Collective anomaly detection in High-dimensional VAR Models

Hyeyoung Maeng[1], Idris A. Eckley[2] and Paul Fearnhead[2]

[1]*Durham University and* [2]*Lancaster University*

Supplementary Material

This document includes the following sections:

S1. Extensions to VAR(q) model and multiple anomaly detection

S2. Proofs

S3. Additional simulation results

S4. Additional results for yellow cab demand data

S1 Extensions to VAR(q) and multiple anomaly detection

S1.1 Extension to VAR(q) model

The VAR process of order 1 presented in Section 2.1 can simply be extended to VAR(q) as follows,

$$\boldsymbol{x}_t = \boldsymbol{A}_{t,1}\boldsymbol{x}_{t-1} + \cdots + \boldsymbol{A}_{t,q}\boldsymbol{x}_{t-q} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma_\varepsilon), \quad t = 1, \ldots, T, \qquad \text{(S1.1)}$$

where $\{\boldsymbol{A}_{t,k}\}_{t=1}^{T}$ is a $p \times p$ matrix for all $k = 1, \ldots, q$ and Σ_ε is assumed to be a positive definite matrix. With a slight abuse of notation, the piecewise-constant coefficient matrices are as follows,

$$\boldsymbol{A}^{(1)} = (\boldsymbol{A}_{t',1}, \ldots, \boldsymbol{A}_{t',q}) \in \mathbb{R}^{p \times pq}, \quad \text{for any } t' = 1, \ldots, \eta_1 - 1, \eta_2 + 1, \ldots, T$$

$$\boldsymbol{A}^{(2)} = (\boldsymbol{A}_{t',1}, \ldots, \boldsymbol{A}_{t',q}) \in \mathbb{R}^{p \times pq}, \quad \text{for any } t' = \eta_1, \ldots, \eta_2,$$

and the model (S1.1) can be represented as

$$\begin{pmatrix} \boldsymbol{x}_q' \\ \boldsymbol{x}_{q+1}' \\ \vdots \\ \boldsymbol{x}_T' \end{pmatrix}_{K \times p} = \begin{pmatrix} \boldsymbol{x}_{q-1}' & \cdots & \boldsymbol{x}_0' & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \boldsymbol{x}_{\eta_1+q-3}' & \cdots & \boldsymbol{x}_{\eta_1-2}' & 0 & \cdots & 0 \\ \boldsymbol{x}_{\eta_1+q-2}' & \cdots & \boldsymbol{x}_{\eta_1-1}' & \boldsymbol{x}_{\eta_1+q-2}' & \cdots & \boldsymbol{x}_{\eta_1-1}' \\ \vdots & & \vdots & \vdots & & \vdots \\ \boldsymbol{x}_{\eta_2+q-2}' & \cdots & \boldsymbol{x}_{\eta_2-1}' & \boldsymbol{x}_{\eta_2+q-2}' & \cdots & \boldsymbol{x}_{\eta_2-1}' \\ \boldsymbol{x}_{\eta_2+q-1}' & \cdots & \boldsymbol{x}_{\eta_2}' & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \boldsymbol{x}_{T-1}' & \cdots & \boldsymbol{x}_{T-q}' & 0 & \cdots & 0 \end{pmatrix}_{K \times 2pq} \begin{pmatrix} \boldsymbol{\theta}^{(1)\prime} \\ \boldsymbol{\theta}^{(2)\prime} \end{pmatrix}_{2pq \times p} + \begin{pmatrix} \boldsymbol{\varepsilon}_q' \\ \boldsymbol{\varepsilon}_{q+1}' \\ \vdots \\ \boldsymbol{\varepsilon}_T' \end{pmatrix}_{K \times p}, \tag{S1.2}$$

where $K = T - q + 1$, $\boldsymbol{\theta}^{(1)} = \boldsymbol{A}^{(1)}$, $\boldsymbol{\theta}^{(2)} = \boldsymbol{A}^{(2)} - \boldsymbol{A}^{(1)}$. With the larger dimension of the parameters, the same argument for the VAR process of order q can be achieved by following the logic from (2.4).

S1.2 Extension to multiple anomaly detection

Following the ideas in Fryzlewicz (2014) and Kovács et al. (2020), to deal with multiple anomalies, we repeatedly update the candidate set by removing the intervals that overlap with any detected anomalies. The detailed procedure is given in Algorithm 3.

Algorithm 3: Multiple anomaly detection

INPUT: $\boldsymbol{X}$ matrix in (2.4), L, λ^{thr}

Step 1: Set a collection of intervals $\mathbb{J}_{T,p}(L)$ where L is the minimum length of intervals.

Step 2: For any interval $J \in \mathbb{J}_{T,p}(L)$, calculate $T^{\text{lasso}}(J)$ as in (2.7).

Step 3: Using a pre-specified threshold λ^{thr}, pick the candidate set

$$\mathbb{I}^* = \left\{J \in \mathbb{J}_{T,p}(L) : T^{\text{lasso}}(J) > \lambda^{\text{thr}}\right\}.$$

If $\mathbb{I}^* \neq \emptyset$, reject the null hypothesis (no anomaly exist). Set $\mathbb{I}^{(1)} = \mathbb{I}^*$, $j = 1$ and proceed the following steps.

while $\mathbb{I}^{(j)} \neq \emptyset$ **do**

Step 4: Save the estimator of the anomaly interval,

$$\hat{I}_j = \underset{J \in \mathbb{I}^{(j)}}{\operatorname{argmax}}\, T^{\text{lasso}}(J),$$

and update the candidate set as

$$\mathbb{I}^{(j+1)} = \mathbb{I}^{(j)} \setminus \{J : J \in \mathbb{I}^{(j)}, J \cap \hat{I}_j \neq \emptyset\}$$

Step 5: Set $j = j + 1$.

end

OUTPUT: $\hat{I} = \{\hat{I}_1, \hat{I}_2, \cdots\}$.

S2 Proofs

We first give a preparatory lemma and then move onto the proofs of main theorems and corollary presented in Section 3.

Lemma 1. *Let Assumptions 1-5 hold. For any* $J \in \mathbb{J}_{T,p}(L)$ *such that* $J \subseteq [\eta_1, \eta_2]$ *and any* $|J| \gtrsim c_1^{\mathfrak{m},\mathcal{M}} \log p$*, with probability at least* $1 - T^{-6}$*, we have*

$$\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \boldsymbol{X}_J \boldsymbol{\Theta} \geq c_2^{\mathfrak{m},\mathcal{M}} |J| \cdot \|\boldsymbol{\Theta}\|_2^2 - c_3^{\mathfrak{m},\mathcal{M}} \log(p) \cdot \|\boldsymbol{\Theta}\|_1^2$$

where $c_1^{\mathfrak{m},\mathcal{M}}, c_2^{\mathfrak{m},\mathcal{M}}, c_3^{\mathfrak{m},\mathcal{M}} > 0$ *are some constants depending on* $\mathfrak{m}$ *and* $\mathcal{M}$*.*

Proof of Lemma 1 The argument follows the proof of Lemma 13-(b) of Wang et al. (2019).

Proof of Theorem 1 By the construction of the candidate set $\mathbb{I}^*$, it is sufficient to show that $T^{\text{lasso}}(J) \to 0$ under the null, where $T^{\text{lasso}}(J)$ is as in (2.11). The KKT conditions for the lasso problem in (2.11) is that any $\hat{\boldsymbol{\Theta}}$ is optimal if and only if there exists a subgradient $\hat{s}$ such that

$$\boldsymbol{X}_J^\top \left(\boldsymbol{Y}_J - \boldsymbol{X}_J \hat{\boldsymbol{\Theta}} \right) = \lambda \hat{s}_J, \tag{S2.3}$$

where $\hat{s}_J = \partial\left|\hat{\boldsymbol{\Theta}}\right|_1$ is a subgradient of the l_1 norm evaluated at $\hat{\boldsymbol{\Theta}}$ which takes the form

$$\hat{s}_J = \text{sgn}(\hat{\boldsymbol{\Theta}}) \text{ for } \hat{\boldsymbol{\Theta}} \neq 0, \quad |\hat{s}_J| \leq 1 \text{ otherwise.} \tag{S2.4}$$

As $\boldsymbol{Y} = \boldsymbol{E}$ under the null, (S2.3) and (S2.4) give a condition on $\boldsymbol{X}$ and $\boldsymbol{E}$ to ensure that we estimate $\boldsymbol{\Theta} = \boldsymbol{0}$ as follows: for any $J \in \mathbb{J}_{T,p}(L)$ such that $J \cap [\eta_1, \eta_2] = \emptyset$,

$$\max_{J: J \in \mathbb{J}_{T,p}(L), J \cap [\eta_1, \eta_2] = \emptyset} \left\| \boldsymbol{X}_J^\top \boldsymbol{E}_J \right\|_\infty \leq \lambda. \tag{S2.5}$$

We remind that $\mathcal{X}_J$ is the unvectorised covariates as follows:

$$\mathcal{X}_J = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \boldsymbol{x}'_{t+1} \\ \vdots \\ \boldsymbol{x}'_{t+h-1} \end{pmatrix}_{(2h-1)\times p}$$

and note that

$$\begin{aligned} \max_{J: J \in \mathbb{J}_{T,p}(L), J \cap [\eta_1, \eta_2] = \emptyset} \left\| \frac{\boldsymbol{X}_J^\top \boldsymbol{E}_J}{|J|} \right\|_\infty &= \max_{\{J: J \in \mathbb{J}_{T,p}(L), J \cap [\eta_1, \eta_2] = \emptyset\}, 1 \leq i, j \leq p} \left| e_i' \left(\frac{\mathcal{X}_J' E_J}{|J|} \right) e_j \right|, \\ &\leq \max_{\{J: J \in \mathbb{J}_{T,p}(L), J \cap [\eta_1, \eta_2] = \emptyset\}, 1 \leq i, j \leq p} \left| e_i' \left(\frac{\mathcal{X}_J' E_J}{L} \right) e_j \right|, \end{aligned} \tag{S2.6}$$

where $e_i \in \mathbb{R}^p$ with the i-th element equals to 1 and zero otherwise. Similar to the argument used in Proposition 2.4(b) of Basu and Michailidis (2015), for fixed i, j, J, there exist $k_1, k_2 > 0$ such that for all $\gamma > 0$:

$$P\left(\left|e_i'\left(\mathcal{X}_J' E_J\right)e_j\right| > k_1 L\gamma\right) \leq 6\exp(-k_2 L\min(\gamma, \gamma^2)). \tag{S2.7}$$

As the number of intervals contained in $\mathbb{J}_{T,p}(L)$ is of the order $O(T)$ when they are constructed through the seeded interval idea in Kovács et al. (2020), we consider the union over $p^2 \cdot T$ possible choices of i, j, J in (S2.6). Then the result follows by setting $\gamma = k_3\sqrt{\frac{2\log p + \log T}{L}}$ for a large enough $k_3 > 0$. Therefore, with probability at least $1 - C_4\exp(-C_5(2\log p + \log T))$, we have

$$\max_{J: J\in\mathbb{J}_{T,p}(L), J\cap[\eta_1,\eta_2]=\emptyset} \left\|\boldsymbol{X}_J^\top \boldsymbol{E}_J\right\|_\infty \leq C_3\sqrt{L\log(T\vee p)}, \tag{S2.8}$$

where $C_4 > 0$ and $C_5 > 0$. Having the condition $\lambda = C_3\sqrt{L\log(T\vee p)}$ with a large enough $C_3 > 0$ in (S2.5), we obtain $\hat{\boldsymbol{\Theta}} = \boldsymbol{0}$ with probability at least $1 - C_4\exp(-C_5(2\log p + \log T))$. Therefore, under the null, the probability that $T^{\text{lasso}}(J) \to 0$ for $J \in \mathbb{J}_{T,p}(L)$ such that $J \cap [\eta_1, \eta_2] = \emptyset$ is at least $1 - C_4\exp(-C_5(2\log p + \log T))$ where $C_4, C_5 > 0$.

We emphasise that (S2.8) can be applied to any serially uncorrelated Gaussian errors $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \Sigma_\varepsilon)$ as the constant k_1 in (S2.7) presented in Proposition

2.4(b) of Basu and Michailidis (2015) has a form of

$$k_1 = 2\pi\Lambda_{\max}(\Sigma_\varepsilon)\Big(1 + \frac{1+\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})}\Big),$$

where $\Lambda_{\max}(\Sigma_\varepsilon)$ is the maximum eigenvalue of Σ_ε and

$$\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \Lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)), \quad \mu_{\min}(\mathcal{A}) = \min_{|z|=1} \Lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z)).$$

Note that $\mathcal{A}(z) = I_p - \mathbf{A}^{(1)}z$ for the VAR(1) model and $\mathcal{A}(z) = I_p - \sum_{d=1}^{q} \mathbf{A}_d^{(1)} z^d$ for the VAR(q) model. Therefore, even if Σ_ε is not an identity matrix, we can have (S2.8) with a different constant C_3 which depends on the maximum eigenvalue of Σ_ε.

Proof of Theorem 2 It is sufficient to prove that for any $J \in \mathbb{J}_{T,p}(L)$ such that $J \subseteq [\eta_1, \eta_2]$, with probability approaching to 1 as $T \to \infty$,

$$\left\|\boldsymbol{Y}_J\right\|_2^2 - \left\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\right\|_2^2 > \lambda||\boldsymbol{\Theta}||_1 + \lambda^{\text{thr}}. \tag{S2.9}$$

This is because the other part in equation (17),

$$\left\{\left\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\right\|_2^2 + \lambda||\boldsymbol{\Theta}||_1 - \left\|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\right\|_2^2 - \lambda||\hat{\boldsymbol{\Theta}}||_1\right\}, \tag{S2.10}$$

is always positive and the left-hand side of (S2.9) dominates (S2.10). We can simplify (S2.9) as

$$\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \boldsymbol{X}_J \boldsymbol{\Theta} + 2\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \boldsymbol{E}_J > \lambda ||\boldsymbol{\Theta}||_1 + \lambda^{\text{thr}}. \tag{S2.11}$$

The left-hand side of (S2.11) is a Gaussian variable that can be written as $\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J + 2\boldsymbol{\nu}_J^\top \boldsymbol{E}_J \sim N(\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J, 4\boldsymbol{\nu}_J^\top \Sigma_\varepsilon \boldsymbol{\nu}_J)$, where $\boldsymbol{\nu}_J = \boldsymbol{X}_J \boldsymbol{\Theta}$ and $\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J \to \infty$. Then for any $g(J) = o(\sqrt{\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J})$ that goes to ∞, we have the following bound with probability approaching to 1,

$$\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J + 2\boldsymbol{\nu}_J^\top \boldsymbol{E}_J \geq \boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J - g(J)\sqrt{4\gamma \boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J}, \tag{S2.12}$$

where γ is the maximum eigenvalue of Σ_ε. The right-hand side of (S2.12) is of order $\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J$, thus we now show that $\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \boldsymbol{X}_J \boldsymbol{\Theta}$ is bounded by $\lambda ||\boldsymbol{\Theta}||_1 + \lambda^{\text{thr}}$ with probability tending to 1. From Lemma 1, with probability approaching to 1, we have

$$\begin{aligned} \boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \boldsymbol{X}_J \boldsymbol{\Theta} &\geq c_2^{\mathfrak{m},\mathcal{M}} |J| \cdot ||\boldsymbol{\Theta}||_2^2 - c_3^{\mathfrak{m},\mathcal{M}} \log(p) \cdot ||\boldsymbol{\Theta}||_1^2 \\ &\geq c_2^{\mathfrak{m},\mathcal{M}} L \cdot ||\boldsymbol{\Theta}||_2^2 - c_3^{\mathfrak{m},\mathcal{M}} \log(p) \cdot ||\boldsymbol{\Theta}||_1^2, \end{aligned}$$

where $c_2^{\mathfrak{m},\mathcal{M}}, c_3^{\mathfrak{m},\mathcal{M}} > 0$, thus we now show

$$c_2^{\mathfrak{m},\mathcal{M}}\|\boldsymbol{\Theta}\|_2^2 > c_3^{\mathfrak{m},\mathcal{M}}\frac{\log(p)}{L}\cdot\|\boldsymbol{\Theta}\|_1^2 + \frac{\lambda}{L}\|\boldsymbol{\Theta}\|_1 + \frac{\lambda^{\text{thr}}}{L}, \tag{S2.13}$$

as $T, p \to \infty$. We can obtain (S2.13) as $T, p \to \infty$ from combining

$$\begin{aligned}
&(a)\ c_2^{\mathfrak{m},\mathcal{M}}\|\boldsymbol{\Theta}\|_2^2 > c_3^{\mathfrak{m},\mathcal{M}}\frac{\log(p)}{L}\cdot\|\boldsymbol{\Theta}\|_1^2,\\
&(b)\ c_2^{\mathfrak{m},\mathcal{M}}\|\boldsymbol{\Theta}\|_2^2 > \frac{\lambda}{L}\|\boldsymbol{\Theta}\|_1,\\
&(c)\ c_2^{\mathfrak{m},\mathcal{M}}\|\boldsymbol{\Theta}\|_2^2 > \frac{\lambda^{\text{thr}}}{L},
\end{aligned}$$

where (a) can be shown by using $d_0\|\boldsymbol{\Theta}\|_2^2 \geq \|\boldsymbol{\Theta}\|_1^2$ from Assumption 4 and $\frac{\log p}{L} \to 0$ from Assumption 3. By using $d_0\|\boldsymbol{\Theta}\|_2^2 \geq \|\boldsymbol{\Theta}\|_1^2$, (b) becomes $c_2^{\mathfrak{m},\mathcal{M}}\|\boldsymbol{\Theta}\|_2 > \frac{\lambda}{L}\sqrt{d_0}$ that can be achieved from $\frac{\lambda}{L} = \sqrt{\frac{C_3\log(T\vee p)}{L}}$ and $\|\boldsymbol{\Theta}\|_2^2 > C_2\cdot\frac{\log^{1+\xi}(T\vee p)}{L}$ in Assumption 5. Similarly (c) can be obtained from $\frac{\lambda^{\text{thr}}}{L} = O\left(\sqrt{\frac{\log(T\vee p)}{L}}\right)$ and Assumption 5.

We now consider the case Σ_ε is not an identity matrix. In that case, (S2.9) becomes

$$\boldsymbol{Y}_J^\top\Sigma_\varepsilon^{-1}\boldsymbol{Y}_J - (\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta})^\top\Sigma_\varepsilon^{-1}(\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}) > \lambda\|\boldsymbol{\Theta}\|_1 + \lambda^{\text{thr}},$$

thus (S2.11) becomes

$$\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \Sigma_\varepsilon^{-1} \boldsymbol{X}_J \boldsymbol{\Theta} + 2\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \Sigma_\varepsilon^{-1} \boldsymbol{E}_J > \lambda \|\boldsymbol{\Theta}\|_1 + \lambda^{\text{thr}}, \tag{S2.14}$$

which holds by following the same argument used above with $\boldsymbol{\nu}_J = \Sigma_\varepsilon^{-1/2} \boldsymbol{X}_J \boldsymbol{\Theta}$ and different constants, as the left-hand side of (S2.14) is a Gaussian random variable bounded by a component that is of order $\boldsymbol{\nu}_J^\top \boldsymbol{\nu}_J$.

Lastly, without repeating all the proofs, we argue that the theory we present for the known Σ_ε can be applied to the case when an estimate of Σ_ε is used. If $\hat{\Sigma}_\varepsilon$ is used instead of Σ_ε, the left-hand side of (S2.14) can be rewritten as

$$\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \Sigma_\varepsilon^{-1} \boldsymbol{X}_J \boldsymbol{\Theta} + 2\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \Sigma_\varepsilon^{-1} \boldsymbol{E}_J + \boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top (\hat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}) \boldsymbol{X}_J \boldsymbol{\Theta} + 2\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top (\hat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}) \boldsymbol{E}_J, \tag{S2.15}$$

thus the test depends on the eigenvalues of the measure of the distance between $\hat{\Sigma}_\varepsilon^{-1}$ and Σ_ε^{-1}. If $\hat{\Sigma}_\varepsilon^{-1}$ converges to Σ_ε^{-1} as observation increases, the last two terms in (S2.15) become under control, thus we can obtain the same argument with extra constant terms.

Proof of Theorem 3 It is straightforward that the test statistic of the OLS method in (10) has a $\chi^2_{p^2}$ distribution, where the degrees of freedom p^2 comes from the difference in dimensionality of Θ_0 and $\hat{\Theta}$. Therefore, we get an asymptotic level

α test if the null hypothesis is rejected for $T(J) > \chi^2_{p^2;(1-\alpha)}$, where $\chi^2_{p^2;(1-\alpha)}$ is the $(1-\alpha)$-quantile of chi-square distribution with p^2 degrees of freedom. Using the threshold established above, under the alternative, an upper bound on the power of the OLS method can be obtained as

$$
\begin{aligned}
&P\Big(T(J) > \chi^2_{p^2;(1-\alpha)}\Big) \\
&= P\Big(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \Big\{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2\Big\} \geq \chi^2_{p^2;(1-\alpha)}\Big) \\
&= P\Big(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \Big\{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2\Big\} - p^2 \geq \chi^2_{p^2;(1-\alpha)} - p^2\Big) \qquad \text{(S2.16)} \\
&\leq \frac{E(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2)}{\chi^2_{p^2;(1-\alpha)} - p^2} \qquad \text{(S2.17)} \\
&\approx \frac{E(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2)}{\frac{1}{2}\Big(z_{1-\alpha} + \sqrt{2p^2-1}\Big)^2 - p^2}, \qquad \text{(S2.18)}
\end{aligned}
$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of Gaussian distribution. The equality in (S2.16) is obtained by subtracting $E\Big\{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2\Big\} = p^2$ from both sides, the inequality in (S2.17) is obtained by using Markov's inequality and (S2.18) is achieved as the quantile of chi-square distribution has an approximation, $\chi^2_{p^2;(1-\alpha)} \approx \frac{1}{2}\Big(z_{1-\alpha} + \sqrt{2p^2-1}\Big)^2$. Therefore, the upper bound on the power of the OLS method can be obtained as in (15), which implies that $E(\|\boldsymbol{Y}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_2^2)$ needs to be at least $O_p(p)$ to have power approaching to 1.

Proof of Corollary 1 We first first give the proof corresponding to Theorem 1. As we have $\boldsymbol{Y} = \boldsymbol{E} + \text{vec}(\mathcal{X}^{(1)}(\boldsymbol{\theta}^{(1)} - \hat{\boldsymbol{\theta}}^{(1)\prime}))$ under the null rather than $\boldsymbol{Y} = \boldsymbol{E}$, the right-hand side of the inequality in (S2.6) can be represented as

$$\begin{aligned}
&\max_{\{J:J\in\mathbb{J}_{T,p}(L),J\cap[\eta_1,\eta_2]=\emptyset\},1\le i,j\le p}\left|e_i'\left(\frac{\mathcal{X}_J^{(2)\prime}\left(E_J+\mathcal{X}_J^{(1)}(\boldsymbol{\theta}^{(1)\prime}-\hat{\boldsymbol{\theta}}^{(1)\prime})\right)}{L}\right)e_j\right| \\
&\le \max_{\{J:J\in\mathbb{J}_{T,p}(L),J\cap[\eta_1,\eta_2]=\emptyset\},1\le i,j\le p}\left|e_i'\left(\frac{\mathcal{X}_J^{(2)\prime}E_J}{L}\right)e_j\right| + \max_{\{J:J\in\mathbb{J}_{T,p}(L),J\cap[\eta_1,\eta_2]=\emptyset\},1\le i,j\le p}\left|e_i'\left(\frac{\mathcal{X}_J^{(2)\prime}\mathcal{X}_J^{(1)}(\boldsymbol{\theta}^{(1)\prime}-\hat{\boldsymbol{\theta}}^{(1)\prime})}{L}\right)e_j\right|,
\end{aligned} \tag{S2.19}$$

It is sufficient to show that both terms in (S2.19) are less than or equal to $C_3\lambda$ with probability approaching 1. The condition for the first term is obtained from the proof of Theorem 1 and the one for the second term is obtained from Assumption 6 and from the fact that $\frac{\mathcal{X}_J^{(2)\prime}\mathcal{X}_J^{(1)}}{|J|}$ converges as $T\to\infty$.

Now we move onto the proof corresponding to Theorem 2. It is sufficient to prove that (S2.9) still holds where $\boldsymbol{Y}_J = \text{vec}(\mathcal{Y}_J - \mathcal{X}_J^{(1)}\boldsymbol{\theta}^{(1)\prime})$ is replaced by $\boldsymbol{Y}_J' = \text{vec}(\mathcal{Y}_J - \mathcal{X}_J^{(1)}\hat{\boldsymbol{\theta}}^{(1)\prime})$. The left-hand side of (S2.9) can be simplified as

$$\boldsymbol{\Theta}^\top\boldsymbol{X}_J^\top\boldsymbol{X}_J\boldsymbol{\Theta}+2\boldsymbol{\Theta}^\top\boldsymbol{X}_J^\top\boldsymbol{E}_J+2\boldsymbol{\Theta}^\top\boldsymbol{X}_J^\top\text{vec}\left(\mathcal{X}_J^{(1)}(\boldsymbol{\theta}^{(1)\prime}-\hat{\boldsymbol{\theta}}^{(1)\prime})\right) > \lambda\|\boldsymbol{\Theta}\|_1+\lambda^{\text{thr}}. \tag{S2.20}$$

As shown in the proof of Theorem 2, it is sufficient to show that $\boldsymbol{\Theta}^\top\boldsymbol{X}_J^\top\boldsymbol{X}_J\boldsymbol{\Theta}$ is bounded by $\lambda\|\boldsymbol{\Theta}\|_1+\lambda^{\text{thr}}$ with probability tending to 1 as the last component in

left-hand side of (S2.20),

$$2\boldsymbol{\Theta}^\top \boldsymbol{X}_J^\top \mathrm{vec}(\mathcal{X}_J^{(1)}(\boldsymbol{\theta}^{(1)\prime} - \hat{\boldsymbol{\theta}}^{(1)\prime})) = 2\boldsymbol{\Theta}^\top \mathrm{vec}\Big(\mathcal{X}_J^{(2)\prime}\mathcal{X}_J^{(1)}(\boldsymbol{\theta}^{(1)\prime} - \hat{\boldsymbol{\theta}}^{(1)\prime})\Big),$$

is less than $\lambda\|\boldsymbol{\Theta}\|_1 + \lambda^{\text{thr}}$ with probability approaching to 1 from Assumption 6 and also from the fact that $\frac{\mathcal{X}_J^{(2)\prime}\mathcal{X}_J^{(1)}}{|J|}$ converges as $T \to \infty$. Following the same logic presented in the proof Theorem 2, it can be shown that the first component in left-hand side of (S2.20) is greater than $\lambda\|\boldsymbol{\Theta}\|_1 + \lambda^{\text{thr}}$ with probability approaching to 1 which completes the proof.

S3 Additional Simulation Results

We now report additional simulation results. In Section S3.1, we compare the performance of our method to the ones proposed in Safikhani and Shojaie (2020) and Bai et al. (2020) available from R package `VARDetect`. Regarding the tuning parameter selection, we follow the recommendation of their papers by using the default ones. In Section S3.2, we examine how our method works under the classical change point scenarios, especially when the changes in the VAR coefficient matrix has a monotonically increasing form. Lastly, in Section S3.3, we assume Σ_ε is estimated and present the simulation results for those examined in Section 4 of the main paper.

S3.1 Stronger signal-to-noise ratio

	T	p	$[\eta_1, \eta_2]$	$\eta_2 - \eta_1$	$\text{nzr}(A^{(1)}_{\text{sps}})$	$\text{nzr}(A^{(2)}_{\text{sps}})$	$\|\boldsymbol{\Theta}\|_0$	Σ_ε
M9	500	20	[166, 333]	167	−0.6	0.75	19	$0.01\boldsymbol{I}_p$
M10	300	20	[100, 200]	100	−0.8683	0.8683	19	$0.01\boldsymbol{I}_p$

Table S3.1: Simulation settings for M9 and M10 described in Section S3.1, where $\text{nzr}(A^{(1)})$ and $\text{nzr}(A^{(2)})$ are the non-zero elements in the sparse part of $A^{(1)}$ and $A^{(2)}$, respectively and $\|\boldsymbol{\Theta}\|_0$ is the number of non-zero elements of $\boldsymbol{\Theta}$.

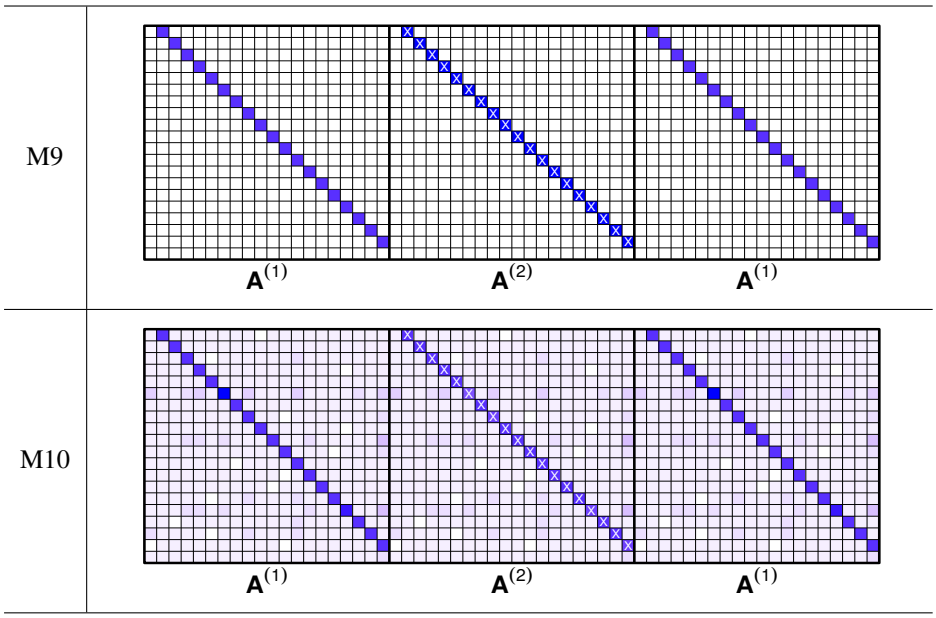

Table S3.2: The underlying coefficient matrices, $(A^{(1)}, A^{(2)}, A^{(1)})$, for the simulation settings M9 and M10 described in Section S3.1. X marks indicate which elements undergo change.

In this section, we borrow the simulation settings used in Safikhani and Shojaie (2020) and (Bai et al., 2020). M9 is the case both the underlying coefficient matrix $A^{(1)}$ and the one corresponding to an anomaly $A^{(2)}$ are sparse, while M10 assumes the low rank + sparse structure for $A^{(1)}$ and the change at anomaly is sparse. The model parameters of two settings can be found in Table S3.1 and the

			Empirical power (# ($[\hat{\eta}_1, \hat{\eta}_2] \subseteq [\eta_1, \eta_2]$))		mean (sd) of d_H	
			$A^{(1)}$ known	$\hat{A}^{(1)}$	$A^{(1)}$ known	$\hat{A}^{(1)}$
M9 (Σ_ε known)	R (s=499)	OLS	100 (22)	100 (26)	1.33 (0.80)	1.43 (0.85)
		LSS	100 (22)	100 (28)	1.31 (0.79)	1.42 (0.83)
	D (s=499)	OLS	100 (36)	100 (54)	0.80 (0.00)	0.80 (0.00)
		LSS	100 (52)	100 (67)	0.80 (0.00)	0.80 (0.00)
M9 (Σ_ε unknown)	R (s=469)	OLS	100 (33)	100 (37)	1.48 (0.84)	1.59 (0.99)
		LSS	100 (31)	100 (35)	1.49 (0.84)	1.61 (1.00)
	D (s=469)	OLS	100 (96)	100 (98)	0.40 (0.00)	0.40 (0.00)
		LSS	100 (96)	100 (98)	0.40 (0.00)	0.40 (0.00)
M10 (Σ_ε known)	R (s=296)	OLS	100 (28)	100 (21)	2.03 (1.03)	2.05 (1.08)
		LSS	100 (34)	100 (24)	2.04 (1.04)	2.13 (1.22)
	D (s=296)	OLS	100 (88)	100 (87)	0.71 (0.11)	0.71 (0.11)
		LSS	100 (96)	100 (96)	0.68 (0.07)	0.68 (0.07)
M10 (Σ_ε unknown)	R (s=271)	OLS	100 (59)	100 (51)	1.94 (1.18)	1.93 (1.20)
		LSS	100 (61)	100 (53)	2.01 (1.21)	1.92 (1.22)
	D (s=271)	OLS	100 (100)	100 (100)	0.33 (0.00)	0.33 (0.00)
		LSS	100 (100)	100 (100)	0.33 (0.00)	0.33 (0.00)

Table S3.3: Empirical power (%), the percentage of estimated anomalies those are within the true anomaly and the mean (standard deviation) of Hausdorff distance from 100 simulation runs for two methods in M9 and M10 described in Section S3.1, where s is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

		# (estimated change points)			mean (sd) of d_H
		0	1	**2**	
M9	SS	0	1	**99**	0.25 (2.42)
	LrS	0	1	**99**	0.26 (2.42)
M10	SS	1	1	**98**	0.74 (4.75)
	LrS	1	0	**99**	0.40 (3.43)

Table S3.4: Distribution of the number of estimated change-points and the mean (standard deviation) of Hausdorff distance from 100 simulation runs obtained by SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020) under M9 and M10 described in Section S3.1.

true coefficient matrices are shown in Table S3.2. To make the single anomaly setting, based on the simulation setting of scenario 1 used in Safikhani and Shojaie (2020), M9 is obtained by modifying the size of non-zero coefficients to $(-0.6, 0.75, -0.6)$ for those intervals divided by an anomaly, whereas Safikhani

and Shojaie (2020) consider the two change points with the corresponding size of non-zero coefficients $(-0.6, 0.75, -0.8)$. For M10, we borrow the simulation scenario A.2 of Bai et al. (2020) and the range of non-zero elements in low-rank part is $(-0.217, 0.212)$, where $A^{(1)}$ has a low rank + sparse structure. From Table S3.1, we can see that the anomalies presented in M9 and M10 are easier to be detected than those presented in Section 4.2 of the main paper in the sense that the size of change is larger, the width of anomaly is longer and the noise variance is smaller.

Tables S3.3-S3.4 show that all methods detect one anomaly most of the time from 100 runs (this is shown as "one" anomaly for the OLS and the Lasso methods and "two" change-points for SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020)). In terms of the localisation, both SS and LrS work better than the OLS and the Lasso method for M9 but not always for M10. Comparing SS and LrS, as expected, SS gives a better result in M9 (i.e. when $A^{(1)}$ is sparse) and LrS outperforms for M10 when $A^{(1)}$ has a low rank + sparse structure.

S3.2 Change point scenarios

In this section, we explore how our method works under two classical change settings whose true coefficient matrices are shown in Table S3.5. Both scenarios, M11 and M12, include three changepoints with the corresponding coefficient

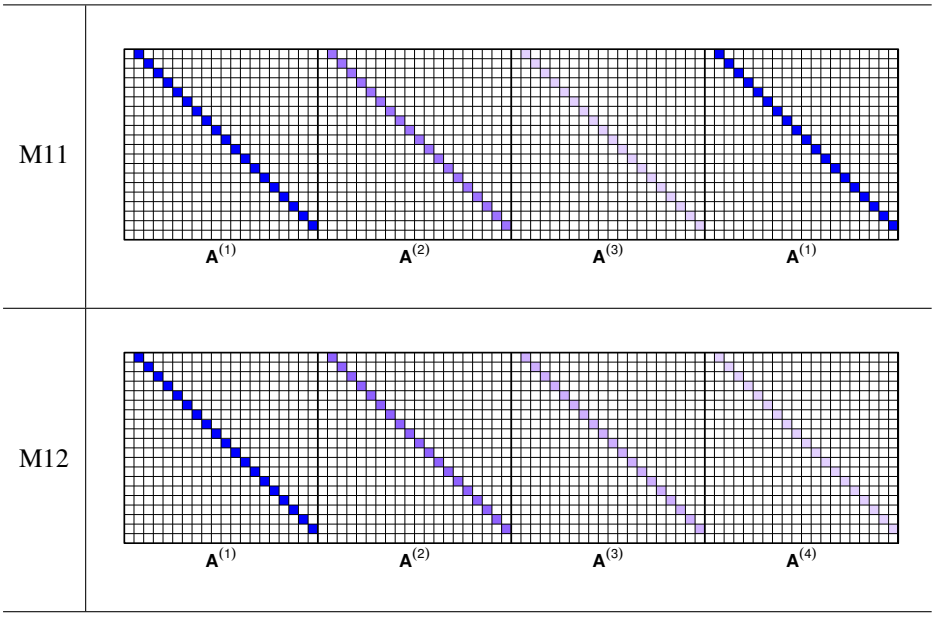

Table S3.5: The underlying coefficient matrices with three change points for the simulation settings M11 and M12 described in Section S3.2.

	T	p	$(\eta_1^*, \eta_2^*, \eta_3^*)$	$(\Delta_1^*, \Delta_2^*, \Delta_3^*)$	$\|\|A^{(j+1)} - A^{(j)}\|\|_{0, j=1,\ldots,3}$
M11	300	20	(125, 150, 175)	(0.25, 0.25, -0.5)	19
M12	100	20	(25, 50, 75)	(0.2, 0.2, 0.1)	19

Table S3.6: The simulation setting for M11 and M12 described in Section S3.2, where $(\eta_1^*, \eta_2^*, \eta_3^*)$ are change points, $(\Delta_1^*, \Delta_2^*, \Delta_3^*) = (A^{(2)} - A^{(1)}, A^{(3)} - A^{(2)}, A^{(4)} - A^{(3)})$ and $\|B\|_0$ is the number of non-zero elements of B.

matrices $(A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)})$. The setting M12 has monotonically increasing coefficient matrices, $(A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)})$, with three change points. Unlike M12, M11 has partially-monotonic coefficient matrices in the sense that the first three matrices, $(A^{(1)}, A^{(2)}, A^{(3)})$, have a monotonically increasing form but $A^{(4)}$ is set as the same with $A^{(1)}$ (i.e. $A^{(1)} = A^{(4)}$) which makes M11 still keeps the epidemic change framework.

Under the single anomaly framework, we let the interval behaving differ-

			Empirical power (# ($[\hat{\eta}_1, \hat{\eta}_2] \subseteq [\eta_1, \eta_2]$))		mean (sd) of Hausdorff distance	
			$A^{(1)}$ known	$\hat{A}^{(1)}$	$A^{(1)}$ known	$\hat{A}^{(1)}$
M11 (Σ_ε known)	R (s=296)	OLS	97 (28)	4 (0)	4.41 (7.17)	40.45 (7.64)
		LSS	**100** (36)	**34** (11)	2.35 (2.05)	28.28 (19.22)
	D (s=296)	OLS	98 (65)	6 (3)	1.90 (5.97)	39.53 (9.81)
		LSS	**100** (80)	**54** (51)	0.58 (0.36)	19.74 (20.66)
M11 (Σ_ε unknown)	R (s=271)	OLS	1 (0)	2 (0)	41.63 (3.70)	41.73 (2.47)
		LSS	**100** (14)	**18** (10)	2.60 (2.05)	34.69 (15.68)
	D (s=271)	OLS	16 (11)	19 (10)	36.34 (13.64)	35.67 (13.83)
		LSS	**100** (68)	**34** (24)	0.62 (0.36)	27.93 (19.70)
M12 (Σ_ε known)	R (s=161)	OLS	99 (74)	3 (3)	11.05 (8.08)	23.84 (2.23)
		LSS	**100** (82)	**52** (48)	12.12 (8.82)	19.72 (7.88)
	D (s=161)	OLS	99 (80)	5 (4)	11.57 (9.75)	23.88 (1.85)
		LSS	**100** (92)	**60** (57)	12.03 (9.34)	21.60 (7.57)
M12 (Σ_ε unknown)	R (s=63)	OLS	10 (10)	9 (7)	24.16 (1.20)	24.23 (2.01)
		LSS	**100** (78)	**42** (36)	11.83 (8.07)	19.31 (7.58)
	D (s=63)	OLS	8 (7)	7 (4)	24.49 (2.12)	24.25 (2.00)
		LSS	**100** (91)	**49** (46)	10.32 (8.60)	20.92 (6.92)

Table S3.7: Empirical power (%), the percentage of estimated anomalies those are within the true anomaly and the mean (standard deviation) of Hausdorff distance from 100 simulation runs for two methods in the settings M11 and M12 described in Section S3.2, where s is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

		# (estimated change points)				mean (sd) of d_H
		0	1	2	**3**	
M11	SS	**66**	32	1	1	37.87 (16.81)
	LrS	**63**	34	2	1	36.83 (17.14)
M12	SS	**93**	7	0	0	47.33 (6.12)
	LrS	**94**	6	0	0	47.56 (5.73)

Table S3.8: Distribution of the number of estimated change-points and the mean (standard deviation) of Hausdorff distance from 100 simulation runs obtained by SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020) under M11 and M12 described in Section S3.2.

ently from $A^{(1)}$ an anomaly, where the corresponding interval is $[125, 175]$ for M11 and $[25, 100]$ for M12. Those tweaked true anomaly is used for computing the Hausdorff distance summarised in Table S3.7 and similar interpretations are obtained with those in Section 4.3 of the main paper. As M11 and M12 are the

classical changepoint settings, we also report the estimation results of Safikhani and Shojaie (2020) and Bai et al. (2020) in Table S3.8 where both methods underestimate changepoint.

S3.3 Simulation results when Σ_ε is estimated

			Empirical power ($\#([\hat{\eta}_1, \hat{\eta}_2] \subseteq [\eta_1, \eta_2])$)		mean (sd) of d_H	
			$A^{(1)}$ known	$\hat{A}^{(1)}$	$A^{(1)}$ known	$\hat{A}^{(1)}$
M1	R (s=1969)	OLS	13 (5)	18 (8)	40.28 (14.38)	40.49 (13.35)
		LSS	**100** (10)	**93** (8)	3.56 (4.18)	10.07 (14.07)
	D (s=1969)	OLS	26 (14)	42 (28)	35.88 (17.39)	30.85 (19.77)
		LSS	**100** (39)	**93** (28)	1.84 (4.32)	8.16 (14.86)
	D (s=981)	OLS	25 (14)	38 (28)	36.61 (16.97)	32.36 (19.32)
		LSS	**100** (37)	**93** (28)	2.00 (4.33)	9.42 (16.05)
M2	R (s=1969)	OLS	6 (2)	7 (3)	44.35 (10.56)	45.31 (7.97)
		LSS	**100** (3)	**62** (2)	3.65 (4.46)	26.48 (20.43)
	D (s=1969)	OLS	13 (8)	19 (8)	41.67 (14.19)	40.80 (15.07)
		LSS	**100** (30)	**63** (12)	2.50 (5.15)	26.46 (21.78)
	D (s=981)	OLS	13 (8)	16 (7)	42.02 (13.80)	41.64 (14.24)
		LSS	**100** (34)	**62** (11)	2.28 (4.70)	26.72 (21.66)
M5	R (s=469)	OLS	4 (2)	11 (5)	43.38 (7.40)	41.16 (11.66)
		LSS	**100** (16)	**93** (27)	1.79 (1.24)	4.53 (11.13)
	D (s=469)	OLS	21 (17)	31 (15)	36.37 (16.95)	35.28 (17.06)
		LSS	**100** (67)	**100** (79)	0.32 (0.15)	0.32 (0.15)
M6	R (s=469)	OLS	3 (0)	10 (0)	45.11 (6.61)	42.74 (11.62)
		LSS	**100** (1)	**18** (1)	3.08 (2.30)	38.09 (17.40)
	D (s=469)	OLS	13 (0)	25 (0)	40.83 (14.53)	39.16 (15.06)
		LSS	**100** (0)	**65** (0)	0.36 (0.16)	16.37 (22.00)

Table S3.9: Empirical power (%), the percentage of estimated anomalies those are within the true anomaly and the mean (standard deviation) of Hausdorff distance from 100 simulation runs for the settings described in Section 4.2 of the main paper, where s is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

In this section, we repeat the simulations for the case when Σ_ε is unknown and the maximum likelihood estimator $\hat{\Sigma}_\varepsilon$ is used. The repeated simulation set-

tings include the ones used in Section 4 of the main paper and we obtain similar interpretations. Considering that a single collective anomaly can be modelled as two classical changepoints, we compare the performance of our method with Safikhani and Shojaie (2020) and Bai et al. (2020) and their methods tend to underestimate. We emphasise that the anomalies presented in this section are harder to detect as the size of change in coefficient matrix is smaller and the noise has a larger variance compared to the ones presented in Section S3.1.

		# (estimated change points)				mean (sd) of d_H
		0	1	**2**	3	
M1	SS	**64**	32	4	0	32.90 (17.55)
	LrS	**64**	31	5	0	33.13 (17.27)
M2	SS	**69**	24	7	0	36.19 (16.80)
	LrS	**68**	23	8	1	35.87 (16.75)
M5	SS	**66**	34	0	0	33.05 (16.50)
	LrS	**66**	34	0	0	33.05 (16.49)
M6	SS	**75**	24	1	0	36.91 (16.25)
	LrS	**74**	25	1	0	36.40 (16.69)

Table S3.10: Distribution of the number of estimated change-points and the mean (standard deviation) of Hausdorff distance from 100 simulation runs obtained by SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020) under the simulation settings described in Section 4.2 of the main paper.

		# (estimated change points)					mean (sd) of d_H
		0	1	2	3	**4**	
M3	SS	**66**	15	15	2	2	28.58 (9.78)
	LrS	**67**	16	13	3	1	29.16 (9.20)
M4	SS	**99**	0	1	0	0	13.69 (0.86)
	LrS	**100**	0	0	0	0	13.60 (0.00)

Table S3.11: Distribution of the number of estimated change-points and the mean (standard deviation) of Hausdorff distance from 100 simulation runs obtained by SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020) under the simulation settings described in Section 4.2 of the main paper.

			#(detected anomalies)								mean (sd) of d_H	
			$A^{(1)}$ known				$\hat{A}^{(1)}$				$A^{(1)}$ known	$\hat{A}^{(1)}$
			0	1	**2**	3	0	1	**2**	3		
M3	R	OLS	**52**	48	0	0	**51**	49	0	0	31.24 (6.49)	31.34 (4.91)
	(s=1969)	LSS	0	**70**	30	0	0	**95**	5	0	8.12 (5.21)	18.35 (9.74)
	D	OLS	19	**59**	22	0	24	**60**	14	2	25.67 (13.16)	27.87 (11.25)
	(s=1969)	LSS	0	**65**	33	2	0	**92**	8	0	5.95 (3.33)	16.10 (11.93)
	D	OLS	25	**59**	16	0	33	**60**	7	0	27.83 (11.54)	29.69 (9.15)
	(s=981)	LSS	0	**66**	34	0	0	**90**	10	0	5.72 (3.16)	15.91 (12.02)
M4	R	OLS	**97**	2	1	0	**91**	9	0	0	13.49 (1.10)	13.53 (0.34)
	(s=1969)	LSS	0	5	**93**	2	15	**80**	5	0	3.83 (5.23)	11.71 (2.49)
	D	OLS	**61**	32	7	0	**57**	35	8	0	13.06 (2.97)	13.08 (3.91)
	(s=1969)	LSS	0	5	**93**	2	8	**81**	11	0	2.87 (4.62)	11.50 (3.22)
	D	OLS	**73**	24	3	0	**66**	29	5	0	13.27 (1.76)	13.16 (3.38)
	(s=981)	LSS	0	6	**93**	1	13	**76**	11	0	2.66 (3.39)	11.47 (3.31)

Table S3.12: Distribution of the number of detected anomalies and the mean (standard deviation) of Hausdorff distance from 100 simulation runs for two methods under the simulation settings described in Section 4.2 of the main paper, where s is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

S4 Additional Results for Yellow cab demand data

This section gives the details of the estimated anomalies and changepoints from different methods for the yellow taxi trip data presented in Section 5.1 of the main paper.

	Estimated anomaly or change-points
LSS, OLS	[2019-12-31 22:30, 2020-01-01 04:00]
SS	2019-09-03 01:30, 2019-10-27 14:30, 2019-11-19 22:30, 2019-12-03 09:30, 2019-12-15 04:30, 2020-01-03 04:30, 2020-01-16 12:00, 2020-02-10 11:30
LrS	2019-08-09 11:00, 2019-09-12 07:30, 2019-09-24 04:30, 2019-10-10 19:30, 2019-10-29 05:30, 2019-11-17 22:30, 2019-12-04 18:00, 2019-12-15 19:30, 2020-01-01 00:30, 2020-01-16 12:00, 2020-02-09 17:30

Table S4.13: The estimated anomaly of LSS (Lasso) and OLS and the estimated change-points of SS (Safikhani and Shojaie, 2020) and LrS (Bai et al., 2020) from the yellow cab demand data described in Section 5.1 of the main paper.

Bibliography

Bai, P., A. Safikhani, and G. Michailidis (2020). Multiple change points detection in low rank and sparse high dimensional vector autoregressive models. *IEEE Trans. Signal Process. 68*, 3074–3089.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist 43*, 1535–1567.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist 42*, 2243–2281.

Kovács, S., H. Li, P. Bühlmann, and A. Munk (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv:2002.06633*.

Safikhani, A. and A. Shojaie (2020). Joint structural break detection and parameter estimation in high-dimensional nonstationary var models. *J. Amer. Statist. Assoc.*, 1–14.

Wang, D., Y. Yu, A. Rinaldo, and R. Willett (2019). Localizing changes in high-dimensional vector autoregressive processes. *arXiv:1909.06359*.