Statistica Sinica **33** (2023), 1271-1294
doi:https://doi.org/10.5705/ss.202020.0056

REGULARIZED ESTIMATION IN HIGH-DIMENSIONAL VECTOR AUTO-REGRESSIVE MODELS USING SPATIO-TEMPORAL INFORMATION

Zhenzhong Wang[1], Abolfazl Safikhani[2], Zhengyuan Zhu[1] and David S. Matteson[3]

[1]*Iowa State University,* [2]*University of Florida and* [3]*Cornell University*

Abstract: The vector auto-regressive (VAR) model is commonly used to model multivariate time series, and there are many penalized methods to handle high dimensionality. However for spatio-temporal data, most of these methods do not consider the spatial and temporal structure of the data, which may lead to unreliable network detection and inaccurate forecasts. This paper proposes a data-driven weighted l_1 regularized approach for spatio-temporal VAR models. Extensive simulation studies compare the proposed method with five existing methods for high-dimensional VAR models, demonstrating advantages of our method over others in terms of parameter estimation, network detection, and out-of-sample forecasts. We also apply our method to a traffic data set to evaluate its performance in a real application. In addition, we explore the theoretical properties of the l_1 regularized estimation of the VAR model under a weakly sparse scenario, in which exact sparsity can be viewed as a special case. To the best of our knowledge, this is the first study to do so. For a general stationary VAR process, we derive the nonasymptotic upper bounds on the l_1 regularized estimation errors, provide the conditions for estimation consistency, and further simplify these conditions for a special VAR(1) case.

Key words and phrases: l_1 regularization, spatio-temporal structure, vector auto-regressive model, weak sparsity.

1. Introduction

The vector auto-regressive (VAR) model is a popular tool for simultaneously modeling and forecasting a number of time series, and has been widely applied in scientific fields such as econometrics (Sims (1980)), finance (Tsay (2015)), and ecology (Hampton et al. (2013)), among others. Recent developments in computing have made high-dimensional time series increasingly common. As the number of time series components increases, the number of parameters in the VAR model increases dramatically, leading to unreliable or even infeasible estimations. The

Corresponding author: Zhengyuan Zhu, Department of Statistics, Iowa State University, Ames, IA 50011-1090, USA. E-mail: zhuz@iastate.edu.

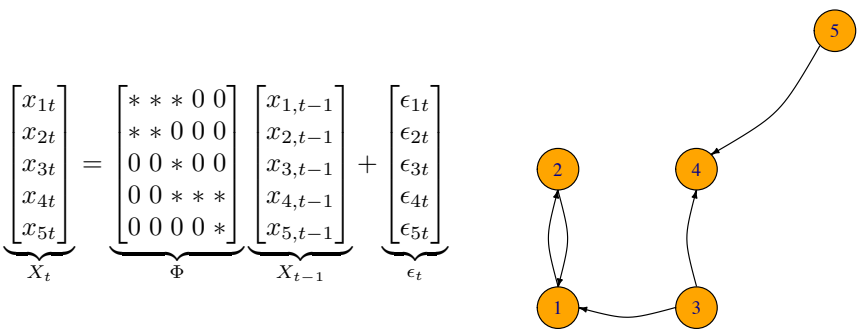

Figure 1. The left panel illustrates the sparsity (zero/nonzero) pattern for the transition matrix Φ in a VAR(1) process, with $*$ denoting nonzero entries. The right panel illustrates the network structure implied by this VAR(1) process. For example, Φ_{13} is nonzero, which indicates a directed connection from the third site to the first site.

usual way to handle high dimensionality is to impose sparsity or a low-rank structure on the transition matrices. Many estimation procedures have been proposed, including l_1 regularization (Basu and Michailidis (2015)), two-stage l_1 regularization (Davis, Zang and Zheng (2016)), the sparse seasonal VAR (Baek, Davis and Pipiras (2017)), the low-rank structured VAR (Basu, Li and Michailidis (2019)), hierarchical lag sparsity (Nicholson, Matteson and Bien (2017); Nicholson et al. (2020); Safikhani et al. (2018)), the banded VAR (Guo, Wang and Yao (2016)), and nonconcave penalization (Zhu (2020)). Other methods assume a factor structure on the time series data to reduce the dimensionality; see Lam and Yao (2012) and Tu, Yao and Rongmao (2020). Such high-dimensional techniques have become very popular in applications such as econometrics (Matteson and Tsay (2011)), genetics (Michailidis and d'Alché Buc (2013)), biology (Hu, Fortin and Ombao (2019)), and ecology (Reyes, Zhu and Aukema (2012)), among others.

For spatio-temporal data, each component of a multivariate time series contains the observations in one spatial location (site). Here, the parameters in the transition matrices can naturally capture the spatial and temporal interrelationships between the sites. At the same time, the zero–nonzero patterns of the transition matrices reflect the network structure in the data set. Figure 1 shows a simple example of a VAR(1) model on five sites. There exists a directed connection from site 3 to site 1, indicating that X_{1t} is dependent on $X_{3,t-1}$ and thus Φ_{13} is nonzero. Furthermore, $\Phi_{31} = 0$ means there is no directed connection from site 1 to site 3. Thus, for spatio-temporal data, the spatial structure and temporal information should be incorporated in the modeling procedure. If such information is ignored, high-dimensional methods may lead to inaccurate net-

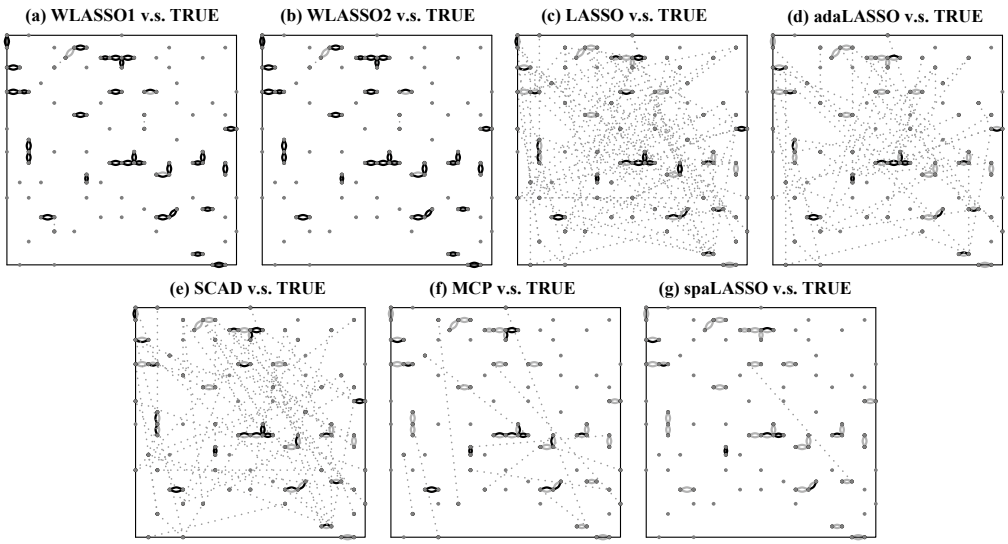

Figure 2. Comparison of the proposed methods (WLASSO1 and WLASSO2) with five existing methods (the LASSO, adaptive LASSO, SCAD, MCP and spaLASSO of Schweinberger, Babkin and Ensor (2017)) in terms of the network estimation of one simulated VAR(1) process from Section 3.3. Specifically, if the true value $\Phi_{ss'}$ and its estimator $\hat{\Phi}_{ss'}$ are nonzero, a black edge is drawn to connect site s and site s'. If $\Phi_{ss'}$ is not zero, but $\hat{\Phi}_{ss'}$ is zero, the edge is grey solid. If $\Phi_{ss'}$ is zero, but $\hat{\Phi}_{ss'}$ is not zero, the edge is grey dotted.

work estimations and unreasonable scientific conclusions. Figure 1 illustrates the drawback of ignoring the spatial and temporal information, based on a simulation study discussed in Section 3.3, in which the grey solid edges and grey dotted edges represent false negatives and false positives, respectively. By not considering the spatial and temporal information, the five existing methods studied in the simulation underestimate the true connections. In addition, the LASSO, adaptive LASSO (adaLASSO), and SCAD also overestimate wrong connections. In contrast, our proposed methods (WLASSO1 and WLASSO2) recover the network very well and significantly reduce false positives and false negatives.

In this paper, we propose a data-driven weighted l_1 regularized approach that constructs the penalty based on the spatial distances between sites and the temporal lags in the VAR model. We derive the nonasymptotic upper bounds of the estimation error, which hold with high probability, and show that these bounds are smaller than those of the LASSO (Remark 1(c) in Section 2.2 and Remark 2(c) in Section 2.3). The simulation studies compare the proposed approach with five existing methods, namely the LASSO (Basu and Michailidis

(2015)), SCAD and MCP (Zhu (2020)), spaLASSO (Schweinberger, Babkin and Ensor (2017)), and adaptive LASSO (Zou (2006); Wang, Li and Tsai (2007a)). The proposed approach shows a significant advantage in terms of model fitting, network detection, and forecasting performance (see Tables 1–5 and Figures 3–11 in the Supplemental Material). We also apply our method to a traffic network data set from the Des Moines, Iowa. Here, the network structure detected by the LASSO is not meaningful, whereas the proposed method provides a more reasonable estimated network and better forecasting results.

Several studies focus on high-dimensional VAR in a spatio-temporal setting. The most relevant work is that of Schweinberger, Babkin and Ensor (2017), who developed the spaLASSO, which incorporates the spatial structure in the VAR model estimation. Their approach assumes the spatial dependence exists only within a specific distance ρ, whereas ρ is either known or estimated in an initial step by the LASSO within sub-sampled sites. After ρ is specified, only parameters associated with distances smaller than the given ρ are estimated; others are fixed as zero. Assuming that the distance ρ is known is usually unrealistic in real data sets. By estimating ρ using an initial LASSO estimator, the inaccuracy of the initial estimator can produce an unreliable estimation of ρ, thus contaminating the final estimation of the model. As shown in Figure 2(c), the LASSO cannot identify the true network, and therefore, delivers an inaccurate estimation of ρ and an inaccurate estimation from the spaLASSO (Figure 2(g)). Furthermore, the assumption of no spatial dependence beyond the distance ρ is restrictive, and may not be true in some real cases, such as the more general weakly sparse scenario considered in this study. In addition, this approach does not incorporate the lag order of the temporal dependence. In contrast, the proposed method incorporates both the spatial and the temporal information in a smooth way, rather than truncating the parameters at a certain distance. Furthermore, the penalty weights are data driven so that no prior information is needed. The algorithm of the proposed method requires only a single step and is easy to implement using existing algorithms.

In a real application, spatial and temporal dependence may still exist, even for a long distance or temporal lag. In such cases, the transition matrices in the VAR model have many small nonzero elements, and thus are not sparse; hence this is the so-called "weakly sparse" scenario. The second goal of this study is to investigate the theoretical properties of the l_1 regularized estimation of the VAR model under the weakly sparse scenario. Weak sparsity is pursued mostly for independent data; see Negahban et al. (2009) and Raskutti, Wainwright and Yu (2011). However, no existing studies investigate the properties of l_1 regularized

estimators for high-dimensional VAR models under the weakly sparse scenario. Our contribution is to fill this gap. In addition, the "weak sparsity" defined here is more general than the l_r ball constraint commonly used in the literature; we discuss the advantages of our weak sparsity in detail in Section 2.3. We first derive the upper bounds of the l_1 regularized estimation error for a general stationary VAR process (Theorem 2) and provide the weak sparsity constraint (2.9) that guarantees the estimation consistency. Then, we further explore the weak sparsity constraint and simplify it for a special case of the VAR(1) process. Moreover, the results in Theorem 2 can also be used directly to derive the error bound under the l_r ball setting (Corollary 1), and we prove that our weak sparsity constraint is more relaxed than the l_r ball setting (Remark 3). Finally, we examine the proposed method under the weakly sparse scenario using simulation studies, demonstrating impressive advantages over other existing methods.

The remainder of the paper is structured as follows. Section 2 introduces the weighted l_1 regularized approach for the high-dimensional spatio-temporal VAR, as well as its theoretical properties. Section 3 presents the implementation of the proposed method and compares its performance with that of the LASSO, SCAD, MCP, adaptive LASSO (adaLASSO), and spaLASSO using simulation studies. In Section 4, we apply the proposed method to a traffic network data set. Section 5 concludes the paper.

Notation. Throughout this paper, we denote the cardinality of a set J by $|J|$, and use J^C to denote its complementary set. For a vector $\boldsymbol{v}$, we use $\boldsymbol{v}_J := (\boldsymbol{v}_i)_{i\in J}$ to denote the sub-vector with support J, and use $\|\boldsymbol{v}\|_q := (\sum_{i=1}^n |v_i|^q)^{1/q}$ to denote its l^q norm. For a matrix A, we use A_j to denote its jth column. $\mathrm{vec}(A)$, A', and A^H are its vectorization, transpose, and conjugate transpose, respectively. $A \circ B$ and $A \otimes B$ are the element-wise product and Kronecker product, respectively, of matrices A and B. $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ are the largest and smallest eigenvalues, respectively, of a symmetric or Hermitian matrix A. For a squared matrix A, $\|A\|_F$, $\rho(A)$, and $\|A\|_2$ are its Frobenius norm $\sqrt{tr(A^H A)}$, spectral radius $\max\{|\lambda_i| : \lambda_i \text{ are eigenvalues of } A\}$, and spectral norm $\sqrt{\Lambda_{\max}(A^H A)}$, respectively. We write $x \gtrsim y$ if there exists a positive constant c such that $x \geq cy$. If we have both $x \gtrsim y$ and $y \gtrsim x$, we use $x \asymp y$ to denote their relationship.

2. High-Dimensional Spatio-Temporal Vector Autoregression

Suppose x_{st} is the observation on site s at time t ($s = 1, \ldots, m$; $t = 1, \ldots, T$), and we assume $X_t = (x_{1t}, \ldots, x_{mt})'$ is generated by a pth-order VAR process:

$$X_t = \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma), \tag{2.1}$$

Here, $\Phi_1, \ldots, \Phi_p$ are $m \times m$ transition matrices encoding dependence across space and temporal lags. We use $\Phi_{l,ss'}$ to denote the ss'th entry of Φ_l, so that $\Phi_{l,ss'}$ represents the l-lagged influence of site s' on site s. We express this VAR(p) model in the following multivariate regression form:

$$\underbrace{\begin{bmatrix} X_T' \\ \vdots \\ X_{p+1}' \end{bmatrix}}_{\boldsymbol{Y}_{(T-p)\times m}} = \underbrace{\begin{bmatrix} X_{T-1}' & \cdots & X_{T-p}' \\ \vdots & \ddots & \vdots \\ X_p' & \cdots & X_1' \end{bmatrix}}_{\boldsymbol{X}_{(T-p)\times pm}} \underbrace{\begin{bmatrix} \Phi_1' \\ \vdots \\ \Phi_p' \end{bmatrix}}_{\boldsymbol{B}_{pm\times m}} + \underbrace{\begin{bmatrix} \varepsilon_T \\ \vdots \\ \varepsilon_{p+1} \end{bmatrix}}_{\boldsymbol{E}_{(T-p)\times m}}.$$

In the high-dimensional case, the LASSO can recover the sparseness of the transition matrices and reduce the forecasting error (Basu and Michailidis (2015)). However, the regular LASSO uses the same penalty for different $\Phi_{l,ss'}$ components, which may be inappropriate for spatio-temporal data. Instead, we propose the following weighted l_1 regularized LS, which penalizes $\Phi_{l,ss'}$ according to the spatial distance between site s and s', say $d_{ss'}$, as well as the temporal lag l:

$$\text{weighted } l_1\text{-LS:} \quad \widehat{\boldsymbol{B}} = \min_{\boldsymbol{B}} \frac{1}{N} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda_N \Omega(\boldsymbol{B}), \tag{2.2}$$

where $N = T - p$ and $\Omega(\boldsymbol{B}) = \sum_{l=1}^{p} \sum_{s,s'=1}^{m} w_{l,ss'} |\Phi_{l,ss'}|$, with $w_{l,ss'} \geq 0$ being the penalty weight for $\Phi_{l,ss'}$. Because $\Phi_{l,ss'}$ quantifies the dependence between site s and site s' across the temporal lag l, it is more likely to be zero if $d_{ss'}$ and l are large. Therefore, the weight $w_{l,ss'}$ is set to be an increasing function of the distance $d_{ss'}$ and the temporal lag l. Using this construction of the penalty weights, we impose a spatio-temporal structure on the data in which the conditional dependence between two sites across the temporal lag l (represented by $\Phi_{l,ss'}$) decays as the spatial distance $d_{ss'}$ and temporal lag l increase. There are several ways to define the weights, for example,

$$w_{l,ss'}^{(1)} = \exp\left(c_1 \frac{l\, d_{ss'}}{p\, d_{max}}\right) \quad \text{or} \quad w_{l,ss'}^{(2)} = \left(1 + \frac{l\, d_{ss'}}{p\, d_{max}}\right)^{c_2}, \tag{2.3}$$

where d_{max} is the maximum of $d_{ss'}$, and $c_1, c_2 > 0$ are universal constants determined using cross-validation. The inclusion of c_1 and c_2 ensures that the weights are data driven and adds flexibility to this method. Other weight functions can be defined as well based on the context of the data set under investigation. A special case is that $w_{l,ss'}$ is a function of $d_{ss'}$ only, such as $w_{l,ss'}^{(3)} = \exp(c_3 d_{ss'}/d_{max})$,

which means that the magnitudes of the parameters are influenced only by the distance. We examine the performances of various weight functions in our simulation studies and real data application.

Utilizing weighted penalty functions such as those above significantly improves the model performance, without sensitivity to the exact choice of weight functions. This is mainly because we include the data-informed constants c_i in all weight functions, which are selected via cross-validation. Including such data-driven constants optimizes the weights to some extent, and reduces the reliance of the model performance on the choice of the weight functions, demonstrating the robustness of the proposed method with respect to changes in the weight functions.

2.1. Model assumption

In the following, we provide the nonasymptotic bounds on the estimation error of the weighted l_1-LS estimation (2.2), and show that under certain conditions, the proposed estimator is consistent. We rewrite the VAR model as

$$\mathrm{vec}(\boldsymbol{Y}) = \mathrm{vec}(\boldsymbol{X}\boldsymbol{B}) + \mathrm{vec}(\boldsymbol{E}) = (I_m \otimes \boldsymbol{X})\mathrm{vec}(\boldsymbol{B}) + \mathrm{vec}(\boldsymbol{E}) := \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$ is an $mN \times 1$ vector, $\boldsymbol{Z} = I_m \otimes \boldsymbol{X}$ is an $mN \times q$ matrix, and $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B})$ is a $q \times 1$ vector with $q = m^2p$. The proposed estimation (2.2) can be expressed as the following M-estimation:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ -2\boldsymbol{\beta}'\hat{\boldsymbol{\gamma}} + \boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}\boldsymbol{\beta} + \lambda_N \Omega(\boldsymbol{\beta}) \right\}, \tag{2.4}$$

where $\hat{\boldsymbol{\gamma}} = (I_m \otimes \boldsymbol{X}')\boldsymbol{y}/N$ and $\hat{\boldsymbol{\Gamma}} = (I_m \otimes \boldsymbol{X}'\boldsymbol{X})/N$. Throughout this paper, we denote the true parameter as $\boldsymbol{\beta}^*$ and the corresponding true transition matrices as $\Phi_1^*, \ldots, \Phi_p^*$. We consider two scenarios: (1) $\boldsymbol{\beta}^*$ is exactly sparse; and (2) $\boldsymbol{\beta}^*$ is not exactly sparse, but can be well approximated by a sparse vector, which is called "weakly sparse." Both scenarios need the following assumption.

Assumption 1. *The VAR(p) process is stationary, that is, the roots of* $|I_m - \sum_{l=1}^{p} \Phi_l z| = 0$ *are lying outside the unit circle. In addition,* Σ *is positive definite.*

Assumption 1 is fundamental in high-dimensional time series analyses. Because the key to analyzing the M-estimation (2.4) is the dependence shown in $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Gamma}}$, this assumption guarantees that the spectral density of $\{X_t\}$ exists. Under such an assumption, Basu and Michailidis (2015) used the spectral density to construct a measure of dependence, and proved that $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Gamma}}$ satisfy two important conditions. More specifically, Propositions (4.2) and (4.3) in

Basu and Michailidis (2015) state that, under Assumption 1, there exist constants b_i, such that for $N \gtrsim \max\{\omega^2, 1\}(\log p + 2\log m)$, the restricted eigenvalue (RE) condition (2.5) and Derivation condition (2.6) hold with probability at least $1 - b_1 \exp(-b_2 N \min\{\omega^{-2}, 1\}) - b_3 \exp(-b_4(\log p + 2\log m))$:

$$\text{RE condition:} \quad \theta' \hat{\mathbf{\Gamma}} \theta \geq \alpha \|\theta\|_2^2 - \tau \|\theta\|_1^2, \quad \forall \theta \in R^q, \tag{2.5}$$

$$\text{Derivation condition:} \quad \|\hat{\boldsymbol{\gamma}} - \hat{\mathbf{\Gamma}} \boldsymbol{\beta}^*\|_\infty \leq \mathbb{Q} \sqrt{\frac{\log p + 2\log m}{N}}. \tag{2.6}$$

Here, ω, α, τ, and $\mathbb{Q}$ are determined by the transition matrices $\{\Phi_l^*\}_{l=1}^p$ and the covariance matrix of the innovation Σ. Specially, we first define

$$\mu_{\min}(\mathbf{\Phi}) = \min_{|z|=1} \Lambda_{\min}(\mathbf{\Phi}^H(z)\mathbf{\Phi}(z)), \quad \mu_{\max}(\mathbf{\Phi}) = \max_{|z|=1} \Lambda_{\max}(\mathbf{\Phi}^H(z)\mathbf{\Phi}(z)),$$

where $\mathbf{\Phi}(z) = I - \sum_{l=1}^p \Phi_l^* z^l$ ($z \in \mathbb{C}$) is the characteristic polynomial of the VAR process and $\mathbf{\Phi}^H(z)$ is its conjugate transpose. Furthermore, we set

$$\tilde{\Phi} = \begin{bmatrix} \Phi_1 & \cdots & \Phi_{p-1} & \Phi_p \\ I_m & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & I_m & \mathbf{0} \end{bmatrix}, \qquad \begin{aligned} \tilde{\mathbf{\Phi}}(z) &= I_{pm} - \tilde{\Phi} z \ (z \in \mathbb{C}), \\ \mu_{\min}(\tilde{\mathbf{\Phi}}) &= \min_{|z|=1} \Lambda_{\min}(\tilde{\mathbf{\Phi}}^H(z)\tilde{\mathbf{\Phi}}(z)). \end{aligned}$$

Then, ω, α, τ, and $\mathbb{Q}$ are defined as follows:

$$\begin{aligned} \omega &= a_1 \frac{\Lambda_{\max}(\Sigma)/\mu_{\min}(\tilde{\mathbf{\Phi}})}{\Lambda_{\min}(\Sigma)/\mu_{\max}(\mathbf{\Phi})}, \quad \alpha = \frac{\Lambda_{\min}(\Sigma)}{2\mu_{\max}(\mathbf{\Phi})}, \quad \tau = \alpha \max\{\omega^2, 1\} \frac{\log p + \log m}{N} \\ \mathbb{Q} &= a_2 \left[\Lambda_{\max}(\Sigma) + \frac{\Lambda_{\max}(\Sigma)}{\mu_{\min}(\mathbf{\Phi})} + \frac{\Lambda_{\max}(\Sigma)\mu_{\max}(\mathbf{\Phi})}{\mu_{\min}(\mathbf{\Phi})} \right], \end{aligned} \tag{2.7}$$

where a_1 and a_2 are positive constants. Refer to Basu and Michailidis (2015) for more detail. The RE condition (2.5) and Derivation condition (2.6) are the key to deriving the convergence rate of the M-estimation (2.4).

2.2. Convergence rate under exact sparsity

In this section, we assume the true parameter $\boldsymbol{\beta}^*$ has many zero entries, and we set its support to be $J = \{(l, ss') : \Phi_{l,ss'}^* \neq 0\}$, with $|J| = k$. In addition, we need the following constraint for the penalty weights.

Assumption 2. *$w_{l,ss'} > 0$, for all $(l, ss') \in J^C$.*

This assumption states that the parameters with true values equal to zero should have nonzero penalties. This assumption can be guaranteed by setting all penalty weights to be positive. In addition, any choice of $(\lambda_N, \{w_{l,ss'}\})$ is equivalent to $(\tilde{\lambda}_N, \{\tilde{w}_{l,ss'}\})$, with $\tilde{\lambda}_N = a\lambda_N$ and $\tilde{w}_{l,ss'} = w_{l,ss'}/a$, for any arbitrary positive number a. Without loss of generality, we can set $\min\{w_{l,ss'} : (l, ss') \in J^C\} = 1$. Furthermore we set $r_w = \max\{w_{l,ss'} : (l, ss') \in J\}$, which is indeed the ratio between the maximum weight of the nonzero parameters and the minimum weight of the zero parameters, that is, $r_w = \max\{w_{l,ss'} : (l, ss') \in J\}/\min\{w_{l,ss'} : (l, ss') \in J^C\}$. In the following theorem, this ratio is the key quantity for the proposed method to achieve smaller error bounds than those of the LASSO.

Theorem 1. *Consider the weighted l_1-LS estimator in* (2.4)*. If Assumptions* 1 *and* 2 *hold, there exist constants $b_i > 0$ not depending on the data or the model parameters, such that for any $N \gtrsim (1 + r_w)^2 \max\{\omega^2, 1\}k(\log p + 2\log m)$ and $\lambda_N \geq 4\mathbb{Q}\sqrt{(\log p + 2\log m)/N}$, with probability at least:*

$$1 - b_1 \exp(-b_2 N \min\{\omega^{-2}, 1\}) - b_3 \exp(-b_4(\log p + 2\log m)),$$

the estimation error $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is bounded as follows:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{1 + 2r_w}{\alpha}\sqrt{k}\lambda_N, \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{2 + 6r_w + 4r_w^2}{\alpha}k\lambda_N,$$
$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{(1 + 2r_w)^2}{2\alpha}k\lambda_N^2.$$

If we set $s_0 = \min\{|\beta_j^*| : j \in J\}$, the number of false zeros is bounded by

$$\left|\operatorname{supp}(\boldsymbol{\beta}^*)\backslash\operatorname{supp}(\hat{\boldsymbol{\beta}})\right| \leq \frac{2 + 6r_w + 4r_w^2}{s_0\alpha}k\lambda_N.$$

If we consider a threshold version $\tilde{\boldsymbol{\beta}} := \{\hat{\beta}_j I(|\hat{\beta}_j| > \lambda_N)\}$, with $I(.)$ being the indicator function, the number of false nonzeros in $\tilde{\boldsymbol{\beta}}$ is bounded by

$$\left|\operatorname{supp}(\tilde{\boldsymbol{\beta}})\backslash\operatorname{supp}(\boldsymbol{\beta}^*)\right| \leq (1 + 2r_w)^2\frac{k}{\alpha}.$$

Remark 1.

(a) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \sqrt{\sum_{l=1}^p \|\hat{\Phi}_l - \Phi_l^*\|_F^2}$ is the error of the transition matrices under the Frobenius norm, and $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \sum_{t=1}^T \|\sum_{l=1}^p (\hat{\Phi}_l - \Phi_l^*)X_{t-l}\|_2^2/T$ is the in-sample prediction error under the l_2 norm.

(b) If we set $r_w = 1$, which corresponds to the LASSO, we will get the following upper bounds, which are similar to those in Basu and Michailidis (2015):

$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 3\sqrt{k}\lambda_N/\alpha$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 12k\lambda_N/\alpha$, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq 9k\lambda_N^2/\alpha$, $|\text{supp}(\boldsymbol{\beta}^*)\backslash\text{supp}(\hat{\boldsymbol{\beta}})| \leq 12k\lambda_N/(s_0\alpha)$, $|\text{supp}(\tilde{\boldsymbol{\beta}})\backslash\text{supp}(\boldsymbol{\beta}^*)| \leq 9k/\alpha$.

(c) Compared with the LASSO ($r_w = 1$), if the weights $\{w_{l,ss'}\}$ are properly specified, the ratio r_w should be much smaller than one. In the ideal case when r_w is close to zero, our upper bounds for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, $|\text{supp}(\boldsymbol{\beta}^*)\backslash\text{supp}(\hat{\boldsymbol{\beta}})|$ and $|\text{supp}(\tilde{\boldsymbol{\beta}})\backslash\text{supp}(\boldsymbol{\beta}^*)|$ are nearly 1/3, 1/6, 1/9, 1/6, and 1/9 respectively, of that of the LASSO.

(d) Condition of Consistency: If λ_N is selected as $\lambda_N \asymp \mathbb{Q}\sqrt{(\log p + 2\log m)/N}$, the upper bounds of the estimation errors become

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{(1+2r_w)\mathbb{Q}}{\alpha}\sqrt{\frac{k(\log p + 2\log m)}{N}},$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{(2+6r_w+4r_w^2)\mathbb{Q}}{\alpha}\sqrt{\frac{k(\log p + 2\log m)}{N}},$$
$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{1}{2}(1+2r_w)^2\alpha^{-1}\mathbb{Q}^2\frac{k(\log p + 2\log m)}{N},$$

where α and $\mathbb{Q}$ are related to unknown parameters, as shown in equation (2.7). When $\mathbb{Q}/\alpha$ has a finite upper bound, we will have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{k(\log p + 2\log m)/N}$. In this case, the consistency of the proposed estimator requires only that N increases at a faster rate than $k(\log p + 2\log m)$.

2.3. Convergence rate under weak sparsity

In real applications, the conditional dependence quantified by $\Phi^*_{l,ss'}$ may not be zero, even for a large distance $d_{ss'}$ and/or lag l. For instance, $\Phi^*_{l,ss'} \neq 0$ may occur for a large distance $d_{ss'}$, especially when the sites are located on an irregular lattice. This example motivates us to consider a scenario called "weak sparsity", in which the true parameter vector $\boldsymbol{\beta}^*$ does not have many zeros (i.e. not exactly sparse), but can be well approximated by a sparse vector. Only a few studies touch on weak sparsity, and almost all of them focus on independent data (Negahban et al. (2009), Raskutti, Wainwright and Yu (2011)). An exception is the work of Sun et al. (2018), which focuses on estimating the spectral density matrix of high-dimensional time series. Moreover, they all define weak sparsity under the so-called "l_r ball" setting. Specifically, they assume the true parameter vector is within the l_r ball: $\mathbb{B}_r(R) := \{\boldsymbol{\beta}^* : \sum_{j=1}^q |\beta_j^*|^r \leq R\}$, where $r \in [0,1]$ is fixed. Under this setting, a constraint on the radius R is required to achieve the estimation consistency. For example, in independent data, the LASSO estimator

is consistent if R satisfies

$$l_r \text{ ball constriant:} \quad R = o\left(\left(\frac{N}{\log q}\right)^{1-r/2}\right), \tag{2.8}$$

where q is the number of parameters (Negahban et al. (2009), Raskutti, Wainwright and Yu (2011)). However, how "sparsifiable" $\boldsymbol{\beta}^*$ is depends on the relative magnitude of each element in $\boldsymbol{\beta}^*$, rather than its overall l_r length. Thus, the l_r ball setting does not clearly describe the "sparsifiability" of $\boldsymbol{\beta}^*$. A special case in which all β_j^* have the same magnitude could still fit in the l_r ball setting. However, in this case, $\boldsymbol{\beta}^*$ cannot be approximated by a sparse vector, and is not suitable for an l_1 regularized estimation. In general, the l_r ball setting may not be a reasonable way to relax the sparsity assumption.

Instead of using the l_r ball setting, we define "weak sparsity" from another perspective: most entries of $\boldsymbol{\beta}^*$ are small enough such that $\boldsymbol{\beta}^*$ can be well approximated by its hard thresholding version, say $\boldsymbol{\beta}_\eta^*$, the jth entry of which is $\beta_j^* I(|\beta_j^*| > \eta)$. For any given threshold η, we use $J_\eta = \{j : |\beta_j^*| > \eta\}$ to denote the support of $\boldsymbol{\beta}_\eta^*$. The formal definition of our proposed weak sparsity is as follows.

Definition 1 (Weak Sparsity Constraint)**.** If there exists an η such that the following two conditions hold:

$$|J_\eta| = o\left(\left(\frac{\alpha}{\mathbb{Q}}\right)^2 \frac{N}{\log p + 2\log m}\right) \quad \text{and}$$
$$\|\boldsymbol{\beta}_{J_\eta^C}^*\|_1 = o\left(\min\left\{\frac{\alpha}{\mathbb{Q}}, 1, \frac{1}{\omega}\right\}\sqrt{\frac{N}{\log p + 2\log m}}\right), \tag{2.9}$$

where $J_\eta^C := \{j : |\beta_j^*| \le \eta\}$, we say $\boldsymbol{\beta}^*$ satisfies the weak sparsity constraint.

This constraint means, with a proper choice of η, $\boldsymbol{\beta}_\eta^*$ is sparse and is a good approximation of $\boldsymbol{\beta}^*$ in the sense that its difference from $\boldsymbol{\beta}^*$, denoted as $\boldsymbol{\beta}_{J_\eta^C}^*$, is sufficiently small. In this way, our weak sparsity constraint quantifies how sparsifiable the true parameter vector $\boldsymbol{\beta}^*$ is, so that its l_1 regularized estimation remains consistent. In the following theorem, first without this constraint, we give a general result of the upper bound of the estimation error. Then, under this weak sparsity constraint, with a proper choice of λ_N, we show that the proposed estimator is consistent. Furthermore, we simplify the weak sparsity constraint in a special case of VAR(1) in Proposition 1. Finally, we directly apply Theorem 2 to derive the upper bound of the estimation error under the l_r ball setting,

and prove that our weak sparsity constraint (2.9) is more relaxed than the l_r ball constraint (Corollary 1). Also note that Theorem 2 and Corollary 1 also hold for the LASSO, because LASSO can be viewed as a special case of the proposed method in which all $w_{l,ss'}$s are the same. To state our theorem, we require the following notation: for any η, set $w_1(\eta) = \min\{w_{l,ss'} : (l, ss') \in J_\eta^C\}$, $w_2(\eta) = \max\{w_{l,ss'} : (l, ss') \in J_\eta\}$, and $r_w(\eta) = w_2(\eta)/w_1(\eta)$.

Assumption 3. *$w_{l,ss'} > 0$, for all (l, ss').*

Theorem 2. *Consider the weighted l_1-LS estimator in (2.4), and assume Assumptions 1 and 3 hold. Then, there exist constants $b_i > 0$, such that for any η, if $N \gtrsim (1 + r_w(\eta))^2 |J_\eta| \max\{\omega^2, 1\}(\log p + 2\log m)$ and $\lambda_N = \widetilde{\lambda}_N / w_1(\eta)$ with $\widetilde{\lambda}_N = 4\mathbb{Q}\sqrt{(\log p + 2\log m)/N}$, with probability at least*

$$1 - b_1 \exp(-b_2 N \min\{\omega^{-2}, 1\}) - b_3 \exp(-b_4(\log p + 2\log m)),$$

the estimation error $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is bounded as follows:

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &\le \frac{1 + 2r_w(\eta)}{\alpha} \sqrt{|J_\eta|} \widetilde{\lambda}_N + 2\sqrt{\frac{r_w(\eta)\widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1}{\alpha}} + \\
&\qquad \frac{4r_w(\eta)\max\{\omega, 1\}}{\mathbb{Q}} \widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1, \\
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &\le (2 + r_w(\eta))\sqrt{|J_\eta|} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + 4r_w(\eta)\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1, \\
(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) &\le \frac{1 + 2r_w(\eta)}{2} \sqrt{|J_\eta|} \widetilde{\lambda}_N \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + 2r_w(\eta)\widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1.
\end{aligned}$$

Second, if there exists an η such that $\boldsymbol{\beta}^*$ satisfies the weak sparsity constraint (2.9), the proposed estimator is consistent, that is, for any arbitrary $\epsilon > 0$, $Pr(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \epsilon) \to 0$ as $T, m \to \infty$.

Remark 2.

(a) Theorem 2 includes exact sparsity as a special case. If $\boldsymbol{\beta}^*$ is exactly sparse with k nonzero entries, by setting $\eta = 0$, we can obtain $|J_\eta| = k$ and $\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1 = 0$. Then, the above three upper bounds are the same as those in Theorem 1. For the weakly sparse scenario, we approximate $\boldsymbol{\beta}^*$ by its hard thresholding version $\boldsymbol{\beta}^*_\eta$. As a result, extra terms containing $\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1$ occur in the upper bounds.

(b) By setting $r_w = 1$, we obtain the upper bounds of the LASSO:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{3}{\alpha}\sqrt{|J_\eta|}\widetilde{\lambda}_N + 2\sqrt{\frac{\widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1}{\alpha}} + \frac{4\max\{\omega, 1\}}{\mathbb{Q}}\widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1,$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 3\sqrt{|J_\eta|}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + 4\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1,$$
$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{3}{2}\sqrt{|J_\eta|}\widetilde{\lambda}_N \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + 2\widetilde{\lambda}_N \|\boldsymbol{\beta}^*_{J_\eta^C}\|_1.$$

Furthermore, if the weak sparsity constraint (2.9) holds, the LASSO estimator is also consistent.

(c) If the weights $\{w_{l,ss'}\}$ are properly specified, the ratio r_w should be smaller than one and implies smaller error bounds than those of the LASSO. In the ideal case, when r_w is close to zero, the error bounds of the proposed method approach:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{1}{\alpha}\sqrt{|J_\eta|}\widetilde{\lambda}_N, \quad \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 2\sqrt{|J_\eta|}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2,$$
$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{1}{2}\sqrt{|J_\eta|}\widetilde{\lambda}_N \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2,$$

which are less than 1/3, 2/9, and 1/9, respectively, of those of the LASSO.

The meaning of the weak sparsity constraint (2.9) is straightforward. However, it is difficult to verify in practice, because it contains α, $\mathbb{Q}$, and ω, which depend on unknown model parameters. When α is bounded away from zero, and $\mathbb{Q}$ and ω are bounded away from infinity, the weak sparsity constraint can be simplified as $|J_\eta| = o\left(N/(\log p + 2\log m)\right)$ and $\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1 = o\left(N/(\log p + 2\log m)\right)$, which depends only on the number of observations and the parameter dimension. For a general stationary VAR process, the behaviors of α, $\mathbb{Q}$, and ω are complex, and cannot be guaranteed to be bounded. Here, we consider a simple case of VAR(1) process with symmetric transition matrix, and explore the properties of α, $\mathbb{Q}$, and ω in Proposition 1.

Proposition 1. *For any stationary VAR(1) process $X_t = \Phi X_{t-1} + \epsilon_t$ with symmetric transition matrix Φ, we have*

$$|\lambda_i| < 1 \text{ for any } i, \quad \rho(\Phi) = \max_{1\leq i\leq m} |\lambda_i|,$$
$$\mu_{\max}(\boldsymbol{\Phi}) = (1 + \rho(\Phi))^2, \quad \mu_{\min}(\boldsymbol{\Phi}) = (1 - \rho(\Phi))^2,$$

where $\{\lambda_i\}_{i=1}^m$ *are the eigenvalues of* Φ*. Furthermore,* α *is bounded away from zero,* $\mathbb{Q}$ *and* ω *are bounded away from infinity if and only if* $\Lambda_{\max}(\Sigma)$ *is bounded away from infinity,* $\Lambda_{\min}(\Sigma)$ *is bounded away from zero and* $\rho(\Phi)$ *is bounded away from one.*

This proposition implies the following: for a VAR(1) process with a symmetric transition matrix, if the eigenvalues of Σ and Φ behave properly and there exists an $\eta > 0$ satisfying $|J_\eta| = o(N/(\log p + 2\log m))$ and $\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1 = o(N/(\log p + 2\log m))$, we achieve the consistency of $\hat{\boldsymbol{\beta}}$. The symmetry of Φ is not required for the general case to build consistency. However, it helps to simplify the weak sparsity constraint, and makes it more informative for real applications. In addition, because $d_{ss'}$ equals to $d_{s's}$, a symmetric Φ can happen in reality when $\Phi_{ss'}$ is a function of the distance $d_{ss'}$.

l_r Ball Setting. Negahban et al. (2009) and Raskutti, Wainwright and Yu (2011) investigate the LASSO estimation of linear regression in independent data under the l_r ball setting. Under some conditions, they build up the upper bound of the l_2 estimation error and provide the condition for consistency (i.e. the l_r ball constraint (2.8)). Based on Theorem 2, we obtain a similar error bound under the l_r ball constraint for the proposed method (Corollary 1). Moreover, we prove that our constraint (2.9) is more relaxed than the l_r ball constraint, and thus is more general.

Corollary 1. *Consider the weighted* l_1*-LS estimator in* (2.4) *with true parameter* $\boldsymbol{\beta}^*$ *within the* l_r *ball:* $\mathbb{B}_r(R) := \{\boldsymbol{\beta}^* : \sum_{j=1}^q |\beta_j^*|^r \le R\}$*. Assume Assumptions* 1 *and* 3 *hold. Furthermore, set* $w_1 = \min\{w_{l,ss'}\}$*,* $w_2 = \max\{w_{l,ss'}\}$*,* $r_w = w_2/w_1$*,* $\lambda_N = 4w_1^{-1}\mathbb{Q}\sqrt{(\log p + 2\log m)/N}$*, and* $\eta = \lambda_N/\alpha$*. Then, there exist constants* $b_i > 0$*, such that for any* $N \gtrsim (1 + r_w)^2|J_\eta|\max\{\omega^2, 1\}(\log p + 2\log m)$*, with probability at least*

$$1 - b_1\exp(-b_2 N\min\{\omega^{-2}, 1\}) - b_3\exp(-b_4(\log p + 2\log m)),$$

the estimation error is bounded as follows:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{w_1 + 2w_2 + 2\sqrt{w_2}}{\alpha^{(2-r)/2}} R^{1/2}\lambda_N^{(2-r)/2} + \frac{4w_2\max\{\omega, 1\}}{\mathbb{Q}\alpha^{1-r}} R\lambda_N^{2-r}. \quad (2.10)$$

Remark 3.

(a) Corollary 1 implies that $\alpha^{(r-2)/2}R^{1/2}\lambda_N^{(2-r)/2} = o(1)$ and $\alpha^{r-1}R\lambda_N^{2-r}/Q = o(1)$ are required to obtain the estimation consistency in the l_r ball setting.

After plugging in the choice of λ_N, we obtain the following l_r ball constraint:

$$\alpha^{r-2}\mathbb{Q}^{2-r}R\left(\frac{\log p + 2\log m}{N}\right)^{(2-r)/2} = o(1), \quad \text{and}$$
$$\max\{\omega, 1\}\alpha^{r-1}\mathbb{Q}^{1-r}R\left(\frac{\log p + 2\log m}{N}\right)^{(2-r)/2} = o(1). \tag{2.11}$$

In the Supplemental Material, we prove this constraint is stricter than our weak sparsity constraint (2.9).

(b) Note that, in the special case when α is bounded away from zero and $\mathbb{Q}$ and ω are bounded away from infinity, the second term in (2.10) is of higher order than the first term. Thus, the convergence rate becomes $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O\left(R^{1/2}(\log q/N)^{1/2-r/4}\right)$. with $q = pm^2$ being the number of parameters. This rate is the same as that in the regression of independent data (Raskutti, Wainwright and Yu (2011); Negahban et al. (2009)).

3. Simulation Studies

In this section, we first describe the implementation of the proposed weighted l_1-LS approach (2.2). Then, we present several simulation studies that compare the proposed method with five existing penalized estimations of high-dimensional VAR: the LASSO (Basu and Michailidis (2015)), SCAD and MCP (Zhu (2020)), adaLASSO, and spaLASSO (Schweinberger, Babkin and Ensor (2017)). Several settings of the VAR order, dimension of the time series and sparse scenarios are considered. We find that, in all settings and scenarios, the proposed method achieves substantial improvements over the existing methods in terms of parameter estimation, network detection, and out-of-sample forecast. Because we have consistent findings across the different settings, we describe simulation of VAR (1) with $m = 100$ in detail and summarize the findings for other simulation settings.

3.1. Practical implementation

The objective function in the minimization problem (2.2) can be decomposed as a sum of independent objectives:

$$\sum_{i=1}^{m}\left[\frac{1}{N}\|\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{B}_i\|_2^2 + \lambda_N\Omega_i(\boldsymbol{B}_i)\right],$$

where $\boldsymbol{Y}_i$ and $\boldsymbol{B}_i$ are the ith columns of matrices $\boldsymbol{Y}$ and $\boldsymbol{B}$, respectively, and $\Omega_i(\boldsymbol{B}_i) = \sum_{l=1}^{p}\sum_{j=1}^{m} w_{l,ss'}|\Phi_{l,ss'}|$. Therefore, the optimization (2.2) can be solved

in parallel by solving the following sub-objectives:

$$\min_{\boldsymbol{B}_i} \frac{1}{N} \|\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{B}_i\|^2 + \lambda_N \Omega_i(\boldsymbol{B}_i), \quad i = 1, \ldots, m. \tag{3.1}$$

By defining $\widetilde{\Phi}_{l,ss'} = w_{l,ss'}\Phi_{l,ss'}$, $\widetilde{\boldsymbol{B}} = [\widetilde{\Phi}_1, \ldots, \widetilde{\Phi}_p]'$ and, correspondingly, $\widetilde{\boldsymbol{X}}^{(i)} = [\widetilde{\boldsymbol{X}}_1^{(i)}, \ldots, \widetilde{\boldsymbol{X}}_{mp}^{(i)}]$ the jth column of which is $\widetilde{\boldsymbol{X}}_j^{(i)} = \boldsymbol{X}_j \circ w^{(i)}$, with $w^{(i)} = (1/w_{1,i1}, \ldots, 1/w_{1,im}, \ldots, 1/w_{p,i1}, \ldots, 1/w_{p,im})'$, objective (3.1) is transformed into a LASSO optimization,

$$\min_{\widetilde{\boldsymbol{B}}_i} \frac{1}{N} \|\boldsymbol{Y}_i - \widetilde{\boldsymbol{X}}^{(i)}\widetilde{\boldsymbol{B}}_i\|_2^2 + \lambda_N \|\widetilde{\boldsymbol{B}}_i\|_1, \quad i = 1, \ldots, m,$$

which can be easily solved by existing LASSO algorithms.

In practice, we need to select the VAR order p, the penalty parameter λ_N, and the universal constant c_i in the penalty weights (2.3). The parameter selection can follow the forward cross-validation approach, which is commonly used in high-dimensional VAR model estimations (Bańbura, Giannone and Reichlin (2010); Song and Bickel (2011); Nicholson et al. (2020)) and provides good performance for finite samples as shown in the following simulation studies and real-data analysis. First, we separate the data into two sets: a training data set $\{1, \ldots, T_0\}$ and a validation data set $\{T_0 + 1, \ldots, T\}$. Here, T_0 is prespecified such as $T_0 = \lfloor 0.6T \rfloor$. Then, we specify potential values of p and c, such as $p \in \{1, \ldots, 4\}$ and $c_i \in \{0.5, 5, 10, 15, 20, 25, 30\}$. For each given pair of (p, c_i), we follow Friedman, Hastie and Tibshirani (2010) to perform a grid search of λ_N, which starts from λ_N^{max}, the smallest value that shrinks all parameters to zero, and then decreases in log-linear increments until the value of $\lambda_N^{max}/1000$ is reached. We take 30 values along this grid, and obtain $4 \times 7 \times 30$ triples of (p, c_i, λ_N). For each triple of (p, c_i, λ_N), we optimize (2.2) using the training data set and then calculate one-step-ahead forecast $\hat{X}_{t+1}^{(p,c_i,\lambda_N)}$ for the validation data set $(t = T_0, \ldots, T-1)$. Then, we select the values of $(p, c_i, \lambda_N) = (p^{opt}, c_i^{opt}, \lambda_N^{opt})$ by minimizing the following root mean squared forecast error (RMSFE):

$$RMSFE = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0}^{T-1} \frac{1}{m} \left\| \hat{X}_{t+1}^{(p,c_i,\lambda_N)} - X_{t+1} \right\|_2^2}.$$

Finally, we optimize (2.2) based on the selected $(p^{opt}, c_i^{opt}, \lambda_N^{opt})$ and data till T.

3.2. Simulation setting

In each simulation setting, we simulate the VAR process 100 times, and each simulated process has 150 observations. The last 80 points ($t = 71, \ldots, 150$) are preserved as a test data set for out-of-sample forecast comparison. For the LASSO, adaLASSO, SCAD, MCP, and proposed method, we apply the aforementioned forward cross-validation to select the tuning parameters, and set the data within $t = 1, \ldots, 40$ as training data set, and the data within $t = 41, \ldots, 70$ as the validation data set. We use the LASSO estimator to derive the penalty weight for the adaLASSO, that is, $\lambda_i = \lambda/|\tilde{\beta}_i|$ with the LASSO estimator $\tilde{\beta}_i$. For the spaLASSO, we use the code provided in the online supplemental materials of Schweinberger, Babkin and Ensor (2017) to carry out the model estimation and prediction. This method uses stability selection (Meinshausen and Bühlmann (2010)) to sidestep the selection of the tuning parameters. Two weight functions are considered in the proposed method:

$$\text{WLASSO1: } w^{(1)}_{l,ss'} = \exp\left(c_1 \frac{l\, d_{ss'}}{p\, d_{max}}\right); \quad \text{WLASSO2: } w^{(2)}_{l,ss'} = \left(1 + \frac{l\, d_{ss'}}{p\, d_{max}}\right)^{c_2}.$$

We consider the following criteria to compare the various methods:

- l_1 estimation error: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \sum_{l,s,s'} |\hat{\Phi}_{ss',l} - \Phi^*_{ss',l}|$.
- l_2 estimation error: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \sqrt{\sum_{l,s,s'} |\hat{\Phi}_{ss',l} - \Phi^*_{ss',l}|^2}$.
- Percentage of false zeros: $\text{PFZ} = \sum_{l,s,s'} I(\hat{\Phi}_{ss',l} = 0, \Phi^*_{ss',l} \neq 0)/m^2 p$.
- Percentage of false nonzeros: $\text{PFNZ} = \sum_{l,s,s'} I(\hat{\Phi}_{ss',l} \neq 0, \Phi^*_{ss',l} = 0)/m^2 p$.
- RMSFE for h-step out-of-sample forecast, with $h = 1, \ldots, 5$.

To simply the presentation of the results, we treat the LASSO as a benchmark, and report the ratio of each method over the LASSO. Ratio less than one means the method outperforms the LASSO.

3.3. Simulation of VAR(1) with dimension m = 100

First, we construct 21×21 lattices with coordinates $\{(x_i, y_j)\}_{i,j=1}^{20}$ as $x_i = 0.05i + \delta_i$ and $y_i = 0.05i + \delta_i$, where δ_i and δ'_i are independently generated from unif(-0.01,0.01). Then, we randomly select 100 sites from all 441 vertices in the lattice. Four sparse scenarios are considered:

(a) Exactly sparse: Generate $|\tilde{\Phi}^*_{ss'}| \sim \text{unif}(0.1, 0.5)$; then, set $|\Phi^*_{ss'}| = |\tilde{\Phi}^*_{ss'}|\, I(d_{ss'} \leq 0.05)$.

(b) Weakly sparse (fast decay): $|\Phi^*_{ss'}| = 0.55/\exp(20\, d_{ss'})$.

(c) Weakly sparse (slow decay): $|\Phi^*_{ss'}| = 0.25/\exp(5\, d_{ss'})$.

(d) Exactly sparse with zero parameters within a small distance: Generate $|\tilde{\Phi}^*_{ss'}| \sim \text{unif}(0.1, 0.5)$ and set $|\Phi^*_{ss'}| = |\tilde{\Phi}^*_{ss'}| I(d_{ss'} \leq 0.06)$. Then, randomly select 33% nonzero parameters to be zero.

The sign of $\Phi^*_{ss'}$ is selected randomly. Scenarios (a) and (d) represent exact sparsity, and scenario (d) less favor the proposed method because some parameters associated with the small distance are zero. This scenario is specifically designed to investigate the performance of the proposed method under an unfavorable scenario. Scenarios (b) and (c) represent weak sparsity. Compared with scenario (c), $|\Phi^*_{ss'}|$ in scenario (b) decays much faster, and thus is more sparsifiable. To guarantee that the VAR(1) process is stationary, the above generation procedure is repeated until all eigenvalues of Φ^* are within (-1,1). We set $\Sigma = 0.01I$.

Simulation results. The empirical results are reported in Table 1 and Figure 3 in the Supplemental Material. In terms of model fitting, the proposed method achieves a considerable improvement over the other five competing methods in all four scenarios, highlighting the advantage of incorporating spatial and temporal information. In particular, in scenario (a), the l_1 error, l_2 error, PFZ, and PFNZ of the proposed method are only 35%, 41%, 5%, and 20%, respectively, of those of the LASSO. In contrast, the other four methods do not outperform the LASSO and underestimate the nonzero parameters. The only exception is PFNZ. This is because the other competing methods are too conservative and severely underestimate the nonzero parameters. Thus, their PFNZ are low, but their PFZ are high. Meanwhile, it is not surprising that the proposed method has a high PFNZ in scenario (d), because there are some zero parameters associated with small distances. We further explored the true zero parameters with distances less than 0.06, and summarize their WLASSO estimations in Table 2 in the Supplemental Material. The zero parameters are estimated well by the proposed method, even though their WLASSO penalties are small. Specifically, more than 70% of the zero parameters are estimated as zero, and further 20% of them are estimated to be within $(-0.03, 0.03)$, which is negligible compared with the true nonzero parameters. This is because the estimation becomes a low-dimensional problem after the proposed method forces the parameters with larger distances to be zero. Thus, their estimations are close to the true value, that is, zero, even without large penalties. On the other hand, the penalty weights of the adaLASSO and spaLASSO are derived from an initial estimator (LASSO). Inaccuracy of

the initial estimator produces unreliable penalty weights, thus contaminating the estimation.

Figure 2 in Section 1 depicts the network detection results of one randomly selected replicate in scenario (a), and the results are consistent with what we observe in PFZ and PFNZ: the proposed method performs the best and provides desirable network estimation. In contrast, the other five methods severely underestimate the true connections. Meanwhile, the LASSO, adaLASSO and SCAD also overestimate wrong connections.

Figure 3 in the Supplementary Material plots the RMSFE ratio between each method and the benchmark (LASSO). The proposed method demonstrates significant improvement over the LASSO at $h = 1, 2, 3, 4$ in all scenarios. In contrast, the other four methods do not show obvious advantages over the LASSO, and sometimes even perform worse. In addition, the performance of WLASSO1 and WLASSO2 are very close, which confirms that the proposed method is not sensitive to the choice of weight function. The following simulation studies and real-data analysis also confirm this robustness.

3.4. Simulation for VAR(1) with m = 200, VAR(2) and VAR(3)

The simulation results are reported in Tables 3–5 and Figures 6–11 in the Supplementary Material, which also indicates the superiority of proposed method over other competing methods. Furthermore, the improvement over the other methods is more significant for $m = 200$ than that it is for $m = 100$. For example, the proposed method has a significantly better forecast than the others even at horizon $h = 5$, and its reduction in the RMSE is larger for $m = 200$ than that it is for $m = 100$. In addition, the improvement over the LASSO in terms of forecasting becomes more obvious as p increases (Figure 11). This is because our method penalizes the parameters based on both the spatial distance and the temporal lags.

4. Traffic Data Analysis

The real data contain the hourly traffic volumes recorded on 79 sites on highways around Des Moines, Iowa. The records are hourly data from Sepmtember 20, 2014 to November 2, 2014 (six weeks and two days), with a total of 1,056 observations for each site. The 79 sites are shown in Figure 17 in the Supplementary Material.

For each site s, the volume series $\{z_{st}\}$ $(s = 1, \ldots, 79;\ t = 1, \ldots, 1056)$ has strong weekly periodicity, that is, its weekly trend is repeated every 168 time

points. For each time point t, we use $d = t \bmod (168)$ to denote the hour of the time point t in one week. We model $\{z_{st}\}$ as follows:

$$z_{st} = \mu_{sd} + \sigma_{sd} x_{st}, \quad E(x_{st}) = 0 \text{ and } E(x_{st}^2) = 1, \quad \log(\sigma_{sd}) = a_s + b_s \log(\mu_{sd}),$$
$$X_t = (x_{1t}, \ldots, x_{mt}), \quad X_t = \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \varepsilon_t. \tag{4.1}$$

Here, $\{\mu_{sd}\}_{d=1}^{168}$ is the weekly trend of $\{z_{st}\}$, and $\{x_{st}\}_{t=1}^{1056}$ is the series after subtracting the trend and standardization, which is assumed to be stationary. In addition, $E(x_{st}) = 0$ and $E(x_{st}^2) = 1$ guarantee that σ_{sd} and x_{st} are identifiable. The following two-stage procedure is carried out for the estimation and forecasting.

Stage 1: Estimate $\boldsymbol{\mu_{sd}}$, $\boldsymbol{\sigma_{sd}}$, and series $\boldsymbol{\{x_{st}\}}$. We first use the local linear kernel regression (Fan, Heckman and Wand (1995)) to estimate $\{\mu_{sd}\}_{d=1}^{168}$ and obtain the detrended series $y_{st} := z_{st} - \hat{\mu}_{sd}$. Because we have multiple y_{st} at each d, we can approximate σ_{sd} using the standard error of these y_{st} (i.e., $\hat{\sigma}_{sd}$ is the standard error of $\{y_{st} : t \bmod (168) = d\}$). Then, we regress $\log(\hat{\sigma}_{sd})$ on $\log(\hat{\mu}_{sd})$ to estimate a_s and b_s. Finally, the estimate of the series $\{x_{st}\}$ is obtained as $\hat{x}_{st} = (z_{st} - \hat{\mu}_{sd}) / \exp(\hat{a}_s + \hat{b}_s \log(\hat{\mu}_{sd}))$. Figure 12 in the Supplemental Material illustrates the result for one site in Stage 1.

Note that some stretches of observations in $\{z_{st}\}$ are zero. This may be the result of road construction or maintenance at that time. These zero observations are considered outliers, and are excluded when estimating $\{\mu_{sd}\}$ and $\{\sigma_{sd}\}$. The following procedure is applied for outlier screening. For a given d, we have six to seven z_{st}. If the median of these z_{st} is above 30, but one of them, say z_{st_0}, is zero, we mark z_{st_0} as an outlier. In addition, we used boxplot to identify outliers: if z_{st_0} is below the interquantile of the 25% quantile or above the interquantile of the 75% quantile, z_{st_0} is marked as an outlier. We exclude these outliers when estimating $\{\mu_{sd}\}$ and $\{\sigma_{sd}\}$, but attribute them to component $\{x_{st}\}$.

Stage 2: Modeling $\boldsymbol{\{\hat{X}_t\}}$. Set $\hat{X}_t = (\hat{x}_{1t}, \ldots \hat{x}_{mt})'$. We apply the VAR, LASSO, and the proposed method to estimate model (4.1) and carry out the forecasting. Here, we divide the time period into four sub-periods: (1) weekday peak time (6 a.m. to 8 p.m.); (2) weekday off-peak time (9 p.m. to next day 5 a.m.); (3) weekend peak time (8 a.m. to 8 p.m.); (4) weekend off-peak time (9 p.m. to next day 7 a.m.). We carry out one-step- to four-steps-ahead forecasting for the last two weeks. To incorporate the spatial location information, we calculate the road distances between the 79 sites. If there is a highway path from site s to site s', $d_{ss'}$ is the road distance of this path, otherwise, we set $d_{ss'} = d_{max}$, where

$d_{max} := \max\{d_{ss'} : \text{there is a road path from } s \text{ to } s'\}$. The following four weight functions are considered:

$$\text{WLASSO1: } w^{(1)}_{l,ss'} = \exp\left(c_1 \frac{l\, d_{ss'}}{p\, d_{max}}\right), \quad \text{WLASSO2: } w^{(2)}_{l,ss'} = \left(1 + \frac{l\, d_{ss'}}{p\, d_{max}}\right)^{c_2},$$
$$\text{WLASSO3: } w^{(3)}_{l,ss'} = \left(\frac{l}{p} \exp\left(\frac{d_{ss'}}{d_{max}}\right)\right)^{c_3}, \quad \text{WLASSO4: } w^{(4)}_{l,ss'} = \exp\left(c_4 \frac{d_{ss'}}{d_{max}}\right).$$

We also tried a setting in which $d_{ss'} = \infty$ if there is no road path between site s and site s'. This setting forces the corresponding $\Phi_{ss',l}$ to be zero. In practice, these two distance settings provide very similar network detection and forecasting performance. For both the LASSO and the proposed method, the VAR order p is selected from $\{1, \ldots, 6\}$. Table 6 in the Supplemental Material lists the partition of the training data set, validation data set, and test data set. In summary, the last two weeks are the test data, and the last third and fourth weeks are the validation data. It turns out WLASSO1, WLASSO2, and WLASSO3 performs similarly and WLASSO4 behaves slightly worse. Thus, we report the result for WLASSO1 only.

Summary of fitting and forecasting results. Table 7 in the Supplemental Material lists the selected orders of the LASSO and WLASSO1 using forward cross-validation. For the VAR without any penalty, we fix $p = 1$, which gives the best forecast. WLASSO1 selects p as one or two for all sub-periods, but LASSO selects $p = 5$ for the weekend peak time. Here, $p = 5$ means one site may be influenced by another site even after five hours, which seems to be unreasonable. This is because the LASSO penalizes the parameters equally regardless of the temporal lag. The forecasting RMSFEs are listed in Table 8 in the Supplemental Material. Unsurprisingly, the LASSO and WLASSO1 outperform the VAR. In addition, WLASSO1 is superior to the LASSO for all scenarios, except the weekday peak time, with $h = 1$. In particular, for the weekend peak time, WLASSO1 outperforms the LASSO by reducing the RMSFE by 17%, 9%, 8%, and 6% for h=1, 2, 3, and 4 respectively. It also reduces the RMSFE by 8% for the weekend off-peak time, with $h = 1$. To examine the significance of such improvements, we carry out the Diebold–Mariano (DM) test (Diebold and Mariano (2002)) for each sub-period. The test results show that WLASSO1 is significantly better than the LASSO for the weekend peak time.

In addition, WLASSO1 yields a more reasonable network estimation than that of the LASSO in all sub-periods, as shown in Figures 13–16 in the Supplemental Material. The LASSO connects some sites that are far from each other, or

even in opposite directions, which is counter-intuitive, whereas WLASSO1 only connects sites that are close to each other.

5. Conclusion

We have introduced a data-driven weighted l_1 regularized estimation of a high-dimensional VAR model for spatio-temporal data. This method incorporates spatial distance and temporal lags to construct penalty weights. Its optimization is straightforward and easy to implement using existing algorithms. Its theoretical properties are explored for both the exactly sparse scenario and the weakly sparse scenario. We also explore the conditions for consistency, which shows the proposed method achieves smaller error bounds than those of the LASSO. The theoretical results of the l_1 regularization in the weakly sparse scenario are new, and have not been addressed previously in a time series framework. Our definition of weak sparsity is also more general than the l_r ball setting used in the literature. To evaluate the model performance, we compare the proposed method with five existing penalized VAR estimation methods using simulation studies, showing that the proposed method yields more reasonable network detection and a substantial improvement in terms of model fitting and forecasting. A real-data application on a traffic data set also demonstrates the advantages of the proposed method over the LASSO.

In this study, the tuning parameters are selected using cross-validation, yielding reasonable performance in the numerical analysis. Another popular approach is to use the BIC (Wang, Li and Tsai (2007a,b)). However, the BIC requires estimating the covariance matrix Σ, which can be infeasible when the number of observations T is smaller than the dimension m. A feasible solution in this case is to apply a penalized estimation for Σ. However, it involves another tuning parameter and is more expensive in terms of computation. An optimal procedure of selecting the tuning parameters for high-dimensional time series and the corresponding theoretical properties are beyond the scope of this study, and are left to future research.

Supplementary Material

The online Supplementary Material contains three parts: (1) proofs of the theorems, propositions, and corollaries; (2) simulation settings; and (3) tables and figures.

Acknowledgments

This research was supported by NSF 1455172, 1934985, 1940124, 1940276, USAID, and Atkinson's Center for a Sustainable Future.

References

Baek, C., Davis, R. A. and Pipiras, V. (2017). Sparse seasonal and periodic vector autoregressive modeling. *Computational Statistics & Data Analysis* **106**, 103 – 126.

Bańbura, M., Giannone, D. and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of applied Econometrics* **25**, 71–92.

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

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

Davis, R. A., Zang, P. and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics* **25**, 1077–1096.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics* **20**, 134–144.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141–150.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles* **33**, 1–22.

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

Hampton, S. E., Holmes, E. E., Scheef, L. P., Scheuerell, M. D., Katz, S. L., Pendleton, D. E. et al. (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (mar) models. *Ecology* **94**, 2663–2669.

Hu, L., Fortin, N. J. and Ombao, H. (2019). Modeling high-dimensional multichannel brain signals. *Statistics in Biosciences* **11**, 91–126.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics* **40**, 694–726.

Matteson, D. S. and Tsay, R. S. (2011). Dynamic orthogonal components for multivariate time series. *Journal of the American Statistical Association* **106**, 1450–1463.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

Michailidis, G. and d'Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical Biosciences* **246**, 326–334.

Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 1348–1356.

Nicholson, W. B., Wilms, I., Bien, J. and Matteson, D. S. (2020). High-dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research* **21**, 1–52.

Nicholson, W. B., Matteson, D. S. and Bien, J. (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* **33**, 627–651.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57**, 6976–6994.

Reyes, P. E., Zhu, J. and Aukema, B. H. (2012). Selection of spatial-temporal lattice models: Assessing the impact of climate conditions on a mountain pine beetle outbreak. *Journal of Agricultural, Biological, and Environmental Statistics* **17**, 508–525.

Safikhani, A., Kamga, C., Mudigonda, S., Faghih, S. S. and Moghimi, B. (2018). Spatio-temporal modeling of yellow taxi demands in New York City using generalized STAR models. *International Journal of Forecasting* **36**, 1138–1148.

Schweinberger, M., Babkin, S. and Ensor, K. B. (2017). High-dimensional multivariate time series with additional structure. *Journal of Computational and Graphical Statistics* **26**, 610–622.

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

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

Sun, Y., Li, Y., Kuceyeski, A. and Basu, S. (2018). Large spectral density matrix estimation by thresholding. *arXiv preprint arXiv:1812.00532*.

Tsay, R. S. (2015). *Financial Time Series*. American Cancer Society, Atlanta.

Tu, Y., Yao, Q. and Rongmao, Z. (2020). Error correction factor models for high-dimensional cointegrated time series. *Statistica Sinica* **30**, 1463–1484.

Wang, H., Li, G. and Tsai, C.-L. (2007a). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 63–78.

Wang, H., Li, R. and Tsai, C.-L. (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Zhu, X. (2020). Nonconcave penalized estimation in sparse vector autoregression model. *Electronic Journal of Statistics* **14**, 1413–1448.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zhenzhong Wang

Department of Statistics, Iowa State University, Ames, IA 50011-1090, USA.

E-mail: zhenzhong.wang77@gmail.com

Abolfazl Safikhani

Department of Statistics, University of Florida, Gainesville, FL 32611, USA.

E-mail: a.safikhani@ufl.edu

Zhengyuan Zhu

Department of Statistics, Iowa State University, Ames, IA 50011-1090, USA.

E-mail: zhuz@iastate.edu

David S. Matteson

Department of Statistics and Data Science, Cornell University, Ithaca, NY 14853, USA.

E-mail: matteson@cornell.edu

(Received February 2020; accepted October 2021)