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, Idris A. Eckley and Paul Fearnhead
Corresponding Authors	Hyeyoung Maeng
E-mails	h.maeng4@lancaster.ac.uk

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 during which the features of data change, before reverting back to normal behavior. We propose a new method for detecting a collective anomaly in the vector autoregressive (VAR) models. We focus on situations in which the change in the VAR coefficient matrix at an anomaly is sparse, that is, 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 the 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 the type-I error and has asymptotic power tending to one. The practicality of our approach is demonstrated using simulations and two data examples, involving New York taxi trip data and EEG data, respectively.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

1. Introduction

There is a growing need for the modeling and analysis of high-dimensional time series, because such series have become increasingly common in many application areas. Here, applications include estimating the 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 analyzing the network structure of volatility interconnections in S&P 100 data (Barigozzi and Hallin, 2017).

Most existing methods assume stationary and stable time series. However, if there is either a structural change or a period of anomalous behavior in a time series, detecting its location is an important task. High-dimensional changepoint analysis is receiving increasing attention, but is still in its early stage. The types of changes that are of interest vary by application, and we mention only a selection. Detecting a change in mean is the most widely studied area, with early works including that of Bai (2010), who studies the consistency of the least squares estimator of a single changepoint. The CUSUM procedure is popular in changepoint analysis, with Zhang et al. (2010) and Horváth and Hušková (2012) presenting test statistics for detecting a change in multivariate data that are based on an l_2-aggregation of the CUSUM values for the individual series. Jirak (2015)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

proposes an l_∞-aggregation of CUSUM statistics, and Enikeeva and Harchaoui (2013) propose using a combination of two chi-square-type test statistics to detect changes that affect many or only a few series. Other recent works on cross-sectionally sparse changes include Cho and Fryzlewicz (2015), Cho (2016), and Wang and Samworth (2018). Related topics for 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, estimating the coefficient matrix becomes challenging, because the number of parameters increases quadratically with the number of time series. To overcome this, structured sparsity of the VAR coefficients is often assumed, because this assumption dramatically reduces the number of model parameters. For example, Song and Bickel (2011) use lasso-type penalties, that is, ℓ_1-penalties, to encourage sparsity in the estimates of the VAR coefficients. Basu and Michailidis (2015) investigate the theoretical properties of ℓ_1-penalized estimators for a Gaussian VAR model and show consistency results, and Lin and Michailidis (2017) generalize the results by considering a general norm instead of being restricted to the ℓ_1-norm for the penalty. Recently, more complex structures

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

In contrast to the aforementioned earlier works, we focus on settings in

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

which we have plenty of information about the current or normal behavior of our time series, and wish to detect periods of different or anomalous behavior. First, this can arise when detecting collective anomalies or epidemic changepoints. Here, we have a potentially short period during which the behavior of our model changes, before it reverts back to its pre-change behavior. Note that both collective anomaly and epidemic change can be modeled as two classical changepoints, for ease of presentation, we use the terminology collective anomaly from now on.

Collective anomaly detection is a problem of significant interest in 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 is also widely studied in the machine learning literature; see Chandola et al. (2009) for an extensive review. Second, the settings we focus on have a lot of information about the normal behavior of the time series. Such settings also arise with sequential change detection (Lai, 1995), when we observe data in real time and wish to detect a change away from the current behavior as quickly as possible. Although our primary focus is on a posteriori collective anomaly

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 behavior than we do about the anomaly. This suggests that we should potentially use different procedures to estimate the parameters of the VAR model for the normal behavior than we use for the anomaly. We do this by assuming that it is the change in the VAR parameters that is sparse.

We focus on improving the detection power when the difference between the coefficient matrices at the 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 that is built on a lasso estimator of the change in the model parameters. This enables us to detect a sparse change more efficiently, because the sparsity of the change is considered when establishing the test statistic. Moreover, our lasso-based approach becomes more advantageous over, say, the standard likelihood-ratio test statistic for shorter anomalous intervals, because we have fewer observations with which to estimate the new VAR coefficient matrix. Conversely, our approach becomes more like a high-dimensional problem in which the number of observations is similar to or less than the number of parameters that need to be estimated.

In Section 4, we compare our approach with a method built on estimating

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

the change in the VAR matrix using an ordinary least squares (OLS) estimator, finding that our method outperforms the other method when detecting a sparse change. As we consider a setting in which a relatively longer region exhibits normal behavior than anomalous behavior, it is reasonable to assume that the underlying VAR coefficient matrix is sufficiently well estimated. Thus, we first develop our method when the normal behavior is assumed to be known, and then extend it to the case in which we use an appropriate estimator for the VAR coefficient 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 in Section 2.1 that the new method can be extended to detect multiple anomalies.

Among the relevant works introduced earlier in this section, those of Safikhani and Shojaie (2020) and Bai et al. (2020) are most closely related to our work in that they also control the change in the VAR parameters using a lasso penalty. However, their approaches differ from ours in several ways. To obtain the initial estimate of the changepoints before screening, Safikhani and Shojaie (2020) use a fused lasso penalty on a full model that considers all time points as changepoint candidates. Thus, their objective function controls the sparsity of the VAR parameters and the sparsity of its difference at the same time. Bai et al. (2020) follow a similar procedure to Safikhani and Shojaie (2020) under a multiple

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

changepoint framework. They use a block fused lasso penalty by assuming that the model parameters in a block are fixed, whereas our objective function controls only the sparsity of the change when building a test statistic, and we search many segments to find an anomalous interval. In addition, Safikhani and Shojaie (2020) and Bai et al. (2020) assume that the l_2-norm of a change in a VAR parameter is bounded away from zero, whereas our assumption on the l_2-norm of a change is related to the sparsity of the change, which is in line with the assumptions used in Wang et al. (2019). Although those changepoint detection methods were not designed for the anomaly setting we consider here, we compare our performance with theirs and present the results in the Supplementary Material. Our method works best, especially when the underlying VAR coefficient matrix is dense, but the change is sparse and, surprisingly, even when 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 paper is organized 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, and we demonstrate our methodology using two datasets in Section 5. Section 6 concludes the paper. The proofs of our main theoretical results are provided in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

2. Methodology

2.1 Problem setting

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

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

where $\{\boldsymbol{A}_t\}_{t=1}^T$ is a $p \times p$ matrix and Σ_ε is a positive-definite matrix. We assume that the high-dimensional VAR model shows an anomalous behavior at $t \in [\eta_1, \eta_2]$, such that $0 < \eta_1 < \eta_2 < T$. Then, the sequence $\{\boldsymbol{A}_t\}_{t=1}^T$ 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:

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 scenarios in which we have a substantial amount of information about the normal or pre-change behavior of the data. Thus, for ease of presentation, we first assume that $\boldsymbol{\theta}^{(1)}$ in (2.2) is known. In practice, we use an estimate of $\boldsymbol{\theta}^{(1)}$, and our theory shows that our approach exhibits 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 a small number of nonzero entries, which are formulated in a later section. Assuming the base coefficient matrix $\boldsymbol{A}^{(1)}$ is known, we can rewrite the model as

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

which can be represented as $\mathcal{Y} - \mathcal{X}^{(1)}\boldsymbol{\theta}^{(1)\prime} = \mathcal{X}^{(2)}\boldsymbol{\theta}^{(2)\prime} + E$. With a slight abuse of

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 the data in an interval of time are anomalous. Then, we apply this test to data from a set of suitably chosen intervals, $\mathbb{J}_{T,p}(L)$. To help with the presentation of the theory in Section 3, we parameterize 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 $\mathcal{Y}_J - \mathcal{X}_J^{(1)}\boldsymbol{\theta}^{(1)\prime} = \mathcal{X}_J^{(2)}\boldsymbol{\theta}^{(2)\prime} + E_J$, which can be vectorized as $\boldsymbol{Y}_J = \boldsymbol{X}_J\boldsymbol{\Theta} + \boldsymbol{E}_J$, as in (2.4).

One of the standard ways of detecting a 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 calculates the likelihood ratio statistic for each interval $J \in \mathbb{J}_{T,p}(L)$ as

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

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

As we consider only $\mathbf{\Theta}$ varying, the first two terms are constant and cancel out in the test statistic. It is common to assume Σ_ε is the identity matrix, in which case the maximum likelihood estimator of $\mathbf{\Theta}$ is the OLS estimator. Alternatively, we can estimate the variance from the residuals obtained when estimating the parameters of the VAR model on the training data. For ease of presentation, we assume Σ_ε 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 in which we either assume that Σ_ε is any positive identity matrix, or we use an estimate of Σ_ε. We now present the likelihood ratio statistic and our suggested improvement based on a penalized estimation of the change in the VAR coefficients.

The OLS method Before introducing the lasso-based approach, we consider a test statistic based on the least squares estimator, which we refer to as the OLS method. The OLS estimator is popular in the changepoint detection literature. For example, 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),

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

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

which is the same as the likelihood ratio statistic in (2.5) when Σ_ε is the identity matrix. Here, $T(J)$ has a $\chi^2_{p^2}$ distribution under the null, $\boldsymbol{\Theta} = \mathbf{0}$. We cannot use 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) 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 more effectively handle the case when $\boldsymbol{\Theta}$ is sparse, we propose the following 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 predetermined 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

Algorithm 1: Single 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 prespecified 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 exists) and save the estimator of the anomaly interval,

$$\hat{I} = \underset{J \in \mathbb{J}_{T,p}(L)}{\operatorname{argmax}} T^{\text{lasso}}(J). \tag{2.8}$$

OUTPUT: $\hat{I}$.

There are two general ways to set the collection of intervals $\mathbb{J}_{T,p}(L)$ in Step 1: randomly generated intervals (Fryzlewicz, 2014; Baranowski et al., 2019), and a deterministic construction of intervals (Kovács et al., 2020). 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 a 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 behavior of the proposed method. We show that our method controls the familywise error under the null (i.e., when no

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

anomaly exists) with an appropriate threshold, and give conditions under which the asymptotic power of the method tends to one. These results are based on the following assumptions.

Assumption 1. *For each* $j = 1, 2$, *let* $\Gamma_j(\ell)$ *be the population version of the lag-*ℓ *covariance matrix of* $\mathfrak{x}_j$, *where* $\mathfrak{x}_j$ *is* $\mathfrak{x}_1 = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\eta_1 - 1}\}$, *and* $\mathfrak{x}_2 = \{\boldsymbol{x}_{\eta_1}, \ldots, \boldsymbol{x}_{\eta_2}\}$. *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 smallest eigenvalues, respectively, of the symmetric matrix A.*

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 a VAR model, but also holds for a large class of general linear processes. Basu and Michailidis (2015) use the same assumption, but for a stable VAR setting, without considering anomalies. We extend it to the single collective anomaly setting by assuming a spectral density function for each common and anomalous segment separately.

In order to bound the power of our method, we need conditions on the size

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

and length of any anomaly and the set of intervals we use. Essentially, we 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 exists 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 a 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 changepoint problems 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 when a single changepoint is considered. Safikhani and Shojaie (2020) assume $\|\mathbf{\Theta}\|_2$ is bounded away from zero.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 one 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 we estimate $\boldsymbol{\theta}^{(1)}$. 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 one, where $\hat{\boldsymbol{\theta}}^{(1)}$ is obtained from a sample of size T. This estimation error bound in the ℓ_2-norm implies our Assumption 6 presented in the ℓ_∞-norm when the sparsity k is fixed.

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

Theorem 1. *Let Assumptions 1–3 hold. If no anomalies exist, 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* λ *is a tuning parameter in* (2.7).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

In Theorem 1, it is clear that our result applies to any positive threshold λ^{thr}. In the proof of Theorem 1 in the Supplementary Material, we show that the familywise error is controlled under an appropriate tuning parameter λ, and the argument still holds if we use $\lambda_J = C_3\sqrt{|J|(2\log p + \log T)}$ instead of λ, where λ_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 an anomaly exists, 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*

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

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

Theorem 2 states that the test statistic corresponding to the intervals in the candidate set is greater than the prespecified 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 λ vary with the interval J by replacing L with $|J|$ in the definition of λ. The following theorem shows that our method has large power when detecting a sparse collective anomaly.

Theorem 3. *Assume that* $\boldsymbol{x}_t$ *follows* (2.3) *and let Assumptions 1–5 hold. Let the*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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-*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 a chi-squared 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 by*

$$\frac{E(\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2)}{W_p}, \tag{3.1}$$

where $W_p = O_p(p)$.

Note that W_p in (3.1) is linked to the false positive rate, because it is an 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 states the asymptotic behaviors of the test statistic of the OLS method under the null and alternative hypotheses. Furthermore, Theorem 3 implies that, when the change is sparse, the test statistic built on the lasso estimator can detect weaker anomalies than that based on the OLS estimator can. Intuitively, the test statistic of the OLS method in (2.6) can be written as

$$\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2\}, \tag{3.2}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

statistic of the lasso method in (2.7) can be written as

$$\|\boldsymbol{Y}_J\|_2^2 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 - \lambda\|\boldsymbol{\Theta}\|_1 + \{\|\boldsymbol{Y}_J - \boldsymbol{X}_J\boldsymbol{\Theta}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_1 - \|\boldsymbol{Y}_J - \boldsymbol{X}_J\hat{\boldsymbol{\Theta}}\|_2^2 - \lambda\|\hat{\boldsymbol{\Theta}}\|_1\}. \tag{3.3}$$

Noting that the term in {}s in (3.3) is positive, the lasso-based test statistic requires that $\|\boldsymbol{Y}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_2^2$ should be 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 and 2 remain true if we replace $\boldsymbol{\theta}^{(1)}$ with an estimator $\hat{\boldsymbol{\theta}}^{(1)}$ that satisfies the condition in Assumption 6.

Corollary 1. *Theorems 1 and 2 hold with a different constant if we use $\hat{\boldsymbol{\theta}}^{(1)}$ to calculate the test statistic instead of using 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 that of 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 performance with ours; the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

details can be found in Section S3 of the Supplementary Material. Perhaps because they are not designed for the collective anomaly setting, we find that these alternative methods perform substantially worse than the proposed method does, 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, because the accuracy of our method depends on how accurately we 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 using a ridge or lasso penalty, depending on the given sparsity of $A^{(1)}$. In the latter case, the training data set is the same size as the test data set that we examine to detect an anomaly.

The threshold of each test is selected by choosing the 99% quantile of the test statistics obtained from the 100 simulation runs performed under the null. This can be done easily 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 using 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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

under the null used to obtain the threshold.

For the error variance, we set Σ_ε as the identity matrix. In the following sections, we report the results when Σ_ε is known. The results for when Σ_ε is estimated can be found in Section S3 of the Supplementary Material.

As presented in Theorems 1 and 2, the performance of the lasso-based method depends on the selection of the tuning parameter. 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 λ is a fixed tuning parameter for all intervals of different lengths and λ_J varies with the length of each interval J. Based on our empirical experience, we use λ_J with the default constant $C = 0.15$, because this 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 optimizing λ_J for each interval (e.g., by minimizing cv), because doing so makes the algorithm faster, especially when both T and p are large, and leads to stable performance, especially when $|J|$ is substantially small.

We also examine how the choice of the set of intervals, $\mathbb{J}_{T,p}(L)$, affects the 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 predetermined minimum length of interval. For the deterministic construction of the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 how we choose the intervals, we force the minimum length of the intervals to be greater than p in order to compare our approach with the OLS method. To deal with high-dimensional settings (such as M7 and M8 in Table 1), we set the minimum length of the 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 eight settings, presented in Table 1. The true coefficient matrices of some settings are shown in Table 2. The settings are categorized into two scenarios: (1) $A^{(1)}$ is dense (M1–M4), and (2) $A^{(1)}$ is sparse (M5–M8); where the number of nonzero elements is large in (1) and small in (2).

	T	p	$[\eta_1, \eta_2]$	$[\eta_3, \eta_4]$	$\eta_2 - \eta_1$	$\eta_4 - \eta_3$	Δ_1	Δ_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 nonzero elements of $\mathbf{\Theta}$.

In the settings M1–M4, we consider the case in which all entries of $A^{(1)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

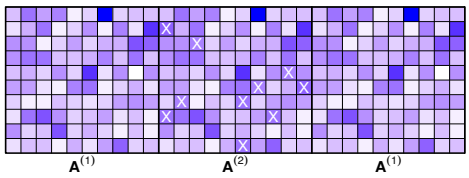
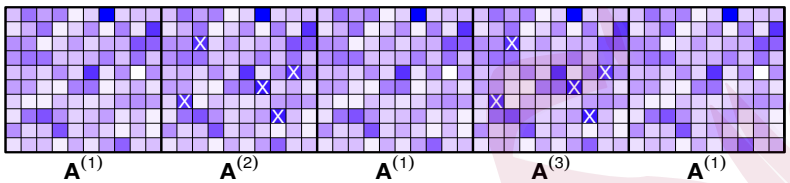
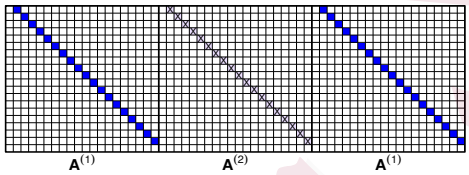
M1, M2	
M3, M4	
M5, M6	

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's indicate which elements change.

are nonzero. The coefficient matrix is generated randomly using the algorithm proposed by Ansley and Kohn (1986), and implemented using the R package `gmvarkit`, 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, we consider the single anomaly $[\eta_1, \eta_2]$, with the corresponding coefficient matrix $A^{(2)}$. However, we assume two collective anomalies, $[\eta_1, \eta_2]$ and $[\eta_3, \eta_4]$, for both M3 and M4, with coefficient matrices $A^{(2)}$ and $A^{(3)}$, respectively. To detect multiple anomalies, we use Algorithm 3, presented in Section S1 of the Supplementary Material. As stated in Assumption 4, only a few (10 for M1–M2

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

and 5 for M3–M4) entries in the VAR coefficient matrix change in an anomalous interval; see Table 1.

Under the settings M5–M8, we consider the case in which $A^{(1)}$ is sparse, that is, a smaller number of entries are nonzero. Similar to the settings used in Safikhani and Shojaie (2020), the one-off diagonal values of the coefficient matrix are nonzero as shown in Table 2. M7 and M8 are high-dimensional settings in the sense that the width of an anomaly ($\eta_2 - \eta_1$) is less than the dimension (p). When $A^{(1)}$ is unknown, we estimate it from the training data, using a ridge penalty for M1–M4 and 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 summarize the 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 does in all settings, regardless of the sparsity of $A^{(1)}$, the way we choose the intervals, or whether $A^{(1)}$ is known or estimated. The lasso-based method also outperforms the OLS-based approach in terms of the distance between the estimated and the true anomaly and its variance. As expected, compared with the results when the true $A^{(1)}$ is known, both methods perform less well when $\hat{A}^{(1)}$ is used. The

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

			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=1029)	OLS	100 (19)	93 (15)	1.59 (1.31)	5.43 (12.08)
		LSS	100 (19)	**99** (17)	1.47 (0.93)	1.95 (4.62)
	D (s=1029)	OLS	100 (43)	94 (33)	0.39 (0.25)	3.55 (11.64)
		LSS	100 (46)	**99** (40)	0.35 (0.16)	0.81 (4.55)
	D (s=540)	OLS	100 (25)	94 (19)	0.39 (0.33)	3.52 (11.65)
		LSS	100 (26)	**99** (27)	0.32 (0.22)	0.78 (4.55)
M2	R (s=1029)	OLS	98 (12)	69 (7)	2.85 (6.52)	17.67 (21.43)
		LSS	98 (18)	**89** (10)	2.63 (6.46)	7.31 (14.79)
	D (s=1029)	OLS	98 (44)	74 (31)	1.32 (6.57)	14.30 (21.31)
		LSS	**99** (50)	**90** (50)	0.82 (4.67)	5.63 (14.68)
	D (s=540)	OLS	98 (69)	72 (52)	1.27 (6.58)	15.14 (21.78)
		LSS	**99** (76)	**87** (71)	0.76 (4.68)	7.24 (16.77)
M5	R (s=499)	OLS	100 (20)	68 (12)	1.67 (1.21)	15.51 (20.22)
		LSS	100 (29)	**99** (33)	1.54 (0.99)	2.08 (4.41)
	D (s=499)	OLS	100 (46)	87 (48)	0.43 (0.24)	6.21 (15.00)
		LSS	100 (63)	**100** (75)	0.34 (0.12)	0.37 (0.13)
M6	R (s=499)	OLS	99 (14)	16 (5)	2.59 (4.66)	39.06 (16.45)
		LSS	**100** (21)	**68** (32)	1.90 (1.48)	15.66 (21.07)
	D (s=499)	OLS	100 (34)	34 (21)	0.45 (0.48)	30.63 (21.80)
		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. Random, deterministic, lasso are shortened to R, D, and LSS, respectively.

number of estimated anomalies located within the true anomaly tends to be proportional to the empirical power, and is larger when the segments are chosen deterministically, rather than randomly. Although it is not shown in the table,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

			#(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	OLS	27	**73**	0	1	**56**	43	3.28 (3.10)	8.29 (9.01)
	(s=1944)	LSS	21	**78**	1	0	39	**61**	2.87 (3.03)	5.02 (5.86)
	D	OLS	24	**76**	0	0	**53**	47	2.46 (2.62)	6.73 (9.07)
	(s=1944)	LSS	12	**86**	2	0	32	**68**	1.97 (2.68)	3.44 (5.16)
	D	OLS	26	**74**	0	1	**54**	45	2.59 (2.57)	7.13 (9.27)
	(s=1029)	LSS	21	**77**	2	0	35	**65**	2.43 (2.66)	3.50 (4.36)
M4	R	OLS	6	**93**	1	11	**64**	25	2.97 (4.53)	9.91 (4.69)
	(s=1944)	LSS	1	**96**	3	0	29	**71**	2.57 (5.43)	4.76 (5.22)
	D	OLS	4	**95**	1	6	**59**	35	1.92 (4.19)	8.59 (5.72)
	(s=1944)	LSS	1	**96**	3	0	21	**79**	1.50 (3.80)	3.53 (5.08)
	D	OLS	4	**95**	1	8	**59**	33	1.98 (4.22)	8.72 (5.61)
	(s=1029)	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. Random, deterministic, lasso are shortened to R, D, and LSS, respectively.

the mean of the Hausdorff distance computed from the estimated anomalies located within the true anomaly tends to be smaller than the one computed from the estimated anomalies that are not exactly located within the true anomaly. Comparing randomly and deterministically chosen segments of the same size, the deterministic way tends to give similar or slightly larger power for both methods, regardless of whether or not $A^{(1)}$ is known. Note that when $A^{(1)}$ is estimated in Table 3, the deterministically chosen intervals with a smaller sample size ($s = 540$) show larger power than those chosen randomly with a larger sample size ($s = 1029$), for both methods. Furthermore, the difference becomes

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

larger as the length of the anomalous interval becomes shorter (from M1 to M2, as presented in Table 1). Table 4 shows similar interpretations. Other simulation settings that include 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 turn to real data applications. In our first example, we apply our method to data on yellow taxi cab trips, previously analyzed 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`), and include the number of yellow taxi pickups recorded from 10 randomly selected zones in Manhattan, a borough in New York City. We aggregate the number of yellow taxi pickups every 30 minutes from March 11, 2019, to February 29, 2020, which results in 17088 time points. The raw data have an anomaly on November 3, 2019, caused by a daylight-saving time adjustment (Wu and Keogh, 2021), because the data for two hours are combined into a single hour when the time change occurred. To solve this, we simply divide the number of observations by two for the corresponding

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

hour and use the adjusted data. To prevent the detection procedure from 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 using a differenced version of the time series. The first 6835 data points are 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 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 we use the same minimum length $\lceil p/4 \rceil$ to analyze 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 that 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 10:30p.m. on December 31, 2019, and 4:00a.m. on January 1, 2020. Note that this anomaly is detected even after

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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, [Dec 31, 2019, 22 : 30, Jan 1, 2020, 04 : 00].

removing the holiday effect of 10 federal holidays from the period March 11, 2019, to February 29, 2020, which includes January 1, 2020. From Figure 2, we see that a sudden high demand occurred at the second and seventh zones located near to Times Square, but there was no such change for the third zone, which is located far from Times Square. Therefore, we interpret this to mean 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

The OLS method gives the same estimated anomaly as that identified by the lasso-based method, although it uses a larger $L = p = 10$. Compared with other methods designed to detect changes in a VAR model, Safikhani and Shojaie (2020) estimate eight changes including 4:30a.m. on Jan 3, 2020, and Bai et al. (2020) return eleven changes, including 00:30a.m. on Jan 1, 2020, which coincides with the anomaly estimated by our method. All estimated changepoints can be found in Section S4 of the the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

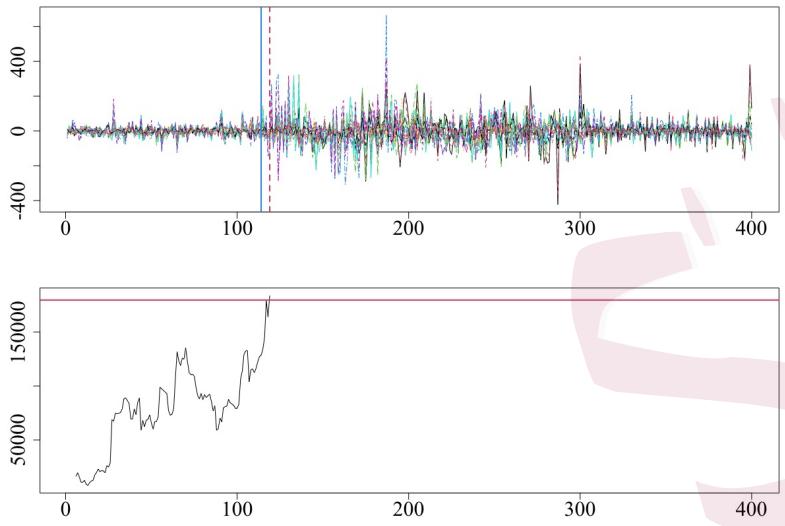


Figure 3: (Top) EEG data recorded at 18 different channels. The blue solid vertical line is the time at which the neurologist thinks the seizure starts, and the red dashed vertical line is the anomaly detected in the online setting. (Bottom) The maximum test statistics at each time point obtained using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

5.2 EEG Data

We now show how our method can be used as an online changepoint detection method. We demonstrate this using electroencephalogram (EEG) data collected from an epileptic patient. Other ways of analyzing this data set can be found in Ombao et al. (2001), Ombao et al. (2005), and Schröder and Ombao (2019). The data consist of brain electrical potentials recorded by placing electrodes on 18 locations on the scalp of the patient. The EEG signals are recorded during an epileptic seizure, and so there is a visible change in the data, as shown in Figure 3. The brain wave patterns are recorded over 500 seconds, with a sampling rate

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

of 100 Hz (i.e., 100 points per second). As in Safikhani and Shojaie (2020), to speed up computation, we use two observations per second, which reduces the number of time points to $T = 1000$.

We separate the data into a training set of size $T_1 = 600$ and a test set of 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: $\boldsymbol{X}$, λ^{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, we show how our method can be applied to the online framework; refer to Fisch et al. (2020) and Yu et al. (2021) for recent works on online detection algorithms for changepoints or anomalies. In the on-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

line setting, we make sequential decisions about the occurrence of an anomaly whenever a new observation 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 five 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 eight time points.

6. Discussion

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

Supplementary Material

The online Supplementary Material contains the technical proofs and additional simulation results.

Acknowledgments

We thank Hernando Ombao and Abolfazl Safikhani for providing access to the EEG data, and Yi Yu for the 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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 de-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

tection. *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*.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0181

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