Supplementary Materials For "Joint Structural Break Detection and Parameter Estimation in High-Dimensional Non-Stationary VAR Models"

Abolfazl Safikhani* and Ali Shojaie†

University of Florida and University of Washington

May 12, 2020

Appendix

In Appendix A, we collect technical lemmas needed to prove the main results. Proofs of the main results are given Appendix B. Details of the algorithm for solving the optimization problem (5) are given in Appendix C. Further details on simulations settings, and additional simulation results are reported in Appendix D. In Appendix E,

*a.safikhani@ufl.edu

†ashojaie@uw.edu

we propose a secondary analysis which sharpens the consistency rate for locating the break points. An alternative procedure to our proposed third step on consistent parameter estimation is discussed in Appendix F. Finally, a data-driven method to select the tuning parameter ω_n is presented in Appendix G.

Appendix A: Technical Lemmas

Lemma 1. *There exist constants* $c_i > 0$ *such that for* $n \geq c_0 \left(\log(n) + 2\log(p) + \log(q)\right)$*, with probability at least* $1 - c_1 \exp\left(-c_2\left(\log(n) + 2\log(p) + \log(q)\right)\right)$*, we have*

$$\left|\left|\frac{\boldsymbol{Z}'\boldsymbol{E}}{n}\right|\right|_{\infty} \leq c_3 \sqrt{\frac{\log(n) + 2\log(p) + \log(q)}{n}} \tag{1}$$

Proof. Note that $\frac{1}{n}\mathbf{Z}'\mathbf{E} = \frac{1}{n}(I_p \otimes \mathcal{X}')\mathbf{E} = \text{vec}(\mathcal{X}'E)/n$. Let $\mathcal{X}(h,.)$ and $\mathcal{X}(h,l)$ be the h-th block column and the l-th column of the h-th block column of $\mathcal{X}$, respectively, for $1 \leq h \leq n$, $1 \leq l \leq d$. More specifically,

$$\mathcal{X}(h,.) = \begin{pmatrix} & 0 & \\ & \vdots & \\ & 0 & \\ y'_{q+h-2} & \cdots & y'_{h-1} \\ & \vdots & \\ y'_{T-1} & \cdots & y'_{T-q} \end{pmatrix}_{n \times pq}, \qquad \mathcal{X}(h,l) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y'_{q+h-l-1} \\ \vdots \\ y'_{T-l} \end{pmatrix}_{n \times p}. \tag{2}$$

Now,

$$\left\|\frac{\mathbf{Z}'\mathbf{E}}{n}\right\|_{\infty} = \max_{1\le h\le n, 1\le l\le d, 1\le i,j\le p} \left| e_i' \left(\frac{\mathcal{X}'(h,l)E}{n}\right) e_j \right|, \tag{3}$$

where $e_i \in \mathbb{R}^p$ with the i-th element equals to 1 and zero on the rest. Note that,

$$\frac{\mathcal{X}'(h,l)E}{n} = \frac{1}{n}\sum_{t=h-l-1}^{T-q-l} y_{q+t}\varepsilon_{q+t+l}'.$$

Now, since $\mathrm{cov}(y_{q+t}, \varepsilon_{q+t+l}) = 0$ for all t, l, h, an argument similar to Proposition 2.4(b) of Basu & Michailidis (2015) shows that for fixed i, j, h, l, there exist $k_1, k_2 > 0$ such that for all $\eta > 0$:

$$\mathbb{P}\left(\left| e_i' \left(\frac{\mathcal{X}'(h,l)E}{n}\right) e_j \right| > k_1\eta\right) \le 6\exp\left(-k_2 n \min(\eta, \eta^2)\right).$$

The result follows by setting $\eta = k_3\sqrt{\frac{\log(n)+2\log(p)+\log(q)}{n}}$ for a large enough $k_3 > 0$, and taking the union over the $\pi = np^2q$ possible choices of i, j, h, l. □

Lemma 2. *Let* $\widehat{\mathbf{\Theta}}$ *be defined as in* (5)*. Then, under the assumptions of Theorem 1:*

$$\sum_{l=\widehat{t}_j}^{n} Y_{l-1}\left(y_l' - Y_{l-1}'\sum_{i=1}^{l}\widehat{\theta}_i'\right) = \frac{n\lambda_n}{2} sign(\widehat{\theta}_{\widehat{t}_j}'), \qquad \text{for } j = 1,2,...,\widehat{m}, \tag{4}$$

where $Y_l' = \left(y_l' \dots y_{l-q+1}'\right)_{1\times pq}$*, and*

$$\left\|\sum_{l=j}^{n} Y_{l-1}\left(y_l' - Y_{l-1}'\sum_{i=1}^{l}\widehat{\theta}_i'\right)\right\|_{\infty} \le \frac{n\lambda_n}{2}, \qquad \text{for } j = q-1,2,...,n. \tag{5}$$

Moreover, $\sum_{i=1}^{t}\widehat{\theta}_i = \widehat{\Phi}^{(.,j)}$ *for* $\widehat{t}_{j-1} \le t \le \widehat{t}_j - 1$, $j = 1,2,...,|\mathcal{A}_n|$.

Proof. The result follows directly from the KKT condition of the optimization problem (5). □

Lemma 3. *Under assumption A1, there exist constants* $c_i > 0$ *such that with probability at least* $1 - c_1 \exp(-c_2(\log(q) + 2\log(p)))$*,*

$$\sup_{1 \le j \le m_0, s \ge t_j, |t_j - s| > n\gamma_n} \left|\left| (t_j - s)^{-1} \left(\sum_{l=s}^{t_j - 1} Y_{l-1} Y'_{l-1} - \Gamma_j^q(0) \right) \right|\right|_\infty \le c_3 \sqrt{\frac{\log(q) + 2\log(p)}{n\gamma_n}}, \tag{6}$$

where $\Gamma_j^q(0) = \mathbb{E}(Y_{l-1}Y'_{l-1})$*. Moreover,*

$$\sup_{1 \le j \le m_0, s \ge t_j, |t_j - s| > n\gamma_n} \left|\left| (t_j - s)^{-1} \sum_{l=s}^{t_j - 1} Y_{l-1} \varepsilon'_l \right|\right|_\infty \le c_3 \sqrt{\frac{\log(q) + 2\log(p)}{n\gamma_n}}. \tag{7}$$

Proof. The proof of this lemma is similar to that of Proposition 2.4 in Basu & Michailidis (2015). Here we briefly outline the main steps of the proof, while omitting the details. For (6), note that using an argument similar to Proposition 2.4(a) in Basu & Michailidis (2015), there exist $k_1, k_2 > 0$ such that for each fixed $k, l = 1, \cdots, pq$,

$$\mathbb{P}\left(\left| e'_k \frac{\sum_{l=s}^{t_j - 1} Y_{l-1} Y'_{l-1} - \Gamma_j^q(0)}{t_j - s} e_l \right| > k_1 \eta \right) \le 6 \exp(-k_2 n\gamma_n \min(\eta, \eta^2)). \tag{8}$$

Setting $\eta = k_3 \sqrt{\frac{\log(qp^2)}{n\gamma_n}}$, and taking union over all possible values of k, l, we obtain (6).

The proof for (7), is similar to Lemma 1. Again, there exist $k_1, k_2 > 0$ such that for each fixed $k = 1, ..., pq$, $l = 1, ..., p$,

$$\mathbb{P}\left(\left|e_k' \frac{\sum_{l=s}^{t_j-1} Y_{l-1}\varepsilon_l'}{t_j - s} e_l\right| > k_1\eta\right) \leq 6\exp(-k_2 n\gamma_n \min(\eta, \eta^2)). \tag{9}$$

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

$$\left|\left|(t_j - s)^{-1}\left(\sum_{l=s}^{t_j-1} Y_{l-1}Y_{l-1}' - \Gamma_j^q(0)\right)\right|\right|_\infty \leq c_3\sqrt{\frac{\log(q) + 2\log(p)}{n\gamma_n}}, \tag{10}$$

and

$$\left|\left|(t_j - s)^{-1}\sum_{l=s}^{t_j-1} Y_{l-1}\varepsilon_l'\right|\right|_\infty \leq c_3\sqrt{\frac{\log(q) + 2\log(p)}{n\gamma_n}}, \tag{11}$$

with high probability converging to 1 for any $j = 1, 2, \cdots, m_0$, as long as $|t_j - s| > n\gamma_n$ and $s \geq t_{j-1}$. 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 $1 - c_1\exp(-c_2(\log(q) + 2\log(p)))$. □

Lemma 4. *Under the assumptions of Theorem 3, for $m < m_0$, there exist constants $c_1, c_2 > 0$ such that:*

$$\mathbb{P}\left(\min_{(s_1,...,s_m)\subset\{1,...,T\}} L_n(s_1, s_2, ..., s_m; \eta_n) > \sum_{t=q}^{T} ||\varepsilon_t||_2^2 + c_1\Delta_n - c_2 m n\gamma_n d_n^{\star 2}\right) \to 1, \tag{12}$$

where $\Delta_n = \min_{1\leq j\leq m_0+1}|t_j - t_{j-1}|$.

Proof. Since $m < m_0$, there exists a point t_j such that $|s_i - t_j| > \Delta_n/4$. In order to find a lower bound on the sum of the least squares, we consider three different cases: (a) $|s_i - s_{i-1}| \leq n\gamma_n$; (b) there exist two true break points t_j, t_{j+1} such that $|s_{i-1} - t_j| \leq n\gamma_n$ and $|s_i - t_{j+1}| \leq n\gamma_n$; and (c) otherwise. The idea is to find a lower bound for the sum of squared errors plus the penalty term for each case. Here, we consider only one candidate for each case. The general case can be argued similarly, but omitted here to avoid complex notations.

Denote the estimated parameter in each of the estimated segments below by $\widehat{\theta}$. For case (a), consider the case where the interval (s_{i-1}, s_i) is inside a true segment. In other words, suppose there exists j such that $t_j < s_{i-1} < s_i < t_{j+1}$. Now,

$$\begin{aligned}
\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta}Y_{t-1}||_2^2 &= \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 + \sum_{t=s_{i-1}}^{s_i-1} ||(\Phi^{(.,j+1)} - \widehat{\theta})Y_{t-1}||_2^2 \\
&+ 2\sum_{t=s_{i-1}}^{s_i-1} Y_{t-1}'(\Phi^{(.,j+1)} - \widehat{\theta})'\varepsilon_t \\
&\geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 - \left| 2\sum_{t=s_{i-1}}^{s_i-1} Y_{t-1}'(\Phi^{(.,j+1)} - \widehat{\theta})'\varepsilon_t \right| \\
&\geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 - c\sqrt{n\gamma_n \log p}||\Phi^{(.,j+1)} - \widehat{\theta}||_1. \qquad (13)
\end{aligned}$$

Therefore, given the tuning parameter selected based on Assumption A4, we have:

$$\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta}Y_{t-1}||_2^2 + \eta_{(s_{i-1},s_i)}||\widehat{\theta}||_1 \geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 - c\sqrt{n\gamma_n \log p}||\Phi^{(.,j+1)}||_1. \qquad (14)$$

For case (b), consider the case where $s_{i-1} < t_j$ and $s_i < t_{j+1}$. Now, similar arguments as in Proposition 4.1 of Basu & Michailidis (2015) show that by the tuning parameter selected based on A4(b), we have:

$$||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \leq 4\sqrt{d_n^\star}||\Phi^{(.,j+1)} - \widehat{\theta}||_2, \text{ and } ||\Phi^{(.,j+1)} - \widehat{\theta}||_2 \leq c_3\sqrt{d_n^\star}\eta_{(s_{i-1},s_i)}. \quad (15)$$

To see this, observe that $\widehat{\theta}$ in (9) minimizes the least squares plus the ℓ_1 norm loss function. Therefore, the value of this objective function for $\widehat{\theta}$ will be smaller than any other choice of parameters, including $\Phi^{(.,j+1)}$. Hence,

$$\begin{aligned} \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta}Y_{t-1}||_2^2 + \eta_{(s_{i-1},s_i)}||\widehat{\theta}||_1 &\leq \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} ||y_t - \Phi^{(.,j+1)}Y_{t-1}||_2^2 \\ &+ \eta_{(s_{i-1},s_i)}||\Phi^{(.,j+1)}||_1. \quad (16) \end{aligned}$$

Some rearrangements lead to:

$$\begin{aligned}
0 \leq c'||\Phi^{(.,j+1)} - \widehat{\theta}||_2^2 \quad &\leq \quad \frac{1}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} Y_{t-1}' \left(\Phi^{(.,j+1)} - \widehat{\theta}\right)' \left(\Phi^{(.,j+1)} - \widehat{\theta}\right) Y_{t-1} \\
&\leq \quad \frac{2}{s_i - s_{i-1}} \sum_{t=s_{i-1}}^{s_i-1} Y_{t-1}' \left(\Phi^{(.,j+1)} - \widehat{\theta}\right)' \left(y_t - \Phi^{(.,j+1)} Y_{t-1}\right) \\
&\quad + \quad \eta_{(s_{i-1},s_i)} \left(||\Phi^{(.,j+1)}||_1 - ||\widehat{\theta}||_1\right) \\
&\leq \quad \left(c\sqrt{\frac{\log p}{s_i - s_{i-1}}} + M_\Phi d_n^\star \frac{n\gamma_n}{s_i - s_{i-1}}\right) ||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \\
&\quad + \quad \eta_{(s_{i-1},s_i)} \left(||\Phi^{(.,j+1)}||_1 - ||\widehat{\theta}||_1\right) \\
&\leq \quad \frac{\eta_{(s_{i-1},s_i)}}{2} ||\Phi^{(.,j+1)} - \widehat{\theta}||_1 + \eta_{(s_{i-1},s_i)} \left(||\Phi^{(.,j+1)}||_1 - ||\widehat{\theta}||_1\right) \\
&\leq \quad \frac{3\eta_{(s_{i-1},s_i)}}{2} ||\Phi^{(.,j+1)} - \widehat{\theta}||_{1,\mathcal{I}} - \frac{\eta_{(s_{i-1},s_i)}}{2} ||\Phi^{(.,j+1)} - \widehat{\theta}||_{1,\mathcal{I}^c} \\
&\leq \quad 2\eta_{(s_{i-1},s_i)} ||\Phi^{(.,j+1)} - \widehat{\theta}||_1. \qquad (17)
\end{aligned}$$

This ensures that $||\Phi^{(.,j+1)} - \widehat{\theta}||_{1,\mathcal{I}^c} \leq 3||\Phi^{(.,j+1)} - \widehat{\theta}||_{1,\mathcal{I}}$, and hence $||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \leq 4||\Phi^{(.,j+1)} - \widehat{\theta}||_{1,\mathcal{I}} \leq 4\sqrt{d_n^\star}||\Phi^{(.,j+1)} - \widehat{\theta}||_2$. This comparison between L_1 and L_2 norms of the error term together with the bound in Equation 17 will get the desired consistency rates in (15).

Similar to case (a), using lemma (3), we have:

$$\begin{aligned}
\sum_{t=t_j}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 &\geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + c|s_i - t_j| \, ||\Phi^{(.,j+1)} - \widehat{\theta}||_2^2 - c' \sqrt{|s_i - t_j| \log p} \, ||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \\
&\geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + c|s_i - t_j| \, ||\Phi^{(.,j+1)} - \widehat{\theta}||_2 \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_2 - \frac{c'}{c} \sqrt{\frac{d_n^\star \log p}{|s_i - t_j|}} \right) \\
&\geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 - c' d_n^\star \log p. \qquad (18)
\end{aligned}$$

Also, for the interval (s_{i-1}, t_j), we have:

$$\begin{aligned}
\sum_{t=s_{i-1}}^{t_j-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 &\geq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 - c' \sqrt{n\gamma_n \log p} \, ||\Phi^{(.,j)} - \widehat{\theta}||_1 \\
&\geq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 - c' \sqrt{n\gamma_n \log p} \, \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_1 + ||\Phi^{(.,j+1)} - \Phi^{(.,j)}||_1 \right) \\
&\geq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 - c' \sqrt{n\gamma_n \log p} \, \left(d_n^\star \eta_{(s_{i-1}, s_i)} + ||\Phi^{(.,j+1)} - \Phi^{(.,j)}||_1 \right) \\
&\geq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 - c' d_n^\star \sqrt{n\gamma_n \log p}. \qquad (19)
\end{aligned}$$

Combining equations (18) and (19) gives:

$$\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 - c' d_n^\star \sqrt{n\gamma_n \log p}. \qquad (20)$$

For case (c), consider the case where $s_{i-1} < t_j < s_i$ with $|s_{i-1} - t_j| > \Delta_n/4$ and $|s_i - t_j| > \Delta_n/4$. Similar arguments as in Proposition 4.1 of Basu & Michailidis

(2015) shows that :

$$||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \leq 4\sqrt{d_n^\star}||\Phi^{(.,j+1)} - \widehat{\theta}||_2, \text{ and } ||\Phi^{(.,j)} - \widehat{\theta}||_1 \leq 4\sqrt{d_n^\star}||\Phi^{(.,j)} - \widehat{\theta}||_2. \quad (21)$$

Note that in this case, the restricted eigenvalue condition does not hold. Therefore, the convergence of the $\widehat{\theta}$ cannot be verified. The reason is that in this case, the two parts of the true segments which intersect with the estimated segment have large lengths. If the length of one of them was negligible as compared to the other segment, one could still verify the restricted eigenvalue, but that's not the case here. However, the deterministic part of the deviation bound argument holds with the suitable choice of the tuning parameter. Now, similar to case (b), on both intervals (s_{i-1}, t_j) and (t_j, s_i):

$$\begin{aligned}
\sum_{t=s_{i-1}}^{t_j-1} ||y_t - \widehat{\theta}Y_{t-1}||_2^2 \quad &\geq \quad \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 + c|t_j - s_{i-1}|\, ||\Phi^{(.,j)} - \widehat{\theta}||_2^2 \\
&- \quad c'\sqrt{|t_j - s_{i-1}|\log p}\, ||\Phi^{(.,j)} - \widehat{\theta}||_1 \\
&\geq \quad \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 \\
&+ c|t_j - s_{i-1}|\, ||\Phi^{(.,j)} - \widehat{\theta}||_2 \left(||\Phi^{(.,j)} - \widehat{\theta}||_2 - \frac{c'}{c}\sqrt{\frac{d_n^\star \log p}{|t_j - s_{i-1}|}} \right) (22)
\end{aligned}$$

and

$$\begin{aligned}\sum_{t=t_j}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 &\geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + c|s_i - t_j|\, ||\Phi^{(.,j+1)} - \widehat{\theta}||_2^2 - c' \sqrt{|s_i - t_j| \log p}\, ||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \\ &\geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + c|s_i - t_j|\, ||\Phi^{(.,j+1)} - \widehat{\theta}||_2 \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_2 - \frac{c'}{c} \sqrt{\frac{d_n^\star \log p}{|s_i - t_j|}} \right). \quad (23)\end{aligned}$$

Since $||\Phi^{(.,j+1)} - \Phi^{(.,j)}||_2 \geq v > 0$, either $||\Phi^{(.,j+1)} - \widehat{\theta}||_2 \geq v/4$ or $||\Phi^{(.,j)} - \widehat{\theta}||_2 \geq v/4$. Assume that $||\Phi^{(.,j)} - \widehat{\theta}||_2 \geq v/4$. Then, based on Equation 22, for some $c_1 > 0$,

$$\sum_{t=s_{i-1}}^{t_j-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 \geq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 + c_1 \Delta_n. \quad (24)$$

For the second interval we have:

$$\sum_{t=t_j}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 \geq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 - c' d_n^\star \log p. \quad (25)$$

Combining (24) and (25), leads to:

$$\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 + c_1 \Delta_n - c' d_n^\star \log p. \quad (26)$$

Note that another situation may arise in this case, where $|s_{i-1} - t_j| > n\gamma_n$ and $|s_i - t_j| > n\gamma_n$. Using similar augments as above in this situation, we get the following lower bound:

$$\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 \geq \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 - c' {d_n^\star}^2 n\gamma_n. \quad (27)$$

Putting all of the cases together will yield the result.

□

Appendix B: Proof of Main Results

Proof of Theorem 1. By definition of $\widehat{\Theta}$, we get

$$\frac{1}{n}||\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Theta}}||_2^2+\lambda_{1,n}\sum_{i=1}^{n}||\widehat{\theta}_i||_1+\lambda_{2,n}\sum_{k=1}^{n}\left\|\sum_{j=1}^{k}\widehat{\theta}_j\right\|_1 \leq \frac{1}{n}||\mathbf{Y}-\mathbf{Z}\mathbf{\Theta}||_2^2+\lambda_{1,n}\sum_{i=1}^{n}||\theta_i||_1+\lambda_{2,n}\sum_{k=1}^{n}\left\|\sum_{j=1}^{k}\theta_j\right\|_1 \tag{28}$$

Denoting $\mathcal{A} = \{t_1, t_1, \cdots, t_{m_0}\}$, we have:

$$\begin{aligned}
\frac{1}{n}\left|\left|\mathbf{Z}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\right)\right|\right|_2^2 &\leq \frac{2}{n}\left(\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\right)'\mathbf{Z}'\mathbf{E} + \lambda_{1,n}\sum_{i=1}^{n}||\theta_i||_1 - \lambda_{1,n}\sum_{i=1}^{n}||\widehat{\theta}_i||_1 \\
&+ \lambda_{2,n}\sum_{k=1}^{n}\left|\left|\sum_{j=1}^{k}\theta_j\right|\right|_1 - \lambda_{2,n}\sum_{k=1}^{n}\left|\left|\sum_{j=1}^{k}\widehat{\theta}_j\right|\right|_1 \\
&\leq 2\left|\left|\frac{\mathbf{Z}'\mathbf{E}}{n}\right|\right|_\infty \sum_{i=1}^{n}||\theta_i-\widehat{\theta}_i||_1 + \lambda_{1,n}\sum_{i\in\mathcal{A}}\left(||\theta_i||_1 - ||\widehat{\theta}_i||_1\right) - \lambda_{1,n}\sum_{i\in\mathcal{A}^c}||\widehat{\theta}_i||_1 \\
&+ \lambda_{2,n}\sum_{j=1}^{m_0+1}(t_j - t_{j-1})\|\Phi^{(.,j)}\|_1 \\
&\leq \lambda_{1,n}\sum_{i\in\mathcal{A}}||\theta_i-\widehat{\theta}_i||_1 + \lambda_{1,n}\sum_{i\in\mathcal{A}}\left(||\theta_i||_1 - ||\widehat{\theta}_i||_1\right) + \lambda_{2,n}n\, d_n^\star \\
&\leq 2\lambda_{1,n}\sum_{i\in\mathcal{A}}||\theta_i||_1 + \lambda_{2,n}n\, d_n^\star \\
&\leq 2\lambda_{1,n}m_n \max_{1\leq j\leq m_0+1}\left|\left|\Phi^{(.,j)}-\Phi^{(.,j-1)}\right|\right|_1 + \lambda_{2,n}n\, d_n^\star \\
\leq\ 4Cm_n \max_{1\leq j\leq m_0+1}&\left\{\sum_{k=1}^{p}\left(d_{kj}+d_{k(j-1)}\right)\right\} M_\Phi\sqrt{\frac{\log(n)+2\log(p)+\log(q)}{n}} \\
&+ \lambda_{2,n}n\, d_n^\star, \qquad (29)
\end{aligned}$$

with high probability approaching to 1 due to Lemma 1. □

Proof of Theorem 2. The proof is different from Theorem 2.2 in Chan *et al.* (2014) and Proposition 5 in Harchaoui & Lévy-Leduc (2010) due to the additional penalty added in equation (5). For a matrix $A \in \mathbb{R}^{pq\times p}$, let $||A||_{1,\mathcal{I}} = \sum_{(j,k)\in\mathcal{I}}|a_{jk}|$.

First, we focus on the second part. Suppose for some $j = 1, \cdots, m_0$, $|\widehat{t}_j - t_j| > n\gamma_n$. Then, there exists a true break point t_{j_0} which is isolated from all the estimated points, i.e., $\min_{1\leq j\leq m_0}|\widehat{t}_j - t_{j_0}| > n\gamma_n$. 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 with 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 (5), the value of the function defined in (5) is minimized exactly at $\widehat{\Theta}$. This means that any other choice of parameters yields to higher value in (5). First, we focus on the interval $[t_{j_0-1} \vee \widehat{t}_j, t_{j_0}]$. Define a new parameter sequence ψ_k's, $k = 1, ..., n$ with $\psi_k = \widehat{\theta}_k$ except for two time points $k = \widehat{t}_j$ and $k = t_{j_0}$. For these two points set $\psi_{\widehat{t}_j} = \Phi^{(.,j_0)} - \widehat{\Phi}_j$ and $\psi_{t_{j_0}} = \widehat{\Phi}_{j+1} - \Phi^{(.,j_0)}$ where $\widehat{\Phi}_j = \sum_{k=1}^{t_{j_0-1} \vee \widehat{t}_j - 1} \widehat{\theta}_k$ and $\widehat{\Phi}_{j+1} = \sum_{k=1}^{t_{j_0} \vee \widehat{t}_j} \widehat{\theta}_k$, i.e. $\widehat{\theta}_{t_{j_0} \vee \widehat{t}_j} = \widehat{\Phi}_{j+1} - \widehat{\Phi}_j$. Denoting $\Psi = \text{vector}(\psi_1, ..., \psi_n) \in \mathbb{R}^{\pi \times 1}$, we have

$$
\begin{aligned}
\frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Theta}}\|_2^2 + \lambda_{1,n}\|\widehat{\boldsymbol{\Theta}}\|_1 + \lambda_{2,n}\sum_{k=1}^{n}\left\|\sum_{j=1}^{k}\widehat{\theta}_j\right\|_1 \quad &\leq \quad \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\Psi}\|_2^2 + \lambda_{1,n}\|\boldsymbol{\Psi}\|_1 \\
&+ \quad \lambda_{2,n}\sum_{k=1}^{n}\left\|\sum_{j=1}^{k}\psi_j\right\|_1. \qquad (30)
\end{aligned}
$$

Some rearrangement of equation (30) leads to

$$\begin{aligned}
0 &\leq c\,\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_2^2 \\
&\leq \frac{1}{t_{j_0} - t_{j_0-1} \vee \widehat{t}_j} \sum_{l=t_{j_0-1}\vee\widehat{t}_j}^{t_{j_0}-1} \left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right)' Y_{l-1}Y_{l-1}' \left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right) \\
&\leq \frac{1}{t_{j_0} - t_{j_0-1} \vee \widehat{t}_j} \sum_{l=t_{j_0-1}\vee\widehat{t}_j}^{t_{j_0}-1} Y_{l-1}' \left(\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\right) \varepsilon_l \\
&+ \frac{n\lambda_{1,n}}{t_{j_0} - t_{j_0-1} \vee \widehat{t}_j} \left(\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + \|\Phi^{(.,j_0)} - \widehat{\Phi}_j\|_1 - \|\widehat{\Phi}_{j+1} - \widehat{\Phi}_j\|_1\right) \\
&+ n\lambda_{2,n} \left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\
&\leq \left(\frac{2n\lambda_{1,n}}{t_{j_0} - t_{j_0-1} \vee \widehat{t}_j} + C\sqrt{\frac{\log p}{n\gamma_n}}\right) \|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + n\lambda_{2,n} \left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\
&\leq \frac{1}{2} n\lambda_{2,n} \|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_1 + n\lambda_{2,n} \left(\|\Phi^{(.,j_0)}\|_1 - \|\widehat{\Phi}_{j+1}\|_1\right) \\
&\leq \frac{3}{2} n\lambda_{2,n} \|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}} - \frac{1}{2} n\lambda_{2,n} \|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_{1,\mathcal{I}^c}. \qquad (31)
\end{aligned}$$

In equation (31), the second inequality holds with high probability converging to 1 due to first part of Lemma 3 and the fact that $t_{j_0} - t_{j_0-1} \vee \widehat{t}_j \geq n\gamma_n$. The fourth inequality holds with high probability converging to 1 due to second part of Lemma 3 and triangular inequality. The fifth inequality is based on the assumption A3 and the selection for $\lambda_{2,n}$ in the statement of the theorem. The last inequality holds by sparsity assumption. This implies that

$$\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_2 = o_p\left(d_n^\star \sqrt{\frac{\log p}{n\gamma_n}}\right), \qquad (32)$$

which means that $\|\Phi^{(.,j_0)} - \widehat{\Phi}_{j+1}\|_2$ converges to zero in probability based on assumption A3. Similarly, the same procedure can be applied to the interval $[t_{j_0}, t_{j_0+1} \wedge \widehat{t}_{j+1}]$ which leads to $\|\Phi^{(.,j_0+1)} - \widehat{\Phi}_{j+1}\|_2$ converges to zero in probability as well. This yields to a contradiction to the 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 $[t_{j_0-1} \vee \widehat{t}_j, t_{j_0}]$ and $[t_{j_0}, t_{j_0+1} \wedge \widehat{t}_{j+1}]$ which leads to $\|\Phi^{(.,j_0+1)} - \Phi^{(.,j_0)}\|_2$ converges to zero and therefore contradicts with assumption A3. This completes the proof.

□

Proof of Theorem 3. For the first part we show that (a) $\mathbb{P}(\widetilde{m} < m_0) \to 0$, and (b) $\mathbb{P}(\widetilde{m} > m_0) \to 0$. For (a), we know, from Theorem 2, that there exists points $\widehat{t}_j \in \mathcal{A}_n$ such that $\max_{1 \leq j \leq m_0} |\widehat{t}_j - t_j| \leq n\gamma_n$. By similar arguments as in Lemma 4, we get that there exists a constant $K > 0$ such that:

$$L(\widehat{t}_1, ..., \widehat{t}_{m_0}; \eta_n) \quad \leq \quad \sum_{t=q}^{T} ||\varepsilon_t||_2^2 + K m_0 n \gamma_n {d_n^\star}^2. \qquad (33)$$

To see this, we only show the calculations for one of the estimated segments. Suppose $s_{i-1} < t_j < s_i$ with $|t_j - s_{i-1}| \leq n\gamma_n$. Denote the estimated coefficient in the segment

(s_{i-1}, s_i) by $\widehat{\theta}$. Similar to case (b) in the proof of Lemma 4, we have:

$$\begin{aligned}
\sum_{t=t_j}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 &\leq \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + c_3 |s_i - t_j| \, ||\Phi^{(.,j+1)} - \widehat{\theta}||_2^2 \\
&+ c' \sqrt{|s_i - t_j| \log p} \, ||\Phi^{(.,j+1)} - \widehat{\theta}||_1 \\
&\equiv \sum_{t=t_j}^{s_i-1} ||\varepsilon_t||_2^2 + I + II. \qquad (34)
\end{aligned}$$

Now, by the convergence rate of the error (see, e.g., case (b) in the proof of Lemma 4),

$$\begin{aligned}
I &\leq 4c_3 |s_i - t_j| d_n^\star \left(c \sqrt{\frac{\log p}{|s_i - t_j|}} + M_\Phi d_n^\star \frac{n\gamma_n}{|s_i - t_j|} \right)^2 \\
&= O_p \left(n\gamma_n {d_n^\star}^2 \right), \qquad (35)
\end{aligned}$$

and

$$\begin{aligned}
II &\leq c' \sqrt{|s_i - t_j| \log p} \, d_n^\star \left(c \sqrt{\frac{\log p}{|s_i - t_j|}} + M_\Phi d_n^\star \frac{n\gamma_n}{|s_i - t_j|} \right) \\
&= O_p \left(n\gamma_n {d_n^\star}^2 \right). \qquad (36)
\end{aligned}$$

Applying a similar argument to the smaller sub-segment (s_{i-1}, t_j), we get:

$$\begin{aligned}
\sum_{t=s_{i-1}}^{t_j-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 &\leq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 + c_3 |t_j - s_{i-1}| \, ||\Phi^{(.,j)} - \widehat{\theta}||_2^2 \\
&+ c' \sqrt{|t_j - s_{i-1}| \log p} \, ||\Phi^{(.,j)} - \widehat{\theta}||_1 \\
&\leq \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 + 2c_3 |t_j - s_{i-1}| \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_2^2 + ||\Phi^{(.,j+1)} - \Phi^{(.,j)}||_2^2 \right) \\
&+ c' \sqrt{|t_j - s_{i-1}| \log p} \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_1 + ||\Phi^{(.,j+1)} - \Phi^{(.,j)}||_1 \right) \\
&= \sum_{t=s_{i-1}}^{t_j-1} ||\varepsilon_t||_2^2 + O_p \left(n \gamma_n {d_n^\star}^2 \right). \qquad (37)
\end{aligned}$$

Finally,

$$\begin{aligned}
\eta_{(s_{i-1}, s_i)} ||\widehat{\theta}||_1 &\leq \eta_{(s_{i-1}, s_i)} \left(||\Phi^{(.,j+1)} - \widehat{\theta}||_1 + ||\Phi^{(.,j+1)}||_1 \right) \\
&= O_p(d_n^\star). \qquad (38)
\end{aligned}$$

Combining (34)–(38) leads to:

$$\sum_{t=s_{i-1}}^{s_i-1} ||y_t - \widehat{\theta} Y_{t-1}||_2^2 + \eta_{(s_{i-1}, s_i)} ||\widehat{\theta}||_1 = \sum_{t=s_{i-1}}^{s_i-1} ||\varepsilon_t||_2^2 + O_p \left(n \gamma_n {d_n^\star}^2 \right). \qquad (39)$$

Adding these equations over all $m_0 + 1$ segments leads to equation 33.

Now, applying Lemma 4, we get:

$$\begin{aligned}
IC(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}) &= L_n(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}; \eta_n) + \widetilde{m}\omega_n \\
&> \sum_{t=q}^{T} ||\varepsilon_t||_2^2 + c_1\Delta_n - c_2\widetilde{m}n\gamma_n {d_n^\star}^2 + \widetilde{m}\,\omega_n \\
&\geq L(\widehat{t}_1, ..., \widehat{t}_{m_0}; \eta_n) + m_0\omega_n + c_1\Delta_n - c_2 m_0 n\gamma_n d_n^\star - (m_0 - \widetilde{m})\omega_n \\
&\geq L(\widehat{t}_1, ..., \widehat{t}_{m_0}; \eta_n) + m_0\omega_n, \qquad (40)
\end{aligned}$$

since $\lim_{n\to\infty} n\gamma_n {d_n^\star}^2/\omega_n \leq 1$, and $\lim_{n\to\infty} m_0\omega_n/\Delta_n = 0$. This proves part (a). To prove part (b), note that a similar argument as in Lemma 4 shows that

$$L_n(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}; \eta_n) \geq \sum_{t=q}^{T} ||\varepsilon_t||_2^2 - c_2\widetilde{m}n\gamma_n {d_n^\star}^2. \qquad (41)$$

A comparison between $IC(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}})$ and $IC(\widehat{t}_1, ..., \widehat{t}_{m_0})$ leads to:

$$\begin{aligned}
\sum_{t=q}^{T} ||\varepsilon_t||_2^2 - c_2\widetilde{m}n\gamma_n {d_n^\star}^2 + m\omega_n &\leq IC(\widetilde{t}_1, ..., \widetilde{t}_{\widetilde{m}}) \\
&\leq IC(\widehat{t}_1, ..., \widehat{t}_{m_0}) \\
&\leq \sum_{t=q}^{T} ||\varepsilon_t||_2^2 + K m_0 n\gamma_n {d_n^\star}^2 + m_0\omega_n, \qquad (42)
\end{aligned}$$

which means

$$(\widetilde{m} - m_0)\omega_n \leq c_2\widetilde{m}n\gamma_n {d_n^\star}^2 + K m_0 n\gamma_n {d_n^\star}^2. \qquad (43)$$

However, (43) contradicts with the fact that $m_0 n\gamma_n {d_n^\star}^2/\omega_n \to 0$. This completes the first part of the theorem.

For the second part, let $B = 2K/c$, and suppose there exists a point t_i such that $\min_{1\le j\le m_0} |\widetilde{t}_j - t_j| \ge Bm_0 n\gamma_n {d_n^\star}^2$. Then, by similar argument as in Lemma 4, we can show that:

$$\begin{aligned} \sum_{t=d}^{T} ||\varepsilon_t||_2^2 + cBm_0 n\gamma_n {d_n^\star}^2 &< L_n(\widetilde{t}_1, ..., \widetilde{t}_{m_0}) \\ &\le L_n(\widehat{t}_1, ..., \widehat{t}_{m_0}) \\ &\le \sum_{t=q}^{T} ||\varepsilon_t||_2^2 + Km_0 n\gamma_n {d_n^\star}^2, \end{aligned} \tag{44}$$

which contradicts with the way B was selected. This completes the proof of the theorem. □

Proof of Theorem 4. The proof of this theorem is similar to that of Proposition 4.1 in Basu & Michailidis (2015). The two main components of the proof is (i) verifying the restricted eigenvalue (RE) for $\hat{\Gamma} = I_p \otimes (\mathcal{X}_{\mathbf{r}}' \mathcal{X}_{\mathbf{r}}/N)$, and (ii) verifying the deviation bound for $\left|\left|\hat{\gamma} - \hat{\Gamma}\Phi\right|\right|_\infty$ 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 & Michailidis (2015). Therefore, here we proof (i) and (ii) only.

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

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

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 & 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

in Basu & Michailidis (2015), we have for any $v \in \mathbb{R}^{pq}$ with $||v||_2 \leq 1$, and any $\eta > 0$:

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

Now, to make the above probability hold uniformly on all the vectors v, we apply the discretization Lemma F2 in Basu & Michailidis (2015) and also Lemma 12 in the supplementary materials of Loh & Wainwright (2012) to get:

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

with high 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 $\Delta_n \geq \varepsilon n$ implies that $N_i/N \geq \varepsilon \geq 2\alpha$.

The deviation condition (DC) here means that there exist a large enough constant $C' > 0$ such that

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

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

$$\left|\left|\hat{\gamma}-\hat{\Gamma}\Phi\right|\right|_{\infty}=\max_{1\le k,l\le p;1\le h\le(m_0+1)d}\left|e_k'\mathcal{X}_{\mathbf{r},(h)}'E_{\mathbf{r}}e_l\right|.$$

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

$$\mathbb{P}\left(\left|e_k'\mathcal{X}_{\mathbf{r},(h)}'E_{\mathbf{r}}e_l\right|>k_1\eta\right)\le 6\exp(-k_2N\min(\eta^2,\eta)),$$

for large enough $k_1, k_2 > 0$, and any $\eta > 0$. Now, setting $\eta = C'\sqrt{\frac{\tilde{\pi}}{N}}$, 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 Theorem 1 in Appendix F. The proof is similar to the proof of Theorem 4, and therefore, most of the details are omitted here. The main difference is on finding a new deviation bound for the misspecified model in intervals of type $[\tilde{t}_j, t_j]$ or $[t_j, \tilde{t}_j]$, $j = 1, 2, \ldots, m_0$, and that would potentially affect the optimal selection of tuning parameter from theoretical perspective. In other words, we need a higher value for the tuning parameter to account for the model misspecification. More specifically, the additional term will be of the form

$$\begin{aligned}
& \frac{1}{n}\sum_{t=t_j}^{\tilde{t}_j-1} Y_{t-1}'\Big(\Phi^{(.,j)}-\widehat{\mathbf{B}}_{f,j}\Big)'\left(y_t-\Phi^{(.,j)}Y_{t-1}\right) \\
= \;& \frac{1}{n}\sum_{t=t_j}^{\tilde{t}_j-1} Y_{t-1}'\Big(\Phi^{(.,j)}-\widehat{\mathbf{B}}_{f,j}\Big)'\left(\varepsilon_t+\left(\Phi^{(.,j+1)}-\Phi^{(.,j)}\right)Y_{t-1}\right) \\
= \;& O_p\left(\sqrt{\frac{\log\tilde{\pi}}{n}}+d_n^{\star}M_{\Phi}\frac{R_n}{n}\right)\left\|\Phi^{(.,j)}-\widehat{\mathbf{B}}_{f,j}\right\|_1 . \qquad (45)
\end{aligned}$$

Therefore, the tuning parameter $\rho_{n,f}$ must be of the same order as in (45) so that the deterministic arguments used in Proposition 4.1 of Basu & Michailidis (2015) can go through here as well. Once the tuning parameter selection is adjusted, the rest of the proof is the same as the proof of Theorem 4.

□

Appendix C: Details of Estimation Algorithms

In this section, we provide details of the algorithm for solving the optimization problem (5), as well as the proposed backward elimination algorithm (BEA) for the second-stage screening.

To describe the algorithm for solving the optimization problem (5), 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 $|x|\le\lambda$. Recall that throughout the paper, for a $m\times n$ matrix A, $\|A\|_\infty=\max_{1\le i\le m,1\le j\le n}|a_{ij}|$. The algorithm is as follows:

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

(ii) For each $i = 1, \dots, n$, calculate the $(h+1)$–th iteration of the parameters $\theta_i^{(h+1)}$ using the KKT conditions of problem (5), presented in Lemma 2 of Appendix A. More specifically,

$$\theta_i'^{(h+1)} = \left(\sum_{l=i}^{n} Y_{l-1} Y_{l-1}' \right)^{-1} S \left(\sum_{l=i}^{n} Y_{l-1} y_l' - \sum_{j \neq i} \left(\sum_{l=\max(i,j)}^{n} Y_{l-1} Y_{l-1}' \right) \theta_j'^{(h)}; \lambda_{1,n} \right), \tag{46}$$

where $Y_l' = \left(y_l' \dots y_{l-q+1}' \right)_{1 \times pq}$.

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

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

(iv) Apply soft-thresholding to the partial sums of $\Theta^{(\text{intermediate})}$, i.e. $\sum_{i=1}^{k} \theta_i^{(\text{intermediate})}$ to find the optimizer in equation (5). In other words, $\widehat{\theta}_1 = S\left(\theta_1^{(\text{intermediate})}; \lambda_{2,n}\right)$ and $\widehat{\theta}_k = S\left(\sum_{i=1}^{k} \theta_i^{(\text{intermediate})}; \lambda_{2,n}\right) - S\left(\sum_{i=1}^{k-1} \theta_i^{(\text{intermediate})}; \lambda_{2,n}\right)$ for $k = 2, 3, \dots, n$. Finally $\widehat{\Theta} = \left(\widehat{\theta}_1, \dots, \widehat{\theta}_n\right)$.

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

Our backward elimination algorithm (BEA) for the second-stage screening is as follows:

(i) Set $m = |\widehat{\mathcal{A}}_n|$. Let $\mathbf{s} = \{s_1, \dots, s_m\}$ be the selected points and define $W_m^\star = \text{IC}(s_1, \dots, s_m; \eta_n)$.

(ii) For each $i = 1, \ldots, m$, calculate $W_{m,i} = \mathrm{IC}(\mathbf{s} \backslash \{s_i\}; \eta_n)$. Define $W^\star_{m-1} = \min_i W_{m,i}$.

(iii) (a) If $W^\star_{m-1} > W^\star_m$, then no further reduction is needed. Return $\widehat{\mathcal{A}}_n$ as the estimated change points.

(b) If $W^\star_{m-1} \leq W^\star_m$, and $m > 1$, set $j = \mathrm{argmin}_i W_{m,i}$, set $\mathbf{s} = \mathbf{s} \backslash \{s_j\}$ and $m = m - 1$. Go to step (ii).

(c) If $W^\star_{m-1} \leq W^\star_m$ and $m = 1$, all selected points are removed. Return the empty set.

Appendix D: Additional Simulation Results

In this section, two additional simulation scenarios are described and the empirical results are reported.

Simulation Scenario 4 (Randomly structured Φ and break points close to the center). As in Scenario 1, in this case we set $t_1 = 100$ and $t_2 = 200$. However, the coefficients matrices are chosen to be randomly structured. The autoregressive coefficients for simulation scenarios 1 and 2 are displayed in Figure 1. The 1-off diagonal values for the three segments are -0.6, 0.75, and -0.8, respectively.However, the autoregressive coefficients for this scenario are chosen to be randomly structured as displayed in Figure 2.

The selected break points in this scenario are shown in the middle part of Figure 3. The mean and standard deviation of locations of the selected break points, relative to the sample size T, as well as the percentage of simulation runs where break points are correctly identified are shown in Table 1. The results suggest that, among all simulation scenarios, this setting, with randomly structured Φ's, is the most challenging for our method in terms of detecting the number of break points. In this setting, the detection rate drops to 99% compared to 100% in the previous scenarios, and the standard deviation of the selected break point locations are higher than the first scenario. The percentage of runs where true break points are within R_n-radius of the estimated points also drops to 96% compared to 100% in Scenarios 1, 2 and 3.

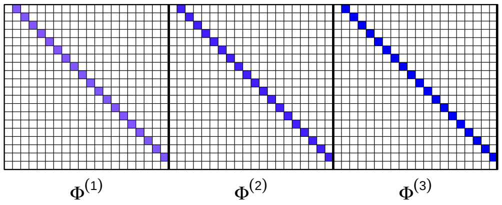

Figure 1: True autoregressive coefficients for the three segments in Simulation Scenarios 1 and 2.

The inferior performance of the proposed method in the fourth simulation scenario could be due to the fact that the L_2-distance between the consecutive autoregressive coefficients are less than the previous two cases. The L_2 norm of the consecutive differences of the VAR parameters in simulation 1 and 2 are 5.88 and 6.76 whereas in simulation 4, they are 4.44 and 4.49. This 35% reduction in the L_2-distance between the consecutive autoregressive coefficients would make it harder to identify the exact location of the break points. In contrast, the sparse changes in coefficient matrices makes this setting more favorable for SBS-MVTS. Nonetheless, estimates from our method are as good or better than those from SBS-MVTS.

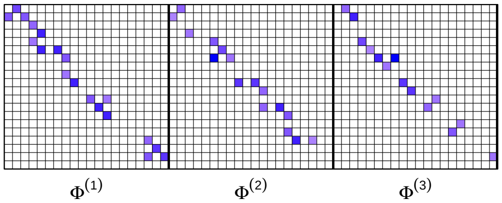

Figure 2: True autoregressive coefficients for the three segments in Simulation Scenario 4.

Table 1: Results for Simulation Scenario 4. The table shows mean and standard deviation of estimated break point locations, the percentage of simulation runs where break points are correctly detected (selection rate), and the percentage of simulation runs where true break points are within the R_n-radius of the estimated break points (R_n-selection rate).

method	break points	truth	mean	std	selection rate	R_n-selection rate
SBS-MVTS	1	0.3333	0.3238	0.0206	0.98	–
	2	0.6667	0.6569	0.0324	0.92	–
Our method	1	0.3333	0.3323	0.0124	0.99	0.98
	2	0.6667	0.6620	0.0200	0.99	0.96

Table 2 summarizes the results for autoregressive parameter estimation in this simulation scenarios. The table shows mean and standard deviation of relative estimation error (REE), as well as true positive (TPR) and false positive rates (FPR) of the estimates. The results suggest that the proposed method performs well in terms of parameter estimation. In simulation scenario 4, the performance of SBS–MVTS is better in estimation and in true positive rate. One reason for this good performance is that in this scenario, the selected break points of SBS–MVTS method are close enough to the true break points which makes it unnecessary to remove the R_n–radius of them in order to ensure stationarity. However, in real data applications, since the ground truth is unknown, this removal becomes necessary.

Table 2: Results of parameter estimation for simulation scenario 4. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

Method	REE	SD(REE)	TPR	FPR
Our Method	0.5263	0.0558	0.94	0.03
SBS-MVTS	0.2757	0.1099	1.00	0.04

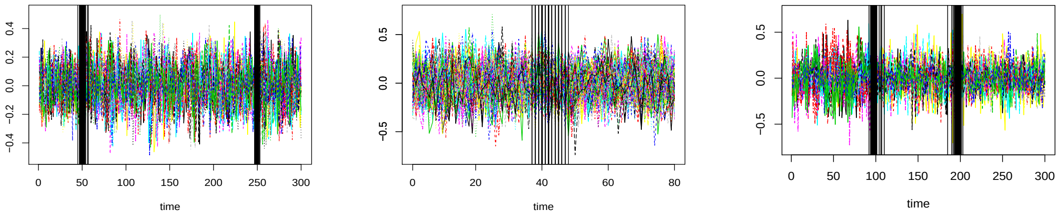

Figure 3: Estimated break points using our method for all 100 runs from Simulation Scenarios 2 (left), 3 (middle) and 4 (right).

Simulation Scenario 5 (Simple Φ and correlated error term). As in Scenario 1, in this case we set $t_1 = 100$ and $t_2 = 200$ with $T = 300$. The coefficients matrices are chosen to be the same as in simulation scenario 1 as displayed in Figure 1. However, the covariance matrix of the error terms is dense. More specifically, $\Sigma_\varepsilon = 0.01\left((\sigma_{ij})\right)_{T\times T}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The reason to add this simulation scenario is to see the effect of additional correlation structure of the noise term on the performance of our method both in terms of detection and estimation.

Table 3 reports the performance of our method and the SBS-MVTS developed in (Cho & Fryzlewicz 2015) in the simulation scenario 5 in terms of detection of break points. More specifically, the mean and standard deviation of locations of the selected break points, relative to the sample size T, as well as the percentage of simulation runs where break points are correctly identified are shown in Table 3. As

Table 3: Results for Simulation Scenario 5. The table shows mean and standard deviation of estimated break point locations, the percentage of simulation runs where break points are correctly detected (selection rate), and the percentage of simulation runs where true break points are within the R_n-radius of the estimated break points (R_n-selection rate).

method	break points	truth	mean	std	selection rate	R_n-selection rate
SBS-MVTS	1	0.3333	0.3688	0.0413	0.68	–
	2	0.6667	0.6119	0.0945	0.80	–
Our method	1	0.3333	0.3251	0.0139	1	1
	2	0.6667	0.6507	0.0213	1	1

Table 4: Results of parameter estimation for simulation scenario 5. The table shows mean and standard deviation of relative estimation error (REE), true positive rate (TPR), and false positive rate (FPR) for estimated coefficients.

Method	REE	SD(REE)	TPR	FPR
Our Method	0.6012	0.0699	0.93	0.04
SBS-MVTS	0.8005	0.1693	0.70	0.01

seen from this table, our method performs very well in this scenario which confirms the applicability of our method in the case of correlated error terms.

Table 4 summarizes the results for autoregressive parameter estimation in the simulation scenario 5. The table shows mean and standard deviation of relative estimation error (REE), as well as true positive (TPR) and false positive rates (FPR) of the estimates. The results suggest that the proposed method performs well in terms of parameter estimation and is superior to the naive approach using the detected points of SBS-MVTS method and applying regularization method for parameter estimation.

Appendix E: Sharpening the Consistency Rate

A potential area of improvement in the asymptotic analysis presented in the main paper is that the consistency rate for locating the break points stated in Theorem 3 depends on the the total sparsity of the model, $d_n^\star$, given by the sum of nonzero parameters in each segment. This number depends on the number of changed coefficients in consecutive segments, and can grow with the number of segments m_0. In this section, we propose an alternative analysis that reduces this rate by replacing the total sparsity with the sparsity of two consecutive segments, thus breaking the dependence on m_0.

To improve the rate of consistency, we take advantage of the fact that our method also gives some form of confidence intervals for the location of the break points. As discussed in Section 5, the interval $[\tilde{t}_j - \omega_n, \tilde{t}_j + \omega_n]$ includes the true break point t_j with high probability. We use this fact to repeat the proposed algorithm locally in order to achieve a better rate of consistency. More specifically, consider the intervals $F_j = [\tilde{t}_{j-1} + \omega_n, \tilde{t}_{j+1} - \omega_n]$, $j = 1, 2, \ldots, m_0$ with $\tilde{t}_0 = 1$ and $\tilde{t}_{m_0+1} = T$. These intervals not only include the true break points t_j but also t_j's are far from the boundaries. Thus, it is possible to apply our method on these intervals to locate the break points. In fact, we can just run the first step of our method to locate t_j since we know there is only *one* true break point in the interval F_j, and the tuning parameter can be selected in such a way that the method gives only one break point. Now, based on our results in Theorem 3, the consistency rate for locating t_j would be of order $O\left(d_j^2 n_j \gamma_{n_j}\right)$ where $n_j = \text{length}(F_j)$ and d_j is the sum of number of non-zero elements in the two consecutive segments j and $j+1$, $j = 1, 2, \ldots, m_0$. The new sparsity parameter d_j can be much smaller than $d_n^\star$ depending on the number of true break points m_0, and the number of changed coefficients in each break point. This result is summarized in the corollary below.

Corollary 1. *Suppose A1–A5 hold. Denote* $\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_{m_0}$ *the selected break points using the two-stage method as in* (5) *and* (12) *applied on the intervals* F_j, $j = 1, 2, \ldots, m_0$. *Then, as* $n \to +\infty$, *there exists a positive constant* $B > 0$ *such that*

$$\mathbb{P}\left(|\bar{t}_j - t_j| \leq B n_j \gamma_{n_j} {d_j}^2\right) \to 1,$$

for $j = 1, 2, \ldots, m_0$.

Appendix F: Alternative Procedure for Consistent Parameter Estimation

An alternative to our third step is to use the full data and estimate the parameters by using the selected break points as the end points in the piecewise high-dimensional regression model. More specifically, we can form the following linear regression:

$$\begin{pmatrix} y'_q \\ \vdots \\ y'_{\tilde{t}_1-1} \\ y'_{\tilde{t}_1} \\ \vdots \\ y'_{\tilde{t}_2-1} \\ \vdots \\ y'_{\tilde{t}_{m_0}-1} \\ \vdots \\ y'_T \end{pmatrix} = \begin{pmatrix} Y'_{q-1} & & & \\ \vdots & 0 & \dots & 0 \\ Y'_{\tilde{t}_1-2} & & & \\ & Y'_{\tilde{t}_1-1} & & \\ 0 & \vdots & \dots & 0 \\ & Y'_{\tilde{t}_2-2} & & \\ \vdots & \vdots & \ddots & \vdots \\ & & & Y'_{\tilde{t}_{m_0}-2} \\ 0 & 0 & & \vdots \\ & & & Y'_{T-1} \end{pmatrix} \begin{pmatrix} \beta'_1 \\ \beta'_2 \\ \vdots \\ \beta'_{m_0+1} \end{pmatrix} + \begin{pmatrix} \zeta'_q \\ \vdots \\ \zeta'_{\tilde{t}_1-1} \\ \zeta'_{\tilde{t}_1} \\ \vdots \\ \zeta'_{\tilde{t}_2-1} \\ \vdots \\ \zeta'_{\tilde{t}_{m_0}-1} \\ \vdots \\ \zeta'_T \end{pmatrix}. \qquad (47)$$

This regression can be written in compact form as

$$\mathcal{Y}_{\mathbf{f}} = \mathcal{X}_{\mathbf{f}} B + E_{\mathbf{f}}$$

or, in a vector form, as

$$\mathbf{Y_f} = \mathbf{Z_f B} + \mathbf{E_f} \qquad (48)$$

where $\mathbf{Y_f} = \mathrm{vec}(\mathcal{Y}_\mathbf{f})$, $\mathbf{Z_f} = I_p \otimes \mathcal{X}_\mathbf{f}$, $\mathbf{B} = \mathrm{vec}(B)$, $\mathbf{E_f} = \mathrm{vec}(E_\mathbf{f})$. Not that, $\mathbf{Y_f} \in \mathbb{R}^{np\times 1}$, $\mathbf{Z_f} \in \mathbb{R}^{np\times\tilde{\pi}}$, $\mathbf{B} \in \mathbb{R}^{\tilde{\pi}\times 1}$, and $\mathbf{E_f} \in \mathbb{R}^{np\times 1}$. Estimating the VAR parameters by solving

$$\widehat{\mathbf{B}}_\mathbf{f} = \mathrm{argmin}_\mathbf{B} N^{-1} \left\| \mathbf{Y_f} - \mathbf{Z_f B} \right\|_2^2 + \rho_{n,f} \left\| \mathbf{B} \right\|_1, \tag{49}$$

we obtain the following consistency result.

Theorem 1. *Suppose A1–A5 hold, m_0 is unknown and $R_n = Bm_0 n \gamma_n {d_n^\star}^2$. Assume also that $\Delta_n > \varepsilon n$ for some large positive $\varepsilon > 0$ and $\rho_{n,f} = C\left(\sqrt{\frac{\log\tilde{\pi}}{n}} + d_n^\star M_\Phi \frac{R_n}{n}\right)$ for large enough $C > 0$. Then, as $n \to +\infty$, the minimizer $\widehat{\boldsymbol{B}}_\boldsymbol{f}$ of* (49) *satisfies*

$$\left\| \widehat{\mathbf{B}}_\boldsymbol{f} - \Phi \right\|_\ell = O_p\left((d_n^\star)^{1/\ell} \rho_{n,f} \right) \text{ for } \ell = 1, 2.$$

As seen in Theorem 1, the consistency rate when using the full data has an additional term $d_n^\star M_\Phi \frac{R_n}{n}$. Depending on the magnitude of the sparsity level $d_n^\star$, this additional term may dominate the first term $\sqrt{\frac{\log\tilde{\pi}}{n}}$ asymptotically, and can lead to a worse consistency rate than the proposed estimator $\widehat{\mathbf{B}}$ in Equation (15).

The rate of consistency for the above procedure suggests that removing the R_n-radius of estimated break points in the third step of our procedure can play a crucial role. The simulation results in Section 7.3 corroborate this theoretical finding.

Appendix G: Data-Driven Method For Selecting ω_n

In this section, we describe, in details, the data-driven method used in Section 7.4 for choosing ω_n. As mentioned, the idea is to first finish the backward elimination algorithm (BEA) until no break points are left. Then, we cluster the *jumps* in the objective function L_n in Equation (10) into two subgroups, small and large. Intuitively, if removing a break point leads to a small jump in L_n, then the break point is likely redundant. In contrast, larger jumps correspond to true break points. The smallest jump in the second group is thus a reasonable candidate for ω_n. More specifically, we apply the BEA to the candidate break points in the first step, i.e. $\{\widehat{t}_1, \ldots, \widehat{t}_{\widehat{m}}\}$, and remove them one by one to the end. Then, we rank the break points based on the level they are removed in the BEA algorithm, say $\widehat{t}_{i_1}, \widehat{t}_{i_2}, \ldots, \widehat{t}_{i_{\widehat{m}}}$. Now, we compute the jumps in the objective function at each step of the BEA algorithm denoting them by v_k, with $v_k = \left|L_n(\widehat{t}_{i_k}, \ldots, \widehat{t}_{i_{\widehat{m}}}; \eta_n) - L_n(\widehat{t}_{i_{k-1}}, \ldots, \widehat{t}_{i_{\widehat{m}}}; \eta_n)\right|$, $k = 1, 2, \ldots, \widehat{m}$. Then, we cluster $V = \{v_1, v_2, \ldots, v_{\widehat{m}}\}$ into two subsets using k-means clustering (Hartigan & Wong 1979). Denote the subset with smaller center as the small subgroup, V_S, and the other subset as the large subgroup, V_L.

Intuitively, the points which their removal leads to jumps in V_S are redundant, and the rest, which made large jumps in the sum of squared error (SSE) in the BEA algorithm, correspond to estimated break points. Therefore, $\min V_L$—which is the smallest jump in the SSE occurred by removing a true estimated break point—is a reasonable candidate for ω_n. Note that if $m_0 = 0$, then all $\widehat{m}$ selected break points in the first step are redundant, and their removal make no significant jumps in the SSE. To avoid keeping any points after the screening step in this case, we recommend comparing the fit in the k-means clustering with two and one cluster. If k-means with two clusters has a high ratio of between-group SS/total SS, then we proceed as discussed. Otherwise, we remove all the $\widehat{m}$ selected break points in the first step, and claim that $m_0 = 0$. In the latter, $\max V$ is considered as the optimal value for ω_n. The proposed algorithm is summarized as follow:

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

(ii) For each $k = 1, 2, \ldots, \widehat{m}$, set $v_k = \left| L_n(\widehat{t}_{i_k}, \ldots, \widehat{t}_{i_{\widehat{m}}}; \eta_n) - L_n(\widehat{t}_{i_{k-1}}, \ldots, \widehat{t}_{i_{\widehat{m}}}; \eta_n) \right|$. Define $V = \{v_1, v_2, \ldots, v_{\widehat{m}}\}$.

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

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

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

References

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

Chan, Ngai Hang, Yau, Chun Yip, & Zhang, Rong-Mao. 2014. Group LASSO for structural break time series. *Journal of the American Statistical Association*, **109**(506), 590–599.

Cho, Haeran, & Fryzlewicz, Piotr. 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.

Harchaoui, Zaıd, & Lévy-Leduc, Céline. 2010. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, **105**(492), 1480–1493.

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

Loh, Po-Ling, & Wainwright, Martin J. 2012. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, **40**(3), 1637–1664.