4-5, one can observe that the linear relationship between the given channels. For panel A, we consider channel Fp1$(t-1)$ versus Fp1(t), while in panel B, we examine channel Fp1$(t-1)$ versus AF7(t).

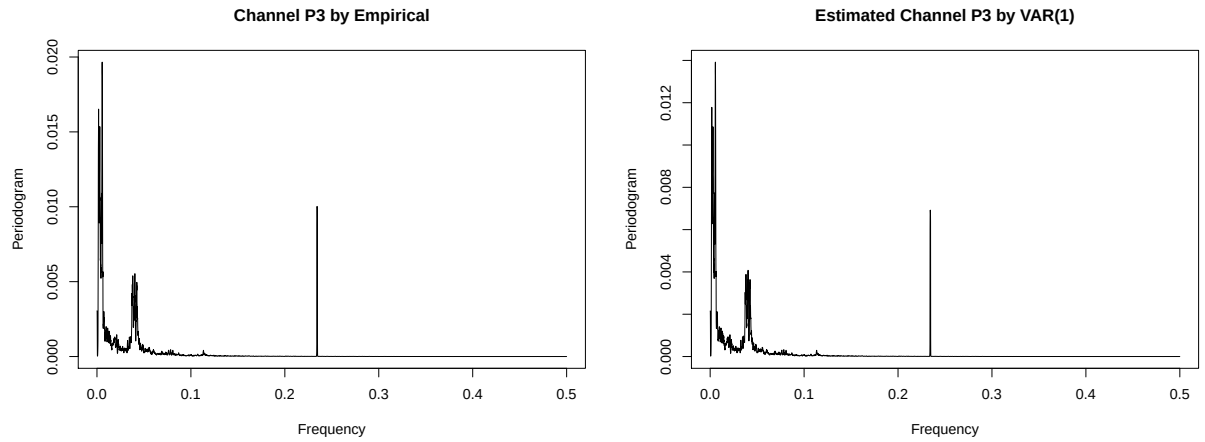


A B

Figure 4-5. Scatter plots for verifying the sufficiency of VAR model.

Moreover, we also compare the periodogram functions for the real data with the estimated data via a VAR(1) model. The following Figure 4-6 illustrates the periodogram functions plots for channel P3 in the real EEG signal data (panel A), and the ones for

estimated signal data by using fitted VAR(1) model (panel B).



A B

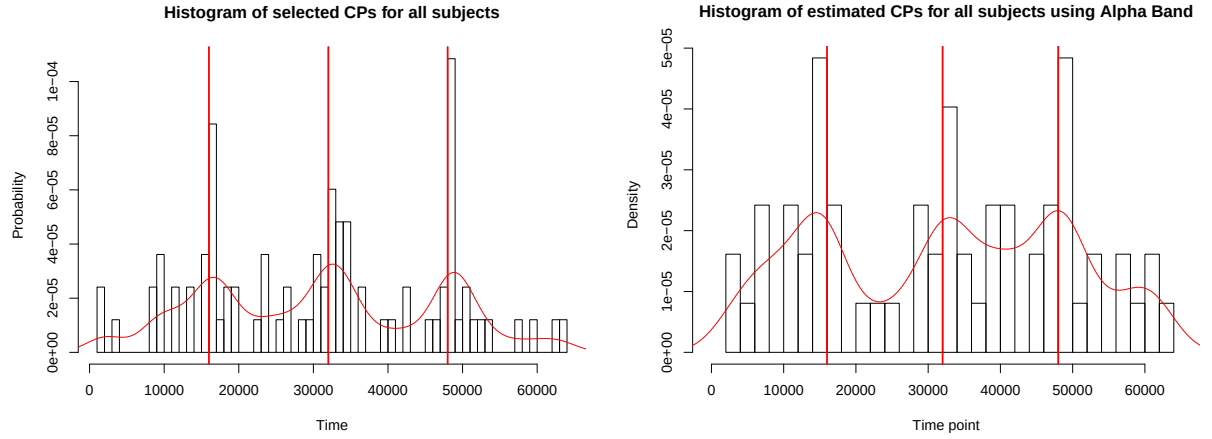
Figure 4-6. Periodogram functions for real EEG channel and estimated P3 by using fitted VAR(1).

4.5.2 Results

To estimate break points for the EEG data of each of the 21 subjects, the block size used in the TBSS algorithm is set to $b_T \in (\lfloor 0.8\sqrt{T} \rfloor, \lfloor 1.2\sqrt{T} \rfloor)$ depending on the subject under consideration, while the other tuning parameters were set according to the guidelines presented in Section 4.4.1. Further, based on the verification in previous Section 4.5.1, VAR model with single time lag exhibited a better fit in terms of mean squared error than a model with more time lags for the majority of the subjects, and thus it was fitted to all subjects.

Figure 4-7 presents the histogram of all estimated break points from all 21 subjects by using the all frequency bands (left panel) and using the specific alpha-band (right panel). The red curve depicts the estimated density of the histogram. It can be clearly seen that the majority of the break points (local maxima) are approximately located at time points: 16000, 32000, and 48000, which are in accordance with the true changes in the stimulus. Specifically, the estimated change points by using all frequency bands are: $\widehat{t}_1 = 15709$, $\widehat{t}_2 = 31555$, and $\widehat{t}_3 = 48499$, and the estimated change points by using alpha-band are: $\widehat{t}_1^{\alpha} = 16000$, $\widehat{t}_2^{\alpha} = 32000$, and $\widehat{t}_3^{\alpha} = 48000$. The estimated locations are

satisfactory and close to the truth.



A B

Figure 4-7. Histogram of estimated break points for all 21 subjects; the red curve depicts their estimated density function and the red vertical lines represent the location of the time points when the stimulus switched.

Hence, four stationary segments are obtained for most of the subjects and the corresponding Granger causal networks for open/close-eye intervals are estimated using Step 5 of the TBSS algorithm. In Figure 4-8, we presented the averaged estimated Granger causal networks over all open (left column) and close (right column) segments for all 21 subjects by using the *all frequency bands* data (top row), *alpha band* data (middle row), and *beta band* data (bottom row), respectively. For ease of presentation of the network structure, we only plot the edges with adequately large magnitudes, and removed the edges with small weights.

To be more precise, we provided the following Table 4-3 to present the averaged number of edges for each segments together with the top five most connected channels by using different frequency bands.

We can see from Table 4-3 that the networks for closed-eye segments are denser than the ones for open-eye segments, which can be observed in Figure 4-8 as well. Besides the sparsity levels, the most connected channels are different for open-eye and close-eye segments and different for frequency bands. Note that channels identified as strongly connected by the analysis have been mentioned in vision-related tasks in the literature

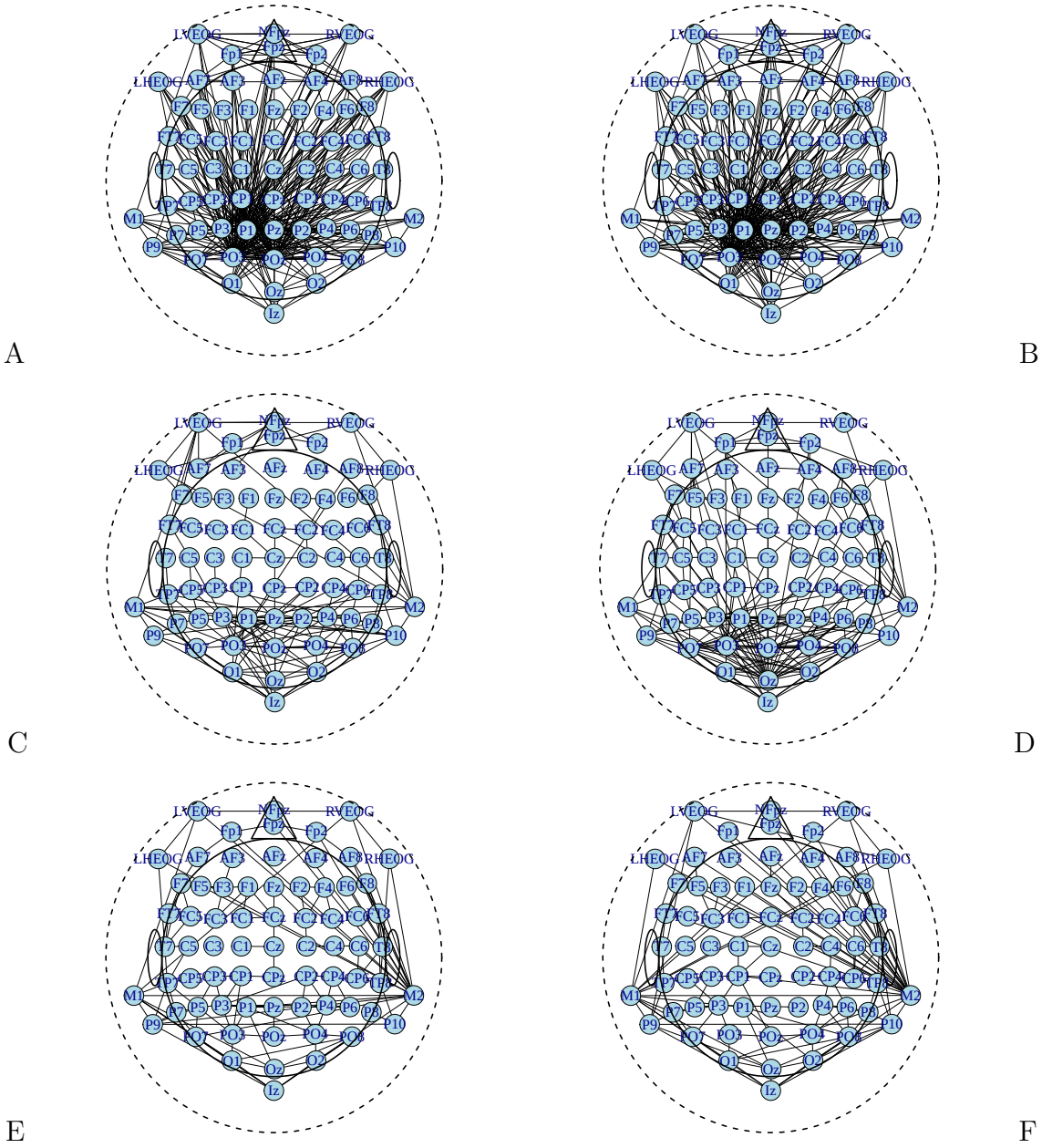


Figure 4-8. Estimated Granger networks for the open and close segments using different frequency band data.

Nezamfar et al. (2011), Das et al. (2016). For example, channel Oz exhibits a high connectivity pattern overall while it has higher connectivity in closed-eye segments. This channel is located within the primary visual cortex (V1) which is the most studied visual area in the brain Arden et al. (2003), Sharon et al. (2007).

In addition, it can be seen that the average degree in the closed-eye state is

Table 4-3. Results for the estimated networks for each segments for all 21 subjects.

Data	Networks	No. edges	Top 5 most connected channels (degrees)
All frequency bands	1(C)	401	PO3(63), P1(56), P2(56), PO4(37), P6(30)
	2(O)	337	P1(54), P2(54), PO3(50), POz(18), P4(18)
	3(C)	408	PO3(69), P2(55), PO4(37), P1(36), P6(33)
	4(O)	370	PO3(53), P1(43), P2(42), P6(38), CP1(25)
Alpha band	1(C)	161	Pz(43), Oz(24), P2(10), CP5(6), CP1(6)
	2(O)	247	P9(41), Pz(28), PO4(22), PO7(19), P1(15)
	3(C)	160	Oz(52), P5(13), P3(7), PO4(7), Iz(4)
	4(O)	267	Pz(44), PO4(40), Oz(35), POz(28), CPz(17)
Beta band	1(C)	151	CP2(21), CPz(16), F1(12), FC1(12), C1(7)
	2(O)	117	F2(8), CP4(7), FC1(6), CP6(6), F3(5)
	3(C)	131	PO4(18), P3(10), AFz(7), P2(6), M1(5)
	4(O)	97	P3(19), Oz(4), M1(3), Iz(2), TP8(2)

significantly higher than in the open-eye state. Specifically, in the estimated networks by using all frequency bands and alpha band, the channels at the visual cortex (PO3, P1, P2, Oz, Pz, etc.) Nezamfar et al. (2011) have high connectivity overall while the beta band captures more signals from the channels at the forehead (F1, FC1, F2, AFz, etc.) Klimesch (1999), Chen et al. (2008), Tan et al. (2013). Moreover, we provided boxplots in Figure 4-9 for all EEG channels in their close/open segments over all subjects.

Figure 4-9 presents a thorough detailed comparison for all EEG channels. It can be seen that the eye-related channels have large changes during these close/open changes. In the central cortex, channels CP1, CP2, FC1, FC2, and Cz shift most between close and open segments. While in the central left and right cortex regions, we can see that all EEG channels located in these two regions don't change extremely. According to Figure 4-8, we have known that the EEG channels located at post head are varying most, which is consistent to the boxplots for the post head channels in the bottom panel.

These findings indicate that the estimated Granger causal networks on the corresponding detected stationary segments are interpretable and can provide useful insights to neuroscientists.

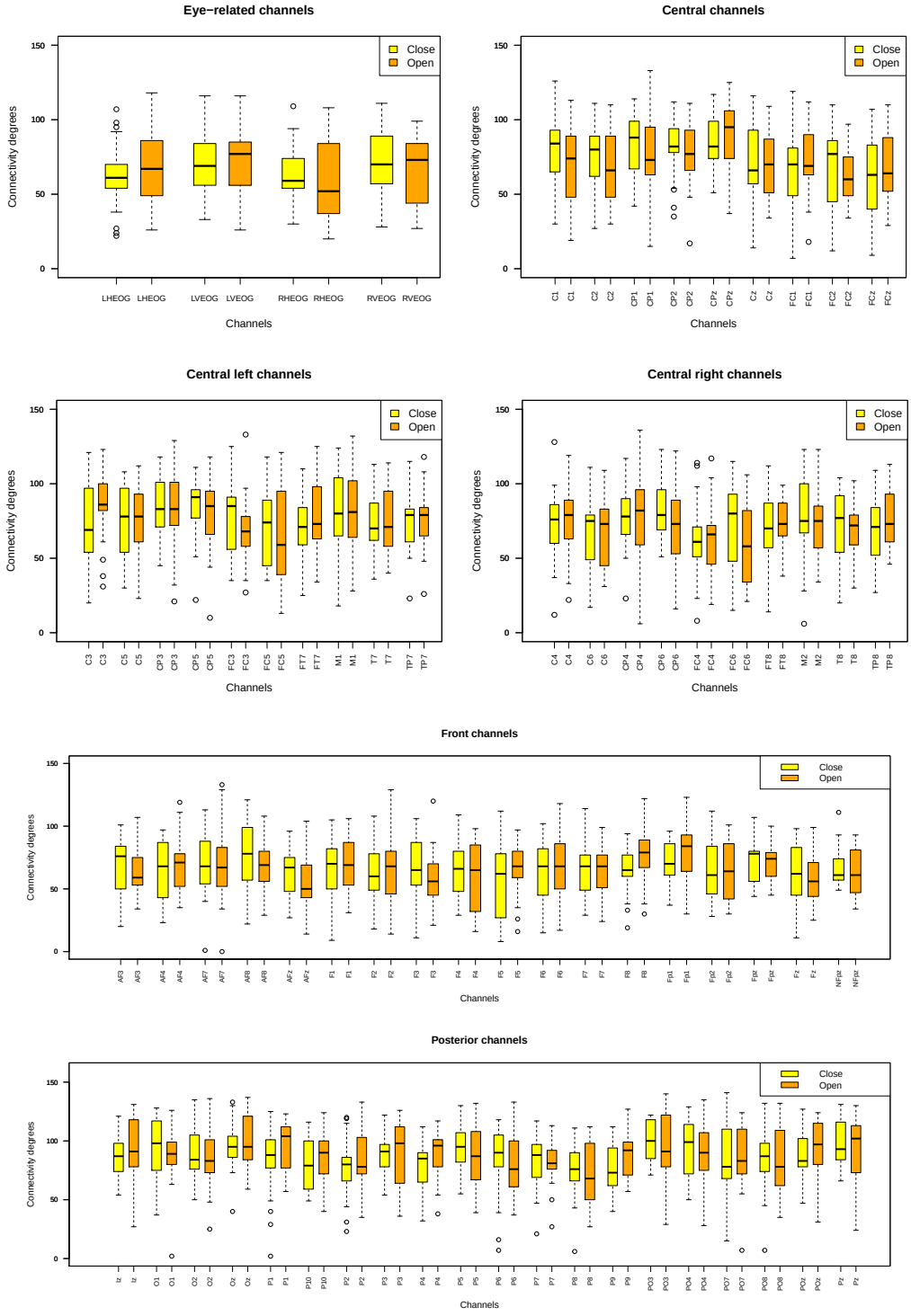


Figure 4-9. Boxplots for all channels in close/open segments over all 21 subjects.

4.6 Discussion

In this section, we presents a five step procedure (TBSS) for break point detection and model parameter estimation in high-dimensional piece-wise stationary VAR models.

TBSS is scalable to exceedingly large data sets, since with appropriate choice of the block size $b_T \sim \sqrt{T}$ achieves *sublinear* computational complexity in the number of time points. Further, it obtains very accurate estimation of both the number and the location of break points. The current presentation has focused on detection and estimation of changes in the auto-regressive parameters and ignored the impact of the contemporaneous covariance matrix Σ_j that may be changing across time segments (see model formulation in (4.1)). Its impact can be easily accommodated by modifying the least squares criterion in (4-4) by a generalized least squares criterion (for details and algorithms, see Lin & Michailidis (2017)).

The key idea underlying TBSS is to examine blocks of time points which leads to substantial computational gains. However, the selection of the block size b_T is a critical tuning parameter. On many occasions, like the EEG data application presented in Section 4.5, the analyst has information about the spacing between consecutive break points that help select b_T. In absence of such information, a data-driven approach can be adopted as follows: apply TBSS with block sizes of order $T^{0.5-\frac{j}{2K}}$ for large enough K, starting from $j = 0$, and continue increasing j until the number of selected break points in Step 4 for j and $j + 1$ coincide (obviously, $j < K$).

Another advantage of TBSS is that it provides regularized estimates of the underlying Granger causal networks (the $\Phi^{(.,j)}$ autoregressive parameters) in high-dimensional regime across all stationary segments. It is of interest to develop *testing procedures* for identifying significant changes in individual entries (or groups of entries) of these networks.

4.7 Technical Proofs for Main Theorems in Chapter 4

In this section, we provide the sketch of the proofs for proposed Theorem 4-1, 4-2, and 4-3.

Proof of Theorem 4-1. The proof is based on arguing by contradiction. Suppose there exists an estimated break point $\widehat{t}_j$ such that $|\widehat{t}_j - t_j| > T\gamma_T$. In other words, there exists a true break point t_{j_0}, which is isolated from all the estimated points (there are no estimated break points close to t_{j_0}).

Based on the optimization problem in (4-4), the value of the function is minimized exactly at $\widehat{\Theta}$. Therefore, for the interval $[t_{j_0-1}, t_{j_0}]$, we denote the closest block end-point r_i to the right side of t_{j_0-1} by s_{j_0-1}, and similarly, denote the closest r_i to the left side of t_{j_0} by s_{j_0}. Define a parameter sequence ψ_k's, $k = 1, 2, \cdots, k_T$ with $\psi_k = \widehat{\theta}_k$ except for two time points $k = \widehat{t}_j$ and $k = t_{j_0}$. In this proof, we focus on the interval $[s_{j_0-1} \vee \widehat{t}_j, s_{j_0}]$. Define the following parameter sequence ψ_k, for $k = 1, 2, \cdots, k_T$ with $\psi_k = \widehat{\theta}_k$ except for two points $k = \widehat{i}_j \vee i_{j_0-1}$ and $k = i_{j_0}$. For these two points, if $\widehat{t}_j > s_{j_0-1}$, then we set $\psi_{\widehat{i}_j} = \Phi^{(\cdot,j_0)} - \widehat{\Phi}_j$ and $\psi_{i_{j_0}} = \widehat{\Phi}_{j+1} - \Phi^{(\cdot,j_0)}$; else if $\widehat{t}_j \leq s_{j_0-1}$, then we set $\psi_{i_{j_0-1}} = \Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}$ and $\psi_{i_{j_0}} = \widehat{\Phi}_{j+1} - \Phi^{(\cdot,j_0)}$. Then, we have

$$\begin{aligned} &\frac{1}{T}\|\mathbf{Y} - \mathbf{Z}\widehat{\Theta}\|_2^2 + \lambda_{1,T}\|\widehat{\Theta}\|_1 + \lambda_{2,T}\sum_{i=1}^{k_T}\left\|\sum_{j=1}^{i}\widehat{\theta}_j\right\|_1 \\ &\qquad \leq \frac{1}{T}\|\mathbf{Y} - \mathbf{Z}\Psi\|_2^2 + \lambda_{1,T}\|\Psi\|_1 + \lambda_{2,T}\sum_{i=1}^{k_T}\left\|\sum_{j=1}^{i}\psi_j\right\|_1. \end{aligned} \tag{4-9}$$

After some algebraic rearrangements for (4-9), we obtain that

$$\begin{aligned} 0 &\leq c_0\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_2^2 \\ &\leq \frac{2}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\sum_{l=s_{j_0-1}\vee\widehat{t}_j}^{s_{j_0}} Y'_{l-1}\left(\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\right)\epsilon_l \\ &+ \frac{T\lambda_{1,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\left(\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_1\right) + \frac{T\lambda_{2,T}}{b_T}\left(\|\Phi^{(\cdot,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\ &\leq \left(\frac{2T\lambda_{1,T}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} + C\sqrt{\frac{\log p}{T\gamma_T}}\right)\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \frac{T\lambda_{2,T}}{b_T}\left(\|\Phi^{(\cdot,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\ &\leq \frac{3}{2}\frac{T\lambda_{2,T}}{b_T}\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}} - \frac{1}{2}\frac{T\lambda_{2,T}}{b_T}\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}}, \end{aligned} \tag{4-10}$$

which implies that

$$\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_F = \mathcal{O}_p\left(\sqrt{\frac{d^\star_{\max}\log p}{T\gamma_T}}\right), \tag{4-11}$$

which indicates that $\|\Phi^{(\cdot,j_0)} - \widehat{\Phi}_{j+1}\|_F$ converges to zero in probability. Similarly, we can prove that in the interval $[s_{j_0}, s_{j_0+1} \wedge \widehat{t}_{j+1}]$, the quantity $\|\Phi^{(\cdot,j_0+1)} - \widehat{\Phi}_{j+1}\|_F$ converges to

zero as well. This contradicts Assumption A3, which completes the proof. □

Proof of Theorem 4-2. The main idea of this proof is to show that the solution $\widetilde{t}_j$ to (4-7) minimizes the objective function (4-12) defined below in the given search interval (l_j, u_j). Based on the step 2 and 3, we know that there exists a true break point t_j in the search interval (l_j, u_j). Denote the objective function $\mathcal{L}(\tau)$ as follows:

$$\mathcal{L}(\tau) = \frac{1}{u_j - l_j} \left(\sum_{t=l_j}^{\tau-1} \|y_{t+1} - \widetilde{\Phi}^{(\cdot,j)} Y_t\|_2^2 + \sum_{t=\tau}^{u_j-1} \|y_{t+1} - \widetilde{\Phi}^{(\cdot,j+1)} Y_t\|_2^2 \right). \tag{4-12}$$

According to the definition of $\widetilde{t}_j$, we have the basic inequality $\mathcal{L}(\widetilde{t}_j) \leq \mathcal{L}(t_j)$, then by using this basic inequality and the similar argument as Proposition 4.1 in Basu & Michailidis (2015), we are able to derive that:

$$\|\widetilde{\Phi}^{(\cdot,j)} - \Phi^{(\cdot,j)}\|_1 \leq 4\sqrt{d^\star_{\max}} \|\widetilde{\Phi}^{(\cdot,j)} - \Phi^{(\cdot,j)}\|_F, \ \|\widetilde{\Phi}^{(\cdot,j+1)} - \Phi^{(\cdot,j+1)}\|_1 \leq 4\sqrt{d^\star_{\max}} \|\widetilde{\Phi}^{(\cdot,j+1)} - \Phi^{(\cdot,j+1)}\|_F. \tag{4-13}$$

Moreover, we obtain that

$$\|\widetilde{\Phi}^{(\cdot,j)} - \Phi^{(\cdot,j)}\|_F = \mathcal{O}_p\left(\sqrt{\frac{d^\star_{\max} \log p}{u_j - l_j}}\right), \ \|\widetilde{\Phi}^{(\cdot,j+1)} - \Phi^{(\cdot,j+1)}\|_F = \mathcal{O}_p\left(\sqrt{\frac{d^\star_{\max} \log p}{u_j - l_j}}\right). \tag{4-14}$$

Using the results in (4-13) and (4-14) leads to the following result: assume that $\widetilde{t}_j > t_j$, then we denote

$$(u_j - l_j)\mathcal{L}(\widetilde{t}_j) = \sum_{t=l_j}^{t_j-1} \|y_t - \widetilde{\Phi}^{(\cdot,j)} Y_t\|_2^2 + \sum_{t=t_j}^{u_j-1} \|y_t - \widetilde{\Phi}^{(\cdot,j+1)} Y_t\|_2^2 \stackrel{\text{def}}{=} I_1 + I_2.$$

Now, we first provide the lower bounds for both I_1 and I_2 by using similar arguments in Lemma 5 case (c) in Safikhani & Shojaie (2020):

$$\begin{aligned} I_1 &\geq \sum_{t=l_j}^{t_j-1} \|\epsilon_t\|_2^2 + c_1|\widetilde{t}_j - t_j| - c_2 d^\star_{\max} \log p, \\ I_2 &\geq \sum_{t=t_j}^{u_j-1} \|\epsilon_t\|_2^2 - c_3 d^\star_{\max} \log p. \end{aligned} \tag{4-15}$$

Hence, combining two parts I_1 and I_2 in (4-15) implies that

$$\mathcal{L}(\widetilde{t}_j) \geq \sum_{t=l_j}^{u_j-1} \|\epsilon_t\|_2^2 + K_1|\widetilde{t}_j - t_j| - K_2 d^{\star}_{\max} \log p, \tag{4-16}$$

where K_1 and K_2 are some positive constants.

Next, we claim the upper bound for the objective function at true break point t_j:

$$\mathcal{L}(t_j) \leq \sum_{t=l_j}^{u_j-1} \|\epsilon_t\|_2^2 + K d^{\star}_{\max} \log p. \tag{4-17}$$

To show this result, we use the similar procedure as of the proof of Theorem 3 in Safikhani & Shojaie (2020). Therefore, by using (4-16) and (4-17) together with the basic inequality, we obtain:

$$\sum_{t=l_j}^{u_j-1} \|\epsilon_t\|_2^2 + K_1|\widetilde{t}_j - t_j| - K_2 d^{\star}_{\max} \log p \leq \mathcal{L}(\widetilde{t}_j) \leq \mathcal{L}(t_j) \leq \sum_{t=l_j}^{u_j-1} \|\epsilon_t\|_2^2 + K d^{\star}_{\max} \log p,$$

which leads to the final result. □

Proof of Theorem 4-3. Theorem 4-3 provides consistent estimators for the VAR model parameters. This proof is similar to that of Proposition 4.1 in Basu & Michailidis (2015). The key steps in the proof that require verification are: (1) the restricted strong convexity (RSC) condition; (2) the deviation bounds for the estimation intervals. For (1), we use analogous arguments as in the proof of Theorem 4 in Safikhani & Shojaie (2020) to establish the result. Further, (2) follows Lemma 1 in Safikhani & Shojaie (2020). □

CHAPTER 5
CONCLUSIONS

In this dissertation, the multiple change points detection problem in high dimensional time series models is studied in various contexts, as well as some relating detection algorithms proposed.

In Chapter 2, we investigate multiple change points detection in a specific structure high dimensional VAR model, where the transition matrices are decomposed into a fixed low rank component and time-varying sparse components. We developed a three-step (BFL) detection algorithm together with estimating the transition matrices in VAR model. We provide theoretical properties of estimated change points and model parameters. Besides the theories, we also evaluated the proposed algorithm in both synthetic data sets and real-world data sets.

In Chapter 3, we generalized the proposed model in Chapter 2 to both time-varying low rank plus sparse components. We started from single change point scenario and provided a detection procedure by utilizing the minimizer of SSE, then we extended to the multiple change points scenario and established a two-step rolling window strategy. From theoretical perspective, we show the error bound for the estimated change points using Hausdorff distance, as well as the consistency in the estimated model parameters. In order to reduce the computational complexity, we proposed a misspecified surrogate weakly sparse model, we also provided the change points estimation procedure together with theoretical properties. At last, we measured the performance of the algorithm in numerical experiments and applied to application scenarios in macroeconomics and neuroscience, respectively.

In Chapter 4, we focused on developing a fast change point detection algorithm to handle exceedingly long time series. We presented a five-step algorithm (TBSS) to detect change points as well as estimate model parameters in high dimensional piece-wise stationary VAR models. TBSS is scalable to large data sets, while it ensures accurate estimation of both the number and the location of change points. We provided the TBSS algorithm paradigm in this chapter and we evaluated the performance by some designed

numerical experiments and we applied the algorithm to a EEG signals data set. We illustrated the strength of TBSS in detecting change points and estimating Granger causal networks for each detected segments compared with other existed algorithms and strategies.

In general, the change points detection in a structured high dimensional vector autoregressive (VAR) models has been thoroughly studied in this work. The reduced rank and sparse structure provides appealing statistical performance in theoretical perspective, and also illustrates wide range of solving application problems. The proposed fast and scalable algorithms is suitable for many application scenarios from macroeconomics to neuroscience.

Adams, R. P. & MacKay, D. J. (2007), 'Bayesian online changepoint detection', *arXiv preprint arXiv:0710.3742* .

Agarwal, A., Negahban, S., Wainwright, M. J. et al. (2012), 'Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions', *The Annals of Statistics* **40**(2), 1171–1197.

Agcaoglu, O., Wilson, T. W., Wang, Y.-P., Stephen, J. & Calhoun, V. D. (2019), 'Resting state connectivity differences in eyes open versus eyes closed conditions', *Human Brain Mapping* **40**(8), 2488–2498.

Arden, G., Wolf, J. & Messiter, C. (2003), 'Electrical activity in visual cortex associated with combined auditory and visual stimulation in temporal sequences known to be associated with a visual illusion', *Vision Research* **43**(23), 2469–2478.

Bai, J. (1997), 'Estimating multiple breaks one at a time', *Econometric theory* **13**(3), 315–352.

Bai, J. & Ng, S. (2008), 'Large dimensional factor analysis', *Foundations and Trends® in Econometrics* **3**(2), 89–163.

Bai, P., Safikhani, A. & Michailidis, G. (2020), 'Multiple change points detection in low rank and sparse high dimensional vector autoregressive models', *IEEE Transactions on Signal Processing* **68**, 3074–3089.

Bańbura, M., Giannone, D. & Reichlin, L. (2010), 'Large bayesian vector auto regressions', *Journal of applied Econometrics* **25**(1), 71–92.

Bardsley, P., Horváth, L., Kokoszka, P. & Young, G. (2017), 'Change point tests in functional factor models with application to yield curves', *The Econometrics Journal* **20**(1), 86–117.

Barigozzi, M., Cho, H. & Fryzlewicz, P. (2018), 'Simultaneous multiple change-point and factor analysis for high-dimensional time series', *Journal of Econometrics* **206**(1), 187–225.

Basseville, M. & Nikiforov, I. V. (1993), *Detection of abrupt changes: theory and application*, Vol. 104, Prentice Hall Englewood Cliffs.

Basu, S., Li, X. & Michailidis, G. (2019), 'Low rank and structured modeling of high-dimensional vector autoregressions', *IEEE Transactions on Signal Processing* **67**(5), 1207–1222.

Basu, S. & Michailidis, G. (2015), 'Regularized estimation in sparse high-dimensional time series models', *The Annals of Statistics* **43**(4), 1535–1567.

Basu, S., Shojaie, A. & Michailidis, G. (2015), 'Network granger causality with inherent grouping structure', *The Journal of Machine Learning Research* **16**(1), 417–453.

Billio, M., Getmansky, M., Lo, A. W. & Pelizzon, L. (2012), 'Econometric measures of connectedness and systemic risk in the finance and insurance sectors', *Journal of financial economics* **104**(3), 535–559.

Bleakley, K. & Vert, J.-P. (2011), 'The group fused lasso for multiple change-point detection', *arXiv preprint arXiv:1106.4199* .

Bordo, M. D. & Eichengreen, B. (2007), *A retrospective on the Bretton Woods system: lessons for international monetary reform*, University of Chicago Press.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chan, N. H., Ing, C.-K., Li, Y. & Yau, C. Y. (2017), 'Threshold estimation via group orthogonal greedy algorithm', *Journal of Business & Economic Statistics* **35**(2), 334–345.

Chan, N. H., Yau, C. Y. & Zhang, R.-M. (2014), 'Group lasso for structural break time series', *Journal of the American Statistical Association* **109**(506), 590–599.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. & Willsky, A. S. (2011), 'Rank-sparsity incoherence for matrix decomposition', *SIAM Journal on Optimization* **21**(2), 572–596.

Chen, A. C., Feng, W., Zhao, H., Yin, Y. & Wang, P. (2008), 'Eeg default mode network in the human brain: spectral regional field powers', *Neuroimage* **41**(2), 561–574.

Cho, H. & Fryzlewicz, P. (2015), 'Multiple-change-point detection for high dimensional time series via sparsified binary segmentation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(2), 475–507.

Das, R., Maiorana, E. & Campisi, P. (2016), 'Eeg biometrics using visual stimuli: A longitudinal study', *IEEE Signal Processing Letters* **23**(3), 341–345.

de Cheveigné, A. & Arzounian, D. (2018), 'Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data', *Neuroimage* **172**, 903–912.

Eichengreen, B. (2014), *Hall of mirrors: The great depression, the great recession, and the uses-and misuses-of history*, Oxford University Press.

Fama, E. F. & French, K. R. (1992), 'The cross-section of expected stock returns', *the Journal of Finance* **47**(2), 427–465.

Fama, E. F. & French, K. R. (1996), 'Multifactor explanations of asset pricing anomalies', *The journal of finance* **51**(1), 55–84.

Fama, E. F. & French, K. R. (2015), 'A five-factor asset pricing model', *Journal of financial economics* **116**(1), 1–22.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

Frisén, M. (2008), *Financial surveillance*, Vol. 71, John Wiley & Sons.

Friston, K. J., Bastos, A. M., Oswal, A., van Wijk, B., Richter, C. & Litvak, V. (2014), 'Granger causality revisited', *Neuroimage* **101**, 796–808.

Friston, K., Moran, R. & Seth, A. K. (2013), 'Analysing connectivity with granger causality and dynamic causal modelling', *Current opinion in neurobiology* **23**(2), 172–178.

Harchaoui, Z. & Lévy-Leduc, C. (2010), 'Multiple change-point estimation with a total variation penalty', *Journal of the American Statistical Association* **105**(492), 1480–1493.

Hartigan, J. A. & Wong, M. A. (1979), 'Algorithm as 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

Hsu, D., Kakade, S. M. & Zhang, T. (2011), 'Robust matrix decomposition with sparse corruptions', *IEEE Transactions on Information Theory* **57**(11), 7221–7234.

Jao, P.-K., Chavarriaga, R. & Millán, J. d. R. (2018), Using robust principal component analysis to reduce eeg intra-trial variability, *in* '40th Annual Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 1956–1959.

Kareken, J. H. (1978), 'Inflation: an extreme view', *Quarterly Review* (win).

Kilian, L. & Lütkepohl, H. (2017), *Structural vector autoregressive analysis*, Cambridge University Press.

Killick, R., Fearnhead, P. & Eckley, I. A. (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association* **107**(500), 1590–1598.

Klimesch, W. (1999), 'Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis', *Brain research reviews* **29**(2-3), 169–195.

Koepcke, L., Ashida, G. & Kretzberg, J. (2016), 'Single and multiple change point detection in spike trains: Comparison of different cusum methods', *Frontiers in systems neuroscience* **10**, 51.

Lam, C., Yao, Q. & Bathia, N. (2011), 'Estimation of latent factors for high-dimensional time series', *Biometrika* **98**(4), 901–918.

Li, G., Qin, S. J. & Zhou, D. (2014), 'A new method of dynamic latent-variable modeling for process monitoring', *IEEE Transactions on Industrial Electronics* **61**(11), 6438–6445.

Lin, J. & Michailidis, G. (2017), 'Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models', *The Journal of Machine Learning Research* **18**(1), 4188–4236.

Liu, F., Wang, S., Qin, J., Lou, Y. & Rosenberger, J. (2018), Estimating latent brain sources with low-rank representation and graph regularization, *in* 'International Conference on Brain Informatics', Springer, pp. 304–316.

Liu, X., Wu, X., Zhong, M., Huang, H., Weng, Y., Niu, M., Zhao, L. & Huang, R. (2020), 'Dynamic properties of human default mode network in eyes-closed and eyes-open', *Brain Topography* **33**(6), 720–732.

Loh, P.-L. & Wainwright, M. J. (2012), 'High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity', *The Annals of Statistics* pp. 1637–1664.

Lütkepohl, H. (2013), *Introduction to multiple time series analysis*, Springer Science & Business Media.

McCracken, M. W. & Ng, S. (2016), 'Fred-md: A monthly database for macroeconomic research', *Journal of Business & Economic Statistics* **34**(4), 574–589.
URL: *https://doi.org/10.1080/07350015.2015.1086655*

McGlohon, M., Bay, S., Anderle, M. G., Steier, D. M. & Faloutsos, C. (2009), Snare: a link analytic system for graph labeling and risk detection, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1265–1274.

Meucci, A. (2009), *Risk and asset allocation*, Springer Science & Business Media.

Michailidis, G. & d'Alché Buc, F. (2013), 'Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues', *Mathematical biosciences* **246**(2), 326–334.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012), 'A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers', *Statistical Science* **27**(4), 538–557.

Nezamfar, H., Orhan, U., Purwar, S., Hild, K., Oken, B. & Erdogmus, D. (2011), 'Decoding of multichannel eeg activity from the visual cortex in response to pseudorandom binary sequences of visual stimuli', *International Journal of Imaging Systems and Technology* **21**(2), 139–147.

Nobre, F. F. & Stroup, D. F. (1994), 'A monitoring system to detect changes in public health surveillance data', *International journal of epidemiology* **23**(2), 408–418.

Orphanides, A. (2004), 'Monetary policy rules, macroeconomic stability, and inflation: A view from the trenches', *Journal of Money, Credit and Banking* pp. 151–175.

Qiu, P. (2013), *Introduction to statistical process control*, Chapman and Hall/CRC.

Rinaldo, A. et al. (2009), 'Properties and refinements of the fused lasso', *The Annals of Statistics* **37**(5B), 2922–2952.

Roy, S., Atchadé, Y. & Michailidis, G. (2017), 'Change point estimation in high dimensional markov random-field models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(4), 1187–1206.

Safikhani, A. & Shojaie, A. (2020), 'Joint structural break detection and parameter estimation in high-dimensional non-stationary var models', *Journal of the American Statistical Association (Theory and Methods), to appear* .

Schröder, A. L. & Ombao, H. (2019), 'Fresped: Frequency-specific change-point detection in epileptic seizure multi-channel eeg data', *Journal of the American Statistical Association* **114**(525), 115–128.

Sharon, D., Hämäläinen, M. S., Tootell, R. B., Halgren, E. & Belliveau, J. W. (2007), 'The advantage of combining meg and eeg: comparison to fmri in focally stimulated visual cortex', *Neuroimage* **36**(4), 1225–1235.

Stock, J. H. & Watson, M. W. (2002), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Stock, J. H. & Watson, M. W. (2016), Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics, *in* 'Handbook of macroeconomics', Vol. 2, Elsevier, pp. 415–525.

Tan, B., Kong, X., Yang, P., Jin, Z. & Li, L. (2013), 'The difference of brain functional connectivity between eyes-closed and eyes-open using graph theoretical analysis', *Computational and mathematical methods in medicine* **2013**.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005), 'Sparsity and smoothness via the fused lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

Trujillo, L. T., Stanfield, C. T. & Vela, R. D. (2017), 'The effect of electroencephalogram (eeg) reference choice on information-theoretic measures of the complexity and integration of eeg signals', *Frontiers in neuroscience* **11**, 425.

Velu, R. P., Reinsel, G. C. & Wichern, D. W. (1986), 'Reduced rank models for multiple time series', *Biometrika* **73**(1), 105–118.

Wang, D., Yu, Y., Rinaldo, A. & Willett, R. (2019), 'Localizing changes in high-dimensional vector autoregressive processes', *arXiv preprint arXiv:1909.06359* .

Wang, D., Yu, Y., Rinaldo, A. et al. (2021), 'Optimal change point detection and localization in sparse dynamic networks', *Annals of Statistics* **49**(1), 203–232.

Wang, X.-H., Li, L., Xu, T. & Ding, Z. (2015), 'Investigating the temporal patterns within and between intrinsic connectivity networks under eyes-open and eyes-closed resting states: a dynamical functional connectivity study based on phase synchronization', *PloS one* **10**(10), e0140300.

Wang, Z. & Bessler, D. A. (2004), 'Forecasting performance of multivariate time series models with full and reduced rank: An empirical examination', *International Journal of Forecasting* **20**(4), 683–695.

Weng, Y., Liu, X., Hu, H., Huang, H., Zheng, S., Chen, Q., Song, J., Cao, B., Wang, J., Wang, S. et al. (2020), 'Open eyes and closed eyes elicit different temporal properties of brain functional networks', *NeuroImage* **222**, 117230.

Yu, Y., Padilla, O. H. M., Wang, D. & Rinaldo, A. (2020), 'A note on online change point detection', *arXiv preprint arXiv:2006.03283* .

Zameni, M., Sadri, A., Ghafoori, Z., Moshtaghi, M., Salim, F. D., Leckie, C. & Ramamohanarao, K. (2020), 'Unsupervised online change point detection in high-dimensional time series', *Knowledge and Information Systems* **62**(2), 719–750.

Zhang, C.-H. & Huang, J. (2008), 'The sparsity and bias of the lasso selection in high-dimensional linear regression', *The Annals of Statistics* **36**(4), 1567–1594.

BIOGRAPHICAL SKETCH

Peiliang Bai was born and raised in Urumqi, Xinjiang, China. He received his bachelor's degree in mathematics and statistics from Peking University in China in July 2015. After that, he enrolled in a graduate program in the Department of Statistics at the University of Florida. He earned his master's degree in statistics in the spring of 2017, and Ph.D. in statistics in the spring of 2021.