JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2022.2075369

Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models

Peiliang Bai[a], Abolfazl Safikhani[a], and George Michailidis[a,b,c]

[a]Department of Statistics, University of Florida, Gainesville, FL; [b]Computer and Information Science Engineering, University of Florida, Gainesville, FL; [c]Informatics Institute, University of Florida, Gainesville, FL

ABSTRACT

We study the problem of detecting and locating change points in high-dimensional Vector Autoregressive (VAR) models, whose transition matrices exhibit low rank plus sparse structure. 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. Next, we extend the results to the case of multiple change points that can grow as a function of the sample size. Their detection is based on a two-step algorithm, wherein the first step, an exhaustive search for a candidate change point is employed for overlapping windows, and subsequently a backward elimination procedure is used to screen out redundant candidates. The two-step strategy yields consistent estimates of the number and the locations of the change points. To reduce computation cost, we also investigate conditions under which a surrogate VAR model with a weakly sparse transition matrix can accurately estimate the change points and their locations for data generated by the original model. This work also addresses and resolves a number of novel technical challenges posed by the nature of the VAR models under consideration. The effectiveness of the proposed algorithms and methodology is illustrated on both synthetic and two real datasets. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2020
Accepted May 2022

KEYWORDS
Algorithms; Consistency; Time series data and their applications

1. Introduction

High dimensional time series analysis and their applications have become increasingly important in diverse domains, including macroeconomics (Stock and Watson 2016; Kilian and Lütkepohl 2017), financial economics (Billio et al. 2012; Lin and Michailidis 2017), molecular biology (Michailidis and d'Alché Buc 2013) and neuroscience (Friston et al. 2014; Schröder and Ombao 2019). Such data are usually both cross-correlated and auto-correlated. There are two broad modeling paradigms for capturing these features in the data: (i) dynamic factor and latent models (Stock and Watson 2002, 2016; Bai and Ng 2008; Lam, Yao, and Bathia 2011; Li, Qin, and Zhou 2014), and (ii) vector autoregressive (VAR) models (Lütkepohl 2013; Kilian and Lütkepohl 2017). The basic premise 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, the latter evolving over time. VAR models aim to capture the self and cross auto-correlation structure in the time series, but the number of parameters to be estimated grows quadratically in the number of time series under consideration. Various structural assumptions have been proposed in the literature to accommodate a large number of time series in the model, with that of *sparsity* (Basu and Michailidis 2015) being a very popular one. However, in many applications the autoregressive dynamics of the time series exhibit also low dimensional structure, which gave rise to the introduction of reduced rank autoregressive models (Box and Tiao 1977; Velu, Reinsel, and Wichern 1986; Ahn and Reinsel 1988; Wang and Bessler 2004). For example, brain activity data (see Example 1 in Section 6) exhibit low dimensional structure (Schröder and Ombao 2019) and so do macroeconomic data (Stock and Watson 2016, Example 2 in Section 6). Reduced rank auto-regressive models for stationary high-dimensional data were studied in Basu, Li, and Michailidis (2019). The key idea of such reduced rank models is that the lead-lagged relationships between the time series cannot simply be described by a few sparse components, as is the case for sparse VAR models. Instead, all the time series influence these relationships and some of them are particularly pronounced (those in the sparse component). Applications in economics/finance, neuroimaging, and environmental science are important candidates for these models.

In many application areas including those mentioned above, nonstationary time series data are commonly observed. The simplest, but realistic departure from stationarity, that also leads to interpretable models for the underlying time series, is *piecewise-stationarity*. Under this assumption, the time series data are modeled as approximately stationary between neighboring change-points, whereas their distribution changes at these change points. The literature on change point analysis for the two classes of modeling paradigms previously mentioned is rather sparse. Bardsley et al. (2017) developed tests for the presence of change points in functional factor models motivated

CONTACT George Michailidis gmichail@ufl.edu Informatics Institute, University of Florida, Gainesville, FL.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

by modeling the yield curve of interest rates, while Barigozzi, Cho, and Fryzlewicz (2018) employed the binary segmentation procedure for detecting and identifying the locations of multiple change points in factor models. Change point detection for *sparse VAR* models has been investigated in Wang et al. (2019), Safikhani and Shojaie (2020), and Bai, Safikhani, and Michailidis (2020).

The objective of this study is to investigate the problem of change point detection in a reduced rank VAR model, whose transition matrices exhibit *low-rank and sparse* structure. The problem poses a number of technical challenges that we address in the sequel.

Formally, a piece-wise stationary VAR model of lag-1 (for introducing the basic issues related to it) for a p-dimensional time series $\{X_t\}$ with m_0 change points $1 \leq \tau_1^\star < \tau_2^\star < \cdots < \tau_{m_0}^\star \leq T$ is given by

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

where $A_j^\star$ is a $p \times p$ coefficient matrix for the jth segment, $j = 1, 2, \ldots, m_0 + 1$, $\mathbf{I}(\tau_{j-1}^\star \leq t < \tau_j^\star)$ presents the indicator function of the jth interval, and ϵ_t^js are $m_0 + 1$ independent zero mean Gaussian noise processes. It is assumed that that the coefficient matrix $A_j^\star$ can be decomposed into a low-rank component plus a sparse component: namely, $A_j^\star = L_j^\star + S_j^\star$, where $L_j^\star$ is a low-rank matrix with rank $r_j^\star$ ($r_j^\star \ll p$), and $S_j^\star$ is a sparse matrix with $d_j^\star$ ($d_j^\star \ll p^2$) nonzero entries.

The modeling framework differs vis-a-vis the one considered in Bai, Safikhani, and Michailidis (2020), since in the current work, *both the low rank and the sparse* components of the transition matrices are allowed to *exhibit changes* at break points. This flexibility rules out the use of a fused lasso based detection algorithm that is suitable for the case wherein *only* the sparse component is allowed to exhibit changes, which was the setting in Bai, Safikhani, and Michailidis (2020). As a result, a novel rolling window detection algorithm is introduced and its theoretical properties studied in the current work.

Next, we outline novel technical challenges, not present in change point analysis of *sparse VAR* (Wang et al. 2019; Safikhani and Shojaie 2020) and other sparse high dimensional models (Roy, Atchadé, and Michailidis 2017):

(i) The change in the transition matrix may be due to a change in the low-rank component, in the sparse component or in both. To that end, we introduce a novel *sufficient identifiability condition* for both detecting a *single* change point and decomposing the transition matrix into its low rank plus sparse components (Assumptions H1 and H2 in the sequel); then, it is *extended* to the case of *multiple* change points (Assumptions H1' and H2').

(ii) For the case of multiple change points, commonly used procedures, such as binary segmentation (Cho and Fryzlewicz 2015) or fused type penalties (Safikhani and Shojaie 2020) are not directly applicable due to the presence of the low rank component. Specifically, the former method would lead to effectively performing singular value decompositions on misspecified models involving mixtures of piece-wise low-rank and sparse models, which may lead to the imposition of very stringent conditions for ensuring detectability of the change points (see discussion on related issues in Bhattacharjee, Banerjee, and Michailidis (2020). Further, it is unclear how to design fused penalties that accommodate low-rank matrices. On the other hand, dynamic programming based algorithms are applicable. However, their time complexity is $\mathcal{O}(T^2 C(T))$, where $C(T)$ indicates the computational cost of estimating the model parameters over the entire observation sequence. This is significantly higher complexity than the previously mentioned methods (which is $\mathcal{O}(TC(T))$, see numerical comparisons and discussion in Remark 6 and Appendix F.7, supplementary materials).

To overcome these challenges, we develop a novel procedure based on *rolling windows*, wherein a *single candidate* change point is identified in each window and then only those exhibiting screened based on certain properties (see Section 3) are retained. This allows to leverage the theoretical results developed for the single change point. The proposed procedure based on rolling windows is *naturally parallelizable*, thus, speeding up computations.

Note that the developed rolling window strategy is applicable to any complex statistical model exhibiting multiple change points. One needs to establish consistency properties for a single change point in a time interval and then appropriately select the length of the rolling window, to ensure that at most a single change point falls within. Hence, this development is of general interest for change point analysis.

(iii) Note that the procedure of estimating change points in low-rank plus sparse VAR models is computationally expensive, even in the presence of a single change point, since it requires performing numerous singular value decompositions. We consider a surrogate model that comes with significant computational savings and under *certain regularity conditions* exhibits *similar accuracy* to the posited model. Specifically, we posit a lag-1 VAR model, wherein the transition matrices $A_j^\star$ are assumed to be *weakly sparse* (see, e.g., Negahban et al. 2012), as an alternative modeling framework. The main reason is that the presence of low rank structure renders the autoregressive parameters in the original model dense. The weak sparse assumption adequately accommodates dense structures under certain conditions and hence can prove useful in certain settings (carefully discussed in the sequel) for change point detection problems. Further, the theoretical properties of exhaustive-search based anomaly detection for weakly sparse VAR models have not been investigated in the literature, and hence this development is of independent interest.

(iv) To establish nonasymptotic error bounds on the model parameters of stationary sparse models, one needs to verify that the commonly imposed (see, e.g., Loh and Wainwright 2012) restricted strong convexity and deviation bound conditions hold (see Propositions 4.2 and 4.3 in Basu and Michailidis 2015).

Verifying these assumptions in the presence of change points in the posited reduced rank VAR model—which technically

is equivalent to working with a misspecified model (see also discussion in Roy, Atchadé, and Michailidis 2017)—represents a nontrivial challenge. This issue is rigorously and successfully addressed in the sequel, together with the introduction of a new version of the deviation bound condition that allows working with misspecified models (technical details presented in Appendix A, supplementary materials).

(v) Finally, obtaining consistent model parameters for each segment identified after detecting the change points requires some care, given the nonstationary nature of the posited model above. This is successfully addressed for the case of a single and multiple change points in Sections 2 and 3, respectively, and for the surrogate model in Section 4.

The remainder of the article is organized as follows. In Section 2, we formulate the model with a single change point, provide a detection procedure based on exhaustive search, and establish theoretical properties for the change point and model parameter estimates. Section 3 discusses the case of multiple change points. It introduces a two-step detection algorithm and establishes consistency of the obtained estimates for the change points and model parameters, leveraging results from Section 2. To reduce computations for detecting the change point(s) in the reduced rank VAR model, we introduce a weakly sparse surrogate model in Section 4 and establish that under certain regularity conditions on the structure of the transitions matrices $A_j^\star$ of the reduced rank model, the estimated change points from the surrogate model are consistent ones for data generated by the former. Section 5 presents a number of numerical experiments to illustrate and assess the performance of the estimates obtained from the single and multiple change points detection procedures. Two real datasets (one on EEG and the other on macroeconomics data) are analyzed using the proposed detection procedures in Section 6. Some concluding remarks are drawn in Section 7. Additional technical conditions, proofs of the main results and additional numerical work are available in the supplementary materials.

Notation: Throughout this article, we denote with a superscript "$\star$" the true value of the model parameters. For any $p \times p$ matrix, we use $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$ to represent the spectral, Frobenius, and nuclear norm, respectively. For any matrix A, A' denotes its transpose, and $A^\dagger$ denotes the conjugate transpose of A, while the ℓ_0, ℓ_1, and ℓ_∞ norms of the vectorized form of A are denoted by: $\|A\|_0 = \text{Card}(\text{vec}(A))$, $\|A\|_1 = \|\text{vec}(A)\|_1$, and $\|A\|_\infty = \|\text{vec}(A)\|_\infty$, respectively. We use $\Lambda_{\max}(\mathbf{X})$ and $\Lambda_{\min}(\mathbf{X})$ to represent the maximum and minimum eigenvalue of the realization matrix $\mathbf{X}$.

2. Single Change Point Model Formulation and Detection Procedure

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

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

where $X_t \in \mathbb{R}^p$ is a vector of observed time series at time t, and $A_1^\star$ and $A_2^\star$ are the $p \times p$ transition matrices for the corresponding models in the two time intervals, and the p dimensional error processes ϵ_t^1 and ϵ_t^2 are independent and identically drawn from Gaussian distributions with mean zero and covariance matrix $\sigma^2 I$ for some fixed σ. It is further assumed that the transition matrices comprise of two time-varying components, a low-rank and a sparse one:

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

The rank of the low-rank components and the density (number of nonzero elements) of the sparse components are denoted by $\text{rank}(L_1^\star) = r_1^\star$, $\text{rank}(L_2^\star) = r_2^\star$, $d_1^\star = \|S_1^\star\|_0$ and $d_2^\star = \|S_2^\star\|_0$, respectively, and satisfy $r_1^\star, r_2^\star \ll p$, $d_1^\star, d_2^\star \ll p^2$.

2.1. Detection Procedure

Let $\{X_0, X_1, \ldots, X_T\}$ be a sequence of observations generated from the VAR model posited in (1) with the structure of the transition matrices given by (2). Then, for any time point $\tau \in \{1, \ldots, T\}$ the corresponding objective functions for estimating the model parameters in the intervals $[1, \tau)$ and $[\tau, T)$ are given by:

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

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

where $\mathbf{X}^{[b:e)}$ denotes the data $\{X_t\}$ from time points b to e, and the nonnegative 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 change point: for any time point $\tau \in \{1, 2, \ldots, T-1\}$,

$$\begin{aligned} \ell(\tau; L_1, L_2, S_1, S_2) \stackrel{\text{def}}{=} \frac{1}{T - 1} \left(\sum_{t=1}^{\tau-1} \|X_t - (L_1 + S_1)X_{t-1}\|_2^2 \right. \\ \left. + \sum_{t=\tau}^{T-1} \|X_t - (L_2 + S_2)X_{t-1}\|_2^2 \right). \end{aligned} \tag{3}$$

The estimator $\widehat{\tau}$ of the change point $\tau^\star$ is given by

$$\widehat{\tau} \stackrel{\text{def}}{=} \arg\min_{\tau \in \mathcal{T}} \ell(\tau; \widehat{L}_{1,\tau}, \widehat{L}_{2,\tau}, \widehat{S}_{1,\tau}, \widehat{S}_{2,\tau}), \tag{4}$$

for the search domain $\mathcal{T} \subset \{1, 2, \ldots, T\}$, where, 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 from the optimization program (4) with tuning parameters $\mu_{1,\tau}$, $\mu_{2,\tau}$, $\lambda_{1,\tau}$, and $\lambda_{2,\tau}$, respectively. Algorithm 1 in Appendix B, supplementary materials describes in detail the key steps in estimating the change point $\tau^\star$ together with the model parameters.

2.2. Theoretical Properties

Next, we address the issue of identifiability of model parameters due to the posited decomposition of the transition matrices into low rank and sparse components. The key idea is to restrict the "spikiness" of the low rank component, so that it can be distinguished from the sparse component. Agarwal, Negahban, and Wainwright (2012) introduced the space Ω defined as

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

wherein the universal parameter α_L defines the *radius of non-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 component to absorb most of the signal, thus, making it harder to identify the sparse component, and vice versa.

Thus, the estimators of the decomposition of the transition matrices A_j are defined as follows, for any fixed time point τ:

$$\begin{aligned} (\widehat{L}_{1,\tau}, \widehat{S}_{1,\tau}) &\stackrel{\text{def}}{=} \arg\min_{\substack{L_1\in\Omega \\ L_1,S_1\in\mathbb{R}^{p\times p}}} \ell(L_1, S_1; \mathbf{X}^{[1:\tau)}), \\ (\widehat{L}_{2,\tau}, \widehat{S}_{2,\tau}) &\stackrel{\text{def}}{=} \arg\min_{\substack{L_2\in\Omega \\ L_2,S_2\in\mathbb{R}^{p\times p}}} \ell(L_2, S_2; \mathbf{X}^{[\tau:T)}). \end{aligned} \tag{5}$$

Next, we introduce an important quantity for future developments, the *information ratio* that measures the relative strength of the maximum signal in the transition matrix $A_j^\star$ generated by the low-rank component vis-a-vis its sparse counterpart, defined as

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

Remark 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

$$\begin{aligned} \|A_2^\star - A_1^\star\|_2 &= \|(L_2^\star - L_1^\star) + (S_2^\star - S_1^\star)\|_2 \\ &\ge \|L_2^\star - L_1^\star\|_2 - \|S_2^\star - S_1^\star\|_2 \\ &\ge \|L_2^\star - L_1^\star\|_2 - p\left(\|S_2^\star\|_\infty + \|S_1^\star\|_\infty\right) \\ &\ge \|L_2^\star - L_1^\star\|_2 - \alpha_L\left(\frac{1}{\gamma_2} + \frac{1}{\gamma_1}\right) \\ &\ge \nu_L - \frac{\alpha_L(\gamma_1+\gamma_2)}{\gamma_1\gamma_2}. \end{aligned}$$

Analogous derivations for the sparse component yield: $\|A_2^\star - A_1^\star\|_2 \ge \|S_2^\star - S_1^\star\|_2 - 2\alpha_L/p \ge \nu_S - 2\alpha_L/p$, where $\nu_L \equiv \|L_2^\star - L_1^\star\|_2 \ge 0$, $\nu_S \equiv \|S_2^\star - S_1^\star\|_2 \ge 0$ are norm differences for the low-rank and the sparse components, respectively.

Based on Remark 1, it can be seen that: (1) when $\gamma_1 \le 1$ or $\gamma_2 \le 1$, we have that $(\gamma_1+\gamma_2)/\gamma_1\gamma_2 \ge 2 > 2/p$ (since $p \gg 2$ in a high dimensional setting). The latter fact implies that in order for changes in the transition matrices $A_j^\star$ to be identifiable—and consequently $\tau^\star$—the difference in the ℓ_2 norm of the low-rank components must significantly exceed that of the sparse components; (2) when both $\gamma_1 > 1$ and $\gamma_2 > 1$, then the quantity $(\gamma_1+\gamma_2)/\gamma_1\gamma_2$ is strictly decreasing with respect to γ_1 and γ_2. Note that in case $1 < \gamma_1 \le p$ and $1 < \gamma_2 \le p$, $(\gamma_1+\gamma_2)/\gamma_1\gamma_2 \ge 2/p$. Combining these two cases leads to the conclusion that when $\gamma_1 \le p$ and $\gamma_2 \le p$, the difference in the ℓ_2 norm ν_L between the low-rank components must be larger than ν_S, the norm difference between the sparse components to guarantee that the change between the transition matrices is detectable.

The following remark discusses an extreme case, wherein the signal in the low-rank components is dominant, but their ℓ_2 norm difference is negligible.

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

$$\begin{aligned} \|A_2^\star - A_1^\star\|_2 &\ge \|S_2^\star - S_1^\star\|_\infty - \epsilon \ge \|S_2^\star\|_\infty - \|S_1^\star\|_\infty - \epsilon \\ &\ge \frac{1}{\gamma_2}\|L_2^\star\|_\infty - \frac{\alpha_L}{p\gamma_1} - \epsilon \\ &= \frac{1}{\gamma_2}\left(\|L_2^\star\|_\infty - \frac{\alpha_L}{p}\frac{\gamma_2}{\gamma_1}\right) - \epsilon. \end{aligned}$$

Note that since the low rank components are constrained to be in the Ω space -$\|L_2^\star\|_\infty \le \alpha_L/p$- 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 $\gamma_2 \ne \gamma_1$ and $\|S_2^\star\|_\infty \ne \|S_1^\star\|_\infty$, is the change in the full transition matrices $A_j^\star$ identifiable, which is intuitive.

The derivations in the two Remarks provide insights into the necessary assumptions needed to establish the theoretical results, presented next.

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

$$\Delta_T(\nu_S^2 + \nu_L^2) \ge C_0\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star(p \vee \log T)\right),$$

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

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

Further, at least one of ν_S, ν_L is strictly positive.

(H2) (Identifiability conditions) Consider low rank matrices $L_1^\star$, $L_2^\star$, and their corresponding Singular Value Decompositions: $L_j^\star = U_j^\star D_j^\star V_j^{\star'}$, where $D_j^\star = \text{diag}(\sigma_1^j, \ldots, \sigma_{r_j}^j, 0, \ldots, 0)$, 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: $\|S_j^\star\|_\infty \le M_S < +\infty, j = 1, 2$;
2. there exists a large enough constant $c > 0$, such that the diagonal matrices $D_j^\star$ satisfy: $\max_{j=1,2}\|D_j^\star\|_\infty \le c < +\infty$; 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 $d^\star_{\max} = \max\{d^\star_1, d^\star_2\}$ satisfies: $d^\star_{\max} \le \frac{1}{C_{\max}}\sqrt{\frac{T}{\log(pT)}}$, 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 $\mathcal{T} \overset{\text{def}}{=} [a, b]$. Assume that, $a = \lfloor (d^\star_{\max} + \sqrt{r^\star_{\max}})^{1+\eta} \rfloor$ and $b = \lfloor T - (d^\star_{\max} + \sqrt{r^\star_{\max}})^{1+\eta} \rfloor$, and denote $|\mathcal{T}|$ as the length of the search domain, then:

$$\frac{|\mathcal{T}|}{d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)} \to +\infty,$$

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

Remark 3. Assumption H1 specifies the relationship between the minimum spacing between the change point and the boundaries of the observation time period and the jump sizes for the low rank and sparse components, analogously to the signal-to-noise assumption in Wang et al. (2019). Assumptions H2-(1) and H2-(2) define the restricted space for the low rank components $L^\star_j$: $\Omega \overset{\text{def}}{=} \left\{L : \|L^\star_j\|_\infty \le \frac{\alpha_L}{p}\right\}$; see analogous definitions and discussion in Agarwal, Negahban, and Wainwright (2012), Basu, Li, and Michailidis (2019), and Bai, Safikhani, and Michailidis (2020) for identifying low rank and sparse matrices. Assumptions H2-(1-3) are sufficient for satisfying the identifiability condition in Hsu, Kakade, and Zhang (2011), the latter implying that the decomposition $A^\star_j = L^\star_j + S^\star_j$ is *unique*. This condition is motivated by the so-called "rank-sparsity" incoherence concept (Chandrasekaran et al. 2011), with further refinements along the lines of results in Hsu, Kakade, and Zhang (2011). This assumption ensures identifiability of model parameters by putting certain conditions on the singular values, and left/right orthonormal singular vectors of the low rank component. Specifically, the new assumption controls the maximum number of nonzeros in any row or column of the sparse component, while ensuring that the low rank part has singular vectors far from the coordinate bases. Note that the new conditions do not put any additional constrains on the dimensionality p and further ensure the *uniqueness* of the low rank plus sparse decomposition of the segment specific transition matrices.

Note that Agarwal, Negahban, and Wainwright (2012) allow α_L to be any constant, whereas we require α_L/p to be vanishing to obtain consistent estimates, due to the presence of misspecification, since the location of the change points is unknown. Assumption H3 reflects the restrictions on the boundary of the search domain $\mathcal{T}$ and connects the estimation rate to the length of the search domain (see analogous condition in Roy, Atchadé, and Michailidis 2017).

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:

$$\begin{aligned}(\lambda_{1,\tau}, \mu_{1,\tau}) &= \left(4c_0\sqrt{\frac{\log p + \log(\tau - 1)}{\tau - 1}},\ 4c'_0\sqrt{\frac{p + \log(\tau - 1)}{\tau - 1}}\right),\\ (\lambda_{2,\tau}, \mu_{2,\tau}) &= \left(4c_0\sqrt{\frac{\log p + \log(T - \tau)}{T - \tau}},\ 4c'_0\sqrt{\frac{p + \log(T - \tau)}{T - \tau}}\right),\end{aligned} \tag{6}$$

for constants $c_0, c'_0 > 0$.

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

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

The proof of Theorem 1 is provided in Appendix E, supplementary materials. Note that the Theorem provides an upper bound for the change point estimation error based on the total sparsity level and the total rank of the model.

Next, we establish estimation consistency for the model parameters. First, given the estimated change point $\widehat{\tau}$, we remove it together with its R-radius neighborhoods $\mathcal{U}(\widehat{\tau}, R)$, to ensure that the remaining time points form two stationary segments. According to Theorem 1, the radius R can be of the order $d^\star_{\max}\log(p \vee T) + r^\star_{\max}(p \vee \log T)$.

Let N_j be the length of the jth segments after removing the R-radius neighborhoods; then, we select another pair of tuning parameters:

$$(\lambda_j, \mu_j) = \left(4c_1\sqrt{\frac{\log p}{N_j}} + \frac{4c_1\alpha_L}{p},\ 4c'_1\sqrt{\frac{p}{N_j}}\right), \quad j = 1, 2, \tag{7}$$

for constants c_1, c'_1 that can selected using cross-validation. The procedure for selecting them, as well as c_0, c'_0 in (6), is provided in Section 5.

Note that the tuning parameters provided in (7) are different from the tuning parameters in (6); the $\log T$ terms are eliminated, since on the selected stationary segments the optimal tuning parameters are always feasible. Based on analogous results in Agarwal, Negahban, and Wainwright (2012) and Basu, Li, and Michailidis (2019) for models whose parameters admit a low rank and sparse decomposition, the optimal tuning parameters in (7) lead to the optimal estimation rate given in the next Theorem.

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

$$\|\widehat{L}_j - L^\star_j\|_F^2 + \|\widehat{S}_j - S^\star_j\|_F^2 \le C_1\left(\frac{d^\star_j\log p + r^\star_j p}{N_j}\right) + C_2\frac{d^\star_j\alpha_L^2}{p^2}, \quad j = 1, 2.$$

The proof of Theorem 2 is provided in Appendix E, supplementary materials.

Remark 4. Notice that Theorem 2 provides the joint estimation rate for the low-rank and the sparse component. It comprises

of two terms, wherein the first one involves the dimensions of the model parameters and converges to zero as the sample size increases, whereas the second term represents the error due to possible unidentifiability of the model parameters. However, in conjunction with Assumption H2 that restricts the space for the low rank component, the second term also converges to zero as the sample size (and hence the dimensionality of the model) increases.

3. The Case of Multiple Change Points

Section 2.2 introduced the technical framework and established the consistency rate for detecting a single change point. Next, these technical developments are leveraged to address the more relevant in practice problem of detecting multiple change points consistently.

We start by formulating the piece-wise VAR model with multiple change points. Consider the p-dimensional VAR(1) process $\{X_t\}$ with m_0 change points $1 = \tau_0^\star < \tau_1^\star < \cdots < \tau_{m_0}^\star < \tau_{m_0+1}^\star = T$; then, the model under consideration is written as

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

where $L_j^\star$ and $S_j^\star$ represent the decomposition of the jth transition matrix into its low-rank and sparse components, and $\mathbf{I}(\tau_{j-1}^\star \le t < \tau_j^\star)$ denotes the indicator function for the jth 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 Assumptions H1–H3 are required:

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

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

where Δ_T is the minimum spacing defined as $\Delta_T \stackrel{\text{def}}{=} \min_{1\le j\le m_0} |\tau_{j+1}^\star - \tau_j^\star|$, 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: $L_j^\star = U_j^\star D_j^\star V_j^{\star'}$, 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: $\|S_j^\star\|_\infty \le M_S < +\infty, j = 1, \ldots, m_0 + 1$;
2. there exists a large enough constant $c > 0$, such that the diagonal matrices $D_j^\star$ satisfy: $\max_{j=1,2} \|D_j^\star\|_\infty \le c < +\infty$, and the orthonormal matrices $U_j^\star$ and $V_j^\star$ such that: $\max_{1\le j\le m_0+1} \left\{\|U_j^\star\|_\infty, \|V_j^\star\|_\infty\right\} = \mathcal{O}\left(\sqrt{\frac{\alpha_L}{r_{\max} p}}\right)$, where $r_{\max} = \max_{1\le j\le m_0+1} r_j^\star$. In addition, we assume that $\alpha_L = \mathcal{O}\left(p\sqrt{\frac{\log(pT)}{T}}\right)$.
3. the maximal sparsity level $d_{\max}^\star = \max_{1\le j\le m_0+1} d_j^\star$ satisfies: $d_{\max}^\star \le \frac{1}{C_{\max}}\sqrt{\frac{T}{\log(pT)}}$, for a large enough positive constant $C_{\max} > 0$.

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

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

for a positive constant $C > 0$.

Assumptions H1' and H2' are direct extensions of Assumptions H1 and H2 to the multiple change points setting. Assumption H3' provides a minimum distance requirement on the consecutive change points and connects the estimation rate and the minimum spacing between change points.

Our detection algorithm will leverage results from the single change point case, and thus, we introduce additional assumptions next. As mentioned in the introduction, the use of fused type penalties is not applicable to the low-rank component and hence an entire different detection procedure is required.

3.1. A Two-step Algorithm for Detecting Multiple Change Points and its Asymptotic Properties

- *Step 1:* It is based on Algorithm 1 provided in Appendix B, supplementary materials 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 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 1 that *not* every window includes a single change point.

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

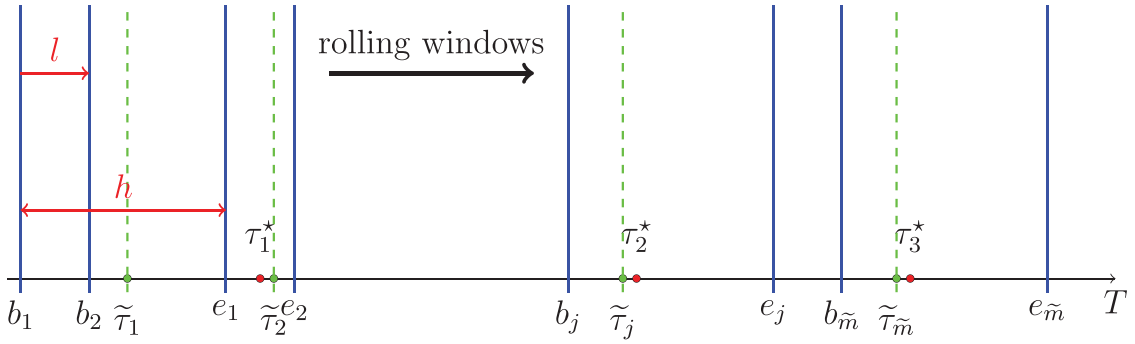

Figure 1. Depiction of the rolling windows strategy. There are three true change points: $\tau_1^\star$, $\tau_2^\star$, and $\tau_3^\star$ (red dots); the boundaries of the rolling-window are represented in blue lines; the estimated change points in each window are plotted in green dashed lines, where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

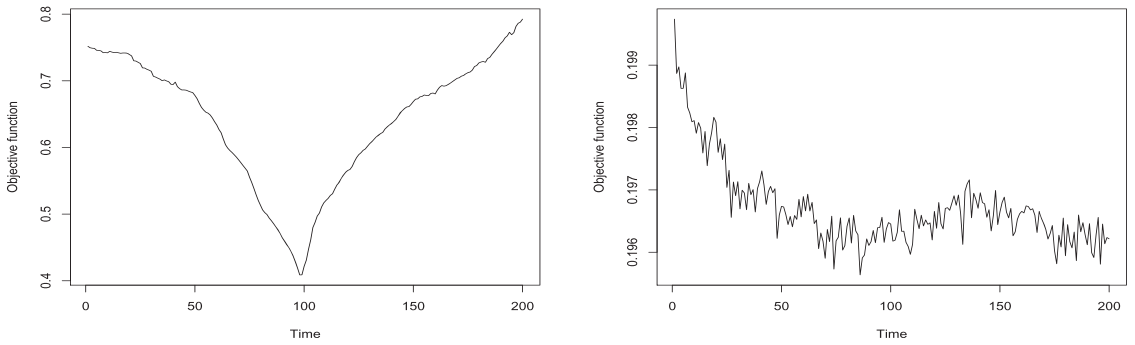

Figure 2. Plots of the objective functions obtained by an application of Algorithm 1, in the presence (left panel) and absence (right panel) of a true change point.

$h = 200$, one containing a change point at $t = 100$ and another not containing a change point. Plots of the objective function (3) used in Algorithm 1 for these two windows are depicted in the left and right panels of Figure 2, respectively. It can be seen that in the presence of a change point, a 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 cannot 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 (see Appendix A, supplementary materials). 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

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

where $[b_i, e_i)$ is the ith rolling-window. Following Chan, Yau, and Zhang (2014), we define the Hausdorff distance between two countable sets on the real line as

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

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

Proposition 1. Suppose Assumptions H1'–H3' and H4 hold, and select the tuning parameters for each rolling window according to (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 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 $\{s_j\}$, $j = 1, 2, \ldots, \widetilde{m}$. Then, model (8) can be rewritten in the following form:

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

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:

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

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 $(s_1, s_2, \ldots, s_m)$:

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

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

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

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

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

Algorithm 2 in Appendix B, supplementary materials describes in detail the key steps for screening the candidate change points by minimizing the information criterion.

The following two additional assumptions on the minimum spacing Δ_T and the selection of tuning parameters are required to establish the main theoretical results.

(H5) Assume that $m_0 T \xi_T (d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}})/\omega_T \to 0$ 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 $\tau_j^\star$ and $\tau_{j+1}^\star$ such that $|s_{i-1} - \tau_j^\star| \le T\xi_T$ and $|s_i - \tau_{j+1}^\star| \le T\xi_T$, select $\lambda_i = 4\left(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_S d_{\max}^\star \frac{T\xi_T}{s_i - s_{i-1}} \right)$ and $\mu_i = 4\left(c\sqrt{\frac{p}{s_i - s_{i-1}}} + \alpha_L \sqrt{r_{\max}^\star} \frac{T\xi_T}{s_i - s_{i-1}} \right)$; (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.

Assumption H5 connects the screening penalty term ω_T, defined with the information criterion (10), and the minimum spacing Δ_T allowed between the change points. Assumption H6 provides the specific rate of the tuning parameters used in the regularized optimization problem formulated in (9). Note that Assumption H6 is required even in standard lasso regression problems for independent and identically distributed data and in the absence of change points (Zhang and Huang 2008). In the literature on change points analysis with misspecified models, a more complex selection of the tuning parameters is needed (Chan, Yau, and Zhang 2014; Roy, Atchadé, and Michailidis 2017). Then, the following Theorem establishes the main result of estimating consistently the number of change points and their locations.

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

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

Remark 5. For a finite number of change points m_0, the sequence $\{\xi_T\}$ can be selected as $\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right)^{1+\frac{\rho}{2}}/T$ 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 $d_{\max}^{\star 2} + r_{\max}^{\star \frac{3}{2}} = o\left(\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right)^{\frac{\rho}{2}} \right)$, then the order of detecting the relative location $-\tau_j^\star/T$- becomes $\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right)^{1+\rho}/T$ in Theorem 3. Finally, one can choose the penalty tuning parameter ω_T to be of order $\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right)^{1+2\rho}$ in this setting, and the minimum spacing Δ_T to be at least of order $\left(d_{\max}^\star \log(p \vee T) + r_{\max}^\star (p \vee \log T)\right)^{2+\rho}$ in accordance to Assumption H3'. Comparing the consistency rates provided in Theorem 3 with those in Safikhani and Shojaie (2020), the additional term $r_{\max}^\star (p \vee \log T)$ reflects the complexity of estimating the low-rank components in the model.

Remark 6 (Computational cost of the rolling windows strategy). 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 $\mathcal{O}(TC(T))$, and the second screening step is of order $\mathcal{O}(T^{1-\delta} C(T))$, where $C(T)$ is the computational cost for model parameters estimation for every search. Hence, the overall complexity is $\mathcal{O}(TC(T))$.

The following corollary provides the error bound for consistent estimation of the low-rank and the sparse components, which is directly extended from Theorem 2 to the multiple change points scenario. To obtain the stationary time series for each segments, we employ the exact same technique of removing R-radius neighborhoods for every estimated change point. In accordance to Theorem 3, the radius R should be at least of order $Bm_0T\xi_T(d^{\star 2}_{\max} + r^{\star\frac{3}{2}}_{\max})$ for some large constant $B > 0$. Denote the length of the jth stationary segment by N_j, after removing the R-radius neighborhoods for each estimated change point.

Corollary 1. 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

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

- *Step 3 (Optional):* After the second Step, the results in Theorem 3 and Corollary 1 ensure accurate estimation of the number of change points and their locations, as well as of the underlying model parameters across the stationary segments. However, a further refinement and hence a tighter bound on the result provided in Theorem 3 can be obtained through the following re-estimation procedure (see also discussion on this point in Wang et al. 2019). Specifically, the conclusions in Theorem 3 ensure that $\widehat{m} = m_0$ almost surely and also provide good estimates of the boundaries of the stationary segments. Then, for an estimated change point $\widehat{\tau}_j$, consider a "refined" interval $(s_j, e_j) \stackrel{\text{def}}{=} (2\widehat{\tau}_{j-1}/3 + \widehat{\tau}_j/3, 2\widehat{\tau}_j/3 + \widehat{\tau}_{j+1}/3)$ for $j = 1, 2, \ldots, \widehat{m}$, where $\tau_0 = 0$. Then, we define the objective function:

$$\ell(\tau; s_j, e_j, A_{j,1}, A_{j,2}) \stackrel{\text{def}}{=} \frac{1}{e_j - s_j}\left(\sum_{\tau=s_j}^{\tau-1}\|X_t - A_{j,1}X_{t-1}\|_2^2 + \sum_{t=\tau}^{e_j}\|X_t - A_{j,2}X_{t-1}\|_2^2\right),$$

  and a "refined" change point together with the refitted model parameters corresponds to:

$$(\widetilde{\tau}_j, \widetilde{A}_{j,1}, \widetilde{A}_{j,2}) = \arg\min_{\tau\in(s_j,e_j)}\ell(\tau; s_j, e_j, A_{j,1}, A_{j,2}) \quad (11)$$

According to the proposed refinement, we derive the following corollary:

Corollary 2. Suppose Assumptions H1'–H3', and H4–H6 hold. As $T \to +\infty$, the minimizer $(\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widehat{m}})$ of (11) satisfies:

$$\mathbb{P}\left(\max_{1\le j\le m_0}|\widetilde{\tau}_j - \tau_j^\star| \le K\frac{d^\star_{\max}\log(p\vee h) + r^\star_{\max}(p\vee\log h)}{\min_{1\le j\le m_0}\{\nu_{j,S}^2 + \nu_{j,L}^2\}}\right) \to 1.$$

Remark 7. Note that in the bound of Corollary 2, the maximum density across all sparse components $d^\star_{\max}$ appears as a linear term, instead of a quadratic one in Theorem 3. This refinement is primarily of theoretical interest, since as the numerical work in Section 5.2 indicates the detection procedure based on Steps 1 and 2 achieves very accurate estimates of the change points and the model parameters.

Remark 8. Corollary 2 indicates that the high probability finite sample bound on the estimation error depends on the maximum sparsity level $d^\star_{\max}$ among the sparse components, the maximum rank $r^\star_{\max}$ among the low rank components, the dimension p, and the signal strength ν_S, ν_L of the sparse and low rank components. Note that the issue of obtaining asymptotic distributions for the estimated change points is a rather complicated task and has not been addressed in the literature even for much simpler models, including sparse mean shift models.

4. A Fast Procedure Based on a Surrogate Model

Remark 6 shows 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 presents plots of the objective functions of the original and the surrogate weakly sparse 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 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.

4.1. Formulation of the Surrogate Weakly Sparse VAR Model

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

$$\mathbb{B}_q(R_q) := \left\{A \in \mathbb{R}^{p\times p} : \sum_{i=1}^p\sum_{j=1}^p |a_{ij}|^q \le R_q\right\}, \quad (12)$$

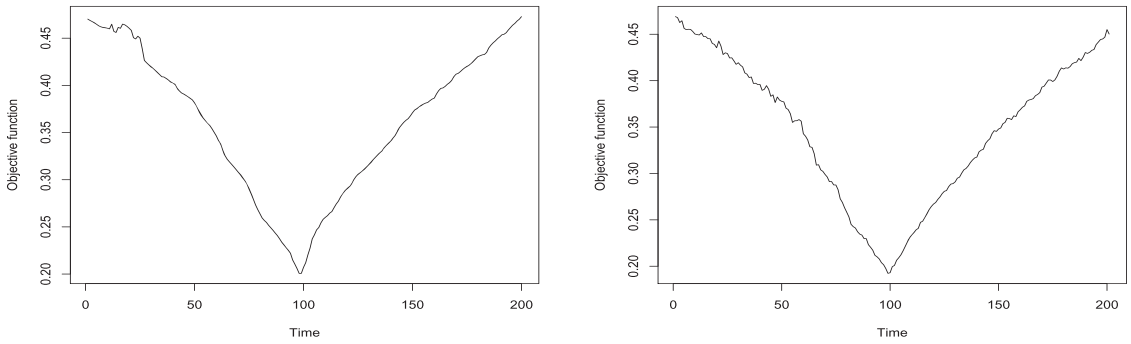

Figure 3. Left: the curve of the objective function of the full low-rank plus sparse model; Right: the curve of the objective function of the alternative weakly sparse model.

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 2 requires some modification, whose details are given in Appendix C, supplementary materials.

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 2 and 2.2. Based on Remark 1 and using the same notation as in the results in Sections 2.2 and 3, 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 9. Assumption W1 is based on Remark 1. 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.

4.2. 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 2. Let $q \in (0,1)$ be fixed and $R_q > 0$ be the radius of $\mathbb{B}_q(R_q)$ defined in (12). 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

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

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

Before we extend Theorem 1 to the surrogate weakly sparse model, a modification to the selection of tuning parameters is required. Recall that (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 (6), the tuning parameters are given by

$$\begin{aligned} \lambda_{1,\tau}^w &= 4c_0^w \sqrt{\frac{\log p + \log(\tau - 1)}{\tau - 1}}, \\ \lambda_{2,\tau}^w &= 4c_0^{w'} \sqrt{\frac{\log p + \log(T - \tau)}{T - \tau}}, \end{aligned} \tag{13}$$

where $c_0^w, c_0^{w'} > 0$ are some positive constants selected by the similar method as c_0 and c_0' in (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 1 in the following proposition, whose proof is provided in Appendix E, supplementary materials.

Proposition 3. Suppose Assumptions W1 and W2 hold and the transition matrices $A_1^\star$ and $A_2^\star$ in (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 (13), 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, for the weakly sparse model.

Proposition 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 3.1. Similarly, we also extend Assumptions H3', H5, and H6 to the weakly sparse scenario—Assumptions W3 and W4 given in Appendix C, supplementary materials—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 5. Suppose Assumptions W1–W4 hold and denote the minimizer of (7) in Appendix C, supplementary materials by $(\widehat{\tau}_1^w, \ldots, \widehat{\tau}_{\widehat{m}^w}^w)$. 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 10. Proposition 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 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 $T^{1+q}\xi_T^w R_q^2 \left(\log(p \vee T)\right)^{2\rho-q}$.

An analogue of Corollary 1 and a comparison of the error bounds established in Theorem 3 and Proposition 5 are given in Appendix C, supplementary materials.

5. Performance Evaluation

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 (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 lth 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 (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 jth 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), we can use the theoretical values of λ_j and μ_j provided in (6) and (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:

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

- *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.

All numerical experiments are run in R 3.6.0 on the uf HiPerGator Computing platform with 4 Intel E5 2.30 GHz Cores and 16 GB memory. The code and scripts for simulation examples and applications are available at *https://github.com/peiliangbai92/LSVAR_cpd*.

5.1. Performance for Detecting A 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, ν_L and ν_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 full summary is provided in the form of a Table in Appendix F.1, supplementary materials.

(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; that is, $\gamma_j \geq 1$ 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 results for these settings over 50 replications are given in Table 1. 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 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 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, supplementary materials.

Figure 4 depicts boxplots based on 50 replicates of the distance between the location of the true change point and its estimate, that is, $|\widehat{\tau} - \tau^\star|$. 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

Table 1. 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)

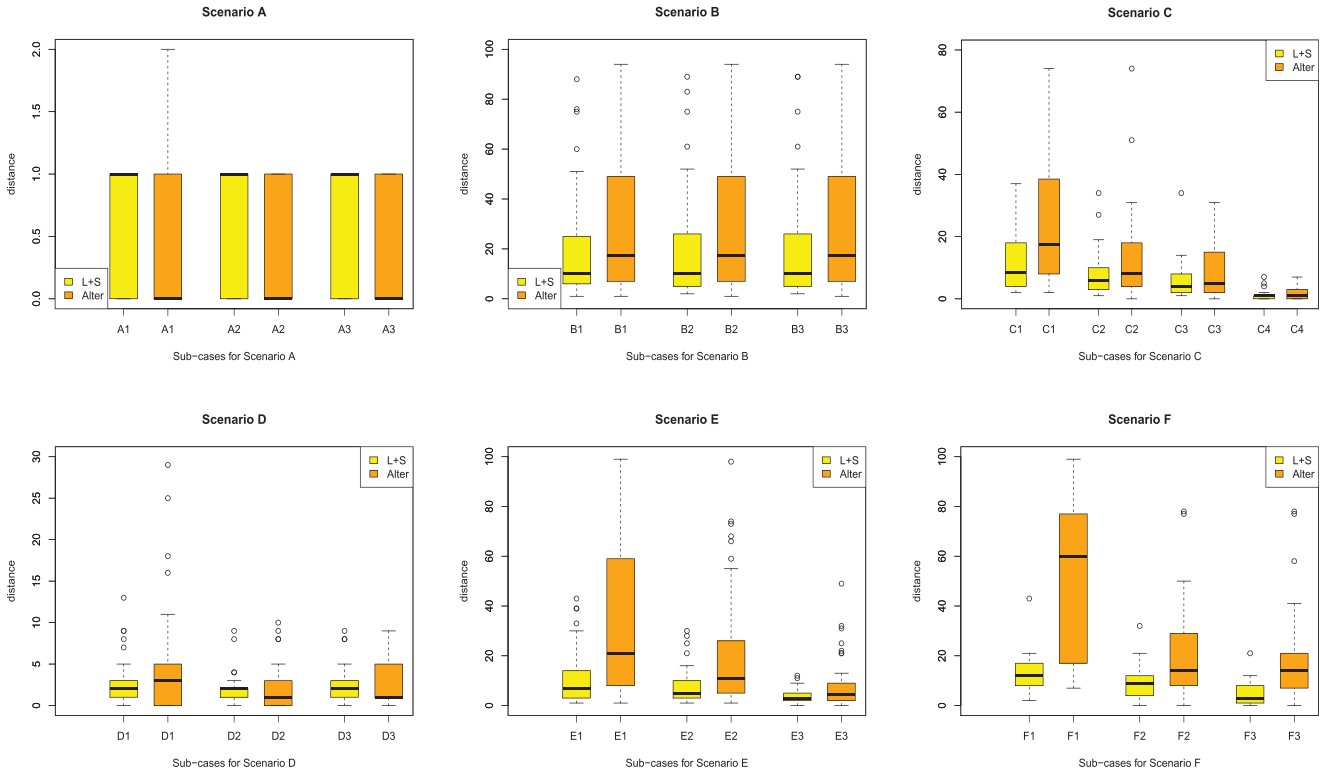

Figure 4. Boxplots for $|\widehat{\tau} - \tau^\star|$ under settings A–F with the full model and the surrogate weakly sparse model.

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.

5.2. Performance for Detecting Multiple Change Points

We consider the same settings for each change point, as in case A in Section 5.1 with modified T and p, respectively. The specific scenarios under consideration are as follows:

(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.

Table 2. Results for multiple change point selection by full L+S model.

	Points	Truth	Mean	SD	Selection rate		Points	Truth	Mean	SD	Selection rate
L.1	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
	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
L.2	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
	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
L.3	1	0.1000	0.1000	0.0000	1.00						
	2	0.3000	0.3000	0.0000	1.00						
	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						

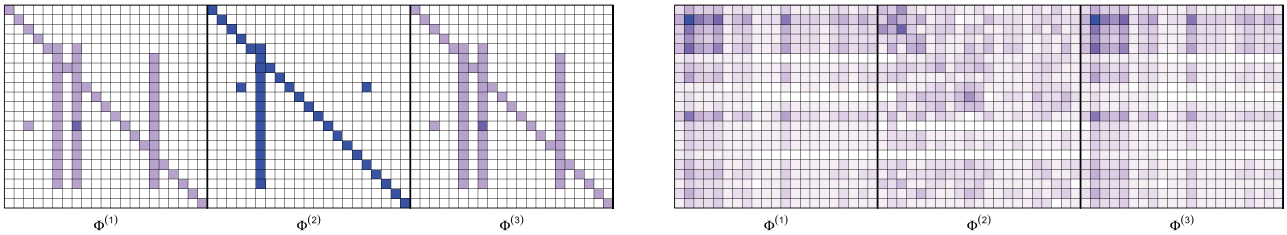

Figure 5. Left: Estimated sparse brain connectivity structure; Right: Estimated low rank brain connectivity structure.

The detailed model parameters are listed in the Table 5 in the Appendix F.2, supplementary materials.

Table 2 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 boxplots of $|\widehat{\tau}_j - \tau_j^\star|$ for scenario N over 50 replicates are depicted in the Appendix F.2, supplementary materials.

5.3. A Simulation Scenario Based on a EEG Dataset

For this scenario, the sparsity structure is extracted from the EEG dataset analyzed in Section 6.1. Specifically, the setting under consideration is as follows: $T = 300$, $p = 21$, with two change points located at $\lfloor T/3 \rfloor$ and $\lfloor 2T/3 \rfloor$, respectively. The structure of the transition matrices is obtained by using the results presented in the application section (see Figure 6 in the Section G.2. in the supplementary materials). We keep the nonzero elements (see Figure 5) and set their magnitudes at random to 0.4, −0.6, and 0.4, respectively. The low rank components are generated by using the spectral decomposition with ranks equal to 1, 3, and 1. The estimated *sparse* and *low rank* structures are illustrated in Figure 5:

The results are summarized in Table 3. It can be seen that based on a low rank and sparse structure motivated by real data, the proposed algorithm exhibits a very satisfactory performance.

Table 3. Results of simulation scenario based on an EEG dataset.

	Points	Truth	Mean	SD	Selection rate
General sparsity pattern	1	0.3333	0.3328	0.002	1.00
	2	0.6667	0.6663	0.007	1.00

5.4. Impact of the Signal-to-noise Ratio on the Detection Rate

The signal-to-noise ratio (SNR) is defined as (see also Wang, Yu, and Rinaldo 2020; Rinaldo et al. 2021):

$$\text{SNR} = \frac{\Delta_T \nu}{T},$$

wherein $\nu \stackrel{\text{def}}{=} \min_j \nu_j$ the minimum jump size, and Δ_T is the minimum spacing, that is, $\Delta_T = \min_{1 \le j \le m_0} |\tau_j^\star - \tau_{j+1}^\star|$. We set $T = 300$ and $p = 20$ with two change points located at $\lfloor T/3 \rfloor = 100$ and $\lfloor 2T/3 \rfloor = 200$, respectively. Further, we set the minimum jump size to $\nu = 0.8$, 1.0, and 1.6, and the resulting SNR takes the values 0.27, 0.33, and 0.53. The results are given Table 4.

Table 4. Extra simulation performance for different signal-to-noise ratios.

SNR	points	truth	mean	sd	selection rate
0.27	1	0.3333	0.3412	0.017	0.90
	2	0.6667	0.6702	0.012	0.94
0.33	1	0.3333	0.3330	0.002	1.00
	2	0.6667	0.6687	0.004	1.00
0.57	1	0.3333	0.3332	0.002	1.00
	2	0.6667	0.6665	0.001	1.00

Figure 6. Left: heat map of Hamming distances between the estimated low-rank components; Right: heat map of Hamming distances between the estimated sparse components.

As expected, for small SNR the detection accuracy deteriorates, both in terms of the selection rate of change points, as well as their locations. However, for SNR around or greater than 1, it becomes very satisfactory. Additional results are provided in Section F.3 in the supplementary materials.

Remark 11 (*Additional numerical results and comparisons*). Additional numerical results including (i) for the surrogate model, (ii) for additional scenarios for multiple change points, (iii) for run times between the low rank plus sparse and the surrogate models, (iv) with a factor model exhibiting change points, (v) between a factor and the low rank plus sparse models under a misspecified data generating mechanism, (vi) comparison between the proposed two-step algorithm and the TSP algorithm in Bai, Safikhani, and Michailidis (2020), and (vi) between the two-step rolling window strategy and a dynamic programming algorithm are presented in Appendices F.1–F.7, respectively.

6. Applications

6.1. Change Point Detection in EEG Signals

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, Chavarriaga, and Millán 2018). Next, we employ the full low-rank plus sparse model to detect change points in data from Trujillo, Stanfield, and Vela (2017). This dataset 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 122,880 over 480 sec. The stimulus procedure is that after a resting state (eliminated from the dataset) 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 detrended, 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, Stanfield, and Vela (2017), there are five open/closed eyes segments in the selected time period with four change points approximately at locations: $\tau_1^\star \cong 11{,}650$, $\tau_2^\star \cong 27{,}750$, $\tau_3^\star \cong 44{,}000$, and $\tau_4^\star \cong 60{,}000$. The data are plotted in Figure 2 in Appendix G.2, supplementary materials. Selection of the tuning parameters is based on the guidelines given in Appendix G.1, supplementary materials. 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 $\widehat{\tau}_1 = 9633$, $\widehat{\tau}_2 = 28{,}529$, $\widehat{\tau}_3 = 43{,}361$ and $\widehat{\tau}_4 = 60{,}209$. 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 6 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. We also provide the estimated low-rank and the sparse patterns for five segments in Figure 3, and the correlation networks for the sparse components in Figure 4 in Appendix G.2, supplementary materials.

6.2. An Application to Macroeconomics Data

We consider the macroeconomics data obtained from the FRED database McCracken and Ng (2016). This dataset comprises of 19 key macroeconomic variables, corresponding to the

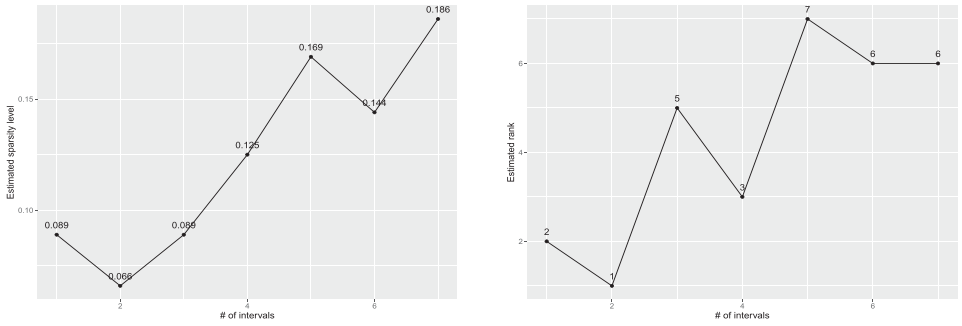

Figure 7. Left panel: Estimated sparsity level for each selected interval; Right panel: Estimated rank for each selected interval.

Table 5. 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, supplementary materials
09/01/2008	Recession following collapse of Lehman Brothers
05/01/2010	Recovery from the Great Financial crisis of 2008

Table 6. Estimated change points by the detection strategy based on a factor model.

Date (mm/dd/yyyy)	Candidate related events
12/01/1979	Rapid increase of interest rates by the Volcker Fed
01/01/1985	Multiple events
11/01/1993	Multiple events
04/01/2008	Prequel to the great financial crisis

"Medium" model analyzed in Bańbura, Giannone, and Reichlin (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 two-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 estimated change points are listed in Table 5, while the sparsity levels and ranks for each segment are plotted in Figure 7. The selected change points are presented in Figure 5 in Appendix G.3, supplementary materials. A detailed discussion (due to space constraints) of related events is also provided in Appendix G.3, supplementary materials.

We also compare the results using the detection strategy based on the static factor model in Barigozzi, Cho, and Fryzlewicz (2018). According to Fama and French (1996), we set the maximum number of factors to three and the estimated change points are listed in Table 6.

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 third quarter, following the collapse of Lehman Brothers in mid-September.

7. Concluding Remarks

The article 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 two-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, that is, 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, Li, and Michailidis (2019).

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

Supplementary Materials

The supplementary materials contain all proofs of the theoretical results, together with auxiliary lemmas, additional details on the detection algorithms, and additional numerical experiments.

Acknowledgments

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

Funding

The work of George Michailidis has been supported in part by NSF grants DMS2124507, DMS1821220, and DMS1830175.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012), "Noisy Matrix Decomposition via Convex Relaxation: Optimal Rates in High Dimensions," *The Annals of Statistics*, 40, 1171–1197. [4,5]

Ahn, S. K., and Reinsel, G. C. (1988), "Nested Reduced-Rank Autoregressive Models for Multiple Time Series," *Journal of the American Statistical Association*, 83, 849–856. [1]

Bai, J., and Ng, S. (2008), "Large Dimensional Factor Analysis," *Foundations and Trends® in Econometrics*, 3, 89–163. [1]

Bai, P., Safikhani, A., and 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. [2,5,15]

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian Vector Auto Regressions," *Journal of Applied Econometrics*, 25, 71–92. [16]

Bardsley, P., Horváth, L., Kokoszka, P., and Young, G. (2017), "Change Point Tests in Functional Factor Models with Application to Yield Curves," *The Econometrics Journal*, 20, 86–117. [1]

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018), "Simultaneous Multiple Change-Point and Factor Analysis for High-Dimensional Time Series," *Journal of Econometrics*, 206, 187–225. [2,16]

Basu, S., Li, X., and Michailidis, G. (2019), "Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions," *IEEE Transactions on Signal Processing*, 67, 1207–1222. [1,5,16]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-Dimensional Time Series Models," *The Annals of Statistics*, 43, 1535–1567. [1,2]

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2020), "Change Point Estimation in a Dynamic Stochastic Block Model," *Journal of Machine Learning Research*, 21, 1–59. [2]

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012), "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," *Journal of Financial Economics*, 104, 535–559. [1]

Box, G. E., and Tiao, G. C. (1977), "A Canonical Analysis of Multiple Time Series," *Biometrika*, 64, 355–365. [1]

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014), "Group Lasso for Structural Break Time Series," *Journal of the American Statistical Association*, 109, 590–599. [7,8]

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011), "Rank-Sparsity Incoherence for Matrix Decomposition," *SIAM Journal on Optimization*, 21, 572–596. [5]

Cho, H., and Fryzlewicz, P. (2015), "Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation," *Journal of the Royal Statistical Society*, Series B, 77, 475–507. [2]

Fama, E. F., and French, K. R. (1996), "Multifactor Explanations of Asset Pricing Anomalies," *The Journal of Finance*, 51, 55–84. [16]

Friston, K. J., Bastos, A. M., Oswal, A., van Wijk, B., Richter, C. and Litvak, V. (2014), "Granger Causality Revisited," *Neuroimage*, 101, 796–808. [1]

Hsu, D., Kakade, S. M., and Zhang, T. (2011), "Robust Matrix Decomposition with Sparse Corruptions," *IEEE Transactions on Information Theory*, 57, 7221–7234. [5]

Jao, P.-K., Chavarriaga, R., and 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. [15]

Kilian, L., and Lütkepohl, H. (2017), *Structural Vector Autoregressive Analysis*, Cambridge: Cambridge University Press. [1]

Lam, C., Yao, Q., and Bathia, N. (2011), "Estimation of Latent Factors for High-Dimensional Time Series," *Biometrika*, 98, 901–918. [1]

Li, G., Qin, S. J., and Zhou, D. (2014), "A New Method of Dynamic Latent-Variable Modeling for Process Monitoring," *IEEE Transactions on Industrial Electronics*, 61, 6438–6445. [1]

Lin, J., and Michailidis, G. (2017), "Regularized Estimation and Testing for High-Dimensional Multi-Block Vector-Autoregressive Models," *The Journal of Machine Learning Research*, 18, 4188–4236. [1]

Liu, F., Wang, S., Qin, J., Lou, Y., and Rosenberger, J. (2018), "Estimating Latent Brain Sources with Low-Rank Representation and Graph Regularization," in International Conference on Brain Informatics, Springer, pp. 304–316. [15]

Loh, P.-L., and Wainwright, M. J. (2012), "High-Dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity," *The Annals of Statistics*, 40, 1637–1664. [2]

Lütkepohl, H. (2013), *Introduction to Multiple Time Series Analysis*, Berlin: Springer. [1]

McCracken, M. W., and Ng, S. (2016), "Fred-md: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589. https://doi.org/10.1080/07350015.2015.1086655 [15]

Michailidis, G., and d'Alché Buc, F. (2013), "Autoregressive Models for Gene Regulatory Network Inference: Sparsity, Stability and Causality Issues," *Mathematical Biosciences*, 246, 326–334. [1]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of m-estimators with Decomposable Regularizers," *Statistical Science*, 27, 538–557. [2,10]

Rinaldo, A., Wang, D., Wen, Q., Willett, R., and Yu, Y. (2021), "Localizing Changes in High-Dimensional Regression Models, in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 2089–2097. [14]

Roy, S., Atchadé, Y., and Michailidis, G. (2017), "Change Point Estimation in High Dimensional Markov Random-Field Models," *Journal of the Royal Statistical Society*, Series B, 79, 1187–1206. [2,3,5,8]

Safikhani, A., and 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 . [2,8]

Schröder, A. L., and Ombao, H. (2019), "Fresped: Frequency-Specific Change-Point Detection in Epileptic Seizure Multi-Channel EEG Data," *Journal of the American Statistical Association*, 114, 115–128. [1]

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [1]

——— (2016), "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics," in *Handbook of Macroeconomics* (Vol. 2), pp. 415–525, Amsterdam: Elsevier. [1]

Trujillo, L. T., Stanfield, C. T., and 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, 1–22. [15]

Velu, R. P., Reinsel, G. C., and Wichern, D. W. (1986), "Reduced Rank Models for Multiple Time Series," *Biometrika*, 73, 105–118. [1]

Wang, D., Yu, Y., and Rinaldo, A. (2020), "Univariate Mean Change Point Detection: Penalization, Cusum and Optimality," *Electronic Journal of Statistics*, 14, 1917–1961. [14]

Wang, D., Yu, Y., Rinaldo, A., and Willett, R. (2019), "Localizing Changes in High-dimensional Vector Autoregressive Processes," arXiv preprint arXiv:1909.06359. [2,5,9,16]

Wang, Z., and Bessler, D. A. (2004), "Forecasting Performance of Multivariate Time Series Models with Full and Reduced Rank: An Empirical Examination," *International Journal of Forecasting*, 20, 683–695. [1]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [8]