Journal of Statistical Planning and Inference 239 (2025) 106277

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Neighborhood VAR: Efficient estimation of multivariate timeseries with neighborhood information

Zhihao Hu [a], Shyam Ranganathan [b,*], Yang Shao [c], Xinwei Deng [a]

[a] Department of Statistics, Virginia Tech, USA
[b] School of Mathematical and Statistical Sciences, Clemson University, USA
[c] Department of Geography, Virginia Tech, USA

ARTICLE INFO

Keywords:
Vector autoregression
High-dimensional timeseries
Model parsimony
Spatio-temporal data
Multivariate timeseries

ABSTRACT

Vector autoregression (VAR) models are popular in modeling multivariate time series in data sciences and other areas. When the number of time series is large, the number of parameters in the VAR model increases dramatically, posing great challenges for proper model estimation and inference. In this work, we propose a so-called neighborhood vector autoregression (NVAR) model to efficiently analyze large-dimensional multivariate time series. We assume that the time series have underlying neighborhood relationships, e.g., spatial or network, among them based on the inherent setting of the problem. When this neighborhood information is available or can be summarized using a distance matrix, we demonstrate that our proposed NVAR method provides a computationally efficient and theoretically sound estimation of model parameters. The performance of the proposed method is compared with other existing approaches in both simulation studies and a real-data application in environmental science.

1. Introduction

Modeling multivariate time series has attracted great attention from different areas such as environmental sciences (e.g. Davis et al. (2016), Cavalcante et al. (2017)), network sciences (e.g. Ma et al. (2015), Valdés-Sosa et al. (2005)), gene expression (e.g. Opgen-Rhein and Strimmer (2007), Fujita et al. (2007)), and economics (e.g. Rubio-Ramirez et al. (2010), Todd (1990), Bernanke et al. (2004), Nicholson et al. (2017)). With the increasing sophistication in data availability and methodology, recent developments in advanced manufacturing such as Steed et al. (2017), Hsu and Liu (2020), Ghahramani et al. (2020) have also started focusing on multivariate time series analysis.

In a number of these multivariate timeseries problems, the underlying source of the timeseries data exhibits dependencies either in the form of a spatial field, which can then be handled by spatiotemporal methods or Gaussian processes e.g., Cressie and Wikle (2015), or neighborhood information, where "nearby" timeseries are more correlated with the focal timeseries than "farther" ones, e.g., Guo et al. (2016). This notion of "neighborhood" is formalized in this paper based on the context of the problem, i.e., spatial timeseries or networked timeseries etc. using a distance matrix and, under the assumption that this distance matrix is known, this paper develops a new methodology to efficiently model multivariate timeseries. In cases where the distance matrix is not given, this paper provides heuristics to compute it based on the structural relationships of the timeseries for the problem at hand. The theoretical justification for the estimation method, a simulation study and an interesting illustration to an important management problem in water systems control using a stream nitrogen study are also demonstrated in this paper.

* Correspondence to: School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, USA
E-mail address: shyamr@clemson.edu (S. Ranganathan).

https://doi.org/10.1016/j.jspi.2025.106277
Received 25 July 2022; Received in revised form 15 March 2024; Accepted 18 February 2025
Available online 31 March 2025

When the number of time series is large, the number of parameters in the conventional vector autoregression model (VAR) increases dramatically, posing great challenges to performing proper estimation and inference on these time series data. Recent work to address this issue in the context of high-dimensional data has resulted in the notion of sparsity for efficient estimation of the VAR model, including the use of lasso penalization (e.g. Hsu et al. (2008), Arnold et al. (2007), Lozano et al. (2009), Song and Bickel (2011)), the group lasso penalization (e.g. Haufe et al. (2010), Shojaie and Michailidis (2010), Basu et al. (2015), Bolstad et al. (2011)), Dantzig-type penalization (e.g. Qiu et al. (2015), Han et al. (2015)), adaptive lasso penalization (e.g. Ren and Zhang (2010), Kock and Callot (2015)), graph regularization (e.g. Jiang et al. (2015)), among many others. Such strategies typically adopt regularization to reduce the number of parameters in the estimated model, achieving model parsimony.

In many applications with multiple time series, e.g., modeling the nitrogen of multiple streams that is considered here, there is a natural dependence structure among time series collected from neighboring locations. The spatial nature of the problem implies that two time series from nearby locations are more likely to correlate with each other than those from far-away locations. Without loss of generality, let us assume that each time series is collected from a sensor at a location. Here we use the metaphor of a "sensor" though the so-called sensor may not have a physical meaning, and just represents the location where the time series originated from. The term "location" is also not restricted to physical location, and instead means that the multiple timeseries are assumed to have a correlation with each other that can be measured in terms of their distances under some distance measure in a metric space. For instance, for time series data over networks, one could consider an embedding of the network in a latent space (e.g. Hoff et al. (2002)) and measure distances between nodes based on the distances in the latent space. This is different from the typical spatial statistics or spatio-temporal statistics problem, where there is an underlying spatial process that results in spatial observations at different locations.

In the work of Guo et al. (2016), the time series are assumed to be located on a one-dimensional lattice at equally spaced intervals. Their so-called banded VAR (BVAR) method has a clear interpretation and can effectively reduce model complexity while outperforming penalization-based algorithms like the LASSO. But the assumption of a sequential ordering of time series is too restrictive for general applications and it would be more reasonable to consider a more generalized notion of location on a vector space rather than on a one-dimensional lattice.

In this article, we propose the Neighborhood Vector AutoRegression model (NVAR), which extends the notion of "band" in Guo et al. (2016) to the notion of a "neighborhood". Having a more general modeling assumption, the proposed method maintains efficient parameter estimation with clear model interpretation. The proposed method also guarantees the convergence rate of the estimated coefficient matrix. In particular, we show that the asymptotic properties proved in the Guo et al. (2016) paper hold for our proposed method. In the case study of modeling the nitrogen content of multiple streams, the proposed NVAR method takes advantage of the notion of neighborhood to model the water quality data from multiple streams in a joint manner. In the United States, excess nutrients are one of the most important causes of impairment for rivers and streams (see EPA (2000)). Increasing rates of nutrient supply fuels accelerating primary production or eutrophication, which leads to discoloration of affected waters (see Paerl et al. (2001)). The proposed method provides a useful tool to understand how the water quality changes over time and also the interactions of water systems at different locations. In comparison with several existing approaches, the case study shows the merits of the proposed method with a reasonable computational cost.

The remainder of this work is organized as follows. Section 2 describes the neighborhood VAR model and the algorithm. Section 3 provides some theoretical results for the proposed method. Section 4 discusses simulation results and we conclude this work with a practical application in Section 5.

2. The proposed model

Denote $\mathbf{y}(t) = [y_1(t), y_2(t), \dots, y_p(t)]^T$ as the p dependent time series, where $y_i(t) \in \mathbb{R}$ and $t = 1, 2, \dots, n$. We assume that the time series $y_1(t), \dots, y_p(t)$ are collected from "sensors" located at $s_1, \dots, s_p$ with $s_i \in (M, d)$, where M is a metric space with distance measure defined by d, say $(\mathbb{R}^m, d)$. Here the d is a pre-defined distance measure with distances $d(s_i, s_j)$ for $s_i, s_j \in M$. Hereafter we will abbreviate $d(s_i, s_j)$ as $d(i, j)$ for notation convenience. Since the sensors have a distance measure, we can define "d_0-neighborhood" of the ith time series as $\mathcal{N}_i^{d_0} = \{j : d(i, j) \le d_0\}$ for some $d_0 \in R$, and is a representation of the set of all time series that have an influence on the ith time series at this particular distance level d_0. Thus we consider the NVAR(q) model in a form similar to the classical VAR(q) model as

$$\mathbf{y}(t) = A_1\mathbf{y}(t-1) + A_2\mathbf{y}(t-2) + \cdots + A_q\mathbf{y}(t-q) + \mathbf{e}(t), t = 1, 2, \dots, n, \tag{1}$$

with $A_s(i, j) = 0, s = 1, \dots, q, j \notin \mathcal{N}_i^{d_0}$ for some "d_0-neighborhood" of i, $\mathcal{N}_i^{d_0}$. Here q is the lag order for the autoregressive process. The coefficient matrices $A_1, \dots, A_q$ are $p \times p$ matrices that represent the dependence between the different time series. We do not use bold typeface for A to enhance readability. The $\mathbf{e}(t)$ is the serially uncorrelated noise with $E(\mathbf{e}) = 0$ and $var(\mathbf{e}) = \Sigma_e$. Note that we need to assume conditions on the coefficient matrices for the time series to be stationary, and we make the assumption that $|I - A_1 z - \cdots - A_q z^q| \neq 0$ for any $|z| \le 1$. Hereafter, we will drop the time index t where possible to enhance readability.

Note that our proposed neighborhood idea generalizes the notion of bandwidth in the banded VAR model and has parallels to work by Besag (1974) among others. The proposed NVAR model considers the time series to be homogeneous in the sense that the same d_0 is sufficient to characterize the neighbors' influence for every time series. It can also be extended to the case with different values of d_0 being estimated for each individual time series y_i but we do not consider that case in this paper. When there is no direct mapping of time series locations to a metric space, we can potentially embed these time series in a latent space, and obtain distances from this embedded space. This adds another layer of uncertainty due to estimation of the distance matrix itself. However,

in the rest of this paper, we will not deal with these technical complications and we assume that the distances are well-specified and given by the symmetric $p \times p$ matrix $\mathbf{D}$. Note that time-varying distances are allowed in this algorithm as the sensors may be allowed to move in time. This can be included in our algorithm by specifying a different matrix $\mathbf{D}(\mathbf{t})$ at each time instant. For easy comprehension though, in this paper, we assume that the coefficient matrices are time-invariant.

2.1. Construction of neighborhood

For every time series y_i, we define a d-neighborhood as the set of all time series that are at most d distance away from it. This is denoted by

$$\mathcal{N}_i^d = \{j : d(i,j) \leq d\}.$$

The neighborhood VAR model assumes that y_i depends only on its $d_0 - neighborhood$ for some value of $d = d_0$. This is a generalization of bandwidth in the banded VAR and allows us to handle more complex time series. Specifically, if we assume that the time series reside on a 1-D lattice at locations $1, 2, \ldots, p$, and distance is measured as $d(i,j) = |i-j|$, we get back the banded VAR formulation with bandwidth $= d_0$.

Our definition of neighborhood VAR model assumes that the elements of the coefficient matrix are non-zero only at locations that are within the $d_0 - neighborhood$ of each of the time series i. That is,

$$A_r(i,j) = 0, \text{ if } j \notin \mathcal{N}_i^{d_0},$$

where $A_r(i,j)$ represents the $(i,j)^{th}$ element of the coefficient matrix A_r corresponding to a lag of r. Thus the maximum number of non-zero elements in row i of the coefficient matrices is given by $\tau_i = |\mathcal{N}_i^{d_0}|$. The sparsity assumption implies that $\tau_i \ll p$.

Below are a few illustrative cases on constructing the neighborhood.

Banded structures: The banded matrix structure of A_r discussed in the banded VAR literature is obtained if the time series are assumed to be present in locations that are arranged along a line segment such that $s_1 < s_2 < \cdots < s_p$ and $d(s_i, s_j) = |i-j|$.

Block-banded structures: A block banded structure is obtained if the locations of the time series correspond to locations arranged in 2-D space (e.g., pixels in an image matrix) with equal distances between neighboring locations and $d(s_i, s_j) = |i-j| \mod \sqrt{p}$. Here the locations are in a $\sqrt{p} \times \sqrt{p}$ lattice and the s_i are obtained via vectorizing the lattices into a p-dimensional set of locations in a row-by-row fashion. Note that this is equivalent to consider each location to be affected by geographically close locations as measured using a city-block distance metric. Other formulations and distance metrics lead to different structures on 2-D data.

Neighborhood structure under spatial data: In spatial–temporal data, the Euclidean distance between the sensor locations can be used for the distance matrix of sensors. The sparsity structure of the coefficient matrices will depend on the actual distances between the sensors. Note that it is important, in this case, to normalize the distances to avoid identifiability problems in terms of the scale, and also to preserve the spatial meaning of neighborhood in terms of the actual problem. A good scaling constant that can be used will approximate the spatial scale to a lattice by scaling distances as $\frac{(N/2)}{(d_{max})^m}$, where N is the number of timeseries, m is the dimension of the space (1, 2 or 3 for typical spatial problems), and d_{max} is the maximum distance among all pairs of sensors.

Neighborhood structure under network data: In a network application, the time series are often from sensors that are connected to each other under a network. The distance matrix can be specified by the adjacency matrix of the network, with the length of the shortest path between two nodes giving the distance between the two nodes. For a more general formulation, we can embed the network in a latent space and compute the distance metric based on distances on the latent space.

2.2. Parameter estimation

Given a particular value of d_0, and subsequently $\mathcal{N}_i^{d_0}$ (which can be computed directly since $\mathbf{D}$ is assumed to be known), the estimation of the NVAR model can be conducted using the ordinary least squares (OLS) estimation of the coefficients corresponding to each time series. For instance, if A_i is the set of all coefficients to be estimated for the ith time series, we can obtain A_i from $\mathcal{N}_i^{d_0}$ and the lag order q by selecting the appropriate elements from the VAR coefficient matrices. In the simple case of VAR(1), where the only coefficient matrix is A_1, the OLS equation for the ith time series is simply

$$\hat{y}_i(t) = \sum_{j \in \mathcal{N}_i^{d_0}, r \in \{1,\ldots,q\}} \hat{A}_1(i,j) y_j(t-r),$$

where $\hat{A}_1(i,j)$ is the $(i,j)^{th}$ element in the estimated coefficient matrix $\hat{A}_1$, and hence the estimates for the coefficient matrices are obtained in a straightforward manner, with all elements of the matrix outside the d_0-neighborhood set to 0.

For NVAR(q) model, we can also use the ordinary least squares (OLS) estimation to estimate parameters separately in each time series variables with respect to its $d_0 - neighborhood$. Let us denote $\mathbf{y}_i = (y_i(n), y_i(n-1), \ldots, y_i(q+1))^T$ and

$$\mathbf{X}_i = \begin{pmatrix} \{\mathbf{y}_j(n-1)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \{\mathbf{y}_j(n-2)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \cdots & \{\mathbf{y}_j(n-q)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} \\ \{\mathbf{y}_j(n-2)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \{\mathbf{y}_j(n-3)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \cdots & \{\mathbf{y}_j(n-q-1)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} \\ \vdots & \vdots & \vdots & \vdots \\ \{\mathbf{y}_j(q)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \{\mathbf{y}_j(q-1)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} & \cdots & \{\mathbf{y}_j(1)\}^T\big|_{j \in \mathcal{N}_i^{d_0}} \end{pmatrix}_{(n-q) \times q |\mathcal{N}_i^{d_0}|},$$

where $\{\mathbf{y}_j(t)\}\big|_{j\in\mathcal{N}_i^{d_0}} = \left(y_{j_1}(t), \quad y_{j_2}(t), \quad \ldots, \quad y_{j_{|\mathcal{N}_i^{d_0}|}}(t)\right)^T$, $j_1, j_2, \ldots, j_{|\mathcal{N}_i^{d_0}|} \in \mathcal{N}_i^{d_0}$ is a column vector, and $|\mathcal{N}_i^{d_0}|$ is the size of $\mathcal{N}_i^{d_0}$. Then we denote the coefficients matrix as

$$\boldsymbol{B} = \left(\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2 \quad \cdots \quad \boldsymbol{\beta}_p\right)_{p\times q|\mathcal{N}_i^{d_0}|},$$

where $\boldsymbol{\beta}_i = \begin{pmatrix} \{A_1(i,j)\}\big|_{j\in\mathcal{N}_i^{d_0}} \\ \{A_2(i,j)\}\big|_{j\in\mathcal{N}_i^{d_0}} \\ \vdots \\ \{A_q(i,j)\}\big|_{j\in\mathcal{N}_i^{d_0}} \end{pmatrix}$, and $\{A_q(i,j)\}\big|_{j\in\mathcal{N}_i^{d_0}} = \begin{pmatrix} A_q(i,j_1) \\ A_q(i,j_2) \\ \vdots \\ A_q(i,j_{|\mathcal{N}_i^{d_0}|}) \end{pmatrix}$, $j_1, j_2, \ldots, j_{|\mathcal{N}_i^{d_0}|} \in \mathcal{N}_i^{d_0}$ is a column vector. The coefficient matrix $\boldsymbol{B}$ are estimated by minimizing the least-squares objective function as

$$\min_{\boldsymbol{B}} \sum_{i=1}^{p} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i).$$

It is easy to see that the optimization can be separated into p independent OLS estimation,

$$\min_{\boldsymbol{\beta}_i} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i), \ i = 1, 2, \ldots, p.$$

Thus we can have

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_j^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{y}_i, \ i = 1, 2, \ldots, p. \tag{2}$$

With the estimates $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_1$, the estimated model can then be expressed as

$$\hat{y}_i(t) = \sum_{j\in\mathcal{N}_i^d} \hat{A}_1(i,j)y_j(t-1) + \cdots \hat{A}_q(i,j)y_j(t-q), \ t = 2, \ldots, n. \tag{3}$$

To choose the optimal value of d_0, we use the Bayesian information criterion (BIC) by computing

$$BIC(d,i) = \log(RSS(d,i)) + \frac{1}{n}q\tau_i C_n \log(p \vee n), \tag{4}$$

for every time series i and for $d = 1, 2, \ldots$. The value of d that minimizes this quantity is the optimal value for time series i, i.e., $d_0(i)$. Since we assume the p time series are homogeneous, we will consider the estimate of d for all time series as the maximum of all the estimated optimal neighborhood distances. That is,

$$\hat{d_0} = \max_{1\le i\le p}\{\operatorname*{argmin}_d BIC(d,i)\} \tag{5}$$

In the next section, we will show that our estimation algorithm for d_0 will lead to optimal estimation of the neighborhood distance. Note that our estimate may result in a number of predictors being given as relevant for any time series depending on the density of arrangement in space, and hence an optional step may be used to compute a BIC among these predictors (or sany other variable selection procedure) to further reduce the number of predictors. The estimation of lag order q can also be wrapped into this same algorithm by searching over a grid of d and q values and choosing the value of $\hat{d_0}$ and $\hat{q}$ that maximize the BIC criterion (see Guo et al. (2016) for a similar idea).

Algorithm 1 Neighborhood VAR Estimation

Input: time series $y_1(t), \ldots y_p(t)$, lag order q, and distance matrix $\mathbf{D}$
Output: Coefficient matrices: $\hat{A}_1, \ldots, \hat{A}_q$
for d *in* $1 : d_{max}$ **do**
  for i *in* $1 : p$ **do**
    Find the d-neighborhood $\mathcal{N}_i^d$ of the i^{th} time series
    Perform regression for the i^{th} time series on $\mathcal{N}_i^d$ and compute coefficients $\beta^{d,i}$
    Compute the marginal BIC as $BIC(d,i) = \log(RSS(d,i)) + \frac{1}{n}q\tau_i C_n \log(p \vee n)$, τ_i - the number of non-zero elements in row i of the coefficient matrices
  end for
end for
Find $\hat{d} = \max_{1\le i\le p}\{\operatorname{argmin}_{1\le d\le d_{max}} BIC(d,i)\}$

Note that the proposed algorithm could be vulnerable to model misspecification. It is worth to remarking that the motivation for the proposed model is to address concerns with model misspecification in the case of the Banded VAR method. The use of the least squares algorithm and the BIC criterion for parsimony suggest the typical problems with model misspecification in terms of linearity assumptions, normality of error variable, homoscedasticity etc. In these cases of severe misspecification, the usual problems of biased coefficients often occur and suitable modifications need to be made to accommodate these situations. If a severe model misspecification such as stationarity assumptions are not met, the VAR model could be severely misspecified and new modeling effort needs to be made.

3. Theoretical properties of estimation consistency

In this section, we show that, under appropriate regularity conditions, the proposed NVAR method can consistently recover the appropriate level of sparsity, in terms of the optimal neighborhood distance. In addition, we establish the convergence rate of the estimated coefficient matrix to the true coefficient matrix. The regularity conditions and the theorems are presented here, while the proofs are in Appendix A.

First we will formulate the NVAR(q) model of order q into a NVAR(1) mode of order 1 as follows.

$$\tilde{\mathbf{y}}(t) = \tilde{A}\tilde{\mathbf{y}}(t-1) + \tilde{\mathbf{e}}(t),$$

where

$$\tilde{\mathbf{y}}(t) = \begin{pmatrix} \mathbf{y}(t) \\ \mathbf{y}(t-1) \\ \vdots \\ \mathbf{y}(t-q+1) \end{pmatrix}, \tilde{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_{q-1} & A_q \\ I_p & 0_p & \cdots & 0_p & 0_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_p & 0_p & \cdots & I_p & 0_p \end{pmatrix}, \tilde{\mathbf{e}}(t) = \begin{pmatrix} \mathbf{e}(t) \\ 0_{p\times 1} \\ \vdots \\ 0_{p\times 1} \end{pmatrix}. \tag{6}$$

Such a reformulation provides a good framework for investigating the theoretical properties. Next we need several regularity conditions that are stated as follows. Note that these regularity conditions are similar to those imposed the banded VAR approach (see Guo et al. (2016)).

- *Condition* 1. For $\tilde{A}$ defined in (6), $\|\tilde{A}\|_2 \le C$ and $\|\tilde{A}^{j_0}\|_2 \le \delta^{j_0}$, where $C > 0$, $\delta \in (0,1)$ and $j_0 \ge 1$ are constants free of n and p, and j_0 is an integer.
- *Condition* 1′. For $\tilde{A}$ defined in (6), $\|\tilde{A}^{j_0}\|_2 \le \delta^{j_0}$, $\|\tilde{A}\|_\infty \le C$ and $\|\tilde{A}^{j_0}\|_\infty \le \delta^{j_0}$, where $C > 0$, $\delta \in (0,1)$ and $j_0 \ge 1$ are constants free of n and p, and j_0 is an integer.
- *Condition* 2. Let $a_{ij}^{(l)}$ be the (i,j)-th element of A_l. For each $i = 1,\dots,p$, at least one $j \in \mathcal{N}_i^{d_0}$, $\{C_n \tau_i^{d_0} n^{-1}\log(p\vee n)\}^{1/2} \ll |a_{ij}^{(l)}|$ for some $1 \le l \le q$, where $C_n \to \infty$ as $n \to \infty$.
- *Condition* 3. The minimal eigenvalue $\lambda_{\min}\{\mathrm{cov}(\mathbf{y}(t))\} \ge \kappa_1$ and $\max_{1\le i/le p}|\sigma_{ii}| \le \kappa_2$ for some positive constants κ_1 and κ_2 free of p, where σ_{ii} is the ith diagonal element of $\mathrm{cov}(\mathbf{y}(t))$, and $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.
- *Condition* 4. The serial noise $\{\mathbf{e}(t) : t = 1,2,\dots,n\}$ is independent and identically distributed with zero mean and covariance Σ_e. Furthermore, one of the two assertions below holds:
  - (i) $\max_{1\le i\le p} E(|\mathbf{e}_i(t)|^{2q}) \le C$ and $p = O(n^\beta)$, where $q > 2$, $\beta \in (0,(q-2)/4)$ and $C > 0$ are some constants free of n and p;
  - (ii) $\max_{1\le i\le p} E\{\exp(\lambda_0|\mathbf{e}_i(t)|^{2\alpha})\} \le C$ and $\log p = o(n^{\alpha/(2-\alpha)})$, where $\lambda_0 > 0$, $\alpha \in (0,1]$ and $C > 0$ are some constants free of n and p.

Briefly, Condition 3 ensures that the covariance matrix is positive definite. Condition 4(i) ensures strict stationarity of the process when the $e_i(t)$ are independent and identically distributed, while Condition 4(ii) is an identifiability condition for the minimum value among the non-zero coefficients.

Now we can show that the estimated size of neighborhood to be consistent.

Theorem 1. *Assume that Conditions 1–4 hold for the proposed neighborhood VAR. Then the estimated size of the neighborhood is consistent, i.e.,* $Pr(\hat{d} = d_0) \to 1$ *as* $n \to \infty$.

This theorem shows that the algorithm to select the optimum neighborhood distance converges to the true neighborhood distance, if it exists, as the length of the time series grows. Moreover, we can also establish the convergence error bounds between the estimated coefficient matrix $\hat{A}_r$ and the true coefficient matrix A_r, when using this optimum neighborhood distance in computing the estimated coefficient matrix. Using either the Frobenius norm or the L_2 norm, we can show bounds on the norm of the error matrix defined as the difference between the estimated coefficient matrix and the true coefficient matrix. The following theorem shows that, for each of the q coefficient matrices, the error norm is bounded and goes to 0 as the length of the time series grows with constant p.

Theorem 2. *Assume that Conditions 1–4 hold for the proposed neighborhood VAR. Then as* $n \to \infty$, *we have the following error bounds of the estimated coefficient matrix as*

$$\|\hat{A}_j - A_j\|_F = O_p\left\{(n^{-1}\log p)^{1/2}\right\},$$
$$\|\hat{A}_j - A_j\|_2 = O_p\left\{(n^{-1}\log p)^{1/2}\right\},$$

which hold for $j = 1,\dots,q$.

From this theorem, one can infer that the estimated coefficient matrix is accurate even when p grows along with n as long as the number of time series grows at a particular rate, but not as fast as n as seen from the theorem. Further details are available in Appendix A.

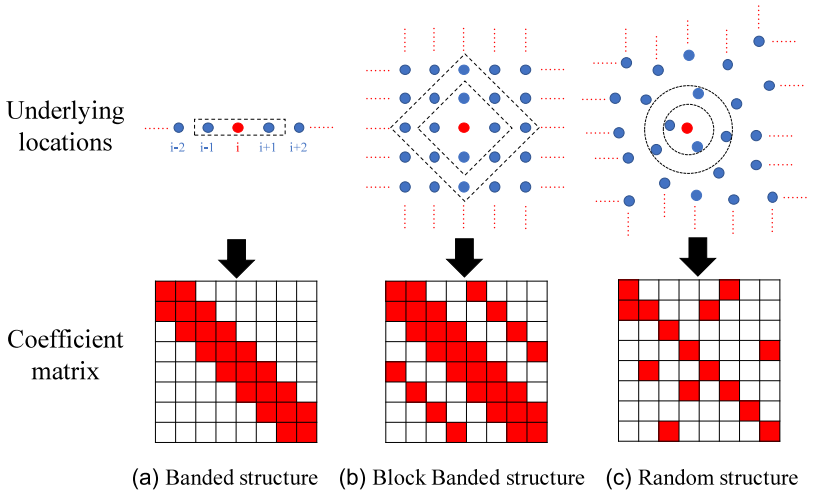

Fig. 1. Three simulation cases: banded, block-banded and random structures.

Here we would like to remark that using $\delta \in (0,1)$ in Condition 1 can be a stringent condition for stationarity as discussed in Basu et al. (2015). Note that we do not use the even more stringent condition of $\|\tilde{A}\|_2 \leq 1$ that is remarked on in that paper. While the relaxation to a condition on the spectral radius is desirable as proposed in that work, we do not address this here. Instead, for practice in the sections of simulation and case study, we consider to scale the transition matrix. In the model, this would introduce a scaling on the error standard deviation, and we show that the algorithm works under different conditions imposed on the error standard deviation. Further tightening of bounds and exploration of ideas around the spectral radius can be a future work.

4. Simulation study

In this section, we evaluate the performance of the proposed method in comparison with some existing methods. Three different simulation cases are conducted with data generated from the NVAR(1) in (1). Fig. 1 illustrates the three simulation cases. In all these cases, we consider different values of bandwidth $d_0 = 1, 2, 3, 4$ and different number of time series $p = 100, 196, 400, 784$. We also assume that the distance matrix is known and that the noise process is $N(0, \sigma_e^2 I_p)$ with $\sigma_e = 0.01, 1$. The sample size n is fixed at $n = 200$. For each case, we create 500 repetitions for the simulation study.

Case 1: 1-D lattice structure. We generate random coefficient matrices A with $A_{ij} = 0, |i-j| \geq d_0$. The non-zero elements of the matrix are chosen randomly from $[-1, 1]$ and to ensure stationarity, we force $\|A\| < 1$ by rescaling the matrix to have a random norm value between $[0.3, 0.9]$ using the operation $(A/\|A\|) \times u$, where $u \sim U[0.3, 0.9]$.

Case 2: 2-D lattice structure. We generate random coefficient matrices A with the 2-D lattice structure shown in the previous section. From the spatial perspective, for any point $[i_1, j_1]$ the surrounding points i_2, j_2 with $|i_1 - i_2| + |j_1 - j_2| < d_0$ are the only ones that are non-zero. When we vectorize the spatial matrix to obtain the coefficient matrix A_r, this results in a block-banded structure where, for the ith row in A_r, the non-zero elements are the matrix positions with $|i-j| \leq d_0 \pm t\sqrt{p}, t = 0, 1, 2, \ldots \sqrt{p}, 1 \leq j \leq p$. This means, as we increase neighborhood distance in steps, multiple time series are included, proportional to $O(d^2)$ rather than to $O(d)$ as in the 1-D lattice case. We ensure stationarity of the timeseries process as before.

Case 3: 2-D spatial structure. We generate a random spatial point process and place the "sensors" in these locations and then generate data based on an NVAR(1) process, where each time series depends only on its neighbors in space. To ensure a fair comparison, we maintain the density of the sensors in any small region to be similar to that of the 2-D lattice on average by suitable scaling. In this case, the number of neighbors is random and as we increase d_0 in steps corresponding to the lattice case, multiple time series are included in the neighborhood. For instance, we first generate a spatial point process in $[0, 1] \times [0, 1]$. We scale this unit square appropriately based on the particular instances of the point processes so that the average number of nearest neighbors for all points is roughly 4. Note that, as the number of time series increases, i.e., as p increases in the simulation setting, to maintain the same average number of neighbors, we need to scale the unit square differently.

The proposed method is compared with the banded VAR method and the LASSO method. In the LASSO method, we use the Lasso regression to estimate the coefficients for each time series independently, which is also chosen as a benchmark method in Guo et al. (2016). The setting of Case 1 is to validate that our proposed method is equivalent to the banded VAR method under the 1-D lattice structure.

Note that error standard deviation $\sigma_e = 0.01$ corresponds to very high signal and we would like to recover the true coefficient matrix exactly in this scenario. On the other hand, error standard deviation $\sigma_e = 1$ corresponds to almost all noise, and hence almost

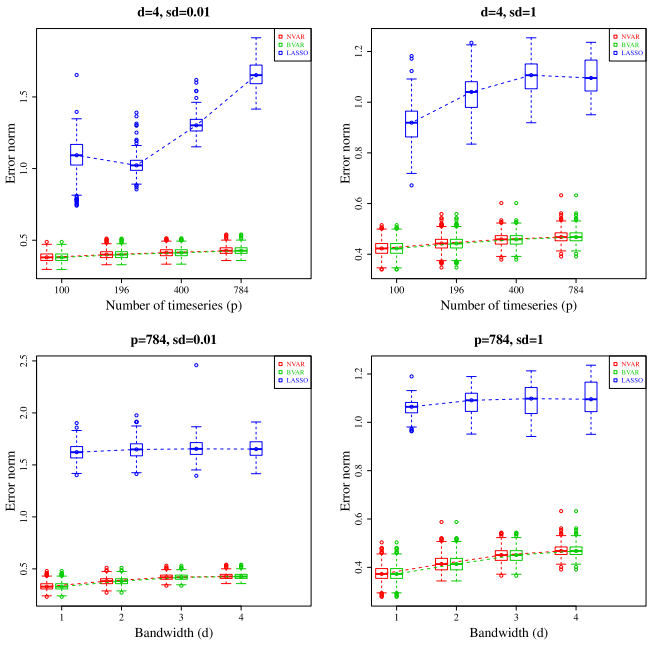

Fig. 2. Performance comparison of methods with different number of time series (p), different bandwidth (d), and different error standard deviation (sd) in Case 1. The error norm is $\|\hat{A} - A\|_2$.

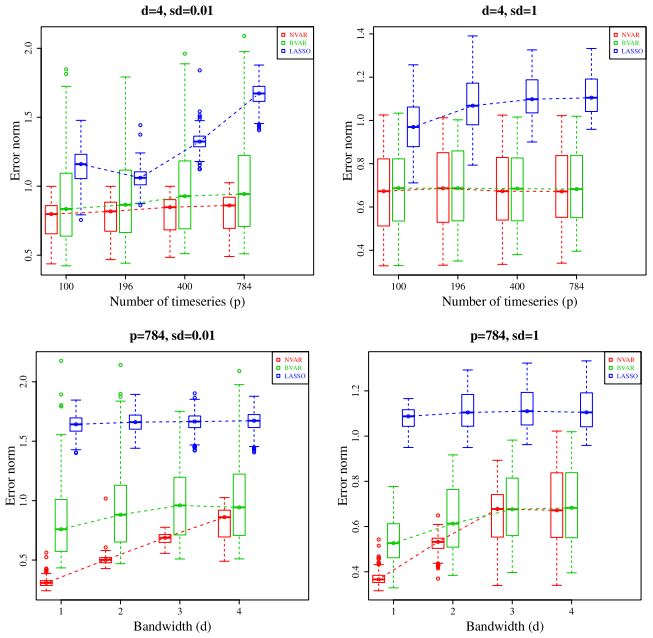

Fig. 3. Performance comparison of methods with different number of time series (p), different bandwidth (d), and different error standard deviation (sd) in Case 2. The error norm is $\|\hat{A} - A\|_2$.

all methods will perform relatively poorly. Here we consider the L_2 error norm (spectral norm), i.e. $\|\hat{A} - A\|_2$, as the performance measure, where $\hat{A}$ is the estimated coefficient matrix, and A is the true coefficient matrix.

The simulation results are reported in Figs. 2–4 and Tables 1–3. In Case 1, it is clear from Figs. 2 and Table 1 that the banded VAR method and the proposed neighborhood VAR method coincide exactly, and both perform much better than the LASSO method, irrespective of the noise variance.

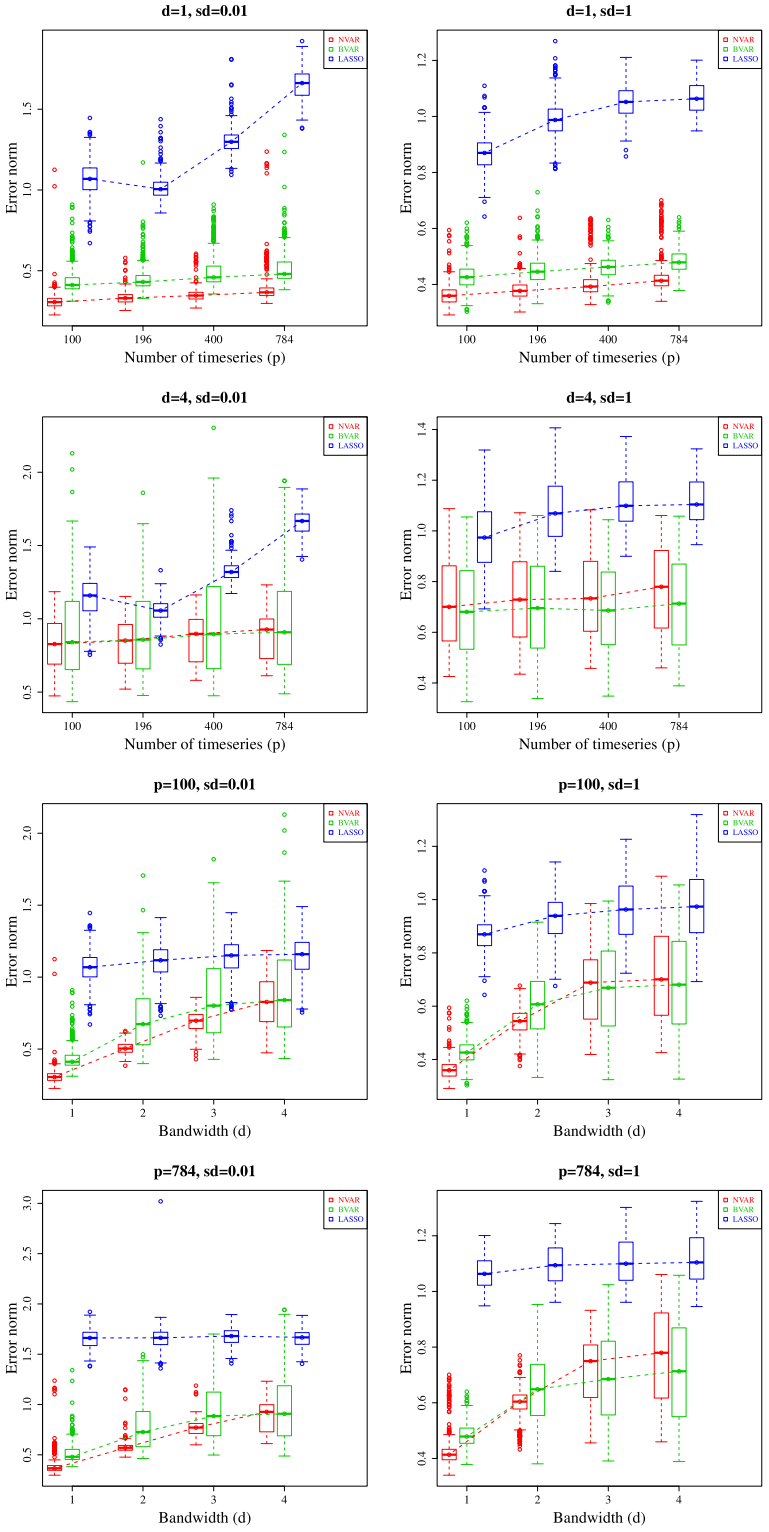

Fig. 4. Performance comparison of methods with different number of time series (p), different bandwidth (d), and different error standard deviation (sd) in Case 3. The error norm is $\|\hat{A} - A\|_2$.

Table 1
Case 1: Banded Structure. Means with their corresponding standard deviations in parentheses of the errors, and the frequency of estimated bandwidth in estimating coefficient matrix.

Error standard deviation = 1

		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm $\|\hat{A}-A\|_2$	Est. bandwidth					L_2 norm $\|\hat{A}-A\|_2$	L_2 norm $\|\hat{A}-A\|_2$
		0	1	2	3	4		0	1	2	3	4		
100	1	0	327	161	10	2	0.30(0.05)	0	327	161	10	2	0.30(0.05)	0.84(0.06)
100	2	0	9	351	133	7	0.35(0.04)	0	9	351	133	7	0.35(0.04)	0.87(0.06)
100	3	0	18	51	340	91	0.39(0.03)	0	18	51	340	91	0.39(0.03)	0.90(0.07)
100	4	1	39	64	49	347	0.42(0.03)	1	39	64	49	347	0.42(0.03)	0.91(0.08)
196	1	0	251	232	16	1	0.32(0.04)	0	251	232	16	1	0.32(0.04)	0.96(0.06)
196	2	0	5	295	183	17	0.37(0.04)	0	5	295	183	17	0.37(0.04)	0.99(0.06)
196	3	0	7	33	312	148	0.41(0.03)	0	7	33	312	148	0.41(0.03)	1.02(0.07)
196	4	0	19	48	51	382	0.44(0.03)	0	19	48	51	382	0.44(0.03)	1.03(0.07)
400	1	0	136	327	36	1	0.35(0.04)	0	136	327	36	1	0.35(0.04)	1.04(0.04)
400	2	0	0	223	252	25	0.39(0.03)	0	0	223	252	25	0.39(0.03)	1.07(0.05)
400	3	0	0	20	282	198	0.43(0.03)	0	0	20	282	198	0.43(0.03)	1.08(0.06)
400	4	0	1	42	39	418	0.46(0.03)	0	1	42	39	418	0.46(0.03)	1.10(0.07)
784	1	0	52	395	49	4	0.38(0.03)	0	52	395	49	4	0.38(0.03)	1.06(0.03)
784	2	0	0	156	297	47	0.42(0.04)	0	0	156	297	47	0.42(0.04)	1.08(0.05)
784	3	0	2	14	242	242	0.45(0.03)	0	2	14	242	242	0.45(0.03)	1.09(0.06)
784	4	0	1	33	50	416	0.47(0.03)	0	1	33	50	416	0.47(0.03)	1.10(0.07)

Error standard deviation = 0.01

		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm $\|\hat{A}-A\|_2$	Est. bandwidth					L_2 norm $\|\hat{A}-A\|_2$	L_2 norm $\|\hat{A}-A\|_2$
		0	1	2	3	4		0	1	2	3	4		
100	1	0	278	207	13	2	0.24(0.16)	0	278	207	13	2	0.24(0.06)	1.01(0.11)
100	2	0	0	314	168	18	0.30(0.05)	0	0	314	168	18	0.30(0.05)	1.06(0.10)
100	3	0	0	0	330	170	0.35(0.04)	0	0	0	330	170	0.35(0.04)	1.08(0.12)
100	4	0	0	1	24	475	0.38(0.03)	0	0	1	24	475	0.38(0.03)	1.09(0.12)
196	1	0	176	297	25	2	0.27(0.05)	0	176	297	25	2	0.27(0.05)	0.98(0.09)
196	2	0	0	223	259	18	0.32(0.04)	0	0	223	259	18	0.32(0.04)	1.01(0.09)
196	3	0	0	0	280	220	0.37(0.04)	0	0	0	280	220	0.37(0.04)	1.01(0.07)
196	4	0	0	0	8	492	0.40(0.03)	0	0	0	8	492	0.40(0.03)	1.03(0.06)
400	1	0	68	398	31	3	0.31(0.04)	0	68	398	31	3	0.31(0.04)	1.28(0.09)
400	2	0	0	103	356	41	0.36(0.04)	0	0	103	356	41	0.36(0.04)	1.28(0.07)
400	3	0	0	0	174	326	0.40(0.03)	0	0	0	174	326	0.40(0.03)	1.30(0.07)
400	4	0	0	0	1	499	0.41(0.03)	0	0	0	1	499	0.41(0.03)	1.31(0.07)
784	1	0	12	412	73	3	0.33(0.04)	0	12	412	73	3	0.33(0.04)	1.62(0.08)
784	2	0	0	39	399	62	0.38(0.04)	0	0	39	399	62	0.38(0.04)	1.65(0.09)
784	3	0	0	0	99	401	0.42(0.03)	0	0	0	99	401	0.42(0.03)	1.66(0.13)
784	4	0	0	0	1	499	0.43(0.03)	0	0	0	1	499	0.43(0.03)	1.65(0.09)

In Case 2, the neighborhood VAR method outperforms the banded VAR method and the LASSO method, especially when the error variance is small and the bandwidth is small (Figs. 3, Table 2). We note here that the banded VAR method is adapted to the 2-D lattice case in a natural manner as the original paper Guo et al. (2016) explicitly defines only a one-dimensional problem. The neighborhood VAR method is significantly better than the banded VAR method at low error variance. This is because banded VAR does not take the possibility of non-zero elements in the coefficient matrix far away from the main diagonal. The banded VAR method and the neighborhood method provide a comparable performance when the error variance is high, and both outperform LASSO significantly.

In Case 3, the simulation results are reported in Figs. 4 and Table 3. From these results, one can see that the neighborhood VAR method outperforms the banded VAR method in this case even for the normalized distances.

Recall from our simulation settings that we have re-scaled the spatial distances so that the average number of neighbors for any time series is roughly the same as in a 2-D lattice in order to facilitate a fair comparison with the banded VAR method. This result indicates that for general spatial problems where the density of neighbors may be very different from that of a lattice process, the neighborhood VAR method can easily outperform the banded VAR method. Note that when the number of time series grows large, or when the bandwidth grows large at fixed number of time series, the neighborhood VAR method and the banded VAR method become comparable. It is clear that the neighborhood VAR method outperforms both the banded VAR method and the LASSO method significantly at low bandwidth, and the neighborhood VAR method and the banded VAR method are comparable at high bandwidth.

We have also checked the performance of prediction accuracy for the proposed method in comparison with the BVAR and the LASSO methods. Specifically, based on the fitted model, we conduct the one-step ahead prediction for 50 steps to calculate the mean

Table 2
Case 2: Block-banded Structure. Means with their corresponding standard deviations in parentheses of the errors, and the frequency of estimated bandwidth in estimating coefficient matrix.

Error standard deviation = 1														
		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm	Est. bandwidth					L_2 norm	L_2 norm
		0	1	2	3	4	$\|\hat{A}-A\|_2$	0	1	2	3	4	$\|\hat{A}-A\|_2$	$\|\hat{A}-A\|_2$
100	1	2	497	1	0	0	0.33(0.03)	0	345	145	8	2	0.50(0.11)	0.87(0.05)
100	2	44	156	300	0	0	0.49(0.05)	6	67	366	60	1	0.62(0.14)	0.94(0.08)
100	3	135	221	92	52	0	0.62(0.14)	21	153	118	201	7	0.65(0.16)	0.95(0.11)
100	4	183	289	25	3	0	0.67(0.18)	33	205	130	78	54	0.68(0.18)	0.97(0.12)
196	1	0	495	5	0	0	0.34(0.03)	0	273	208	18	1	0.51(0.11)	0.99(0.07)
196	2	21	169	310	0	0	0.50(0.04)	0	47	366	82	5	0.63(0.14)	1.05(0.09)
196	3	92	240	109	59	0	0.63(0.14)	4	121	125	237	13	0.67(0.16)	1.07(0.11)
196	4	152	308	38	2	0	0.68(0.19)	12	186	155	91	56	0.69(0.18)	1.08(0.12)
400	1	0	500	0	0	0	0.36(0.02)	0	167	310	22	1	0.54(0.11)	1.07(0.05)
400	2	10	175	315	0	0	0.51(0.04)	0	25	351	116	8	0.63(0.14)	1.10(0.08)
400	3	43	272	122	63	0	0.65(0.13)	0	74	134	270	22	0.68(0.16)	1.12(0.10)
400	4	89	369	40	2	0	0.68(0.18)	1	153	170	96	80	0.69(0.17)	1.11(0.10)
784	1	0	498	2	0	0	0.37(0.03)	0	84	381	34	1	0.54(0.10)	1.08(0.05)
784	2	1	180	319	0	0	0.52(0.04)	0	12	297	174	17	0.64(0.14)	1.11(0.08)
784	3	21	284	120	75	0	0.65(0.12)	0	37	140	270	53	0.69(0.15)	1.13(0.13)
784	4	53	386	57	4	0	0.69(0.17)	0	89	179	126	106	0.70(0.17)	1.12(0.09)
Error standard deviation = 0.01														
		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm	Est. bandwidth					L_2 norm	L_2 norm
		0	1	2	3	4	$\|\hat{A}-A\|_2$	0	1	2	3	4	$\|\hat{A}-A\|_2$	$\|\hat{A}-A\|_2$
100	1	0	498	2	0	0	0.26(0.04)	0	0	12	70	418	0.75(0.23)	1.06(0.10)
100	2	0	20	480	0	0	0.45(0.04)	0	0	11	66	423	0.83(0.27)	1.11(0.11)
100	3	0	14	211	275	0	0.62(0.05)	0	0	3	54	443	0.86(0.28)	1.13(0.13)
100	4	0	23	223	186	68	0.76(0.12)	0	0	2	25	473	0.88(0.28)	1.14(0.13)
196	1	0	495	5	0	0	0.28(0.04)	0	0	11	63	426	0.77(0.25)	1.01(0.08)
196	2	0	11	489	0	0	0.47(0.03)	0	0	1	36	463	0.88(0.28)	1.04(0.06)
196	3	0	10	210	280	0	0.64(0.05)	0	0	0	34	466	0.89(0.29)	1.06(0.07)
196	4	0	8	238	187	67	0.78(0.12)	0	0	1	13	486	0.91(0.29)	1.06(0.07)
400	1	0	498	2	0	0	0.29(0.04)	0	0	0	42	458	0.82(0.28)	1.28(0.08)
400	2	0	4	495	0	1	0.49(0.04)	0	0	0	13	487	0.89(0.30)	1.31(0.06)
400	3	0	4	202	294	0	0.67(0.05)	0	0	0	11	489	0.93(0.30)	1.32(0.06)
400	4	0	5	217	199	79	0.80(0.12)	0	0	0	2	498	0.96(0.32)	1.33(0.06)
784	1	0	498	2	0	0	0.31(0.03)	0	0	0	31	469	0.82(0.28)	1.64(0.08)
784	2	0	1	498	0	1	0.50(0.04)	0	0	0	3	497	0.92(0.31)	1.66(0.08)
784	3	0	0	179	321	0	0.68(0.05)	0	0	0	1	499	0.98(0.31)	1.66(0.08)
784	4	0	2	195	223	80	0.82(0.12)	0	0	0	0	500	0.99(0.33)	1.71(0.092)

squared prediction errors for the methods in comparison. The results show similar merits of the proposed NVAR method as shown in the estimated coefficient matrix $\hat{A}$ in comparison with the BVAR and LASSO methods, thus the results are omitted here.

Finally, we would like to make two remarks on the performance of the proposed method with respect to model misspecification and selection accuracy

Remark 1. Model Misspecification In the case of minor to moderate model misspecification, it will be interesting to conduct a full theoretical exploration of what guarantees would be possible to make, which is beyond the scope of this work.

However, simulations suggest that the proposed model can be robust to certain kinds of model misspecification. For example, consider the case that a 1-D banded structure is the true data structure but the model is misspecified as having a 2-D structure. The proposed algorithm, because it is built to subsume 1-D structures, works as well as Banded VAR as shown in Table 1. Similarly, a reverse model misspecification of the true 2-D data structure as a 1-D banded VAR model does not directly impact the algorithm's performance as it is designed to search through both 1-D and 2-D data structures.

Remark 2. Selection Accuracy The results presented in the tables can also indicate how the algorithm performs in terms of the tradeoff between precision and accuracy. We present the analysis results of false positive and false negative rates in terms of the estimated d_0, which could reflect the corresponding rates in terms of selected coefficients. For instance, when a block-banded structure is considered, under-estimation of d_0 will have a quadratic effect on the number of wrongly omitted coefficients. Whereas in a banded structure, such a relationship will be linear. As the tables show, the false negative rate (FNR) is typically larger than the false positive rate (FPR) in terms of d_0 estimation. For instance, in Table 3 under the error standard deviation = 0.01, it is seen that the FNR for d_0 increases from 3.2% for $d_0 = 2, p = 100$ to 91.6%, for $d_0 = 4, p = 100$. The FPR decreases from 6% to 0% in the

Table 3
Case 3: 2-D Spatial Structure. Means with their corresponding standard deviations in parentheses of the errors, and the frequency of estimated bandwidth in estimating coefficient matrix.

Error standard deviation = 1														
		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm	Est. bandwidth					L_2 norm	L_2 norm
		0	1	2	3	4	$\|\hat{A}-A\|_2$	0	1	2	3	4	$\|\hat{A}-A\|_2$	$\|\hat{A}-A\|_2$
100	1	0	494	6	0	0	0.36(0.04)	0	33	93	107	267	0.43(0.05)	0.87(0.06)
100	2	0	139	361	0	0	0.54(0.05)	6	85	89	81	239	0.60(0.12)	0.93(0.08)
100	3	0	283	174	43	0	0.67(0.13)	24	170	93	107	106	0.66(0.17)	0.96(0.11)
100	4	0	394	101	5	0	0.71(0.17)	47	240	127	61	25	0.68(0.18)	0.98(0.13)
196	1	0	497	3	0	0	0.38(0.04)	0	18	87	90	305	0.45(0.05)	0.99(0.07)
196	2	0	119	381	0	0	0.55(0.05)	2	61	104	72	261	0.60(0.12)	1.03(0.09)
196	3	0	251	210	39	0	0.68(0.14)	2	141	136	95	126	0.67(0.17)	1.05(0.11)
196	4	0	329	168	3	0	0.73(0.17)	10	215	157	87	31	0.69(0.18)	1.08(0.12)
400	1	0	473	27	0	0	0.40(0.05)	0	6	76	115	303	0.46(0.04)	1.05(0.06)
400	2	0	94	406	0	0	0.57(0.05)	0	26	93	84	297	0.62(0.12)	1.09(0.08)
400	3	0	189	245	66	0	0.71(0.13)	0	93	135	91	181	0.69(0.17)	1.11(0.10)
400	4	0	281	215	4	0	0.74(0.16)	2	170	166	103	59	0.70(0.17)	1.11(0.10)
784	1	0	469	31	0	0	0.43(0.06)	0	0	63	107	330	0.48(0.05)	1.06(0.05)
784	2	0	68	432	0	0	0.60(0.05)	0	15	67	79	339	0.64(0.12)	1.10(0.07)
784	3	0	155	269	76	0	0.72(0.12)	0	58	132	101	209	0.69(0.16)	1.15(0.98)
784	4	0	217	278	5	0	0.77(0.17)	0	102	181	111	106	0.71(0.18)	1.12(0.09)
Error standard deviation = 0.01														
		NVAR						BVAR						LASSO
p	d_0	Est. bandwidth					L_2 norm	Est. bandwidth					L_2 norm	L_2 norm
		0	1	2	3	4	$\|\hat{A}-A\|_2$	0	1	2	3	4	$\|\hat{A}-A\|_2$	$\|\hat{A}-A\|_2$
100	1	0	497	1	0	2	0.31(0.06)	0	0	7	70	423	0.44(0.08)	1.07(0.11)
100	2	0	16	484	0	0	0.51(0.04)	0	0	1	33	466	0.71(0.21)	1.11(0.12)
100	3	0	7	264	229	0	0.69(0.07)	0	0	1	23	476	0.84(0.26)	1.14(0.13)
100	4	0	11	283	172	34	0.83(0.15)	0	0	3	40	457	0.89(0.30)	1.14(0.13)
196	1	0	495	5	0	0	0.33(0.04)	0	0	3	84	413	0.45(0.08)	1.01(0.07)
196	2	0	5	493	1	1	0.53(0.05)	0	0	1	21	478	0.73(0.22)	1.04(0.07)
196	3	0	2	239	258	1	0.72(0.07)	0	0	0	14	486	0.87(0.26)	1.06(0.07)
196	4	0	1	267	196	36	0.84(0.15)	0	0	0	17	483	0.91(0.29)	1.06(0.07)
400	1	0	488	12	0	0	0.35(0.05)	0	0	0	44	456	0.50(0.10)	1.30(0.08)
400	2	0	2	496	1	1	0.55(0.05)	0	0	0	9	491	0.77(0.23)	1.32(0.07)
400	3	0	0	224	274	2	0.73(0.07)	0	0	0	5	495	0.89(0.27)	1.32(0.07)
400	4	0	0	230	218	52	0.86(0.15)	0	0	0	7	493	0.96(0.34)	1.32(0.07)
784	1	0	461	33	0	6	0.39(0.10)	0	0	0	25	475	0.52(0.10)	1.65(0.10)
784	2	0	0	494	3	3	0.58(0.06)	0	0	0	3	497	0.77(0.22)	1.65(0.11)
784	3	0	0	199	296	5	0.77(0.08)	0	0	0	1	499	0.92(0.28)	1.67(0.09)
784	4	0	0	217	236	47	0.88(0.15)	0	0	0	1	499	0.97(0.33)	1.66(0.09)

same span for the NVAR method. Similar observations can be made throughout the other tables and suggests that the algorithm reduces false positives consistently.

5. Case study of stream nitrogen data

Excess nitrogen is one of the most important causes of impairment for rivers and streams (see EPA (2000)). Previous studies also showed that nitrogen and phosphorus loads are important driving factors of Harmful algal blooms (HABs) (see Paerl et al. (2001)). For urban and urbanizing watersheds, less developed agricultural and low-density residential (suburban/exurban) areas contribute most in terms of annual loads of nitrogen, mainly through sewage and fertilizers (see Shields et al. (2008)). Thus, it is of great importance to understand the complication and prediction of nitrogen of multiple streams.

In this section, we apply the proposed NVAR method to the case study of surface water quality data, focusing on observed total nitrogen (TN) loads from the U.S. Geological Survey (USGS) stream gauges. The daily TN data (1990 to current) for the states of Virginia and West Virginia were downloaded. The data for this study are openly available in National Water Information System (NWIS) at https://waterdata.usgs.gov/nwis.

We conducted an initial analysis of data availability and temporal consistency. First we transfer the original data to a monthly data by using each month's maximum value as some months have more than one measurement. Next we select a subset of data, which has p time series, and does not contain any missing values over some n consecutive months. Among these selected possible datasets, we choose the one with the largest sample size, pn. As a result, a total of $p = 14$ monitoring sites were selected as the dataset for analysis, and the length of time series is $n = 73$. Fig. 5 reports the locations of the sites of the 14 streams of study.

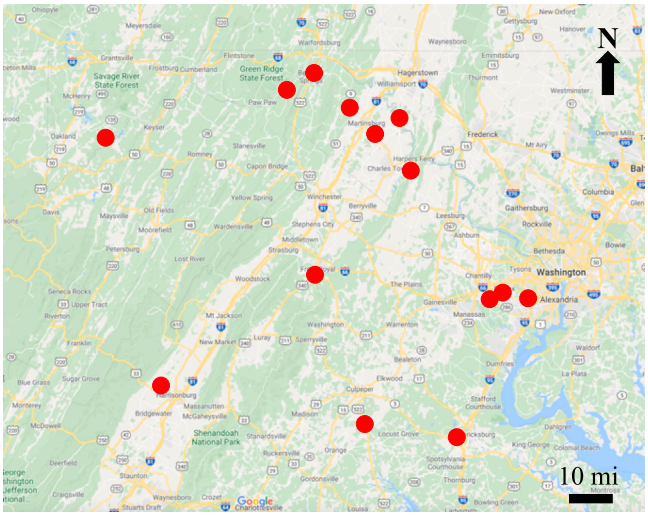

Fig. 5. Location map of the streams in our study.

Table 4
Mean square prediction error (MSPE) of NVAR, BVAR, LASSO for the stream data. A upper bound is set for the number of selected variables, which is $p/2$.

p	n	NVAR			BVAR			LASSO	
		Est. band width	MSPE	Comp. time	Est. band width	MSPE	Comp. time	MSPE	Comp. time
14	73	7	0.746	0.092	3	1.026	0.053	4.787	0.411

Table 5
Mean square prediction error (MSPE) of the BVAR with different index for the stream data. A upper bound is set for the number of selected variables, which is $p/2$.

p	n	BVAR: longitude index		BVAR: latitude index		BVAR: PCA 1 index		BVAR: PCA 2 index	
		Est. band width	MSPE	Est. band width	MSPE	Est. band width	MSPE	Est. band width	MSPE
14	73	3	1.026	3	0.813	3	1.069	3	0.941

We analyze this nitrogen dataset using the proposed NVAR method in comparison with the BVAR method and the LASSO method. To evaluate the performance of the methods in comparison, we partition the data into the training data and test data. For each time series, the beginning 80% data are used as training data and the later 20% data are used as test data. The one-step ahead prediction is used to calculate the mean squared prediction errors for the methods in comparison. For the NVAR method, the definition of bandwidth k_{NVAR} is a little different from the random structure. The k_{NVAR} equals the number of selected neighbors that are closest to a stream of interest.

Table 4 reports the performance of the methods in comparison. It is seen that the proposed NVAR method performs better than the BVAR method in terms of mean squared prediction error, and both the NVAR method and the BVAR method have much lower values of mean squared prediction error in comparison with that the LASSO method.

Note that the order of streams needs to be specified for applying the BVAR method, and the streams are ordered by their longitude for the BVAR method in Table 4. It means that, when using the BVAR method for analyzing the nitrogen data, its performance depends on how to specify the order of streams. In contrast, the proposed NVAR method accommodates the natural distance measure to allow the analysis invariant on the order of the streams.

Table 5 shows the results of the BVAR method using different directions to order the streams. The PCA 1 index chooses the direction which explains the largest variety of the 2-D location (i.e., longitude and latitude). The PCA 2 index uses the direction which is perpendicular to the PCA 1 direction. As shown in the Table 5, the performance of the BVAR method varies among different directions. It is seen that the order of streams has a significant effect on the performance of the BVAR method. The prediction performance of the BVAR method is not as good as the proposed NVAR method. Note that the NVAR method incorporates 2-D information instead of 1-D, thus it can include adequate information and is more robust than the BVAR method.

6. Discussion

In this work, we have presented the neighborhood vector autoregression model, which utilizes the underlying distances among the time series based on the inherent setting of the problem. We have generalized the model assumption in Guo et al. (2016)

by extending the notion of "band" to the notion of a "neighborhood". The notion of neighborhood can be quite general under a distance or dissimilarity measure on where the data of multiple time series are collected. With the aid of the Bayesian information criterion to choose an appropriate neighborhood size, the proposed NVAR method uses least squares for parameter estimation, thus can outperform penalization-based algorithms (e.g. Lozano et al. (2009), Bolstad et al. (2011)) in terms of computing efficiency. We also investigate the theoretical properties of the proposed method under some regularity conditions. Our theoretical studies show that the optimum neighborhood distance selected by the NVAR method converges to the true neighborhood distance, and the estimated coefficient matrix converges to the true coefficient matrix. The simulation study and case study of stream nitrogen data show the NVAR method outperforms the BVAR method and the LASSO method, and the NVAR method is more robust than the BVAR method. In particular, it is seen that the proposed method can gain prediction accuracy by borrowing information from the neighborhood streams.

A potential limitation of the proposed NVAR method is the assumption that the dependency (e.g., spatial dependency) among the time series must exist, and it can be captured by a distance or dissimilarity matrix. When there is no such dependency among the time series or the distance matrix is unknown, then the NVAR method might not show advantages over other methods.

There are several directions for future research. First, we would like to study the NVAR method in terms of the estimation of auto-covariance matrix, and compare the performance with other methods in terms of estimation accuracy, efficiency, and theoretical convergence. Second, our current approach adopts the BIC to choose the size of the neighborhood, and then the parameters are estimated by the ordinary least squares. Alternatively, we can use the penalized least squares for parameter estimation. Under this situation, it will be interesting to investigate what will be appropriate penalty functions for the neighbor vector autoregression model. It will also be interesting to examine the theoretical properties on the estimation accuracy when penalized estimation is involved. Third, the proposed method could be conservative in terms of false positives and could be aggressive on minimizing the false positives preferentially. One possible explanation can be due to the use of the BIC in the setting where d_0 is already small. Alternatively, the use of penalized least squares or other techniques would be interesting in future work to achieve a tradeoff between sensitivity and specificity by creating a different threshold for bandwidth detection. Moreover, in a more general framework, the investigation on the effects of the precision–accuracy or sensitivity–specificity tradeoff would be of practical use, which can make a complete ROC curve for selecting the best bandwidth. Fourth, to address the limitation of relying on a distance matrix, one potential remedy is to combine covariance/precision matrix estimation with the NVAR method. It is reasonable to view the multivariate time series as a graph, then the conditional dependency can be viewed as the notion of "neighborhood". Thus, the neighborhood of time series can be identified with some sparse covariance/precision matrix estimation method, which is a substitute when the distance matrix is unknown.

Acknowledgments

The authors thank the editor, associate editor, and anonymous reviewers for their constructive comments for improving this manuscript. This work has been partially supported by NSF CISE Expeditions (CCF-1918770) and Virginia Tech Center for Coastal Studies.

Appendix A. Supplementary material

A.1. Proof of Theorem 1

Without loss of generality, we consider the NVAR(1) model with $\|A\|_1 \le \delta < 1$. Our goal is to prove that $\mathrm{pr}(\hat{d} = d_0) \to 1$, i.e., $\mathrm{pr}(\hat{d} \neq d_0) \to 0$. If $\hat{d} \neq d_0$, then either $\hat{d} > d_0$ or $\hat{d} < d_0$ holds. Hence it suffices to show that $\mathrm{pr}(\hat{d} < d_0) \to 0$ and $\mathrm{pr}(\hat{d} > d_0) \to 0$. Our proof follows the arguments in Guo. et al. (2016).

Consider the first case. Observe that $\mathrm{pr}(\hat{d} < d_0) \le \mathrm{pr}(\hat{d}_i < d_0)$ for some $i \in \{1, \dots, p\}$ and the event $(\hat{d}_i < d_0)$ imply $\{\min_{d<d_0} \mathrm{BIC}(d,i) < \mathrm{BIC}(d_0,i)\}$. To prove $\mathrm{pr}(\hat{d} \neq d_0) \to 0$, we only need to show that

$$\mathrm{pr}\{\min_{d<d_0} \mathrm{BIC}(d,i) < \mathrm{BIC}(d_0,i)\} \to 0$$

for some i. Suppose that we have shown that there exists a constant $\eta > 0$ and an event $\mathcal{A}_n$ such that $\mathrm{pr}(\mathcal{A}_n) \to 1$ as $n \to \infty$ and on the event $\mathcal{A}_n$,

$$\mathrm{RSS}(d,i) - \mathrm{RSS}(d_0,i) \ge \eta \mathrm{RSS}(d_0,i) \sum_{j \in \mathcal{N}_i^{d_0}} (a_{i,j}^2) \tag{7}$$

for sufficiently large n, where $a_{j,k}$ is the (j,k)-element of A_1. On the event $\mathcal{A}_n$ with large n, $\log\mathrm{RSS}(d,i) - \log\mathrm{RSS}(d_0,i) \ge \log\{1 + \eta \sum_{j \in \mathcal{N}_i^{d_0}} (a_{i,j}^2)\}$. Note that $\log(1+x) \ge \min(0.5x, \log 2)$ for any $x > 0$. consequently, with probability tending to one, $\log\mathrm{RSS}(d,i) - \log \mathrm{RSS}(d_0,i)$ can be further bound below by $\min(0.5\eta \sum_{j \in \mathcal{N}_i^{d_0}} a_{i,j}^2, \log 2)$. Condition 2 implies that for some $i^* \in \{1, \dots, p\}$, $0.5\eta \sum_{j \in \mathcal{N}_{i^*}^{d_0}} a_{i^*,j}^2 \gg C_n \tau_{i^*}^{d_0} n^{-1} \log(p \vee n)$ as $n \to \infty$, where $\tau_{i^*}^{d_0} = |\mathcal{N}_{i^*}^{d_0}|$. Hence, it follows that, with probability tending to 1,

$$\min_{d<d_0} \mathrm{BIC}(d,i^*) - \mathrm{BIC}(d_0,i^*) = \log\mathrm{RSS}(d,i^*) - \log\mathrm{RSS}(d_0,i^*) + C_n(\tau_{i^*}^{d} - \tau_{i^*}^{d_0}) n^{-1} \log(p \vee n)$$

$$> \min(0.5\eta \sum_{j\in\mathcal{N}_{i^*}^{d_0}} a_{i^*,j}^2, \log 2) - C_n \tau_{i^*}^{d_0} n^{-1}\log(p\vee n)$$
$$\gg 0.$$

where $p\vee n = \max(p,n)$, $\mathcal{N}_{i^*}^{d}$ is the length of non-zero elements in the i^*-th row of A_1 with d-neighborhood, $\mathcal{N}_{i^*}^{d_0}$ is the length of non-zero elements in the i^*-th row of A_1 with d_0-neighborhood. Hence, $\mathrm{pr}\{\min_{d<d_0} \mathrm{BIC}(d,i) < \mathrm{BIC}(d_0,i)\} \to 0$ and thus $\mathrm{pr}(\hat{d} < d_0) \to 0$.

Let us prove Eq. (7). For $d < d_0$, denote $H_{i,d} = X_{i,d}(X_{i,d}^T X_{i,d})^{-1} X_{i,d}^T$, $X_{i,d_0} = (S_{i,d}, X_{i,d})$ and $\beta_{i,d_0} = (b_i^T, \beta_{i,d}^T)^T$, where $X_{i,d}$, $\beta_{i,d}$, $S_{i,d}$ are defined as below

$$y_i(t) = \sum_{j\in\mathcal{N}_i^{d_0}} A_1(i,j) y_j(t-1) + e_i(t),$$
$$y_i = X_{i,d}\beta_{i,d} + e_i,$$
Let $\{y_j(t-1)\}|_{j\in\mathcal{N}_i^d}$ be a column vector,
$$\text{then } X_{i,d} = \left\{ \{y_j(n-1)\}|_{j\in\mathcal{N}_i^d}, \{y_j(n-2)\}|_{j\in\mathcal{N}_i^d}, \dots, \{y_j(1)\}|_{j\in\mathcal{N}_i^d} \right\}^T,$$
$$\beta_{i,d} = \left\{ A_1(i,j) \right\}\Big|_{j\in\mathcal{N}_i^d},$$
$$S_{i,d} = \left\{ \{y_j(n-1)\}|_{j\in\mathcal{N}_i^{d_0}\setminus\mathcal{N}_i^d}, \{y_j(n-2)\}|_{j\in\mathcal{N}_i^{d_0}\setminus\mathcal{N}_i^d}, \dots, \{y_j(1)\}|_{j\in\mathcal{N}_i^{d_0}\setminus\mathcal{N}_i^d} \right\}^T,$$
$$\{j\in\mathcal{N}_i^{d_0}\setminus\mathcal{N}_i^d\} = \{j\in\mathcal{N}_i^{d_0} \text{ and } j\notin\mathcal{N}_i^d\}.$$

Then $\mathrm{RSS}(d,i) = y_i^T(I_{n-1} - H_{i,d})y_i$, and by Lemma 1(ii) or Lemma 2(ii), we have

$$\mathrm{RSS}(d,i) - \mathrm{RSS}(d_0,i) = b_i^T S_{i,d}^T (I_{n-1} - H_{i,d}) S_{i,d} b_i + o_P(1).$$

From Lemma 1(ii) or Lemma 2(ii) and Lemma 3, there exists a small constant $\eta > 0$ such that, with probability tending to one,

$$\lambda_{min}\{S_{i,d}^T(I_{n-1} - H_{i,d})S_{i,d}\} > \eta(1+\eta)n\sigma_i^2,$$

and $\mathrm{RSS}(d_0,i) \le (1+\eta)n\sigma_i^2$. Therefore, Eq. (7) follows.

Now let us prove the second case, $\mathrm{pr}(\hat{d} > d_0) \to 0$. For $d > d_0$, set

$$X_{i,d} = (S_{i,d}, X_{i,d_0}),\ \beta_{i,d} = (0^T, \beta_{i,d_0}^T)^T,\ \text{and } \tilde{S}_{i,d} = (I_{n-1} - H_{i,d_0}S_{i,d}).$$

Let η be an arbitrary but fixed positive constant and define

$$\mathcal{B}_n = \left\{ \inf_{d_0\le d\le d_{max}} \inf_{1\le i\le p} \frac{\mathrm{RSS}_i(k)}{n\sigma_i^2} > (1-\eta) \right\},$$

$$\mathcal{C}_n = \bigcup_{1\le i\le p,\ d_0\le d\le d_{max}} \left\{ \lambda_{\min}^{-1}(n^{-1}\tilde{S}_{i,d}^T\tilde{S}_{i,d}) < \kappa_1^{-1}(1+\eta), \sup_{1\le j\le d-d_0} \left| (n^{-1}S_{i,d}^T S_{i,d})_{jj} \right| < \kappa_2(1+\eta) \right\}.$$

We first give an upper bound on $\mathrm{RSS}(d_0,i) - \mathrm{RSS}(d,i)$ for $d > d_0$. For each i, $\mathrm{RSS}(d,i)$ can be rewritten as

$$\mathrm{RSS}(d,i) = \inf_b \|y_i - X_{i,d}b\|^2 = \inf_{b_1,b_2} \|y_i - X_{i,d}b_1 - S_{i,d}b_2\|^2.$$

where b is the estimator for $\beta_{i,d}$. It can be verified that $\mathrm{RSS}(d_0,i) = \|(I_{n-1} - H_{i,d_0})y_i\|^2$ and $\mathrm{RSS}(d,i) = \mathrm{RSS}(d_0,i) - \|\tilde{S}_i^{(d)}\hat{b}_2\|^2$, where $\hat{b}_2 = \left(\tilde{S}_{i,d}^T\tilde{S}_{i,d}\right)^{-1}\tilde{S}_{i,d}^T e_i$, and e_i is the residual for ith time series. Then on the event $\mathcal{C}_n$ we have

$$\begin{aligned}\mathrm{RSS}(d_0,i) - \mathrm{RSS}(d,i) &= e_i^T\tilde{S}_{i,d}(\tilde{S}_{i,d}^T\tilde{S}_{i,d})^{-1}\tilde{S}_{i,d}^T e_i \\ &\le \kappa_1^{-1}(1+\eta)|\tau_i^d - \tau_i^{d_0}| \sup_{j,d\le p} |n^{-1/2}e_j^T(I_{n-1} - H_{i,d_0})x_{(d)}|^2.\end{aligned}$$

Define

$$\mathcal{D}_n = \left\{ \sup_{j,d\le p} \left| n^{-1/2}e_j^T(I_{n-1} - H_{i,d_0})x_{(d)} \right|^2 \sigma_i^{-2} < \frac{\kappa_1(1-\eta)}{1+\eta} C_n \log(p\vee n) \right\}.$$

On the set $\mathcal{B}_n\cap\mathcal{C}_n\cap\mathcal{D}_n$, for all d with $d_0 \le d \le d_{max}$,

$$\begin{aligned}\mathrm{RSS}(d_0,i) - \mathrm{RSS}(d,i) &< \sigma_i^2(1-\eta)|\tau_i^d - \tau_i^{d_0}|C_n\log(p\vee n) \\ &< \mathrm{RSS}(d,i)C_n|\tau_i^d - \tau_i^{d_0}|n^{-1}\log(p\vee n).\end{aligned}$$

Note that $\log(1+x) \le x$ for any $x > 0$. Hence, for all d with $d_0 \le d \le d_{max}$, on the set $\mathcal{B}_n\cap\mathcal{C}_n\cap\mathcal{D}_n$,

$$\begin{aligned}\mathrm{BIC}(d,i) - \mathrm{BIC}(d_0,i) &= \log \mathrm{RSS}(d,i) - \log \mathrm{RSS}(d_0,i) + C_n|d^m(i) - d_0^m(i)|n^{-1}\log(p\vee n) \\ &\ge -\log\left(1 + C_n|\tau_i^d - \tau_i^{d_0}|n^{-1}\log(p\vee n)\right)\end{aligned}$$

$$+ C_n|\tau_i^d - \tau_i^{d_0}|n^{-1}\log(p \vee n)$$

which indicates that over the set $\mathcal{B}_n \cap C_n \cap \mathcal{D}_n$, we have that $\hat{d} \le d_0$. To prove that $\mathrm{pr}(\hat{d} > d_0) \to 0$, it is sufficient to show that $\mathrm{pr}\{(\mathcal{B}_n \cap C_n \cap \mathcal{D}_n)^c\} \to 0$. In fact, it follows from Lemmas 1 and 3 or Lemma 2(i), that $\mathrm{pr}(\mathcal{B}_n^c) \to 0$ and $\mathrm{pr}(C_n^c) \to 0$. It remains to show that $\mathrm{pr}(\mathcal{D}_n^c) \to 0$. Let $\hat{\Sigma}_{i,d} = n^{-1}X_{i,d}^T X_{i,d}$, $\Sigma_{i,d} = n^{-1}E\left(X_{i,d}^T X_{i,d}\right)$, where $E(X)$ denotes the expectation of X. Set $\tilde{H}_{i,d} = n^{-1}X_{i,d}\Sigma_{i,d}^{-1}X_{i,d}^T$, and $\tilde{x}_{(d)} = (I_{n-1} - \tilde{H}_{i,d})x_{(d)}$. On the event $\mathcal{D}_n$, we obtain that

$$\begin{aligned}
\sup_{j,d\le p}|e_j^T(I_{n-1} - H_{i,d_0})x_{(d)}| &\le \sup_{j,d\le p}|e_j^T\tilde{x}_{(d)}| + \sup_{j,d\le p}|e_j^T(I_{n-1} - \tilde{H}_{i,d_0})x_{(d)}| \\
&\le \sup_{j,d\le p}|e_j^T\tilde{x}_{(d)}| \\
&\quad + \sup_{j,d\le p}\|e_j^T X_{i,d_0}\|_2\,\|\Sigma_{i,d_0}^{-1}\|_2\,\|\hat{\Sigma}_{i,d_0}^{-1}\|_2\,\|\hat{\Sigma}_{i,d_0} - \Sigma_{i,d_0}\|_2\,\|X_{i,d_0}^T x_{(d)}\|_2 \\
&\le \sup_{j,d\le p}|e_j^T\tilde{x}_{(d)}| + d_0\kappa_1^{-2}\kappa_2(1+\eta)^2\sup_{j,d\le p}\left|e_j^T x_{(d)}\right|\cdot\|\hat{\Sigma}_{i,d_0} - \Sigma_{i,d_0}\|_2,
\end{aligned}$$

where $\sup_{1\le d\le p}(n^{-1}x_{(d)}x_{(d)}^T) \le \kappa_2(1+\eta)$ is used in the above inequality. Hence, it follows from Lemmas 1 and 2, together with Condition 3, that $\mathrm{pr}(\mathcal{D}_n^c) \to 0$ as $n \to \infty$. Then $\mathrm{pr}(\hat{d} > d_0) \to 0$. □

A.2. Proof of Theorem 2

Without loss of generality, we consider the case of order 1, i.e. NVAR(1) only. It is shown in Theorem 1 that $\mathrm{pr}(\hat{d} = d_0) \to 1$ as $n \to \infty$. Thus it is sufficient to consider the set $\mathcal{A}_n = \{\hat{d} = d_0\}$. Over the set $\mathcal{A}_n$, for each i,

$$\hat{\beta}_i - \beta_i = (X_i^T X_i)^{-1}x_i^T e_i \tag{8}$$

For each i, the law of large numbers for the stationary process case yields that $n^{-1}X_i^T X_i$ converges to a positive matrix almost surely, and furthermore, with probability tending to one, $\lambda_{min}(n^{-1}X_i^T X_i)$ is bounded away from zero. As a matter of fact, if we define

$$\mathcal{B}_n = \bigcap_{1\le i\le p}\left\{\lambda_{min}(n^{-1}X_i^T X_i) > \kappa_1(1-\eta)\right\}$$

with a small constant $\eta \in (0,1)$, then it follows from by Lemmas 1 and 2 under different moment conditions that $P\{\mathcal{B}_n\} \to 1$ as $n \to \infty$. Hence, over the event $\mathcal{A}_n \cup \mathcal{B}_n$,

$$\begin{aligned}
\left\|\hat{\beta}_i - \beta_i\right\|_2^2 &\le \kappa_1^{-2}(1-\eta)^{-2}n^{-2}\|e_i^T X_i\|_2^2, \\
&= C_1 n^{-2}\|e_i^T X_i\|_2^2,
\end{aligned}$$

where $C_1 = \kappa_1^{-2}(1-\eta)^{-2} > 0$. It is not hard to see from Lemma 1(ii) or Lemma 2(ii) that, for all $1 \le i \le p$, $n^{-1}E\|X_i^T e_i\|_2^2 \le C_2$ with some constant $C_2 > 0$. Therefore, for a large positive constant C, we obtain that

$$\begin{aligned}
\mathrm{pr}\left(\left\|\hat{A}_1 - A_1\right\|_F^2 > Cn^{-1}p\right) &= \mathrm{pr}\left(\left\|\hat{A}_1 - A_1\right\|_F^2 > Cn^{-1}p, \mathcal{A}_n \cup \mathcal{B}_n\right) \\
&\quad + \mathrm{pr}\left(\left\|\hat{A}_1 - A_1\right\|_F^2 > Cn^{-1}p, (\mathcal{A}_n \cup \mathcal{B}_n)^c\right) \\
&\le (Cp)^{-1}n(C_1 n^{-2})E\left(\sum_{i=1}^p\|X_i^T e_i\|_2^2\right) \\
&\quad + \mathrm{pr}\left(\left\|\hat{A}_1 - A_1\right\|_F^2 > Cn^{-1}p, (\mathcal{A}_n \cup \mathcal{B}_n)^c\right) \\
&\le C_1 C_2 C^{-1} + o(1)
\end{aligned}$$

For a sufficiently large C, we have $\mathrm{pr}\left(\|\hat{A}_1 - A_1\|_F^2 > Cn^{-1}p\right) \to 0$. Thus the convergence rate of $\|\hat{A}_1 - A_1\|_F$ is established.

Now Let us derive the convergence rate of $\|\hat{A}_1 - A_1\|_2$. For any matrix B, $\|B\|_2^2 \le \|B\|_1\|B\|_\infty$. Hence, on the event $\mathcal{A}_n$,

$$\begin{aligned}
\|\hat{A}_1 - A_1\|_2 &\le \sqrt{\|\hat{A}_1 - A_1\|_1}\sqrt{\|\hat{A}_1 - A_1\|_\infty} \\
&\le \tau_i^{d_0}\sup_{i,j\le p}|\hat{\beta}_{ij} - \beta_{ij}|,
\end{aligned}$$

where $\hat{\beta}_{ij}$ and β_{ij} are the jth element of $\hat{\beta}_i$ and β_i, respectively. Observe from (3) that

$$\sup_{i,j\le p}|\hat{\beta}_{ij} - \beta_{ij}| = \kappa_1^{-1}(1-\eta)^{-1}\tau_i^{d_0}n^{-1}\left(\sup_{i,j\le p}|e_i^T x_{(j)}|\right), i = 1,\ldots,p.$$

Hence, using Lemma 1(ii) or Lemma 2(ii), we have

$$\sup_{i,j\le p}|\hat{\beta}_{ij} - \beta_{ij}| = O_P\left\{\left(n^{-1}\log p\right)^{1/2}\right\},$$

which shows that

$$\|\hat{A}_1 - A_1\|_2 = O_P\left\{ \left(n^{-1}\log p\right)^{1/2} \right\}.$$

Then the proof is done. □

Appendix B. Technical lemmas

The proof of Lemmas 1–3 can be found in Guo. et al. (2016), which are corresponding to Lemma 5 to 7. In addition, the regularity conditions should be replaced with our regularity conditions, and the lemmas still hold.

Lemma 1. *Suppose that Conditions (1)-(3) and 4(i) hold. (i) For $j, d = 1, \ldots, p$, there exist positive constants C_1, C_2, and C_3 free of (j, d, n, p) such that*

$$pr\left(\left| \hat{\Sigma}_{jd} - \Sigma_{jd} \right| > x \right) \leq \frac{C_1 n}{(nx)^q} + C_2 exp(-C_3 n x^2)$$

holds for $x > 0$; consequently, this leads to the following uniform convergence rate:

$$\sup_{1\leq j,d\leq p} \left| \hat{\Sigma}_{jd} - \Sigma_{jd} | \right| = O_P\left\{ (n^{-1} log p)^{1/2} \right\}.$$

(ii) For $j, d = 1, \ldots, p$, there exist positive constants C_1, C_2, and C_3 free of (j, d, n, p) such that

$$pr\left(\left| e_j^T x_{(k)} \right| > x \right) \leq \frac{C_1 n}{x^{2q}} + C_2 exp(-C_3 x^2)$$

holds for $x > 0$; in particular, we have:

$$\sup_{1\leq j,d\leq p} \left| e_j^T x_{(k)} \right| = O_P\left\{ (n log p)^{1/2} \right\}.$$

Lemma 2. *Suppose that Conditions (1)-(3) and 4(ii) hold. (i)*

$$\sup_{1\leq j,d\leq p} \left| \hat{\Sigma}_{jd} - \Sigma_{jd} | \right| = O_P\left\{ (n^{-1} log p)^{1/2} \right\}.$$

(ii)

$$\sup_{1\leq j,d\leq p} \left| e_j^T x_{(k)} \right| = O_P\left\{ (n log p)^{1/2} \right\}.$$

Lemma 3. *Suppose that Conditions (1)-(3) and 4(i) or 4(ii) hold. Then for each finite d with $d \geq d_0$,*

$$\sup_{1\leq i\leq p} \left| \frac{RSS(d, i)}{n\sigma_i^2} - 1 \right| = O_P\left\{ (n^{-1} log p)^{1/2} \right\}.$$

as $n \to \infty$, where $RSS(d, i)$ is defined in the main article and σ_i^2 is the (i, i)-th element of Σ_e.

References

Arnold, A., Liu, Y., Abe, N., 2007. Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 66–75.

Basu, S., Shojaie, A., Michailidis, G., 2015. Network granger causality with inherent grouping structure. J. Mach. Learn. Res. 16 (1), 417–453.

Bernanke, B.S., Boivin, J., Eliasz, P., 2004. Measuring the Effects of Monetary Policy: a Factor-Augmented Vector Autoregressive (FAVAR) Approach. Tech. Rep., National Bureau of Economic Research.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 192–225.

Bolstad, A., Van Veen, B.D., Robert, N., 2011. Causal network inference via group sparse regularization. IEEE Trans. Signal Process. 59, 2628–2641.

Cavalcante, L., Bessa, R.J., Reis, M., Browell, J., 2017. LASSO vector autoregression structures for very short-term wind power forecasting. J. Comput. Graph. Statist. 20, 657–675.

Cressie, N., Wikle, C.K., 2015. Statistics for Spatio-Temporal Data. John Wiley & Sons.

Davis, R.A., Zang, P., Zheng, T., 2016. Sparse vector autoregressive modeling. J. Comput. Graph. Statist. 25 (4), 1077–1096.

EPA, 2000. National Water Quality Inventory: 1998 Report to Congress. US Environmental Protection Agency (EPA), Office of Water. 2000c. EPA 841-R-00-001. Washington, DC.

Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C., Ferreira, C.E., 2007. Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Syst. Biology 1 (1), 39.

Ghahramani, M., Qiao, Y., Zhou, M., Hagan, A.O., Sweeney, J., 2020. AI-based modeling and data-driven evaluation for smart manufacturing processes. IEEE/CAA J. Autom. Sin. 7 (4), 1026–1037.

Guo, S., Wang, Y., Yao, Q., 2016. High dimensional and banded vector autoregressions. Biometrika 103, 889–903.

Han, F., Lu, H., Liu, H., 2015. A direct estimation of high dimensional stationary vector autoregressions. J. Mach. Learn. Res. 16 (1), 3115–3150.

Haufe, S., Müller, K.-R., Nolte, G., Krämer, N., 2010. Sparse causal discovery in multivariate time series. In: Causality: Objectives and Assessment. pp. 97–106.

Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. J. Amer. Statist. Assoc. 97 (460), 1090–1098.

Hsu, N.J., Hung, H.L., Chang, Y.M., 2008. Subset selection for vector autoregressive processes using lasso. Comput. Statist. Data Anal. 52 (7), 3645–3657.

Hsu, C.Y., Liu, W.C., 2020. Multiple time-series convolutional neural network for fault detection and diagnosis and empirical study in semiconductor manufacturing. J. Intell. Manuf. 1–14.

Jiang, X., Hu, X., Xu, W., Park, E.K., 2015. Predicting microbial interactions using vector autoregressive model with graph regularization. IEEE/ ACM Trans. Comput. Biology Bioinform. 12 (2), 254–261.

Kock, A.B., Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. J. Econometrics 186, 325–344.

Lozano, A.C., Abe, N., Liu, Y., Rosset, S., 2009. Grouped graphical granger modeling for gene expression regulatory networks discovery. Bioinformatics 25 (12), i110–i118.

Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C: Emerg. Technol. 54, 187–197.

Nicholson, W.B., Matteson, D.S., Bien, J., 2017. VARX-L: Structured regularization for large vector autoregressions with exogenous variables. Int. J. Forecast. 33, 627–651.

Opgen-Rhein, R., Strimmer, K., 2007. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC Bioinformatics 8.

Paerl, H.W., Fulton, R.S., Moisander, P.H., Dyble, J., 2001. Harmful freshwater algal blooms, with an emphasis on cyanobacteria. Sci. World J. (1), 76–113.

Qiu, H., Xu, S., Han, F., Liu, H., Caffo, B., 2015. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37. pp. 1843–1851.

Ren, Y., Zhang, X., 2010. Subset selection for vector autoregressive processes via adaptive lasso. Statist. Probab. Lett. 80, 1705–1712.

Rubio-Ramirez, J.F., Waggoner, D.F., Zha, T., 2010. Structural vector autoregressions: Theory of identification and algorithms for inference. Rev. Econ. Stud. 77 (2), 665–696.

Shields, C.A., Band, L.E., Law, N., Groffman, P.M., Kaushal, S.S., Savvas, K., Fisher, G.T., Belt, K.T., 2008. Streamflow distribution of non-point source nitrogen export from urban?rural catchments in the Chesapeake Bay watershed. Water Resour. Res. 44 (9).

Shojaie, A., Michailidis, G., 2010. Discovering graphical granger causality using the truncating lasso penalty. Bioinformatics 26 (18), i517–i523.

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

Steed, C.A., Halsey, W., Dehoff, R., Yoder, S.L., Paquit, V., Powers, S., 2017. Falcon: Visual analysis of large, irregularly sampled, and multivariate time series data in additive manufacturing. Comput. Graph..

Todd, R.M., 1990. Improving economic forecasting with Bayesian vector autoregression. Model. Econ. Ser. 214–234.

Valdés-Sosa, P.A., Sánchez-Bornot, J.M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., Canales-Rodríguez, E., 2005. Estimating brain functional connectivity with sparse multivariate autoregression. Philos. Trans. R. Soc. B 360 (1457), 969–981.