# Statistica Sinica Preprint No: SS- 2021-0181

| | |
|---|---|
| **Title** | Collective Anomaly Detection in High-Dimensional Var Models |
| **Manuscript ID** | SS-2021-0181 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202021.0181 |
| **Complete List of Authors** | Hyeyoung Maeng,<br>Idris A. Eckley and<br>Paul Fearnhead |
| **Corresponding Authors** | Hyeyoung Maeng |
| **E-mails** | h.maeng4@lancaster.ac.uk |

Notice: Accepted version subject to English editing.

Statistica Sinica

# Collective anomaly detection in High-dimensional VAR Models

Hyeyoung Maeng, Idris A. Eckley and Paul Fearnhead

*Lancaster University, United Kingdom*

*Abstract:* There is increasing interest in detecting collective anomalies: potentially short periods of time where the features of data change, before reverting back to normal behaviour. We propose a new method for detecting a collective anomaly in VAR models. Our focus is on situations where the change in the VAR coefficient matrix at an anomaly is sparse, i.e. a small number of entries of the VAR coefficient matrix change. To tackle this problem, we propose a test statistic for a local segment that is built on the lasso estimator of the change in model parameters. This enables us to detect a sparse change more efficiently, and our lasso-based approach becomes especially advantageous when the anomalous interval is short. We show that the new procedure controls Type 1 error and has asymptotic power tending to one. The practicality of our approach is demonstrated through simulations and two data examples, involving New York taxi trip data and EEG data.

*Key words and phrases:* Collective anomaly; High-dimensional time series; Lasso; Sparse changes; Epidemic change; Vector autoregressive model.

## 1. Introduction

There is a growing need for modelling and analysis of high-dimensional time series, as such series have become increasingly common in many application areas. Applications include estimating causal relationships among genes and constructing gene regulatory networks (Shojaie and Michailidis, 2010), discovering causal interactions in Neuroimaging (Seth et al., 2015), detecting changes in the network structure of functional magnetic resonance imaging data (Cribben and Yu, 2017) and analysing the network structure of volatility interconnections in the S&P 100 data (Barigozzi and Hallin, 2017).

The majority of existing methods are built on the assumption of stationary and stable time series, however if there is either a structural change or a period of anomalous behaviour in a time series, detecting its location is an important task. High-dimensional changepoint analysis has recently received increasing attention and is still in its early stage. The types of changes that are of interest differ depending on application, and we only mention a selection. Detecting a change in mean is the most popularly studied area and early works include Bai (2010) which studies the consistency of the least squares estimator of a single change-point. The CUSUM procedure has been popularly used in changepoint analysis, with Zhang et al. (2010) and Horváth and Hušková (2012) presenting a test statistic for detecting a change in multivariate data that is based on $l_2$-

aggregation of the CUSUM values for individual series. Jirak (2015) proposes an $l_\infty$-aggregation of CUSUM statistics and Enikeeva and Harchaoui (2013) propose a combination of two chi-square type test statistics so as to be able to detect changes that affect many or only a few series. Other recent works dealing with cross-sectionally sparse changes include Cho and Fryzlewicz (2015), Cho (2016) and Wang and Samworth (2018). Related topics in high-dimensional time series include detecting changes in covariance (Aue et al., 2009; Wang et al., 2017) and in factor models (Chen et al., 2014; Barigozzi et al., 2018).

One of the most popular models for high-dimensional time series is the vector autoregressive (VAR) model (Sims, 1980; Lütkepohl, 2005), due to its ability to capture complex temporal and cross-sectional relationships. However, the estimation of the coefficient matrix becomes challenging as the number of parameters increases quadratically with the number of time series. To overcome this, structured sparsity of the VAR coefficients is often assumed as this assumption dramatically reduces the number of model parameters. For example, Song and Bickel (2011) use lasso type, that is $\ell_1$, penalties to encourage sparsity in the estimates of the VAR coefficients. Basu and Michailidis (2015) investigate the theoretical properties of $\ell_1$-penalised estimators for a Gaussian VAR model and show consistency results, while Lin and Michailidis (2017) generalise the results by considering a general norm instead of being restricted to the $\ell_1$-norm for the penalty. Recently, more complex structures have been studied in the literature: Basu et al. (2019) study the low-rank and structured sparse VAR model and Nicholson et al. (2020) impose a hierarchical structure on VAR coefficient matrices according to the lag order, thus addressing both the dimensionality and the lag selection issues at the same time.

Despite the large body of literature on VAR models, detecting a structural change has rarely been studied. Kirch et al. (2015) consider two scenarios, detecting at-most-one-change and epidemic change in model parameters of multivariate time series which is not restricted to VAR models. Safikhani and Shojaie (2020) consider the multiple change-point setting for the VAR coefficient matrix under a high-dimensional regime and propose a three-stage procedure that returns consistent estimators of both change-points and parameters. Wang et al. (2019) also study the same setting (i.e. when the model parameters have a form of piecewise constant over time) and use a dynamic programming approach for localising change-points and improving the corresponding error rates. Bai et al. (2020) study the multiple change-point setting but assume the low-rank plus sparse structure on the VAR coefficient matrices and consider the case where only the sparse structure changes over time, while the low-rank parts remain constant. We will explain how our proposal is different from these existing works later in this section.

In contrast to these earlier works, we focus on settings where we have plenty of information about the current or normal behaviour of our time-series, and wish to detect periods of different or anomalous behaviour. First, this can arise when detecting collective anomalies or epidemic changepoints – where we have a, potentially short, period of time where the behaviour of our model changes before it reverts back to pre-change behaviour. Note that both collective anomaly and epidemic change can be modelled as two classical changepoints, and for ease of presentation, we use the terminology collective anomaly from now on.

Collective anomaly detection is a problem of significant interest in many applications such as genetics (Siegmund et al., 2011; Jeng et al., 2012; Bardwell and Fearnhead, 2017) and brain science (Aston and Kirch, 2012; Kirch et al., 2015). A selection of existing works include cost function based approaches for univariate (Yao, 1993; Fisch et al., 2018), independent multivariate (Fisch et al., 2021) and cross-correlated multivariate (Tveten et al., 2020) data. Anomaly detection has also been widely studied in the machine learning literature, see Chandola et al. (2009) for an extensive review. Secondly, the settings we focus on, where we have a lot of information about the normal behaviour of the time series, also arises with sequential change detection (Lai, 1995), when we observe data in real-time and wish to detect any change away from the current behaviour as quickly as possible. Although our primary focus is on a posteriori collective anomaly detection, we show how our method can be extended to the online framework in Section 5. The key feature of our detection problem is that we have substantially more information about the current or normal behaviour than about the anomaly. This suggests that we should potentially use different procedures to estimate the parameters of the VAR model for the normal behaviour than for the anomaly. We do this through making an assumption that it is the change in VAR parameters that is sparse.

We focus on improving the detection power when the difference between the coefficient matrices at anomaly point is sparse (i.e. a small number of entries of the VAR coefficient matrix change). To tackle this problem, we propose a test statistic for a local segment which is built on the lasso estimator of the change in model parameters. This enables us to detect a sparse change more efficiently, as the sparsity of change is considered in establishing the test statistic. Moreover, our lasso-based approach become more advantageous over, say, the standard likelihood-ratio test statistic for shorter anomalous intervals: as for shorter intervals we have fewer observations to estimate the new VAR coefficient matrix. Conversely, our approach becomes more like a high-dimensional problem where the number of observations is similar to or less than the number of parameters to estimate.

In Section 4, our approach is compared with a method that is built on estimating the change in VAR matrix using ordinary least squares estimator, and the results show that our method outperforms it when detecting sparse change. As we consider the setting where a relatively longer region has a normal behaviour than the anomalous behaviour, it is reasonable to assume that the underlying VAR coefficient matrix is estimated well enough. Thus, we first develop our method when the normal behaviour is assumed to be known and extend it to the case where an appropriate estimator for the VAR coefficient is used instead. Our theory in Section 3 shows the validity of this approach providing that the estimator for the VAR coefficient is close enough to the true one. Although our main focus is on single anomaly detection, we show that the new method can be extended for detecting multiple anomalies in Section 2.1.

Among those relevant works already introduced earlier in this section, Safikhani and Shojaie (2020) and Bai et al. (2020) are most closely related to our work in that they also control the change in VAR parameters with a lasso penalty, however their approaches are different from our method in several aspects. To obtain the initial estimate of change-points before screening, Safikhani and Shojaie (2020) use a fused lasso penalty on a full model considering all time points being a candidate for change-point. Thus their objective function controls the sparsity of VAR parameters and the sparsity of its difference at the same time. Bai et al. (2020) follows a similar procedure to Safikhani and Shojaie (2020) under the multiple change-point framework. They use a block fused lasso penalty by assuming that the model parameters in a block is fixed, while our objective function controls only the sparsity of change in building a test statistic and search many segments to find an anomalous interval. Also, Safikhani and Shojaie (2020) and Bai et al. (2020) assume that the $l_2$-norm of a change in VAR parameter is bounded away from zero, whereas our assumption on the $l_2$-norm of a change is related to the sparsity of change which is in line with the assumptions used in Wang et al. (2019). Although those change-point detection methods are not exactly designed for the anomaly setting we consider in this paper, we compare our performance with theirs and present results in the supplementary material. Our method works better especially when the underlying VAR coefficient matrix is dense but the change is sparse, and surprisingly even in the case where the VAR coefficient matrix has a low rank plus sparse structure and only a sparse component changes. Full details can be found in the supplementary material.

The remainder of the article is organised as follows. Section 2 gives a full description of our procedure and the relevant theoretical results are presented in Section 3. The supporting simulation studies are described in Section 4. Our methodology is illustrated through two datasets in Section 5 and we end with additional discussion in Section 6. The proofs of our main theoretical results are in the supplementary material.

## 2. Methodology

### 2.1 Problem setting

We consider a zero-mean, stationary, $p$-dimensional multivariate time series $\boldsymbol{x}_t = (x_{1t}, \dots, x_{pt})'$ generated by a VAR(1) model:

$$\boldsymbol{x}_t = \boldsymbol{A}_t \boldsymbol{x}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \Sigma_\varepsilon), \quad t = 1, \dots, T, \tag{2.1}$$

where $\{\boldsymbol{A}_t\}_{t=1}^T$ is a $p \times p$ matrix and $\Sigma_\varepsilon$ is a positive definite matrix. We assume that the high-dimensional VAR model shows an anomalous behaviour at $t \in [\eta_1, \eta_2]$ such that $0 < \eta_1 < \eta_2 < T$. Then the sequence of $\{\boldsymbol{A}_t\}_{t=1}^T$ forms piecewise-constant coefficient matrices as follows

$$\boldsymbol{A}^{(1)} = \boldsymbol{A}_1 = \cdots = \boldsymbol{A}_{\eta_1 - 1}, \quad \boldsymbol{A}^{(2)} = \boldsymbol{A}_{\eta_1} = \cdots = \boldsymbol{A}_{\eta_2}, \quad \boldsymbol{A}^{(1)} = \boldsymbol{A}_{\eta_2+1} = \cdots = \boldsymbol{A}_T,$$

where $\boldsymbol{A}^{(1)} \neq \boldsymbol{A}^{(2)}$ and $\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)} \in \mathbb{R}^{p \times p}$. The model in equation (2.1) can be represented as the following linear regression,

$$\begin{pmatrix} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \vdots \\ \boldsymbol{x}'_T \end{pmatrix}_{T \times p} = \begin{pmatrix} \boldsymbol{x}'_0 & 0 \\ \vdots & \vdots \\ \boldsymbol{x}'_{\eta_1 - 2} & 0 \\ \boldsymbol{x}'_{\eta_1 - 1} & \boldsymbol{x}'_{\eta_1 - 1} \\ \vdots & \vdots \\ \boldsymbol{x}'_{\eta_2 - 1} & \boldsymbol{x}'_{\eta_2 - 1} \\ \boldsymbol{x}'_{\eta_2} & 0 \\ \vdots & \vdots \\ \boldsymbol{x}'_{T-1} & 0 \end{pmatrix}_{T \times 2p} \begin{pmatrix} \boldsymbol{\theta}^{(1)'} \\ \boldsymbol{\theta}^{(2)'} \end{pmatrix}_{2p \times p} + \begin{pmatrix} \varepsilon'_1 \\ \varepsilon'_2 \\ \vdots \\ \varepsilon'_T \end{pmatrix}_{T \times p}, \tag{2.2}$$

where $\boldsymbol{\theta}^{(1)} = \boldsymbol{A}^{(1)}$, $\boldsymbol{\theta}^{(2)} = \boldsymbol{A}^{(2)} - \boldsymbol{A}^{(1)}$. The model, as written in equation (2.2), is a linear regression of the form $\mathcal{Y} = \mathcal{X}\Theta + E$. As such, it can be represented as $\boldsymbol{Y}_{Tp\times 1} = \boldsymbol{X}_{Tp\times 2p^2}\boldsymbol{\Theta}_{2p^2\times 1} + \boldsymbol{E}_{Tp\times 1}$, where $\boldsymbol{X} = I_p \otimes \mathcal{X}$ and $\otimes$ is the tensor product of two matrices.

Now our interest is in estimating the collective anomaly $[\eta_1, \eta_2]$. Our motivation is for scenarios where there is substantial information about the normal or pre-change behaviour of the data. Thus, for ease of presentation, we will first assume that $\boldsymbol{\theta}^{(1)}$ in (2.2) is known. In practice we will use an estimate of $\boldsymbol{\theta}^{(1)}$, and our theory shows that our approach has good asymptotic properties if we plug-in a suitably accurate estimate of $\boldsymbol{\theta}^{(1)}$ in the following procedure. We assume that the change $\boldsymbol{\theta}^{(2)}$ is sparse in that it has small number of nonzero entries which will be formulated in a later section. Assuming the base coefficient matrix $\boldsymbol{A}^{(1)}$ is known, we can rewrite the model as

$$\begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_T' \end{pmatrix}_{T\times p} - \begin{pmatrix} \boldsymbol{x}_0'\boldsymbol{\theta}^{(1)\prime} \\ \vdots \\ \boldsymbol{x}_{\eta_1-2}'\boldsymbol{\theta}^{(1)\prime} \\ \boldsymbol{x}_{\eta_1-1}'\boldsymbol{\theta}^{(1)\prime} \\ \vdots \\ \boldsymbol{x}_{\eta_2-1}'\boldsymbol{\theta}^{(1)\prime} \\ \boldsymbol{x}_{\eta_2}'\boldsymbol{\theta}^{(1)\prime} \\ \vdots \\ \boldsymbol{x}_{T-1}'\boldsymbol{\theta}^{(1)\prime} \end{pmatrix}_{T\times p} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \boldsymbol{x}_{\eta_1-1}' \\ \vdots \\ \boldsymbol{x}_{\eta_2-1}' \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{T\times p} \left(\boldsymbol{\theta}^{(2)\prime}\right)_{p\times p} + \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_T' \end{pmatrix}_{T\times p}, \qquad (2.3)$$

that can be represented as $\mathcal{Y} - \mathcal{X}^{(1)}\boldsymbol{\theta}^{(1)\prime} = \mathcal{X}^{(2)}\boldsymbol{\theta}^{(2)\prime} + E$. With slight abuse of notation by using different definitions of $\boldsymbol{Y}$, $\boldsymbol{X}$ and $\boldsymbol{\Theta}$, we can rewrite (2.3) as

$$\boldsymbol{Y}_{Tp\times 1} = \boldsymbol{X}_{Tp\times p^2}\boldsymbol{\Theta}_{p^2\times 1} + \boldsymbol{E}_{Tp\times 1}, \tag{2.4}$$

where $\boldsymbol{X} = I_p \otimes \mathcal{X}^{(2)}$.

## 2.2 Lasso-based approach

To detect a collective anomaly we derive a test for whether data in an interval of time is anomalous, and then apply this test to data from a set of suitably chosen intervals, $\mathbb{J}_{T,p}(L)$. To help with the presentation of theory in Section 3, we parameterise this set by the length, $L$, of the smallest interval it contains. For any interval $J \in \mathbb{J}_{T,p}(L)$, by extracting the corresponding rows from each matrix in (2.3), the linear regression form can be rewritten as: $\mathcal{Y}_J - \mathcal{X}_J^{(1)}\boldsymbol{\theta}^{(1)\prime} = \mathcal{X}_J^{(2)}\boldsymbol{\theta}^{(2)\prime} + E_J$, that can be vectorised in a form of $\boldsymbol{Y}_J = \boldsymbol{X}_J\boldsymbol{\Theta} + \boldsymbol{E}_J$ as in (2.4).

One of the standard ways to detect change or epidemic changes in regression models is to use a likelihood ratio test (Kim and Siegmund, 1989; Siegmund and Venkatraman, 1995; Yau and Zhao, 2016; Baranowski et al., 2019; Dette and Gösmann, 2020), and these methods can be applied in the VAR setting. To detect a collective anomaly in a set of intervals, our procedure involves calculating the likelihood ratio statistic for each interval $J \in \mathbb{J}_{T,p}(L)$ as

$$-2\Big\{\sum_{s\in J} l_s(\boldsymbol{\Theta} = 0, \Sigma_\varepsilon) - \sum_{s\in J} l_s(\hat{\boldsymbol{\Theta}}, \Sigma_\varepsilon)\Big\}, \tag{2.5}$$

where $\hat{\mathbf{\Theta}}$ is the maximum likelihood estimator and the likelihood function has the form of

$$\sum_{s\in J} l_s(\mathbf{\Theta}, \Sigma_\varepsilon) = -\frac{1}{2}\Big\{|J|p\log(2\pi) + |J|\log|\Sigma_\varepsilon| + (\boldsymbol{Y}_J - \boldsymbol{X}_J\mathbf{\Theta})^\top(\Sigma_\varepsilon^{-1}\otimes I)(\boldsymbol{Y}_J - \boldsymbol{X}_J\mathbf{\Theta})\Big\}.$$

As we consider only $\mathbf{\Theta}$ varying, the first two terms are constant and will cancel in the test statistic. It is common to assume $\Sigma_\varepsilon$ is the identity matrix, in which case the maximum likelihood estimator of $\mathbf{\Theta}$ is the ordinary least squares (OLS) estimator. Alternatively we can estimate the variance from the residuals obtained when estimating the parameters of the VAR model on training data. For ease of presentation, we will assume $\Sigma_\varepsilon$ is an identity matrix from now on, but our theoretical results are still valid if this assumption is not correct. Furthermore, the theory can be extended to situations where we assume either $\Sigma_\varepsilon$ is any positive identity matrix or an estimate of $\Sigma_\varepsilon$ is used. We now give details of the likelihood ratio statistic and our suggested improvement based on penalised estimation of the change in the VAR coefficients.

**The OLS method** Before introducing the lasso-based approach, we consider the test statistic based on the least squares estimator which we refer to as the OLS method. The OLS estimator has been popularly used in the changepoint detection literature e.g. in a linear model setup, CUSUM-type approaches built on the least squares estimator are studied by Horváth et al. (2004), Aue et al.

(2006) and Fremdt (2015). For any interval $J \in \mathbb{J}_{T,p}(L)$, the test statistic of the OLS method takes the form,

$$T(J) = \|\boldsymbol{Y}_J\|_2^2 - \min_{\boldsymbol{\Theta}} \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2\} \tag{2.6}$$

that is the same as the likelihood ratio statistic in (2.5) when $\Sigma_\varepsilon$ is the identity matrix. $T(J)$ has a $\chi^2_{p^2}$ distribution under the null, $\boldsymbol{\Theta} = \boldsymbol{0}$. The classical least squares estimator $\hat{\boldsymbol{\Theta}} = \operatorname{argmin}_{\boldsymbol{\Theta}} \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2\}$ in (2.6) is not able to be used when the dimension $p$ is greater than $T$. Note that $\hat{\boldsymbol{\Theta}}$ also depends on $J$ but this is suppressed in the notation for simplicity.

**The Lasso method** To handle the case when $\boldsymbol{\Theta}$ is sparse more effectively, we propose a test statistic based on a lasso estimator:

$$T^{\text{lasso}}(J) = \|\boldsymbol{Y}_J\|_2^2 - \min_{\boldsymbol{\Theta}} \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_1\}. \tag{2.7}$$

To detect a collective anomaly, we calculate this test statistic for a collection of intervals, $\mathbb{J}_{T,p}(L)$. We detect an anomaly if the maximum value of these test statistics is above a pre-determined threshold. If we detect an anomaly, we estimate its location as the interval in $\mathbb{J}_{T,p}(L)$ with the largest test-statistic value. The detailed procedure is given in Algorithm 1.

**Algorithm 1:** Single anomaly detection

```
INPUT: $X$ matrix in (2.4), $L$, $\lambda^{\text{thr}}$
    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 $\lambda^{\text{thr}}$, 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 exists) and save the
  estimator of the anomaly interval,
        $\hat{I} = \underset{J \in \mathbb{J}_{T,p}(L)}{\text{argmax}}\, T^{\text{lasso}}(J).$    (2.8)
  OUTPUT: $\hat{I}$.
```

For setting the collection of intervals $\mathbb{J}_{T,p}(L)$ in Step 1, there exist two general methods; randomly generated intervals (Fryzlewicz, 2014; Baranowski et al., 2019) and deterministic construction of intervals (Kovács et al., 2020). In this paper, we use both methods and compare their performance in Section 4.

## 2.3 Extensions to VAR(q) model and multiple anomaly detection

Our method can be extended to deal with VAR(q) model and multiple anomaly detection. The details can be found in Section S1 of the supplementary material.

# 3. Theoretical results

In this section, we explore the asymptotic behaviour of the proposed method. We show that our method controls the familywise error under the null (i.e. when there exist no anomaly) with an appropriate threshold and give conditions under which the asymptotic power of the method tends to 1. These results are based upon the following assumptions.

**Assumption 1.** *For each* $j = 1, 2$, *let* $\Gamma_j(\ell)$ *be the population version of the lag-*$\ell$ *covariance matrix of* $\boldsymbol{\mathfrak{x}}_j$ *where* $\boldsymbol{\mathfrak{x}}_j$ *is* $\boldsymbol{\mathfrak{x}}_1 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\eta_1 - 1}\}$, *and* $\boldsymbol{\mathfrak{x}}_2 = \{\boldsymbol{x}_{\eta_1}, \ldots, \boldsymbol{x}_{\eta_2}\}$. *For* $\kappa \in [-\pi, \pi]$, *there exist the spectral density matrices,*

$$f_j(\kappa) = \frac{1}{2\pi} \sum_{l \in \mathbb{Z}} \Gamma_j(l) \exp^{-\sqrt{-1}\kappa l}.$$

*In addition,* $\max_j \mathcal{M}(f_j) = \max_j \left\{ ess \sup_{\kappa \in [-\pi,\pi]} \Lambda_{\max}(f_j(\kappa)) \right\} < +\infty$ *and* $\min_j \mathfrak{m}(f_j) = \min_j \left\{ ess \inf_{\kappa \in [-\pi,\pi]} \Lambda_{\min}(f_j(\kappa)) \right\} > 0$, *where* $\Lambda_{\max}(A)$ *and* $\Lambda_{\min}$ *are the largest and the smallest eigenvalues of the symmetric matrix A, respectively.*

This first condition is needed to control the stability properties of the VAR models. This is a spectral density condition that is not only valid for VAR model but also holds for a large class of general linear process. Basu and Michailidis (2015) use the same assumption but for a stable VAR setting without considering anomalies, while we extend it to the single collective anomaly setting by assuming a spectral density function for each common and anomalous segments separately.

In order to bound the power of our method we need conditions on the size and length of any anomaly and the set of intervals we use – essentially we will need at least one interval of sufficient length to be contained within the anomaly. To this end we introduce the following:

**Assumption 2.** *The dimensionality $p$ satisfies $p \sim T^{\alpha}$ for some fixed $\alpha \in [0, \infty)$.*

**Assumption 3.** *There exist at least one interval $J \in \mathbb{J}_{T,p}(L)$ such that $J \subseteq [\eta_1, \eta_2]$ and the choice of $L$ for a set of intervals $\mathbb{J}_{T,p}(L)$ satisfies $\frac{log(T \vee p)}{L} \to 0$ as $T \to \infty$, where any interval $J \in \mathbb{J}_{T,p}(L)$ has length at least $L$.*

**Assumption 4.** *The sparsity of change is fixed; $\|\mathbf{\Theta}\|_0 = d_0$.*

**Assumption 5.** *For any $\xi > 0$, $L \cdot \|\mathbf{\Theta}\|_2^2 > C_2 \cdot d_0^2 \cdot \log^{1+\xi}(T \vee p)$, where $C_2 > 0$ is a constant.*

Assumption 4 gives the condition on the number of nonzero entries of the coefficient matrix, where the sparsity parameter $d_0$ affects the signal-to-noise ratio condition in Assumption 5. Our Assumption 5 is similar to the conditions required in other change-point problem in high-dimensional VAR models. For example, Wang et al. (2019) study a multiple changepoint setting and their signal-to-noise ratio assumption becomes equal to ours in the case when single change-point is considered, while Safikhani and Shojaie (2020) assume $\|\mathbf{\Theta}\|_2$ is bounded away from zero.

**Assumption 6.** *For the estimator $\hat{\boldsymbol{\theta}}^{(1)}$, $\left\|\boldsymbol{\theta}^{(1)} - \hat{\boldsymbol{\theta}}^{(1)}\right\|_\infty < C\sqrt{\frac{\log(T \vee p)}{L}}$ with probability approaching $1$ as $T \to \infty$ and $p \to \infty$, where $C > 0$ is a constant.*

Assumption 6 states the necessary condition on the estimation error bound on $\hat{\boldsymbol{\theta}}^{(1)}$ and is only used to extend our theoretical results to the case when $\boldsymbol{\theta}^{(1)}$ is estimated. Such error bounds are presented in Proposition 4.1 of Basu and Michailidis (2015) and Lemma 15 of Wang et al. (2019): when $\boldsymbol{\theta}^{(1)}$ is assumed to be sparse with the condition $||\boldsymbol{\theta}^{(1)}||_0 = k$, then its lasso estimator, $\hat{\boldsymbol{\theta}}^{(1)}$, satisfies $\left\|\boldsymbol{\theta}^{(1)} - \hat{\boldsymbol{\theta}}^{(1)}\right\|_2 \leq c\sqrt{k}\sqrt{\frac{\log(T \vee p)}{T}}$ with probability tending to 1, where $\hat{\boldsymbol{\theta}}^{(1)}$ is obtained from a sample of size $T$. This estimation error bound in $\ell_2$-norm implies our Assumption 6 presented in $\ell_\infty$-norm when the sparsity $k$ is fixed.

We now present our main theoretical results where the proofs can be found in Section S2 of the supplementary material. The following theorem gives conditions on the lasso penalty to ensure the procedure asymptotically controls the familywise error when there is no anomaly.

**Theorem 1.** *Let Assumptions 1-3 hold. If there exist no anomaly, for a tuning parameter* $\lambda = C_3\sqrt{L(2\log p + \log T)}$ *with a constant* $C_3$ *large enough, we have*

$$P\Big(\max_{J \in \mathbb{J}_{T,p}(L)} T^{lasso}(J) \leq \lambda^{thr}\Big) \geq P\Big(\max_{J \in \mathbb{J}_{T,p}(L)} T^{lasso}(J) = 0\Big)$$

$$\geq 1 - C_4 \exp(-C_5(2\log p + \log T)),$$

*where* $C_4, C_5 > 0$, $\lambda^{thr} > 0$ *and* $\lambda$ *is a tuning parameter in* (2.7).

In Theorem 1, it is clear that our result applies to any positive threshold $\lambda^{\mathrm{thr}}$. In the proof of Theorem 1 in the supplementary material, we show that the familywise error is controlled under an appropriate tuning parameter $\lambda$, and the argument still holds if we use $\lambda_J = C_3\sqrt{|J|(2\log p + \log T)}$ instead of $\lambda$, where $\lambda_J$ varies with each interval $J$. We now turn to the asymptotics of the test statistic under the alternative.

**Theorem 2.** *Let Assumptions 1-5 hold. If there exist an anomaly, with a tuning parameter $\lambda = C_2\sqrt{L(2\log p + \log T)}$ for a large enough $C_2$, as $T \to \infty$, then*

$$P\Big(\max_{J\in\mathbb{J}_{T,p}(L)} T^{lasso}(J) \leq \lambda^{thr}\Big) \to 0 \quad \textit{and} \quad P(\hat{I}\cap[\eta_1,\eta_2] \neq \emptyset) \to 1,$$

*where $\lambda^{thr}$ has the order of $\sqrt{L\cdot\log(p\vee T)}$, the estimated anomaly $\hat{I}$ is as in* (2.8) *and $\lambda$ is a tuning parameter in lasso regression in* (2.7).

Theorem 2 states that the test statistic corresponding to the intervals in the candidate set is greater than the pre-specified threshold if the interval is located within the true anomaly. In other words, it shows that the individual test has asymptotic power one. The argument in the proof of Theorem 2 still applies if we make $\lambda$ vary with interval $J$ by replacing $L$ by $|J|$ in the definition of $\lambda$. The following theorem shows that our method has larger power to detect a sparse collective anomaly.

**Theorem 3.** *Assume that $\boldsymbol{x}_t$ follows* (2.3) *and let Assumptions 1-5 hold. Let the null hypothesis hold, then for any $\{J : J \in \mathbb{J}_{T,p}(L), J\cap[\eta_1,\eta_2] = \emptyset\}$, the test statistic of the OLS method in* (2.6) *follows a $\chi^2_{p^2}$ distribution. Consequently,* *we have an asymptotic level $\alpha$ 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. Under the alternative, for any $J \in \mathbb{J}_{T,p}(L)$ such that $J \subseteq [\eta_1, \eta_2]$, the upper bound on the power of the OLS method is given as*

$$\frac{E(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2)}{W_p}, \tag{3.9}$$

*where $W_p = O_p(p)$.*

Note that $W_p$ in (3.9) is linked to the false positive rate as it is the approximation of $\chi^2_{p^2;(1-\alpha)} - p^2$. See the proof in Section S2 of the supplementary material for further details.

Theorem 3 shows the asymptotic behaviours of the test statistic of the OLS method under both the null and the alternative hypotheses. Furthermore, Theorem 3 implies that the test statistic built on the lasso estimator can detect weaker anomalies than using the OLS estimator when the change is sparse. The intuition behind this is that the test statistic of the OLS method in (2.6) can be written as

$$\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2\}, \tag{3.10}$$

and $E(\|\boldsymbol{Y}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_2^2)$ needs to be at least as large as $O_p(p)$ to have high power. By comparison, if we denote the lasso estimator of $\boldsymbol{\Theta}$ by $\hat{\boldsymbol{\Theta}}$, then the test statistic of the lasso method in (2.7) can be written as

$$\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \lambda\|\boldsymbol{\Theta}\|_1 + \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_1 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2 - \lambda\|\hat{\boldsymbol{\Theta}}\|_1\}. \tag{3.11}$$

Noting that the term in {}s in (3.11) is positive, the lasso-based test statistic requires that $\|\boldsymbol{Y}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_2^2$ should at least as large as $O_p(\lambda\|\boldsymbol{\Theta}\|_1)$ and $\lambda = C_2\sqrt{L(2\log p + \log T)}$. The following corollary states that the assertions in Theorems 1-2 remain true if the $\boldsymbol{\theta}^{(1)}$ is replaced by an estimator $\hat{\boldsymbol{\theta}}^{(1)}$ that satisfies the condition in Assumption 6.

**Corollary 1.** *Theorems 1-2 hold with a different constant if $\hat{\boldsymbol{\theta}}^{(1)}$ is used in calculating the test statistic instead of the true parameter $\boldsymbol{\theta}^{(1)}$, where $\hat{\boldsymbol{\theta}}^{(1)\prime}$ is an estimator fulfilling Assumption 6.*

## 4. Simulation study

### 4.1 Parameter choice and setting

We compare the performance of our lasso-based approach with the OLS method described in Section 2.2. Whilst there are other methods for detecting changes in a VAR model, such as those of Safikhani and Shojaie (2020) and Bai et al. (2020), they are not designed for the collective anomaly setting that we consider. For completeness, we compare their performances with ours, and the details can be found in Section S3 of the supplementary material. Perhaps due to not being designed for the collective anomaly setting, we find these alternative methods perform substantially worse than ours, particularly when the underlying matrix $A^{(1)}$ is dense but the change is sparse.

In practice, the underlying parameter $A^{(1)}$ is often unknown and needs to be estimated. In this case, as the accuracy of our method depends on how accurately we can estimate $A^{(1)}$, considering two extreme cases gives upper and lower bounds on our method: $A^{(1)}$ is known and $A^{(1)}$ is estimated from a relatively small amount of data with ridge or lasso penalty depending on the given sparsity of $A^{(1)}$. In the latter case, the training data is the same size as the test data which we examine for detecting an anomaly.

The threshold of each test is selected by choosing the 99% quantile of the test statistics obtained through the 100 simulation runs performed under the null. This can be easily done when $A^{(1)}$ is known. A naïve approach when $A^{(1)}$ is unknown is to simulate data from the model with the estimator $\hat{A}^{(1)}$ obtained from the training set. However this ignores the estimation error in $A$ and consequently leads to thresholds that are too low. To overcome this we use a two stage simulation procedure. We simulate a data set with the estimator $\hat{A}^{(1)}$ obtained from the training set and re-estimate $A$ from this data set. This estimate is denoted by $\tilde{A}^{(1)}$. Then we use data simulated from $\tilde{A}^{(1)}$ as the data simulated under the null that is used to obtain the threshold.

For the error variance, we set $\Sigma_\varepsilon$ to be the identity matrix. In the following sections, we report the results when $\Sigma_\varepsilon$ is known. The results for when $\Sigma_\varepsilon$ is estimated can be found in Section S3 of the supplementary material.

As presented in Theorems 1-2, the performance of the lasso-based method depends on the tuning parameter selection. Our theoretical results hold under both $\lambda = C\sqrt{L(2\log p + \log T)}$ and $\lambda_J = C\sqrt{|J|(2\log p + \log T)}$, where $\lambda$ is a fixed tuning parameter for all intervals of different lengths and $\lambda_J$ varies with the length of each interval $J$. Based on our empirical experience, in practice we use $\lambda_J$ with the default constant $C = 0.15$, as it achieves stable performance across the different settings as presented in the following section. In practice, similar performance is obtained for any $C \in [0.05, 0.25]$. Using a fixed constant $C$ is advantageous over optimising $\lambda_J$ for each interval (e.g. by minimising cv), as it makes the algorithm faster especially when both $T$ and $p$ are large and it leads to the stable performance especially when $|J|$ is substantially small.

We also look at how the choice of the set of intervals, $\mathbb{J}_{T,p}(L)$, affects performance. We vary both the number of intervals which we denote by $s$, and the way we choose the intervals, randomly or deterministically, with a pre-determined minimum length of interval. For the deterministic construction of intervals, we use the technique proposed in Definition 1 of Kovács et al. (2020) with the decay parameter $1/a = 1.1, 1.2$. Regardless of the way of choosing the intervals, we force the minimum length intervals to be greater than $p$ in order to compare our approach with the OLS method. To deal with the high-dimensional settings (such as M7 and M8 in Table 1), we set the minimum length intervals to be greater than $\lceil p/10 \rceil$ and report only the results of the lasso-based method.

## 4.2 Simulation settings

We simulate data from 8 settings presented in Table 1 and the true coefficient matrices of some settings are shown in Table 2. Those settings are categorised into two scenarios: (1) $A^{(1)}$ is dense (M1-M4) and (2) $A^{(1)}$ is sparse (M5-M8); where the number of non-zero elements is large in (1) and small in (2).

| | T | p | $[\eta_1, \eta_2]$ | $[\eta_3, \eta_4]$ | $\eta_2 - \eta_1$ | $\eta_4 - \eta_3$ | $\Delta_1$ | $\Delta_2$ | $\|\|\mathbf{\Theta}\|\|_0$ |
|---|---|---|---|---|---|---|---|---|---|
| M1 | 500 | 10 | $[227, 272]$ | | 45 | | 0.35 | | 10 |
| M2 | 500 | 10 | $[233, 266]$ | | 33 | | 0.35 | | 10 |
| M3 | 500 | 10 | $[133, 166]$ | $[333, 366]$ | 33 | 33 | 0.6 | 0.6 | 5 |
| M4 | 500 | 10 | $[33, 66]$ | $[433, 466]$ | 33 | 33 | 0.5 | 0.5 | 5 |
| M5 | 500 | 20 | $[222, 277]$ | | 55 | | 0.55 | | 19 |
| M6 | 500 | 20 | $[229, 270]$ | | 41 | | 0.55 | | 19 |
| M7 | 100 | 50 | $[44, 55]$ | | 11 | | 1.1 | | 49 |
| M8 | 100 | 70 | $[40, 60]$ | | 20 | | 1.1 | | 69 |

Table 1: Simulation settings, where $\Delta_1 = |A^{(2)} - A^{(1)}|$, $\Delta_2 = |A^{(3)} - A^{(1)}|$ and $||\mathbf{\Theta}||_0$ is the number of non-zero elements of $\mathbf{\Theta}$.

In the settings M1-M4, we consider the case when all entries of $A^{(1)}$ are non-zero. The coefficient matrix is randomly generated by using the algorithm proposed by Ansley and Kohn (1986) and implemented in R package `gmvarkit`

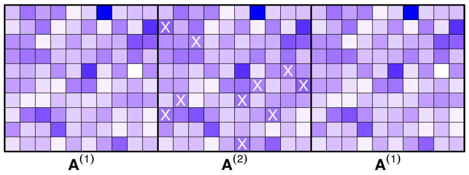
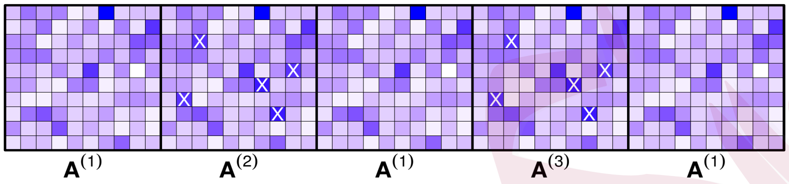
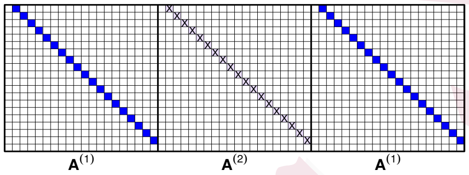
| | |
|---|---|
| M1, M2 |  |
| M3, M4 |  |
| M5, M6 |  |

Table 2: The underlying coefficient matrices for some of the simulation settings described in Section 4.2, where $A^{(2)}$ and $A^{(3)}$ correspond to anomalies and X marks indicate which elements undergo change.

which forces the resulting VAR model to be stationary, where the range of the entries of $A^{(1)}$ is obtained as $[-0.67, 0.58]$. Under the settings M1 and M2, the single anomaly, $[\eta_1, \eta_2]$, is considered with the corresponding coefficient matrix $A^{(2)}$, while two collective anomalies, $[\eta_1, \eta_2]$ and $[\eta_3, \eta_4]$, are assumed for both M3 and M4, where the corresponding coefficient matrix is $A^{(2)}$ and $A^{(3)}$, respectively. To detect multiple anomalies, we use Algorithm 3 presented in Section S1 of the supplementary material. As given in Assumption 4, only a few (ten for M1-M2 and five for M3-M4) entries in the VAR coefficient matrix undergo change in anomalous interval where the details are presented in Table 1.

Under the settings M5-M8, we consider the case when $A^{(1)}$ is sparse i.e. only a smaller number of entries are non-zero. Similar to the settings used in Safikhani and Shojaie (2020), the 1-off diagonal values of the coefficient matrix are non-zero as shown in Table 2. M7 and M8 are the high dimensional settings in the sense that the width of anomaly $(\eta_2 - \eta_1)$ is less than the dimension (p). When $A^{(1)}$ is unknown, it is estimated from the training data with a ridge penalty for M1-M4 and with a lasso penalty for M5-M8. In the following section, we present the simulation results for all settings.

### 4.3 Results

Tables 3 and 4 show the summary of simulation results for the single and multiple anomaly cases, respectively. As shown in Table 3, the lasso-based method tends to detect an anomaly more often than the OLS-based approach in all settings regardless of the sparsity of $A^{(1)}$, the way of choosing intervals to investigate and whether $A^{(1)}$ is known or estimated. The lasso-based method also outperforms in terms of distance between the estimated and the true anomaly and its variance. As expected, compared to the results when the true $A^{(1)}$ is known, both OLS and lasso methods perform less well when $\hat{A}^{(1)}$ is used. The number of estimated anomalies located within the true anomaly tends to be proportional to empirical power, and to be larger when segments are chosen deterministically

| | | | Empirical power<br>(# ($[\hat{\eta}_1, \hat{\eta}_2] \subseteq [\eta_1, \eta_2]$) ) | | mean (sd)<br>of $d_H$ | |
|---|---|---|---|---|---|---|
| | | | $A^{(1)}$ known | $\hat{A}^{(1)}$ | $A^{(1)}$ known | $\hat{A}^{(1)}$ |
| M1 | R | OLS | 100 (19) | 93 (15) | 1.59 (1.31) | 5.43 (12.08) |
| | (s=1029) | LSS | 100 (19) | **99** (17) | 1.47 (0.93) | 1.95 (4.62) |
| | D | OLS | 100 (43) | 94 (33) | 0.39 (0.25) | 3.55 (11.64) |
| | (s=1029) | LSS | 100 (46) | **99** (40) | 0.35 (0.16) | 0.81 (4.55) |
| | D | OLS | 100 (25) | 94 (19) | 0.39 (0.33) | 3.52 (11.65) |
| | (s=540) | LSS | 100 (26) | **99** (27) | 0.32 (0.22) | 0.78 (4.55) |
| M2 | R | OLS | 98 (12) | 69 (7) | 2.85 (6.52) | 17.67 (21.43) |
| | (s=1029) | LSS | 98 (18) | **89** (10) | 2.63 (6.46) | 7.31 (14.79) |
| | D | OLS | 98 (44) | 74 (31) | 1.32 (6.57) | 14.30 (21.31) |
| | (s=1029) | LSS | **99** (50) | **90** (50) | 0.82 (4.67) | 5.63 (14.68) |
| | D | OLS | 98 (69) | 72 (52) | 1.27 (6.58) | 15.14 (21.78) |
| | (s=540) | LSS | **99** (76) | **87** (71) | 0.76 (4.68) | 7.24 (16.77) |
| M5 | R | OLS | 100 (20) | 68 (12) | 1.67 (1.21) | 15.51 (20.22) |
| | (s=499) | LSS | 100 (29) | **99** (33) | 1.54 (0.99) | 2.08 (4.41) |
| | D | OLS | 100 (46) | 87 (48) | 0.43 (0.24) | 6.21 (15.00) |
| | (s=499) | LSS | 100 (63) | **100** (75) | 0.34 (0.12) | 0.37 (0.13) |
| M6 | R | OLS | 99 (14) | 16 (5) | 2.59 (4.66) | 39.06 (16.45) |
| | (s=499) | LSS | **100** (21) | **68** (32) | 1.90 (1.48) | 15.66 (21.07) |
| | D | OLS | 100 (34) | 34 (21) | 0.45 (0.48) | 30.63 (21.80) |
| | (s=499) | LSS | 100 (65) | **93** (76) | 0.29 (0.40) | 3.55 (11.76) |
| M7 | R (s=367) | LSS | 100 (14) | 91 (23) | 2.53 (1.22) | 6.35 (12.58) |
| | D (s=367) | LSS | 100 (13) | 88 (33) | 1.36 (1.32) | 7.01 (14.55) |
| M8 | R (s=270) | LSS | 100 (25) | 100 (28) | 2.67 (1.43) | 2.64 (1.25) |
| | D (s=270) | LSS | 100 (61) | 100 (84) | 1.61 (0.49) | 1.84 (0.37) |

Table 3: Empirical power (%), the number of estimated anomalies located within the true anomaly and the mean (standard deviation) of $d_H$ (Hausdorff distance) from 100 simulation runs for two methods under M1, M2 and M5-M8, where $s$ is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

rather than randomly. Although it is not shown in the table, the mean of Hausdorff distance computed from the estimated anomalies located within the true anomaly tends to be smaller than the one computed from the estimated anoma-

| | | | #(detected anomalies) | | | | | | mean (sd) of $d_H$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $A^{(1)}$ known | | | $\hat{A}^{(1)}$ | | | $A^{(1)}$ known | $\hat{A}^{(1)}$ |
| | | | 1 | **2** | 3 | 0 | 1 | **2** | | |
| M3 | R<br>(s=1944) | OLS | 27 | **73** | 0 | 1 | **56** | 43 | 3.28 (3.10) | 8.29 (9.01) |
| | | LSS | 21 | **78** | 1 | 0 | 39 | **61** | 2.87 (3.03) | 5.02 (5.86) |
| | D<br>(s=1944) | OLS | 24 | **76** | 0 | 0 | **53** | 47 | 2.46 (2.62) | 6.73 (9.07) |
| | | LSS | 12 | **86** | 2 | 0 | 32 | **68** | 1.97 (2.68) | 3.44 (5.16) |
| | D<br>(s=1029) | OLS | 26 | **74** | 0 | 1 | **54** | 45 | 2.59 (2.57) | 7.13 (9.27) |
| | | LSS | 21 | **77** | 2 | 0 | 35 | **65** | 2.43 (2.66) | 3.50 (4.36) |
| M4 | R<br>(s=1944) | OLS | 6 | **93** | 1 | 11 | **64** | 25 | 2.97 (4.53) | 9.91 (4.69) |
| | | LSS | 1 | **96** | 3 | 0 | 29 | **71** | 2.57 (5.43) | 4.76 (5.22) |
| | D<br>(s=1944) | OLS | 4 | **95** | 1 | 6 | **59** | 35 | 1.92 (4.19) | 8.59 (5.72) |
| | | LSS | 1 | **96** | 3 | 0 | 21 | **79** | 1.50 (3.80) | 3.53 (5.08) |
| | D<br>(s=1029) | OLS | 4 | **95** | 1 | 8 | **59** | 33 | 1.98 (4.22) | 8.72 (5.61) |
| | | LSS | 1 | **98** | 1 | 0 | 22 | **78** | 1.44 (3.68) | 3.54 (5.00) |

Table 4: Distribution of the number of detected anomalies and the mean (standard deviation) of $d_H$ (Hausdorff distance) from 100 simulation runs for two methods under M3-M4, where $s$ is the number of intervals examined. Note that Random, Deterministic, Lasso are shortened to R, D, LSS, respectively.

lies those are not exactly located within the true anomaly. Comparing the randomly and the deterministically chosen segments with the same size, for both the OLS and the lasso methods, the deterministic way tends to give a similar or a slightly larger power regardless of whether $A^{(1)}$ is known or not. Note, when $A^{(1)}$ is estimated in Table 3, the deterministically chosen intervals with smaller sample size ($s = 540$) shows a larger power than those chosen randomly with a larger sample size ($s = 1029$) for both methods, and the difference becomes larger as the length of anomalous interval becomes shorter (from M1 to M2 as presented in Table 1). Table 4 shows similar interpretations. Other simulation settings including stronger signal-to-noise ratio scenarios (M9-M10) and changepoint scenarios (M11-M12) are explored in Section S3 of the supplementary material.

## 5. Data analysis

### 5.1 Yellow cab demand in New York City

To demonstrate the usefulness of our method, we now turn to real data applications. In our first example, we apply our method to the yellow taxi trip data that has previously been analysed by Safikhani and Shojaie (2020). The data can be downloaded from the New York City Taxi and Limousine Commission (TLC) Database (`https://www1.nyc.gov`). This data consists of the number of yellow taxi pick-ups recorded from 10 randomly selected zones in Manhattan, a borough in New York City. We aggregate the number of yellow taxi pick-ups every 30 minutes from March 11, 2019 to February 29, 2020 which results in 17088 time points. The raw data has an anomaly on November 3, 2019 that is caused by a daylight-saving time adjustment (Wu and Keogh, 2021) as two hours of data are placed into a single hour when the time change occurred. To solve this, we simply divided the number of observations by two for the corresponding hour and used the adjusted data. To prevent the detection procedure being affected by other effects, we remove weekly, seasonal and bank holiday effects by regressing the raw time series onto the corresponding indicator variables and using the residuals. We also remove the first order nonstationarity from the data by having the differenced version of the time series. The first 6835 data points is used to estimate the underlying VAR coefficient $A^{(1)}$ by applying a lasso penalty. As the true $A^{(1)}$ is unknown in practice, to determine the threshold, we basically use the same technique proposed in Section 4.1 and choose the 99% quantile of the test statistics from 100 deterministically chosen intervals. The remaining 10252 data points are used to detect a single anomaly, where the length of the smallest interval is set to $L = \lceil p/4 \rceil = 3$. Note that the same minimum length $\lceil p/4 \rceil$ is also applied for analysing the EEG data under the online change detection framework in Section 5.2.

The top plot in Figure 1 shows that a few spikes are observed between December 30, 2019 and January 2, 2020, where the interval within green vertical lines is enlarged in the bottom plot. From the middle plot, we see that the largest test statistic is obtained for a small interval which includes the spikes shown in the top plot. The bottom plot shows that the spikes occur around New Year's Eve and our method detects an anomaly between 10:30pm on December 31, 2019 and 4am on January 1, 2020. We emphasise that this anomaly is detected even after removing the holiday effect for 10 federal holidays from the period between March 11, 2019 to February 29, 2020 that includes January 1, 2020. From Figure 2, we see that a sudden high demand occurred at the 2nd and 7th

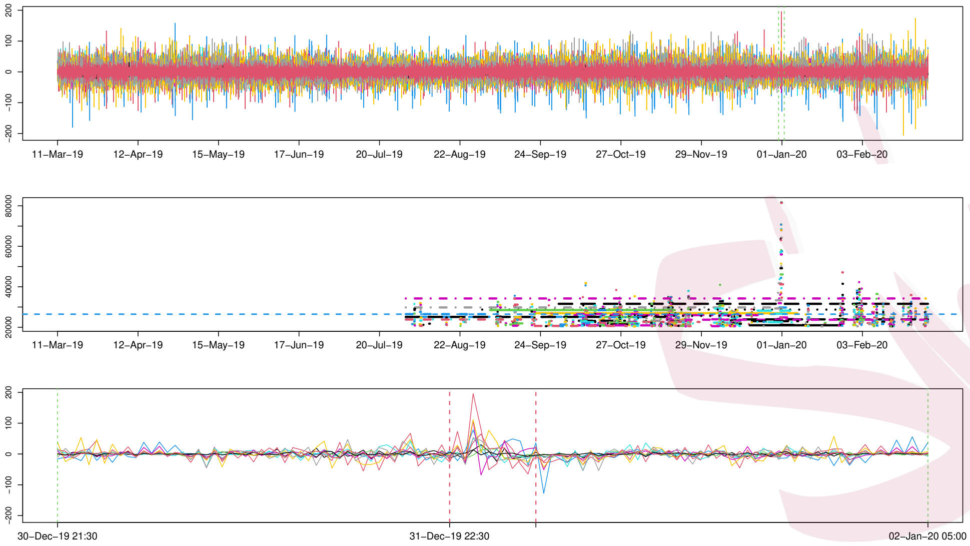


Figure 1: (Top) The differenced yellow taxi pickups recorded from March 11, 2019 to February 29, 2020 in Manhattan. (Middle) The 50 largest test statistics with the corresponding interval. The blue horizontal dashed line indicates the threshold. (Bottom) The portion of the top plot indicated with dashed green vertical lines. Red vertical lines show the estimated anomaly, $[\text{Dec } 31, 2019\ 00:00, \text{Jan } 1, 2020\ 04:00]$.

zones located near to Times Square, while there was no such change for the 3rd zone that is located far from Times Square. Therefore, we can interpret that there was a sudden high demand near Times Square where the annual New Year's Eve celebration takes place, and this changes the relationship between the 10 zones we investigate.

The OLS method gives the same estimated anomaly with the lasso-based method although the larger $L = p = 10$ is used. Comparing with other methods designed for detecting changes in a VAR model, Safikhani and Shojaie (2020)

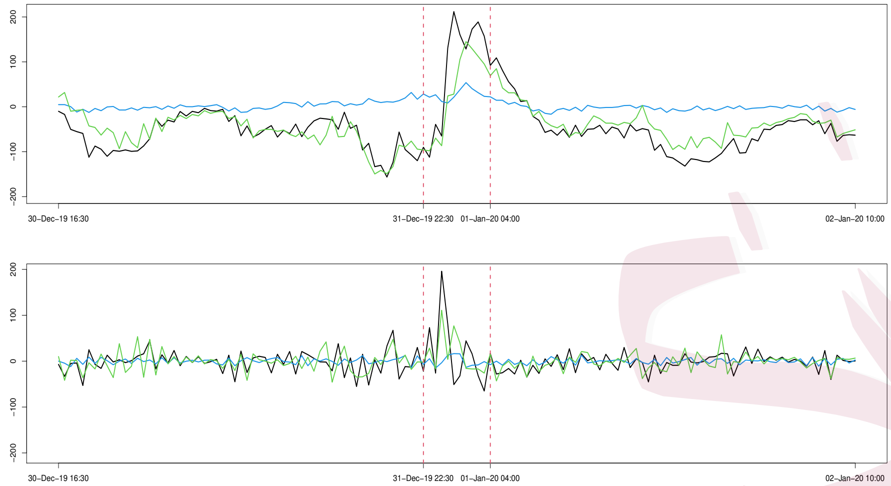


Figure 2: Taxi demand (Top) and differenced Taxi demand (Bottom) for 2nd (black), 3rd (blue) and 7th (green) zones in Manhattan recorded from December 30, 2019 to January 2, 2020. Red vertical lines show the estimated anomaly.

estimates 8 changes including 4:30am on Jan 3, 2020 while Bai et al. (2020) returns 11 changes including 00:30am on Jan 1, 2020 that coincides with the estimated anomaly by our method. All the estimated changepoints can be found in Section S4 of the the supplementary material.

## 5.2 EEG Data

We now show how our method can be used in as an online changepoint detection method. We demonstrate this on electroencephalogram (EEG) data collected from an epileptic patient. Other ways of analysing this dataset can be found in Ombao et al. (2001), Ombao et al. (2005) and Schröder and Ombao (2019). The data consists of brain electrical potentials recorded by placing electrodes on

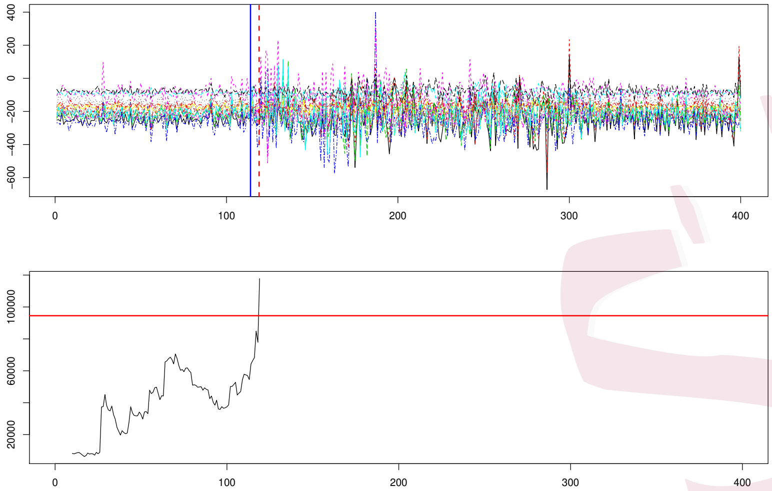


Figure 3: (Top) EEG data recorded at 18 different channels. Blue solid vertical line is the time at which the neurologist thinks seizure starts and the red dashed vertical line is the anomaly detected in the online setting. (Bottom) The maximum test statistics at each time point obtained through Algorithm 2. The horizontal red solid line presents the pre-specified threshold.

18 locations on the scalp of a patient. The EEG signals are recorded during an epileptic seizure, thus these exists a visible change in the data as shown in Figure 3. The brain wave patterns are recorded over 500 seconds with the sampling rate 100 Hz (i.e. 100 points per second). As done in Safikhani and Shojaie (2020), to speed up computation, we use 2 observations per second which reduces the number of time points to $T = 1000$.

We separate the data into a training set of the size $T_1 = 600$ and a test set of the size $T_2 = 400$. The first half of the training set is used to estimate the underlying VAR coefficient $A^{(1)}$ by applying a lasso penalty and the second half is used to have a threshold that is chosen as the 99% quantile of the test statistics computed from 327 deterministically chosen intervals. Then we perform the single anomaly detection using a test set.

**Algorithm 2:** Online anomaly detection

```
INPUT: $X$, $\lambda^{\text{thr}}$, $t_0$
$t \leftarrow t_0$
FLAG $\leftarrow$ 0
while FLAG = 0 do
    $t \leftarrow t + 1$
    $K \leftarrow \left\lfloor \frac{\log t}{\log 2} \right\rfloor$
    $j \leftarrow 1$
    while FLAG = 0 and $j \leq K$ do
        $s_j \leftarrow t - \max(2^{(j-1)}, \lceil p/4 \rceil)$
        $J \leftarrow [s_j, t]$
        FLAG $\leftarrow \mathbb{1}\{T^{\text{lasso}}(J) > \lambda^{\text{thr}}\}$
        $j \leftarrow j + 1$
    end
end
OUTPUT: $t$.
```

As mentioned in Section 1, here we show how our method can be applied to the online framework. We refer the reader to Fisch et al. (2020) and Yu et al. (2021) for the recent works on online detection algorithm for change-points or anomalies. In the online setting, we make sequential decisions about the occurrence of an anomaly whenever each new observations is obtained. Our algorithm for online anomaly detection is similar to Algorithm 2 of Yu et al. (2021). The detailed procedure is given in Algorithm 2 where we set $t_0 = 10$. As shown in Figure 3, an anomaly is estimated at $t = 119$, giving a detection delay of 5 time points compared to the time at which the neurologist states a seizure occurred. When a different lower bound of $\max(2^{(j-1)}, \xi)$ is used in Algorithm 2 with $\xi = \lceil p/2 \rceil, \lceil p/3 \rceil, \lceil p/5 \rceil$ instead of $\lceil p/4 \rceil$, it still detects an anomaly at $t = 119$. If a larger lower bound is set with $\xi = \lceil p \rceil, \lceil 1.5p \rceil$ in which case the OLS method can also be used, an anomaly is estimated at $t = 122$, giving a detection delay of 8 time points.

## 6. Discussion

Our lasso-based approach is motivated for data where we have substantially more data about the normal behaviour of the time series than for any anomaly or epidemic change. We provide a numerical evidence that our method outperforms existing competitors in detecting sparse change when $A^{(1)}$ is either dense or sparse. Our method searches a set of local segments to detect an anomalous interval, whereas the existing change detection methodologies for the VAR model perform global optimisation. The local optimisation aspect of our method permits the extension to the online setting.

## Supplementary Materials

The Supplementary material for "Collective anomaly detection in High-dimensional VAR Models" contains the technical proofs and the additional simulation results.

## Acknowledgements

We thank Hernando Ombao and Abolfazl Safikhani for providing access to the EEG data, and Yi Yu for helpful conversations. This work was support by EPSRC grant EP/N031938/1.

## References

Ansley, C. F. and R. Kohn (1986). A note on reparameterizing a vector autoregressive moving average model to enforce stationarity. *J. Stat. Comput. Simul. 24*, 99–106.

Aston, J. A. and C. Kirch (2012). Evaluating stationarity via change-point alternatives with applications to fmri data. *Ann. Appl. Stat.*, 1906–1948.

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

Aue, A., L. Horváth, M. Hušková, and P. Kokoszka (2006). Change-point monitoring in linear models. *Econom. J. 9*, 373–403.

Bai, J. (2010). Common breaks in means and variances for panel data. *J. Econometrics 157*, 78–92.

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.

Baranowski, R., Y. Chen, and P. Fryzlewicz (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 81*, 649–672.

Bardwell, L. and P. Fearnhead (2017). Bayesian detection of abnormal segments in multiple time series. *Bayesian Anal. 12*, 193–218.

Barigozzi, M., H. Cho, and P. Fryzlewicz (2018). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *J. Econometrics 206*, 187–225.

Barigozzi, M. and M. Hallin (2017). A network analysis of the volatility of high dimensional financial series. *J. R. Stat. Soc. Ser. C. Appl. Stat. 66*, 581–605.

Basu, S., X. Li, and G. Michailidis (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Trans. Signal Process. 67*, 1207–1222.

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

Chandola, V., A. Banerjee, and V. Kumar (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR) 41*, Article 15.

Chen, L., J. J. Dolado, and J. Gonzalo (2014). Detecting big structural breaks in large factor models. *J. Econometrics 180*, 30–48.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electron. J. Stat. 10*, 2000–2038.

Cho, H. and P. Fryzlewicz (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc.*

*Ser. B. Stat. Methodol. 77*, 475–507.

Cribben, I. and Y. Yu (2017). Estimating whole-brain dynamics by using spectral clustering. *J. R. Stat. Soc. Ser. C. Appl. Stat. 66*, 607–627.

Dette, H. and J. Gösmann (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *J. Amer. Statist. Assoc. 115*, 1361–1377.

Enikeeva, F. and Z. Harchaoui (2013). High-dimensional change-point detection with sparse alternatives. *arXiv preprint arXiv:1312.1900*.

Fisch, A., L. Bardwell, and I. A. Eckley (2020). Real time anomaly detection and categorisation. *arXiv preprint arXiv:2009.06670*.

Fisch, A., I. A. Eckley, and P. Fearnhead (2018). A linear time method for the detection of point and collective anomalies. *arXiv preprint arXiv:1806.01947*.

Fisch, A. T., I. A. Eckley, and P. Fearnhead (2021). Subset multivariate collective and point anomaly detection. *Journal of Computational and Graphical Statistics* (just-accepted), 1–31.

Fremdt, S. (2015). Page's sequential procedure for change-point detection in time series regression. *Statistics 49*, 128–155.

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

Horváth, L. and M. Hušková (2012). Change-point detection in panel data. *J. Time Series Anal. 33*, 631–648.

Horváth, L., M. Hušková, P. Kokoszka, and J. Steinebach (2004). Monitoring changes in linear models. *J. Statist. Plann. Inference 126*, 225–251.

Jeng, X. J., T. T. Cai, and H. Li (2012). Simultaneous discovery of rare and common segment variants. *Biometrika 100*, 157–172.

Jirak, M. (2015). Uniform change point tests in high dimension. *Ann. Statist 43*, 2451–2483.

Kim, H.-J. and D. Siegmund (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika 76*, 409–423.

Kirch, C., B. Muhsal, and H. Ombao (2015). Detection of changes in multivariate time series with application to eeg data. *J. Amer. Statist. Assoc. 110*, 1197–1216.

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*.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 57*, 613–644.

Lin, J. and G. Michailidis (2017). Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *The J. Mach. Learn. Res. 18*, 4188–4236.

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

Nicholson, W. B., I. Wilms, J. Bien, and D. S. Matteson (2020). High dimensional forecasting via interpretable vector autoregression. *J. Mach. Learn. Res. 21*, 1–52.

Ombao, H., R. Von Sachs, and W. Guo (2005). SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc. 100*, 519–531.

Ombao, H. C., J. A. Raz, R. von Sachs, and B. A. Malow (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc. 96*, 543–560.

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.

Schröder, A. L. and H. Ombao (2019). Fresped: Frequency-specific change-point detection in epileptic seizure multi-channel EEG data. *J. Amer. Statist. Assoc. 114*, 115–128.

Seth, A. K., A. B. Barrett, and L. Barnett (2015). Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci 35*, 3293–3297.

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

Siegmund, D. and E. Venkatraman (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist*, 255–271.

Siegmund, D., B. Yakir, and N. R. Zhang (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat. 5*, 645–668.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 1–48.

Song, S. and P. J. Bickel (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

Tveten, M., I. A. Eckley, and P. Fearnhead (2020). Scalable changepoint and anomaly detection in cross-correlated data with an application to condition monitoring. *arXiv preprint arXiv:2010.06937*.

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

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

Wang, T. and R. J. Samworth (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 80*, 57–83.

Wu, R. and E. Keogh (2021). Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Trans. Knowl. Data Eng.*.

Yao, Q. (1993). Tests for change-points with epidemic alternatives. *Biometrika 80*, 179–191.

Yau, C. Y. and Z. Zhao (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 895–916.

Yu, Y., O. H. M. Padilla, D. Wang, and A. Rinaldo (2021). Optimal network online change point localisation. *arXiv preprint arXiv:2101.05477*.

Zhang, N. R., D. O. Siegmund, H. Ji, and J. Z. Li (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika 97*, 631–645.

Department Of Mathematics And Statistics, Lancaster University, Lancaster LA1 4YR, United Kingdom

E-mail: h.maeng4@lancaster.ac.uk / i.eckley@lancaster.ac.uk / p.fearnhead@lancaster.ac.uk