Statistica Sinica Preprint No: SS-2020-0056	
Title	Regularized Estimation in High-Dimensional Vector Auto-Regressive Models Using Spatio-Temporal Information
Manuscript ID	SS-2020-0056
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0056
Complete List of Authors	Zhenzhong Wang, Abolfazl Safikhani, Zhengyuan Zhu and David S. Matteson
Corresponding Author	Zhengyuan Zhu
E-mail	zhuz@iastate.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Regularized Estimation in High-Dimensional Vector Auto-Regressive Models using Spatio-Temporal Information

Zhenzhong Wang, Abolfazl Safikhani, Zhengyuan Zhu and David S. Matteson

Iowa State University, University of Florida and 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 in terms of spatio-temporal data, most methods do not take the spatial and temporal structure of the data into consideration, 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 model. Extensive simulation studies are carried out to compare the proposed method with five existing methods of high-dimensional VAR model, demonstrating improvements of our method over others in parameter estimation, network detection and out-of-sample forecasts. We also apply our method on a traffic data set to evaluate its performance in real application. In addition, we explore the theoretical properties of l_1 regularized estimation of VAR model under the weakly sparse scenario, in which the exact sparsity can be viewed as a special case. To the best of our knowledge, this direction has not been considered yet in the literature. For general stationary VAR process, we derive the non-asymptotic upper bounds on l_1 regularized estimation errors, provide the conditions of es-

timation consistency, and further simplify these conditions for a special VAR(1) case.

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

1. Introduction

The Vector Auto-regressive (VAR) model, a popular tool to simultaneously model and forecast a number of time series, has been widely applied in different scientific fields such as econometrics (Sims, 1980), finance (Tsay, 2015), ecology (Hampton et al., 2013) and so on. Recent developments in computing have made high-dimensional time series increasingly common in many studies. As the number of time series component increases, the number of parameters in VAR model increases dramatically, which leads to unreliable or even infeasible estimation. The usual way to handle the high dimensionality is to impose sparsity or low rank structure on the transition matrices. Many estimation procedures have been proposed including but not limited to l_1 regularization (Basu and Michailidis, 2015), two-stage l_1 regularization (Davis et al., 2016), sparse seasonal VAR (Baek et al., 2017), low rank structured VAR (Basu et al., 2019), hierarchical lag sparsity (Nicholson et al., 2017; Nicholson and Matteson, 2020; Safikhani et al., 2018), banded VAR (Guo et al., 2016) and nonconcave penalization (Zhu

et al. 2020). Another group of methods assume a factor structure on the time series data to reduce the dimensionality, e.g., Lam and Yao (2012) and Tu et al. (2020). Meanwhile, such high-dimensional techniques become very popular in many applications, such as econometrics (Matteson and Tsay, 2011), genetics (Michailidis and d'Alché Buc, 2013), biology (Hu et al., 2019), ecology (Reyes et al., 2012) and so on.

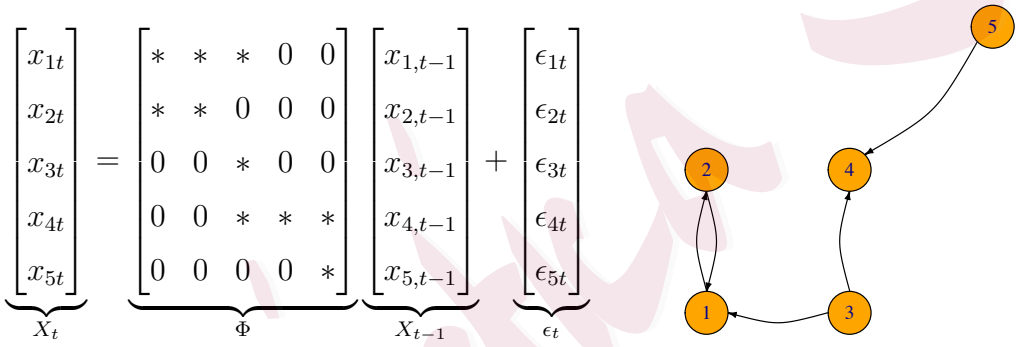

Figure 1: The left panel illustrates the sparsity (zero/non-zero) pattern for the transition matrix Φ in a VAR(1) process with $*$ denoting non-zero entries. The right panel illustrates the network structure implied by this VAR(1) process. For example, Φ_{13} is non-zero, which indicates a directed connection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

As for spatio-temporal data, each component of the multivariate time series contains the observations in one spatial location (site). Parameters

in the transition matrices can naturally capture the spatial and temporal interrelationship among the sites. Meanwhile the zero-nonzero patterns of the transition matrices reflect the network structure in the dataset. Figure 1 shows a simple example of VAR(1) model on five sites. We can see there exists a directed connection from site 3 to site 1, indicating that X_{1t} is dependent on $X_{3,t-1}$, so Φ_{13} is nonzero. Meanwhile $\Phi_{31} = 0$ 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 network estimation and unreasonable scientific conclusion. Figure 1 illustrates the drawback of ignoring the spatial and temporal information based on a simulation study in Section 3.3, in which the blue edges and red edges stand for false negatives and false positives respectively. Without considering the spatial and temporal information, the five existing methods studied in the simulation underestimate true connections. Meanwhile, LASSO, adaptive LASSO (adaLASSO) and SCAD also overestimates wrong connections. In contrast, our proposed method (WLASSO1 and WLASSO2) recovers the network very well and significantly reduces false positives and false negatives.

In this paper, we proposed a data-driven weighted l_1 regularized ap-

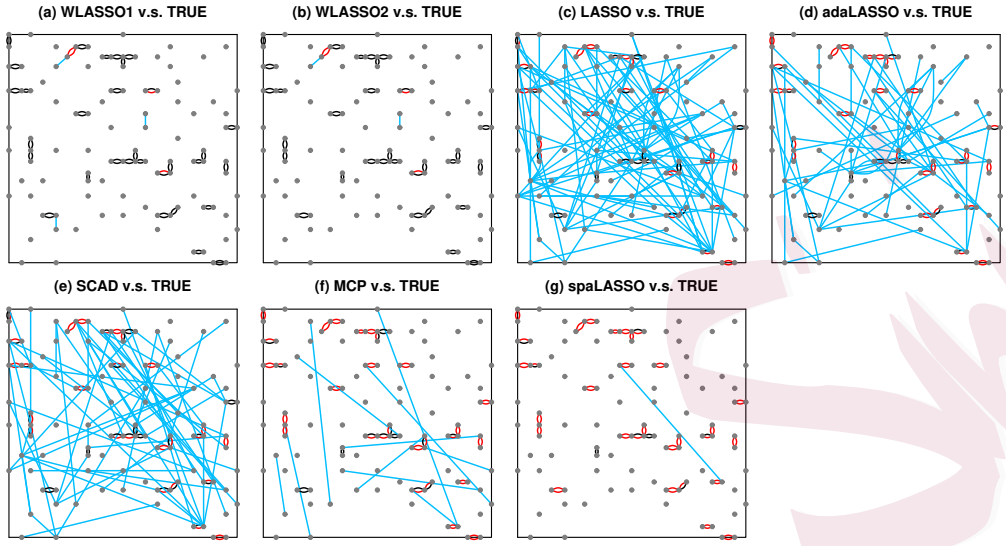

Figure 2: Comparison of the proposed methods (WLASSO1 and WLASSO2) to five existing methods (LASSO, adaptive LASSO, SCAD, MCP and spaLASSO from Schweinberger et al. (2017)) in network estimation of one simulated VAR(1) process from Section 3.3. In detail, if both true value $\Phi_{ss'}$ and its estimator $\hat{\Phi}_{ss'}$ 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 red. If $\Phi_{ss'}$ is zero but $\hat{\Phi}_{ss'}$ is not zero, the edge is blue.

proach that constructs the penalty according to spatial distance among sites and temporal lags in the VAR model. We derived non-asymptotic upper bounds of the estimation error which hold with high probability, and showed these bounds are smaller than those from LASSO (remark (c) in

Section 2.2 and Section 2.3). The simulation studies compare the proposed approach with five existing methods including LASSO (Basu and Michailidis, 2015), SCAD and MCP (Zhu et al., 2020), spaLASSO (Schweinberger et al., 2017) and adaptive LASSO (Zou, 2006; Wang et al., 2007a). The proposed approach shows significant advantage in model fitting, network detection and forecasting performance (Table 1-5 and Figure 3-11 in the supplemental material). We applied our method to a traffic network dataset in Des Moines, Iowa area. The network structure detected by LASSO is not meaningful while the proposed method provides more reasonable estimated network and better forecasting results.

There are few papers focusing on high-dimensional VAR in the spatio-temporal setting. The most relevant one is Schweinberger et al. (2017), denoted as spaLASSO, which incorporates spatial structure in VAR model estimation. Their approach assumes the spatial dependence only exists within a specific distance ρ, while ρ is either known or estimated in an initial step by LASSO within sub-sampled sites. After ρ is specified, only parameters associated with distances smaller than the given ρ are estimated, and others are fixed as zero. Assuming the distance ρ is known is usually unrealistic in real data sets. In estimating ρ by an initial LASSO estimator, inaccuracy of the initial estimator can produce unreliable estimation of ρ,

thus contaminating the final estimation of the model. As shown in Figure 2 (c), LASSO cannot identify the true network and therefore would deliver inaccurate estimation of ρ and eventually results in an inaccurate estimation from spaLASSO (Figure 2 (g)). Further, the assumption of no spatial dependence beyond distance ρ is restrictive, and may not be true in some real cases, such as the more general weakly sparse scenario considered in this paper. In addition, this approach does not incorporate the lag order of temporal dependence. In contrast, the proposed method incorporates both spatial and temporal information in a smooth way rather than truncating the parameters at a certain distance, and the penalty weights are data-driven so that no prior information is needed. The algorithm of the proposed method is one-step and easy to carry out via existing algorithms.

In 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 will have many small non-zero elements thus are not sparse, which is the so called "weakly sparse" scenario. The second goal of this paper is to investigate the theoretical properties of l_1 regularized estimation of VAR model under weakly sparse scenario. Weak sparsity is pursued mostly for independent data including Negahban et al. (2009) and Raskutti et al. (2011). There is a gap in the literature in investigat-

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

Outline of the Paper: The remainder of the paper is structured as follows. Section 2 introduces the weighted l_1 regularized approach for high-dimensional spatio-temporal VAR and its theoretical properties. Section 3 presents the implementation of the proposed method and compares its

performance with LASSO, SCAD, MCP, adaptive LASSO (adaLASSO) and spaLASSO through several simulation studies. Application on the traffic network dataset is in Section 4, followed by the conclusion in Section 5.

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. $\text{vec}(A)$, 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 of matrices A and B. $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ are the largest and smallest eigenvalue of a symmetric or Hermitian matrix A. For a squared matrix A, $\|A\|_F$, $\rho(A)$ and $\|A\|_2$ are its Frobenius norm $\sqrt{tr(A^HA)}$, spectral radius $\max\{|\lambda_i| : \lambda_i\text{'s are eigenvalues of } A\}$, and spectral norm $\sqrt{\Lambda_{\max}(A^HA)}$ 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 ($s = 1, \cdots, m$; $t = 1, \cdots, T$), and we assume $X_t = (x_{1t}, \cdots, x_{mt})'$ is generated by a p-th order vector auto-regressive (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 $\Phi_1, \cdots, \Phi_p$ 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 as 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, LASSO can recover the sparseness of transition matrices and reduce forecasting error (Basu and Michailidis, 2015). However, regular LASSO uses the same penalty for different $\Phi_{l,ss'}$ components, which may be inappropriate for spatio-temporal data. Instead, we proposed 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 $\Phi_{l,ss'}$. Since $\Phi_{l,ss'}$ quantifies the dependence between site s and site s' across 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 distance $d_{ss'}$ and temporal lag l. Through this construction of penalty weights we impose a spatio-temporal structure on the data in which the conditional dependence among two sites across temporal lag l (represented by $\Phi_{l,ss'}$) decays as 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 to be determined by cross validation. The inclusion of c_1 and c_2 ensures weights are data-driven and adds flexibility to this method. Other weight functions can be defined as well based on the context of the dataset under investigation. A special case is that $w_{l,ss'}$ is only a function of $d_{ss'}$ such as $w_{l,ss'}^{(3)} = \exp\left(c_3 d_{ss'}/d_{max}\right)$, which means the magnitudes of parameters are only influenced by the distance. The performances of different weight

functions are examined in simulation studies and real data application.

Utilizing weighted penalty functions such as those above significantly improves model performance without sensitivity to the exact choice of weight functions. This is mainly due to including a data-informed constants c_i in all weight functions, which are selected via cross-validation. Including such data-driven constants optimizes the weight to some extent and reduces the reliance of model performance on the choice of 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 non-asymptotic bounds on the estimation error of weighted l_1-LS estimation (2.2), and show that under certain conditions the proposed estimator is consistent. We rewrite the VAR model as:

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

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

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

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 $\Phi_1^*, \cdots, \Phi_p^*$. We consider two scenarios: (1) $\boldsymbol{\beta}^*$ is exactly sparse; (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. 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. Also Σ is positive definite.

This is a fundamental assumption in high-dimensional time series analysis. Since the key in 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 assumption, Basu and Michailidis (2015) used spectral density to construct measure of dependence and proved that $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Gamma}}$ satisfy two important conditions. More specifically, Proposition (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 RE condition (2.5) and Derivation condition (2.6) hold with probability at least

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

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

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

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

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

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. We refer to Basu and Michailidis (2015) for more details. The RE condition (2.5) and Derivation condition (2.6) are the key to derive 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 $|J| = k$. Further 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 being 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. So without lost of generality, we can set $\min\{w_{l,ss'} : (l, ss') \in J^C\} = 1$. Further we set $r_w = \max\{w_{l,ss'} : (l, ss') \in J\}$, which is indeed the ratio between the maximum weight of nonzero parameters and the minimum weight of zero parameters, i.e. $r_w = \max\{w_{l,ss'} : (l, ss') \in J\}/\min\{w_{l,ss'} : (l, ss') \in J^C\}$. In the following theorem, we can see this ratio is the key quantity for the proposed method to achieve smaller error bounds than LASSO.

Theorem 1. *Consider weighted l_1-LS estimator in (2.4). If Assumption 1 and 2 hold, there exist constants $b_i > 0$ not depending on data and 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 at least probability:*

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

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 zero is bounded by:*

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

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 non-zero in* $\tilde{\boldsymbol{\beta}}$ *is bounded by:*

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

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

(b) If we set $r_w = 1$ which corresponds to LASSO, we will get the following upper bounds that are similar to those in Basu and Michailidis (2015): $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq 3\sqrt{k}\lambda_N/\alpha$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 12k\lambda_N/\alpha$, $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq 9k\lambda_N^2/\alpha$, $|\text{supp}(\boldsymbol{\beta}^*)\backslash\text{supp}(\hat{\boldsymbol{\beta}}) \leq 12k\lambda_N/(s_0\alpha)$, $\left|\text{supp}(\tilde{\boldsymbol{\beta}})\backslash\text{supp}(\boldsymbol{\beta}^*)\right| \leq 9k/\alpha$.

(c) Compared with LASSO ($r_w = 1$), if 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 $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$, $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$, $\left|\text{supp}(\boldsymbol{\beta}^*)\backslash\text{supp}(\hat{\boldsymbol{\beta}})\right|$ and $\left|\text{supp}(\tilde{\boldsymbol{\beta}})\backslash\text{supp}(\boldsymbol{\beta}^*)\right|$ are nearly 1/3, 1/6, 1/9, 1/6 and 1/9 of that from LASSO respectively.

(d) Condition of Consistency: If the global penalty 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 only requires 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 large distance $d_{ss'}$ and/or lag l. For instance, $\Phi^*_{l,ss'} \neq 0$ may occur for large distance $d_{ss'}$ especially when the sites are located on

an irregular lattice. This example motivate 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. There are only few results in the literature discussing weak sparsity, and almost all of them have focused on independent data (Negahban et al., 2009, Raskutti et al., 2011), except Sun et al. (2018) which focuses on estimating the spectral density matrix of high-dimensional time series. Moreover, they each define weak sparsity under the so-called " l_r ball" setting. Specifically, they assumed the true parameter vector is within the l_r ball: $\mathbb{B}_r(R) := \left\{\boldsymbol{\beta}^* : \sum_{j=1}^q |\beta_j^*|^r \leq R\right\}$ 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, 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 et al., 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 "sparsifiablility" of $\boldsymbol{\beta}^*$. A special case in which all β_j^*s have the same magnitude could still fit in the l_r ball setting. While, in this case $\boldsymbol{\beta}^*$ cannot be approximated by a sparse vector and is not

suitable for l_1 regularized estimation. As a consequence, 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 $\boldsymbol{\beta}^*_\eta$, whose jth entry 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 (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 small enough. 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 proper

choice of λ_N we can show the proposed estimator is consistent. Furthermore, we simplified 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 estimation error under the l_r ball setting and prove our weak sparsity constraint (2.9) is more relaxed than the l_r ball constraint (Corollary 1). Also notice that the following Theorem 2 and Corollary 1 also hold for LASSO, since LASSO can be viewed as a special case of the proposed method where all $w_{l,ss'}$s are the same. To state our theorem, we define the following notations: for any η, we 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 weighted l_1-LS estimator in (2.4) and assume Assumption 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 at least probability:*

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

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

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

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

Remarks: (a) Theorem 2 includes the 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 exactly the same as those in Theorem 1. For weakly sparse scenario, we approximate $\boldsymbol{\beta}^*$ by its hard thresholding version $\boldsymbol{\beta}^*_\eta$. As a consequence, extra terms containing $\|\boldsymbol{\beta}^*_{J_\eta^C}\|_1$ occur in the upper bounds.

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

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

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

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

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

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

The meaning of the weak sparsity constraint (2.9) is straightforward. However, it is hard to verify in application since it contains α, $\mathbb{Q}$ and ω which depend on unknown model parameters. When α is bounded away from zero, $\mathbb{Q}$ and ω are bounded away from infinity, 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 only depends on the number of observation and parameter dimension. For 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 whose transition matrix is symmetric, and explore the properties of α, $\mathbb{Q}$ and ω in the following Proposition 1.

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

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

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 1.*

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

l_r **Ball Setting:** Negahban et al. (2009) and Raskutti et al. (2011) investigate LASSO estimation of linear regression in independent data under the l_r ball setting. Under some conditions, they built up the upper bound

of l_2 estimation error and provided the condition of consistency (i.e. l_r ball constraint (2.8)). Based on Theorem 2, we can obtain similar error bound and l_r ball constraint for the proposed method. We present this result as the following corollary. Moreover, we prove that our constraint (2.9) is more relaxed than the l_r ball constraint and thus more general.

Corollary 1. *Consider weighted l_1-LS estimator in (2.4) with true parameter $\boldsymbol{\beta}^*$ within the l_r ball: $\mathbb{B}_r(R) := \left\{\boldsymbol{\beta}^* : \sum_{j=1}^{q} |\beta_j^*|^r \leq R\right\}$. Assume Assumption 1 and 3 hold. Further 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 at least probability:*

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

the estimation error is bounded as follows:

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

Remark: (a) The above corollary implies $\alpha^{\frac{r-2}{2}} R^{\frac{1}{2}} \lambda_N^{\frac{2-r}{2}} = o(1)$ 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:

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

In the supplemental material, we proved this constraint is more strict than our weak sparsity constraint (2.9).

(b) It is also worth noting 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 $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 = O\left(R^{\frac{1}{2}}\left((\log p+2\log m)/N\right)^{1/2-r/4}\right) = O\left(R^{\frac{1}{2}}\left(\log q/N\right)^{1/2-r/4}\right)$ with $q=pm^2$ being the number of parameters. This rate is the same as that in regression of independent data (Raskutti et al., 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 which compare the proposed method with five existing penalized estimations of high-dimensional VAR, which includes LASSO (Basu and Michailidis, 2015), SCAD and MCP (Zhu et al., 2020), adaLASSO and spaLASSO (Schwein-

berger et al., 2017). Several settings of VAR order, dimension of time sereis and sparse scenarios are considered. It turns out in all settings and scenarios the proposed method achieves substantial improvement over the existing methods in parameter estimation, network detection and out-of-sample forecast. Due to consistent findings across different settings, we will describe simulation of VAR (1) with $m = 100$ in detail and only summarize distinct 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 column 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:

$$\min_{\boldsymbol{B}_i} \frac{1}{N} \|\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{B}_i\|^2 + \lambda_N \Omega_i(\boldsymbol{B}_i), \quad i = 1, \cdots, m. \tag{3.12}$$

By defining $\widetilde{\Phi}_{l,ss'} = w_{l,ss'} \Phi_{l,ss'}$, $\widetilde{\boldsymbol{B}} = [\widetilde{\Phi}_1, \cdots, \widetilde{\Phi}_p]'$ and correspondingly $\widetilde{\boldsymbol{X}}^{(i)} = \left[\widetilde{\boldsymbol{X}}_1^{(i)}, \cdots, \widetilde{\boldsymbol{X}}_{mp}^{(i)} \right]$ whose jth column is $\widetilde{\boldsymbol{X}}_j^{(i)} = \boldsymbol{X}_j \circ w^{(i)}$ with $w^{(i)} = (1/w_{1,i1}, \cdots, 1/w_{1,im}, \cdots, 1/w_{p,i1}, \cdots, 1/w_{p,im})'$, objective (3.12) is

transformed into a LASSO optimization:

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

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 estimation (Bańbura et al., 2010; Song and Bickel, 2011; Nicholson and Matteson, 2020) and it provides good performance for finite sample as shown in the following simulation studies and real data analysis. Firstly, we separate data into two sets: training dataset $\{1, \cdots, T_0\}$ and validation dataset $\{T_0 + 1, \cdots, T\}$. Here T_0 is pre-specified such as $T_0 = \lfloor 0.6T \rfloor$. Then we specify the potential values of p and c such as $p \in \{1, \cdots, 4\}$ and $c_i \in \{0.5, 5, 10, 15, 20, 25, 30\}$. For each given pair of (p, c_i), we follow Friedman et al. (2010) to perform the 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 dataset and then calculate 1-step-ahead forecast $\hat{X}_{t+1}^{(p,c_i,\lambda_N)}$ for the validation dataset ($t = T_0, \cdots, T-1$). After that we se-

lect 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 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 ($t = 71, \cdots, 150$) is preserved as test dataset for out-of-sample forecast comparison. For LASSO, adaLASSO, SCAD, MCP and the proposed method, we apply the aforementioned forward cross-validation to select the tuning parameters, and set data within $t = 1, \cdots, 40$ as training dataset and data within $t = 41, \cdots, 70$ as the validation dataset. We use LASSO estimator to derive the penalty weight for adaLASSO, i.e. $\lambda_i = \lambda/|\tilde{\beta}_i|$ with LASSO estimator $\tilde{\beta}_i$. For spaLASSO, we directly use the code in the online supplemental materials of Schweinberger et al. (2017) to carry out model estimation and prediction. This method uses stability selection (Meinshausen and Bühlmann, 2010) to sidestep the selection of 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 method performance:

- 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 $h = 1, \cdots, 5$.

To simply the presentation of results, we treat LASSO as benchmark and report the ratio of each method over LASSO. Ratio less than one means the method outperforms 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 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 0.

The sign of $\Phi^*_{ss'}$ is randomly selected. Scenarios (a) and (d) stand for exact sparsity, while scenario (d) is less favorable to the proposed method since some parameters associated with small distance are zero. This scenario is specifically designed to investigate the performance of the proposed method under unfavorable scenario. Scenarios (b) and (c) are for weak sparsity. Comparing to scenario (c), $|\Phi^*_{ss'}|$ in scenario (b) decays much faster thus more sparsifiable. To guarantee 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 is reported in Table 1 and Figure 3 in the supplemental material. In terms of model fitting, the proposed method achieves considerable improvement over other five competing methods in all four scenarios, highlighting the advantage of incorporating spatial and temporal information. In particular, in scenario (a), l_1 error, l_2 error, PFZ and PFNZ of the proposed method are only 35%, 41%, 5% and 20% of those from LASSO. In contrast, the other four methods do not

outperform LASSO and suffer from underestimation of nonzero parameters. The only exception is PFNZ. This is because other competing methods are too conservative and severely underestimate nonzero parameters, thus their PFNZ are low but their PFZ are high. Meanwhile, it is not surprising that the proposed method has high PFNZ in scenario (d) since there are some zero parameters associated with small distances. We further explored the true zero parameters with distances less than 0.06 and summarized their WLASSO estimation in Table 2 in supplemental material. It demonstrates that the zero parameters are estimated pretty 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 another 20% of them are estimated to be within (-0.03,0.03) which are negligible compared to 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 zero, thus their estimation will be close to the true value, i.e. zero, even without large penalties. On the other hand, the penalty weights of adaLASSO and spaLASSO are derived from an initial estimator (LASSO). Inaccuracy of the initial estimator would produce unreliable penalty weights thus contaminating the estimation.

Figure 2 in Section 1 depicts the network detection results of one ran-

domly selected replicate in scenario (a) and the results are consistent with what we observed in PFZ and PFNZ: the proposed method performs the best and provides desirable network estimation. In contrast, the other five methods severely underestimate true connections, meanwhile 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 significantly improves over LASSO at $h = 1, 2, 3, 4$ in all scenarios. In contrast, the other four methods do not show obvious advantages over LASSO and sometimes are even worse. In addition, the performance of WLASSO1 and WLASSO2 are very close, which confirms the proposed method is not sensitive to the choice of weight functions. The following simulation studies and real data analysis also confirm this robustness.

3.4 Simulation for VAR(1) with $\mathbf{m = 200}$, VAR(2) and VAR(3)

Simulation results are reported in Table 3-5 and Figure 6-11 in supplementary material, which also indicates the superiority of proposed method over other competing methods. Furthermore, the improvement over other methods is more significant in $m = 200$ than that in $m = 100$. For example, the proposed method has significantly better forecast than others even at hori-

zon $h = 5$ and its reduction in RMSE is larger in $m = 200$ than $m = 100$. In addition, the improvement over LASSO on forecasting becomes more obvious as p increases (Figure 11). This is because our method penalizes parameters according to not only spatial distance but also temporal lags.

4. Traffic Data Analysis

The real data contains the hourly traffic volumes recorded on 79 sites on highways around Des Moines, Iowa. The records are hourly data from 2014-09-20 to 2014-11-02 (six weeks and two days), with a total of 1056 observations for each site. These 79 sites are shown in Figure 17 in the supplementary material.

For each site s, the volume series $\{z_{st}\}$ $(s = 1, \cdots, 79;\ t = 1, \cdots, 1056)$ has strong weekly periodicity, i.e. 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 the volume series $\{z_{st}\}$ as follows:

$$z_{st} = \mu_{sd} + \sigma_{sd} x_{st}, \quad E(x_{st}) = 0 \text{ and } E(x_{st}^2) = 1, \quad \log(\sigma_{sd}) = a_s + b_s \log(\mu_{sd}),$$
$$X_t = (x_{1t}, \cdots, x_{mt}), \quad X_t = \Phi_1 X_{t-1} + \cdots + \Phi_p X_{t-p} + \varepsilon_t. \tag{4.13}$$

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 station-

ary. $E(x_{st}) = 0$ and $E(x_{st}^2) = 1$ guarantee σ_{sd} and x_{st} are identifiable. The following two-stage procedure is carried out for 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 et al., 1995) to estimate $\{\mu_{sd}\}_{d=1}^{168}$, and obtain detrended series $y_{st} := z_{st} - \hat{\mu}_{sd}$. Since we have multiple y_{st}'s at each d, we can approximate σ_{sd} by the standard error of these y_{st}'s (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 series $\{x_{st}\}$ can be obtained by $\hat{x}_{st} = (z_{st} - \hat{\mu}_{sd})/\exp(\hat{a}_s + \hat{b}_s \log(\hat{\mu}_{sd}))$. Figure 12 in the supplemental material illustrates the result of one site in Stage 1.

Notice that some stretch of observations in $\{z_{st}\}$ are zero. This may be a result from road construction or maintenance at that time. These zero observations are considered as outliers and 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}'s. If the median of these z_{st}'s is above 30, but one of them, say z_{st_0}, is zero, we mark z_{st_0} as outlier. In addition, we used the idea of boxplot to identify outliers: if z_{st_0} is below the interquantile of 25% quantile or above the interquantile of 75% quantile, z_{st_0} is marked as outlier. We exclude these outliers when estimating $\{\mu_{sd}\}$ and $\{\sigma_{sd}\}$, but

attribute them to component $\{x_{st}\}$.

Stage 2: Modeling $\{\hat{\boldsymbol{X}}_{\boldsymbol{t}}\}$ Set $\hat{X}_t = (\hat{x}_{1t}, \cdots \hat{x}_{mt})'$, we apply VAR, LASSO and the proposed method to estimate model (4.13) and carry out forecasting. Here we divide the time period into four sub-periods: (1) weekday peak time (6am - 8pm); (2) weekday off-peak time (9pm - next day 5am); (3) weekend peak time (8am - 8pm); (4) weekend off-peak time (9pm - next day 7am). We carried out 1 to 4 steps ahead forecasting for the last two weeks. To incorporate the spatial location information, we calculate the road distances among 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 kinds of 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 another 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 LASSO and the proposed method, VAR order p is selected from $\{1, \cdots, 6\}$. Table 6 in the supple-

mental material lists the partition of training dataset, validation dataset and test dataset. In short, the last two weeks are the test data, the last third and forth weeks are the validation data. It turns out the performance of WLASSO1, WLASSO2 and WLASSO3 are very close and WLASSO4 behaves slightly worse, thus we only report the result of WLASSO1.

Summary of Fitting and Forecasting Results Table 7 in the supplemental material lists the selected orders of LASSO and WLASSO1 through forward cross-validation. For VAR without any penalty, we fix $p = 1$ which gives the best forecast. WLASSO1 selects p as 1 or 2 for all sub-periods, but LASSO selects $p = 5$ for weekend peak time. $p = 5$ means one site may be influenced by another site even after five hours, which seems to be unreasonable. This fallacy is because LASSO penalizes parameters equally no matter what the temporal lag is. The forecasting RMSFEs are listed in Table 8 in the supplemental material. Unsurprisingly LASSO and WLASSO1 behave much better than VAR. Meanwhile WLASSO1 is superior than LASSO for all scenarios except weekday peak time with $h = 1$. In particular for weekend peak time, WLASSO1 outperforms LASSO by reducing RMSFE by 17%, 9%, 8% and 6% for h=1, 2, 3 and 4 respectively. It also reduces RMSFE by 8% in weekend off-peak time with $h = 1$. To examine the sig-

nificance of such improvements, we carried out Diebold-Mariano (DM) test (Diebold and Mariano, 2002) for each sub-period. The test results state that WLASSO1 is significantly better than LASSO in weekend peak time.

In addition, WLASSO1 gives more reasonable network estimation than LASSO does in all sub-periods as shown in Figure 13-16 in the supplemental material. LASSO connects some sites far from each other or even in the opposite directions, which is counter-intuitive, while WLASSO1 only connects the sites close to each other.

5. Conclusion and Discussion

In this paper, we introduced a data-driven weighted l_1 regularized estimation of 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 by existing algorithms. Its theoretical properties has been explored in both exactly sparse scenario and weakly sparse scenario, as well as the conditions for consistency, which indicates the proposed method achieves smaller error bounds than LASSO. The theoretical results of l_1 regularization in weakly sparse scenario are new and have not been addressed in the time series framework. Our definition of weak sparsity is also more general than the

l_r ball setting in the literature. To evaluate the model performance, we compare the proposed method with five existing penalized VAR estimation through simulation studies, which demonstrates the proposed method can obtain more reasonable network detection and substantial improvement on model fitting and forecasting. Real application on a traffic dataset also indicates advantages of the proposed method over LASSO.

In this paper, the tuning parameters are selected by cross-validation and it yields reasonable performances in the numerical analysis. Another popular approach is the BIC criterion (Wang et al., 2007a,b). However, BIC requires estimation of covariance matrix Σ, and the traditional estimation of Σ is infeasible when the number of observations T is smaller than dimension m. A feasible solution in this case is to apply penalized estimation for Σ, but it will involve another tuning parameter and is more expensive in computation. The optimal procedure of tuning parameter selection for high-dimensional time series and the corresponding theoretical properties is out of the scope of this paper, but it is an interesting topic for future study.

Supplementary Materials

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

Acknowledgements

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

References

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

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of applied Econometrics 25*(1), 71–92.

Basu, S., X. Li, and G. Michailidis (2019). Low rank and structured modeling of high dimensional vector autoregressions. *IEEE Transactions on Signal Processing 67*(5), 1207–1222.

Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist. 43*(4), 1535–1567.

Davis, R. A., P. Zang, and T. Zheng (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics 25*(4), 1077–1096.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1), 134–144.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*(429), 141–150.

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

Guo, S., Y. Wang, and Q. Yao (2016). High-dimensional and banded vector autoregressions. *Biometrika 103*(4), 889–903.

Hampton, S. E., E. E. Holmes, L. P. Scheef, M. D. Scheuerell, S. L. Katz, D. E. Pendleton, and E. J. Ward (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (mar) models. *Ecology 94*(12), 2663–2669.

Hu, L., N. J. Fortin, and H. Ombao (2019). Modeling high-dimensional multichannel brain signals. *Statistics in Biosciences 11*(1), 91–126.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Matteson, D. S. and R. S. Tsay (2011). Dynamic orthogonal components for multivariate time series. *Journal of the American Statistical Association 106*(496), 1450–1463.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

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

Negahban, S., B. Yu, M. J. Wainwright, and P. K. Ravikumar (2009). A unified framework for

high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pp. 1348–1356.

Nicholson, W. B.; Bien, J. and D. S. Matteson (2020). High-dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research 21*(166), 1–52.

Nicholson, W. B., D. S. Matteson, and J. Bien (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting 33*(3), 627–651.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory 57*(10), 6976–6994.

Reyes, P. E., J. Zhu, and B. H. Aukema (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*(3), 508–525.

Safikhani, A., C. Kamga, S. Mudigonda, S. S. Faghih, and B. Moghimi (2018). Spatio-temporal modeling of yellow taxi demands in new york city using generalized star models. *International Journal of Forecasting*.

Schweinberger, M., S. Babkin, and K. B. Ensor (2017). High-dimensional multivariate time series with additional structure. *Journal of Computational and Graphical Statistics 26*(3), 610–622.

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

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

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

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

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

Wang, H., G. Li, and C.-L. Tsai (2007a). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(1), 63–78.

Wang, H., R. Li, and C.-L. Tsai (2007b, 08). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

Zhu, X. et al. (2020). Nonconcave penalized estimation in sparse vector autoregression model. *Electronic Journal of Statistics 14*(1), 1413–1448.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Department of Statistics, Iowa State University, Ames, IA. E-mail: zwang1@iastate.edu

Department of Statistics, University of Florida, Gainesville, FL. E-mail: a.safikhani@ufl.edu

Department of Statistics, Iowa State University, Ames, IA. E-mail: zhuz@iastate.edu

Department of Statistics and Data Science, Cornell University, Ithaca, NY. E-mail: matteson@cornell.edu