Supplementary Materials for "Fast and Scalable Algorithm for Detection of Structural Breaks in Big VAR Models"

Abolfazl Safikhani
Yue Bai
and
George Michailidis
Department of Statistics and Informatics Institute, University of Florida

June 26, 2021

In Appendix A, we collect technical lemmas needed to prove the main results, which are presented in Appendix B. Details of the algorithm for solving the optimization problem (4) are presented in Appendix C, while the selection of the tuning parameters is discussed in detail in Appendix D. Further, additional simulation results for the comparison with two competing methods are reported in Appendix E. Finally, note that all constants $c_1, c_2, \ldots, k_1, k_2, \ldots$ in the proofs of Lemmas and Theorems are generic constants, and their repetitions on different places in the proofs do not imply they are identical.

Appendix A: Technical Lemmas

Lemma A1. *Under Assumption A1, there exist constants $c_i > 0$ such that with probability at least $1 - c_1 \exp(-c_2(\log(q) + 2\log(p)))$, for any sequence c_n,*

$$\sup_{1\le j\le m_0+1, s\ge t_{j-1}+q, |t_j-s|>c_n} \left|\left|(t_j - s)^{-1} \sum_{l=s}^{t_j-1} Y_{l-1}Y_{l-1}' - \Gamma_j^q(0)\right|\right|_\infty \le c_3\sqrt{\frac{\log(q) + 2\log(p)}{c_n}}, \tag{A.1}$$

where $\Gamma_j^q(0) = \mathbb{E}(Y_{t-1}Y_{t-1}')$ for $t_{j-1} + q \le t < t_j$. Moreover,

$$\sup_{1\le j\le m_0+1, s\ge t_{j-1}+q, |t_j-s|>c_n} \left|\left|(t_j - s)^{-1} \sum_{l=s}^{t_j-1} Y_{l-1}\varepsilon_l'\right|\right|_\infty \le c_3\sqrt{\frac{\log(q) + 2\log(p)}{c_n}}. \tag{A.2}$$

Proof. The proof of this lemma is similar to that of Proposition 2.4 in Basu and Michailidis (2015). Next, we briefly outline the main steps of the proof, while omitting unnecessary

details. For (A.1), note that using an argument similar to Proposition 2.4(a) in Basu and Michailidis (2015), there exist $k_1, k_2 > 0$ such that for each fixed $k, l = 1, \cdots, pq$,

$$\mathbb{P}\left(\left|e_k'\Big(\frac{\sum_{l=s}^{t_j-1} Y_{l-1}Y_{l-1}'}{t_j - s} - \Gamma_j^q(0)\Big)e_l\right| > k_1\eta\right) \leq 6\exp(-k_2 c_n \min(\eta, \eta^2)). \tag{A.3}$$

Setting $\eta = k_3\sqrt{\frac{\log(qp^2)}{c_n}}$, and taking union over all possible values of k, l, we obtain (A.1).

Note that, there exist $k_1, k_2 > 0$ such that for each fixed $k = 1, ..., pq$, $l = 1, ..., p$,

$$\mathbb{P}\left(\left|e_k'\frac{\sum_{l=s}^{t_j-1} Y_{l-1}\varepsilon_l'}{t_j - s}e_l\right| > k_1\eta\right) \leq 6\exp(-k_2 c_n \min(\eta, \eta^2)). \tag{A.4}$$

Setting $\eta = k_3\sqrt{\frac{\log(qp^2)}{c_n}}$, and taking union over all possible values of k, l, we get:

$$\left|\left|(t_j - s)^{-1}\sum_{l=s}^{t_j-1} Y_{l-1}\varepsilon_l'\right|\right|_\infty \leq c_3\sqrt{\frac{\log(q) + 2\log(p)}{c_n}}, \tag{A.5}$$

with high probability converging to 1 for any $j = 1, 2, \cdots, m_0 + 1$, as long as $|t_j - s| > c_n$ and $s \geq t_{j-1} + q$. Note that the constants c_1, c_2 and c_3 can be chosen large enough such that the upper bounds above would be independent of the break point t_i. Therefore, we have the desired upper bounds verified with probability at least $1 - c_1\exp(-c_2(\log(q) + 2\log(p)))$. □

Appendix B: Proof of Main Results

Proof of Theorem 1. The proof is similar to Theorem 2 in Safikhani and Shojaie (2020). For a matrix $A \in \mathbb{R}^{p\times pq}$, let $||A||_{1,\mathcal{I}_j} = \sum_{(k,h)\in\mathcal{I}_j} |a_{kh}|$.

First, we focus on the second part. Suppose there exists a true break point t_{j_0} where $j_0 \in \{1, \cdots, m_0\}$, which is isolated from all estimated points, i.e., $\min_{1\leq j\leq \widehat{m}} |\widehat{t}_j - t_{j_0}| > n\gamma_n$. In other words, there exists an estimated break point $\widehat{t}_j$ such that, $t_{j_0} - t_{j_0-1} \vee \widehat{t}_j \geq n\gamma_n$ and $t_{j_0+1} \wedge \widehat{t}_{j+1} - t_{j_0} \geq n\gamma_n$. The idea of the proof is to show the estimated AR parameter estimated in the interval $[t_{j_0-1} \vee \widehat{t}_j, t_{j_0+1} \wedge \widehat{t}_{j+1}]$ converges in L_2 to both $\Phi^{(.,j_0)}$ and $\Phi^{(.,j_0+1)}$ which contradicts Assumption A3. This is due to the fact that the length of the interval is large enough to verify restricted eigenvalue and deviation bound inequalities needed to show parameter estimation consistency.

Based on the definition of $\widehat{\Theta}$ in (4), the value of the function defined in (4) is minimized exactly at $\widehat{\Theta}$. This means that any other choice of parameters yields a higher value in (4). Denote the closest r_i to the right side of t_{j_0-1} plus q by s_{j_0-1}, i.e, $s_{j_0-1} = r_i + q$, and similarly, denote the closest r_i to the left side of t_{j_0} by s_{j_0}. First, we focus on the interval $[s_{j_0-1} \vee \widehat{t}_j, s_{j_0}]$. Define a new parameter sequence ψ_k's, $k = 1, ..., k_n$ with $\psi_k = \widehat{\theta}_k$ except for two time points $k = \widehat{i}_j \vee i_{j_0-1}$ and $k = i_{j_0}$. For these two points, if $\widehat{t}_j > s_{j_0-1}$, set $\psi_{\widehat{i}_j} = \Phi^{(.,j_0)} - \widehat{\Phi}_j$ and $\psi_{i_{j_0}} = \widehat{\Phi}_{j+1} - \Phi^{(.,j_0)}$; if $\widehat{t}_j \leq s_{j_0-1}$, set set $\psi_{i_{j_0-1}} = \Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}$ and $\psi_{i_{j_0}} = \widehat{\Phi}_{j+1} - \Phi^{(.,j_0)}$, where $\widehat{\Phi}_j = \sum_{k=1}^{\widehat{i}_j-1}\widehat{\theta}_k$ and $\widehat{\Phi}_{j+1} = \sum_{k=1}^{\widehat{i}_j}\widehat{\theta}_k$, i.e. $\widehat{\theta}_{\widehat{i}_j} = \widehat{\Phi}_{j+1} - \widehat{\Phi}_j$,

where i_{j_0-1}, i_{j_0} and $\widehat{i}_j$ are the corresponding indices of candidate points s_{j_0-1}, s_{j_0} and $\widehat{t}_j$. Denoting $\Psi = \text{vector}(\psi_1, ..., \psi_{k_n}) \in \mathbb{R}^{\pi_b \times 1}$, we have

$$\begin{aligned} \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Theta}}\|_2^2 + \lambda_{1,n}\|\widehat{\mathbf{\Theta}}\|_1 + \lambda_{2,n}\sum_{i=1}^{k_n}\left\|\sum_{j=1}^{i}\widehat{\theta}_j\right\|_1 &\leq \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\mathbf{\Psi}\|_2^2 + \lambda_{1,n}\|\mathbf{\Psi}\|_1 \\ &+ \lambda_{2,n}\sum_{i=1}^{k_n}\left\|\sum_{j=1}^{i}\psi_j\right\|_1. \end{aligned} \tag{B.1}$$

When $\widehat{t}_j > s_{j_0-1}$, some rearrangement of equation (B.1) leads to

$$\begin{aligned} 0 &\leq c\,\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_F^2 \\ &\leq \frac{1}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\sum_{l=s_{j_0-1}\vee\widehat{t}_j}^{s_{j_0}-1}\left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right)' Y_{l-1}Y_{l-1}'\left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right) \\ &\leq \frac{2}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\sum_{l=s_{j_0-1}\vee\widehat{t}_j}^{s_{j_0}-1} Y_{l-1}'\left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right)\varepsilon_l \\ &+ \frac{n\lambda_{1,n}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j}\left(\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \|\Phi^{(.,j_0)} - \widehat{\Phi}_j\|_1 - \|\widehat{\Phi}_{j+1} - \widehat{\Phi}_j\|_1\right) \\ &+ \frac{n\lambda_{2,n}}{b_n}\left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\ &\leq \left(\frac{2n\lambda_{1,n}}{s_{j_0} - s_{j_0-1} \vee \widehat{t}_j} + C\sqrt{\frac{\log p}{n\gamma_n}}\right)\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \frac{n\lambda_{2,n}}{b_n}\left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\ &\leq \frac{1}{2}\frac{n\lambda_{2,n}}{b_n}\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \frac{n\lambda_{2,n}}{b_n}\left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\ &\leq \frac{3}{2}\frac{n\lambda_{2,n}}{b_n}\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}_{j_0}} - \frac{1}{2}\frac{n\lambda_{2,n}}{b_n}\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}_{j_0}^c}. \end{aligned} \tag{B.2}$$

When $\widehat{t}_j \leq s_{j_0-1}$, a similar result can be obtained using an analogous rearrangement of equation (B.1). In equation (B.2), the second inequality holds with high probability converging to 1 due to first part of Lemma A1 and the fact that $s_{j_0} - s_{j_0-1} \vee \widehat{t}_j \geq \frac{1}{2}n\gamma_n$ and $b_n \leq \frac{1}{4}n\gamma_n$ by Assumption A3. The fourth inequality holds with high probability converging to 1 due to second part of Lemma A1 and triangular inequality. The fifth inequality is based on Assumption A3 and the selection for $\lambda_{2,n}$ in the statement of the Theorem. The last inequality holds by the sparsity assumption. This implies that

$$\left\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{n\gamma_n}}\right), \tag{B.3}$$

which means that $\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_F$ converges to zero in probability based on Assumption A3. Similarly, the same procedure can be applied to the interval $[s_{j_0}, s_{j_0+1} \wedge \widehat{t}_{j+1}]$, where s_{j_0+1} is the closest r_i to the left side of t_{j_0+1}, which leads to $\|\Phi^{(.,j_0+1)} - \widehat{\Phi}_{j+1}\|_F$ converges to zero in probability as well. This yields a contradiction to Assumption A3, and therefore, the proof is complete.

The proof of the first part is similar to the second part. Hence, a brief sketch is provided. Assume $|\widehat{\mathcal{A}}_n| < m_0$. This means there exist an isolated true break point, say t_{j_0}. More specifically, there exists an estimated break point $\widehat{t}_j$ such that, $t_{j_0} - t_{j_0-1} \vee \widehat{t}_j \geq n\gamma_n/3$ and $t_{j_0+1} \wedge \widehat{t}_{j+1} - t_{j_0} \geq n\gamma_n/3$. Now, similar arguments as explained in details in the second part can be applied to both intervals $[s_{j_0-1} \vee \widehat{t}_j, s_{j_0}]$ and $[s_{j_0}, s_{j_0+1} \wedge \widehat{t}_{j+1}]$ which leads to $\left\|\Phi^{(.,j_0+1)} - \Phi^{(.,j_0)}\right\|_F$ converges to zero and therefore contradicts Assumption A3. This completes the proof.

□

Proof of Theorem 2. Theorem 1 implies that for each t_i, $i = 1, \ldots, m_0$, there exist points $\widehat{t}_j \in \widehat{\mathcal{A}}_n$ such that $|\widehat{t}_j - t_i| \leq n\gamma_n$. First, we show that

$$\left\{\widehat{t}_j \in \widehat{\mathcal{A}}_n \text{ s.t. } |\widehat{t}_j - t_i| \leq n\gamma_n, \text{for some } i = 1, \ldots, m_0\right\} \subseteq \widetilde{\mathcal{A}}_n$$

with high probability converging to one as n tends to infinity. This shows that (a) $\mathbb{P}(\widetilde{m} < m_0) \to 0$; (b) $\mathbb{P}\left(d_H\left(\widetilde{\mathcal{A}}_n, \mathcal{A}_n\right) \leq n\gamma_n\right) \to 1$. We show this by contradiction. Suppose for some j, $\widehat{t}_j$ is not included in the optimal set $\widetilde{\mathcal{A}}_n$ which is a minimizer of (10). We show that adding $\widehat{t}_j$ costs less than removing it based on the LIC price.

Consider the case when $\widehat{t}_j$ is dropped in the screening step, while there exists a true change point t_i such that $|\widehat{t}_j - t_i| \leq n\gamma_n$. Similar arguments as in Proposition 4.1 of Basu and Michailidis (2015) show that the tuning parameter selected based on Assumption A5 (i.e. $\eta_{\widehat{t}_j} = c_1\sqrt{\frac{d_n^\star n\gamma_n}{a_n}} + c_2 d_n^\star$) yield: $\left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j}\right\|_F$ and $\left\|\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j}\right\|_F$. Note that in this case, the restricted eigenvalue condition does not hold. Therefore, the convergence of the $\widehat{\psi}_{\widehat{t}_j}$ cannot be verified. For this scenario, based on similar calculations as in case (c) of proof of Lemma 4 in Safikhani and Shojaie (2020) (see equations (22)-(26) in Safikhani and Shojaie (2020)'s supplement for details), after replacing s_{i-1} and s_i with $\widehat{t}_j - a_n$ and $\widehat{t}_j + a_n - 1$, we have the following lower bound:

$$\sum_{t=\widehat{t}_j - a_n}^{\widehat{t}_j + a_n - 1} \|y_t - \widehat{\psi}_{\widehat{t}_j} Y_{t-1}\|_2^2 \geq \sum_{t=\widehat{t}_j - a_n}^{\widehat{t}_j + a_n - 1} \|\varepsilon_t\|_2^2 + c_1 a_n - c_2 d_n^\star \log p, \tag{B.4}$$

for some positive c_1 and c_2 with high probability converging to one.

Denote by $B = \widetilde{\mathcal{A}}_n \cup \{\widehat{t}_j\}$. We only consider one segment $\left[\widehat{t}_j - a_n, \widehat{t}_j + a_n\right)$ for the case when $\widehat{t}_j$ is kept as a selected break point. Similar arguments as in Lemma 4 of Safikhani and Shojaie (2020) yield:

$$\left\|\widehat{\psi}_{\widehat{t}_j,1} - \Phi^{(.,i)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widehat{\psi}_{\widehat{t}_j,1} - \Phi^{(.,i)}\right\|_F, \ \left\|\widehat{\psi}_{\widehat{t}_j,1} - \Phi^{(.,i)}\right\|_F = O_p\left(d_n^\star\sqrt{\frac{n\gamma_n}{a_n}}\right), \tag{B.5}$$

and

$$\left\|\widehat{\psi}_{\widehat{t}_j,2} - \Phi^{(.,i+1)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widehat{\psi}_{\widehat{t}_j,2} - \Phi^{(.,i+1)}\right\|_F, \ \left\|\widehat{\psi}_{\widehat{t}_j,2} - \Phi^{(.,i+1)}\right\|_F = O_p\left(d_n^\star\sqrt{\frac{n\gamma_n}{a_n}}\right). \tag{B.6}$$

To see this, suppose $t_i + q < \widehat{t}_j$ (the other case is similar). Observe that $\widehat{\psi}_{\widehat{t}_j,1}$ in (6) minimizes the least squares plus the ℓ_1 norm loss function. Therefore, the value of this objective function for $\widehat{\psi}_{\widehat{t}_j,1}$ will be smaller than any other choice of parameters, including $\Phi^{(.,i)}$. Hence,

$$\begin{aligned}\frac{1}{a_n}\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}\left\|y_t-\widehat{\psi}_{\widehat{t}_j,1}Y_{t-1}\right\|_2^2+\eta_{\widehat{t}_j,1}\left\|\widehat{\psi}_{\widehat{t}_j,1}\right\|_1 &\leq \frac{1}{a_n}\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}\left\|y_t-\Phi^{(.,i)}Y_{t-1}\right\|_2^2 \\ &+ \eta_{\widehat{t}_j,1}\left\|\Phi^{(.,i)}\right\|_1. \qquad \text{(B.7)}\end{aligned}$$

Some rearrangements together with the use of proposition 4.2 of Basu and Michailidis (2015) lead to:

$$\begin{aligned}c_1\left\|\Phi^{(.,i+1)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_F^2-c_2\frac{\log p}{a_n}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1^2 &\leq \max\left(0,c_1\left\|\Phi^{(.,i+1)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_F^2-c_2\frac{\log p}{a_n}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1^2\right) \\ &\leq \frac{1}{a_n}\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}Y_{t-1}'\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)'\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)Y_{t-1} \\ &\leq \frac{2}{a_n}\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}Y_{t-1}'\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)'\left(y_t-\Phi^{(.,i)}Y_{t-1}\right)+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right) \\ &= \frac{2}{a_n}\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}Y_{t-1}'\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)'\varepsilon_t+\eta_{\widehat{t}_j,,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\psi}_{j,1}\right\|_1\right) \\ &\quad+\frac{2}{a_n}\sum_{t=t_i}^{\widehat{t}_j-1}Y_{t-1}'\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)'\left(\Phi^{(.,i+1)}-\Phi^{(.,i)}\right)Y_{t-1} \\ &\leq c\sqrt{\frac{\log p}{a_n}}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right) \\ &\quad+2\frac{\widehat{t}_j-t_i}{a_n}\mathrm{tr}\left(\left(\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right)'\left(\frac{1}{\widehat{t}_j-t_i}\sum_{t=t_i}^{\widehat{t}_j-1}Y_{t-1}Y_{t-1}'-\Gamma_{i+1}^q(0)+\Gamma_{i+1}^q(0)\right)\left(\Phi^{(.,i+1)}-\Phi^{(.,i)}\right)\right) \\ &= \mathrm{I},\end{aligned}$$

where the second inequality is due to the restricted eigenvalue property (proposition 4.2 of Basu and Michailidis (2015)), the fourth inequality uses the circular invariance property of the trace function. Now, observe that

$$
\begin{aligned}
\mathrm{I} &\leq c\sqrt{\frac{\log p}{a_n}}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right)\\
&+ 2c'\frac{\widehat{t}_j-t_i}{a_n}\left(\max\left(\sqrt{\frac{\log p}{\widehat{t}_j-t_i}},\frac{\log p}{\widehat{t}_j-t_i}\right)+\lambda_{\max}\left(\Gamma^q_{i+1}(0)\right)\right)\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\left\|\Phi^{(.,i+1)}-\Phi^{(.,i)}\right\|_1\\
&\leq \left(c\sqrt{\frac{\log p}{a_n}}+4c'M_\Phi d_n^\star\frac{n\gamma_n}{a_n}\right)\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\theta}\right\|_1\right)\\
&\leq \max\left(c,4c'M_\Phi\right)\sqrt{\frac{d_n^\star n\gamma_n}{a_n}}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\theta}\right\|_1\right)\\
&\leq \frac{\eta_{\widehat{t}_j,1}}{2}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1+\eta_{\widehat{t}_j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right)\\
&\leq \frac{3\eta_{\widehat{t}_j,1}}{2}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_{1,\mathcal{I}_i}-\frac{\eta_{\widehat{t}_j,1}}{2}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_{1,\mathcal{I}_i^c}\\
&\leq 2\eta_{\widehat{t}_j,1}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1, \qquad \text{(B.8)}
\end{aligned}
$$

where the first inequality is based on the first part of Lemma A1 with $c_n=\widehat{t}_j-t_i$, the third inequality is due to Assumptions A3 and A4 under which $\log p\leq n\gamma_n\leq a_n$ and $a_n>cd_n^{\star\,3}n\gamma_n$, which lead to $\sqrt{\frac{\log p}{a_n}}\leq\sqrt{\frac{n\gamma_n}{a_n}}$ and $d_n^\star\frac{n\gamma_n}{a_n}\leq\sqrt{\frac{d_n^\star n\gamma_n}{a_n}}$; the fourth inequality is due the selection of the tuning parameter $\eta_{\widehat{t}_j,1}=c\sqrt{\frac{d_n^\star n\gamma_n}{a_n}}$; the fifth inequality is due the triangular inequality.

This ensures that $\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_{1,\mathcal{I}_i^c}\leq 3\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_{1,\mathcal{I}_i}$, and hence $\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_1\leq 4\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_{1,\mathcal{I}_i}\leq 4\sqrt{d_n^\star}\left\|\Phi^{(.,i)}-\widehat{\psi}_{\widehat{t}_j,1}\right\|_F$. This comparison between L_1 and L_2 norms of the error term together with the bound in Equation (B.8) will get the desired consistency rates in (B.5). Verifying (B.6) follows exactly from proposition 4.1 of Basu and Michailidis (2015) since there are no break points in the interval $[\widehat{t}_j,\widehat{t}_j+a_n]$. Note that the convergence rate here is of order $\sqrt{\frac{d_n^\star n\gamma_n}{a_n}}$ instead of $\sqrt{d_n^\star}\sqrt{\frac{\log p}{a_n}}$ due to the different selection of tuning parameter in our setup compared to Basu and Michailidis (2015). In fact, one could select $\eta_{\widehat{t}_j,2}$ to be of order $\sqrt{\frac{\log p}{a_n}}$ in order to match the rate here with Basu and Michailidis (2015). However, since we don't know whether the true break point t_i will be on the left hand side or the right hand side of $\widehat{t}_j$, it's better for the tuning parameters $\eta_{\widehat{t}_j,1}$ and $\eta_{\widehat{t}_j,2}$ to have the same order from a practical viewpoint. This justifies setting them to be of the same order here.

Using the results of (B.5) and (B.6), one can find upper bounds for the sum of squared errors in the segment $\left(\widehat{t}_j-a_n,\widehat{t}_j+a_n\right)$ for the case when $\widehat{t}_j$ is kept as a selected break point. Specifically, we have:

$$
\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}\left\|y_t-\widehat{\psi}_{\widehat{t}_j,1}Y_{t-1}\right\|_2^2\leq\sum_{t=\widehat{t}_j-a_n}^{\widehat{t}_j-1}\left\|\varepsilon_t\right\|_2^2+O_p\left(d_n^{\star\,3}n\gamma_n\right), \qquad \text{(B.9)}
$$

and

$$\sum_{t=\widehat{t}_j}^{\widehat{t}_j+a_n-1} \|y_t - \widehat{\psi}_{\widehat{t}_j,2} Y_{t-1}\|_2^2 \leq \sum_{t=\widehat{t}_j}^{\widehat{t}_j+a_n-1} \|\varepsilon_t\|_2^2 + O_p\left(d_n^{\star\,3} n\gamma_n\right). \tag{B.10}$$

To see (B.9), we follow the steps of the proof of Theorem 3 in Safikhani and Shojaie (2020). Observe that:

$$\begin{aligned}
\sum_{t=\widehat{t}_j-a_n}^{t_i-1} \|y_t - \widehat{\psi}_{\widehat{t}_j,1} Y_{t-1}\|_2^2 &\leq \sum_{t=\widehat{t}_j-a_n}^{t_i-1} \|\varepsilon_t\|_2^2 + c\sqrt{(t_i - \widehat{t}_j + a_n)\log p}\left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1 \\
&+ (t_i - \widehat{t}_j + a_n)\mathrm{tr}\left(\left(\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right)'\left(\frac{1}{t_i - \widehat{t}_j + a_n}\sum_{t=\widehat{t}_j-a_n}^{t_i-1} Y_{t-1}Y_{t-1}' - \Gamma_i^q(0) + \Gamma_i^q(0)\right)\left(\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right)\right) \\
&\leq \sum_{t=\widehat{t}_j-a_n}^{t_i-1} \|\varepsilon_t\|_2^2 + cd_n^{\star\,2}\sqrt{n\gamma_n \log p} + (t_i - \widehat{t}_j + a_n)\left(\sqrt{\frac{\log p}{t_i - \widehat{t}_j + a_n}} + \lambda_{\max}\left(\Gamma_i^q(0)\right)\right)\left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1^2 \\
&\leq \sum_{t=\widehat{t}_j-a_n}^{t_i-1} \|\varepsilon_t\|_2^2 + cd_n^{\star\,2}\sqrt{n\gamma_n \log p} + c' d_n^{\star\,3} n\gamma_n \\
&= \sum_{t=\widehat{t}_j-a_n}^{t_i-1} \|\varepsilon_t\|_2^2 + O_p\left(d_n^{\star\,3} n\gamma_n\right),
\end{aligned} \tag{B.11}$$

where the second inequality is based on the fact that $\left\|\widehat{\psi}_{\widehat{t}_j,1} - \Phi^{(.,i)}\right\|_1 \leq cd_n^{\star\,3/2}\sqrt{\frac{n\gamma_n}{a_n}}$ and $t_i - \widehat{t}_j + a_n \leq a_n$. Also, for the segment $[t_i, \widehat{t}_j]$, note that:

$$\begin{aligned}
\sum_{t=t_i}^{\widehat{t}_j-1} \|y_t - \widehat{\psi}_{\widehat{t}_j,1} Y_{t-1}\|_2^2 &\leq \sum_{t=t_i}^{\widehat{t}_j-1} \|\varepsilon_t\|_2^2 + c(\widehat{t}_j - t_i)\max\left(\frac{\log p}{\widehat{t}_j - t_i}, \sqrt{\frac{\log p}{\widehat{t}_j - t_i}}\right)\left\|\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1 \\
&+ (\widehat{t}_j - t_i)\mathrm{tr}\left(\left(\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j,1}\right)'\left(\frac{1}{\widehat{t}_j - t_i}\sum_{t=t_i}^{\widehat{t}_j-1} Y_{t-1}Y_{t-1}' - \Gamma_{i+1}^q(0) + \Gamma_{i+1}^q(0)\right)\left(\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j,1}\right)\right) \\
&\leq \sum_{t=t_i}^{\widehat{t}_j-1} \|\varepsilon_t\|_2^2 + cn\gamma_n\left(2M_\Phi d_n^\star + \left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right) \\
&+ (\widehat{t}_j - t_i)\left(\max\left(\frac{\log p}{\widehat{t}_j - t_i}, \sqrt{\frac{\log p}{\widehat{t}_j - t_i}}\right) + \lambda_{\max}\left(\Gamma_{i+1}^q(0)\right)\right)\left\|\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1^2 \\
&\leq \sum_{t=t_i}^{\widehat{t}_j-1} \|\varepsilon_t\|_2^2 + cd_n^\star n\gamma_n + c' n\gamma_n\left(2M_\Phi d_n^\star + \left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1\right)^2 \\
&= \sum_{t=t_i}^{\widehat{t}_j-1} \|\varepsilon_t\|_2^2 + O_p\left(d_n^{\star\,2} n\gamma_n\right).
\end{aligned} \tag{B.12}$$

where the second inequality is based on $\widehat{t}_j - t_i \leq n\gamma_n$, $\log p \leq n\gamma_n$ and triangular inequality $\left\|\Phi^{(.,i+1)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1 \leq \left\|\Phi^{(.,i+1)} - \Phi^{(.,i)}\right\|_1 + \left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1 \leq 2M_\Phi d_n^\star + \left\|\Phi^{(.,i)} - \widehat{\psi}_{\widehat{t}_j,1}\right\|_1$; the second term in the third inequality is based on Assumption A4 of $a_n > cd_n^{\star\,3} n\gamma_n$ so that $d_n^\star \geq d_n^{\star\,3/2}\sqrt{\frac{n\gamma_n}{a_n}}$.

Property (B.9) is now verified by combining (B.11) and (B.12). Property (B.10) is similar and therefore its proof is omitted.

Recall that $B = \{\widehat{t}_1, \ldots, \widehat{t}_{m_0}\} = \widetilde{\mathcal{A}}_n \cup \{\widehat{t}_j\}$. Applying the results (B.4), (B.9) and (B.10), we have

$$\begin{aligned}
& \mathrm{LIC}(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}) - \mathrm{LIC}(B) \\
= \; & L_n(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}; \eta_n) + \widetilde{m}\omega_n - \left(L_n(\widehat{t}_1, ..., \widehat{t}_{m_0}; \eta_n) + (\widetilde{m}+1)\omega_n\right) \\
= \; & \sum_{t=\widehat{t}_j - a_n}^{\widehat{t}_j + a_n - 1} \|y_t - \widehat{\psi}_j Y_{t-1}\|_2^2 - \sum_{t=\widehat{t}_j - a_n}^{\widehat{t}_j - 1} \|y_t - \widehat{\psi}_{\widehat{t}_j,1} Y_{t-1}\|_2^2 - \sum_{t=\widehat{t}_j}^{\widehat{t}_j + a_n - 1} \|y_t - \widehat{\psi}_{\widehat{t}_j,2} Y_{t-1}\|_2^2 - \omega_n \\
\geq \; & c_1 a_n - c_2 d_n^\star \log p - c_3 n\gamma_n {d_n^\star}^3 - \omega_n \\
\geq \; & c_1 a_n - 2c_2 n\gamma_n {d_n^\star}^3 - \omega_n \\
> \; & 0.
\end{aligned}$$

The second inequality hold due to condition $\log p \leq n\gamma_n$ in Assumption A3. The last inequality holds due to conditions $\omega_n = n\gamma_n {d_n^\star}^3$ and $\lim_{n\to\infty} \omega_n / a_n = 0$ in Assumption A4. This proves that $\mathbb{P}\left(\{\widehat{t}_1, \ldots, \widehat{t}_{m_0}\} \subseteq \widetilde{\mathcal{A}}_n\right) \to 1$.

It remains to establish that $\mathbb{P}\left(d_H\left(\mathcal{A}_n, \widetilde{\mathcal{A}}_n\right) \leq a_n\right) \to 1$. It suffices to show that there exists a true break point in the interval $\left(\widetilde{t}_j - a_n, \widetilde{t}_j + a_n\right)$ for any $\widetilde{t}_j \in \widetilde{\mathcal{A}}_n$. We prove this statement by contradiction. Suppose that there is a time point $s \in \widetilde{\mathcal{A}}_n$, such that there exists no true break point in the interval $(s - a_n, s + a_n)$. Next, define a new set $C = \widetilde{\mathcal{A}}_n \backslash \{s\}$.

Since there is no true break point in the segment, $(s - a_n, s + a_n)$, it belongs to some stationary segment, say the i-th one. Select tuning parameter $\eta_s = \eta_{s,1} = \eta_{s,2} = c\sqrt{\frac{\log p}{a_n}}$. Denote the estimated AR parameters in $[s - a_n, s)$ by $\widehat{\psi}_{s,1}$ and the estimated AR parameters in $[s, s + a_n)$ by $\widehat{\psi}_{s,2}$. Keeping the point s yields to :

$$\left\|\widehat{\psi}_{s,1} - \Phi^{(.,i)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widehat{\psi}_{s,1} - \Phi^{(.,i)}\right\|_F, \; \left\|\widehat{\psi}_{s,1} - \Phi^{(.,i)}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{a_n}}\right), \tag{B.13}$$

and

$$\left\|\widehat{\psi}_{s,2} - \Phi^{(.,i)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widehat{\psi}_{s,2} - \Phi^{(.,i)}\right\|_F, \; \left\|\widehat{\psi}_{s,2} - \Phi^{(.,i)}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{a_n}}\right). \tag{B.14}$$

Therefore, we have the following lower bound:

$$\sum_{t=s-a_n}^{s-1} \|y_t - \widehat{\psi}_{s,1} Y_{t-1}\|_2^2 + \sum_{t=s}^{s+a_n-1} \|y_t - \widehat{\psi}_{s,2} Y_{t-1}\|_2^2 \geq \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 - c_1 d_n^\star n\gamma_n, \tag{B.15}$$

for some positive c_1 with high probability converging to one.

To see (B.15), we follow the case (b) in the proof of Lemma 4 in Safikhani and Shojaie (2020). For interval $[s-a_n, s)$,

$$\begin{aligned}
\sum_{t=s-a_n}^{s-1} \|y_t - \widehat{\psi}_{s,1} Y_{t-1}\|_2^2 &\geq \sum_{t=s-a_n}^{s-1} \|\varepsilon_t\|_2^2 + ca_n \left\|\Phi^{(.,i)} - \widehat{\psi}_{s,1}\right\|_2^2 - c'\sqrt{a_n \log p}\left\|\Phi^{(.,i)} - \widehat{\psi}_{s,1}\right\|_1 \\
&\geq \sum_{t=s-a_n}^{s-1} \|\varepsilon_t\|_2^2 + ca_n \left\|\Phi^{(.,i)} - \widehat{\psi}_{s,1}\right\|_2 \left(\left\|\Phi^{(.,i)} - \widehat{\psi}_{s,1}\right\|_2 - \frac{4c'}{c}\sqrt{\frac{d_n^\star \log p}{a_n}}\right) \\
&\geq \sum_{t=s-a_n}^{s-1} \|\varepsilon_t\|_2^2 - c'' d_n^\star \log p.
\end{aligned}$$

Similarly, for interval $[s, s+a_n)$, we have $\sum_{t=s}^{s+a_n-1} \|y_t - \widehat{\psi}_2 Y_{t-1}\|_2^2 \geq \sum_{t=s}^{s+a_n-1} \|\varepsilon_t\|_2^2 - c'' d_n^\star \log p$.

Next, we remove the break point s and denote the estimated parameters in $[s-a_n, s+a_n)$ by $\widehat{\psi}_s$. We get

$$\sum_{t=s-a_n}^{s+a_n-1} \|y_t - \widehat{\psi}_s Y_{t-1}\|_2^2 \leq \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 + c_2 {d_n^\star}^2 \log p \tag{B.16}$$

To see (B.16), we again follow the steps of the proof of Theorem 3 in Safikhani and Shojaie (2020):

$$\begin{aligned}
\sum_{t=s-a_n}^{s+a_n-1} \|y_t - \widehat{\psi}_s Y_{t-1}\|_2^2 &\leq \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 + c\sqrt{(2a_n)\log p}\left\|\Phi^{(.,i)} - \widehat{\psi}_s\right\|_1 \\
&+ (2a_n)\mathrm{tr}\left(\left(\Phi^{(.,i)} - \widehat{\psi}_s\right)' \left(\frac{1}{2a_n}\sum_{t=s-a_n}^{s+a_n-1} Y_{t-1}Y_{t-1}' - \Gamma_i^q(0) + \Gamma_i^q(0)\right)\left(\Phi^{(.,i)} - \widehat{\psi}_s\right)\right) \\
&\leq \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 + c d_n^\star \log p + (2a_n)\left(\sqrt{\frac{\log p}{2a_n}} + \lambda_{\max}\left(\Gamma_i^q(0)\right)\right)\left\|\Phi^{(.,i)} - \widehat{\psi}_s\right\|_1^2 \\
&\leq \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 + c d_n^\star \log p + c' {d_n^\star}^2 \log p \\
&= \sum_{t=s-a_n}^{s+a_n-1} \|\varepsilon_t\|_2^2 + O_p\left({d_n^\star}^2 \log p\right).
\end{aligned}$$

where $\left\|\widehat{\psi}_s - \Phi^{(.,i)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widehat{\psi}_s - \Phi^{(.,i)}\right\|_2$, $\left\|\widehat{\psi}_s - \Phi^{(.,i)}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{a_n}}\right)$,

Combining (B.15) and (B.16) yields

$$\begin{aligned}
0 &\geq \mathrm{LIC}(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}) - \mathrm{LIC}(C) \\
&\geq -c_1 d_n^\star \log p - c_2 {d_n^\star}^2 \log p + \omega_n \\
&\geq -2c_2 {d_n^\star}^2 \log p + \omega_n
\end{aligned} \tag{B.17}$$

However, (B.17) contradicts Assumption A4 that ${d_n^\star}^2 \log p/\omega_n \to 0$. This completes the proof of the second part.

It remains to show that $\mathbb{P}\left(\left|\mathrm{cluster}\left(\widetilde{\mathcal{A}}_n, 2a_n\right)\right| = m_0\right) \to 1$. For this, note that based on the previous calculations, all points in $\widetilde{\mathcal{A}}_n$ are either very close to a true break point,

i.e. in the $Kn\gamma_n$-neighborhood of a true break point for some positive and finite K, or they are at least in the ca_n-neighborhood of a true break point for a constant $0 < c < 1$. Next, fix one of the true break points, say t_j. Denote by B_j, all the selected break points in $\widetilde{\mathcal{A}}_n$ which are in the $Kn\gamma_n$-neighborhood of t_j for some positive and finite K. Now, if there are any points in the a_n-neighborhood of t_j, add them to B_j as well. Note that since there is at least one estimated break point in the $n\gamma_n$-neighborhood of t_j, B_j is not empty. Also, note that the diameter of each B_j is at most $2a_n$. Finally, based on the choice of a_n in Assumption A4, all B_j's are disjoint, $j = 1, \ldots, m_0$. Therefore, the collection of all B_j's form cluster $\left(\widetilde{\mathcal{A}}_n, 2a_n\right)$ which has cardinality equal to m_0 (note that due to Assumption A4, the collection of sets B_j's is a minimal partitioning of set $\widetilde{\mathcal{A}}_n$ while the diameter of all B_j's are at most $2a_n$). This completes the proof of the last part.

□

Proof of Theorem 3. Suppose for any constant $K > 0$, there exists some change point $t_j \in \mathcal{A}_n$ such that $\left|\widetilde{t}_j^f - t_j\right| > Kd_n^\star \log p$. From Theorem 2, we know that for any subset $B_j \in$ cluster $\left(\widetilde{\mathcal{A}}_n, 2a_n\right)$, there exists a true change point t_j lies in $(\min(B_j) - a_n, \max(B_j) + a_n)$. Therefore, there exists a true change point t_j lying within interval $[\widetilde{t}_j^f - a_n, \widetilde{t}_j^f + a_n)$. Assume, without loss of generality, that $\widetilde{t}_j^f > t_j$.

Based on the local refinement algorithm, we define the loss function $L_n\left(\widetilde{t}_j^f\right)$ as follows:

$$\sum_{t=\min(B_j)-a_n}^{\widetilde{t}_j^f-1} \left\|y_t - \widetilde{\psi}_{j,1} Y_{t-1}\right\|_2^2 + \sum_{t=\widetilde{t}_j^f}^{\max(B_j)+a_n-1} \left\|y_t - \widetilde{\psi}_{j,2} Y_{t-1}\right\|_2^2 \stackrel{\text{def}}{=} I_1 + I_2,$$

and the loss function $L_n(t_j)$ for the true change point t_j as:

$$\sum_{t=\min(B_j)-a_n}^{t_j-1} \left\|y_t - \widetilde{\psi}_{j,1} Y_{t-1}\right\|_2^2 + \sum_{t=t_j}^{\max(B_j)+a_n-1} \left\|y_t - \widetilde{\psi}_{j,2} Y_{t-1}\right\|_2^2 \stackrel{\text{def}}{=} I_3 + I_4.$$

where $\widetilde{t}_j^f \in (l_j, u_j)$, $l_j = \min(B_j) - a_n$ and $u_j = \max(B_j) + a_n$, for $j = 1, \ldots, m_0 + 1$. $\widetilde{\psi}_{j,1}$ and $\widetilde{\psi}_{j,2}$ are the local coefficient parameter estimates given that a time point $s \in B_j$ is the break point.

Suppose $s > t_j$; then, similar arguments as in Lemma 4 of Safikhani and Shojaie (2020) yield:

$$\left\|\widetilde{\psi}_{j,1} - \Phi^{(.,j)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widetilde{\psi}_{j,1} - \Phi^{(.,j)}\right\|_F, \ \left\|\widetilde{\psi}_{j,1} - \Phi^{(.,j)}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{\widetilde{R}_n}}\right), \quad \text{(B.18)}$$

and

$$\left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_1 \leq 4\sqrt{d_n^\star}\left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_F, \ \left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_F = O_p\left(\sqrt{\frac{d_n^\star \log p}{\widetilde{R}_n}}\right). \quad \text{(B.19)}$$

To see this, observe that $\widetilde{\psi}_{j,1}$ in (12) minimizes the least squares plus the ℓ_1 norm loss function. Therefore, the value of this objective function for $\widetilde{\psi}_{j,1}$ will be smaller than any other choice of parameters, including $\Phi^{(.,j)}$. Hence,

$$\begin{aligned}\frac{1}{\widetilde{R}_n}\sum_{t=\widetilde{t}_j^f-\widetilde{R}_n}^{\widetilde{t}_j^f-1}\left\|y_t-\widetilde{\psi}_{j,1}Y_{t-1}\right\|_2^2+\widetilde{\eta}_{j,1}\left\|\widetilde{\psi}_{j,1}\right\|_1 &\leq \frac{1}{\widetilde{R}_n}\sum_{t=\widetilde{t}_j^f-\widetilde{R}_n}^{\widetilde{t}_j^f-1}\left\|y_t-\Phi^{(.,j)}Y_{t-1}\right\|_2^2 \\ &+ \widetilde{\eta}_{j,1}\left\|\Phi^{(.,j)}\right\|_1. \qquad \text{(B.20)}\end{aligned}$$

Note that the segment $[t_j,\widetilde{t}_j^f)$ has at most length a_n, so the length of $[\widetilde{t}_j^f-\widetilde{R}_n,t_j)$ is at least $c\widetilde{R}_n$, for some positive constant $c>0$. In that case, the misspecification part is negligible. Some rearrangements together with the use of proposition 4.2 of Basu and Michailidis (2015) lead to:

$$\begin{aligned}
c_1\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_F^2-c_2\frac{\log p}{\widetilde{R}_n}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1^2 &\leq \max\left(0,c_1\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_F^2-c_2\frac{\log p}{\widetilde{R}_n}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1^2\right)\\
&\leq \frac{1}{\widetilde{R}_n}\sum_{t=\widetilde{t}_j^f-\widetilde{R}_n}^{\widetilde{t}_j^f-1}Y_{t-1}'\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)'\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)Y_{t-1}\\
&\leq \frac{2}{\widetilde{R}_n}\sum_{t=\widetilde{t}_j^f-\widetilde{R}_n}^{\widetilde{t}_j^f-1}Y_{t-1}'\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)'\left(y_t-\Phi^{(.,j)}Y_{t-1}\right)+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&= \frac{2}{\widetilde{R}_n}\sum_{t=\widetilde{t}_j^f-\widetilde{R}_n}^{\widetilde{t}_j^f-1}Y_{t-1}'\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)'\varepsilon_t+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&\quad+\frac{2}{\widetilde{R}_n}\sum_{t=t_j}^{\widetilde{t}_j^f-1}Y_{t-1}'\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)'\left(\Phi^{(.,j+1)}-\Phi^{(.,j)}\right)Y_{t-1}\\
&\leq c\sqrt{\frac{\log p}{\widetilde{R}_n}}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&\quad+2\frac{\widetilde{t}_j^f-t_j}{\widetilde{R}_n}\mathrm{tr}\left(\left(\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right)'\left(\frac{1}{\widetilde{t}_j^f-t_j}\sum_{t=t_j}^{\widetilde{t}_j^f-1}Y_{t-1}Y_{t-1}'-\Gamma_{j+1}^q(0)+\Gamma_{j+1}^q(0)\right)\left(\Phi^{(.,j+1)}-\Phi^{(.,j)}\right)\right)\\
&= \mathrm{I},
\end{aligned}$$

where the second inequality is due to the restricted eigenvalue property (proposition 4.2 of Basu and Michailidis (2015)) and the fourth inequality uses the circular invariance property of the trace function. Next, observe that

$$
\begin{aligned}
\mathrm{I} &\leq c\sqrt{\frac{\log p}{\widetilde{R}_n}}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&+ 2c'\frac{\widetilde{t}_j^f-t_j}{\widetilde{R}_n}\left(\max\left(\sqrt{\frac{\log p}{\widetilde{t}_j^f-t_j}},\frac{\log p}{\widetilde{t}_j^f-t_j}\right)+\lambda_{\max}\left(\Gamma_{j+1}^q(0)\right)\right)\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1\left\|\Phi^{(.,j+1)}-\Phi^{(.,j)}\right\|_1\\
&\leq \left(c\sqrt{\frac{\log p}{\widetilde{R}_n}}+4c'M_\Phi d_n^\star\frac{a_n}{\widetilde{R}_n}\right)\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,i)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&\leq \max\left(c,4c'M_\Phi\right)\sqrt{\frac{\log p}{\widetilde{R}_n}}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&\leq \frac{\widetilde{\eta}_{j,1}}{2}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1+\widetilde{\eta}_{j,1}\left(\left\|\Phi^{(.,j)}\right\|_1-\left\|\widetilde{\psi}_{j,1}\right\|_1\right)\\
&\leq \frac{3\widetilde{\eta}_{j,1}}{2}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_{1,\mathcal{I}_j}-\frac{\widetilde{\eta}_{j,1}}{2}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_{1,\mathcal{I}_j^c}\\
&\leq 2\widetilde{\eta}_{j,1}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1, \qquad \text{(B.21)}
\end{aligned}
$$

where the first inequality is based on the first part of Lemma A1 with $c_n=\widetilde{t}_j^f-t_j$; the third inequality is due to Assumption A3 and Assumption A4 under which $\log p\leq n\gamma_n\leq a_n$ and $\frac{\Delta_n}{4}\geq\widetilde{R}_n=\frac{d_n^{\star\,2}a_n^2}{\log p}$ such that $\sqrt{\frac{\log p}{\widetilde{R}_n}}\geq d_n^\star\frac{a_n}{\widetilde{R}_n}$; the fourth inequality is due the selection of the tuning parameter $\widetilde{\eta}_{j,1}=\sqrt{\frac{\log p}{\widetilde{R}_n}}$; and finally the fifth inequality is due the triangular inequality.

This ensures that $\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_{1,\mathcal{I}_j^c}\leq 3\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_{1,\mathcal{I}_j}$, and hence $\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_1\leq 4\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_{1,\mathcal{I}_j}\leq 4\sqrt{d_n^\star}\left\|\Phi^{(.,j)}-\widetilde{\psi}_{j,1}\right\|_F$. This comparison between L_1 and L_2 norms of the error term together with the bound in Equation (B.21) will get the desired consistency rates in (B.18). Verifying (B.19) follows exactly from proposition 4.1 of Basu and Michailidis (2015) since there are no break points in the interval $[\widetilde{t}_j^f,\widetilde{t}_j^f+\widetilde{R}_n)$.

Using the results from (B.18) and (B.19), we obtain

$$
\begin{aligned}
I_1 &= \sum_{t=\widetilde{t}_j^f - a_n}^{t_j-1} \left\| y_t - \widetilde{\psi}_{j,1} Y_{t-1} \right\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \left\| y_t - \widetilde{\psi}_{j,1} Y_{t-1} \right\|_2^2 \\
&\geq I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \left\| \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j+1)} \right) Y_{t-1} \right\|_2^2 - 2 \left| \sum_{t=t_j}^{\widetilde{t}_j^r-1} Y_{t-1}' \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j+1)} \right) \boldsymbol{\varepsilon}_t \right| \\
&= I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \left\| \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j)} + \Phi^{(.,j)} - \Phi^{(.,j+1)} \right) Y_{t-1} \right\|_2^2 \\
&\quad - 2 \left| \sum_{t=t_j}^{\widetilde{t}_j^r-1} Y_{t-1}' \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j)} + \Phi^{(.,j)} - \Phi^{(.,j+1)} \right) \boldsymbol{\varepsilon}_t \right| \\
&\geq I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \left\| \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right) Y_{t-1} \right\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \left\| \left(\Phi^{(.,j)} - \Phi^{(.,j+1)} \right) Y_{t-1} \right\|_2^2 \\
&\quad - 2 \left| \sum_{t=t_j}^{\widetilde{t}_j^r-1} Y_{t-1}' \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right)' \left(\Phi^{(.,j)} - \Phi^{(.,j+1)} \right) Y_{t-1} \right| - 2 \left| \sum_{t=t_j}^{\widetilde{t}_j^r-1} Y_{t-1}' \left(\widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right) \boldsymbol{\varepsilon}_t \right| \\
&\quad - 2 \left| \sum_{t=t_j}^{\widetilde{t}_j^r-1} Y_{t-1}' \left(\Phi^{(.,j)} - \Phi^{(.,j+1)} \right) \boldsymbol{\varepsilon}_t \right| \\
&\overset{\text{(i)}}{\geq} c' \left| \widetilde{t}_j^r - t_j \right| \left(\left\| \widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right\|_F^2 + \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F^2 \right) - c'' \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F \left\| \widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right\|_F \\
&\quad - c''' \sqrt{\left| \widetilde{t}_j^r - t_j \right| (\log p)} \left(\left\| \widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right\|_1 + \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_1 \right) + I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 \\
&\overset{\text{(ii)}}{\geq} I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + c' \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F^2 - c'' \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F \left\| \widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right\|_F \\
&\quad - c''' \sqrt{\left| \widetilde{t}_j^r - t_j \right| (\log p)} \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_1 \\
&\geq I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + C' \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F \left(\left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F - \left\| \widetilde{\psi}_{j,1} - \Phi^{(.,j)} \right\|_F \right) \\
&\quad - c''' \sqrt{\left| \widetilde{t}_j^r - t_j \right| (\log p)} \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_1 \\
&\overset{\text{(iii)}}{\geq} I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + c_1 \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F \left(\left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F - \sqrt{\frac{d_n^\star (\log p)}{\left| \widetilde{t}_j^r - t_j \right|}} \right) \\
&\geq I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + C \left| \widetilde{t}_j^r - t_j \right| \left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F \left(\left\| \Phi^{(.,j)} - \Phi^{(.,j+1)} \right\|_F - \sqrt{\frac{1}{K}} \right) \\
&\overset{\text{(iv)}}{\geq} I_3 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + K_1 \left| \widetilde{t}_j^r - t_j \right| . \qquad \text{(B.22)}
\end{aligned}
$$

The second term inequality in (i) holds by Hölder's inequality; the third term inequality

in (i) holds by the deviation bound condition; inequality (ii) holds by $\left\|\Phi^{(.,j+1)}-\Phi^{(.,j)}\right\|_1 \leq \sqrt{d_n^\star}\left\|\Phi^{(.,j+1)}-\Phi^{(.,j)}\right\|_F$; inequality (iii) holds by the fact that $\left\|\Phi^{(.,j+1)}-\Phi^{(.,j)}\right\|_F^2 \geq v^2$, where v is a positive constant; and inequality (iv) holds by $\left|\widetilde{t}_j^f - t_j\right| > K d_n^\star \log p$ and choosing large enough constant $K>0$ such that $\sqrt{\frac{1}{K}} < \frac{1}{4}$.

Similarly, we have

$$
\begin{aligned}
I_4 &= \sum_{t=t_j}^{\max(B_j)+a_n-1} \left\|y_t - \widetilde{\psi}_{j,2} Y_{t-1}\right\|_2^2 \\
&= \sum_{t=t_j}^{\widetilde{t}_j^f-1} \left\|y_t - \widetilde{\psi}_{j,2} Y_{t-1}\right\|_2^2 + \sum_{t=\widetilde{t}_j^f}^{\max(B_j)+a_n-1} \left\|y_t - \widetilde{\psi}_{j,2} Y_{t-1}\right\|_2^2 \\
&\leq I_2 + \sum_{t=t_j}^{\widetilde{t}_j^f-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + \sum_{t=t_j}^{\widetilde{t}_j^f-1} \left\|\left(\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right) Y_{t-1}\right\|_2^2 + 2\left|\sum_{t=t_j}^{\widetilde{t}_j^f-1} Y_{t-1}'\left(\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right)\boldsymbol{\varepsilon}_t\right| \\
&\overset{\text{(i)}}{\leq} I_2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + c' a_n \left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_F^2 + c''\sqrt{a_n(\log p)}\left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_1 \\
&\leq I_2 + \sum_{t=t_j}^{\widetilde{t}_j^r-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + c' a_n \left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_F \left(\left\|\widetilde{\psi}_{j,2} - \Phi^{(.,j+1)}\right\|_F + \frac{c''}{c}\sqrt{\frac{d_n^\star(\log p)}{a_n}}\right) \\
&\overset{\text{(ii)}}{\leq} I_2 + \sum_{t=t_j}^{\widetilde{t}_j^f-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + K_2 d_n^\star(\log p),
\end{aligned} \tag{B.23}
$$

where the second term in inequality (i) holds by the upper-RE condition and the fact that $\left\|\Phi^{(.,j+1)} - \widetilde{\psi}_{j,2}\right\|_1^2 \leq 16 d_n^\star \left\|\Phi^{(.,j+1)} - \widetilde{\psi}_{j,2}\right\|_F^2$; the third term in inequality (i) holds by deviation bound condition; inequality (ii) holds by the result (B.19). Note that based on Assumption A4 that $\frac{\Delta_n}{4} \geq \widetilde{R}_n = \frac{{d_n^\star}^2 a_n^2}{\log p}$, we have $\left\|\Phi^{(.,j+1)} - \widetilde{\psi}_{j,2}\right\|_F \leq c\sqrt{\frac{d_n^\star \log p}{a_n}}$.

Next, based on the definition of (11), we have $L_n\left(\widetilde{t}_j^f\right) \leq L_n(t_j)$ and then

$$
I_3 + \sum_{t=t_j}^{\widetilde{t}_j^f-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + K_1\left|\widetilde{t}_j^f - t_j\right| + I_2 \leq L_n(\widetilde{t}_j^f) \leq L_n(t_j) \leq I_3 + I_2 + \sum_{t=t_j}^{\widetilde{t}_j^f-1} \|\boldsymbol{\varepsilon}_t\|_2^2 + K_2 d_n^\star(\log p),
$$

which leads to

$$
\left|\widetilde{t}_j^f - t_j\right| \leq K_j d_n^\star \log p. \tag{B.24}
$$

This contradicts the setting. Under Assumptions A1 and A2, we set $K^\star = \max_{1\leq j\leq m_0} K_j$ and complete the proof.

□

Proof of Theorem 4. The proof of this theorem follows along similar lines to Theorem 4 of Safikhani and Shojaie (2020), where it is also shown that, if m_0 is known, then it is enough to set $R_n = n\gamma_n$.

We need to firstly verify two important conditions: (1) the restricted eigenvalue (RE) condition for $\hat{\Gamma} = I_p \otimes (\mathcal{X}_{\mathbf{r}}'\mathcal{X}_{\mathbf{r}}/N)$; and (2) the deviation bound condition for $\|\hat{\gamma} - \hat{\Gamma}\Phi\|_\infty$ where $\hat{\gamma} = (I_p \otimes \mathcal{X}_{\mathbf{r}}')\mathbf{Y}_{\mathbf{r}}/N$. Once these two are verified, the rest of the proof is applying deterministic arguments used in Proposition 4.1 in Basu and Michailidis (2015).

Condition (1) implies that there exist $\alpha, \tau > 0$ such that for any $\theta \in \mathbb{R}^{\tilde{\pi}}$, we have

$$\theta'\hat{\Gamma}\theta \geq \alpha\|\theta\|_2^2 - \tau\|\theta\|_1^2$$

with probability at least $1 - c_1\exp(-c_2 N)$ for large enough constants $c_1, c_2 > 0$. Based on Lemma B.1 in Basu and Michailidis (2015), it is enough to show the RE for $S = \mathcal{X}_i'\mathcal{X}_i/N$, where $\mathcal{X}_i$ is the ith block component of $\mathcal{X}_{\mathbf{r}}$. Applying Proposition 2.4 of Basu and Michailidis (2015), we have for any $\mathbb{R}^{pq}$ with $\|v\|_2 \leq 1$, and any $\eta > 0$:

$$\mathbb{P}\Big(\Big|v'\big(S - \frac{N_i}{N}\Gamma_i(0)\big)v\Big| > c\eta\Big) \leq 2\exp(-c_3 N\min(\eta^2, \eta)).$$

Next, to make the above probability hold uniformly over all vectors v, we apply the discretization Lemma F2 in Basu and Michailidis (2015) and also Lemma 12 in the supplement of Loh and Wainwright (2012) to get:

$$\Big|v'\big(S - \frac{N_i}{N}\Gamma_i(0)\big)v\Big| \leq \alpha\|v\|_2^2 + \alpha/k\|v\|_1^2,$$

with probability at least $1 - c_1\exp(c_2 N)$, for all $v \in \mathbb{R}^{pq}$, some $\alpha > 0$ and with an integer $k = \lceil c_4 N/\log(pq)\rceil$ with some $c_4 > 0$. This implies that

$$v'Sv \geq v'\frac{N_i}{N}\Gamma_i(0)v - \alpha\|v\|_2^2 - \alpha/k\|v\|_1^2 \geq \alpha\|v\|_2^2 - \alpha/k\|v\|_1^2,$$

since $N_i \geq \Delta_n - 4R_n$, $N = n + q - 1 - 2m_0R_n$, and assuming $\Delta_n \geq \epsilon n$ implies that $N_i/N \geq \epsilon \geq 2\alpha$.

Condition (2) means that there exists a large enough constant $C' > 0$ such that

$$\|\hat{\gamma} - \hat{\Gamma}\Phi\|_\infty \leq C'\sqrt{\frac{\log\tilde{\pi}}{N}},$$

with probability at least $1 - c_1\exp(-c_2\log\tilde{\pi})$. To verify this condition here, observe that $\hat{\gamma} - \hat{\Gamma}\Phi = \mathrm{vec}(\mathcal{X}_{\mathbf{r}}'E_r)/N$. Therefore, denoting the h–th column block of $\mathcal{X}_{\mathbf{r}}$ by $\mathcal{X}_{\mathbf{r},(h)}$, for $h = 1, \ldots, (m_0+1)q$, we have:

$$\|\hat{\gamma} - \hat{\Gamma}\Phi\|_\infty = \max_{1\leq k,l\leq p; 1\leq h\leq (m_0+1)q} \big|e_k'\mathcal{X}_{\mathbf{r},(h)}'E_re_l\big|. \tag{B.25}$$

Now, for a fixed k, l, h, applying Proposition 2.4(b) in Basu and Michailidis (2015) gives:

$$\mathbb{P}\big(\big|e_k'\mathcal{X}_{\mathbf{r},(h)}'E_re_l\big| > k_1\eta\big) \leq 6\exp(-k_2 N\min(\eta^2, \eta)), \tag{B.26}$$

for large enough $k_1, k_2 > 0$, and any $\eta > 0$. Now, setting $\eta = C'\sqrt{\frac{\log\tilde{\pi}}{N}}$, and taking the union over all the $\tilde{\pi}$ cases for k, l, h yield the desired result. This completes the proof of this theorem.

□

Proof of Corollary 1. The proof of this Corollary is similar to that of Proposition 4.1 in Basu and Michailidis (2015). The key steps in this proof are similar to the previous proof of Theorem 4. Hence, we only highlight the main differences compared to Theorem 4; see more details in the proof of Theorem 4.

Instead of estimating the $m_0 + 1$ transition matrices simultaneously, we separately estimate the model parameters in each segment. For the i-the segment, $i = 1, \cdots, m_0 + 1$, let $\hat{\Gamma}_i = (I_p \otimes \mathcal{X}_i'\mathcal{X}_i/N_i)$ and $\hat{\gamma}_i = (I_p \otimes \mathcal{X}_i')\mathcal{Y}_i/N_i$. We need to verify the restricted eigenvalue (RE) condition for $\hat{\Gamma}_i$ and the deviation bound condition for $\left\|\hat{\gamma}_i - \hat{\Gamma}_i\Phi^{(.,i)}\right\|_\infty$. Once these two are verified, the rest of the proof is applying deterministic arguments used in Proposition 4.1 in Basu and Michailidis (2015).

Similar to Theorem 4, it suffices to show the RE for $S = \mathcal{X}_i'\mathcal{X}_i/N_i$. Then, we have

$$v'Sv \geq v'\Gamma_i(0)v - \alpha\|v\|_2^2 - \alpha/k\|v\|_1^2 \geq \alpha\|v\|_2^2 - \alpha/k\|v\|_1^2,$$

with probability at least $1 - c_1\exp(c_2N_i)$, for all $v \in \mathbb{R}^{pq}$, some $\alpha > 0$ and with an integer $k = \lceil c_4N_i/\log(pq)\rceil$ with some $c_4 > 0$

The DB condition implies that there exists a large enough constant $C' > 0$ such that

$$\|\hat{\gamma}_i - \hat{\Gamma}_i\Phi^{(.,i)}\|_\infty \leq C'\sqrt{\frac{\log p^2q}{N_i}},$$

with probability at least $1-c_1\exp(-c_2\log(p^2q))$. To verify this condition here, observe that $\hat{\gamma}_i - \hat{\Gamma}_i\Phi^{(.,i)} = \text{vec}(\mathcal{X}_i'E_i)/N_i$. We replace $\mathcal{X}_\mathbf{r}$, E_r and N with $\mathcal{X}_i$, E_i and N_i in (B.25) and (B.26). Next, setting $\eta = C'\sqrt{\frac{\log p^2q}{N_i}}$, and taking the union over all p^2q cases for k, l, h yields the desired result. This completes the proof. □

Appendix C: Details on the Detection Algorithm

Next, we provide details of the algorithm that solves optimization problem (4).

Let $S(.;\lambda)$ be the element-wise soft-thresholding operator which maps its input x to $x-\lambda$ when $x > \lambda$, $x + \lambda$ when $x < -\lambda$, and 0 when $|x| \leq \lambda$. Recall that throughout the paper, for a $m \times n$ matrix A, $\|A\|_\infty = \max_{1\leq i\leq m, 1\leq j\leq n}|a_{ij}|$. recall that $Y_l' = \big(y_l' \ldots y_{l-q+1}'\big)_{1\times pq}$. For two time points $s < t$, define $Y_{(s,t)} = (Y_s, Y_{s+1}, \ldots, Y_t) \in \mathbb{R}^{pq\times(t-s+1)}$ and $y_{(s,t)} = (y_s, y_{s+1}, \ldots, y_t) \in \mathbb{R}^{p\times(t-s+1)}$.

The main steps of the algorithm are as follows:

(i) Set the initial values for all parameters to zero; i.e. $\theta_i^{(0)} = 0$, for $i = 1, \ldots, k_n$.

(ii) For each $i = 1, \ldots, k_n$, calculate the $(h+1)$-th iteration of the parameters $\theta_i^{(h+1)}$ using the KKT condition as follows:

$$\theta_i'^{(h+1)} = \Big(\sum_{l=i}^{k_n} Y_{(r_{i-1},r_i)}Y_{(r_{i-1},r_i)}'\Big)^{-1} S\Big(\sum_{l=i}^{k_n} Y_{(r_{i-1},r_i)}y_{(r_i,r_{i+1})}' - \sum_{j\neq i}\Big(\sum_{l=\max(i,j)}^{k_n} Y_{(r_{i-1},r_i)}Y_{(r_{i-1},r_i)}'\Big)\theta_j'^{(h)}; \lambda_{1,n}\Big). \tag{C.1}$$

(iii) (a) If $\max_{1\leq i\leq k_n} \|{\theta_i}^{(h+1)} - {\theta_i}^{(h)}\|_\infty < \delta$, where δ is a tolerance threshold set to 10^{-3} in the implementation, stop and denote the final estimate by $\Theta^{(intermediate)}$.

(b) If $\max_{1\leq i\leq k_n} \|{\theta_i}^{(h+1)} - {\theta_i}^{(h)}\|_\infty \geq \delta$, set $h = h + 1$. Go to step (ii).

(iv) Apply soft–thresholding to $\Theta^{(intermediate)}$ to find the optimizer in equation (4). In other words, $\widehat{\Theta} = S(\Theta^{(intermediate)}; \lambda_{2,n})$.

Note that in this algorithm, the whole block of θ_i with p^2q elements is updated once, which reduces the computation time dramatically.

Appendix D: Tuning Parameters Selection

There are a number of tuning parameters in BSS: $\lambda_{1,n}, \lambda_{2,n}$, a_n, η_n, ω_n, ρ_n and R_n. Although their asymptotic values have been presented and discussed in the previous section, their selection in finite samples merits further discussion. Data-driven methods are developed to select them[1]:

$\lambda_{1,n}$, $\lambda_{2,n}$: Both $\lambda_{1,n}$ and $\lambda_{2,n}$ can be selected through cross-validation. For splitting the training set and the validation set, we randomly select 20% of the blocks equally spaced with a random initial point. Denote the last time point in these selected blocks by $\mathcal{T}$. The data without observations in $\mathcal{T}$ can then be used in the first step of our procedure to estimate Θ for a range of values for $\lambda_{1,n}$ and $\lambda_{2,n}$. The parameters estimated in the first step are then used to predict the series at time points in $\mathcal{T}$. The value of $\lambda_{1,n}$ and $\lambda_{2,n}$ which minimize the mean squared prediction error over $\mathcal{T}$ is the cross-validated choice of $\lambda_{1,n}$ and $\lambda_{2,n}$. The sequence for $\lambda_{1,n}$, $\lambda_{2,n}$ are selected as follows. Similar to Friedman et al. (2010), we construct a sequence of K_1 values for $\lambda_{1,n}$, decreasing from $\lambda_{1,\max}$ to $\lambda_{1,\min}$ on the log scale, where the maximum value $\lambda_{1,\max}$ is the smallest value for which the entire estimated parameter $\widehat{\theta}_i = 0$, for all $i = 1, \ldots, k_n$ while the minimum value is set to be $\lambda_{1,\min} = \epsilon\lambda_{1,\max}$. We choose $\lambda_{2,n} = c\sqrt{\frac{\log p}{T}}$, where c is a decreasing sequence of K_2 values. In the simulation study, we choose $K_1 = 10$, $K_2 = 3$, $\epsilon = 10^{-3}$ when the blocks size b_n is smaller than $2p$ and $\epsilon = 10^{-4}$ otherwise.

a_n: This tuning parameter can be selected through a grid search. We apply an exhaustive search procedure on a grid of a_n's ranked from the minimum to the maximum and record the number of selected break points for each a_n, and then stop this process when the number of break points selected does not change any more. In detail, we select $a_n^{(1)}, a_n^{(2)}, \ldots, a_n^{(\ell)}$ as an equally spaced sequence from the interval $[a_n^{(lb)}, a_n^{(ub)}]$ in a increasing order where $a_n^{(lb)} = \max(\lfloor\overline{b_n}\rfloor, \lfloor \log n \log p \rfloor)$, $a_n^{(ub)} = \min(10a_n^{(lb)}, (\widehat{t}_1 - q - 1), (T - q - \widehat{t}_{\widehat{m}} - 1))$, and $\overline{b_n}$ is the mean of block sizes. Denote the number of selected break points using BSS with $a_n^{(i)}$ as the neighborhood size by n_i, for $i = 1, 2, \ldots, \ell$. The optimal neighborhood size can be defined as the first time n_i

[1]The R/Rcpp codes to perform the BSS algorithm are available at the author's GitHub page: https://github.com/abolfazlsafikhani/BSS-ChangePoint-VAR

remains unchanged. In other words, $a_n^{(\ell^\star)}$ is the optimal neighborhood size when $\ell^\star = \min\{1 \le i \le \ell : n_i = n_{i+1} = n_{i+2}\}$. In all simulation scenarios, we set $\ell = 5$.

η_n: Despite the fact that based on Assumption A5, different treatments are needed for selecting tuning parameters $\eta_{\widehat{t}_i}$, $\eta_{\widehat{t}_i,1}$, $\eta_{\widehat{t}_i,2}$, all $\eta_{\widehat{t}_i}$, $\eta_{\widehat{t}_i,1}$ and $\eta_{\widehat{t}_i,2}$ are suggested to be set to $(\log(2a_n)\log p)/(2a_n)$, $i = 1, \ldots, \widehat{m}$. This choice was used in all of the numerical analyses undertaken and provides very good results.

ω_n: The selection of ω_n is the most difficult, since it depends on how large changes in the VAR parameters must be in order to consider them as break points in finite sample applications. Here, we consider using a data-driven method for selecting ω_n. The idea is to cluster the changes in the objective function L_n into two subgroups, small and large. The proposed algorithm is summarized as follow:

- Denote the candidate break points selected in the first step as $\widehat{t}_1, \ldots, \widehat{t}_{\widehat{m}}$. For each $k = 1, 2, \ldots, \widehat{m}$, compute $v_k = L_n(\widehat{\mathcal{A}}_n \backslash \{\widehat{t}_k\}; \eta_n) - L_n(\widehat{\mathcal{A}}_n; \eta_n)$.
- Consider two boundary points $t_1^r = a_n + q$ and $t_2^r = T - a_n$ as reference points. Compute $v_i^r = L_n(\widehat{\mathcal{A}}_n; \eta_n) - L_n(\widehat{\mathcal{A}}_n \cup \{t_i^r\}; \eta_n)$ for $i = 1, 2$, and set the reference value as $v^r = \max(v_1^r, v_2^r)$.
- Combine the jumps v_k for each candidate break points and the reference value v^r (with 2 replicates) into one vector $V = (v_1, v_2, \ldots, v_{\hat{m}}, v^r, v^r)$. Apply k-means clustering algorithm (Hartigan and Wong, 1979) to the vector V with two centers. Denote the sub-vector with smaller center as the small subgroup, V_S, and the other sub-vector as the large subgroup, V_L.
- If (between-group SS/total SS) in k-means clustering is high and the reference value v^r is not in V_L, set $\omega_n = \min V_L$; otherwise, set $\omega_n = \max V$.

One could also combine the k-means clustering method (Hartigan and Wong, 1979) with the BIC criterion (Schwarz et al., 1978) to cluster the changes in the parameter matrix into two subgroups.

ρ_n: We select ρ_n as the minimizer of the combined Bayesian Information Criterion (BIC) for all the segments. Following Lütkepohl (2005) and Zou et al. (2007), for $j = 0, \ldots, \tilde{m}$ we define the BIC on the interval $I_{j+1} = [r_{j2}, r_{(j+1)1}]$ as

$$\mathrm{BIC}(j, \rho_n) = \log(\det\widehat{\Sigma}_{\varepsilon,j}) + \frac{\log(r_{(j+1)1} - r_{j2})}{(r_{(j+1)1} - r_{j2})} \left\| \widehat{\beta}_{j+1} \right\|_0,$$

where $\widehat{\Sigma}_{\varepsilon,j}$ is the residual sample covariance matrix with $\widehat{\mathbf{B}}$ estimated in (16), and $\|\widehat{\beta}_{j+1}\|_0$ is the number of nonzero elements in $\widehat{\beta}_{j+1}$; then ρ_n is selected as

$$\widehat{\rho}_n = \mathrm{argmin}_{\rho_n} \sum_{j=0}^{\tilde{m}} \mathrm{BIC}(j, \rho_n). \tag{D.1}$$

R_n: Recall from Section 3 that we need to remove the selected break points together with their R_n-radius neighborhood before estimating the parameters using (16). In practice, the radius R_n needs to be estimated. However, a closer look into Theorem 2 together with Assumption A4 suggest that a_n can be chosen as the radius R_n. Therefore, in all simulation scenarios and data applications, we set $R_n = a_n$.

Appendix E: Comparison with SBS and DCBS

We record the detection accuracy of the proposed BSS strategy and compare it to two binary segmentation-based methods, the SBS (Cho and Fryzlewicz, 2015) and DCBS (Cho, 2016) methods (both are implemented in the R package "hdbinseg"), geared towards detection of multiple break points in multivariate time series data. Here, we consider the BSS method with three different block size settings: the large block size $b_n = \lfloor n^{\frac{1}{2}} \rfloor$,the medium block size $b_n = \lfloor n^{\frac{2}{5}} \rfloor$ and the small block size $b_n = \lfloor n^{\frac{1}{3}} \rfloor$, where $n = T - q$. Six additional simulation scenarios (G.1 - G.6) are considered for this comparison with model parameter values summarized in Table 1. The true coefficient matrices are similar to simulation B, as depicted in Figure 2 (top panel) with repeated entries -0.6, 0.6, and -0.6 off the main diagonal. Details of the simulation settings are as follows:

Setting G (detection comparison). Similar to scenario E, there are many true break points in the data generating process. In scenario G, $T = 4,000$, $p = 10$, $q = 1$ with break points being equally spaced: $\lfloor \frac{T}{m_0+1} \rfloor$, $\lfloor \frac{2T}{m_0+1} \rfloor, \ldots, \lfloor \frac{m_0 T}{m_0+1} \rfloor$. In scenarios G.1 though G.6, the true number of break points are $m_0 = 2, 4, 6, 8, 10, 12$, respectively.

Table 1: Details of model parameters for simulation settings G.

Sim	T	p	AR order q	block size b_n	m_0	AR structure
G.1	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
G.2	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	4	simple
G.3	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	6	simple
G.4	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	8	simple
G.5	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	10	simple
G.6	4,000	10	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	12	simple

The median of selection rates among m_0 true break points for all scenarios in setting G are summarized in the left panel in Figure 1. This plot clearly shows that in all settings considered, BSS has an advantage over SBS and DCBS in detection performance while this advantage gets more significant by increasing m_0. The selection rate for the BSS method with small and medium block sizes remains above 98%, while the selection rate for the SBS and DCBS methods are $\sim 80\%$ for $m_0 = 4, 6$ and drop to $\sim 10\%$ for $m_0 = 10, 12$. Note that the selection rate for BSS with large block size also remains above $\sim 95\%$ with the exception of $\sim 40\%$ for the case of $m_0 = 12$ (still higher than the selection rate of SBS and DCBS in this case). The Hausdorff distance $d_H\left(\widetilde{\mathcal{A}}_n^f, \mathcal{A}_n\right)$ between the set of estimated break points and the set of true break points is a reasonable measure for estimation accuracy. The right panel in Figure 1 illustrates the performance of all three methods in terms of this estimation accuracy (averaged over 100 replicates) in which the BSS outperforms the SBS

and DCBS methods in almost all settings, while the advantage of BSS is more significant for larger m_0 values. For example, in the case of 6 true break points, the average estimation accuracy for SBS is around 677.29, while the same quantity for BSS (large, medium, small block sizes) is around 3.50, 0.84 and 0.10, respectively. Therefore, even the BSS method with large block sizes reduces the estimation error in locating the break points by around 99.5% compared to SBS.

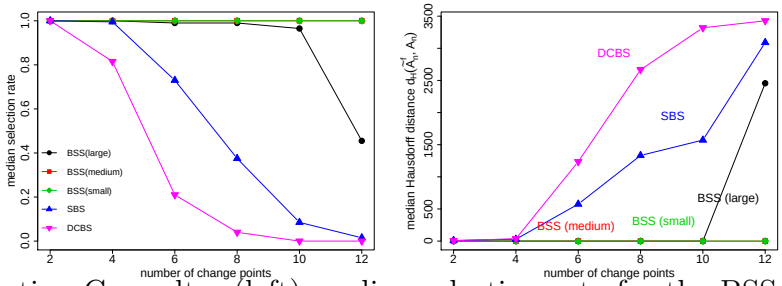

Figure 1: Simulation G results: (left) median selection rate for the BSS, SBS and DCBS methods; (right) median Hausdorff distance $d_H\left(\widetilde{\mathcal{A}}_n^f, \mathcal{A}_n\right)$ for the BSS, SBS and DCBS methods.

Next, we compare the computation time for three methods (BSS, SBS and DCBS). Six additional simulation scenarios (H.1-H.6) are considered for comparison among BSS, SBS and DCBS with model parameter values summarized in Table 2. Details of the simulation settings are as follows:

Setting H (computation time comparison). In scenario H, $p = 20$, $q = 1$,$m_0 = 2$, $t_1 = \lfloor \frac{T}{3} \rfloor$, $t_2 = \lfloor \frac{2T}{3} \rfloor$. The auto-regressive coefficients are chosen to have the same simple 1-off diagonal structure as in Scenario A.1 as shown in the top left panel of Figure 2 with repeated entries -0.6, 0.6, and -0.6 off the main diagonal. The sample size for scenarios H.1 through H.6 are $T = 1000, 2000, 3000, 4000, 5000, 6000$, respectively.

Table 2: Details of model parameters for simulation settings H.

Sim	T	p	AR order q	block size b_n	m_0	AR structure
H.1	1,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
H.2	2,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
H.3	3,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
H.4	4,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
H.5	5,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple
H.6	6,000	20	1	$\lfloor n^{\frac{1}{2}} \rfloor$, $\lfloor n^{\frac{2}{5}} \rfloor$, $\lfloor n^{\frac{1}{3}} \rfloor$	2	simple

The average computation time over 100 replicates for simulation setting H is plotted in Figure 2. BSS with large block sizes is the fastest method overall, while DCBS is the slowest one. It is worth noting that BSS with small block sizes remains faster than both SBS and DCBS, while its estimation accuracy and selection rate are the best over all these three methods as explained in simulation setting G. In this numerical experiment, the reduction in computation time in BSS (small block size) compared to SBS and DCBS is around 30%

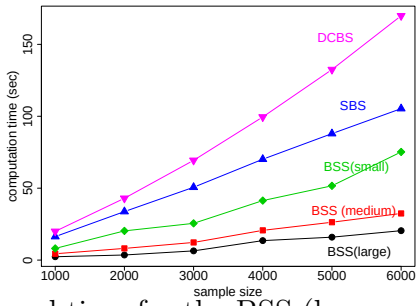

Figure 2: Average computational time for the BSS (large, medium, small block sizes), SBS and DCBS methods .

and 55%, respectively, while BSS with medium and large block sizes achieves even higher time reduction.

References

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*(4), 1535–1567.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics 10*(2), 2000–2038.

Cho, H. and P. Fryzlewicz (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(2), 475–507.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*(1), 100–108.

Loh, P.-L. and M. J. Wainwright (2012, 06). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist. 40*(3), 1637–1664.

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

Safikhani, A. and A. Shojaie (2020). Joint structural break detection and parameter estimation in high-dimensional non-stationary var models. *Journal of American Statistical Association (Theory and methods)*, To Appear.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics 35*(5), 2173–2192.