Localizing Changes in High-Dimensional Vector Autoregressive Processes

Daren Wang¹, Yi Yu², Alessandro Rinaldo³, and Rebecca Willett¹

¹Department of Statistics, University of Chicago
²Department of Statistics, University of Warwick
³Department of Statistics & Data Science, Carnegie Mellon University

Abstract

Autoregressive models capture stochastic processes in which past realizations determine the generative distribution of new data; they arise naturally in a variety of industrial, biomedical, and financial settings. Often, a key challenge when working with such data is to determine when the underlying generative model has changed, as this can offer insights into distinct operating regimes of the underlying system. This paper describes a novel dynamic programming approach to localizing changes in high-dimensional autoregressive processes and associated error rates that improve upon the prior state of the art. When the model parameters are piecewise constant over time and the corresponding process is piecewise stable, the proposed dynamic programming algorithm consistently localizes change points even as the dimensionality, the sparsity of the coefficient matrices, the temporal spacing between two consecutive change points, and the magnitude of the difference of two consecutive coefficient matrices are allowed to vary with the sample size. Furthermore, initial, coarse change point localization estimates can be improved via a computationally efficient refinement algorithm that offers further improvements on the localization error rate. At the heart of the theoretical analysis lies a general framework for high-dimensional change point localization in regression settings that unveils key ingredients of localization consistency in a broad range of settings. The autoregressive model is a special case of this framework. A byproduct of this analysis are new, sharper rates for high-dimensional change point localization in linear regression settings that may be of independent interest.

Keywords: Change point detection; High-dimensional statistics; Vector autoregressive models; Dynamic programming.

1 Introduction

High-dimensional data are routinely collected in both traditional and emerging application areas. Time series data are by no means immune to this high dimensionality trend, and commonly arise in applications from econometrics (e.g. Bai and Perron, 1998; De Mol et al., 2008), finance (e.g. Chen and Gupta, 1997), genetics (e.g. Michailidis and d’Alché Buc, 2013), neuroimaging (e.g. Smith, 2012; Bolstad et al., 2011), predictive maintenance (e.g. Susto et al., 2014; Swanson, 2001; Yam et al., 2001), to name but a few.

Arguably, the most popular tool in modeling high-dimensional time series is the vector autoregressive (VAR) model (see e.g. Lütkepohl, 2005). The recent literature on the high-dimensional VAR models is vast. Hsu et al. (2008), Haufe et al. (2010), Shojaie and Michailidis (2010), Basu and Michailidis
Let \((X_1, \ldots, X_n)\) be a time series with random vectors in \(\mathbb{R}^p\) and let \(0 = \eta_0 < \eta_1 < \ldots < \eta_k < \eta_{k+1} = n\) be an increasing sequence of change points. For any \(k \in \{0, \ldots, K\}\), set \(X_k = \{X_{\eta_k+1}, \ldots, X_{\eta_{k+1}}\}\). We assume the following.

- For each \(k, l \in \{0, \ldots, K\}\), \(k \neq l\), it holds that \(X_k\) and \(X_l\) are independent.
- For each \(k \in \{0, \ldots, K\}\), \(X_k\) is a subset of an infinite stationary and stable time series \(X_k^{\infty}\) and
  \[X_{t+1} = A_t^* X_t + \varepsilon_{t+1}, \quad t = \eta_k + 1, \ldots, \eta_{k+1}-1,\]
  where
  \[A_{\eta_k} = \ldots = A_{\eta_{k+1}-1}^* \in \mathbb{R}^{p \times p}, \quad \|A_t^*\|_{\text{op}} < 1, \quad A_{\eta_k}^* \neq A_{\eta_{k+1}}^*,\]
  the notation \(\|\cdot\|_{\text{op}}\) denotes the operator norm of a matrix, and \(\varepsilon_t \sim \mathcal{N}(0, I_p)\) independent with \(X_t\).

For \(k \in \{0, \ldots, K\}\), \(X_{\eta_k+1}\) is defined as
  \[X_{\eta_k+1} = A_{\eta_k}^* \bar{X}_{\eta_k} + \varepsilon_{\eta_k+1},\]
where \(\bar{X}_{\eta_k}\) is an unobserved latent random vector drawn from \(X_k^{\infty}\).

**Remark 1.** The condition \(\|A_t^*\|_{\text{op}} < 1\) is assumed to guarantee that between two consecutive change points, the time series is stable (see e.g. Chapter 2.1.1 in Lütkepohl, 2005). In addition, \(X_{\eta_k}\)'s are not predictors of \(X_{\eta_k+1}\)'s, and \(\bar{X}_{\eta_k}\)'s are the de facto predictors of \(X_{\eta_k+1}\)'s. For each \(k \in \{1, \ldots, K\}\), \(\bar{X}_{\eta_k} \in X_k^{\infty}\); for each \(k \in \{1, \ldots, K+1\}\), \(X_{\eta_k} \in X_{k-1} \subset X_{k-1}^{\infty}\).

Given data sampled from Model 1, our main task is to develop computationally-efficient algorithms that can consistently estimate both the unknown number \(K\) of change points and the time points \(\{\eta_k\}_{k=1}^{K}\), at which the coefficient matrices change. That is, we seek consistent estimators.
\[ \hat{\eta}_k \] \text{ for } k = 1, \ldots, K, \] such that, as the sample size \( n \) grows unbounded, it holds with probability tending to 1 that
\[
\hat{K} = K \quad \text{and} \quad \frac{\epsilon}{n} = \max_{k=1, \ldots, K} \frac{|\hat{\eta}_k - \eta_k|}{n} = o(1).
\]
In the rest of the paper, we refer to \( \epsilon \) as the localization error rate.

Despite the vast body of literature on different change point detection problems, the study on Model 1 is scarce. Safikhani and Shojaie (2017) is within the very little existing literature dedicated to a similar problem. Detailed comparisons with Safikhani and Shojaie (2017) will be presented after we state our algorithm and main results, at the end of Section 2.3.

**Methods**

To achieve the goal of obtaining consistent change point estimators, we adopt a dynamic programming approach. To be specific, let \( \mathcal{P} \) be an interval partition of \( \{1, \ldots, n\} \) into \( K \) intervals, i.e.
\[
\mathcal{P} = \{\{1, \ldots, i_1\}, \{i_1 + 1, \ldots, i_2\}, \ldots, \{i_{K-1} + 1, \ldots, i_K\}\},
\]
for some integers \( 0 < i_1 < \cdots < i_K = n \), where \( K \geq 1 \). For a positive tuning parameter \( \gamma > 0 \), let
\[
\hat{\mathcal{P}} \in \arg \min_{\mathcal{P}} \left\{ \sum_{I \in \mathcal{P}} \mathcal{L}(I) + \gamma|\mathcal{P}| \right\},
\]
where \( \mathcal{L}(\cdot) \) is a loss function, \( |\mathcal{P}| \) is the cardinality of \( \mathcal{P} \) and the minimization is taken over all possible interval partitions of \( \{1, \ldots, n\} \).

The change point estimator resulting from the solution to (1) is simply obtained by taking all the right endpoints of the intervals \( I \in \hat{\mathcal{P}} \), except \( n \). The optimization problem (1) is known as the minimal partition problem and can be solved using dynamic programming with computational cost of order \( O(n^2 T(n)) \), where \( T(n) \) denotes the computational cost of solving \( \mathcal{L}(I) \) with \( |I| = n \) (see e.g. Algorithm 1 in Friedrich et al., 2008).

We will tackle Model 1 in the framework of (1), by setting
\[
\mathcal{L}((s, e]) = \sum_{t=s+1}^{e-1} \|X_{t+1} - \hat{A}^\lambda_{(s,e]} X_t\|_2^2,
\]
where \( \hat{A}^\lambda_{(s,e]} \) is the Lasso estimator defined as
\[
\hat{A}^\lambda_{(s,e]} = \arg \min_{A \in \mathbb{R}^p \times p, \|A\|_{op} \leq \lambda} \sum_{t=s+1}^{e-1} \|X_{t+1} - AX_t\|_2^2 + \lambda \sqrt{\max\{e - s - 1, \log(n \lor p)\}} \|A\|_1.
\]
\( \lambda \geq 0 \) is a penalty term, and the norms are defined in Section 1.1. The penalty will be chosen as a function of
\[
\max\{e - s - 1, \log(n \lor p)\}.
\]
This is due to a large deviation inequality in Lemma 22(a). Intuitively speaking, it is due to the fact that sums of i.i.d. sub-Exponential random variables behave like sub-Exponential random variables when the sample size is small, and behave like sub-Gaussian random variables when the sample size is large.
Algorithms based on dynamic programming are widely used in the change point detection literature. Friedrich et al. (2008), Killick et al. (2012), Rigaill (2010), Maidstone et al. (2017), Wang et al. (2018b), among others, studied dynamic programming approaches for change point analysis involving a univariate time series with piecewise-constant means. Leonardi and Bühlmann (2016) examined high-dimensional linear regression change point detection problems by using a version of dynamic programming approach. We will provide more comparisons with Leonardi and Bühlmann (2016) in Section 3.

List of contributions

- In this paper, we provide consistent change point estimators for Model 1. We allow for model parameters to change with the sample size $n$, including the dimensionality of the data, the entry-wise sparsity of the coefficient matrices, the number of change points, the smallest distance between two consecutive change points, and the smallest difference between two consecutive different coefficient matrices. To the best of our knowledge, the theoretical results we provide in this paper are the sharpest in the existing literature. Furthermore, the proposed algorithms, based on the general framework described in (1), can be implemented using dynamic programming approaches and are computationally efficient.

- We further devise an additional second step (Algorithm 1), called local refinement, to deliver an even better localization error rate, even though directly optimizing (1) already provides the sharpest rates among the ones existing in the literature.

- We not only thoroughly analyze the VAR(1) Model 1, but also extend those results to a general VAR(L) process. To be specific, we also consider a more general autoregressive model, described next as Model 2, and develop analogous guarantees for it.

**Model 2 (VAR(L) model).** Let $(X_1, \ldots, X_n)$ be a time series of random vectors in $\mathbb{R}^p$ and let $0 = \eta_0 < \eta_1 < \ldots < \eta_K < \eta_{K+1} = n$ be an increasing sequence of change points. For each $k \in \{0, \ldots, K\}$, set $X_k = \{X_{\eta_k+1}, \ldots, X_{\eta_{k+1}}\}$. We assume the following.

- For each $k, l \in \{0, \ldots, K\}, k \neq l$, it holds that $X_k$ and $X_l$ are independent.
- For each $k \in \{0, \ldots, K\}$, $X_k$ is a subset of an infinite stationary and stable time series $X_k^\infty$ and we have

\[
X_{t+1} = A^*_t(X_t^\top, \ldots, X_{t-L+1}^\top)^\top + \varepsilon_{t+1}, \quad t = \eta_k + L, \ldots, \eta_{k+1} - 1, \tag{4}
\]

where

\[
A^*_t = (A^*_t[1], \ldots, A^*_t[L]) \in \mathbb{R}^{p \times pL}, \quad A^*_t[i] \in \mathbb{R}^{p \times p}, \quad i = 1, \ldots, L,
\]

\[
A^*_{\eta_k} = \ldots = A^*_{\eta_{k+1} - 1} \in \mathbb{R}^{p \times p}, \quad A^*_{\eta_k} \neq A^*_{\eta_{k+1}},
\]

\[
\|A^*_t\|_{op} < 1,
\]

and $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_p)$ independent of $X_t$.

- For each $k \in \{0, \ldots, K\}, i \in \{1, \ldots, L\}$, $X_{\eta_k+i}$ is defined as

\[
X_{\eta_k+i} = A^*_t(\tilde{X}_{\eta_k+i-1}^\top, \ldots, \tilde{X}_{\eta_k+i-L}^\top)^\top + \varepsilon_{\eta_k+i},
\]

where $\tilde{X}_{\eta_k+i-1}$’s are unobserved latent random vectors drawn from $X_k^\infty$. 

4
Note that (5) is imposed to guarantee the stability of the time series between two consecutive change points (see e.g. Eq.(2.1.12) in Lütkepohl, 2005).

- Lastly, we provide a general framework for analyzing general regression-type change point localization problems that include the VAR models above as a special case. We present a thorough study on a high-dimensional regression change point detection problem under i.i.d. observations (see Model 3 below), yielding the sharpest rates in the existing literature. This analysis may be utilized as a blueprint for more complex change point localization problems. In our analysis, we develop a new and refined toolbox for the change point detection community to study more complex data generating mechanisms above and beyond VAR models.

The rest of the paper is organized as follows. The main theoretical results are presented in Section 2. We present the consistency result on Model 1 in Section 2.1, with an extra local refinement result in Section 2.2. A general VAR(L) case is studied in Section 2.3. In Section 3, we provide a unified framework of dynamic programming approaches in regression problems. All the technical details are left in the Appendices.

1.1 Notation
Throughout this paper, we adopt the following notation. For any set $S$, $|S|$ denotes its cardinality. For any vector $v$, let $\|v\|_2$, $\|v\|_1$, $\|v\|_0$ and $\|v\|_\infty$ be its $\ell_2$-, $\ell_1$-, $\ell_0$- and entry-wise maximum norms, respectively; and let $v(j)$ be the $j$th coordinate of $v$. For any square matrix $A \in \mathbb{R}^{n \times n}$, let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ be the smallest and largest eigenvalues of matrix $A$, respectively; let $A_S$ be the sub matrix of $A$ taking entries in $S^\otimes 2$, where $S \subset \{1, \ldots, n\}$. For any matrix $B \in \mathbb{R}^{n \times m}$, let $\|B\|_{op}$ be the operator norm of $B$; let $\|B\|_1 = \|\text{vec}(B)\|_1$, $\|B\|_2 = \|\text{vec}(B)\|_2$ and $\|B\|_0 = \|\text{vec}(B)\|_0$, where $\text{vec}(B) \in \mathbb{R}^{nm}$ is the vectorisation of $B$ by stacking the columns of $B$. In fact, $\|B\|_2$ corresponds to the Frobenius norm of $B$. For any pair of integers $s, e \in \{0, 1, \ldots, n\}$ with $s < e$, we let $(s, e) = \{s + 1, \ldots, e\}$ and $[s, e] = \{s, \ldots, e\}$ be the corresponding integer intervals.

2 Main results
2.1 VAR(1) processes
In this subsection, we provide the theoretical guarantees for the change point estimators arising from the dynamic programming approach, based on data from Model 1. We begin by formulating the assumptions we impose to derive consistency guarantees.

Assumption 1. Consider Model 1. We assume the following.

a. (Sparsity). There exists a subset $S \subset \{1, \ldots, p\}^\otimes 2$ such that

\[ A_t(i, j) = 0, \quad \forall t = 1, \ldots, n, \quad \forall (i, j) \in S^c = \{1, \ldots, p\}^\otimes 2 \setminus S. \]

Let $d_0 = |S|$. 

b. (Spectral density conditions). For each \( k \in \{0, \ldots, K\} \), let \( \Sigma_k(h) \) be the population version of the lag-\( h \) autocovariance function of \( X_k \). The spectral density function

\[
f_k(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Sigma_k(\ell)e^{-i\ell\theta}, \quad \theta \in (-\pi, \pi]
\]

exists for each \( k \). In addition,

\[
M = \max_{k=0, \ldots, K} \mathcal{M}(f_k) = \max_{k=0, \ldots, K} \esssup_{\theta \in (-\pi, \pi]} \Lambda_{\max}(f_k(\theta)) < \infty
\]

and

\[
c_x = \min_{k=0, \ldots, K} \essinf_{\theta \in (-\pi, \pi]} \Lambda_{\min}(f_k(\theta)) > 0.
\]

c. (Signal-to-noise ratio). For any \( \xi > 0 \), there exists an absolute constant \( C_{\text{SNR}} > 0 \), dependent on \( M, c_x \) and \( \xi \), such that

\[
\Delta \kappa^2 \geq C_{\text{SNR}} d_0^2 K \log^{1+\xi}(n \vee p),
\]

where \( \kappa \) and \( \Delta \) are the minimal jump size and minimal spacing defined as follows, respectively,

\[
\kappa = \min_{k=1, \ldots, K+1} \|A^*_\eta_k - A^*_\eta_{k-1}\|_2 \quad \text{and} \quad \Delta = \min_{k=1, \ldots, K+1} (\eta_k - \eta_{k-1}).
\]

Assumption 1(a)-(b) are imposed to guarantee that the Lasso estimators in (3) exhibit good performance, while Assumption 1(c) can be interpreted as a signal-to-noise ratio condition for detecting and estimating the location of the change points. We further elaborate on these conditions next.

- **Sparsity.** The set \( S \) appearing in Assumption 1(a) is a superset of the union of all the nonzero entries in all coefficient matrices. If, alternatively, the sparsity parameter is defined as \( d_0 = \max_{t=1, \ldots, n} |S^t| \), where \( S^t \subset \{1, \ldots, p\}^{\otimes 2} \) and \( A^*_t(i, j) = 0 \), for all \((i, j) \in \{1, \ldots, p\}^{\otimes 2} \setminus S^t\), then the signal-to-noise ratio in (6) and the localization error rate in Theorem 1 would change correspondingly, by replacing the sparsity level \( d_0 \) thereof with \( Kd_0 \).

- **Restricted eigenvalue.** Assumption 1(b) corresponds to a restricted eigenvalue condition (e.g. Bickel et al., 2009; Bühlmann and van de Geer, 2011; van de Geer, 2018); see Section 3.2. The statement in Assumption 1(b) is identical to Assumption 2.1 and the assumption in Proposition 3.1 in Basu and Michailidis (2015), which pertained to a stable VAR process without change points. As pointed out in Basu and Michailidis (2015), this holds for a large class of general linear processes, including stable and invertible ARMA processes.

- **Signal-to-noise ratio.** If \( K = d_0 = 1 \), (6) becomes

\[
\Delta \kappa^2 \gtrsim \log^{1+\xi}(n \vee p),
\]

which matches be the minimax optimal signal-to-noise ratio (up to constants and logarithmic terms) for the univariate mean change point detection problem (see e.g. Chan and Walther, 2013; Frick et al., 2014; Wang et al., 2018b).
In addition, we have
\[
\Delta \geq \frac{C_{\text{SNR}}d_0^2K\log^{1+\xi}(n \lor p)}{\kappa^2} \geq \frac{C_{\text{SNR}}d_0^2K\log^{1+\xi}(n \lor p)}{4C^2d_0} \geq \frac{C_{\text{SNR}}}{4C^2d_0}K\log^{1+\xi}(n \lor p),
\]  
(7)

where the second inequality follows from the bound
\[
\kappa^2 = \min_{k=1,\ldots,K+1} \|A^*_{\eta_k} - A^*_{\eta_{k-1}}\|_2^2 \leq d_0(2C_\beta)^2 = 4C_\beta^2d_0.
\]

If \(\Delta = \Theta(n)\) and \(K = O(1)\), then (7) becomes
\[
n \gtrsim d_0\log^{1+\xi}(n \lor p),
\]
which can be interpreted as an effective sample size condition needed in the Lasso estimation literature.

Another way to inspect Assumption 1(c) is to introduce a normalized jump size
\[
\kappa_0 = \kappa/\sqrt{d_0},
\]
which leads to the signal-to-noise ratio condition
\[
\Delta\kappa_0^2 \geq C_{\text{SNR}}d_0K\log^{1+\xi}(n \lor p).
\]

Similar conditions are required in other change point detection problems, including high-dimensional mean change point detection (Wang and Samworth, 2018), high-dimensional covariance change point detection (Wang et al., 2017), sparse dynamic network change point detection (Wang et al., 2018a), high-dimensional regression change point detection (Wang et al., 2019), to name but a few. Note that in these aforementioned papers, when variants of wild binary segmentation (Fryzlewicz, 2014) were deployed, additional knowledge is needed to get rid of \(K\) in the lower bound of the signal-to-noise ratio. We refer readers to Wang et al. (2018a) for more discussions regarding this point.

The constant \(\xi\) is needed to guarantee consistency and can be set to zero if \(\Delta = o(n)\). We may instead replace it by a weaker condition of the form
\[
\Delta\kappa^2 \gtrsim C_{\text{SNR}}d_0^2\{\log(n \lor p) + a_n\},
\]
where \(a_n \to \infty\) as \(n \to \infty\). We stick with the signal-to-noise ratio condition (6) for simplicity.

We are now ready to state one of the main results of the paper.

**Theorem 1.** Assume Model 1 and the conditions in Assumption 1. Then for the change point estimators \(\{\hat{\eta}_k\}_{k=1}^\hat{K}\) obtained as the solution to the dynamic programming optimization problem given in (1), (2) and (3) with tuning parameters
\[
\lambda = C_\lambda \sqrt{d_0\log(n \lor p)} \quad \text{and} \quad \gamma = C_\gamma(K+1)d_0^2\log(n \lor p),
\]
we have that
\[
\mathbb{P}\left\{\hat{K} = K, \max_{k=1,\ldots,K} |\hat{\eta}_k - \eta_k| \leq \frac{KC_\epsilon d_0^2\log(n \lor p)}{\kappa^2}\right\} \geq 1 - Cn^{-c},
\]
where \(C_\lambda,C_\gamma,C_\epsilon,C,c > 0\) are absolute constants depending only on \(M\) and \(c_x\).
The above result implies that, with probability tending to 1 as \( n \) grows,

\[
\max_{k=1, \ldots, K} \frac{|\tilde{\eta}_k - \eta_k|}{n} \leq \frac{KC_d^2 \log(n \vee p)}{\kappa^2 n} \leq \frac{C_\ell \Delta}{C_{\text{SNR}} n \log \xi(n \vee p)} \rightarrow 0,
\]

where in the second inequality we have used Assumption 1(c). Thus, the localization error converges to zero in probability.

The tuning parameter \( \lambda \) affects the performance of the Lasso estimator. This is shown in Lemma 21. The second tuning parameter \( \gamma \) prevents overfitting while searching the optimal partition as a solution to the problem (1). In particular, \( \gamma \) is determined by the squared \( \ell_2 \)-loss of the Lasso estimator and is of order \( \lambda^2 d_0 \). We will elaborate more on this point in Section 3.2.

2.2 Local refinement

The localization error afforded by Theorem 1 is linear in \( K \), the number of change points. Although the corresponding localization rate is already sharper than any other rates previously established in the literature (see the discussion at the end of Section 2.3), it is possible to improve it by removing the dependence on \( K \) through an additional step, which we refer to as local refinement, detailed in Algorithm 1.

**Algorithm 1** Local refinement. LR(\( \{X(t)\}_{t=1}^n \), \( \{\tilde{\eta}_k\}_{k=1}^K \), \( \zeta \))

**INPUT:** Data \( \{X(t)\}_{t=1}^n \), a collection of time points \( \{\tilde{\eta}_k\}_{k=1}^K \), tuning parameter \( \zeta > 0 \).

\( (\tilde{\eta}_0, \tilde{\eta}_{K+1}) \leftarrow (0, n) \)

for \( k = 1, \ldots, K \) do

\( (s_k, e_k) \leftarrow (2\tilde{\eta}_k - 3/3, 2\tilde{\eta}_k / 3 + 3 \tilde{\eta}_k + 1 / 3) \)

\[
\left( \hat{A}_1, \hat{A}_2, \tilde{\eta}_k \right) \leftarrow \arg \min_{\eta \in \{s_k+2 \ldots e_k-2\}, \|A_1\|_p, \|A_2\|_p < 1, \|A_1\|_p \neq 1} \left\{ \sum_{t=s_k + 1}^\eta \|X_{t+1} - A_1 X_t\|_2^2 + \sum_{t=\eta + 1}^{e_k} \|X_{t+1} - A_2 X_t\|_2^2 \right\} + \zeta \sqrt{\log(n \vee p)} \sum_{i,j=1}^p \sqrt{(\eta - s_k)(A_1)_{ij}^2 + (e_k - \eta)(A_2)_{ij}^2} \right\}
\]

end for

**OUTPUT:** The set of estimated change points \( \{\tilde{\eta}_k\}_{k=1}^K \).

The local refinement algorithm takes as input a preliminary collection of change point estimators \( \{\tilde{\eta}_k\}_{k=1}^K \) such that \( \max_{k=1, \ldots, K} |\tilde{\eta}_k - \eta_k| \) is a small enough fraction of the minimal spacing \( \Delta \) (see condition 9 below) and returns an improved collection of change point estimators \( \{\hat{\eta}_k\}_{k=1}^K \) with a vanishing localization error rate of order \( O \left( \frac{d_0 \log(n \vee p)}{n \kappa^2} \right) \). Interestingly, the initial estimators need not be consistent in order for local refinement to work: all that is required is essentially that the each of the working intervals in Algorithm 1 contains one and only one true change point. This fact allows to refine the search within each working intervals separately, yielding better rates.
In particular, if we use the outputs of (1), (2) and (3) as the inputs of Algorithm 1, then it follows from Theorem 1 and Assumption 1(c) that, for any \( k \in \{1, \ldots, K\} \),
\[
s_k - \eta_{k-1} > \frac{2}{3} \tilde{\eta}_{k-1} + \frac{1}{3} \tilde{\eta}_k - \tilde{\eta}_{k-1} - \epsilon = \frac{1}{3} (\tilde{\eta}_k - \tilde{\eta}_{k-1}) - \epsilon > \Delta/3 - 5\epsilon/3 > 0
\]
and
\[
s_k - \eta_k < \frac{2}{3} \tilde{\eta}_{k-1} + \frac{1}{3} \tilde{\eta}_k - \tilde{\eta}_k + \epsilon = -\frac{2}{3} (\tilde{\eta}_{k-1} - \tilde{\eta}_k) + \epsilon < -2\Delta/3 + 5\epsilon/3 < 0.
\]

**Corollary 2.** Assume the same conditions of Theorem 1. Let \( \{\tilde{\eta}_k\}_{k=1}^K \) be a set of time points satisfying
\[
\max_{k=1,\ldots,K} |\tilde{\eta}_k - \eta_k| \leq \Delta/7. \tag{9}
\]
Let \( \{\hat{\eta}_k\}_{k=1}^\hat{K} \) be the change point estimators generated from Algorithm 1, with \( \{\tilde{\eta}_k\}_{k=1}^K \) and the tuning parameter
\[
\zeta = C_\zeta \sqrt{\log(n \lor p)}
\]
as inputs. Then
\[
P\left\{ \hat{K} = K, \max_{k=1,\ldots,K} |\hat{\eta}_k - \eta_k| \leq \frac{C_\zeta d_0 \log(n \lor p)}{\kappa^2} \right\} \geq 1 - Cn^{-c},
\]
where \( C_\zeta, C_\epsilon, C, c > 0 \) are absolute constants depending only on \( \mathcal{M} \) and \( c_x \).

Compared to the localization error given in Theorem 1, the improved localization error obtained by running the local refinement algorithm and using that estimator as input does not have a direct dependence on \( K \), the number of change points. The intuition for this is as follows.

- Due to the nature of the change point detection problem, there is an innate group structure. This justifies the use of the group Lasso-type penalty, which reduces the localization error by bringing down \( d_2^{\eta} \) to \( d_0^{\eta} \).

- Using condition (9), there is one and only one true change point in every working interval used by the local refinement algorithm. The true change points can then be estimated separately using \( K \) independent searches, in such a way that the final localization rate that does not depend on the number of searches, namely \( \hat{K} \).

Algorithm 1 is a simplified version of Algorithm 3 in Wang et al. (2019), who first pointed out how to leverage the implicit group structure of change point problems in high-dimensional regression settings.

### 2.3 General VAR(\( L \)) processes

We now generalize the results presented in the previous section to VAR(\( L \)) processes. It is well known that, for any general \( L \in \mathbb{Z}_+ \), a VAR(\( L \)) process can be rewritten as a VAR(1) process in the following way. Assuming Model 2, we let \( Y_t = (X_t^\top, \ldots, X_{t-L+1}^\top)^\top, \xi_t = (\epsilon_t^\top, \ldots, \epsilon_{t-L+1}^\top)^\top \) and
\[
A_t^* = \begin{pmatrix}
A_t^*[1] & A_t^*[2] & \cdots & A_t^*[L-1] & A_t^*[L] \\
I & 0 & \cdots & 0 & 0 \\
0 & I & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & I & 0
\end{pmatrix}.
\tag{10}
\]
Then we can rewrite (4) as
\[ Y_{t+1} = A_t^* Y_t + \zeta_t, \quad (11) \]
which is now a VAR(1) process.

Now we are at the stage of providing parallel assumptions and conditions for Model 2. They are nearly identical to the ones we have assumed for Model 1.

**Assumption 2.** Consider Model 2. We assume the following.

a. **(Sparsity).** There exists a subset \( S \subset \{1, \ldots, p\}^2 \) such that
\[ (A_t^*[l])(i, j) = 0, \quad t = 1, \ldots, n, \quad l = 1, \ldots, L, \quad (i, j) \in S^c = \{1, \ldots, p\}^2 \setminus S. \]
Let \( d_0 = L|S| \).

b. **(Spectral density conditions).** For \( k \in \{0, \ldots, K\} \), let \( \Sigma_k(h) \) be the population version of the lag-\( h \) autocovariance function of \( X_k \). The spectral density function
\[ f_k(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Sigma_k(\ell) e^{-i\ell\theta}, \quad \theta \in (-\pi, \pi] \]
exists for each \( k \). In addition,
\[ \mathcal{M} = \max_{k=0, \ldots, K} \mathcal{M}(f_k) = \max_{k=1, \ldots, K} \sup_{\theta \in (-\pi, \pi]} \Lambda_{\max}(f_k) < \infty \]
and
\[ c_x = \min_{k=0, \ldots, K} m(f_k) = \min_{k=1, \ldots, K} \inf_{\theta \in (-\pi, \pi]} \Lambda_{\min}(f_k) > 0. \]

c. **(Signal-to-noise ratio).** For any \( \xi > 0 \), there exists an absolute constant \( C_{\text{SNR}} > 0 \), dependent on \( \mathcal{M} \) and \( c_x \) such that
\[ \Delta \kappa^2 \geq C_{\text{SNR}} d_0^2 K \log^{1+\xi}(n \lor p), \quad (12) \]
where \( \kappa \) and \( \Delta \) are the minimal jump size and minimal spacing defined as follows, respectively,
\[ \kappa = \min_{k=1, \ldots, K+1} \| A_{\eta_k}^* - A_{\eta_{k-1}}^* \|_2 \quad \text{and} \quad \Delta = \min_{k=1, \ldots, K+1} (\eta_k - \eta_{k-1}). \]

Our dynamical programming algorithm for change point localization in Model 2 and under Assumption 2 continues to be the solution to the optimization problem in (1), with appropriate modifications to the quantities \( \mathcal{L}((s, e]) \) and \( \widehat{A}^\lambda_{(s, e]} \) in (2) and (3) respectively. In detail, following the notation defined above in (10) and (11), we set
\[ \mathcal{L}((s, e]) = \sum_{t=s+1}^{e-1} \| Y_{t+1} - \widehat{A}^\lambda_{(s, e]} Y_t \|_2^2, \quad (13) \]
where \( \widehat{A}^\lambda_{(s, e]} \) is a Lasso estimator defined as
\[ \widehat{A}^\lambda_{(s, e]} = \arg \min_{A \in \mathbb{R}^{Lp \times Lp}, \|A\|_{op} < 1} \sum_{t=s+1}^{e-1} \| Y_{t+1} - AY_t \|_2^2 \]
\[ + \lambda \sqrt{\max\{e - s - 1, \log(n \lor p)\}} \sum_{i=1}^{pL} \sum_{j=1}^{pL} |(A)_{ij}|, \quad (14) \]
Corollary 3. Assume Model 2 and the conditions in Assumption 2. Then, the change point estimators \( \{ \tilde{\eta}_k \}_{k=1}^K \) obtained as solution to the dynamic programming optimization problem given (1), (13) and (14) with tuning parameters

\[
\lambda = C_\lambda \sqrt{d_0 \log(n \lor p)} \quad \text{and} \quad \gamma = C_\gamma d_0^2 \log(n \lor p),
\]

are such that

\[
P \left\{ \hat{K} = K, \max_{k=1, \ldots, K} |\tilde{\eta}_k - \eta_k| \leq \frac{KC_\lambda d_0^2 \log(n \lor p)}{\kappa^2} \right\} \geq 1 - Cn^{-c},
\]

where \( C_\lambda, C_\gamma, C, c > 0 \) are absolute constants depending only on \( M \) and \( c_x \).

Next, we show that the local refinement Algorithm 1, applied to the estimators \( \{ \tilde{\eta}_k \}_{k=1}^K \) deliver a smaller localization error rate with no direct dependence on \( K \).

Corollary 4. Assume the same conditions of Corollary 3. Let \( \{ \tilde{\eta}_k \}_{k=1}^K \) be a set of time points satisfying

\[
\max_{k=1, \ldots, K} |\tilde{\eta}_k - \eta_k| \leq \Delta/7.
\]

Let \( \{ \tilde{\eta}_k \}_{k=1}^K \) be the change point estimators generated from Algorithm 1, adjusted based on (13) and (14) with \( \{ \tilde{\eta}_k \}_{k=1}^K \) and

\[
\zeta = C_\zeta \log(n \lor p),
\]

as inputs. Then,

\[
P \left\{ \hat{K} = K, \max_{k=1, \ldots, K} |\hat{\eta}_k - \eta_k| \leq \frac{C_\zeta d_0 \log(n \lor p)}{\kappa^2} \right\} \geq 1 - Cn^{-c},
\]

where \( C_\zeta, C, c > 0 \) are absolute constants depending only on \( M \) and \( c_x \).

Remark 2 (Dependence on the lag \( L \)). We treat the lag \( L \) as a constant, independent of the sample size \( n \). If \( L \) is to vary with \( n \), the resulting localization rates are increasing functions of \( L \), because the cardinality of the set \( S \) in Assumption 2(a) becomes a function of \( L \). Keeping track explicitly of such dependence in the proof of Lemma 29, we see that the population average coefficient matrix has sparsity level \( Ld_0 \). Consequently, the localization errors in Corollaries 3 and 4, and the required signal-to-noise ratio condition (12) to guarantee a vanishing (in probability) localization rate are inflated by \( L^2 \), \( L \) and \( L^2 \), respectively.

It would seem that Corollary 3 follows directly from Theorem 1 and Corollary 2, using the identities (10) and (11). In fact, the proof of Corollary 3 is technically more involved and we do not end up using the representation implied by (10) and (11) in the proof. We leave all the technical details in the Appendices.

We now compare our results with Safikhani and Shojaie (2017).

• In terms of the localization error rate, Safikhani and Shojaie (2017) proves consistency for their methods by assuming that the minimal magnitude of the structural changes \( \kappa \) is a sufficiently large constant independent of \( n \), while our dynamic programming approach is valid...
even when $\kappa$ is allowed to decrease with the sample size $n$. In addition, Safikhani and Shojaie (2017) achieve the localization error bound of order $K\tilde{\Delta}d_n^2$, where $\tilde{\Delta}$ satisfies $d_n^2 \log(p) \lesssim \tilde{\Delta} \lesssim \Delta$ and $d_n = Kd_0$. Translating to our notation, their best localization error is larger than $K^5d_0^2\log(p)/\kappa^2$ even in their setting where $\kappa$ is a constant.

• In terms of methodology, Safikhani and Shojaie (2017) adopted a two-stage procedure: first, a penalized least squares estimator with a total variation penalty is utilized to obtain an initial estimator of change points; then an information-type criterion is applied to identify the significant estimators and to remove the false discoveries. The change point estimators in Safikhani and Shojaie (2017) are selected from fused Lasso estimators, which are sub-optimal for change point detection purposes, especially when the size of the structure change $\kappa$ is small (see e.g. Lin et al., 2017). In addition, the theoretically-valid selection criterion proposed in Safikhani and Shojaie (2017) has a computational cost growing exponentially in $\tilde{K}$, where $\tilde{K}$ is the number of change points estimated by the fused Lasso and in general one has that $\tilde{K} \gg K$.

3 A general framework for regression change point detection problems

In this section we present a general framework to study change point problems in linear models that include special cases the VAR Models 1 and 2 analyzed above. Indeed, the autoregression model is essentially a regression model. Classical time series estimation approaches, including Yule–Walker estimators (e.g. Lütkepohl, 2005), essentially treat $(X_t, X_{t+1}, A_t)$ triplets as $(x_t, y_t, \beta_t)$ in the linear representation

$$y_t = x_t^T \beta_t + \varepsilon_t,$$

and then take extra care in handling the dependence involved. We formulate a general high-dimensional regression change point model framework as follows.

Model 3 (High-dimensional regression problems). Let the data be $\{(x_t, y_t)\}_{t=1}^n \subset \mathbb{R}^p \otimes \mathbb{R}$, satisfying

$$y_t = x_t^T \beta_t^* + \varepsilon_t,$$

where $\beta_t^* \in \mathbb{R}^p$ is the unknown coefficient vector, and $\varepsilon_t$ are independent centered sub-Gaussian random variables with parameters $\sigma_t^2 \leq \sigma_\varepsilon^2$ and independent of $\{x_t\}$.

In addition, there exists a collection of change points $\{\eta_k\}_{k=0}^{K+1} \subset \{0, 1, \ldots, n\}$ with $\eta_0 = 0$ and $\eta_{K+1} = n$ such that

$$\beta_{\eta_k}^* = \ldots = \beta_{\eta_{k+1}}^* \quad \text{and} \quad \beta_{\eta_{k+1}}^* \neq \beta_{\eta_k}^*, \quad k = 0, \ldots, K.$$

Note that, for $p = 1$, Model 3 reduces to Model 1 by setting $y_t = x_{t+1}$. If one further assumes that $x_t$’s are independent and identically distributed, then this is a linear regression problem, which
has been studied before. Lee et al. (2016), Kaul et al. (2018), Lee et al. (2018), among others, focused on the cases where there exists at most one true change point. Leonardi and Bühlmann (2016) and Zhang et al. (2015) considered instead multiple change points and devised consistent change point estimators, albeit with localization error rates worse than the one we establish below in Theorem 5. Wang et al. (2019) also allowed for multiple change points in a regression setting and proposed a variant of the wild binary segmentation (WBS, Fryzlewicz, 2014) method, the performances thereof match the one of the procedure we study next. More detailed discussions can be found after Theorem 5.

A general dynamic programming algorithm for change point analysis in high-dimensional regression problems entails solving the core optimization problem (1) with

$$
\mathcal{L}(I) = \sum_{t \in I} (y_t - x_t^\top \hat{\beta}_I^\lambda)^2 
$$

and

$$
\hat{\beta}_I^\lambda = \arg \min_{v \in \mathbb{R}^p} \left\{ \sum_{t \in I} (y_t - x_t^\top v)^2 + \lambda \sqrt{\max\{ |I|, \log(n \vee p) \} \|v\|_1} \right\},
$$

where $\lambda \geq 0$ is a tuning parameter.

In order to obtain the results in Theorem 1 and Corollary 3, we have developed a general strategy for analyzing the performance of the dynamic programming estimator (1) in the high-dimensional linear regression framework of Model 3. We detail the steps of this general approach in Section 3.2, where we lay out the blueprint of our proofs for change point analysis in the specialized setting of high-dimensional linear regression with independent, sub-Gaussian covariates. The overall proof strategy is highly modular, and in order to extend it to non-i.i.d. settings, such as VAR processes, it is enough to only modify its parts involving few high-probability bounds to accommodate for the stochastic dependence. We believe that our approach is quite general and can be applied to other problems as well. Furthermore, as a by-product of our analysis, we are able to establish localization rates for change point analysis in high-dimensional linear regression settings that sharpen existing results.

3.1 A high-dimensional regression problem

In this section we analyze the performance of the change point estimator arising as the solution to the dynamic programming optimization problem (1) for a high-dimensional linear regression problem.

Assumption 3. Consider the model defined in Model 3, where $x_t$’s are independent and identically distributed centred sub-Gaussian random vectors with $\mathbb{E}(x_t x_t^\top) = \Sigma$. We impose the following additional assumptions.

a. (Sparsity). There exists a subset $S \subset \{1, \ldots, p\}$ such that

$$
\beta_t^*(j) = 0, \quad t = 1, \ldots, n, \quad j \in S^c = \{1, \ldots, p\} \setminus S.
$$

Let $d_0 = |S|$.

b. (Boundedness). For some absolute constant $C_\beta > 0$, $\max_{t=1,\ldots,n} \|\beta_t^*\|_\infty \leq C_\beta$. 

13
c. (Minimal eigenvalue). We have that
\[ \Lambda_{\min}(\Sigma) = c_2^2 > 0 \quad \text{and} \quad \max_{j=1,\ldots,p} (\Sigma)_{jj} = C x > 0. \]

d. (Signal-to-noise ratio). Assume for any \( \xi > 0 \), there exists an absolute constant \( C_{\text{SNR}} > 0 \) such that
\[ \Delta \kappa^2 \geq C_{\text{SNR}} d_0^2 K \sigma_x^2 \log^{1+\xi}(n \vee p), \]  
where \( \kappa \) and \( \Delta \) are the minimal jump size and minimal spacing defined as follows, respectively,
\[ \kappa = \min_{k=1,\ldots,K+1} \| \beta_{\eta_k}^* - \beta_{\eta_{k-1}}^* \|_2 \quad \text{and} \quad \Delta = \min_{k=1,\ldots,K+1} (\eta_k - \eta_{k-1}). \)

Assumption 3(a) and (c) are standard conditions required for consistency of Lasso estimators in i.i.d. cases, and Assumption 3(d) is the signal-to-noise ratio condition for the change point detection purpose. In addition, Assumption 3(b) is served as a counterpart of the \( \| A_t \|_{\text{op}} < 1 \) in Models 1 and 2, and it can also be found in Wang et al. (2019) and Leonardi and Bühlmann (2016), among others.

Remark 3. The assumption that the covariates \( \{ x_t \}_{t=1}^n \) have the same covariance matrix \( \Sigma \) can be relaxed to the weaker condition that \( \Sigma_1 \preceq E(x_t x_t^\top) \preceq \Sigma_2 \) for all \( t \) and some positive definite matrices \( \Sigma_1 \) and \( \Sigma_2 \), satisfying appropriate eigenvalue conditions. For simplicity, we will not pursue this extension.

Theorem 5. Assume Model 3 and the conditions in Assumption 3. Then, the change point estimators \( \{ \hat{\eta}_k \}_{k=1}^K \) obtained as solution to the dynamic programming optimization problem given in (1), (15) and (16) with tuning parameters
\[ \lambda = C_\lambda \sigma_\varepsilon \sqrt{d_0 \log(n \vee p)} \quad \text{and} \quad \gamma = C_\gamma \sigma_x^2 (K + 1) d_0^2 \log(n \vee p), \]
are such that
\[ P \left\{ \hat{K} = K, \max_{k=1,\ldots,K} | \hat{\eta}_k - \eta_k | \leq \frac{KC_\zeta d_0^2 \sigma_x^2 \log(n \vee p)}{\kappa^2} \right\} \geq 1 - C (n \vee p)^{-c}, \tag{18} \]
where \( C_\lambda, C_\gamma, C_\zeta, C, c > 0 \) are absolute constants depending only on \( C_\beta, C_x \) and \( c_x \).

As before, it is possible to eliminate the dependence on \( K \) in the localization error (18) by improving on the original change point estimators through a local refinement algorithm.

Corollary 6. Assume the same conditions of Theorem 5. Assume data are generated from Model 3, satisfying Assumption 3. Let \( \{ \widetilde{\eta}_k \}_{k=1}^K \) be a set of time points satisfying
\[ \max_{k=1,\ldots,K} | \widetilde{\eta}_k - \eta_k | \leq \Delta / 7. \tag{19} \]

Let \( \{ \tilde{\eta}_k \}_{k=1}^\hat{K} \) be the change point estimators generated from Algorithm 1 with \( \{ \widetilde{\eta}_k \}_{k=1}^K \) and
\[ \zeta = C_\zeta \sqrt{\log(n \vee p)} \]
as inputs and changing (8) to

$$
\left( \hat{\beta}_1, \hat{\beta}_2, \hat{\eta}_k \right) \leftarrow \arg \min_{\eta \in \{ s_k + 2, \ldots, e_k - 1 \}, \beta_1, \beta_2 \in \mathbb{R}^p, \beta_1 \neq \beta_2} \left\{ \sum_{t=s_k+1}^{e_k} \| y_t - \beta_1^\top x_t \|_2^2 + \sum_{t=n+1}^{e_k} \| y_t - \beta_2 x_t \|_2^2 + c \sum_{i=1}^{p} \sqrt{(\eta - s_k)(\beta_1)_i^2 + (e_k - \eta)(\beta_2)_i^2} \right\},
$$

Then,

$$
P\left\{ \hat{K} = K, \max_{k=1, \ldots, K} | \hat{\eta}_k - \eta_k | \leq \frac{C d_0 \log(n \lor p)}{\kappa^2} \right\} \geq 1 - C n^{-c},
$$

where $C_\zeta, C_\epsilon, C, c > 0$ are absolute constants depending only on $C_\beta, M$ and $c_x$.

We now conclude this subsection by discussing how our contributions compared with the results of Wang et al. (2019) and of Leonardi and Bühlmann (2016), the references that consider exactly the same change point problem as we study in Section 3.1.

- **Wang et al. (2019)** proposed different algorithms, all of which are variants of wild binary segmentation, with or without additional Lasso estimation procedures. Those methods inherit both the advantages and the disadvantages of WBS. Compared with dynamic programming, WBS-based methods require an additional tuning parameter and additional information about the minimal spacing when choosing the random intervals. With this additional piece of information, Theorem 1 in Wang et al. (2019) achieved the same statistical accuracy in terms of the localization error rate as Theorem 5 above.

In terms of computational cost, the methods in Wang et al. (2019) are of order $O(K^2 n \cdot \text{Lasso}(n))$, where $K$, $n$ and $\text{Lasso}(n)$ denote the number of change points, the sample size and the computational cost of Lasso algorithm with sample size $n$, respectively, while the dynamic programming approach of this paper is of order $O(n^2 \cdot \text{Lasso}(n))$. Thus, when $K \lesssim \sqrt{n}$, the algorithm in Wang et al. (2019) is computationally more efficient, but when $K \gtrsim \sqrt{n}$, the method in this paper has smaller complexity.

- **Leonardi and Bühlmann (2016)** analysed two algorithms, one based on a dynamic programming approach, and the other on binary segmentation, and claimed that they both yield the same localization, which is, in our notation\(^1\),

$$
\sum_{k=1}^{K} | \hat{\eta}_k - \eta_k | \lesssim \frac{d_0^2 \sqrt{n \log(n \lor p)}}{\kappa^2},
$$

(20)

It is not immediate to directly compare the sum of all localization errors, used by Leonardi and Bühlmann (2016), with the maximum localization error, which is the target in this paper. Using a worst case upper bound, Theorem 5 yields that

$$
\sum_{k=1}^{K} | \hat{\eta}_k - \eta_k | \lesssim \frac{K^2 d_0^2 \sigma^2 \log(n \lor p)}{\kappa^2}.
$$

\(^1\) The error bound in Leonardi and Bühlmann (2016) is originally of the form $\sum_{k=1}^{K} | \hat{\eta}_k - \eta_k | \lesssim \frac{d_0^2 \sqrt{n \log(n \lor p)}}{\kappa^2}$ under a slightly stronger assumption than ours. In the more general settings of Assumption 3, the localization error bound of Leonardi and Bühlmann (2016) is of the form (20), based on our communication with the authors.
In light of Corollary 6, this error bound can be sharpened, using the local refinement Algorithm 1 to
\[ \sum_{k=1}^{K} |\hat{\eta}_k - \eta_k| \lesssim \frac{K d_0 \sigma^2 \log(n \vee p)}{\kappa^2}. \]

As long as \( K^2 \lesssim \sqrt{\frac{n}{\log(np)}} \), or, using the local refinement algorithm, \( K \lesssim \sqrt{\frac{n}{\log(np)}} \), our localization rates are better than the one implied by (20).

It is not easy to compare directly the assumptions used in Leonardi and Bühlmann (2016) with the ones we formulate due to the slightly different ways we use to present them. For instance, the conditions in Theorem 3.1 of Leonardi and Bühlmann (2016) imply that, in our notation, the following is needed for consistency,
\[ \Delta \gtrsim \sqrt{n \log(p)}, \]
even if the sparsity parameter \( d_0 = \Theta(1) \). However in our case, in view of (17), if we assume \( d_0 = \kappa = \Theta(1) \), then we only require \( \Delta \gtrsim \log^{1+\epsilon}(n \vee p) \) for consistency.

### 3.2 Sketch of the proof of Theorem 5

In this section, we sketch the proof of Theorem 5, which serves as a general template to derive upper bounds on the localization error change point problems in the general regression framework described in Model 3. We will leave all detailed technical arguments to the Appendices.

Theorem 5 is an immediate consequence of Propositions 7 and 8.

**Proposition 7.** Under the same conditions in Theorem 5 and letting \( \hat{\mathcal{P}} \) being the solution to (1), the following hold with probability at least \( 1 - C(n \vee p)^{-c} \).

(i) For each interval \( \hat{I} = (s, e] \in \hat{\mathcal{P}} \) containing one and only one true change point \( \eta \), it must be the case that
\[ \min\{e - \eta, \eta - s\} \leq C_{\epsilon} \left( \frac{d_0 \lambda^2 + \gamma}{\kappa^2} \right), \]
where \( C_{\epsilon} > 0 \) is an absolute constant;

(ii) for each interval \( \hat{I} = (s, e] \in \hat{\mathcal{P}} \) containing exactly two true change points, say \( \eta_1 < \eta_2 \), it must be the case that
\[ \max\{e - \eta_2, \eta_1 - s\} \leq C_{\epsilon} \left( \frac{d_0 \lambda^2 + \gamma}{\kappa^2} \right), \]
where \( C_{\epsilon} > 0 \) is an absolute constant;

(iii) for all consecutive intervals \( \hat{I} \) and \( \hat{J} \) in \( \hat{\mathcal{P}} \), the interval \( \hat{I} \cup \hat{J} \) contains at least one true change point; and

(iv) no interval \( \hat{I} \in \hat{\mathcal{P}} \) contains strictly more than two true change points.

The four cases in Proposition 7 are proved in Lemmas 14, 15, 16 and 17, respectively, and Proposition 7 is proved consequently.
Proposition 8. Under the same conditions in Theorem 5, with \( \hat{P} \) being the solution to (1), satisfying \( K \leq |\hat{P}| \leq 3K \), then with probability at least \( 1 - C(n \lor p)^{-c} \), it holds that \( |\hat{P}| = K \).

Proof of Theorem 5. It follows from Proposition 7 that, \( K \leq |\hat{P}| \leq 3K \). This combined with Proposition 8 completes the proof.

The key ingredients of the proofs of both Propositions 7 and 8 are two types of deviation inequalities.

• Restricted eigenvalues. In the literature on sparse regression, there are several versions of the restricted eigenvalue conditions (see, e.g. Bühlmann and van de Geer, 2011). In our analysis, such conditions amount to controlling the probability of the event

\[
E_I = \left\{ \sqrt{\sum_{t \in I} (x_t^\top v)^2} \geq \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 9C_x \sqrt{\log(p)} \|v\|_1, \quad \forall v \in \mathbb{R}^p \right\},
\]

which is done in Lemma 9.

• Deviations bounds of scaled noise. In addition, we need to control the deviations of the quantities of the form

\[
\left\| \sum_{t \in I} \varepsilon_t x_t \right\|_\infty.
\]

See Lemma 10.

In standard analyses of the performance of the Lasso estimator, as detailed e.g. in Section 6.2 of Bühlmann and van de Geer (2011), the combination of restricted eigenvalues conditions and large probability bounds on the noise lead to oracle inequalities for the estimation and prediction errors in situations in which there exists no change point and the data are independent. We have extended this line of arguments to the present, more challenging settings, to derive analogous oracle inequalities. We emphasize a few points in this regard.

• In standard analyses of the Lasso estimator, where there is one and only one true coefficient vector, the magnitude of \( \lambda \) is determined as a high-probability upper bound to (21). However in our situation, in order to control the \( \ell_1 \)- and \( \ell_2 \)-loss of the estimators \( \hat{\beta}_I^\lambda \), where the interval \( I \) contains more than one true coefficient vectors, the value of \( \lambda \) needs to be inflated by a factor of \( \sqrt{d_0} \). This is detailed in Lemma 13; see, in particular, (32).

• The magnitude of the tuning parameter \( \gamma \) is determined based on an appropriate oracle inequality for the Lasso and on the number of true change points; more precisely, \( \gamma \) can be derived as a high-probability bound for

\[
\left| \sum_{t \in I} \{ (y_t - x_t^\top \hat{\beta}_I^\lambda)^2 - (y_t - x_t^\top \beta_t^*)^2 \} \right|.
\]

See Lemma 12 for details.

The fact that \( \gamma \) is linear in the number of change point \( K \) is to prompt the consistency. This is shown in (52) in the proof of Proposition 8.
The final localization error is obtained by the following calculations. Assume that there exists one and only one true change point $\eta \in I = (s, e]$. Define $I_1 = (s, \eta]$ and $I_2 = [\eta, e]$. Let $\beta^*_I$ and $\beta^*_I$ be the two true coefficient vectors in $I_1$ and $I_2$, respectively. For readability, below we will omit all constants here and use the symbol $\lesssim$ to denote an inequality up to hidden universal constants. We first assume by contradiction that

$$\min\{|I_1|, |I_2|\} \gtrsim d_0 \log(n \lor p),$$

then use oracle inequalities to establish that

$$\sum_{t \in I_1} \{x_t^\top (\hat{\beta}^I_T - \beta^*_I)\}^2 + \sum_{t \in I_2} \{x_t^\top (\hat{\beta}^I_T - \beta^*_I)\}^2$$

$$\lesssim \lambda \max\{|I_1|, \log(n \lor p)\} \{\sqrt{d_0}(\|\hat{\beta}^I_T - \beta^*_I\|_2 + \|\hat{\beta}^I_T(S^c)\|_1) + \lambda \sqrt{\max\{|I_2|, \log(n \lor p)\} \{\sqrt{d_0}(\|\hat{\beta}^I_T - \beta^*_I\|_2 + \|\hat{\beta}^I_T(S^c)\|_1) + \gamma \}

$$

$$\lesssim \lambda \sqrt{|I_1|} \{\sqrt{d_0}(\|\hat{\beta}^I_T - \beta^*_I\|_2 + \|\hat{\beta}^I_T(S^c)\|_1) + \lambda \sqrt{|I_2|} \{\sqrt{d_0}(\|\hat{\beta}^I_T - \beta^*_I\|_2 + \|\hat{\beta}^I_T(S^c)\|_1) + \gamma \}$$

$$\lesssim \frac{\lambda^2 d_0}{c^2} + |I_1| \|\hat{\beta}^I_T - \beta^*_I\|_2^2 + |I_2| \|\hat{\beta}^I_T - \beta^*_I\|_2^2 + \lambda^2 + (|I_1|^2 + |I_2|^2) \|\hat{\beta}^I_T(S^c)\|_2^2 + \gamma,$$

where the second inequality follows (22) and the third inequality follows from $2ab \leq a^2 + b^2$ and from setting

$$a = \lambda \sqrt{d_0} \quad \text{and} \quad b = \sqrt{|I_1|} \|\hat{\beta}^I_T - \beta^*_I\|_2.$$

Next we apply the restricted eigenvalue conditions along with standard arguments from the Lasso literature to establish that

$$\sum_{t \in I_1} \{x_t^\top (\hat{\beta}^I_T - \beta^*_I)\}^2 + \sum_{t \in I_2} \{x_t^\top (\hat{\beta}^I_T - \beta^*_I)\}^2$$

$$\geq c^2_x |I_1| \|\hat{\beta}^I_T - \beta^*_I\|_2^2 + c^2_x |I_2| \|\hat{\beta}^I_T - \beta^*_I\|_2^2 \geq c^2_x \kappa^2 \epsilon,$$

where $\epsilon$ is an upper bound on the localization error. Combining (23) and (24) leads to

$$\epsilon \lesssim \frac{\lambda^2 d_0}{c^2} + \gamma.$$

Finally, the signal-to-noise ratio condition that one needs to assume in order to obtain consistent localization rates is determined by setting $\epsilon \lesssim \Delta$.

Table 1 summarises the schematics of our proofs for the main results of the paper, namely Theorems 5, 1 and 3. For each theorem, the table indicates the Lemmas used to determine the magnitudes of the tuning parameters $\lambda$ and $\gamma$, which in turn jointly determine the localization error rate $\epsilon/n$.

As for the proofs related with Algorithm 1, Corollaries 2, 4 and 6 are all based on an oracle inequality of the group Lasso estimator. In the context of the high-dimensional regression problem Model 3, once it is established that

$$\sum_{t=s+1}^{e} \|\hat{\beta}_t - \beta^*_I\|_2^2 \leq \delta \leq \kappa \sqrt{\Delta},$$

(25)
where $\delta \asymp d_0 \log(n \lor p)$ and where there is one and only one change point in the interval $(s, e]$ for both the sequence $\{\tilde{\beta}_t\}$ and $\{\beta_t^*\}$, then the final claim follows immediately that the refined localization error $\epsilon$ satisfies

$$\epsilon \leq \delta / \kappa^2.$$  

The group Lasso penalty is deployed to prompt (25) and the designs of the algorithm guarantee the desirability of each working interval.

4 Conclusions

This paper considers change point localization in general linear regression settings, which include as special cases both vector autoregressive process models and linear regression models. We have developed several procedures for change point localization that can be characterized as solutions to a common optimization framework and that can be efficiently implemented using a combination of dynamic programming and Lasso-type estimators. We have demonstrated that our methods yield the sharpest localization rates for autoregressive processes and match the best known rates for change point localization in linear regression model. We further conjecture that the rates we obtain are minimax optimal. Both minimax rates and extensions of this framework beyond sparse models to other models of low-dimensional structure remain important open questions for future research.

Appendices

In Sections A and B, we detail the proofs of Theorem 5 and Corollary 6. In Sections C and D, we omit the repetitive parts and only provide the unique techniques needed for the proofs of Theorem 1, Corollaries 2, 3 and 4.

A Proof of Theorem 5

The proof of Theorem 5 proceeds through several steps. For convenience, Figure 1 provides a roadmap for the entire proof. Throughout this section, with some abuse of notation, for any interval $I \subset (0, n]$, we denote with $\beta_I^* = |I|^{-1} \sum_{t \in I} \beta_t^*$.
A.1 Large probability events

Lemma 9. For Model 3, under Assumption 3(c), for any interval $I \subset (0, n]$, it holds that

$$\mathbb{P}\{\mathcal{E}_I\} \geq 1 - c_1 \exp(-c_2 |I|),$$

where $c_1, c_2 > 0$ are absolute constants only depending on the distributions of covariants $\{x_t\}$, and

$$\mathcal{E}_I = \left\{ \left(\sum_{t \in I} (x_t^\top v)^2 \geq \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 9C_x \sqrt{\log(p)} \|v\|_1, \quad v \in \mathbb{R}^p \right) \right\}.$$

This follows from the same proof as Theorem 1 in Raskutti et al. (2010), therefore we omit the proof of Lemma 9. For interval $I$ satisfying $|I| > C d_0 \log(p)$, an immediate consequence of Lemma 9 is a restricted eigenvalue condition (e.g. van de Geer and Bühlmann, 2009; Bickel et al., 2009). It will be used repeatedly in the rest of this paper.

It will become clearer in the rest of the paper, we only deal with intervals satisfying $|I| \gtrapprox d_0 \log(n \lor p)$ when considering the events $\mathcal{E}_I$.

Lemma 10. For Model 3, under Assumption 3(c), for any interval $I \subset (0, n]$, it holds that for any

$$\lambda \geq \lambda_1 := C_\lambda \sigma_x \sqrt{\log(n \lor p)},$$

where $C_\lambda > 0$ is a large enough absolute constant such that, we have

$$\mathbb{P}\{\mathcal{B}_I(\lambda)\} > 1 - 2(n \lor p)^{-c_3},$$

where

$$\mathcal{B}_I(\lambda) = \left\{ \left\| \sum_{t \in I} \varepsilon_t x_t \right\|_\infty \leq \lambda \sqrt{\max\{|I|, \log(n \lor p)|} / 8 \right\},$$

where $c_3 > 0$ is an absolute constant depending only on the distributions of covariants $\{x_t\}$ and $\{\varepsilon_t\}$.

For notational simplicity, we drop the dependence on $\lambda$ in the notation $\mathcal{B}_I(\lambda)$. 
Proof. Since $\varepsilon_t$'s are sub-Gaussian random variables and $x_t$'s are sub-Gaussian random vectors, we have that $\varepsilon_t x_t$'s are sub-Exponential random vectors with parameter $C_\varepsilon \sigma_\varepsilon$ (see e.g. Lemma 2.7.7 in Vershynin, 2018). It then follows from Bernstein’s inequality (see e.g. Theorem 2.8.1 in Vershynin, 2018) that for any $t > 0$,

$$
\mathbb{P}\left\{ \left\| \sum_{t \in I} \varepsilon_t x_t \right\|_\infty > t \right\} \leq 2p \exp \left\{ -c \min \left\{ \frac{t^2}{|I|C_\varepsilon^2 \sigma_\varepsilon^2}, \frac{t}{C_\varepsilon \sigma_\varepsilon} \right\} \right\}.
$$

Taking

$$
t = C_\lambda C_x / 4\sigma_\varepsilon \sqrt{\log(n \lor p)} \sqrt{\max\{|I|, \log(n \lor p)|}
$$

yields that

$$
\mathbb{P}\{B_I\} > 1 - 2(n \lor p)^{-c_3},
$$

where $c_3 > 0$ is an absolute constant depending on $C_\lambda, C_x, \sigma_\varepsilon$. \qed

### A.2 Auxiliary lemmas

**Lemma 11.** For Model 3, under Assumption 3(a) and (c), if there exists no true change point in $I = (s, e]$, with $|I| > 288^2 C_\varepsilon^2 d_0 \log(n \lor p)/C_x^2$ and

$$
\lambda \geq \lambda_1 := C_\lambda \sigma_\varepsilon \sqrt{\log(n \lor p)},
$$

where $C_\lambda > 0$ being an absolute constant, it holds that

$$
\mathbb{P}\left\{ \left\| \hat{\beta}_I^\lambda - \beta_I^* \right\|_2 \leq \frac{c_3 \lambda \sqrt{d_0}}{\sqrt{|I|}}, \left\| \hat{\beta}_I^\lambda - \beta_I^* \right\|_1 \leq \frac{c_3 \lambda d_0}{\sqrt{|I|}} \right\}
$$

$$
\geq 1 - c_1 (n \lor p)^{-288^2 C_\varepsilon^2 d_0/C_x^2} - 2(n \lor p)^{-c_3},
$$

where $C_3 > 0$ is an absolute constant depending on all the other absolute constants, $c_1, c_2, c_3$ are absolute constants defined in Lemmas 9 and 10.

**Proof.** Let $v = \hat{\beta}_I^\lambda - \beta_I^*$. Since $|I| > \log(n \lor p)$, it follows from the definition of $\hat{\beta}_I^\lambda$ that

$$
\sum_{t \in I} (y_t - x_t^\top \hat{\beta}_I^\lambda)^2 + \lambda \sqrt{|I|} \left\| \hat{\beta}_I^\lambda \right\|_1 \leq \sum_{t \in I} (y_t - x_t^\top \beta_I^*)^2 + \lambda \sqrt{|I|} \left\| \beta_I^* \right\|_1,
$$

which leads to

$$
\sum_{t \in I} (x_t^\top v)^2 + \lambda \sqrt{|I|} \left\| \hat{\beta}_I^\lambda \right\|_1 \leq \lambda \sqrt{|I|} \left\| \beta_I^* \right\|_1 + 2 \sum_{t \in I} \varepsilon_t x_t^\top v \leq \lambda \sqrt{|I|} \left\| \beta_I^* \right\|_1 + \frac{\lambda}{2} \sqrt{|I|} \left\| v \right\|_1, \quad (26)
$$

where the last inequality holds on the event $B_I$, with the choice of $\lambda$ and due to Lemma 10. Note that

$$
\left\| \hat{\beta}_I^\lambda \right\|_1 \geq \left\| \beta_I^*(S) \right\|_1 - \left\| v(S) \right\|_1 + \left\| \hat{\beta}_I^\lambda(S^c) \right\|_1 \quad (27)
$$

and

$$
\left\| v \right\|_1 = \left\| v(S) \right\|_1 + \left\| \hat{\beta}_I^\lambda(S^c) \right\|_1. \quad (28)
$$

21
Combining (26), (27) and (28) yields
\[
\sum_{t \in I} (x_t^T v)^2 + \frac{\lambda}{2} \sqrt{|I|} \|\hat{\beta}_I^\lambda(S^c)\|_1 \leq \frac{3\lambda}{2} \sqrt{|I|} \|\hat{\beta}_I^\lambda(S)\|_1,
\] (29)
which in turn implies
\[
\|\hat{\beta}_I^\lambda(S^c)\|_1 \leq 3\|\hat{\beta}_I^\lambda(S)\|_1.
\]

On the event of \(\mathcal{E}_I\), it holds that
\[
\sqrt{\sum_{t \in I} (x_t^T v)^2} \geq \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 9C_x \sqrt{\log(p)} \|v\|_1
\]
\[
= \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 9C_x \sqrt{\log(p)} \|v(S)\|_1 - 9C_x \sqrt{\log(p)} \|v(S^c)\|_1
\]
\[
\geq \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 36C_x \sqrt{\log(p)} \|v(S)\|_1 \geq \frac{c_x \sqrt{|I|}}{4} \|v\|_2 - 36C_x \sqrt{d_0 \log(p)} \|v(S)\|_2
\]
\[
\geq \left( \frac{c_x \sqrt{|I|}}{4} - 36C_x \sqrt{d_0 \log(p)} \right) \|v\|_2 > \frac{c_x \sqrt{|I|}}{8} \|v\|_2,
\] (30)
where the second inequality follows from (29), the third inequality follows from Assumption 3(a) and the last inequality follows from the choice of \(|I|\).

Combining (29) and (30) leads to
\[
\frac{c_x^2 |I|}{64} \|v\|_2^2 \leq \frac{3\lambda}{2} \sqrt{|I|} \|v(S)\|_1 \leq \frac{3\lambda}{2} \sqrt{|I|d_0} \|v\|_2,
\]
therefore
\[
\|v\|_2 \leq \frac{96\lambda \sqrt{d_0}}{\sqrt{|I|c_x^2}}
\]
and
\[
\|v\|_1 = \|v(S)\|_1 + \|v(S^c)\|_1 \leq 4\|v(S)\|_1 \leq 4\sqrt{d_0} \|v\|_2 \leq \frac{384\lambda d_0}{\sqrt{|I|c_x^2}}
\]

\(\square\)

**Lemma 12.** For Model 3, under Assumption 3(a) and (c), if there exists no true change point in \(I = (s, e]\), and
\[
\lambda \geq \lambda_1 := C_\lambda \sigma_x \sqrt{\log(n \vee p)},
\]

where \(C_\lambda > 0\) being an absolute constant, it holds that if \(|I| \geq 288^2 C_x^2 d_0 \log(n \vee p)/c_x^2\), then
\[
P\left\{ \left( \sum_{t \in I} \left( (y_t - x_t^T \beta)^2 - (y_t - x_t^T \beta^*)^2 \right) \right) \leq \lambda^2 d_0 \right\} \geq 1 - c_1 (n \vee p)^{-288^2 C_x^2 d_0 c_2/c_x^2 - 2(n \vee p)^{-c_3}};
\]
if \(|I| < 288^2 C_x^2 d_0 \log(n \vee p)/c_x^2\), then
\[
P\left\{ \left( \sum_{t \in I} \left( (y_t - x_t^T \beta)^2 - (y_t - x_t^T \beta^*)^2 \right) \right) \leq C_4 \lambda \sqrt{\log(n \vee p) d_0^3/2} \right\} \geq 1 - 2(n \vee p)^{-c_3},
\]
where \(C_4 > 0\) is an absolute constant depending on all the other constants.

22
Proof. To ease notation, in this proof, let \( \hat{\beta} = \hat{\beta}_t^0 \) and \( \beta^* = \beta_t^* \).

**Case 1.** If \( |I| \geq 288^2 C^2_x d_0 \log(n \vee p)/c_x^2 \), then \( |I| > \log(n \vee p) \). With probability at least \( 1 - c_1 \exp(-c_2 |I|) - 2(n \vee p)^{-c_3} \), we have that

\[
\sum_{t \in I} \{(y_t - x_t^T \hat{\beta})^2 - (y_t - x_t^T \beta^*)^2\} \leq \lambda \sqrt{|I|} \|\beta^*\|_1 - \lambda \sqrt{|I|} \|\hat{\beta}\|_1 \leq \lambda \sqrt{|I|} \|\beta - \beta^*\|_1 \leq C_3 \lambda^2 d_0,
\]

where the first inequality follows from the definition of \( \hat{\beta} \) and the second is due to Lemma 11.

**Case 2.** If \( |I| < 288^2 C^2_x d_0 \log(n \vee p)/c_x^2 \), then

\[
\sum_{t \in I} \{(y_t - x_t^T \hat{\beta})^2 - (y_t - x_t^T \beta^*)^2\} \leq \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|\beta^*\|_1 \leq C_4 \lambda \sqrt{\log(n \vee p) d_0^{3/2}},
\]

since \( \|\beta^*\|_1 \leq C_\beta d_0 \). In addition, it holds with probability at least \( 1 - 2(n \vee p)^{-c_3} \) that

\[
\sum_{t \in I} \{(y_t - x_t^T \beta^*)^2 - (y_t - x_t^T \hat{\beta})^2\} = -\sum_{t \in I} (x_t^T \beta^* - x_t^T \hat{\beta})^2 + 2 \sum_{t \in I} \varepsilon_t x_t^T (\hat{\beta} - \beta^*)
\]

\[
\leq -\sum_{t \in I} (x_t^T \beta^* - x_t^T \hat{\beta})^2 + \sum_{t \in I} (x_t^T \beta^* - x_t^T \hat{\beta})^2 + \sum_{t \in I} \varepsilon_t^2 \leq \sum_{t \in I} \varepsilon_t^2
\]

\[
\leq \max\{\sqrt{|I|} \log(n \vee p), \log(n \vee p)\} \leq C_4 \lambda \sqrt{\log(n \vee p) d_0^{3/2}},
\]

where the first inequality follow from \( 2ab \leq a^2 + b^2 \) and letting \( a = \varepsilon_t, b = x_t^T (\hat{\beta} - \beta^*), \) the third inequality follows from the sub-Gaussianity of \( \{\varepsilon_t\} \).

\[\square\]

**Lemma 13.** For Model 3, under Assumption \( \beta'(a)-(c) \), for any interval \( I = (s, e) \) and

\[
\lambda \geq \lambda_2 := C_\lambda \sigma \sqrt{d_0 \log(n \vee p)},
\]

where \( C_\lambda > 8C_\beta C_x / \sigma \epsilon \), it holds with probability at least \( 1 - 2(n \vee p)^{-c} \) that,

\[
\|\hat{\beta}_J^0 (S^*)\|_1 \leq 3 \|\hat{\beta}_J^0 (S)\|_1.
\]

If in addition, the interval \( I \) satisfies \( |I| > 288^2 C^2_x d_0 \log(n \vee p)/c_x^2 \), it holds with probability at least \( 1 - c_1 (n \vee p)^{-288^2 C^2_x d_0 c_2 / c_x^2 - 2(n \vee p)^{-c_3}} \) that

\[
\left\| \hat{\beta}_I^0 - \frac{1}{|I|} \sum_{t \in I} \beta_t^* \right\|_2 \leq \frac{C_5 \lambda \sqrt{d_0}}{\sqrt{|I|}} \text{ and } \left\| \hat{\beta}_I^0 - \frac{1}{|I|} \sum_{t \in I} \beta_t^* \right\|_1 \leq \frac{C_5 \lambda d_0}{\sqrt{|I|}},
\]

where \( C_5 > 0 \) is an absolute constant depending on other constants.

**Proof.** Denote \( \hat{\beta} = \hat{\beta}_I^0 \) and \( \beta^* = (|I|)^{-1} \sum_{t \in I} \beta_t^* \). It follows from the definition of \( \hat{\beta} \) that

\[
\sum_{t \in I} (y_t - x_t^T \beta)^2 + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|\hat{\beta}\|_1 \leq \sum_{t \in I} (y_t - x_t^T \beta^*)^2 + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|\beta^*\|_1,
\]

which leads to

\[
\sum_{t \in I} \{(x_t^T (\hat{\beta} - \beta^*))^2 + 2 \sum_{t \in I} (y_t - x_t^T \beta^*) x_t^T (\beta^* - \hat{\beta}) + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|\hat{\beta}\|_1
\]

\[\textit{23}\]
If, in addition, it holds that 

\[ \sum_{t \in I} \left\{ x_t^T (\hat{\beta} - \beta^*) \right\}^2 + 2(\hat{\beta} - \beta^*)^T \sum_{t \in I} x_t x_t^T (\beta^* - \beta_t^*) \]

\[ \leq 2 \sum_{t \in I} \epsilon_t x_t^T (\hat{\beta} - \beta^*) + \lambda \sqrt{\max\{|I|, \log(n \vee p)|\|\beta^*\|_1 - \|\hat{\beta}\|_1\}}. \]  

(31)

We bound

\[ \left\| \sum_{t \in I} x_t x_t^T (\beta^* - \beta_t^*) \right\|_\infty. \]

For any \( k \in \{1, \ldots, p\} \), the \( k \)th entry of \( \sum_{t \in I} x_t x_t^T (\beta^* - \beta_t^*) \) satisfies that

\[ \mathbb{E} \left\{ \sum_{t \in I} \sum_{j=1}^p x_t(k) x_t(j) (\beta^*(j) - \beta_t^*(j)) \right\} = \sum_{t \in I} \sum_{j=1}^p \mathbb{E} \{ x_t(k) x_t(j) \} \{ \beta^*(j) - \beta_t^*(j) \} \]

\[ = \sum_{j=1}^p \mathbb{E} \{ x_1(k) x_1(j) \} \sum_{t \in I} \{ \beta^*(j) - \beta_t^*(j) \} = 0. \]

Note that \( x_t^T (\beta^* - \beta_t^*) \)'s are sub-Gaussian random variables with a common parameter \( 2C_\beta C_x \sqrt{d_0} \), and \( x_t \)'s are sub-Gaussian random vectors with parameter \( C_x \). Therefore due to sub-Exponential inequalities (e.g. Proposition 2.7.1 in Vershynin, 2018), it holds with probability at least of \( 1 - 2(n \vee p)^{-c} \) that,

\[ \left\| \sum_{t \in I} x_t x_t^T (\beta^* - \beta_t^*) \right\|_\infty \leq 2C_x C_\beta \sqrt{d_0} \max\{|I|, \log(n \vee p)|\log(n \vee p)|\} \]

\[ \leq \lambda \sqrt{\max\{|I|, \log(n \vee p)|\|\beta^*\|_1\}} / 4. \]  

(32)

On the event \( B_t \), combining (31) and (32) yields

\[ \sum_{t \in I} \left\{ x_t^T (\hat{\beta} - \beta^*) \right\}^2 + \lambda \sqrt{\max\{|I|, \log(n \vee p)|\|\beta^*\|_1\}} \|\hat{\beta}\|_1 \]

\[ \leq \lambda/2 \sqrt{\max\{|I|, \log(n \vee p)|\|\beta^* - \hat{\beta}\|_1 + \lambda \sqrt{\max\{|I|, \log(n \vee p)|\|\beta^*\|_1\}}. \]

(33)

The final claims follow from the same arguments as in Lemma 11.

\[ \square \]

A.3 All cases in Proposition 7

Lemma 14 (Case (i)). With the conditions and notation in Proposition 7, assume that \( I = (s, e) \in \widehat{\mathcal{P}} \) has one and only one true change point \( \eta \). Denote \( I_1 = (s, \eta] \), \( I_2 = (\eta, e] \) and \( \|\beta^*_{I_1} - \beta^*_{I_2}\|_2 = \kappa \). If, in addition, it holds that

\[ \sum_{t \in I} (y_t - x_t^T \hat{\beta}_{I_1}^T)^2 \leq \sum_{t \in I_1} (y_t - x_t^T \hat{\beta}_{I_1}^T)^2 + \sum_{t \in I_2} (y_t - x_t^T \hat{\beta}_{I_2}^T)^2 + \gamma, \]  

(33)
then with
\[ \lambda \geq \lambda_2 = C_\lambda \sigma_{\epsilon} \sqrt{d_0 \log(n \vee p)}, \]
where \( C_\lambda > 8C_\beta^2C_x/\sigma_{\epsilon} \), it holds with probability at least
\[ 1 - 2c_1(n \vee p)^{-288^2C_2d_0c_2/c_1^2} - 2(n \vee p)^{-c_3} \]
that, that
\[ \min\{|I_1|, |I_2|\} \leq C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right). \]

**Proof.** First we notice that with the choice of \( \lambda \), it holds that
\[ \lambda \geq \max\{\lambda_1, \lambda_2\}, \]
and therefore we can apply Lemmas 11, 12 and 13 when needed.

We prove by contradiction, assuming that
\[ \min\{|I_1|, |I_2|\} > C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right) > 288^2C_\beta^2d_0 \log(n \vee p)/c_x^2, \]
where the second inequality follows from the observation that \( \kappa^2 \leq 4d_0C_\beta^2 \). Therefore we also have
\[ \min\{|I_1|, |I_2|\} > \log(n \vee p). \]

It follows from Lemma 12 and (33) that, with probability at least
\[ 1 - 2c_1(n \vee p)^{-288^2C_2d_0c_2/c_1^2} - 2(n \vee p)^{-c_3} \]
that, that
\[ \sum_{t \in I_1} (y_t - x_t^\top \hat{\beta}_I) + \sum_{t \in I_2} (y_t - x_t^\top \hat{\beta}_I) \leq \sum_{t \in I_1} (y_t - x_t^\top \hat{\beta}_I) + \sum_{t \in I_2} (y_t - x_t^\top \hat{\beta}_I) \]
\[ \leq \sum_{t \in I_1} (y_t - x_t^\top \hat{\beta}_I) + \sum_{t \in I_2} (y_t - x_t^\top \hat{\beta}_I) + \gamma + 2C_3\lambda^2d_0. \]

Denoting \( \Delta_i = \hat{\beta}_I - \beta_I^*, i = 1, 2 \), (35) leads to that
\[ \sum_{t \in I_1} (y_t - x_t^\top \Delta_1) \leq 2 \sum_{t \in I_1} \varepsilon_t x_t^\top \Delta_1 + 2 \sum_{t \in I_2} \varepsilon_t x_t^\top \Delta_2 + \gamma + 2C_3\lambda^2d_0 \]
\[ \leq 2 \left\| \sum_{t \in I_1} \varepsilon_t x_t \right\|_\infty \left\| \Delta_1 \right\|_1 + 2 \left\| \sum_{t \in I_2} \varepsilon_t x_t \right\|_\infty \left\| \Delta_2 \right\|_1 + \gamma + 2C_3\lambda^2d_0 \]
\[ \leq 2 \left\| \sum_{t \in I_1} \varepsilon_t x_t \right\|_\infty (\left\| \Delta_1(S) \right\|_1 + \left\| \Delta_1(S^c) \right\|_1) + 2 \left\| \sum_{t \in I_2} \varepsilon_t x_t \right\|_\infty (\left\| \Delta_2(S) \right\|_1 + \left\| \Delta_2(S^c) \right\|_1) \]
\[ \leq 2 \left\| \sum_{t \in I_1} \varepsilon_t x_t \right\|_\infty (\sqrt{d_0} \left\| \Delta_1(S) \right\|_2 + \left\| \Delta_1(S^c) \right\|_1) + 2 \left\| \sum_{t \in I_2} \varepsilon_t x_t \right\|_\infty (\sqrt{d_0} \left\| \Delta_2(S) \right\|_2 + \left\| \Delta_2(S^c) \right\|_1) \]
\[ + \gamma + 2C_3\lambda^2d_0. \]
On the events $\mathcal{B}_{I_1} \cap \mathcal{B}_{I_2}$, it holds that

\[
(36) \leq \frac{\lambda}{2} (\sqrt{|I_1||d_0||\Delta_1(S)||_2} + \sqrt{|I_1||\Delta_1(S^c)||_1} + \sqrt{|I_2||d_0||\Delta_2(S)||_2} + \sqrt{|I_2||\Delta_2(S^c)||_1}) + \gamma + 2C_3\lambda^2d_0
\]

\[
\leq \frac{32\lambda^2d_0}{c_x^2} + \frac{c_c^2|I_1||\Delta_1||^2_2}{256} + \frac{c_c^2|I_2||\Delta_2||^2_2}{256} + \frac{\lambda (\sqrt{|I_1|} + \sqrt{|I_2|})}{2} ||\tilde{\beta}_1(S^c)||_1 + \gamma + 2C_3\lambda^2d_0
\]

\[
\leq \frac{32\lambda^2d_0}{c_x^2} + \frac{c_c^2|I_1||\Delta_1||^2_2}{256} + \frac{c_c^2|I_2||\Delta_2||^2_2}{256} + \gamma + 4C_3\lambda^2d_0,
\]

where the second inequality follows from $2ab \leq a^2 + b^2$, letting

\[
a = 4\lambda \sqrt{d_0/c_x} \quad \text{and} \quad b = c_x \sqrt{|I_2||\Delta_1||^2_2/16}, \quad j = 1, 2,
\]

and the last inequality follows from Lemma 13.

Note that

\[
||\Delta_1||_1 \leq ||\Delta_1(S)||_1 + ||\Delta_1(S^c)||_1 \leq \sqrt{d_0}||\Delta_1||_2 + \frac{C_5\lambda d_0}{\sqrt{|I_1|}},
\]

which combines with (34), on the event $\mathcal{E}_{I_1}$, leads to

\[
\sqrt{\sum_{i \in I_1} (x_i^T \Delta_1)^2} > \frac{c_c}{4} \sqrt{|I_1||\Delta_1||_2} - 9C_x \sqrt{\log(p)}||\Delta_1||_1 > \frac{c_c \sqrt{|I_1|}}{8} ||\Delta_1||_2 > \frac{9C_5C_x\lambda d_0 \sqrt{\log(p)}}{c_x^2} \frac{\sqrt{|I_1|}}{c_x^2}.
\]

Moreover, we have

\[
\sqrt{|I_1||\Delta_1||_2} + \sqrt{|I_2||\Delta_2||_2} \geq \sqrt{|I_1||\Delta_1||^2_2 + |I_2||\Delta_2||^2_2}
\]

\[
\geq \sqrt{\inf_{v \in \mathbb{R}^p} \{|I_1||\beta^* - v||^2 + |I_2||\beta^* - v||^2\}} = \kappa \sqrt{|I_1||I_2|/|I_1|} \geq \frac{\kappa}{\sqrt{2}} \min \{\sqrt{|I_1|}, \sqrt{|I_2|} \}, \quad \text{(38)}
\]

Therefore, on the event $\mathcal{E}_{I_1} \cap \mathcal{E}_{I_2} \cap \mathcal{B}_{I_1} \cap \mathcal{B}_{I_2}$, combining (36) and (37), we have that

\[
\sqrt{|I_1||\Delta_1||_2} + \sqrt{|I_2||\Delta_2||_2} \leq \frac{8}{c_x} \left( \sqrt{\sum_{i \in I_1} (x_i^T \Delta_1)^2} + \sqrt{\sum_{i \in I_2} (x_i^T \Delta_2)^2} \right) + \frac{8}{c_x} \left( \frac{9C_5C_x\lambda d_0 \sqrt{\log(p)}}{c_x^2} \frac{\sqrt{|I_1|}}{c_x^2} + \frac{9C_5C_x\lambda d_0 \sqrt{\log(p)}}{c_x^2} \frac{\sqrt{|I_2|}}{c_x^2} \right)
\]

\[
\leq \frac{8\sqrt{2}}{c_x} \sqrt{\frac{32\lambda^2d_0}{c_x^2} + \frac{c_c^2|I_1||\Delta_1||^2_2}{256} + \frac{c_c^2|I_2||\Delta_2||^2_2}{256} + \gamma + 4C_3\lambda^2d_0}
\]

\[
\leq \frac{64\lambda \sqrt{d_0}}{c_x^2} + \frac{\sqrt{2}\sqrt{|I_1||\Delta_1||_2}}{2} + \frac{\sqrt{2}\sqrt{|I_2||\Delta_2||_2}}{2} + \frac{8\sqrt{2}}{c_x} + \frac{16\sqrt{2C_3\lambda d_0}}{c_x} + \frac{C_5\lambda \sqrt{d_0}}{2c_x^2},
\]

26
which implies that

$$\frac{2 - \sqrt{2}}{2} \left( \sqrt{|I_1|\|\Delta_1\|_2} + \sqrt{|I_2|\|\Delta_2\|_2} \right) \leq \frac{128 + 32\sqrt{2}c_1\sqrt{C_0} + C_5}{2c_x^2} \sqrt{d_0} + \frac{8\sqrt{2}}{c_x}. \quad (39)$$

Combining (38) and (39) yields

$$\frac{2 - \sqrt{2}}{2\sqrt{2}} \kappa \sqrt{\min\{|I_1|, |I_2|\}} \leq \frac{128 + 32\sqrt{2}c_1\sqrt{C_0} + C_5}{2c_x^2} \sqrt{d_0} + \frac{8\sqrt{2}}{c_x},$$

therefore

$$\min\{|I_1|, |I_2|\} \leq C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right),$$

which is a contradiction with (34).

**Lemma 15 (Case (ii)).** For Model 3, under Assumption 3, with

$$\lambda \geq \lambda_2 = C_\lambda \sigma \sqrt{d_0 \log(n \lor p)},$$

where $C_\lambda > 8C_\beta C_2 / \sigma$, $I = (s, e]$ containing exactly two change points $\eta_1$ and $\eta_2$. Denote $I_1 = (s, \eta_1]$, $I_2 = (\eta_1, \eta_2]$, $I_3 = (\eta_2, e]$, $\|\beta^*_I - \beta^*_J\|_2 = \kappa_1$ and $\|\beta^*_I - \beta^*_J\|_2 = \kappa_2$. If in addition it holds that

$$\sum_{t \in I_1} (y_t - x_t^\top \hat{\beta}^*_I)^2 \leq \sum_{t \in I_1} (y_t - x_t^\top \hat{\beta}^*_I)^2 + \sum_{t \in I_2} (y_t - x_t^\top \hat{\beta}^*_I)^2 + \sum_{t \in I_3} (y_t - x_t^\top \hat{\beta}^*_I)^2 + 2\gamma,$$

then

$$\max\{|I_1|, |I_3|\} \leq C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right),$$

with probability at least $1 - 3c_1(n \lor p)^{-288^2 C_\epsilon^2 d_0 c_2 / c_x^2} - 2(n \lor p)^{-c_3}$.

**Proof.** First we notice that with the choice of $\lambda$, it holds that

$$\lambda \geq \max\{\lambda_1, \lambda_2\},$$

and therefore we can apply Lemmas 11, 12 and 13 when needed.

By symmetry, it suffices to show that

$$|I_1| \leq C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right).$$

We prove by contradiction, assuming that

$$|I_1| > C_\epsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right) > 288^2 C_\epsilon^2 d_0 \log(n \lor p)/c_x^2, \quad (40)$$

where the second inequality follows from the observation that $\kappa^2 \leq 4d_0 C_\beta^2$. Therefore we have

$$|I_1| > \log(n \lor p).$$

Denote $\Delta_i = \hat{\beta}^*_i - \beta^*_i$, $i = 1, 2, 3$. We then consider the following two cases.

**Case 1.** If

$$|I_3| > 288^2 C_\epsilon^2 d_0 \log(n \lor p)/c_x^2,$$
then $|I_3| > \log(n \lor p)$. It follows from Lemma 12 that the following holds with probability at least $1 - 3c_1(n \lor p)^{-288^2C_2d_0c_2/c_1^2} - 2(n \lor p)^{-c_3}$ that,

$$
\sum_{t \in I}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_1}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_2}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_3}(y_t - x_t^\top \hat{\beta}_t) + 2\gamma
\leq \sum_{t \in I_1}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_2}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_3}(y_t - x_t^\top \hat{\beta}_t) + 3C_3\lambda^2d_0 + 2\gamma,
$$

which implies that

$$
\sum_{i=1}^3 \sum_{t \in I_i} (x_t^\top \Delta_i)^2 \leq 2 \sum_{i=1}^3 \sum_{t \in I_i} \varepsilon_t x_t^\top \Delta_i + 3C_3\lambda^2d_0 + 2\gamma
\leq 2 \sum_{i=1}^3 \left\| \frac{1}{\sqrt{|I_i|}} \sum_{t \in I_i} \varepsilon_t x_t \right\| \|\sqrt{|I_i|} \Delta_i\|_1 + 3C_3\lambda^2d_0 + 2\gamma
\leq \lambda/2 \sum_{i=1}^3 \left( \sqrt{d_0} |I_i| \|\Delta_i(S)\|_2 + \sqrt{|I_i|} \|\Delta_i(S^c)\|_1 \right) + 3C_3\lambda^2d_0 + 2\gamma,
$$

where the last inequality follows from Lemma 10.

It follows from identical arguments in Lemma 14 that, with probability at least $1 - 3c_1(n \lor p)^{-288^2C_2d_0c_2/c_1^2} - 2(n \lor p)^{-c_3}$,

$$
\min\{|I_1|, |I_2|\} \leq C_\epsilon \left( \frac{\lambda^2d_0 + \gamma}{\kappa^2} \right).
$$

Since $|I_2| \geq \Delta$ by assumption, it follows from Assumption 3(d) that

$$
|I_1| \leq C_\epsilon \left( \frac{\lambda^2d_0 + \gamma}{\kappa^2} \right),
$$

which contradicts (40).

**Case 2.** If 

$$
|I_3| \leq 288^2C_2d_0 \log(n \lor p)/c_1^2,
$$

then it follows from Lemma 12 that the following holds with probability at least $1 - 2c_1(n \lor p)^{-288^2C_2d_0c_2/c_1^2} - 2(n \lor p)^{-c_3}$ that,

$$
\sum_{t \in I}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_1}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_2}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_3}(y_t - x_t^\top \hat{\beta}_t) + 2\gamma
\leq \sum_{t \in I_1}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_2}(y_t - x_t^\top \hat{\beta}_t) + \sum_{t \in I_3}(y_t - x_t^\top \hat{\beta}_t) + 2C_3\lambda^2d_0 + C_4\lambda \sqrt{\log(p)d_0^{3/2}} + 2\gamma,
$$

which implies that

$$
\sum_{i=1}^3 \sum_{t \in I_i} (x_t^\top \Delta_i)^2 \leq 2 \sum_{i=1}^3 \sum_{t \in I_i} \varepsilon_t x_t^\top \Delta_i + 2C_3\lambda^2d_0 + C_4\lambda \sqrt{\log(p)d_0^{3/2}} + 2\gamma
$$
\[
\leq 2 \sum_{i=1}^{2} \left\| \frac{1}{\sqrt{|I_i|}} \sum_{t \in I_i} \varepsilon_t x_t \right\|_\infty \leq 2C_3 \lambda^2 d_0 + C_4 \lambda \sqrt{\log(p)} d_0^{3/2} + 2\gamma + \sum_{t \in I_3} (x_t^T \Delta_3)^2 + \sum_{t \in I_3} \varepsilon_t^2.
\]

The rest follows from the same arguments as in Case 1.

Lemma 16 (Case (iii) in Proposition 7). For Model 3, under Assumption 3, if there exists no true change point in \( I = (s, e) \), with

\[
\lambda \geq \lambda_2 = C_\lambda \sigma_e \sqrt{d_0 \log(n \vee p)},
\]

where \( C_\lambda > \max \{8C_1 C_x, 8C_3 C_x / \sigma_e \} \), and \( \gamma = \gamma(C_\lambda^2 d_0 \log(n \vee p)) \), where \( C_\gamma > \max \{3C_3 / c_x^2, 3C_4 / c_x \} \), it holds with probability at least \( 1 - 3c_1(n \vee p)^{-288^2C_0^2 d_0 c_2 / c_x^2} - 2(n \vee p)^{-c_3} \) that

\[
\sum_{t \in I} (y_t - x_t^T \hat{\beta}^t)^2 < \min_{b = s + 1, \ldots, e - 1} \left\{ \sum_{t \in (s, b)} (y_t - x_t^T \hat{\beta}^t)^2 + \sum_{t \in [b, e]} (y_t - x_t^T \hat{\beta}^t_{(b, e)})^2 \right\} + \gamma.
\]

Proof. First we notice that with the choice of \( \lambda \), it holds that \( \lambda > \lambda_1 \), therefore we can apply Lemma 12 when needed.

For any \( b = s + 1, \ldots, e - 1 \), let \( I_1 = (s, b) \) and \( I_2 = (b, e) \). It follows from Lemma 12 that with probability at least \( 1 - 3c_1(n \vee p)^{-288^2C_0^2 d_0 c_2 / c_x^2} - 2(n \vee p)^{-c_3} \),

\[
\max_{J \in \{I_1, I_2, I\}} \left\| \sum_{t \in J} (y_t - x_t^T \hat{\beta}_J)^2 - \sum_{t \in J} (y_t - x_t^T \hat{\beta}_J)^2 \right\| \leq \max \left\{ C_3 \lambda^2 d_0, C_4 \lambda \sqrt{\log(n \vee p)} d_0^{3/2} \right\} < \gamma / 3.
\]

Since \( \beta_{J^*}^t = \beta_{I_1}^t = \beta_{I_2}^t \), the final claim holds automatically.

Lemma 17 (Case (iv) in Proposition 7). For Model 3, under Assumption 3, if \( I = (s, e) \) contains \( J \) true change points \( \{\eta_k\}_{k=1}^J \), where \( |J| \geq 3 \), if

\[
\lambda \geq \lambda_2 = C_\lambda \sigma_e \sqrt{d_0 \log(n \vee p)},
\]

where \( C_\lambda > 8C_\beta C_x / \sigma_e \), then with probability at least \( 1 - nc_1(n \vee p)^{-288^2C_0^2 d_0 c_2 / c_x^2} - 2(n \vee p)^{-c_3} \),

\[
\sum_{t \in I} (y_t - x_t^T \hat{\beta}_I^t)^2 > \sum_{j=1}^{J+1} \sum_{t \in I_j} (y_t - x_t^T \hat{\beta}_I^t)^2 + J\gamma,
\]

where \( I_1 = (s, \eta_1) \), \( I_j = (\eta_j, \eta_{j+1}) \) for any \( 2 \leq j \leq J \) and \( I_{J+1} = (\eta_J, e) \).
Proof. First we notice that with the choice of $\lambda$, it holds that
\[ \lambda \geq \max\{\lambda_1, \lambda_2\}, \]
and therefore we can apply Lemmas 11, 12 and 13 when needed.

We prove the claim by contradiction, assuming that
\[ \sum_{t \in I} (y_t - x_t^T \hat{\beta}_t)^2 \leq \sum_{j=1}^{J+1} \sum_{i \in I_j} (y_t - x_t^T \hat{\beta}_{tj})^2 + J\gamma. \]

Let $\Delta_i = \hat{\beta}_t - \beta_{tj}, i = 1, \ldots, J + 1$. It then follows from Lemma 12 that with probability at least $1 - nc_1(n \lor p)^{-2882C_2^2d_0/c_x^2} - 2(n \lor p)^{-c_3},$
\[ \sum_{t \in I} (y_t - x_t^T \hat{\beta}_t)^2 \leq \sum_{j=1}^{J+1} \sum_{i \in I_j} (y_t - x_t^T \hat{\beta}_{tj})^2 + J\gamma \]
\[ \leq \sum_{j=1}^{J+1} \sum_{i \in I_j} (y_t - x_t^T \hat{\beta}_{tj})^2 + J\gamma + (J + 1)C_4 \sigma_x^2 d_0^2 \log(n \lor p), \]
which implies that
\[ \sum_{j=1}^{J+1} \sum_{i \in I_j} (x_t^T \Delta_i)^2 \leq 2 \sum_{j=1}^{J+1} \sum_{i \in I_j} \epsilon_t x_t^T \Delta_i + J\gamma + (J + 1)C_4 \sigma_x^2 d_0^2 \log(n \lor p). \quad (41) \]

Step 1. For any $j \in \{2, \ldots, J\}$, it follows from Assumption 3 that
\[ |I_j| \geq \Delta \geq 2882C_2^2d_0 \log(n \lor p)/c_x^2. \quad (42) \]

Due to Lemma 10, on the event $\mathcal{B}_{[0, n]}$, it holds that
\[ \sum_{t \in I_j} \epsilon_t x_t^T \Delta_j \leq \frac{1}{\sqrt{|I_j|}} \sum_{t \in I_j} \epsilon_t x_t^T \Delta_j \leq \lambda/4 \left( \sqrt{|I_j|} \Delta_j \|S\|_2 + \sqrt{|I_j|} \|\Delta_j(S)\|_1 \right) \]
\[ \leq \frac{4\lambda^2 d_0}{c_x^2} + \frac{c_x^2 |I_j|}{256} \|\Delta_j\|_2^2 + \lambda/4 \sqrt{|I_j|} \|\beta_t(S)\|_1 \]
\[ = \frac{4\lambda^2 d_0}{c_x^2} + \frac{c_x^2 |I_j|}{256} \|\Delta_j\|_2^2 + \lambda/4 \sqrt{|I_j|} \|\beta_t - \beta_t(S)\|_1 \]
\[ \leq \frac{4\lambda^2 d_0}{c_x^2} + \frac{c_x^2 |I_j|}{256} \|\Delta_j\|_2^2 + \lambda/4 \sqrt{|I_j|} \|\beta_t - \beta_t(S)\|_1 \]
\[ \leq \frac{4\lambda^2 d_0}{c_x^2} + \frac{c_x^2 |I_j|}{256} \|\Delta_j\|_2^2 + C_5/4\lambda^2 d_0, \quad (43) \]
where the third inequality follows from $2ab \leq a^2 + b^2$, letting
\[ a = 2\lambda \sqrt{d_0/c_x} \quad \text{and} \quad b = c_x \sqrt{|I_j|} \|\Delta_j\|_2/16, \]
and the last inequality follows from Lemma 13. In addition, on the event of $\mathcal{E}_{I_j}$, due to Lemma 9, it holds that

$$\sqrt{\sum_{t \in I_j} (x_t^\top \Delta_j)^2} \geq \frac{c_x \sqrt{|I_j|}}{4} \|\Delta_j\|_2 - 9C_x \sqrt{\log(p)} \|\Delta_j\|_1$$

$$\geq \frac{c_x \sqrt{|I_j|}}{4} \|\Delta_j\|_2 - 9C_x \sqrt{d_0 \log(p)} \|\Delta_j\|_2 - 9C_x \sqrt{\log(p)} \|\Delta_j(S^c)\|_1$$

$$\geq \frac{c_x \sqrt{|I_j|}}{8} \|\Delta_j\|_2 - 9C_x \sqrt{\log(p)} \|\Delta_j(S^c)\|_1 \geq \frac{c_x \sqrt{|I_j|}}{8} \|\Delta_j\|_2 - \frac{9C \lambda d_0 \sqrt{\log(p)}}{\sqrt{|I|}},$$

(44)

where the third inequality follows from (42) and the last follows from Lemma 13.

**Step 2.** We then discuss the intervals $I_1$ and $I_{J+1}$. These two will be treated in the same way, and therefore for $L \in \{I_1, I_{J+1}\}$ and $l \in \{1, J + 1\}$, we have the following arguments. If $|L| \geq 288^2 C_\varnothing^2 d_0 \log(n \lor p)/c_2^2$, then due to the same arguments in **Step 1**, (43) and (44) hold. If instead, $|L| < 288^2 C_\varnothing^2 d_0 \log(n \lor p)/c_2^2$ holds, then

$$\sum_{t \in L} \varepsilon_t x_t^\top \Delta_t \leq 2^{-1} \sum_{t \in L} (x_t^\top \Delta_t)^2 + 4 \sum_{t \in L} \varepsilon_t^2.$$ 

Therefore, it follows from (41) that

$$\sum_{j=2}^J |I_j| \|\Delta_j\|_2^2 \leq JC \max \left\{ \lambda^2 d_0, \lambda \sqrt{\log(n \lor p) d_0^{3/2}} \right\} + J \gamma.$$

**Step 3.** Since for any $j \in \{2, \ldots, J - 1\}$, it holds that

$$|I_j| \|\Delta_j\|_2^2 + |I_{j+1}| \|\Delta_{j+1}\|_2^2 \geq \inf_{v \in \mathbb{R}^k} \left\{ |I_j| \|\beta_j^* - v\|_2^2 + |I_{j+1}| \|\beta_{j+1}^* - v\|_2^2 \right\}$$

$$\geq \frac{|I_j| |I_{j+1}| \kappa^2}{|I_j| + |I_{j+1}|} \geq \min\{|I_j|, |I_{j+1}|\} \kappa^2/2.$$

It then follows from the same arguments in Lemma 14 that

$$\min_{j=2, \ldots, J-1} |I_j| \leq C_\varepsilon \left( \frac{\lambda^2 d_0 + \gamma}{\kappa^2} \right),$$

which is a contradiction to (42).

### A.4 Proof of Proposition 8

**Lemma 18.** Under the assumptions and notation in Proposition 7, suppose there exists no true change point in the interval $I$. For any interval $J \supset I$, with

$$\lambda \geq \lambda_2 = C_\lambda \sigma \sqrt{d_0 \log(n \lor p)},$$

where $C_\lambda > \max\{8C_1 C_x, 8C_\beta C_x/\sigma \varepsilon\}$, it holds that with probability at least $1 - c_1 (n \lor p)^{-288^2 C_\varnothing^2 d_0 c_2/c_2^2 - 2(n \lor p)^{-c_3}}$

$$\sum_{t \in I} (y_t - x_t^\top \beta_t^*)^2 - \sum_{t \in I} (y_t - x_t^\top \hat{\beta}_t^*)^2 \leq C_6 \lambda^2 d_0.$$
Proof. Case 1. If
\[ |I| \geq 288^2 C_x^2 d_0 \log(n \lor p) / c_x^2, \quad (45) \]
then letting \( \Delta_I = \beta_I^* - \tilde{\beta}_I \), on the event \( \mathcal{E}_I \), we have
\[
\sqrt{\sum_{t \in I} (x_t^\top \Delta_I)^2} \geq \frac{c_x \sqrt{|I|}}{4} \|\Delta_I\|_2 - 9C_x \sqrt{\log(p)} \|\Delta_I\|_1
\]
\[
= \frac{c_x \sqrt{|I|}}{4} \|\Delta_I\|_2 - 9C_x \sqrt{\log(p)} \|\Delta_I\|_1 - 9C_x \sqrt{\log(p)} \|\Delta_I(S^c)\|_1
\]
\[
\geq \frac{c_x \sqrt{|I|}}{8} \|\Delta_I\|_2 - 9C_x \sqrt{d_0 \log(p)} \|\beta_I^* \|_1
\]
\[
geq \frac{c_x \sqrt{|I|}}{8} \|\Delta_I\|_2 - 9C_x \sqrt{d_0 \log(p)} \|\beta_I^* \|_1 \geq \frac{c_x \sqrt{|I|}}{8} \|\Delta_I\|_2 - 9C_5 C_x d_0 \lambda \log^{1/2}(p), \quad (46)\]
where the last inequality follows from Lemma 13. We then have on the event \( \mathcal{B}_I \),
\[
\sum_{t \in I} (y_t - x_t^\top \beta_I^*)^2 - \sum_{t \in I} (y_t - x_t^\top \beta_I^*)^2 = 2 \sum_{t \in I} \epsilon_t x_t^\top \Delta_I - \sum_{t \in I} (x_t^\top \Delta_I)^2
\]
\[
\leq 2 \left\| \sum_{t \in I} x_t \epsilon_t \right\|_{\infty} \left( \sqrt{d_0} \|\Delta_I\|_2 + \|\beta_I^* \|_1 \right)
\]
\[
- \frac{c_x^2 |I|}{64} \|\Delta_I\|_2^2 - \frac{81 C_x^2 C_x^2 \lambda^2 d_0 \log(p)}{c_x^2 |I|} + \frac{9C_5 C_x d_0 \lambda \log^{1/2}(p)}{4} \|\Delta_I\|_2
\]
\[
\leq \frac{\lambda}{2} \sqrt{d_0 \log(p)} \|\Delta_I\|_2 + \frac{\lambda^2 d_0 C_5}{2 c_x^2 \sqrt{|I|}} - \frac{c_x^2 |I|}{64} \|\Delta_I\|_2^2 + \frac{9C_5 C_x d_0 \lambda \log^{1/2}(p)}{4} \|\Delta_I\|_2
\]
\[
\leq \frac{\lambda}{2} \sqrt{d_0 \log(p)} \|\Delta_I\|_2 + \frac{\lambda^2 \sqrt{d_0 C_5}}{576c_x \sqrt{\log(n \lor p) C_x}} - \frac{36^2 C_x^2 d_0 \log(n \lor p)}{2 c_x^2 \sqrt{|I|}} - \frac{36^2 C_x^2 d_0 \log(n \lor p)}{c_x^2 |I|} \|\Delta_I\|_2
\]
\[
\leq \frac{\lambda^2}{16 C_x^2} + \frac{d_0 C_x^2 \|\Delta_I\|_2^2}{2} + \frac{\lambda^2 \sqrt{d_0 C_5}}{576c_x \sqrt{\log(n \lor p) C_x}} - \frac{36^2 C_x^2 d_0 \log(n \lor p)}{2 c_x^2 \sqrt{|I|}} + \frac{36^2 C_x^2 d_0 \log(n \lor p)}{c_x^2 |I|} \|\Delta_I\|_2
\]
\[
\leq \frac{181 C_x^2 C_x^2 \lambda^2}{64}
\]
\[
\leq C_6 \lambda^2 d_0.
\]
where the first inequality follows from (46), the second inequality follows from event \( \mathcal{B}_I \) and Lemma 13, the third follows from the (45), the fourth follows from \( 2ab \leq a^2 + b^2 \), first letting
\[
a = \lambda/(4C_x) \quad \text{and} \quad b = \sqrt{d_0 C_x} \|\Delta_I\|_2,
\]
then letting
\[
a = C_x \sqrt{d_0 \log(p)} \|\Delta_I\|_2 \quad \text{and} \quad b = 9C_5 \sqrt{d_0 \lambda}/8,
\]
and the last inequality follows from Lemma 13.

Case 2. If \( |I| \leq 288^2 C_x^2 d_0 \log(n \lor p) / c_x^2, \) then with probability at least \( 1 - 2(n \lor p)^{-c} \),
\[
\sum_{t \in I} (y_t - x_t^\top \beta_I^*)^2 - \sum_{t \in I} (y_t - x_t^\top \beta_I^*)^2 = 2 \sum_{t \in I} \epsilon_t x_t^\top (\tilde{\beta}_I^* - \beta_I^*) - \sum_{t \in I} (x_t^\top (\beta_I^* - \tilde{\beta}_I^*))^2
\]
\[ \leq \sum_{i \in I} \varepsilon_i^2 \leq \max\{ \sqrt{|I| \log(n \lor p)}, \log(n \lor p) \} \leq C_6 \lambda^2 d_0. \]

\[ \Box \]

**Proof of Proposition 8.** Denote \( S_n^* = \sum_{t=1}^n (y_t - x_t^\top \beta_t)^2 \). Given any collection \( \{t_1, \ldots, t_m\} \), where \( t_1 < \cdots < t_m \), and \( t_0 = 0, t_{m+1} = n \), let

\[ S_n(t_1, \ldots, t_m) = \sum_{k=1}^m \sum_{t=t_k+1}^{t_{k+1}} (y_t - x_t^\top \hat{\beta}_{(t_k, t_{k+1})})^2. \]

(47)

For any collection of time points, when defining (47), the time points are sorted in an increasing order.

Let \( \{\hat{\eta}_k\}_{k=1}^{\hat{K}} \) denote the change points induced by \( \hat{P} \). If one can justify that

\[ S_n^* + K \gamma \geq S_n(\eta_1, \ldots, \eta_K) + K \gamma - C_3(K + 1)d_0 \lambda^2 \]

\[ \geq S_n(\tilde{\eta}_1, \ldots, \tilde{\eta}_{\hat{K}}) + \hat{K} \gamma - C_3(K + 1)d_0 \lambda^2 \]

\[ \geq S_n(\tilde{\eta}_1, \ldots, \tilde{\eta}_{\hat{K}}, \eta_1, \ldots, \eta_K) + \hat{K} \gamma - 2C(K + 1)d_0 \lambda^2 - C_3(K + 1)d_0 \lambda^2 \]

(50)

and that

\[ S_n^* - S_n(\tilde{\eta}_1, \ldots, \tilde{\eta}_{\hat{K}}, \eta_1, \ldots, \eta_K) \leq C(K + \hat{K} + 2) \lambda^2 d_0, \]

(51)

then it must hold that |\( \hat{P} \)| = \( \hat{K} \), as otherwise if \( \hat{K} \geq K + 1 \), then

\[ C(K + \hat{K} + 2) \lambda^2 d_0 \geq S_n^* - S_n(\tilde{\eta}_1, \ldots, \tilde{\eta}_{\hat{K}}, \eta_1, \ldots, \eta_K) \]

\[ \geq -3C(K + 1)\lambda^2 d_0 + (\hat{K} - K) \gamma \geq C_\gamma(K + 1) \lambda^2 d_0. \]

Therefore due to the assumption that \( |\hat{P}| = \hat{K} \leq 3K \), it holds that

\[ C(5K + 3) \lambda^2 d_0 \geq (\hat{K} - K) \gamma \geq \gamma, \]

(52)

Note that (52) contradicts the choice of \( \gamma \).

Note that (48) is implied by

\[ |S_n^* - S_n(\eta_1, \ldots, \eta_K)| \leq C_3(K + 1)d_0 \lambda^2, \]

(53)

which is immediate consequence of Lemma 12. Since \( \{\hat{\eta}_k\}_{k=1}^{\hat{K}} \) are the change points induced by \( \hat{P} \), (49) holds because \( \hat{P} \) is a minimiser.

For every \( \tilde{I} = (s, e) \in \hat{P} \) denote

\[ \tilde{I} = (s, \eta_{p+1}] \cup \ldots \cup (\eta_{p+q}, e] = J_1 \cup \ldots \cup J_{q+1}, \]

where \( \{\eta_{p+1}\}_{l=1}^{q+1} = \tilde{I} \cap \{\eta_k\}_{k=1}^{\hat{K}} \). Then (50) is an immediate consequence of the following inequality

\[ \sum_{t \in \tilde{I}} (y_t - x_t^\top \hat{\beta}_{\tilde{I}}^\lambda)^2 \geq \sum_{l=1}^{q+1} \sum_{t \in J_l} (y_t - x_t^\top \hat{\beta}_{J_l}^\lambda)^2 - C(q + 1) \lambda^2 d_0. \]

(54)
By Lemma 12, it holds that
\[
\sum_{l=1}^{q+1} \sum_{t \in J_t} (y_t - x_t^\top \beta_l^\star)^2 \leq \sum_{l=1}^{q+1} \sum_{t \in J_t} (y_t - x_t^\top \beta_l^\star)^2 + (q + 1) \max \left\{ C_3 d_0 \lambda^2, C_4 \lambda \log(n \vee p) d_0^{3/2} \right\}
\]
\[
= \sum_{t \in \tilde{I}} (y_t - x_t^\top \beta_l^\star)^2 + (q + 1) \max \left\{ C_3 d_0 \lambda^2, C_4 \lambda \log(n \vee p) d_0^{3/2} \right\}. \tag{55}
\]
Then for each \( l \in \{1, \ldots, q + 1\} \),
\[
\sum_{t \in J_t} (y_t - x_t^\top \beta_l^\star)^2 \geq \sum_{t \in J_t} (y_t - x_t^\top \beta_l^\star)^2 - C_6 \lambda^2 d_0,
\]
where the inequality follows from Lemma 18. Therefore the above inequality implies that
\[
\sum_{t \in \tilde{I}} (y_t - x_t^\top \beta_l^\star)^2 \geq \sum_{t \in \tilde{I}} (y_t - x_t^\top \beta_l^\star)^2 - C_6 (q + 1) \lambda^2 d_0. \tag{56}
\]
Note that (55) and (56) implies (54).

Finally, to show (51), observe that from (53), it suffices to show that
\[
S_n(\eta_1, \ldots, \eta_K) - S_n(\hat{\eta}_1, \ldots, \hat{\eta}_K, \eta_1, \ldots, \eta_K) \leq C(K + \hat{K}) \lambda^2;
\]
the analysis of which follows from a similar but simpler argument as above. \( \square \)

## B Proof of Corollary 6

**Lemma 19.** Let \( S \) be any linear subspace in \( \mathbb{R}^n \) and \( \mathcal{N}_{1/4} \) be a 1/4-net of \( S \cap B(0, 1) \), where \( B(0, 1) \) is the unit ball in \( \mathbb{R}^n \). For any \( u \in \mathbb{R}^n \), it holds that
\[
\sup_{v \in S \cap B(0, 1)} \langle v, u \rangle \leq 2 \sup_{v \in \mathcal{N}_{1/4}} \langle v, u \rangle,
\]
where \( \langle \cdot, \cdot \rangle \) denotes the inner product in \( \mathbb{R}^n \).

**Proof.** Due to the definition of \( \mathcal{N}_{1/4} \), it holds that for any \( v \in S \cap B(0, 1) \), there exists a \( v_k \in \mathcal{N}_{1/4} \), such that \( \|v - v_k\|_2 < 1/4 \). Therefore,
\[
\langle v, u \rangle = \langle v - v_k + v_k, u \rangle = \langle x_k, u \rangle + \langle v_k, u \rangle \leq \frac{1}{4} \langle v, u \rangle + \frac{1}{4} \langle v^\perp, u \rangle + \langle v_k, u \rangle,
\]
where the inequality follows from \( x_k = v - v_k = \langle x_k, v \rangle v + \langle x_k, v^\perp \rangle v^\perp \). Then we have
\[
\frac{3}{4} \langle v, u \rangle \leq \frac{1}{4} \langle v^\perp, u \rangle + \langle v_k, u \rangle.
\]
It follows from the same argument that
\[
\frac{3}{4} \langle v^\perp, u \rangle \leq \frac{1}{4} \langle v, u \rangle + \langle v_l, u \rangle,
\]
where \( v_l \in \mathcal{N}_{1/4} \) satisfies \( \|v^\perp - v_l\|_2 < 1/4 \). Combining the previous two equation displays yields
\[
\langle v, u \rangle \leq 2 \sup_{v \in \mathcal{N}_{1/4}} \langle v, u \rangle,
\]
and the final claims holds. \( \square \)
Lemma 20 is an adaptation of Lemma 3 in Wang et al. (2019).

**Lemma 20.** For data generated from Model 3, for any interval \( I = (s, e] \subset \{1, \ldots, n\} \), it holds that for any \( \delta > 0 \), \( i \in \{1, \ldots, p\} \),

\[
\mathbb{P} \left\{ \sup_{v \in \mathbb{R}^{(e-s)}, \|v\|_2 = 1} \left| \sum_{t=s+1}^{e} v_t \xi_t x(i) \right| > \delta \right\} \leq C(e-s-1)^m g^{m+1} \exp \left\{ -c \min \left\{ \frac{\delta^2}{4C_x^2}, \frac{\delta}{2C_x \|v\|_\infty} \right\} \right\}.
\]

**Proof.** For any \( v \in \mathbb{R}^{(e-s)} \) satisfying \( \sum_{t=1}^{e-s-1} 1 \{ v_i \neq v_{i+1} \} = m \), it is determined by a vector in \( \mathbb{R}^{m+1} \) and a choice of \( m \) out of \( (e-s-1) \) points. Therefore we have,

\[
\mathbb{P} \left\{ \sup_{v \in \mathbb{R}^{(e-s)}, \|v\|_2 = 1} \left| \sum_{t=s+1}^{e} v_t \xi_t x(i) \right| > \delta \right\} \leq \left( \frac{(e-s-1)}{m} \right)^{g^{m+1}} \sup_{v \in N_{1/4}} \mathbb{P} \left\{ \left| \sum_{t=s+1}^{e} v_t \xi_t x(i) \right| > \delta/2 \right\}
\]

\[
\leq \left( \frac{(e-s-1)}{m} \right)^{g^{m+1}} C \exp \left\{ -c \min \left\{ \frac{\delta^2}{4C_x^2}, \frac{\delta}{2C_x \|v\|_\infty} \right\} \right\}
\]

\[
\leq C(e-s-1)^m g^{m+1} \exp \left\{ -c \min \left\{ \frac{\delta^2}{4C_x^2}, \frac{\delta}{2C_x \|v\|_\infty} \right\} \right\}.
\]

**Proof of Corollary 6.** For each \( k \in \{1, \ldots, K\} \), let

\[
\hat{\beta}_t = \begin{cases} 
\tilde{\beta}_1, & t \in \{s_k + 1, \ldots, \hat{\eta}_k\}, \\
\tilde{\beta}_2, & t \in \{\hat{\eta}_k + 1, \ldots, e_k\}.
\end{cases}
\]

Without loss of generality, we assume that \( s_k < \eta_k < \hat{\eta}_k < e_k \). We proceed the proof discussing two cases.

**Case (i).** If

\[
\hat{\eta}_k - \eta_k < \max\{288^2 C_x^2 d_0 \log(n \lor p)/\epsilon^2, C_\epsilon \log(n \lor p)/\kappa^2\},
\]

then the result holds.

**Case (ii).** If

\[
\hat{\eta}_k - \eta_k \geq \max\{288^2 C_x^2 d_0 \log(n \lor p)/\epsilon^2, C_\epsilon \log(n \lor p)/\kappa^2\},
\]

then we first to prove that with probability at least \( 1 - C(n \lor p)^{-c} \),

\[
\sum_{t=s_k+1}^{e_k} \| \hat{\beta}_t - \hat{\beta}_t^* \|_2^2 \leq C_1 d_0 \kappa^2 = \delta.
\]

35
Due to (8), it holds that

$$\sum_{t=s_k+1}^{e_k} \|y_t - x_t^\top \hat{\beta}_t\|_2^2 + \zeta \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\hat{\beta}_t)_i^2} \leq \sum_{t=s_k+1}^{e_k} \|y_t - x_t^\top \beta_t^*\|_2^2 + \zeta \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\beta_t^*)_i^2}. \quad (58)$$

Let $\Delta_t = \hat{\beta}_t - \beta_t^*$. It holds that

$$\sum_{t=s_k+1}^{e_k-1} \mathbb{1}(\Delta_t \neq \Delta_{t+1}) = 2.$$

Eq. (58) implies that

$$\sum_{t=s_k+1}^{e_k} \|\Delta_t^\top x_t\|_2^2 + \zeta \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\hat{\beta}_t)_i^2} \leq 2 \sum_{t=s_k+1}^{e_k} (y_t - x_t^\top \beta_t^*) \Delta_t^\top x_t + \zeta \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\beta_t^*)_i^2}. \quad (59)$$

Note that

$$\sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\beta_t)_i^2} - \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\hat{\beta}_t)_i^2} = \sum_{i \in S} \sqrt{\sum_{t=s_k+1}^{e_k} (\beta_t)_i^2} - \sum_{i \in S} \sqrt{\sum_{t=s_k+1}^{e_k} (\hat{\beta}_t)_i^2} \leq \sum_{i \in S} \sqrt{\sum_{t=s_k+1}^{e_k} (\Delta_t)_i^2}.$$

(60)

We then examine the cross term, with probability at least $1 - C(n \lor p)^{-c}$, which satisfies the following

$$\left| \sum_{t=s_k+1}^{e_k} (y_t - x_t^\top \beta_t^*) \Delta_t^\top x_t \right| = \sum_{t=s_k+1}^{e_k} \varepsilon_t \Delta_t^\top x_t = \sum_{i=1}^p \left\{ \frac{\sum_{l=s_k+1}^{e_k} \varepsilon_l \Delta_l(i) x_l(i)}{\sqrt{\sum_{l=s_k+1}^{e_k} (\Delta_l(i))^2}} \right\} \leq \sup_{i=1, \ldots, p} \left| \sum_{l=s_k+1}^{e_k} \varepsilon_l \Delta_l(i) x_l(i) \right| \leq \frac{\zeta}{4} \sum_{i=1}^p \sqrt{\sum_{t=s_k+1}^{e_k} (\Delta_t(i))^2}, \quad (61)$$

where the second inequality follows from Lemma 20 and (57).

Combining (58), (59), (60) and (61) yields

$$\sum_{t=s_k+1}^{e_k} \|\Delta_t^\top x_t\|_2^2 \leq \frac{3\zeta}{2} \sum_{i \in S} \sqrt{\sum_{t=s_k+1}^{e_k} (\Delta_t)_i^2}. \quad (62)$$

Now we are to explore the restricted eigenvalue inequality. Let

$$I_1 = (s_k, \eta_k], \quad I_2 = (\eta_k, \tilde{\eta}_k], \quad I_3 = (\tilde{\eta}_k, e_k].$$

We have that with probability at least $1 - C(n \lor p)^{-c}$, on the event $\cap_{i=1,3} E_{I_i}$,

$$\sum_{t=s_k+1}^{e_k} \|\Delta_t^\top x_t\|_2^2 \leq \sum_{i=1}^p \sum_{t \in I_i} \|\Delta_{I_i}^\top x_t\|_2^2 \geq \sum_{i=1,3} \sum_{t \in I_i} \|\Delta_{I_i}^\top x_t\|_2^2.$$
\[
\sum_{i=1,3} \left( \frac{c_x \sqrt{|I_i|}}{4} ||\Delta_{I_i}||_2 - 9C_x \sqrt{\log(p)} ||\Delta_{I_i}||_1 \right)^2
\geq \sum_{i=1,3} \left( \frac{c_x \sqrt{|I_i|}}{8} ||\Delta_{I_i}||_2 - 9C_x \sqrt{\log(p)} ||\Delta_{I_i}(S^c)||_1 \right)^2,
\]
where the last inequality follows from (19) and Assumption 3, that
\[
\min\{|I_1|, |I_3|\} > (1/3)\Delta > 288^2C_x^2d_0 \log(n \vee p)/\epsilon_x^2.
\]
Since \(|I_2| > 288^2C_x^2d_0 \log(n \vee p)/\epsilon_x^2\), we have
\[
\sqrt{\sum_{t \in I_2} ||\Delta_{I_2} x_t||_2^2} \geq \frac{c_x \sqrt{|I_2|}}{8} ||\Delta_{I_2}||_2 - 9C_x \sqrt{\log(p)} ||\Delta_{I_2}(S^c)||_1.
\]
Note that
\[
\sqrt{\sum_{i=1}^{3} \left( \sum_{j \in S^c} |\Delta_{I_i}(j)| \right)^2} \leq \sqrt{\sum_{i=1}^{3} \left( \frac{|I_i|}{I_{0}} \sum_{j \in S^c} |\Delta_{I_i}(j)| \right)^2}
\leq \sum_{i=1}^{3} \frac{I_{0}^{-1/2}}{2} \sqrt{\sum_{j=1}^{e} \frac{I_{0}^{-1/2}}{2} \sqrt{\sum_{t=s_k+1}^{e} (\Delta_t(i))^2} \leq \sum_{j=1}^{e} \frac{I_{0}^{-1/2}}{2} \sqrt{\sum_{t=s_k+1}^{e} (\Delta_t(i))^2}
\leq I_{0}^{-1/2} 3 d_0 \sum_{j=1}^{e} \frac{I_{0}^{-1/2}}{2} \sqrt{\sum_{t=s_k+1}^{e} (\Delta_t(i))^2} \leq \frac{c_x}{96C_x \sqrt{\log(n \vee p)}} \sqrt{\sum_{t=s_k+1}^{e} ||\Delta_t||_2^2}.
\]
Therefore,
\[
\frac{c_x}{8} \sqrt{\sum_{t=s_k+1}^{e} ||\Delta_t||_2^2} - \frac{3c_x}{32C_x \sqrt{\log(n \vee p)}} \sqrt{\sum_{t=s_k+1}^{e} ||\Delta_t||_2^2}
\leq \frac{3}{2} \left( \frac{c_x \sqrt{|I_i|}}{8} ||\Delta_{I_i}||_2 - \frac{3c_x}{32C_x \sqrt{\log(n \vee p)}} \sum_{t=s_k+1}^{e} ||\Delta_t||_2^2 \leq \sqrt{3} \sum_{t=s_k+1}^{e} ||\Delta_t||_2^2 \right)^{1/4}
\leq \frac{3\sqrt{2}}{\sqrt{2}} d_0^{1/4} \left( \sum_{t=s_k+1}^{e} ||\Delta_t||_2^2 \right)^{1/4} \leq \frac{18\zeta d_0^{1/2}}{c_x} + \frac{c_x}{16} \sqrt{\sum_{t=s_k+1}^{e} ||\Delta_t||_2^2}
\]
where the last inequality follows from (62) and which implies
\[
\sum_{t=s_k+1}^{e} \frac{c_x}{32} \sqrt{\sum_{t=s_k+1}^{e} ||\Delta_t||_2^2} \leq \frac{18\zeta d_0^{1/2}}{c_x}
\]
Therefore,
\[
\sum_{t=s_k+1}^{e} ||\hat{\beta}_t - \beta_t||_2^2 \leq 576^2 \zeta^2 d_0/c_x^4.
\]
Let $\beta_1^* = \beta_{\eta_k}^*$ and $\beta_2^* = \beta_{\eta_k + 1}^*$. We have that
\[
\sum_{t=s_k+1}^{c_k} \|\beta_t - \beta_1^*\|_2^2 = |I_1| \|\beta_1^* - \tilde{\beta}_1\|_2^2 + |I_2| \|\beta_2^* - \tilde{\beta}_1\|_2^2 + |I_3| \|\beta_2^* - \tilde{\beta}_2\|_2^2.
\]
Since
\[
\eta_k - s_k = \eta_k - \frac{2}{3} \bar{\eta}_k - \frac{1}{3} \bar{\eta}_k
\]
\[
= \frac{2}{3} (\eta_k - \eta_{k-1}) + \frac{2}{3} (\bar{\eta}_k - \eta_k) - \frac{2}{3} (\bar{\eta}_{k-1} - \eta_{k-1}) + (\eta_k - \bar{\eta}_k)
\]
\[
\geq \frac{2}{3} \Delta - \frac{1}{3} \Delta = \frac{1}{3} \Delta,
\]
where the inequality follows from Assumption 3 and (19), we have that
\[
\Delta \|\beta_1^* - \tilde{\beta}_1\|_2^2 / 3 \leq |I_1| \|\beta_1^* - \tilde{\beta}_1\|_2^2 \leq \frac{C_1 C_2^2 \Delta \kappa^2}{C_{\text{SNR}} d_0 K \sigma^2 \log^2 (n \lor p)} \leq c_1 \kappa^2,
\]
where $1/4 > c_1 > 0$ is an arbitrarily small positive constant. Therefore we have
\[
\|\beta_1^* - \tilde{\beta}_1\|_2^2 \leq c_1 \kappa^2.
\]
In addition we have
\[
\|\beta_2^* - \tilde{\beta}_1\|_2 \geq \|\beta_2^* - \beta_1^*\|_2 - \|\beta_1^* - \tilde{\beta}_1\|_2 \geq \kappa/2.
\]
Therefore, it holds that
\[
\kappa^2 |I_2| / 4 \leq |I_2| \|\beta_2^* - \tilde{\beta}_1\|_2 \leq \delta,
\]
which implies that
\[
|\bar{\eta}_k - \eta_k| \leq \frac{4 C_1 d_0 \zeta^2}{\kappa^2}.
\]

\section{Proofs in Model 1}

In view of Appendix A, in order to prove Theorem 1, we only need to provide the counterparts of Lemmas 9, 10, 11, 12 and 13, which are Lemmas 22(b), 22(a), 23, 24 and 26, respectively. The final results can be traced in Table 1.

For convenience, we also denote
\[
A_i^* = (A_i^*[1], \ldots, A_i^*[L]) \in \mathbb{R}^{p \times p_L}.
\]

We introduce some additional notation. For any $k \in \{0, \ldots, K\}$ and $J \subset \{1, \ldots, p\}$, define
\[
f_{k,J} = \frac{1}{2 \pi} \sum_{l=\infty}^{\infty} \Sigma_{k,J}(l) \cos \theta, \quad \theta \in (-\pi, \pi]
\]
and
\[
M(f_k, s) = \max_{J \subset \{1, \ldots, p\}, \ |J| \leq s} M(f_{k,J}), \quad s \in \{1, \ldots, p\}.
\]
It follows from the definitions that
\[
M(f_k, 1) \leq \ldots \leq M(f_k, p) = M(f_k).
\]
C.1 Deviation bounds

**Lemma 21** (Proposition 2.4 in Basu and Michailidis (2015)). For Model 1, under Assumption 1, the following holds.

(a) There exists an absolute constant \( c > 0 \), such that for any \( u, v \in \{ w \in \mathbb{R}^p : \| w \|_0 \leq s, \| w \|_2 \leq 1 \} \) and any \( \xi > 0 \), it holds that

\[
P \left\{ \left| v^T \sum_{t \in I} \left( X_tX_t^T - \mathbb{E}\{X_tX_t^T\} \right) v \right| \geq 2\pi \max_{k=0, \ldots, K} \mathcal{M}(f_k, s) \xi \right\} \leq 2 \exp\left\{ -c \min\{\xi^2 / |I|, \xi \} \right\}
\]

and

\[
P \left\{ \left| v^T \sum_{t \in I} \left( X_tX_t^T - \mathbb{E}\{X_tX_t^T\} \right) u \right| \geq 6\pi \max_{k=0, \ldots, K} \mathcal{M}(f_k, 2s) \xi \right\} \leq 6 \exp\left\{ -c \min\{\xi^2 / |I|, \xi \} \right\};
\]

in particular, for any \( i, j \in \{1, \ldots, p\} \), it holds that

\[
P \left\{ \left| \sum_{t \in I} \left( X_tX_t^T - \mathbb{E}\{X_tX_t^T\} \right) \right|_{ij} \geq 6\pi \max_{k=0, \ldots, K} \mathcal{M}(f_k, 2) \xi \right\} \leq 6 \exp\left\{ -c \min\{\xi^2 / |I|, \xi \} \right\}.
\]

(b) Let \( \{Y_t\}_{t=1}^n \) be a \( p \)-dimensional, centred, stationary process. Assume that for any \( t \in \{1, \ldots, n\} \), \( \text{Cov}(X_t, Y_t) = 0 \). The joint process \( \{(X_t^T, Y_t^T)^T\}_{t=1}^n \) satisfies Model 1(c). Let \( f_Y \) be the spectral density function of \( \{Y_t\}_{t=1}^n \), and \( f_{k,Y} \) be the cross spectral density function of \( X_k \) and \( \{Y_t\}_{t=1}^n, k \in \{0, \ldots, K\} \). There exists an absolute constant \( c > 0 \), such that for any \( u, v \in \{ w \in \mathbb{R}^p : \| w \|_2 \leq 1 \} \) and any \( \xi > 0 \), it holds that

\[
P \left\{ \left| v^T \sum_{t \in I} \left( X_tY_t^T - \mathbb{E}\{X_tY_t^T\} \right) u \right| \geq 2\pi \max_{k=0, \ldots, K} \left( \mathcal{M}(f_k) + \mathcal{M}(f_Y) + \mathcal{M}(f_{k,Y}) \right) \xi \right\} \leq 6 \exp\left\{ -c \min\{\xi^2 / |I|, \xi \} \right\}.
\]

Although there exists one difference between Lemma 21 and Proposition 2.4 in Basu and Michailidis (2015) that we have \( K + 1 \) different spectral density distributions, while in Basu and Michailidis (2015), \( K = 0 \), the proof can be conducted in the identical way, only noticing that the largest eigenvalue should be taken as the largest over all \( K + 1 \) different spectral density functions.

**Lemma 22.** For Model 1, under Assumption 1, the following holds.

(a) For any interval \( I \subset \{1, \ldots, n\} \), it holds that

\[
P \left\{ \left\| \sum_{t \in I} \varepsilon_{t+1}X_t^T \right\|_{\infty} \leq C \max\{\sqrt{|I| \log(n \vee p)}, \log(n \vee p)\} \right\} \geq 1 - 6(n \vee p)^{-c},
\]

where \( c > 0 \) is a constant defined in Lemma 21 and \( C > 0 \) is a constant depending on \( \mathcal{M}(f_k), \mathcal{M}(f_Y), \mathcal{M}(f_{k,e}), k = 0, \ldots, K \).
(b) For any interval \( I \subset \{1, \ldots, p\} \) satisfying
\[
|I| > \left( \frac{6 \times 54 \pi M}{c_x^2} \right)^2 \frac{\log(p)}{c}, \tag{64}
\]
with probability at least \( 1 - 6n^{-c} \), it holds that for any \( B \in \mathbb{R}^{p \times p} \),
\[
\sum_{t \in I} \|BX_t\|_2^2 \geq \frac{|I|c_x^2}{2} \|B\|_2^2 - C_x \log(p) \|B\|_2^1,
\]
where \( C_x > 0 \) is an absolute positive constant depending on all the other constants.

Proof. Part (a) is a direct application of Lemma 21(b), by setting \( Y_t = \varepsilon_t \). Part (b) is as follows.

Let \( \hat{\Sigma}_I = (|I|)^{-1} \sum_{t \in I} X_t X_t^\top \) and \( \Sigma^*_I = \mathbb{E}(\hat{\Sigma}_I) \). It is due to (63) that with probability at least \( 1 - 6n^{-c} \), it holds that
\[
(\log(p))^{-1} \sum_{t \in I} \|(X_t^\top \otimes I) \text{vec}(B)\|_2^2 = (\text{vec}(B))^\top (\hat{\Sigma}_I \otimes I_p) \text{vec}(B)
\]
\[
\geq (\text{vec}(B))^\top (\Sigma^*_I \otimes I_p) \text{vec}(B) - \left| (\text{vec}(B))^\top \left\{ (\hat{\Sigma}_I - \Sigma^*_I) \otimes I_p \right\} \text{vec}(B) \right| \geq \frac{c_x^2}{2} \|B\|_2^2 - \frac{\log(p)}{|I|} \|B\|_2^1.
\]

The last inequality in (65) follows the proof of Lemmas 12 and 13 in the Supplementary Materials in Loh and Wainwright (2011), and the proof of Proposition 4.2 in Basu and Michailidis (2015), by taking
\[
\delta = \frac{6\pi M \sqrt{\log(p)}}{\sqrt{|I|}} \leq \frac{c_x^2}{54},
\]
where the inequality holds due to (64), and by taking
\[
s = \left\lceil 2 \times \left( 27 \times 6\pi M \right)^2 \frac{C_x c_3}{C_x c_2} \right\rceil.
\]

\( \square \)

C.2 Additional technical lemmas

Lemma 23. For Model 1, under Assumption 1, if there exists no true change point in \( I \in (s, e] \), which satisfies that
\[
|I| > \frac{4C_x \log(n \lor p)d_0}{c_x^2},
\]
with
\[
\lambda \geq \lambda_1 = C_\lambda \sqrt{\log(n \lor p)}
\]
and \( C_\lambda \) being an absolute constant, then with probability at least \( 1 - 6(n \lor p)^{-c} \),
\[
\|\hat{A}_I^\lambda - A_I^\lambda\|_2 \leq \frac{C_1 \lambda \sqrt{d_0}}{\sqrt{|I|}} \quad \text{and} \quad \|\hat{A}_I^\lambda - A_I^\lambda\|_1 \leq \frac{C_1 d_0 \lambda}{\sqrt{|I|}},
\]
where \( C_1 > 0 \) is an absolute constant depending on all the other constants.
Proof. Denote \( A^* = A^* I \) and \( \hat{A} = \hat{A}^I \). Note that
\[
|I| > \frac{4C_x \log(n \vee p)d_0}{c^2_x},
\]
therefore \(|I| > \log(n \vee p)\).
Since
\[
\sum_{t \in I} \|X_{t+1} - A X_{t}\|_2^2 + \lambda \sqrt{|I|}\|\hat{A}\|_1 \leq \sum_{t \in I} \|X_{t+1} - A^* X_t\|_2^2 + \lambda \sqrt{|I|}\|A^*\|_1,
\]
we have the following holds with probability at least \( 1 - n^{-c} \),
\[
\sum_{t \in I} \|\hat{A} X_{t} - A^* X_t\|_2^2 + \lambda \sqrt{|I|}\|\hat{A}\|_1 \leq 2 \sum_{t \in I} \varepsilon_{t+1}^T (\hat{A} - A^*) X_t + \lambda \sqrt{|I|}\|A^*\|_1 \leq \lambda/2 \sqrt{|I|}\|A^*\|_1 + \lambda \sqrt{|I|}\|A^*\|_1.
\]
(66)
The final claims follow from (66) and standard calculations in Lemma 11. \( \square \)

**Lemma 24.** For Model 1, under Assumption 1, if there exists no true change point in \( I \in [s, e] \), with
\[
\lambda \geq \lambda_1 = C_\lambda \sqrt{\log(n \vee p)}
\]
and \( C_\lambda \) being an absolute constant, then with probability at least \( 1 - 6(n \vee p)^{-c} \),
\[
\sum_{t \in I} \|X_{t+1} - A^*_I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - \hat{A}^*_I X_t\|^2 \leq C_2 d_0 \lambda^2, \text{ if } |I| > \frac{4C_x \log(n \vee p)d_0}{c^2_x}, \text{ and}
\]
\[
\sum_{t \in I} \|X_{t+1} - A^*_I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - \hat{A}^*_I X_t\|^2 \leq C_2 \lambda d_0^{3/2} \sqrt{\log(n \vee p)}, \text{ if } |I| \leq \frac{4C_x \log(n \vee p)d_0}{c^2_x}.
\]

Proof. If
\[
|I| > \frac{4C_x \log(n \vee p)d_0}{c^2_x},
\]
then it follows Lemma 23 that with probability at least \( 1 - 6(n \vee p)^{-c} \),
\[
\sum_{t \in I} \|X_{t+1} - \hat{A}_I^I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - A^*_I X_t\|^2 \leq -\lambda \sqrt{|I|}\|\hat{A}^I_I\|_1 + \lambda \sqrt{|I|}\|A^*_I\|_1 \leq \lambda \sqrt{|I|}\|\hat{A}^I_I - A^*_I\|_1 \leq C_1 d_0 \lambda^2
\]
and
\[
\sum_{t \in I} \|X_{t+1} - A^*_I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - \hat{A}^*_I X_t\|^2 = -\sum_{t \in I} \|\hat{A}^I_I X_t - A^*_I X_t\|^2 + 2 \sum_{t \in I} \varepsilon_{t+1}^T (A^*_I - \hat{A}^*_I) X_t \leq 2 \|\hat{A} - A^*_I\|_1 \sum_{t \in I} \varepsilon_{t+1}^T X_t^T \leq \lambda \sqrt{|I|}\|\hat{A}^I_I - A^*_I\|_1 \leq C_1 d_0 \lambda^2.
\]
where the second inequality is due to Lemma 22(a).
If
\[ |I| < \frac{4C_x \log(n \vee p)d_0}{c_x^2}, \]
then with probability at least \(1 - 6(n \vee p)^{-c}\),
\[ \sum_{t \in I} \|X_{t+1} - \hat{A}_t^I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - A_t^I X_t\|^2 \leq \lambda \sqrt{\max\{|I|, \log(n \vee p)|A_t^I\|_1} + \lambda \sqrt{\max\{|I|, \log(n \vee p)|A_t^I\|_1} \leq \lambda d_0^{3/2} C \beta \sqrt{\log(n \vee p)}.
\]
In addition, with probability at least \(1 - 6(n \vee p)^{-c}\), it follows from standard Lasso estimation arguments that
\[ \|\hat{A}_t^I(S^c)\|_1 \leq 3\|\hat{A}_t^I - A_t^I\|(S)\|_1 \leq Cd_0, \]
which implies that
\[ \|\hat{A}_t^I - A_t^I\|_1 \leq 4Cd_0. \]
Therefore,
\[ \sum_{t \in I} \|X_{t+1} - A_t^I X_t\|^2 - \sum_{t \in I} \|X_{t+1} - \hat{A}_t^I X_t\|^2 = -\sum_{t \in I} \|\hat{A}_t^I X_t - A_t^I X_t\|^2 + 2\sum_{t \in I} \varepsilon_t^\top (A_t^I - \hat{A}_t^I) X_t \leq 2\|\hat{A}_t^I - A_t^I\|_1 \left\| \sum_{t \in I} \varepsilon_t X_t^\top \right\|_\infty \leq 8Cd_0 \sigma \epsilon C \max\{\sqrt{|I| \log(n \vee p)}, \log(n \vee p)\} \leq \lambda d_0^{3/2} C \beta \sqrt{\log(n \vee p)}. \]

\[ \text{Lemma 25. For Model 1, suppose Assumption 1 holds. Let } A_t^* \text{ be defined as} \]
\[ \left( \sum_{t \in I} E (X_t X_t^\top) \right) (A_t^*)^\top = \sum_{t \in I} E (X_t X_t^\top) (A_t^*)^\top. \]  
\[ \text{(67)} \]
Then we have that \(\|A_t^*\|_0 \leq 4d_0^2\) and
\[ \|A_t^*\|_\infty \leq \frac{\Lambda\max \left( \sum_{t \in I} E (X_t X_t^\top) A_t^* \right)}{\Lambda\min \left( \sum_{t \in I} E (X_t X_t^\top) \right)} \leq \frac{\max \left( \sum_{k=0,\ldots,K} \Lambda\min (\Sigma_k(0)) \right)}{\Lambda\min (\Sigma_k(0))}. \]  
\[ \text{(68)} \]

\[ \text{Proof. Let } S \text{ be the common support of } A_t^* \text{ defined in Assumption 1(a),} \]
\[ S_1 = \{i : (i, j) \in S\} \subset \{1, \ldots, p\} \quad \text{and} \quad S_2 = \{j : (i, j) \in S\} \subset \{1, \ldots, p\}. \]
Therefore by assumption, \(\max\{|S_1|, |S_2|\} \leq d_0\). With a permutation if necessary, without loss of generality, we have that \((S_1 \cup S_2) \subset \{1, \ldots, 2d_0\}\), which implies that each \(A_t^*\) has the block structure
\[ A_t^* = \begin{pmatrix} a_t^* & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{p \times p}, \]  
\[ \text{(69)} \]
where $a_i^* \in \mathbb{R}^{2d_0 \times 2d_0}$. Denote $S = (S_1 \cup S_2)^{\otimes 2} \subset \{1, \ldots, 2d_0\}^{\otimes 2}$. Note that $|S| \leq 4d_0^2$ and that $A_t^*(i, j) = 0$ if $(i, j) \in S^c$. For any realisation $X_t$ of such VAR(1) process with transition matrix $A_1$, the covariance of $X_t$ is of the form

$$\mathbb{E}(X_t X_t^\top) = \begin{pmatrix} \sigma_t & 0 \\ 0 & 0 \end{pmatrix},$$

where $\sigma_t \in \mathbb{R}^{2d_0 \times 2d_0}$. Since $\sigma_t$ is invertible, the matrix $A_t^*$ is unique and is of the same form as in (69). Since $\|A_t^*\|_{op} < 1$ for all $t \in I$, by assumption it holds that

$$\Lambda_{\max} \left( \sum_{t \in I} \mathbb{E}(X_t X_t^\top) (A_t^*)^\top \right) \leq \sum_{t \in I} \Lambda_{\max} \left( \mathbb{E}(X_t X_t^\top) \right)$$

and

$$\Lambda_{\min} \left( \sum_{t \in I} \mathbb{E}(X_t X_t^\top) \right) \geq \sum_{t \in I} \Lambda_{\min} \left( \mathbb{E}(X_t X_t^\top) \right).$$

Combining (67), (70) and (71) leads to

$$\|A_t^*\|_{op} \leq \frac{\Lambda_{\max} \left( \sum_{t \in I} \mathbb{E}(X_t X_t^\top) A_t^* \right)}{\Lambda_{\min} \left( \sum_{t \in I} \mathbb{E}(X_t X_t^\top) \right)} \leq \max_{k=0, \ldots, K} \frac{\Lambda_{\max}(\Sigma_k(0))}{\Lambda_{\min}(\Sigma_k(0))}.$$  

\[\Box\]

**Lemma 26.** For Model 1, under Assumption 1, if

$$\lambda \geq \lambda_2 = C_\lambda \sqrt{\log(n \vee p)}$$

with $C_\lambda$ an absolute constant, then with probability at least $1 - n^{-c}$,

$$\|\hat{A}_I^*(S^c)\|_1 \leq 3 \|\hat{A}_I^*(S)\|_1;$$

if in addition, the interval $I$ satisfies

$$|I| > \frac{4C_\lambda \log(n \vee p)d_0^2}{c_x^2},$$

then

$$\|A_I^* - \hat{A}_I^*\|_2 \leq \frac{C\lambda d_0}{\sqrt{|I|}} \quad \text{and} \quad \|A_I^* - \hat{A}_I^*\|_1 \leq \frac{C\lambda d_0^2}{\sqrt{|I|}},$$

where $C > 0$ is an absolute constant and $A_I^*$ is any matrix such that

$$\left( \sum_{t \in I} \mathbb{E}(X_t X_t^\top) \right) (A_I^*)^\top = \sum_{t \in I} \mathbb{E}(X_t X_t^\top) (A_t^*)^\top.$$

**Proof.** Due to Model 1, we have that $X_{\eta_k+1}$ and $X_{\eta_k}$ are independent, therefore

$$X_{\eta_k+1} - A_{\eta_k} X_{\eta_k} \not\sim \varepsilon_{\eta_k+1}.$$
As a remedy, let $\tilde{X}_{\eta_k}$ to be a latent random vector, which is a real predictor of $X_{\eta_k+1}$. Thus one can write

$$X_{\eta_k+1} - A_{\eta_k} \tilde{X}_{\eta_k} = \varepsilon_{\eta_k+1}.$$ 

The existence of $\tilde{X}_{\eta_k}$ is due to Model 1. Lemma 25 implies that $\|A_t^*\|_0 \leq 4d_0^2$, and that the support $S$ of $A_t^*$ is such that $S \subset S$.

Let $\Delta_t = A_t^* - \hat{A}_t^*$. From standard Lasso calculations, we have

$$\sum_{t \in I} \|\Delta_t X_t\|^2 + 2 \sum_{t \in I} (X_{t+1} - A_t^* X_t)^\top (\Delta_t X_t) + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|\hat{A}_t^*\|_1$$

$$\leq \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|A_t^*\|_1. \quad (72)$$

Note that

$$\sum_{t \in I} (X_{t+1} - A_t^* X_t)^\top (\Delta_t X_t) = \sum_{t \in I} (X_{t+1} - A_t^* X_t)^\top (\Delta_t X_t) + \sum_{t \in I} \{(A_t^* - A_t^*)_t\}^\top (\Delta_t X_t)$$

$$= \sum_{t \in I \setminus \{\eta_k\}} \varepsilon_t \Delta_t X_t + \sum_{t \in \{\eta_k\} \cap I} \sum_{j,l} \eta_t \Delta_t X_t + \sum_{t \in I} \{(A_t^* - A_t^*)_t\}^\top (\Delta_t X_t)$$

$$= (I) + (II) + (III).$$

As for $(I)$, by Lemma 22(a), with probability at least $1 - 6(n \vee p)^{-c}$,

$$|(I)| \leq \|\Delta_t\|_1 C \max\{\sqrt{|I| \log(n \vee p)}, \log(n \vee p)\}.$$

As for $(III)$, we have

$$|(III)| \leq \|\Delta_t\|_1 \left\| \sum_{t \in I} X_{t}X_{t}^\top (A_t^* - A_t^*)_t \right\| \leq \max_{j,l \in \{1, \ldots, p\}} \left\| \sum_{t \in I} X_{t}(j)X_{t}^\top (A_t^* - A_t^*)_t \right\|.$$

In addition, it holds that

$$\mathbb{E} \left( \sum_{t \in I} X_{t}X_{t}^\top (A_t^* - A_t^*)_t \right) = 0,$$

due to (67). Let $v_t^l$ to be the $l$-th column of $(A_t^* - A_t^*_t)$. Then $\|v_t^l\|_2 \leq \|A_t^* - A_t^*_t\|_{\text{op}} \leq 2$.

Consider the process $\{V_t\} = \{(X_t^\top, X_t^\top v_t^l)\} \in \mathbb{R}^{p+1}$, $v = e_j$ for any $j = 1, \ldots, p$ and $u = e_{p+1}$, where $e_k \in \mathbb{R}^p$ with $e_{kl} = 1\{k = l\}$. Observe that

$$v^\top \sum_{t \in I} V_t V_t^\top u = \sum_{t \in I} X_t(j)X_t^\top (A_t^* - A_t^*)_t,$$

and

$$\text{Var}(X_t^\top v_t^l) \leq v_t^l E(X_t^\top X_t^\top) v_t^l \leq 8\pi \mathcal{M},$$

where $\|v_t^l\|_2 \leq 2$ and $\|E(X_t^\top X_t^\top)\|_{\text{op}} \leq 2\pi \mathcal{M}$ are used in the last inequality. It follows from Lemma 21(a) that with probability at least $1 - 6(n \vee p)^{-c}$,

$$\left| v \left( \sum_{t \in I} V_t V_t^\top \right) u \right| \leq 6\pi \mathcal{M} C_\beta d_0 \max\{\sqrt{|I| \log(n \vee p)}, \log(n \vee p)\},$$

44
therefore
\[(III) \leq \|\Delta_I\|_1 6\pi M C_{\beta} \sqrt{d_0} \max\{\sqrt{|I| \log(n \lor p), \log(n \lor p)}\}.
\]

As for \((II)\), we have
\[(II) \leq \|\Delta_I\|_1 \left\| \sum_{t \in I} X_t (\tilde{X}_t - X_t)^\top (A_t^*)^\top \right\|_\infty
\]

For any row \(A_t^*(i)\), it holds that \(\|A_t^*(i)\|_0 \leq \|A_t^*\|_0 \leq d_0\), and \(\|A_t^*(i)\|_2 \leq \|A_t^*\|_{op} \leq 1\). It follows from Lemma 21(a) that with probability at least \(1 - 6(n \lor p)^{-c}\),
\[
\max_{i,j=1,\ldots,p} \left| \sum_{t \in \{\eta_k\} \cap I} X_t(i) X_t^\top A_t^*(j) \right| = \max_{i,j=1,\ldots,p} \left| \sum_{t \in \{\eta_k\} \cap I} X_t X_t^\top A_t^*(j) \right|
\leq 6\pi M \max\{\sqrt{|I| \log(n \lor p), \log(n \lor p)}\};
\]

and it follows from Lemma 21(b) that with probability at least \(1 - 6(n \lor p)^{-c}\),
\[
\max_{i,j=1,\ldots,p} \left| \sum_{t \in \{\eta_k\} \cap I} X_t(i) \tilde{X}_t^\top A_t^*(j) \right| = \max_{i,j=1,\ldots,p} \left| \sum_{t \in \{\eta_k\} \cap I} X_t \tilde{X}_t^\top A_t^*(j) \right|
\leq 6\pi M \max\{\sqrt{|I| \log(n \lor p), \log(n \lor p)}\}.
\]

Therefore, we have
\[(II) \leq 12\pi M \max\{\sqrt{|I| \log(n \lor p), \log(n \lor p)}\} \|\Delta_I\|_1.
\]

Thus (72) leads to
\[
\begin{align*}
\sum_{t \in I} \|\Delta_t X_t\|^2 + \lambda \max\{|I|, \log(n \lor p)\} \|\tilde{A}_t\|_1 & \leq \lambda \sqrt{\max\{|I|, \log(n \lor p)\}} \|A_t^*\|_1 \\
+ \|\Delta_I\|_1 (2C + 12\pi M + 24\pi M) \max\{\sqrt{|I| \log(n \lor p), \log(n \lor p)}\} & \leq \lambda \sqrt{\max\{|I|, \log(n \lor p)\}} \|A_t^*\|_1 + \lambda/2 \sqrt{\max\{|I|, \log(n \lor p)\}} \|\Delta_I\|_1.
\end{align*}
\]

which leads to the final claims combining the fact that \(\|A_t^*\|_0 \leq 4d_0^2\) and the standard treatments on Lasso estimation procedures as in Lemma 11. \(\square\)

### D Proof in Model 2

In view of Appendix A, in order to prove Corollary 3, we only need to provide the counterparts of Lemmas 9, 10, 11, 12 and 13, which are Lemmas 27, 22(a), 23, 24 and 29, respectively.

Note that we directly inherit Lemmas 22(a), 23 and 24 from the VAR(1) case, since they are identical in the VAR(L) case. Note that the final results can be traced in Table 1.

**Lemma 27.** For Model 2, under Assumption 2, for any interval \(I \subset \{1, \ldots, p\}\) satisfying
\[
|I| \geq 2 \max \left\{ \left( \frac{6 \times 54\pi M}{c^2} \right) \frac{\log(p)}{c}, KL \right\}, \quad (73)
\]


with probability $1 - 6n^{-c}$, it holds that for any matrix sequence $\{B[1], \ldots, B[L]\} \in \mathbb{R}^{p \times p}$,

$$\sum_{t \in I} \left\| \sum_{l=1}^{L} B[l] X_{t-l+1} \right\|_2^2 \geq \frac{|I\setminus J| c_x^2}{2} \|B\|_2^2 - C_x \log(p) \left( \sum_{l=1}^{L} \|B[l]\|_1 \right)^2,$$

where $C_x > 0$ is an absolute constant depending on all the other constants.

**Proof.** For any sequence of matrices $\{B[1], \ldots, B[L]\} \subset \mathbb{R}^{p \times p}$, let $B = (B[1], \ldots, B[L]) \in \mathbb{R}^{p \times p L}$. Let $Y_t = (X_t^T, \ldots, X_{t-L+1}^T)^T \in \mathbb{R}^{p L}$.

It follows from Lemma 22(b) that with probability at least $1 - 6n^{-c}$,

$$\sum_{t \in I} \|BY_t\|_2^2 \geq \frac{|I\setminus J| c_x^2}{2} \|B\|_2^2 - C_x \log(p) \|B\|_1^2,$$

where

$$J = \cup_{k=0}^{K} \{\eta_k + 1, \ldots, \eta_k + L\},$$

as long as

$$|I\setminus J| > \left( \frac{6 \times 54 \pi M}{c_x^2} \right) \log(p).$$

In addition, since $|J| \leq KL$ and

$$\sum_{t \in I} \|BY_t\|_2^2 \geq \sum_{t \in I\setminus J} \|BY_t\|_2^2,$$

the final result holds. \qed

**Lemma 28.** For Model 2, under Assumption 2, let $\Sigma_t$ be the covariance matrix of

$$Y_t = (X_t^T, \ldots, X_{t-L+1}^T)^T \in \mathbb{R}^{p L},$$

for each $t$. Let $m = |S|$. With a permutation if needed, suppose that each $A_t^*[l] \in \mathbb{R}^{p \times p}$, $t \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, L\}$, has the block structure

$$A_t^*[l] = \begin{pmatrix} a_t[l] & 0 \\ 0 & 0 \end{pmatrix}, \quad (74)$$

where $a_t[l] \in \mathbb{R}^{2m \times 2m}$. Let the matrix $A_t^* \in \mathbb{R}^{p \times p L}$ satisfy

$$\sum_{t \in I} (A_t^*[1], \ldots, A_t^*[L]) \Sigma_t = A_t^* \sum_{t \in I} \Sigma_t. \quad (75)$$

Then the solution $A_t^*$ exists and is unique. It holds that

$$\|A_t^*\|_{op} \leq \max_{k=0, \ldots, K} \frac{\Lambda_{\max}(\Sigma_t)}{\Lambda_{\min}(\Sigma_t)}.$$

In addition, let

$$A_t^* = (A_t^*[1], \ldots, A_t^*[L]) \in \mathbb{R}^{p \times p L},$$

where $A_t^*[l] \in \mathbb{R}^{p \times p}$. Then $\|A_t^*[l]\|_0 \leq m^2$, $l \in \{1, \ldots, L\}$, and consequently $\|A_t^*\|_0 \leq Lm^2$. 46
Proof. It follows from (74), the covariance of $Y_t$ is of the form

$$
\Sigma_t = \begin{pmatrix}
\Sigma_t(1, 1) & \ldots & \Sigma_t(1, L) \\
\vdots & \ddots & \vdots \\
\Sigma_t(L, 1) & \ldots & \Sigma_t(L, L)
\end{pmatrix},
$$

where for $i \in \{1, \ldots, L\}$,

$$
\Sigma_t(i, i) = \begin{pmatrix}
\sigma_t(i, i) & 0 \\
0 & I
\end{pmatrix} \in \mathbb{R}^{p \times p},
$$

for some $\sigma_t(i, i) \in \mathbb{R}^{2m \times 2m}$; for $i, j \in \{1, \ldots, L\}$ with $i < j$,

$$
\Sigma_t(i, j) = \Sigma_t(j, i) = \begin{pmatrix}
\sigma_t(i, j) & 0 \\
0 & 0
\end{pmatrix},
$$

for some $\sigma_t(i, j) \in \mathbb{R}^{2m \times 2m}$. Since

$$
\Lambda_{\min} \left( \sum_{t \in I} \Sigma_t \right) \geq \sum_{t \in I} \Lambda_{\min}(\Sigma_t),
$$

the matrix $A^*_t$ exits and is unique. The bounds on the operator norm of $A^*_t$ follows from the same argument used in (68). By matching coordinates, (75) is equivalent to

$$
\left( \sum_{t \in I} \sum_{i=1}^L a_t[i] \sigma_t(i, j) \right) = \sum_{i=1}^L A^*_t[i] \left( \sum_{t \in I} \sigma_t(i, j) \right), \quad j = 1, \ldots, L.
$$

Let

$$
A^*_t[i] = \begin{pmatrix}
A^*_t[i](1, 1) & A^*_t[i](1, 2) \\
A^*_t[i](2, 1) & A^*_t[i](2, 2)
\end{pmatrix},
$$

where $A^*_t[i](1, 1) \in \mathbb{R}^{2m \times 2m}$. It suffices to show that in the above block structure, only $A^*_t[i](1, 1) \neq 0$, which implies that $\|A^*_t[i]\|_0 \leq 4m^2$. Since

$$
\sum_{t \in I} (a_t[1] \ldots a_t[L]) \begin{pmatrix}
\sigma_t(1, 1) & \ldots & \sigma_t(1, L) \\
\vdots & \ddots & \vdots \\
\sigma_t(L, 1) & \ldots & \sigma_t(L, L)
\end{pmatrix}
= \begin{pmatrix}
A^*_t[1](1, 1) & \ldots & A^*_t[L](1, 1)
\end{pmatrix} \sum_{t \in I} \begin{pmatrix}
\sigma_t(1, 1) & \ldots & \sigma_t(1, L) \\
\vdots & \ddots & \vdots \\
\sigma_t(L, 1) & \ldots & \sigma_t(L, L)
\end{pmatrix},
$$

by the uniqueness of $A^*_t$, $A^*_t[i](k, l) = 0$ for any $k = 2$. Since the matrix

$$
\sigma_t = \begin{pmatrix}
\sigma_t(1, 1) & \ldots & \sigma_t(1, L) \\
\vdots & \ddots & \vdots \\
\sigma_t(L, 1) & \ldots & \sigma_t(L, L)
\end{pmatrix}
$$

is the covariance matrix of $(X_t[1 : 2s]^{\top}, X_{t-1}[1 : 2s]^{\top}, \ldots, X_{t-L+1}[1 : 2s]^{\top})^{\top}$, we have that

$$
e_x \leq \Lambda_{\min}(\sigma_t) \leq \Lambda_{\max}(\sigma_t) \leq M, \quad \forall t,
$$

and that the matrix $\sum_{t \in I} \sigma_t$ is invertible. We therefore complete the proof. \qed
Recall that $S$ is the common support of $A_t^*[l]$, $l \in \{1, \ldots, L\}$, defined in Assumption 1(a). Let

$$S_1 = \{i : (i, j) \in S\} \subset \{1, \ldots, p\} \quad \text{and} \quad S_2 = \{j : (i, j) \in S\} \subset \{1, \ldots, p\}.$$  

By assumption, it holds that $\max\{|S_1|, |S_2|\} \leq |S| = d_0/L$. With a permutation on coordinates if necessary, without loss of generality, we assume that $S_1 \cup S_2 \subset \{1, \ldots, 2|S|\}$, which implies that for each $t$, $A_t^*$ has the block structure

$$A_t^* = \begin{pmatrix} a_t^* & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{p \times p},$$

where $a_t^* \in \mathbb{R}^{2|S| \times 2|S|}$. Denote

$$S = (S_1 \cup S_2)^{\otimes 2} \subset \{1, \ldots, 2|S|\}^{\otimes 2}. \tag{76}$$

Note that $|S| \leq 4|S|^2$ and that $A_t^*(i, j) = 0$ if $(i, j) \in S^c$. Let $X_t$ and $X_{t'}$ be two realisations generated from Model 2, then it holds that

$$\mathbb{E}(X_tX_{t'}^\top) = \begin{pmatrix} \sigma_{t,t'} & 0 \\ 0 & 0 \end{pmatrix}.$$

**Lemma 29.** Under Model 2 and Assumption 2, for any interval $I \in (s,e]$ satisfying (73), with

$$\lambda \geq \lambda_2 = C_\lambda \log(n \vee p),$$

and $C_\lambda > 0$ an absolute constant, it holds that with probability at least $1 - n^{-c}$,

$$\| (\hat{A}_1^*[1](S^c), \ldots, \hat{A}_L^*[1](S^c)) \|_1 \leq 3 \| (\hat{A}_1^*[1](S), \ldots, \hat{A}_L^*[1](S)) \|_1,$$

where $S$ is defined in (76). If in addition, the interval $I$ satisfies

$$|I| > \frac{4C_x \log(n \vee p) d_0^2}{C_x},$$

then

$$\| (A_1^*[1], \ldots, A_L^*[L]) - (\hat{A}_1^*[1], \ldots, \hat{A}_L^*[1]) \|_2 \leq \frac{C \lambda d_0}{\sqrt{|I|}},$$

and

$$\| (A_1^*[1], \ldots, A_L^*[L]) - (\hat{A}_1^*[1], \ldots, \hat{A}_L^*[1]) \|_1 \leq \frac{C \lambda d_0}{\sqrt{|I|}},$$

where $C > 0$ is an absolute constant and $(A_1^*[1], \ldots, A_L^*[L]) \in \mathbb{R}^{p \times pL}$ is any matrix such that

$$(A_1^*[1], \ldots, A_L^*[L]) \left( \sum_{l \in I} \mathbb{E}(Y_lY_l^\top) \right) = \sum_{l \in I} (A_1^*[1], \ldots, A_L^*[L]) \mathbb{E}(Y_lY_l^\top). \tag{77}$$

**Proof.** Let $A_t^*$ be defined as (77). By Lemma 28, $A_t^*[l]$, $l \in \{1, \ldots, L\}$, is supported on $S$, and it suffices to show that

$$\sum_{l=1}^L \| \hat{A}_t[l] - A_t^*[l] \|_1 \leq \frac{CL^2|S|^2 \lambda}{\sqrt{|I|}}.$$
where \( C > 0 \) is an absolute constant.

**Step 1.** For \( l \in \{1, \ldots, L\} \), let \( \Delta[l] = A^*_{t}[l] - \hat{A}_{t}[l] \). Standard calculations lead to

\[
\sum_{t \in I} \left\| \sum_{i=1}^{L} \Delta[l]X_{t+1-l} \right\|_2^2 + 2 \sum_{t \in I} (\Delta[l]X_{t+1-l})^T \left( X_{t+1} - \sum_{i=1}^{L} A^*_{t}[l]X_{t+1-l} \right) + \lambda \sqrt{|I|} \sum_{l=1}^{L} \| \hat{A}[l] \|_1 \leq \lambda \sqrt{|I|} \sum_{l=1}^{L} \| A^*_{t}[l] \|_1, \tag{78}
\]

Denote \( \mathcal{I}_k = [\eta_k, \eta_k + L - 1] \) and \( \mathcal{I} = \bigcup_{k=1}^{K} \mathcal{I}_k \). Let \( \tilde{Y}_t = (\tilde{X}^T_t, \ldots, \tilde{X}^T_{t-L+1})^T \) be the predictor of \( X_{t+1} \). Note that \( \tilde{Y}_t \neq Y_t = (X^T_t, \ldots, X^T_{t-L+1})^T \) only at \( t \in \mathcal{I} \). Observe that (78) gives

\[
\sum_{t \in I} (\Delta Y_t)^T (X_{t+1} - A^*_{t}Y_t) = \sum_{t \in I} (\Delta Y_t)^T \left\{ (X_{t+1} - A^*_{t}\tilde{Y}_t) + A^*_{t}(\tilde{Y}_t - Y_t) + (A^*_{t} - \hat{A}^*_{t})Y_t \right\} = \sum_{t \in I} (\Delta Y_t)^T A^*_{t}(\tilde{Y}_t - Y_t) + \sum_{t \in I} (\Delta Y_t)^T (A^*_{t} - \hat{A}^*_{t})Y_t = (I) + (II) + (III).
\]

**Step 2.** As for term \((I)\), by Lemma 21(b) and the assumption that \( \lambda \geq C_{\lambda} \sqrt{\log(p)} \), it holds that

\[
| (I) | \leq \| \Delta \|_1 \max_{i=1, \ldots, L} \left\| \sum_{t \in I} X_{t+1-l} \epsilon_t^T \right\|_\infty \leq \frac{\lambda}{10} \sqrt{|I|} \| \Delta \|_1.
\]

**Step 3.** As for term \((II)\), denote \( A_{i}^* \) as the \( i \)-th row of \( A_{t}^* \). It holds that

\[
| (II) | \leq \| \Delta \|_1 \max_{i=1, \ldots, p} \max_{j=1, \ldots, Lp} \left| \sum_{t \in \mathcal{I}} A_{i}^*(t)(\tilde{Y}_t - Y_t)Y_t(j) \right| \\
\leq \| \Delta \|_1 L \max_{i=1, \ldots, p} \max_{j=1, \ldots, Lp} \max_{\eta_k \in \mathcal{I}} \left| \sum_{\eta_k \in \mathcal{I}} A_{i}^*_{\eta_k-1+l}(i)(\tilde{Y}_{\eta_k-1+l} - Y_{\eta_k-1+l})Y_{\eta_k-1+l}(j) \right|.
\]

Observe that \( (\tilde{Y}_{\eta_k-1+l} - Y_{\eta_k-1+l})Y_{\eta_k-1+l}(j) \) and \( (\tilde{Y}_{\eta_{k'}-1+l} - Y_{\eta_{k'}-1+l})Y_{\eta_{k'}-1+l}(j) \) are independent if \( |k - k'| > 1 \). Therefore with probability at least \( 1 - n^{-6} \),

\[
\sum_{\eta_k \in \mathcal{I}} A_{i}^*_{\eta_k-1+l}(i)(\tilde{Y}_{\eta_k-1+l} - Y_{\eta_k-1+l})Y_{\eta_k-1+l}(j) \\
= \left( \sum_{k: k \text{ is odd}} + \sum_{k: k \text{ is even}} \right) A_{i}^*_{\eta_k-1+l}(i)(\tilde{Y}_{\eta_k-1+l} - Y_{\eta_k-1+l})Y_{\eta_k-1+l}(j) \\
\leq 2C_{\lambda} \max \{ \sqrt{K \log(Lp)} , \log(Lp) \},
\]

where the last inequality follows from standard tail bounds for the sum of independent sub-exponential random variables together with the observations that \( \| A_{t}(i)^* \|_2 \leq \| A_{t}^* \|_{op} \leq 1 \) for all \( t \). Therefore

\[
| (II) | \leq C_{\lambda} \| \Delta \|_1 L \max \{ \sqrt{K \log(p)} , \log(p) \} \leq \left( \frac{\lambda}{10} \right) \sqrt{|I|} \| \Delta \|_1,
\]

}\]
where $|I| \geq 2CL^2|S|^2K\log(p)$ and $\lambda \geq C\lambda\sqrt{\log(n \vee p)}$ are used in the last inequality.

**Step 4.** As for term $(III)$, we have that

$$
|(III)| \leq \|\Delta\|_1 \max_{i=1,\ldots,p} \max_{j=1,\ldots,L_p} \left| \sum_{t \in I} (A^*_t(i) - A^*_t(i))Y_tY_t(j) \right|
$$

$$
\leq \|\Delta\|_1 \max_{i=1,\ldots,p} \max_{j=1,\ldots,L_p} \left| \sum_{t \in I} (A^*_t(i) - A^*_t(i))\tilde{Y}_t\tilde{Y}_t(j) \right|
$$

$$
+ \|\Delta\|_1 \max_{i=1,\ldots,p} \max_{j=1,\ldots,L_p} \left| \sum_{t \in I} (A^*_t(i) - A^*_t(i))(\tilde{Y}_t - Y_t)\tilde{Y}_t(j) \right|
$$

$$
+ \|\Delta\|_1 \max_{i=1,\ldots,p} \max_{j=1,\ldots,L_p} \left| \sum_{t \in I} (A^*_t(i) - A^*_t(i))(\tilde{Y}_t - Y_t)\tilde{Y}_t(j) \right|
$$

$$
=(III.1) + (III.2) + (III.3).
$$

Using the same arguments as in **Step 3**, we have that

$$
|(III.3)| = \|\Delta\|_1 \max_{i=1,\ldots,p} \max_{j=1,\ldots,L_p} \left| \sum_{t \in I} (A^*_t(i) - A^*_t(i))(\tilde{Y}_t - Y_t)\tilde{Y}_t(j) \right| \leq (\lambda/10) \sqrt{|I|}\|\Delta\|_1
$$

and $|(III.1)| \leq (\lambda/10) \sqrt{|I|}\|\Delta\|_1$. Due to the construction of $A^*_t$, it holds that

$$
\mathbb{E} \left( \sum_{t \in I} (A^*_t - A^*_t)\tilde{Y}_t\tilde{Y}_t^T \right) = 0.
$$

Denote $v_t[i] = A^*_t(i) - A^*_t(i)$. Observe that

$$
\|v_t[i]\|_2 \leq \|A^*_t(i) - A^*_t(i)\|_{\text{op}} \leq 2.
$$

Consider the VAR process $V_t = (\tilde{Y}_t^T, \tilde{Y}_t^Tv_t[i])^T \in \mathbb{R}^{p+1}$. Since

$$
e_{p+1} \sum_t V_tV_t^T e_j = \sum_t v_t[i]\tilde{Y}_t\tilde{Y}_t(j),
$$

and that $\{\tilde{Y}_t\}$ is a VAR(1) change point process, Lemma 21(a) gives

$$
\mathbb{P} \left( \left| \sum_{t \in I} v_t[i]\tilde{Y}_t\tilde{Y}_t(j) - \mathbb{E} \left( \sum_{t \in I} v_t[i]\tilde{Y}_t\tilde{Y}_t(j) \right) \right| \geq 12\mathcal{M}\sqrt{|I|\log(pm)} \right) \leq \frac{1}{n^3p^3}.
$$

Therefore if $\lambda \geq C\mathcal{M}\sqrt{\log(pm)}$, then with probability less than $1/(p^3n^3)$,

$$
\left| \sum_{t \in I} v_t[i]\tilde{Y}_t\tilde{Y}_t(j) \right| \geq \frac{\lambda\sqrt{|I|}}{10}.
$$

So it holds that

$$
(III) \leq \frac{\lambda}{10} \sqrt{|I|}\|\Delta\|_1.
$$
Step 5. The previous calculations give

\[ \sum_{t \in I} (\Delta Y_t)^2 + \frac{\lambda}{2} \sqrt{T} \sum_{l=1}^{L} \| \Delta[l] (S^c) \|_1 \leq (3\lambda/10) \sqrt{T} \| \Delta(S) \|_1 \leq (3\lambda/5) \sqrt{T} \sum_{l=1}^{L} \| \Delta[l] (S) \|_1, \]

where \(|S| \leq 4L|S|^2\). With the restricted eigenvalue condition in Lemma 27, standard Lasso calculations yields the desired results.

References

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. The Annals of Statistics, 37 4046–4087.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. Econometrica 47–78.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics, 43 1535–1567.

Bickel, P. J. and Gel, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 711–728.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics, 37 1705–1732.

Bolstad, A., Van Veen, B. D. and Nowak, R. (2011). Causal network inference via group sparse regularization. IEEE transactions on signal processing, 59 2628–2641.

Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. Statistica Sinica, 1 409–428.

Chang, J., Guo, B. and Yao, Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. Journal of econometrics, 189 297–312.

Chang, J., Guo, B. and Yao, Q. (2018). Principal component analysis for second-order stationary vector time series. The Annals of Statistics, 46 2094–2124.

Chang, J., Yao, Q. and Zhou, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. Biometrika, 104 111–127.

Chen, J. and Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. Journal of the American Statistical association, 92 739–747.

Chen, X., Xu, M. and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. The Annals of Statistics, 41 2994–3021.
CHO, H. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics, 10* 2000–2038.

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 (Statistical Methodology), 77* 475–507.

DE MOL, C., GIANNONE, D. and REICHLIN, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics, 146* 318–328.

DETTLE, H. and GÖSMANN, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics, 12* 2578–2636.

FIECAS, M., LENG, C., LIU, W. and YU, Y. (2018). Spectral analysis of high-dimensional time series. *arXiv preprint arXiv:1810.11223*.

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association, 100* 830–840.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76* 495–580.

FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized m-estimation: Fast computation. *Journal of Computational and Graphical Statistics, 17* 201–204.

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics, 42* 2243–2281.

GUO, S., WANG, Y. and YAO, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* asw046.

HAN, F., LU, H. and LIU, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research, 16* 3115–3150.

HAUFE, S., MÜLLER, K.-R., NOLTE, G. and KRÄMER, N. (2010). Sparse causal discovery in multivariate time series. In *Causality: Objectives and Assessment*. 97–106.

HSU, N.-J., HUNG, H.-L. and CHANG, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis, 52* 3645–3657.

KAUL, A., JANDHYALA, V. K. and FOTOPoulos, S. B. (2018). Parameter estimation for high dimensional change point regression models without grid search. *arXiv preprint arXiv:1805.03719*.

KILICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107* 1590–1598.

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics, 40* 694–726.
Lee, S., Liao, Y., Seo, M. H. and Shin, Y. (2018). Oracle estimation of a change point in high-dimensional quantile regression. *Journal of the American Statistical Association, 113* 1184–1194.

Lee, S., Seo, M. H. and Shin, Y. (2016). The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 193–210.

Leonardi, F. and Bühlmann, P. (2016). Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*.

Lin, K., Sharpnack, J. L., Rinaldo, A. and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*. 6884–6893.

Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*. 2726–2734.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maidstone, R., Hocking, T., Rigail, G. and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing, 27* 519–533.

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.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research, 11* 2241–2259.

Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*.

Safikhani, A. and Shojaie, A. (2017). Joint structural break detection and parameter estimation in high-dimensional non-stationary var models. *arXiv preprint arXiv:1711.07357*.

Schneider-Luftman, D. and Walden, A. T. (2016). Partial coherence estimation via spectral matrix shrinkage under quadratic loss. *IEEE Transactions on Signal Processing, 64* 5767–5777.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics, 26* i517–i523.

Smith, S. M. (2012). The future of FMRI connectivity. *Neuroimage, 62* 1257–1266.

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S. and Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics, 11* 812–820.

Swanson, D. C. (2001). A general prognostic tracking algorithm for predictive maintenance. In *2001 IEEE Aerospace Conference Proceedings (Cat. No. 01TH8542)*, vol. 6. IEEE, 2971–2977.
Tank, A., Foti, N. and Fox, E. (2015). Bayesian structure learning for stationary time series. *arXiv preprint arXiv:1505.03131*.

Tu, Y., Yao, Q. and Zhang, R. (2017). Error-correction factor models for high-dimensional cointegrated time series.

Van de Geer, S. (2018). On tight bounds for the lasso. *The Journal of Machine Learning Research*, 19:1767–1814.

Van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge University Press.

Wang, D., Lin, K. and Willett, R. (2019). Statistically and computationally efficient change point localization in regression settings. *arXiv preprint arXiv:1906.11364*.

Wang, D., Yu, Y. and Rinaldo, A. (2017). Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*.

Wang, D., Yu, Y. and Rinaldo, A. (2018a). Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*.

Wang, D., Yu, Y. and Rinaldo, A. (2018b). Univariate mean change point detection: Penalization, cusum and optimality. *arXiv preprint arXiv:1810.09498*.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:57–83.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379.

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics*, 40:466–493.

Yam, R., Tse, P., Li, L. and Tu, P. (2001). Intelligent predictive decision support system for condition-based maintenance. *The International Journal of Advanced Manufacturing Technology*, 17:383–391.

Zhang, B., Geng, J. and Lai, L. (2015). Change-point estimation in high dimensional linear regression models via sparse group lasso. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*. IEEE, 815–821.

Zhang, R., Robinson, P. and Yao, Q. (2019). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, 114:916–927.