Journal of Machine Learning Research 0 (0000) 0-0 Submitted 00/00; Published 00/00

Sparse Nonlinear Vector Autoregressive Models

Yuefeng Han YUEFENG.HAN@RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854-8019, USA

Likai Chen LIKAI.CHEN@WUSTL.EDU
Department of Mathematics and Statistics
Washington University in St. Louis
St. Louis, MO 63130-4899, USA

Wei Biao Wu WBWU@GALTON.UCHICAGO.EDU
Department of Statistics
The University of Chicago
Chicago, IL 60637, USA

Editor:

Abstract

Vector autoregressive (VAR) models have a wide range of scientific applications in econometrics, computational biology, climatology, and so on. Prior work has focused on linear VAR models. However, linear VAR approaches are somewhat restrictive in practice. This paper introduces the non-parametric sparse additive model, a more flexible framework to address this challenge. Our method uses basis expansions to construct nonlinear VAR models. We provide convergence rates and model selection consistencies of the estimators in terms of the dependence measures of the processes, the moment condition of the errors, the sparsity condition and basis expansions. Our theory substantially extends earlier linear VAR models by allowing non-Gaussianity and non-linearity structures. As our main technical tools, we derive sharp Bernstein-type inequalities for tail probabilities for non-sub-Gaussian linear and nonlinear VAR processes. Modulo some constants, our exponential inequalities coincide with the classical Bernstein inequality for independent random variables. We also provide numerical experiments that support our theoretical results and display advantages of using nonlinear VAR model for a time series gene expression data set.

Keywords: Vector autoregressive (VAR) model, Bernstein inequality, Sparsity, Basis expansion, Time series

1. Introduction

Driven by a diversity of contemporary scientific applications, high dimensional data with network structure play a key role in statistics. The demand for modelling and forecasting such data arises from genomics, panel studies in economics, environmental studies, and communication engineering, among others. For example, reconstruction of gene regulatory networks from expression data has become a canonical problem in computational system biology (Lawrence et al., 2010); analysis of roll call of legislative bodies is essential in

political science (Morton and Williams, 2010); understanding climate changes implies to be able to predict the behavior of climate variables and their relationship (Liu et al., 2010). The inference of networks that describe how variables influence each other has emerged simultaneously from all these fields.

Over the past decade, a number of statistical models have been developed for estimating networks from high dimensional data. Graphical models have emerged as a powerful class of models and a large amount of theoretical advances have been introduced for independent and identically distributed (i.i.d.) data under structural assumptions; e.g., see Bühlmann and van de Geer (2011). Under time series setting, there also exists a substantial literature on network inference based on sparse linear models or Granger causality concepts. Friedman (2004) and Lèbre (2009) applied dynamic Bayesian networks to time series data. Basu and Michailidis (2015) investigated theoretical properties of Lasso penalized high dimensional linear vector autoregressive (VAR) models for Gaussian processes. This was further extended to multi-block VAR models in Lin and Michailidis (2017). Guo et al. (2016) proposed a class of VAR models with banded coefficient matrices. Gao et al. (2019) extend the idea of banded coefficient matrices to study spatio-temporal VAR models. Hall et al. (2018) studied regularized high-dimensional autoregressive generalized linear models. Ghosh et al. (2019) developed a Bayesian VAR model with multivariate stochastic volatility.

Despite many mechanisms (e.g. regulatory methods in biology, cf. Sima et al. (2009) for a survey) involve nonlinear dynamics, very limited work focused on network inference for variables in the presence of such dynamics. Mazur et al. (2009) and Äijö and Lähdesmäki (2009) applied Bayesian learning to deal with the stochasticity of biological data. Lim et al. (2015) introduced a family of VAR models based on different operator-valued kernels to identify the nonlinear dynamic system. Zhou and Raskutti (2018) provided a framework of autoregressive models under the generalized linear models by exploiting reproducing kernel Hilbert spaces, and analyzed the convex penalized sparse and smooth estimator. In this paper, we aim at extending the framework of sparse linear VAR models to that of sparse non-parametric nonlinear VAR models.

The goal of this paper is two folds: (i) to develop sharp inequalities for tail probabilities for non-sub-Gaussian nonlinear VAR processes; (ii) to propose a new class of methods for high dimensional non-parametric VAR models and to apply our inequalities to obtain theoretical properties of ℓ_1 regularized estimators. It is expected that our framework, inequalities and tools will be useful in other high-dimensional linear and nonlinear VAR problems.

In our theoretical framework, we shall consider the following nonlinear VAR models

$$X_i = h^{(1)}(X_{i-1}) + h^{(2)}(X_{i-2}) + \cdots + h^{(d)}(X_{i-d}) + \epsilon_i, \tag{1}$$

where $\epsilon_i \in \mathbb{R}^p, i \in \mathbb{Z}$, are i.i.d random vectors, $X_i = (X_{i,1}, ..., X_{i,p})^\top \in \mathbb{R}^p$, $h^{(j)} = (h_1^{(j)}, \ldots, h_p^{(j)})^\top$ and $h_k^{(j)} : \mathbb{R}^p \to \mathbb{R}$, $1 \le j \le d, 1 \le k \le p$, are real-valued functions. By stacking lagged vectors, we can let $d = 1$ in (1) and consider the nonlinear VAR(1) model. Then (1) can be rewritten as

$$X_i = h(X_{i-1}) + \epsilon_i. \tag{2}$$

Based on model (2), we shall develop sharp Bernstein-type inequalities. Establishing exponential-type tail probability inequalities for temporal dependent processes is a chal-

lenging problem. There has been some effort to derive concentration inequalities for non-i.i.d. processes. For example, generalizations of Bernsteins inequality to α-mixing and ϕ-mixing random variables have been studied in Bosq (1993), Modha and Masry (1996), Samson (2000) and Merlevède et al. (2009, 2011), among others. Zhang (2021) provided Bernstein-type inequality for dependent random variables under geometric moment contraction. Exponential-type inequalities were also derived for sums of Markov chains in Douc et al. (2008) under some drift condition and in Adamczak (2008) under the minorization condition. Unfortunately, all these inequalities involve extra non-constant factors to account for weak dependence, and are not as sharp as Bernsteins inequality for independent random variables. Recently, Fan et al. (2018) and Jiang et al. (2018) established sharp Hoeffding-type inequality and Bernstein-type inequality for stationary Markov dependent random variables. Chen and Wu (2018) derived exponential inequalities and Nagaev-type inequalities for one dimensional linear (or moving average) processes under both short- and long-range dependence. Due to the interactions between temporal and cross-sectional dependence, tail probabilities of high dimensional time series is much more complicated than one dimensional ones. In this work, we establish Bernstein-type inequalities for nonlinear VAR processes. Modulo some constants, our Bernstein-type inequalities are as sharp as the classical Bernstein inequality for i.i.d. random variables. To the best of our knowledge, we are among the first to develop sharp Bernstein-type inequalities for high dimensional time series.

To study nonlinear dynamical systems from high dimensional time series data, in this paper, we introduce sparse additive non-parametric VAR models. Our method combines ideas from sparse linear modelling, additive non-parametric regression and VAR models. Each nonlinear function h_j, $1 \leq j \leq p$, in model (2) can be expressed as:

$$h_j(x) = \sum_{k=1}^{p} h_{jk}(x_k),$$

where $x = (x_1, ..., x_p)^\top \in \mathbb{R}^p$ and $h_{jk}(\cdot)$ are functions of one dimensional variables. The underlying VAR model is similar to sparse linear regression, but we impose a sparsity constraint on the index set $\{(j,k) : h_{jk} \neq 0\}$ of functions h_{jk} that are not identically zero. Then we estimate each nonlinear function h_{jk} in terms of a truncated set of basis functions. Ravikumar et al. (2009) proposed a sparse additive linear models using a basis expansion and LASSO type penalty under i.i.d. data. Meier et al. (2009) considered a sparsity-smoothness penalty for high-dimensional generalized additive models. Koltchinskii and Yuan (2010), Raskutti et al. (2012) and Yuan and Zhou (2016) studied a different framework, sparse additive kernel regression, for the cases where the component functions belong to a reproducing kernel Hilbert spaces. They penalized the sum of the reproducing kernel Hilbert space norms of the component functions. Their sparse additive linear models are extended to autoregressive generalized linear models in Zhou and Raskutti (2018). Lim et al. (2015) introduced operator-valued kernel-based VAR models, and developed proximal gradient descent algorithms. However, their paper does not provide any theoretical guarantees.

In this work, our method has the nice feature that it decouples smoothness and sparsity. This leads to a simple block coordinate descent algorithm (cf. Ravikumar et al. (2009)) that can be carried out with any non-parametric smoother and scales easily to high dimensions.

Besides, with our new probability inequalities as primary tools, we can analyze the properties of ℓ_1 regularized estimators under non-Gaussian errors in the context where p is much larger than n. Roughly speacking, p can be as large as e^{n^c} for some constant $0 < c < 1$ if ϵ_i has finite exponential moments, and the power constant c is related to the truncated number of basis expansion. We shall give a detailed description on how the dependence measures of the processes, the moment condition of the errors, the sparsity of functions and basis expansion affect the rate of convergence and the model selection consistency of the estimator.

The rest of the paper is structured as follows. Section 2 presents Bernstein-type inequalities for nonlinear VAR processes in (2) under Lipschitz condition and different types of moment conditions for the error processes. In Section 3, we first formulate an ℓ_1 regularized optimization problem for nonlinear VAR models in the population level that induces sparsity. Then we derive a sample version of the problem using basis expansion. Theoretical properties that analyze the effectiveness of the estimators in the high dimensional setting are also presented. Simulation studies and real data analysis are carried out in Section 4 and 5, respectively. Proofs of Theorems in Section 3 and technical lemmas are contained in Section 6.

We now introduce some notation. For a vector $x = (x_1, ..., x_p)^\top$, define $|x|_q = (|x_1|^q + ... + |x_p|^q)^{1/q}$, $q \geq 1$, $|x| = |x|_2$, and $\text{abs}(x) := (|x_1|, \ldots, |x_p|)^\top$. For a matrix $A = (a_{ij})$, write $|A|_\infty = \max_{i,j} |a_{ij}|$, the Frobenius norm $\|A\|_F = (\sum_{ij} a_{ij}^2)^{1/2}$, the spectral norm $\|A\|_2 = \max_{|x|_2 \leq 1} |Ax|_2$ and the matrix infinity norm $\|A\|_\infty = \max_i \sum_j |a_{ij}|$. Let $\lambda_{\min}(A)$ (resp. $\lambda_{\max}(A)$) be the minimum (resp. maximum) eigenvalue of A. Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_p)^\top$ be a random vector. Write $\boldsymbol{\xi} \in \mathcal{L}^m$, $m \geq 1$, if the m-norm $\|\boldsymbol{\xi}\|_m := (\mathbb{E}|\boldsymbol{\xi}|^m)^{1/m} < \infty$. Denote $\|\boldsymbol{\xi}\| := \|\boldsymbol{\xi}\|_2$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ (resp. $a_n \asymp b_n$) if there exists a constant C such that $|a_n| \leq C|b_n|$ (resp. $1/C \leq a_n/b_n \leq C$) holds for all sufficiently large n, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

Let $\epsilon_i, i \in \mathbb{Z}$, be i.i.d. random vectors and $\mathcal{F}_k = (\ldots, \epsilon_{k-1}, \epsilon_k)$. Define projection operator P_k, $k \in \mathbb{Z}$, by $P_k(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_k) - \mathbb{E}(\cdot|\mathcal{F}_{k-1})$. Let (ϵ_k') be an i.i.d copy of (ϵ_k). For $X_i = \mathcal{H}(\ldots, \epsilon_{i-1}, \epsilon_i)$, where $\mathcal{H}$ is some measurable function, we define the coupled version $X_{i,\{k\}} = \mathcal{H}(\ldots, \epsilon_{k-1}, \epsilon_k', \epsilon_{k+1}, \ldots, \epsilon_i)$, which has the same distribution as X_i with ϵ_k in the latter replaced by an i.i.d. copy ϵ_k'.

2. Bernstein Inequalities for Nonlinear VAR Processes

Exponential inequalities play a fundamental role in high dimensional inference. Differently from i.i.d data, directly applying concentration inequalities for dependent random variables to high dimensional time series problems may lead to suboptimal results in many cases, due to the interrelationship between temporal and cross sectional dependencies. Zhang and Wu (2017) and Zhang and Wu (2020) introduced new dependence measures to describe temporal and cross-sectional dependence of high dimensional time series, then derived Fuk-Nagaev type inequalities for heavy tailed random vectors to study statistical properties of sample mean vector and spectral density matrix estimation, respectively. In this section, we shall present new and powerful inequalities for tail probabilities of nonlinear vector autoregressive (VAR) processes. The processes can be non-Gaussian. In Theorems 1 and 4, we provide Bernstein-type inequalities for nonlinear VAR process under finite moment condition and

exponential moment condition, respectively. In contrast, exponential inequalities provided in Basu and Michailidis (2015) are only applicable to Gaussian processes and linear VAR models with Gaussian innovation vectors (cf. Proposition 2.4 therein).

To establish exponential inequalities, we introduce the following assumptions on function h and errors ϵ_i in model (2). Recall $\|\cdot\|_\infty$ is the matrix infinity norm.

Assumption 1 *Consider model* (2)*, let* $h = (h_1, ..., h_p)^\top$ *and* $h_j : \mathbb{R}^p \to \mathbb{R}$*,* $1 \le j \le p$ *be real valued functions. Assume that componentwise Lipschitz condition holds for each* h_j*. That is, for any* $x = (x_1, ..., x_p)^\top, y = (y_1, .., y_p)^\top \in \mathbb{R}^p$*,* $1 \le j \le p$*, there exists coefficients* $H_{jk} \ge 0$ *such that*

$$|h_j(x) - h_j(y)| \le \sum_{k=1}^{p} H_{jk}|x_k - y_k|. \tag{3}$$

Write $H = (H_{jk})_{p\times p}$ *and* $\|H\|_\infty = \max_{1\le j\le p}\sum_{k=1}^{p} H_{jk}$*. Assume there exists a constant* $0 < \rho < 1$ *such that* $\|H\|_\infty \le \rho$*.*

The above assumption requires componentwise Lipschitz condition for nonlinear VAR processes. If Assumption 1 fails with $\|H\|_\infty = 1$, then X_i may not have a stationary solution. A prominent example is the random walk $X_i = X_{i-1} + \epsilon_i$ which has $\|H\|_\infty = 1$. This assumption can be easily extended to nonlinear VAR(d) processes. See also Chen and Tsay (1993), Diaconis and Freedman (1999), Jarner and Tweedie (2001), Shao and Wu (2007), Fan and Yao (2008) and Chen and Wu (2016) for nonlinear autoregressive processes. Intuitively, ρ quantifies the strength of dependence. For example, in one dimensional AR(1) model, $X_i = \rho X_{i-1} + \epsilon_i$. Larger ρ suggests stronger dependence.

Assumption 2 *For i.i.d. random vectors* $\epsilon_i \in \mathbb{R}^p$*,* $i \in \mathbb{Z}$*, assume*

(i) (finite moment) $\mu_q := \max_{1\le j\le p}\|\epsilon_{ij}\|_q < \infty$ *for some* $q \ge 2$*.*

(ii) (exponential moment) $\mu_e := \max_{1\le j\le p}\mathbb{E}\big(\exp(c_0|\epsilon_{ij}|)\big)$*, for some* $c_0 > 0$*.*

We first consider the finite moment case of the error vectors ϵ_i (cf. Assumption 2(i)). The following theorem provides a Bernstein-type inequality for bounded Lipschitz continuous functions.

Theorem 1 *Assume that function* $g : \mathbb{R}^p \to \mathbb{R}$*, is Lipschitz continuous with* $|g(x) - g(y)| \le \sum_{i=1}^{p} G_i|x_i - y_i|$*, for any* $x = (x_1, ..., x_p)^\top, y = (y_1, ..., y_p)^\top \in \mathbb{R}^p$*, where* G_j *are Lipschitz coefficients. Denote* $G = (G_1, ..., G_p)^\top$ *and* $\tau := |G|_1 = \sum_{j=1}^{p} G_j$*. Further assume* g *is bounded with* $|g|_\infty \le M$*. For the VAR process* (2)*, under Assumption 1 and Assumption 2(i), we have, for all* $z \ge 0$*,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\big(g(X_i) - \mathbb{E}g(X_i)\big)\Big| \ge z\Big) \le 2e^{-\frac{z^2}{c_1\tau^2 n + c_2\tau M z}}, \tag{4}$$

where c_1 *and* c_2 *are positive constants only depending on* q*,* ρ *and* μ_q*.*

Based on the proof of Theorem 1, we can have the explicit form for coefficients c_1 and c_2 as $c_1 = 32e^2(-\rho^2\log\rho)^{-2}\mu_2^2$ and $c_2 = 8e(-\rho^2\log\rho)^{-1}$. If function g is bounded by an absolute constant, then we can simplify above tail inequality and obtain the following Hoeffding type inequality.

Corollary 2 *If g is bounded with $|g|_\infty \leq 1$, then we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\big(g(X_i) - \mathbb{E}g(X_i)\big)\Big| \geq z\Big) \leq 2e^{-c_1 z^2/(\tau^2 n)}, \tag{5}$$

where c_1 is a positive constant depending only on q, ρ and μ_q.

Remark 3 *Note that up to a multiplicative constant, our Bernstein-type inequality* (4) *coincides with classical Bernsteins inequality for i.i.d. random variables. Thus one can expect sharper convergence rates for estimators of such processes. We remark that majority of the previous inequalities for temporal dependent processes do not recover Bernstein's inequality. For example, under geometric moment contraction with decay coefficient $0 < \rho < 1$ (see Wu and Shao (2004)) and assume $|X_i| \leq M$, Zhang (2021) provided the following Bernstein-type inequality,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\big(X_i - \mathbb{E}X_i\big)\Big| \geq z\Big) \leq \exp\left\{-\frac{z^2}{4c_1(c_3 n + M^2) + 2c_2 M(\log(n))^2 z}\right\},$$

where c_1, c_2 are some constants only depending on ρ, and $c_3 < \infty$ is a positive constant measuring the temporal dependence. Similarly, Merlevède et al. (2009) obtained a Bernstein-type inequality for a class of exponentially decay α-mixing and bounded random variables,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\big(X_i - \mathbb{E}X_i\big)\Big| \geq z\Big) \leq \exp\left\{-\frac{c_1 z^2}{nM^2 + M\log(n)\mathrm{loglog}(n)z}\right\},$$

where $c_1 > 0$ and $|X_i| \leq M$. Both involve an extra $\log(n)$ factor. Our sharp Bernstein-type inequality is of independent interest. We expect our sharp inequality can be useful for other high dimensional linear and nonlinear time series problems.

Proof (Proofs of Theorem 1) Without loss of generality, assume $|G|_1 = 1$. Recall $\mathcal{F}_k = (\ldots, \epsilon_{k-1}, \epsilon_k)$ and the projection operator $P_k(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_k) - \mathbb{E}(\cdot|\mathcal{F}_{k-1})$, $k \in \mathbb{Z}$. For $X_i = \mathcal{H}(\ldots, \epsilon_{i-1}, \epsilon_i)$, where $\mathcal{H}$ is some measurable function, we define the coupled version

$$X_{i,\{k\}} = \mathcal{H}(\ldots, \epsilon_{k-1}, \epsilon_k', \epsilon_{k+1}, \ldots, \epsilon_i).$$

For $x = (x_1, \ldots, x_p)^\top$, write $\mathrm{abs}(x) = (|x_1|, \ldots, |x_p|)^\top$. Since g and h_j are both Lipschitz continuous,

$$\begin{aligned}
|P_k g(X_i)| &= \big|\mathbb{E}(g(X_i) - g(X_{i,\{k\}})|\mathcal{F}_k)\big| \\
&\leq \mathbb{E}\left(G^\top \mathrm{abs}(X_i - X_{i,\{k\}})\Big|\mathcal{F}_k\right) \\
&\leq \mathbb{E}\left(G^\top H^{i-k}\mathrm{abs}(\epsilon_k - \epsilon_k')\Big|\mathcal{F}_k\right). \qquad (6)
\end{aligned}$$

Let $S_n(g) = \sum_{i=1}^n \big(g(X_i) - \mathbb{E}g(X_i)\big)$. For $k \le n$, denote $\xi_k = P_k(S_n(g))$. Then $S_n(g) = \sum_{k\le n} \xi_k$. Note we have

$$\mathbb{P}(S_n(g) \ge 2z) \le \mathbb{P}\Big(\sum_{-n<k\le n} \xi_k \ge z\Big) + \mathbb{P}\Big(\sum_{k\le -n} \xi_k \ge z\Big) =: \mathrm{I}_1 + \mathrm{I}_2.$$

By Assumption 1 and $|G|_1 \le 1$, $|H^{i-k\top}G|_1 \le \|H\|_\infty^{i-k}|G|_1 \le \rho^{i-k}$. Denote $v_i = H^{i\top}G$. Since $|g|_\infty \le M$, $|P_k g(X_i)| \le 2M$. Thus by (6) we have

$$|\xi_k| \le \sum_{i=1}^n |P_k g(X_i)| \le \sum_{i=k\vee 1}^n \min\Big\{ v_{i-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')|\mathcal{F}_k), 2M\Big\}, \quad \text{with} \quad |v_i|_1 \le \rho^i. \tag{7}$$

For I_1, let $h^* := -\rho^2(\log\rho)/(4eM)$. By Lemma 10 and (7) for any $0 < h \le h^*$, $\mathbb{E}(e^{|\xi_k|h}) < \infty$. Note that $\mathbb{E}(\xi_k|\mathcal{F}_{k-1}) = 0$. Then

$$\begin{aligned}\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) &= 1 + \mathbb{E}(e^{\xi_k h} - \xi_k h - 1|\mathcal{F}_{k-1}) \\ &\le 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big]h^2, \end{aligned}\tag{8}$$

in view of $e^x - x \le e^{|x|} - |x|$ for any x. Note that for any fixed $x > 0$, $(e^{tx} - tx - 1)/t^2$ is increasing in $t \in (0,\infty)$. By Lemma 10,

$$\mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big] \le \mathbb{E}\Big[\frac{e^{|\xi_k|h^*} - |\xi_k|h^* - 1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big] \le (h^*)^{-2}\mu_2^2(2M)^{-2} \le c_3 < \infty, \tag{9}$$

where $c_3 = 4e^2(-\rho^2\log\rho)^{-2}\mu_2^2$. Hence for any $h \le h^*$, by (8) and (9),

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \le 1 + c_3 h^2. \tag{10}$$

By Markov's inequality we have $\mathrm{I}_1 \le e^{-zh}\mathbb{E}[\exp(\sum_{-n<k\le n}\xi_k h)]$. Let $h = \min\{z(4c_3 n)^{-1}, h^*\}$, then by recursively applying (10),

$$\begin{aligned}\mathrm{I}_1 \le& e^{-zh}\mathbb{E}\Big(e^{\sum_{k=-n+1}^{n-1}\xi_k h}\mathbb{E}(e^{\xi_n h}|\mathcal{F}_{n-1})\Big) \\ \le& e^{-zh}(1 + c_3h^2)^{2n} \\ \le& \exp\big(-zh + 2nc_3h^2\big) \\ \le& \exp\Big\{-\frac{z^2}{8c_3 n + c_4 M z}\Big\}, \end{aligned}\tag{11}$$

where the third inequality is due to $1 + x \le e^x$ for $x > 0$, and $c_4 = 8e/(-\rho^2\log\rho)$.

For I_2, by (7), $\|\xi_k\|_q \le \sum_{i=1}^n \rho^{i-k}\mu_q \le \rho^{1-k}(1-\rho)^{-1}\mu_q$, for $k \le 0$. Then by Lemma 9,

$$\begin{aligned}\mathrm{I}_2 &\le z^{-q}\Big((q-1)\sum_{k\le -n}\|\xi_k\|_q^2\Big)^{q/2} \\ &\le (q-1)^{q/2}z^{-q}\Big(\sum_{k\le -n}\|\xi_k\|_q^2\Big)^{q/2} \\ &\le c_5\rho^{qn}/z^q = c_5 e^{-qn\log(\rho^{-1})}/z^q, \end{aligned}\tag{12}$$

where $c_5 = (q-1)^{q/2}\mu_q^q(1-\rho)^{-3q/2}$ only depends on ρ, q and μ_q.

Combining I_1 and I_2 parts, the desired result follows by noticing $z \le 2Mn$. ■

If the error vectors $\epsilon_i, i \in \mathbb{Z}$, satisfy stronger moment condition than the existence of finite qth moment, we expect that a stronger form than (4) exist. Indeed, when ϵ_i has subexponenial tail (Asumption 2(ii)), we are able to obtain an improved Bernstein-type inequality. Different from the above Theorem 1, in the following Theorem 4, function g can be unbounded.

Theorem 4 *Assume that function $g : \mathbb{R}^p \to \mathbb{R}$, is Lipschitz continuous with $|g(x)-g(y)| \le \sum_{i=1}^p G_i|x_i - y_i|$, for any $x, y \in \mathbb{R}^p$. Denote $G = (G_1, ..., G_p)^\top$ and $\tau := |G|_1$. For the VAR process (2), under Assumption 1 and Assumption 2(ii), we have, for all $z \ge 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n \big(g(X_i) - \mathbb{E}g(X_i)\big)\Big| \ge z\Big) \le 2e^{-\frac{z^2}{c_1\tau^2 n + c_2\tau z}}, \tag{13}$$

where c_1 and c_2 are positive constants only depending on ρ and μ_e.

Proof (Proof of Theorem 4) Without loss of generality, assume $|G|_1 = 1$. Similar to the proof of Theorem 1, let $S_n(g) = \sum_{i=1}^n \big(g(X_i) - \mathbb{E}g(X_i)\big)$, and $\xi_k = P_k(S_n(g))$. Then $S_n(g) = \sum_{k\le n}\xi_k$, and

$$\mathbb{P}(S_n(g) \ge 2z) \le \mathbb{P}\Big(\sum_{-n<k\le n}\xi_k \ge z\Big) + \mathbb{P}\Big(\sum_{k\le -n}\xi_k \ge z\Big) =: \mathrm{I}_1 + \mathrm{I}_2.$$

Denote $v_i = H^{i\top}G$ and $\omega_k = \sum_{i=1\vee k}^n v_{i-k}$. Since (6) still holds, we have

$$|\xi_k| \le \sum_{i=k\vee 1}^n v_{i-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')\big|\mathcal{F}_k) = \omega_k^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')\big|\mathcal{F}_k). \tag{14}$$

For I_2, $k \le -n$, $|w_k|_1 \le \rho^{1-k}/(1-\rho)$. Let $h^* := c_0(1-\rho)/\rho$. By (8) and (9), for any $0 \le h \le h^*$,

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \le 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h^*} - |\xi_k|h^* - 1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big]h^2 \le 1 + \frac{\mathbb{E}(e^{|\xi_k|h^*} - 1|\mathcal{F}_{k-1})}{h^{*2}}h^2. \tag{15}$$

Let $a_k = \rho^{1-k}/(1-\rho)$ and $u_k = w_k/a_k$, then

$$\mathbb{E}(e^{|\xi_k|h^*} - 1|\mathcal{F}_{k-1}) \le \mathbb{E}\Big(e^{w_k^\top \mathrm{abs}(\epsilon_k - \epsilon_k')h^*} - 1\Big) = \mathbb{E}\Big(e^{c_0 u_k^\top \mathrm{abs}(\epsilon_k - \epsilon_k')\rho^{-k}} - 1\Big).$$

If $f(0) = 0$, then $E(f(X)) = \int_0^\infty f'(t)\mathbb{P}(X \ge t)\mathrm{d}t$. Therefore we further obtain

$$\begin{aligned}\mathbb{E}(e^{|\xi_k|h^*} - 1|\mathcal{F}_{k-1}) &\le \int_0^\infty e^{t\rho^{-k}}\rho^{-k}\mathbb{P}(c_0 u_k^\top \mathrm{abs}(\epsilon_k - \epsilon_k') \ge t)\mathrm{d}t \\ &\le \rho^{-k}\int_0^\infty e^{-t(1-\rho^{-k})}\mu_e^2\mathrm{d}t \le \rho^{-k}(1-\rho)^{-1}\mu_e^2.\end{aligned} \tag{16}$$

Since $1 + x \leq e^x$, by (15) and (16),

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \leq 1 + \rho^{-k}(1-\rho)^{-1}\mu_e^2(h^*)^{-2}h^2 \leq e^{c_3\rho^{-k}h^2}, \tag{17}$$

where $c_3 = \mu_e^2(1-\rho)^{-3}\rho^2 c_0^{-2}$. Recursively applying (17), we can obtain

$$\mathrm{I}_2 \leq e^{-zh^*}\mathbb{E}\Big(e^{\sum_{k\leq -n}\xi_k h^*}\Big) \leq \exp(-zh^* + c_4\rho^n h^{*2}),$$

where $c_4 = c_3/(1-\rho)$. Similar to (11), we can bound the I_1 part and we complete the proof. ■

It should be emphasized that our Bernstein-type concentration inequalities are sharp, and does not contain any unpleasant extra logarithmic terms. These inequalities are useful for handling non-Gaussian VAR problems. They also suggest an interesting phenomenon, that the effect of dependence is captured by the multiplicative constants in the tail bounds.

3. Sparse additive nonlinear VAR models

3.1 The model

Assume that we are given observed time series data $X_1, ..., X_n \in \mathbb{R}^p$ sampled from a dynamical system comprising p variables. We are interested in inferring direct influences of a variable j on other variables $k \neq j$, $1 \leq k \leq p$. For example, in linear VAR models, $X_i = GX_{i-1} + \epsilon_i$, where G is $p \times p$ coefficient matrix. The set of influences among variables can be captured by a network matrix $A = (a_{jk})$ of size $p \times p$ for which each coefficient $a_{jk} = 1$ if variables k influences the variable j, and 0 otherwise. For simplicity, we assume that a first-order stationary model is adequate to encode the temporal dependence of the system. In other words, we consider nonlinear VAR model (2),

$$X_i = h(X_{i-1}) + \epsilon_i,$$

where the dynamics is captured by a possibly nonlinear function h.

Linear VAR models ($h(X_{i-1}) = GX_{i-1}$) or other parametric models explicitly involve a matrix that can be interpreted as network matrix, and its estimation (also possibly sparse) can be directly accomplished, see, for example, Basu and Michailidis (2015) and Hall et al. (2018). However, for nonlinear models, this becomes a more involved task and estimation of function h can be very challenge. In this work, we propose to use a new class of high dimensional sparse additive non-parametric VAR models. Specifically, we assume an additive model for each function h_j:

$$h_j(x) = \sum_{k=1}^{p} h_{jk}(x_k), \tag{18}$$

where $h_{jk} : \mathbb{R} \to \mathbb{R}, 1 \leq j, k \leq p$, $x = (x_1, ..., x_p)^\top \in \mathbb{R}^p$. Then, each function h_{jk} gives a score to the potential influence of variable k on variable j. Our strategy is to use the empirical mean of the estimated functions h_{jk} in terms of the data $X_1, ..., X_n$:

$$\frac{1}{n-1}\sum_{i=2}^{n}\Big|\hat{h}_{jk}(X_{i-1})\Big|,$$

where $\hat{h}_{jk}$ is estimated by the penalized least squares procedure (19). To provide a final estimate of A, these coefficients can be sorted and thresholded in some way.

Let Π denote the distribution of X_i and let Π_k denote the marginal distribution of $X_{i,k}$ for each $1 \leq k \leq p$. Denote $L_2(\Pi_k)$ norm of h_{jk} by

$$\|h_{jk}\|_{\Pi_k,2} = \sqrt{\int h_{jk}^2(x)\mathrm{d}\Pi_k(x)} = \sqrt{\mathbb{E}h_{jk}(X_{i,k})^2}.$$

Our estimator in the population level is given by the following penalized least squares problem:

$$(\hat{h}_{jk}, 1 \leq j,k \leq p) := \underset{h_{jk}\in\mathcal{H}_{jk}, 1\leq j,k\leq p}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f(X_i - h(X_{i-1})) + \lambda \sum_{j,k=1}^{p} \|h_{jk}\|_{\Pi_k,2} \right\}, \quad (19)$$

where h is some additive function in (18), $\mathcal{H}_{jk}$ is a suitable class of functions, f is the loss function. Empirical version of $\|h_{jk}\|_{\Pi_k,2}$ is given by $(n^{-1}\sum_{i=1}^{n} h_{jk}^2(X_{i-1,k}))^{1/2}$. In our analysis, we let f be the ℓ_2 loss function.

Assumption 3 (Basis function) *For each $1 \leq j,k \leq p$, let $(\psi_{j,k,l} : l = 1,2,...)$ be an orthonormal basis such that $\sup_x |\psi_{j,k,l}(x)| \leq B$. Assume that our function class H_{jk} satisfies*

$$H_{jk} = \left\{ h_{jk} : h_{jk}(\cdot) = \sum_{l=1}^{\infty} b_{jkl}^{*}\psi_{j,k,l}(\cdot), \quad \sum_{l=1}^{\infty} b_{jkl}^{*2} l^{2\beta} \leq C^2 \right\}, \quad 1 \leq j,k \leq p,$$

for some $0 < C < \infty$, $\beta \geq 1$.

Note that it implies that $\sum_{l=L+1}^{\infty} b_{jkl}^{*2} \leq C^2 L^{-2\beta}$. This condition corresponds to the functional class condition in Ravikumar et al. (2009). Such condition is standard, commonly imposed for basis expansion. Order β identifies the level of smoothness in the Sobolev space. It is also possible to set adaptive β for different j,k, although we do not pursue that direction here.

Let $L = L_n$ be a truncation parameter and $h_{jk}^{(L)}$ be an approximation of h_{jk} satisfying

$$h_{jk}^{(L)}(\cdot) = \sum_{l=1}^{L} b_{jkl}^{*}\psi_{j,k,l}(\cdot).$$

In this setting, $h_{jk}^{(L)}$ can be thought as the projection onto the truncated set of basis functions $\{\psi_{j,k,1}, ..., \psi_{j,k,L}\}$. Then, for $1 \leq j,k \leq p$,

$$X_{i,j} = \sum_{k=1}^{p} h_{jk}^{(L)}(X_{i-1,k}) + r_{ij} + \epsilon_{ij}, \text{ where } r_{ij} = \sum_{k=1}^{p}[h_{jk}(X_{i-1,k}) - h_{jk}^{(L)}(X_{i-1,k})] \quad (20)$$

is the reminder term, representing the bias of the basis expansion.

Define the oracle coefficients in basis expansion and the design matrix as follows

$$\begin{aligned} b^*_{j,k,\cdot} &= (b^*_{j,k,1}, ..., b^*_{j,k,L})^\top, \\ b^*_{j,\cdot,\cdot} &= (b^{*\top}_{j,1,\cdot}, ..., b^{*\top}_{j,p,\cdot})^\top, \\ b^* &= (b^{*\top}_{1,\cdot,\cdot}, ..., b^{*\top}_{p,\cdot,\cdot})^\top, \\ \psi_{j,k,\cdot}(\cdot) &= (\psi_{j,k,1}(\cdot), ..., \psi_{j,k,L}(\cdot))^\top, \\ \psi_{j,\cdot,\cdot}(x) &= (\psi_{j,1,\cdot}(x_1), ..., \psi_{j,p,\cdot}(x_p))^\top, \end{aligned} \tag{21}$$

where $x = (x_1, ..., x_p)^\top \in \mathbb{R}^p$. Let $r_i = (r_{i1}, ..., r_{ip})^\top$. Then (20) can be rewritten as

$$\begin{aligned} X_i &= \begin{pmatrix} \psi_{1,\cdot,\cdot}(X_{i-1})^\top & 0 & 0 & \cdots & 0 \\ 0 & \psi_{2,\cdot,\cdot}(X_{i-1})^\top & 0 & \cdots & 0 \\ 0 & 0 & \psi_{3,\cdot,\cdot}(X_{i-1})^\top & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \psi_{p,\cdot,\cdot}(X_{i-1})^\top \end{pmatrix} b^* + r_i + \epsilon_i. \\ &:= \Psi(X_{i-1})^\top b^* + r_i + \epsilon_i. \end{aligned} \tag{22}$$

By (21), for vector $b = (b_{j,k,\cdot})_{1 \le j,k \le p}$ and $b_{j,k,\cdot} \in \mathbb{R}^L$, define the $(2, \alpha)$ group structure norm

$$|b|_{2,\alpha} := ||b_{j,k,\cdot}|_2|_\alpha = \left(\sum_{j,k=1}^{p} \left(\sum_{l=1}^{L} b^2_{j,k,l} \right)^{\alpha/2} \right)^{1/\alpha}, \tag{23}$$

where $\alpha \ge 1$. For instance, with the choice $\alpha = 1$, this norm corresponds to the regularizer that underlies the group Lasso. For $\alpha = \infty$,

$$|b|_{2,\infty} := ||b_{j,k,\cdot}|_2|_\infty = \max_{1 \le j,k \le p} \left(\sum_{l=1}^{L} b^2_{j,k,l} \right)^{1/2}.$$

Then the solution to the optimization problem (19) can be approximately estimated through

$$\hat{b} := \operatorname*{argmin}_b \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i - \Psi(X_{i-1})^\top b) + \lambda \sum_{j,k=1}^{p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\psi_{j,k,\cdot}(X_{i-1,k})^\top b_{j,k,\cdot})^2} \right\}. \tag{24}$$

Note that it can be viewed as a functional version of the group lasso. Standard convexity theory implies the existence of a minimizer. Using empirical norm

$$\|f\|_{\Pi_k,2,n} = \left(\frac{1}{n} \sum_{i=1}^{n} f^2(X_{i-1,k}) \right)^{1/2},$$

the minimizer (24) can be written as

$$\hat{b} := \operatorname*{argmin}_b \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i - \Psi(X_{i-1})^\top b) + \lambda \sum_{j,k=1}^{p} \|\psi^\top_{j,k,\cdot} b_{j,k,\cdot}\|_{\Pi_k,2,n} \right\}. \tag{25}$$

Lim et al. (2015) introduced operator-valued reproducing kernel-based VAR models. The advantage of our formulation is that it decouples smoothness and sparsity, and thus we are able to apply block coordinate descent algorithm (cf. Ravikumar et al. (2009)) to construct the estimator. In the following section, using the Bernstein-type inequalities developed in Section 2, we provide theoretical properties of our ℓ_1 regularized estimators by assuming that this particular smoother in (25) is being used.

3.2 Asymptotic properties

To facilitate the theoretical analysis, we impose the following assumptions on the functions h_{jk} $(1 \le j, k \le p)$ and the basis expansions. For a function $f : \mathbb{R}^d \to \mathbb{R}$, denote $\|f\|_2 := (\int_{\mathbb{R}^d} f^2(x)dx)^{1/2}$ and $\|f\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x)|$.

Assumption 4 *There exist constants $\phi_U, \phi_L > 0$, so that*

$$\lambda_{\min}\Big\{\mathbb{E}\Psi(X_{i-1})\Psi(X_{i-1})^\top\Big\} \ge \phi_L, \tag{26}$$

and

$$\max_{1\le j,k\le p} \lambda_{\max}\Big\{\mathbb{E}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top\Big\} \le \phi_U. \tag{27}$$

Assumption 4 is similar to Assumption 3.3 in Fan et al. (2016) on the basis functions. When in the population level $\mathbb{E}\Psi(X_{i-1})\Psi(X_{i-1})^\top$ is well-conditioned, we provide the following proposition in relation to the sample version of the minimum restricted eigenvalue and maximum eigenvalue. Note that $L \ll n$. Then the sample version of the maximum eigenvalue (27) can follow from the strong law of large numbers.

Proposition 5 *Suppose Assumptions 1 and 2(ii) hold. Assume* $\sup_x |\psi_{j,k,l}(x)| \le B$ *for any* $1 \le j, k \le p, 1 \le l \le L$.

(i). Assume that (26) *hold and that for some constant* $c > 0$, *for all* $w \in \mathbb{R}^{p^2L}$,

$$\mathbb{E}(w^\top\Psi(X_i)\Psi(X_i)^\top w)^2 \le c\big(w^\top\mathbb{E}(\Psi(X_i)\Psi(X_i)^\top)w\big)^2.$$

Then, with probability at least $1 - p^{-c_1} - p^2e^{-c_2n/\log(n)}$, *for all* $u \in \mathbb{R}^{pL}$ *with* $|u|_2 = 1$,

$$\min_{1\le j\le p} \frac{1}{n}\sum_{i=1}^n u^\top\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top u \ge \frac{\phi_L}{2} - \frac{1}{n} - c_3\frac{\log(n)\log(pL)\cdot|u|_1^2}{n}, \tag{28}$$

where $c_1, c_2, c_3 > 0$ *are constants independent of* n, p, L.

(ii). Assume that (27) *hold. Then, with probability at least* $1 - p^{-c_4} - e^{-c_5n/\log(n)}$, *for all* $u \in \mathbb{R}^L$ *with* $|u|_2 = 1$,

$$\max_{1\le j,k\le p} \frac{1}{n}\sum_{i=1}^n u^\top\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top u \le \phi_U + c_6L\sqrt{\frac{\log(n)(\log p + \log L)}{n}}, \tag{29}$$

where $c_4, c_5, c_6 > 0$ *are constants independent of* n, p, L.

Assumption 5 *Let* $S := \{(j,k) : h_{jk} \not\equiv 0, 1 \le j,k \le p\}$ *and* $S_j := \{k : h_{jk} \not\equiv 0, 1 \le k \le p\}$, $1 \le j \le p$. *Assume that nonzero indices* $s_0 := \max_{1\le j\le p} \sum_{k=1}^{p} \mathbf{1}_{\{h_{jk}\not\equiv 0\}} = \max_{1\le j\le p} Card(S_j) = o(p)$, *and* $s := \sum_{j,k=1}^{p} \mathbf{1}_{\{h_{jk}\not\equiv 0\}} = Card(S) = o(p^2)$.

Assumption 5 imposes a sparsity condition on the nonlinear functions. Structural sparsity condition is often used in high dimensional setting, for example, Cai and Liu (2011) in covariance matrix estimation.

The following Proposition 6 provides an upper bound of the reminder part $|r_i|_\infty$ in terms of smoothness level β, the number of the basis functions L, and sparsity level s_0.

Proposition 6 *Under Assumptions 3 and 5, we have*

$$|r_i|_\infty \le BC(2\beta-1)^{-1} s_0 L^{1/2-\beta}.$$

Formally, we have the following asymptotic properties for the ℓ_1 regularized estimators. Theorem 7 shows how the rate of convergence of $\hat{b} - b^*$ and the errors of the estimated functions $\hat{h}_{jk}$ depend on the sparsity of functions, basis expansions, the dependence strength of the processes and the moment condition.

Theorem 7 *Suppose Assumptions 1, 2(ii), 3, 4 and 5 hold. Let* $\hat{b}$ *be the corresponding LASSO solution given in the optimization problem* (24). *Consider the estimator*

$$\hat{h}_{jk}(x) = \sum_{l=1}^{L} \psi_{j,k,l}(x)\hat{b}_{j,k,l}, \qquad 1 \le j,k \le p. \tag{30}$$

Assume that there exists a constant $c_1 > 0$, *such that for all* $u \in \mathbb{R}^{p^2 L}$,

$$\mathbb{E}(u^\top \Psi(X_i)\Psi(X_i)^\top u)^2 \le c_1\big(u^\top \mathbb{E}(\Psi(X_i)\Psi(X_i)^\top)u\big)^2.$$

Assume that

$$\lambda \ge c_2\Big(\sqrt{\frac{L\log(pL)}{n}} + s_0 L^{1-\beta}\Big), \tag{31}$$

for some $c_2 > 0$. *Also suppose that*

$$n \ge c_3 s_0 L \cdot \log(n)\log(pL) + c_3 L^2 \cdot \log(n)\log(pL)$$

for some sufficiently large constant c_3. *We have, with probability approaching one (as* $n, p \to \infty$*),*

$$|\hat{b} - b^*|_2 \le c_4\sqrt{s}\lambda, \tag{32}$$

$$\sum_{j=1}^{p}\sum_{k=1}^{p} \|\hat{h}_{jk} - h_{jk}\|_2^2 \le c_5 s\lambda^2 + c_5 s L^{-2\beta}, \tag{33}$$

where $c_2, c_3 > 0$ *are constants depending on* ρ *and* μ_e.

Note that as $s \leq s_0 p$, (32) and (33) imply that

$$\max_{1\leq j\leq p} |\hat{b}_{j,\cdot,\cdot} - b^*_{j,\cdot,\cdot}|_2 \leq c_4\sqrt{s_0}\lambda,$$
$$\max_{1\leq j\leq p} \sum_{k=1}^{p} \|\hat{h}_{jk} - h_{jk}\|_2^2 \leq c_5 s_0\lambda^2 + c_5 s_0 L^{-2\beta},$$

where b^* and $b^*_{j,\cdot,\cdot}$ is defined in (21) and (22). The quantity ρ indicate the strength of dependence of the processes, and the constant μ_e correspond to the moment condition. Theorem 7 indicates the dependence measures of the processes and the moment condition do not affect the convergence rate if both Assumptions 1 and 2(ii) hold with $\rho \leq \rho_0 < 1$ and ρ_0 is a constant. Besides, the second term in (33) reveals the bias of basis expansion in the estimated functions. Theorem 7 implies that p can be as large as e^{n^c} for some constant $0 < c < 1$ if ϵ_i has finite exponential moments, and the power constant c is related to the truncated number L of basis expansion.

It is interesting to compare the two terms in the requirement of λ (31). In the case with relative low dimensional $\log(p) \lesssim s_0^2 n L^{1-2\beta}$ and low basis number $L \lesssim s_0^{2/(2\beta-1)}(n/\log n)^{1/(2\beta-1)}$, the part of basis expansion bias, which corresponds to $s_0 L^{1-\beta}$, dominates. On the other hand, if the dimension p is large such that $\log(p) \gtrsim s_0^2 n L^{1-2\beta}$ or basis number L is large with $L \gtrsim s_0^{2/(2\beta-1)}(n/\log n)^{1/(2\beta-1)}$, then the dominating term is the first part $(n^{-1}L\log(pL))^{1/2}$.

The setting in our Theorem 7 is very general as it allows a wide class of non sub-Gaussian nonlinear VAR processes. Han et al. (2015) and Basu and Michailidis (2015) considered the special case of the estimation of transition matrices of linear VAR model under the assumption that errors ϵ_i are i.i.d. Gaussian. Our setting also allows a large number of parameters in the context that the dimension p can be much larger than sample size n. Moreover, (24) leads to sparse solution $\hat{b}$, that is $\hat{b}_{j,k,\cdot} = 0$ for some $1 \leq j, k \leq p$. By checking non-zero vectors of $\hat{b}_{j,k,\cdot} = 0$, $1 \leq j, k \leq p$, we can construct the network matrix A. A theory-free principle was advocated in Sims (1980) for inferring economic relations between variables of linear VARs. Theorem 8 provides theoretical guarantee for model selection consistency.

Instead of Assumptions 4, we shall consider model selection consistency under the following condition. To simplify the notation, let $\Psi_{S_j}(X_i) = (\psi_{j,k,\cdot}(X_{i,k})^\top, k \in S_j)$, be a vector in $\mathbb{R}^{L\cdot \mathrm{Card}(S_j)}$, where $\psi_{j,k,\cdot}$ is defined in (21). Denote

$$\Psi_S(X_i) = \begin{pmatrix} \Psi_{S_1}(X_i)^\top & 0 & 0 & \cdots & 0 \\ 0 & \Psi_{S_2}(X_i)^\top & 0 & \cdots & 0 \\ 0 & 0 & \Psi_{S_3}(X_i)^\top & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Psi_{S_p}(X_i)^\top \end{pmatrix}.$$

Assumption 6 *There are some constants $\phi_{\max}, \phi_{\min} > 0, 0 < \delta \leq 1$, so that with probability approaching one (as $n, p \to \infty$), we have*

$$\lambda_{\min}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\Psi_S(X_{i-1})^\top\Big\} \geq \phi_{\min} > 0, \tag{34}$$

$$\lambda_{\max}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\Psi_S(X_{i-1})^\top\Big\} \leq \phi_{\max} < \infty, \tag{35}$$

and

$$\max_{1\leq j\leq p}\max_{k\in S_j^c}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\Psi_{S_j}(X_{i-1})^\top\right)\left(\frac{1}{n}\sum_{l=1}^{n}\Psi_{S_j}(X_{l-1})\Psi_{S_j}(X_{l-1})^\top\right)^{-1}\right\|_2 \leq \sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}. \tag{36}$$

This assumption corresponds to the condition of Ravikumar et al. (2009). Similar to Assumption 4, (34) and (35) are also standard, and are commonly imposed for high-dimensional regression analysis. Besides, (36) relates to the incoherence condition, see e.g. Wainwright (2009).

In Theorem 8, we show that, under certain conditions, our method recovers the sparsity pattern asymptotically. Recall $S := \{(j,k) : h_{jk} \not\equiv 0, 1 \leq j,k \leq p\}$. Then $S = \{(j,k) : b^*_{j,k,\cdot} \neq 0, 1 \leq j,k \leq p\}$. Let $\hat{S}_n := \{(j,k) : \hat{b}_{j,k,\cdot} \neq 0, 1 \leq j,k \leq p\}$.

Theorem 8 *Suppose Assumptions 1, 2(ii), 3, 5 and 6 hold. Let $\hat{b}$ be the corresponding LASSO solution given in the optimization problem* (24)*. Let $\beta > 3/2$. Assume that*

$$\frac{s_0L^2\cdot\log(pL)}{n} + s_0L^{1-2\beta/3} \to 0, \tag{37}$$

and

$$\lambda\sqrt{s_0}L + \lambda^{-1}\sqrt{\frac{L\log(n)}{n}} + \lambda^{-1}s_0L^{1-\beta} \to 0. \tag{38}$$

Then the solution $\hat{b}$ to problem (24) *is unique and satisfies $\hat{S}_n = S$, with probability approaching one (as $n, p \to \infty$).*

We set elements of estimated network matrix $\hat{a}_{jk} = 1$ if $\hat{b}_{j,k,\cdot} \neq 0$ (ignoring the sign of $\hat{b}_{j,k,\cdot}$), otherwise, set $\hat{a}_{jk} = 0$. As the estimated network matrix $\hat{A} = (\hat{a}_{jk})$ is not symmetric, it is an adjacency matrix for a directed graph. Our Theorem 8 provides model selection consistency for the estimated network matrix $\hat{A}$, which is also proposed in section 3.1.

4. Simulation Studies

In this section, we shall evaluate the numerical performance of the proposed estimation procedures of nonlinear VAR models.

We design three different patterns of the binary transition matrix (network matrix, see Section 3.1) A: random, band, cluster. Typical realizations of these patterns are illustrated in Figure 1. The pattern "cluster" has block diagonal structure, where each block is of dimension 10×10 and satisfies the pattern "random". In each dimension j, $1 \le j \le p$, we randomly assign 5 nonzero functions, according to the pattern of the transition matrix. The relevant nonzero component functions are given by

$$\begin{aligned} f_1(x) &= 0.2x, \\ f_2(x) &= -0.15\sin(1.5x), \\ f_3(x) &= -0.5\Phi(x, 0.5, 1), \\ f_4(x) &= 0.2xe^{-0.5x^2}, \\ f_5(x) &= 0.15\log(|x| + 2), \end{aligned}$$

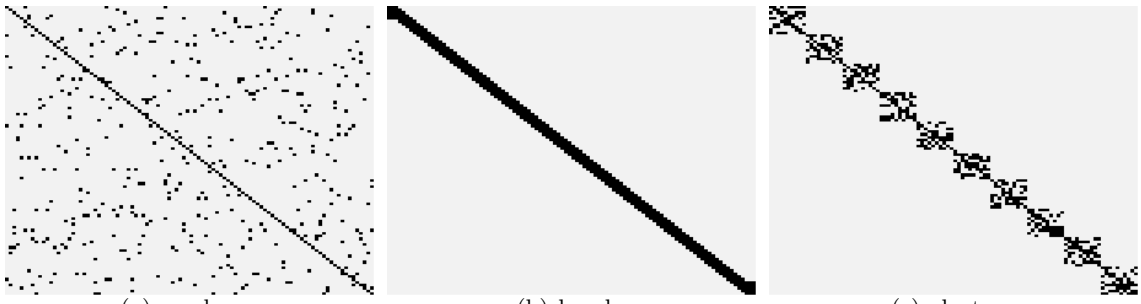
where $\Phi(\cdot, 0.5, 1)$ is the Gaussian probability distribution function with mean 0.5 and standard deviation 1. In other words, for each j with $1 \le j \le p$, we randomly select 5 functions h_{jk} $(1 \le k \le p)$ to be the above nonzero functions. The rest $p - 5$ functions of h_{jk} $(1 \le k \le p)$ are all zeros. Elementary calculation shows that this nonlinear VAR process is stable and satisfies Assumption 1. In order to ensure reasonable signal to noise ratio, the error processes ϵ_t are generated from $0.2N(0, 1)$.

In all the conducted experiments, we assess the model selection performance of our model using the area under the receiver operating characteristic curve (AUROC) and the area under the Precison-Recall curve (AUPR) ignoring the sign (positive negative influence), where the ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) and the precision-recall curve is a plot of the precision against the recall. Define TPR, FPR, precision and recall as follows

$$\text{TPR} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Here TP and TN stand for true positives and true negatives, respectively, and FP and FN stand for false positives/negatives. We choose a set of data dimensions $p = 20, 50, 100$ while the sample size is $n = 50, 100, 200, 500$, respectively. The empirical values reported in Tables 1 are averages over 1000 replications.

It can be seen from Table 1 that the proposed estimation procedure of nonlinear VAR model performs fairly well as reflected in both AUROC and AUPR. In particular, when the sample size is moderate ($n \ge 100$), our method provides pretty good AUROC in all cases. As expected, when the sample size n increases, our method performs better. And both AUROC and AUPR decreases as the dimension p increase. Besides, our proposed method makes no significant differences in terms of 3 patterns of transition matrix.



(a) random (b) band (c) cluster

Figure 1: Three different network matrix patterns used in the simulation studies. Here gray points represent the zero entries and black points represent nonzero entries.

Table 1: Model selection performance of the proposed nonlinear VAR method with three different patterns of the transition matrix, "random", "band", "cluster", based on 1000 replications.

p		AUROC				AUPR			
	n	50	100	200	500	50	100	200	500
	Pattern "random"								
20		0.633	0.744	0.851	0.924	0.443	0.651	0.856	0.937
50		0.611	0.720	0.842	0.920	0.230	0.458	0.753	0.904
100		0.591	0.696	0.830	0.918	0.132	0.320	0.666	0.883
	Pattern "band"								
20		0.647	0.753	0.858	0.928	0.469	0.681	0.864	0.938
50		0.610	0.720	0.841	0.920	0.234	0.464	0.758	0.905
100		0.592	0.698	0.830	0.918	0.143	0.339	0.672	0.881
	Pattern "cluster"								
20		0.642	0.746	0.855	0.922	0.464	0.667	0.861	0.933
50		0.609	0.718	0.839	0.920	0.231	0.454	0.744	0.905
100		0.591	0.696	0.827	0.918	0.138	0.328	0.661	0.883

5. Real Data Analysis

We now apply our nonlinear VAR model to the analysis of a real biological gene regulatory network time series expression data. The network is an *E. coli* SOS DNA repair system, which has been well studied in biology, see e.g, Ronen et al. (2002). The main function of the SOS signaling pathway is to regulate cellular immunity and repair DNA damage. We consider an eight gene network, part of the SOS DNA repair network in the bacteria *E. coli*. The time series gene expression data set of the network was collected by Ronen et al. (2002). The data are kinetics of 8 genes, that is, lexA, recA, ruvA, polB, umuDC, uvrA, uvrD, uvrY, where lexA and recA are the key genes in the pathway. The 8 genes were measured at 50 instants which are evenly spaced by 6 min intervals.

We compare the performance of our method with the Lasso regularized linear VAR method (Basu and Michailidis (2015)). The tuning parameter λ in both methods are chosen by time series cross-validation procedure (see Han et al. (2015)). Figure 2 represents the bacterial SOS DNA repair system. Figure 3 shows the real SOS DNA repair network, which contains 9 edges. Figures 4 and 5 show the inferred gene regulatory networks using our nonlinear VAR model and the ℓ_1 regularized linear VAR model, respectively. In Figure 4, one can see that our method finds 6 out of the 9 edges in the target network and identifies lexA as the hub gene for this network. Our method identifies most interactions except lexA→ruvA, lexA→uvrY and recA→lexA. In comparison, in the Figure 5, the ℓ_1 regularized linear VAR model recognizes only 4 out of the 9 true edges, and predicts a wrong edge. Furthermore, our proposed method gives the area under ROC curve 0.8116 and the area under Precison-Recall curve 0.6836. While, the ℓ_1 regularized linear VAR model gives AUROC 0.7222 and AUPR 0.6036. In summary, our proposed method has a better performance than the regularized linear VAR model on the SOS DNA repair network, although none of these two methods can faithfully recover all of the edges. This phenomenon also confirms that there exists nonlinear dynamics in the gene regulatory networks.

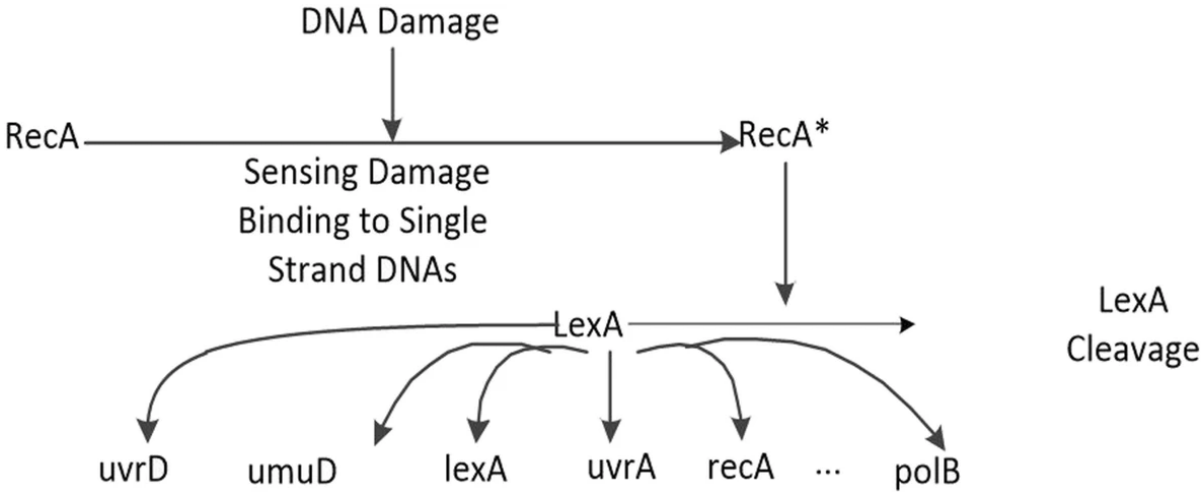


Figure 2: The bacterial SOS DNA repair system

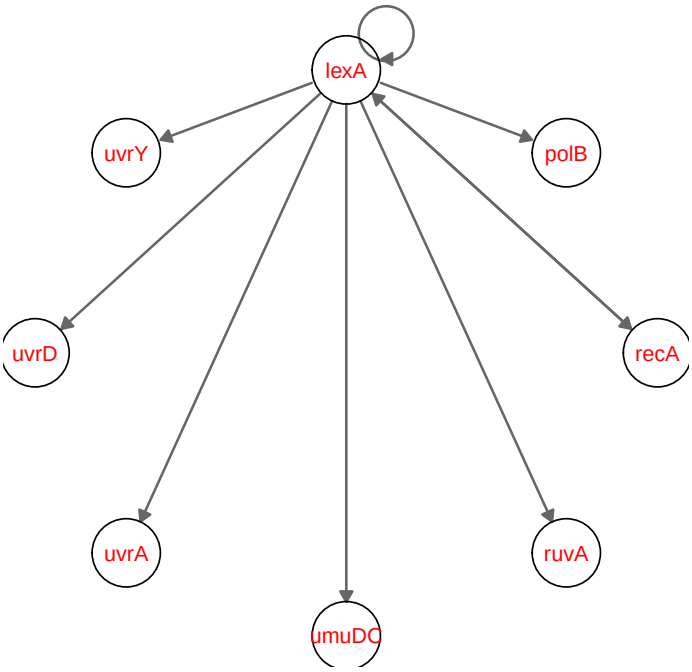


Figure 3: The target SOS DNA repair network

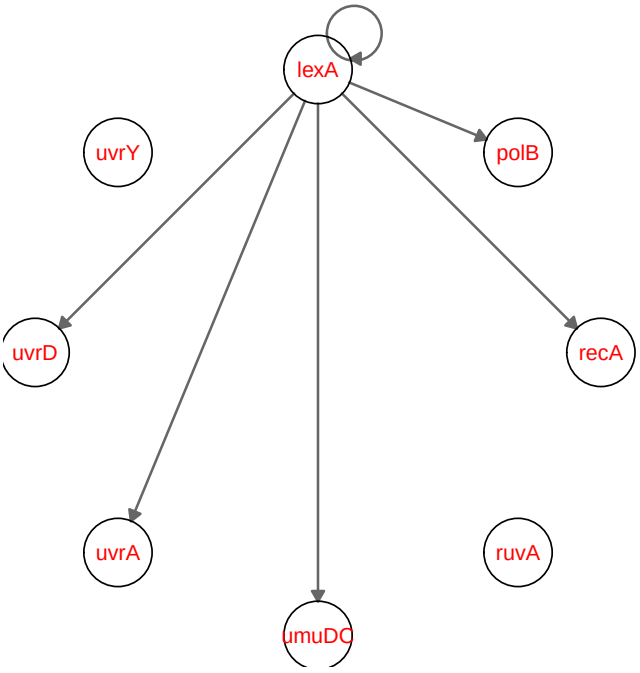


Figure 4: Reconstruction of SOS DNA repair network by nonlinear VAR model

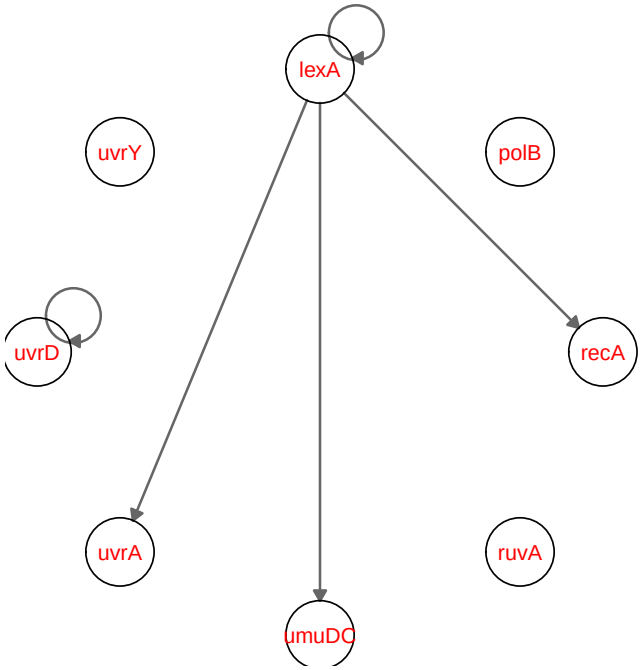


Figure 5: Reconstruction of SOS DNA repair network by linear VAR model

6. Appendix

Let $\mathcal{F}_k = (\ldots, \epsilon_{k-1}, \epsilon_k)$, $\mathcal{F}_k^n = \{\epsilon_k, \ldots, \epsilon_n\}$, and $\mathbb{E}_0(X) = X - \mathbb{E}X$. Define projection operator $P_k(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_k) - \mathbb{E}(\cdot|\mathcal{F}_{k-1})$, $k \in \mathbb{Z}$. Let $(\epsilon_k')_{k\in\mathbb{Z}}$ be an i.i.d. copy of $(\epsilon_k)_{k\in\mathbb{Z}}$. For any $X_i = \mathcal{H}(\ldots, \epsilon_{i-1}, \epsilon_i)$, where $\mathcal{H}$ is a measurable function, we define the coupled version $X_{i,\{k\}} = \mathcal{H}(\ldots, \epsilon_{k-1}, \epsilon_k', \epsilon_{k+1}, \ldots, \epsilon_i)$. If $k > i$, then $X_{i,\{k\}} = X_i$.

Lemma 9 (Burkholder (1988), Rio (2009)) *Let $q > 1$, $q' = \min\{q, 2\}$. Let $D_T = \sum_{t=1}^T \xi_t$, where $\xi_t \in \mathcal{L}^q$ are martingale differences. Then*

$$\|D_T\|_q^{q'} \le K_q^{q'} \sum_{t=1}^T \|\xi_t\|_q^{q'}, \text{ where } K_q = \max\left\{(q-1)^{-1}, \sqrt{q-1}\right\}.$$

Lemma 10 *Let $\epsilon \in \mathbb{R}^p$ be a random vector with non-negative entries, satisfying Assumption 2(i) with $\mu_q < \infty$, for some $q \ge 2$. For non-negative vectors $v_i \in \mathbb{R}^p$, assume $|v_i|_1 \le \rho^i$ where $\rho < 1$. Consider*

$$X := \sum_{i=0}^{\infty} \min\left\{v_i^\top \epsilon, M\right\}.$$

Take $c_0 = -\rho^2 \log\rho/(2e)$. Then for any $c \le c_0/M$, $\mathbb{E}(e^{cX})$ exists and

$$\mathbb{E}e^{c_0 X/M} - \mathbb{E}(c_0 X/M) - 1 \le \mu_2^2 M^{-2} < \infty.$$

Proof Note we have the decomposition

$$X = M\sum_{i=0}^{\infty}\mathbf{1}_{v_i^\top\epsilon\geq M} + \sum_{i=0}^{\infty}v_i^\top\epsilon\mathbf{1}_{v_i^\top\epsilon<M} =: \mathrm{I}_1 + \mathrm{I}_2.$$

For I_1 part, by Markov's inequality,

$$\mathbb{P}(v_i^\top\epsilon \geq M) \leq M^{-2}\|v_i^\top\epsilon\|_2^2 \leq \rho^{2i}\mu_2^2 M^{-2}.$$

Hence for $m \geq 1$, we have

$$\mathbb{E}|\mathrm{I}_1|^m \leq M^m\Big(\sum_{i=0}^{\infty}\mathbb{P}(v_i^\top\epsilon\geq M)^{1/m}\Big)^m \leq M^m\Big(\mu_2^{2/m}M^{-2/m}\sum_{i=0}^{\infty}\rho^{2i/m}\Big)^m \leq \mu_2^2(1-\rho^{2/m})^{-m}M^{m-2}.$$

Since for any $m \geq 1$,

$$1-\rho^{2/m} \geq (1-\rho^2)/m \geq -2\rho^2\log(\rho)/m, \tag{39}$$

we further obtain

$$\mathbb{E}|\mathrm{I}_1|^m \leq \mu_2^2(-2\rho^2\log(\rho)/m)^{-m}M^{m-2}.$$

Choose $c_{1,M} = -\rho^2\log(\rho)/(eM)$, then by $m! \geq (2\pi)^{1/2}m^{m+1/2}e^{-m}$ (Robbins (1955)), we have

$$\sum_{m\geq 2}\frac{\mathbb{E}((c_{1,M}\mathrm{I}_1)^m)}{m!} \leq \frac{1}{2}\mu_2^2M^{-2}.$$

For I_2 part, for any $m \geq 2$,

$$\begin{aligned}\mathbb{E}|\mathrm{I}_2|^m &\leq \left(\sum_{i=0}^{\infty}\|v_i^\top\epsilon\mathbf{1}_{v_i^\top\epsilon<M}\|_m\right)^m \leq \left(\sum_{i=0}^{\infty}(M^{m-2}\mathbb{E}|v_i^\top\epsilon|^2)^{1/m}\right)^m \leq \mu_2^2\Big(M^{1-2/m}\sum_{i=0}^{\infty}\rho^{iq/m}\Big)^m \\ &\leq \mu_2^2(-2\rho^2\log(\rho)/m)^{-m}M^{m-2},\end{aligned}$$

where the last inequality is by (39). Therefore

$$\sum_{m\geq 2}\frac{\mathbb{E}((c_{1,M}\mathrm{I}_2)^m)}{m!} \leq \frac{1}{2}\mu_2^2M^{-2} < \infty,$$

We complete the proof by combining the two parts and setting $c_0 = Mc_{1,M}/2$,

$$\mathbb{E}e^{c_0X/M} - 1 - \mathbb{E}(c_0X/M) = \sum_{m\geq 2}\frac{\mathbb{E}((c_0X/M)^m)}{m!} \leq \sum_{m\geq 2}\frac{\mathbb{E}((c_{1,M}\mathrm{I}_1)^m)}{m!} + \sum_{m\geq 2}\frac{\mathbb{E}((c_{1,M}\mathrm{I}_2)^m)}{m!} \leq \mu_2^2M^{-2}.$$

■

Proof [Proof of Proposition 6] Note that since basis functions are orthonormal, $\|h_{jk}\|_2 = (\sum_{l=1}^{\infty} b_{jkl}^{*2})^{1/2}$. Since basis functions are bounded by B, by Assumption 3, we have

$$\begin{aligned}\|h_{jk} - h_{jk}^{(L)}\|_\infty &\le \sum_{l\ge L+1} |b_{jkl}^*| B = B \sum_{l\ge L+1} \frac{|b_{jkl}^*| l^\beta}{l^\beta} \\ &\le B \sqrt{\sum_{l\ge L+1} b_{jkl}^{*2} l^{2\beta}} \sqrt{\sum_{l\ge L+1} l^{-2\beta}} \\ &\le BC(2\beta-1)^{-1} L^{1/2-\beta}.\end{aligned}$$

Hence, as $s_0 = \max_{1\le j\le p} \mathrm{Card}(S_j)$ with $S_j := \{k : h_{jk} \neq 0, 1 \le k \le p\}$,

$$|r_i|_\infty \le \sum_{k=1}^{p} \|h_{jk} - h_{jk}^{(L)}\|_\infty \le BC(2\beta-1)^{-1} s_0 L^{1/2-\beta}.$$

Then we obtain the desired result. ■

Proof [Proof of Proposition 5] We first prove part (i). By (26), we have, for any $u \in \mathbb{R}^{pL}$ with $|u|_2 = 1$,

$$\mathbb{E} u^\top \psi_{j,\cdot,\cdot}(X_i) \psi_{j,\cdot,\cdot}(X_i)^\top u \ge \phi_L.$$

Let $m = 4(-\log\rho)^{-1}\log(n)$. Recall $\mathcal{F}_k^n = \{\epsilon_k, \ldots, \epsilon_n\}$. By Lemma 11, we have, for any $1 \le j \le p$, with probability at least $1 - mp^{-c_1}/12 - 2mp^2 L e^{-3n/(10m)}$, for any $u \in \mathbb{R}^{pL}$,

$$\frac{1}{n}\sum_{i=1}^{n} u^\top \mathbb{E}\big(\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top | \mathcal{F}_{i-m+1}^n\big) u \ge \frac{1}{2} u^\top \mathbb{E}\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top u - \frac{c_2 \log(n)\log(pL)}{n} |u|_1^2.$$

Note that $L = o(n)$. Let $z = 1$ in Lemma 12, we can obtain, with probability at least $1 - mp^{-c_1}/12 - 2mp^2 L e^{-3n/(10m)} - e^{-c_3 n}$, for any $u \in \mathbb{R}^{pL}$,

$$\frac{1}{n}\sum_{i=1}^{n} u^\top \big(\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top\big) u \ge \frac{1}{2} u^\top \mathbb{E}\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top u - \frac{c_2 \log(n)\log(pL)}{n} |u|_1^2 - \frac{1}{n}|u|_2^2.$$

Then (28) follows.

For part (ii), denote $\Omega_{j,k} = \mathbb{E}(\psi_{j,k,\cdot}(X_{i,k})\psi_{j,k,\cdot}(X_{i,k})^\top)$. For $m = o(n)$, let $N = [(n-1)/m]$ and $\mathcal{N} = \{1, m+1, 2m+1, ..., (N-1)m+1\}$. Then there exists constant $c_3 > 0$ such that for any $1 \le l_1, l_2 \le L$, $z > 0$, we have

$$\mathbb{P}\left(\Big|\frac{1}{N}\sum_{i\in\mathcal{N}} \mathbb{E}\Big((\psi_{j,k,\cdot}(X_{i,k})\psi_{j,k,\cdot}(X_{i,k})^\top)_{l_1,l_2} | \mathcal{F}_{i-m+1}^n\Big) - \Omega_{j,k,l_1,l_2}\Big| \ge z\right) \le 2\exp\big\{-c_3 N z^2\big\}.$$

Therefore with probability at least $1 - 2L^2\exp\{-c_3 N z^2\}$, for any $u \in \mathbb{R}^L$ with $|u|_2 = 1$,

$$\Big|\frac{1}{N}\sum_{i\in\mathcal{N}} \mathbb{E}\Big(u^\top \psi_{j,k,\cdot}(X_{i,k})\psi_{j,k,\cdot}(X_{i,k})^\top u | \mathcal{F}_{i-m+1}^n\Big) - u^\top \Omega_{j,k} u\Big| \le Lz.$$

Take $z = c_4(\log(p) + \log(L))/N$ some constant c_4 large enough. Then we have with probability greater than $1 - m(pL)^{-c_4}$, for any $u \in \mathbb{R}^L$, $|u|_2 = 1$, $1 \le j, k \le p$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big(u^\top \psi_{j,k,\cdot}(X_{i,k})\psi_{j,k,\cdot}(X_{i,k})^\top u | \mathcal{F}_{i-m+1}^n \Big) \le \phi_U + c_5 L\sqrt{\frac{\log(p)+\log(L)}{N}}.$$

Then (29) follows by combining above and Lemma 12 with $z = 1$ and $m = 4(-\log\rho)^{-1}\log(n)$. ■

For $m = o(n)$, denote $N = [(n-1)/m]$ and $\mathcal{N} = \{1, m+1, 2m+1, ..., (N-1)m+1\}$.

Lemma 11 *Consider the VAR process* (2)*, suppose Assumptions 1 and 2(ii) hold. Assume that there exists constant $c > 0$, such that for all $u \in \mathbb{R}^{p^2L}$, $\mathbb{E}[(u^\top\Psi(X_i)\Psi(X_i)^\top u)^2] \le c\big(u^\top\mathbb{E}(\Psi(X_i)\Psi(X_i)^\top)u\big)^2$. Let $N \ge C\log(p^2L)$, where $C > 0$ is a sufficiently large constant. Then, we have, with probability at least $1 - p^{-c_1}/12 - 2p^2Le^{-3N/10}$,*

$$\forall u \in \mathbb{R}^{p^2L}, \frac{1}{N}\sum_{i\in\mathcal{N}} u^\top\mathbb{E}\big(\Psi(X_i)\Psi(X_i)^\top|\mathcal{F}_{i-m+1}^n\big)u \ge \frac{1}{2}u^\top\mathbb{E}\Psi(X_i)\Psi(X_i)^\top u - \frac{c_2\log(p^2L)}{N}|u|_1^2,$$

where $c_1 > 0$ is a sufficiently large constant and c_2 depends only on c and B.

Proof Recall for any $1 \le j, k \le p, 1 \le l \le L$, $\sup_x|\psi_{jkl}(x)| \le B$, some $B \ge 1$, and $\mathcal{F}_k^n = \{\epsilon_k, \ldots, \epsilon_n\}$. Denote $\Sigma = \mathbb{E}(\Psi(X_i)\Psi(X_i)^\top)$ and

$$\tilde{\Sigma}_N = N^{-1}\sum_{i\in\mathcal{N}}\mathbb{E}\big(\Psi(X_i)\Psi(X_i)^\top|\mathcal{F}_{i-m+1}^n\big).$$

Let $\tilde{\Sigma}_{\text{diag}}$ be the diagonal of $\tilde{\Sigma}_N$. Note that $\mathbb{E}\big(\Psi(X_i)\Psi(X_i)^\top|\mathcal{F}_{i-m+1}^n\big) = \mathbb{E}\big(\Psi(X_i)\Psi(X_i)^\top|\mathcal{F}_{i-m+1}^i\big)$ are independent for all $i \in \mathcal{N}$. By Jensen's inequality,

$$\mathbb{E}\left[\big(\mathbb{E}\big(u^\top\Psi(X_i)\Psi(X_i)^\top u|\mathcal{F}_{i-m+1}^n\big)\big)^2\right] \le \mathbb{E}[(u^\top\Psi(X_i)\Psi(X_i)^\top u)^2] \le c\big(u^\top\mathbb{E}(\Psi(X_i)\Psi(X_i)^\top)u\big)^2.$$

Then, employing similar arguments as in the proof of Lemmas 5.1 and 5.2 in Oliveira (2013), we can obtain, for $N \ge 1568c(c_3+1)\log(p^2L)$ and $c_3 > 0$,

$$\mathbb{P}\left(\forall u \in \mathbb{R}^{p^2L}, u^\top\tilde{\Sigma}_N u \ge \frac{1}{2}u^\top\Sigma u - \frac{1568c(c_3+1)\log(p^2L)}{N}\left|\tilde{\Sigma}_{\text{diag}}^{1/2}u\right|_1^2\right) \ge 1 - \frac{1}{12}p^{-c_3}. \quad (40)$$

Since for any $1 \le j, k \le p, 1 \le l \le L$, $|\psi_{jkl}|_\infty \le B$, then, by Bernstein's inequality, we have,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i\in\mathcal{N}}(\psi_{jkl}(X_{ik})^2 - \mathbb{E}\psi_{jkl}(X_{ik})^2)\right| \ge z\right) \le 2\exp\left(-\frac{Nz^2}{2B^4 + 4B^2z/3}\right).$$

Hence, we have

$$\mathbb{P}\left(\max_{1\le j,k\le p, 1\le l\le L}\left|\frac{1}{N}\sum_{i\in\mathcal{N}}\psi_{jkl}(X_{ik})^2\right| \ge 2B^2\right) \le 2p^2L\exp\left(-10N/3\right).$$

Combining the above inequality with (40), it follows that, with probability at least $1 - p^{-c_3}/12 - 2p^2Le^{-3N/10}$, for any $u \in \mathbb{R}^{p^2L}$,

$$u^\top \tilde{\Sigma}_N u \geq \frac{1}{2} u^\top \Sigma u - \frac{3136B^2c(c_3+1)\log(p^2L)}{N} |u|_1^2 .$$

■

Lemma 12 *(m-approximation) Considering the VAR process* (2)*, suppose Assumptions 1 and 2 (ii) hold. Let $z\rho^{-m}/(s_0L) > Cn$, where $C > 0$ is a sufficient large constant. For any $1 \leq j \leq p$, we have*

$$\mathbb{P}\left(\sup_{|u|_2=1,\ |u|_1^2=s_0L} \Big| \sum_{i=1}^n u^\top \big[\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top - \mathbb{E}\big(\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top | \mathcal{F}_{i-m+1}^n\big)\big] u \Big| \geq z \right) \leq s_0^2L^2e^{-cn},$$

for some constant $c > 0$.

Proof For matrix A, denote by A_{k_1,k_2} the $(k_1,k_2)th$ entry of A, and let $\mathbb{E}_{i-m+1}(\cdot) = (\cdot) - \mathbb{E}(\cdot|\mathcal{F}_{i-m+1}^n)$, then we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|u|_2=1,\ |u|_1^2=s_0L} \Big| u^\top \sum_{i=1}^n \mathbb{E}_{i-m+1}\big(\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top\big) u \Big| \geq z \right) \\
\leq &\mathbb{P}\left(\sup_{|u|_2=1,\ |u|_1^2=s_0L} |u|_1^2 \max_{1\leq k_1,k_2\leq p} \Big| \sum_{i=1}^n \mathbb{E}_{i-m+1}\big((\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top)_{k_1,k_2}\big)\mathbf{1}_{u_{k_1},u_{k_2}\neq 0} \Big| \geq z \right) \\
\leq &s_0^2L^2 \max_{1\leq k_1,k_2\leq p} \mathbb{P}\left(\Big| \sum_{i=1}^n \mathbb{E}_{i-m+1}\big((\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top)_{k_1,k_2}\big) \Big| \geq z/(s_0L) \right).
\end{aligned}$$

By construction, for any indices i, j, k_1, k_2, there exist functions

$$\phi_1, \phi_2 \in \{f : \mathcal{R}^p \to \mathcal{R} | f(x) = \psi_{j,k,l}(x_i) \text{ for some } 1 \leq j,k \leq p, 1 \leq l \leq L, 1 \leq i \leq p\}$$

such that $(\psi_{j,\cdot,\cdot}(X_i)\psi_{j,\cdot,\cdot}(X_i)^\top)_{k_1,k_2} = \phi_1(X_i)\phi_2(X_i)$. Since function $\psi_{j,k,l}$ satisfies conditions in Lemma 13, we complete the proof. ■

Lemma 13 *Consider the VAR process* (2)*, suppose Assumption 1 and 2(ii) hold. Assume functions $\phi_1, \phi_2 : \mathbb{R}^p \to \mathbb{R}$ are both bounded with $|\phi_i|_\infty \leq B$, $i = 1, 2$. For any $x, y \in \mathbb{R}^p$, assume $|\phi_i(x) - \phi_i(y)| \leq \beta^\top |x - y| = \sum_{j=1}^p \beta_j |x_j - y_j|$, where $|\beta|_1 \leq 1$. Then we have*

$$\mathbb{P}\Big(\Big| \sum_{i=1}^n \big[\phi_1(X_i)\phi_2(X_i) - \mathbb{E}\big(\phi_1(X_i)\phi_2(X_i)|\mathcal{F}_{i-m+1}^n\big)\big] \Big| \geq z \Big) \leq e^{-c\min\{n, z\rho^{-m}, z^2\rho^{-2m}/n\}}, \quad (41)$$

where constant c only depends on ρ, μ_2, μ_e and B.

Proof Recall $\mathcal{F}_k^n = \{\epsilon_k, \ldots, \epsilon_n\}$. Denote

$$S_n = \sum_{i=1}^{n} \big[\phi_1(X_i)\phi_2(X_i) - \mathbb{E}\big(\phi_1(X_i)\phi_2(X_i)|\mathcal{F}_{i-m+1}^n\big)\big] \quad \text{and} \quad \xi_k = \mathbb{E}(S_n|\mathcal{F}_{k-1}^n) - \mathbb{E}(S_n|\mathcal{F}_k^n).$$

Then $S_n = \sum_{k\le n-m+1} \xi_k$ and

$$\begin{aligned} |\xi_k| \le & \sum_{i=(k+m-1)\vee 1}^{n} \mathbb{E}\Big(|\phi_1(X_{i,\{k\}}) - \phi_1(X_i)||\phi_2(X_i)|\big|\mathcal{F}_k^n\Big) \\ & + \sum_{i=(k+m-1)\vee 1}^{n} \mathbb{E}\Big(|\phi_1(X_{i,\{k\}})||\phi_2(X_{i,\{k\}}) - \phi_2(X_i)|\big|\mathcal{F}_k^n\Big) =: \xi_{1k} + \xi_{2k}. \end{aligned} \tag{42}$$

Since $|\phi_1(X_{i,\{k\}}) - \phi_1(X_i)| \le \beta^\top H^{i-k}\mathrm{abs}(\epsilon_k' - \epsilon_k)$ and $|\phi_1|_\infty \le B$, we have

$$\xi_{1k} \le \sum_{i=(k+m-1)\vee 1}^{n} B \cdot \mathbb{E}\Big(\beta^\top H^{i-k}\mathrm{abs}(\epsilon_k' - \epsilon_k)\big|\mathcal{F}_k^n\Big).$$

A similar bound can be derived for ξ_{2k}. Hence

$$|\xi_k| \le \mathbb{E}\big(\omega_k^\top \mathrm{abs}(\epsilon_k' - \epsilon_k)|\mathcal{F}_k^n\big), \text{ where } \omega_k^\top = 2B\beta^\top \sum_{i=(k+m-1)\vee 1}^{n} H^{i-k}.$$

Then $|\omega_k|_1 \le 2B(1-\rho)^{-1}\rho^{m-1}$ for $k > -n$ and $|\omega_k|_1 \le 2B(1-\rho)^{-1}\rho^{1-k}$ if $k \le -n$. For $k \le -n$, since ξ_k are martingale differences, by Burkholder's inequality (Lemma 9), we have, for any $q \ge 2$,

$$\Big\|\sum_{k\le -n} \xi_k\Big\|_q^2 \le (q-1)^{q/2}\Big(\sum_{k\le -n} \|\xi_k\|_q^2\Big)^{q/2} \le (q-1)^{q/2}(2B)^q\mu_q^q(1-\rho)^{-q}(1-\rho^2)^{-q/2}\rho^q\rho^{nq}.$$

Thus by Markov's inequality

$$\mathbb{P}\Big(\big|\sum_{k\le -n} \xi_k\big| \ge z\Big) \le z^{-2}4B^2(1-\rho)^{-2}(1-\rho^2)^{-1}\mu_2^2\rho^2 \cdot \rho^{2n} \le z^{-2}4B^2(1-\rho)^{-4}\mu_2^2\rho^2 \cdot e^{-(-2\log\rho)n}.$$

For $k > -n$, let $h^* = (2B)^{-1}(1-\rho)\rho c_0$ and $\xi_k' = \xi_k/\rho^m$ Then $\mathbb{E}\exp(h^*|\xi_k'|) \le 2\mu_e < \infty$. By (8), (9) and (10), we have for any $h \le h^*$,

$$\mathbb{E}(e^{\xi_k' h}|\mathcal{F}_{k-1}) \le 1 + c_1 h^2,$$

where $c_1 = 2\mu_e h^{*-2}$. Similar as (11), we have

$$\mathbb{P}\Big(\big|\sum_{k=-n+1}^{n} \xi_k/\rho^m\big| \ge z\Big) \le \inf_{h\le h^*} \exp\big(-zh + 2c_1 n h^2\big) \le \exp\big\{-z^2/(c_2 z + c_3 n)\big\},$$

for some constants c_2, c_3 depending on ρ, μ_2, μ_e and B. Then the desired result follows. ■

Proof [Proof of Theorem 7] Let

$$F(b) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \Psi(X_{i-1})^\top b)^2 + \lambda \sum_{j,k=1}^{p} \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\psi_{j,k,\cdot}(X_{i-1,k})^\top b_{j,k,\cdot})^2}.$$

Define

$$\nabla_n = \frac{1}{n}\sum_{i=1}^{n}\Psi(X_{i-1})(X_i - \Psi(X_{i-1})^\top b^*). \tag{43}$$

Recall the definition of $|\cdot|_{2,\alpha}$ in (23). Then

$$\begin{aligned}
|\nabla_n|_{2,\infty} &= \big|\frac{1}{n}\sum_{i=1}^{n}\Psi(X_{i-1})(\epsilon_i + r_i)\big|_{2,\infty} \\
&\le \frac{1}{n}\sum_{i=1}^{n} L^{1/2}\big|\Psi(X_{i-1})\big|_\infty |r_i|_\infty + \big|\frac{1}{n}\sum_{i=1}^{n}\Psi(X_{i-1})\epsilon_i\big|_{2,\infty} \\
&:= \mathrm{I}_1 + \mathrm{I}_2.
\end{aligned} \tag{44}$$

For I_1 part, by (22) and Proposition 6, we have $|\Psi(X_{i-1})|_\infty \le B$ and thus $\mathrm{I}_1 \le B^2 C(2\beta - 1)^{-1} s_0 L^{1-\beta}$. For I_2 part, by Lemma 14, with probability at least $1-(pL)^{-c'}$, $\mathrm{I}_2 \le c\sqrt{L\log(pL)/n}$, for some constants $c, c' > 0$.

For $c_2 \ge 12(c + CB^2(2\beta-1)^{-1})/\phi_L$, by Proposition 5, we have

$$\lambda \ge (12/\phi_L)\big(c\sqrt{L\log(pL)/n} + B^2 C(2\beta-1)^{-1} s_0 L^{1-\beta}\big) \ge 12|\nabla_n|_{2,\infty}/\phi_L.$$

Let

$$\widetilde{\phi}_L = \frac{\phi_L}{2} - \frac{1}{n} - \frac{c_4(s_0 L)\log(n)\log(pL)}{n},$$

and

$$\widetilde{\phi}_U = \phi_U + c_5 L\sqrt{\frac{\log(n)\log(pL)}{n}},$$

where $|u|_1 = s_0 L$ in Proposition 5, and c_4, c_5 are the constants in (28) and (29). Then, for $n \ge c_3(s_0 L)\log(n)\log(pL) + c_3 L^2\log(n)\log(pL)$ with sufficient large constant $c_3 > 0$, we have

$$\widetilde{\phi}_L \ge \frac{\phi_L}{3} \text{ and } \widetilde{\phi}_U \le 2\phi_U.$$

Denote

$$\Sigma_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top \quad \text{and} \quad J_n = \frac{1}{n}\sum_{i=1}^{n}\Psi(X_{i-1})\Psi(X_{i-1})^\top.$$

Hence, by Assumption 4 and Proposition 5, with probability approaching one, we have

$$\begin{aligned}
F(b) - F(b^*) =& -2\nabla_n^\top (b-b^*) + (b-b^*)^\top J_n (b-b^*) + \lambda \sum_{j,k=1}^{p} \big(|\Sigma_{j,k}^{1/2} b_{j,k,\cdot}|_2 - |\Sigma_{j,k}^{1/2} b^*_{j,k,\cdot}|_2 \big) \\
\geq& -2|\nabla_n|_{2,\infty} |b-b^*|_{2,1} + \widetilde{\phi}_L |b-b^*|_2^2 + \lambda \sum_{j,k \notin S} |\Sigma_{j,k}^{1/2} b_{j,k,\cdot}|_2 - \lambda \sum_{j,k \in S} |\Sigma_{j,k}^{1/2} (b_{j,k,\cdot} - b^*_{j,k,\cdot})|_2 \\
\geq& \widetilde{\phi}_L |b-b^*|_2^2 - \lambda(\phi_L/6 + \widetilde{\phi}_U) \sum_{j,k \in S} |b_{j,k,\cdot} - b^*_{j,k,\cdot}|_2, \\
\geq& (\phi_L/3) |b-b^*|_2^2 - \lambda(\phi_L/6 + 2\phi_U) \sum_{j,k \in S} |b_{j,k,\cdot} - b^*_{j,k,\cdot}|_2.
\end{aligned}$$

Since $\mathrm{Card}(S) = |S|_0 = s$, we have

$$\sum_{j,k \in S} |b_{j,k,\cdot} - b^*_{j,k,\cdot}|_2 \leq \sqrt{s} \sqrt{\sum_{j,k \in S} |b_{j,k,\cdot} - b^*_{j,k,\cdot}|_2^2} \leq s^{1/2} |b-b^*|_2.$$

Hence $|\hat{b} - b^*|_2 \leq (1/2 + 6\phi_U/\phi_L)\sqrt{s}\lambda$ in view of $F(\hat{b}) - F(b^*) \leq 0$.

Furthermore,

$$\sum_{j,k=1}^{p} \|\hat{h}_{jk} - h_{jk}\|_2^2 \leq \sqrt{2} \sum_{j,k=1}^{p} \Big\| \sum_{l=1}^{L} (\hat{b}_{j,k,l} - b^*_{j,k,l}) \psi_{j,k,l} \Big\|_2^2 + \sqrt{2} \sum_{j,k=1}^{p} \Big\| \sum_{l=L+1}^{\infty} b^*_{j,k,l} \psi_{j,k,l} \Big\|_2^2.$$

Since $(\psi_{j,k,l})_{j,k,l}$ are orthonormal basis functions, we have

$$\begin{aligned}
\sum_{j,k=1}^{p} \|\hat{h}_{jk} - h_{jk}\|_2^2 &\leq \sqrt{2} \sum_{j,k=1}^{p} \sum_{l=1}^{L} (\hat{b}_{j,k,l} - b^*_{j,k,l})^2 + \sqrt{2} \sum_{j,k=1}^{p} \sum_{l=L+1}^{\infty} b^{*2}_{j,k,l} \\
&\lesssim s\lambda^2 + \sum_{j,k=1}^{p} \sum_{l=L+1}^{\infty} b^{*2}_{j,k,l} l^{2\beta} l^{-2\beta} \\
&\lesssim s\lambda^2 + sL^{-2\beta},
\end{aligned}$$

which also implies (33). ■

Lemma 14 *For function $g : \mathbb{R}^p \to \mathbb{R}$, assume $|g|_\infty \leq B$. Under Assumption 2(ii), we have*

$$\mathbb{P}\Big(\big| \frac{1}{n} \sum_{i=1}^{n} g(X_{i-1}) \epsilon_{ij} \big| \geq z \Big) \leq \begin{cases} 2\exp\Big(-\frac{nz^2}{4c_1} \Big), & \text{if } z \leq 2c_0 c_1 B^{-1}, \\ 2\exp\big(-c_0 nz/(2B) \big), & \text{if } z > 2c_0 c_1 B^{-1}, \end{cases} \tag{45}$$

where $c_1 = \mu_e c_0^{-2} B^2$.

Proof Let $\xi_i = g(X_{i-1})\epsilon_{ij}$. Then $\xi_i, 1 \le i \le n$, are martingale differences with respect to $\mathcal{F}_i$. Let $h^* = c_0/B$. By Assumption 2 (ii), for any $0 < h \le h^*$, $\mathbb{E}(e^{|\xi_k|h}) < \infty$. Since $\mathbb{E}(\xi_k|\mathcal{F}_{k-1}) = 0$ and $e^x - x \le e^{|x|} - |x|$ for any x, we have

$$\begin{aligned}\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) &= 1 + \mathbb{E}(e^{\xi_k h} - \xi_k h - 1|\mathcal{F}_{k-1}) \\ &\le 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big]h^2. \end{aligned} \tag{46}$$

Note that for any fixed $x > 0$, $(e^{tx} - tx - 1)/t^2$ is increasing in $t \in (0, \infty)$. Hence

$$\mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big] \le \mathbb{E}\Big[\frac{e^{|\xi_k|h^*} - |\xi_k|h^* - 1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big] \le \frac{\mathbb{E}(e^{Bh^*|\epsilon_{ij}|})}{h^{*2}} \le c_1, \tag{47}$$

where $c_1 = \mu_e B^2 c_0^{-2}$. Combining (46) and (47), we can obtain

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \le 1 + c_1 h^2.$$

Then, by recursively applying the above inequality, we have

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i \ge z\Big) &\le e^{-nzh}\mathbb{E}\Big(e^{\sum_{i=1}^{n-1}\xi_i h}\mathbb{E}(e^{\xi_n h}|\mathcal{F}_{n-1})\Big) \\ &\le e^{-nzh}(1 + c_1 h^2)^n \\ &\le \exp\big(-nzh + nc_1 h^2\big).\end{aligned}$$

Take $h = \min\{h^*, z/(2c_1)\}$, we further obtain

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i \ge z\Big) \le \exp\Big(-\frac{nz^2}{4c_1}\Big)\mathbf{1}_{\{h^* \ge z/(2c_1)\}} + \exp\big(-c_0 nz/(2B)\big)\mathbf{1}_{\{h^* < z/(2c_1)\}}.$$

Similar argument can be applied to $\mathbb{P}(n^{-1}\sum_{i=1}^n \xi_i \le -z)$ and the desired result follows. ■

Proof [Proof of Theorem 8] Let $b_S = (b_{j,k,\cdot}, (j,k) \in S) \in \mathbb{R}^{sL}$, and

$$\Omega(b) = \sum_{j,k=1}^p \sqrt{\frac{1}{n}\sum_{i=1}^n (\psi_{j,k,\cdot}(X_{i-1,k})^\top b_{j,k,\cdot})^2}.$$

Denote

$$\hat{\Sigma}_{S,S} = \frac{1}{n}\sum_{i=1}^n \Psi_S(X_{i-1})\Psi_S(X_{i-1})^\top,$$

and

$$\hat{\Sigma}_{S_j,S_j} = \frac{1}{n}\sum_{i=1}^n \Psi_{S_j}(X_{i-1})\Psi_{S_j}(X_{i-1})^\top.$$

By Assumption 6, (34), (35) and (36) hold on some event $\mathcal{Z}$ with $\mathbb{P}(\mathcal{Z}) \to 1$. In the following, we shall only work on $\mathcal{Z}$.

A vector $\hat{b} \in \mathbb{R}^{p^2L}$ is an optimum of the objective function in (19) if and only if there is a subgradient $\hat{g} \in \partial\Omega(\hat{b})$, such that

$$\frac{2}{n}\sum_{i=1}^{n}\Psi(X_{i-1})(\Psi(X_{i-1})^\top\hat{b} - X_i) + \lambda\hat{g} = 0. \tag{48}$$

The subdifferential $\partial\Omega(b)$ is the set of vectors $g = (g_{jk}, 1 \le j,k \le p)$, with $\hat{g}_{jk} \in \mathbb{R}^L$, satisfying

$$g_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top b_{j,k,\cdot}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\psi_{j,k,\cdot}(X_{i-1,k})^\top b_{j,k,\cdot})^2}}, \tag{49}$$

$$g_{jk}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top\right)^{-1} g_{jk} \le 1. \tag{50}$$

Following the primal dual witness argument in Ravikumar et al. (2009) and Wainwright (2009), it suffices to set $\hat{b}_{S^c} = 0$ and $\hat{g}_S = \partial\Omega(b^*)_S$, and then show

$$\hat{b}_{j,k,\cdot} \ne 0, \quad \text{for } (j,k) \in S, \tag{51}$$

$$\hat{g}_{jk}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top\right)^{-1}\hat{g}_{jk} < 1, \quad \text{for } (j,k) \in S^c, \tag{52}$$

hold with probability approaching 1.

(i). Proof of (51).

Since $\hat{b}_{S^c} = b^*_{S^c} = 0$, (48) reduces to

$$\frac{2}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})(\Psi_S(X_{i-1})^\top\hat{b}_S - X_i) + \lambda\hat{g}_S = 0. \tag{53}$$

It implies that

$$\hat{b}_S - b^*_S = \hat{\Sigma}^{-1}_{S,S}\cdot\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\epsilon_i + \hat{\Sigma}^{-1}_{S,S}\cdot\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})r_i - \frac{\lambda}{2}\hat{\Sigma}^{-1}_{S,S}\cdot\hat{g}_S := \mathrm{I}_1 + \mathrm{I}_2 - \mathrm{I}_3. \tag{54}$$

We now proceed to bound $\mathrm{I}_1, \mathrm{I}_2$ and I_3. Recall the definition of $|\cdot|_{2,\alpha}$ in (23). Also recall that $\|A\|_\infty$ is the matrix ∞ norm of $A = (a_{ij})_{n\times m}$ with $\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{m}|a_{ij}|$.

For I_1, we have

$$\begin{aligned}|\mathrm{I}_1|_{2,\infty} &\le \sqrt{L}\left\|\hat{\Sigma}^{-1}_{S,S}\right\|_\infty\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\epsilon_i\right|_\infty \\ &= \sqrt{L}\max_{1\le j\le p}\left\|\hat{\Sigma}^{-1}_{S_j,S_j}\right\|_\infty\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\epsilon_i\right|_\infty.\end{aligned}$$

By Lemma 14, with probability at least $1-(pL)^{-c_1}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\Psi_S(X_{i-1})\epsilon_i\right|_\infty \le c_2\sqrt{\frac{\log(pL)}{n}}. \tag{55}$$

Note that

$$\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_\infty = \max_{1\le j\le p}\left\|\hat{\Sigma}_{S_j,S_j}^{-1}\right\|_\infty \le \max_{1\le j\le p}\left\|\hat{\Sigma}_{S_j,S_j}^{-1}\right\|_2 \cdot \sqrt{s_0 L} = \sqrt{s_0 L}\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_2.$$

Then by (34), with probability at least $1-(pL)^{-c_1}$,

$$|\mathrm{I}_1|_{2,\infty} \le c_2\sqrt{L}\cdot\frac{\sqrt{s_0 L}}{\phi_{\min}}\cdot\sqrt{\frac{\log(pL)}{n}} = c_2\phi_{\min}^{-1}\frac{L\sqrt{s_0\log(pL)}}{\sqrt{n}}. \tag{56}$$

For I_2, by (22) and Proposition 6, we have

$$|\mathrm{I}_2|_{2,\infty} \le \sqrt{L}\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_\infty |\Psi_S(X_{i-1})|_\infty |r_i|_\infty \le B^2 C(2\beta-1)^{-1}\phi_{\min}^{-1}s_0^{3/2}L^{3/2-\beta}. \tag{57}$$

For I_3 part, note that for all $(j,k)\in S$,

$$\frac{1}{\phi_{\max}}|\hat{g}_{jk}|_2^2 \le \hat{g}_{jk}^\top\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top\right)^{-1}\hat{g}_{jk} \le 1.$$

It follows that

$$|\hat{g}_S|_\infty = \max_{(j,k)\in S}|\hat{g}_{jk}|_\infty \le \max_{(j,k)\in S}|\hat{g}_{jk}|_2 \le \sqrt{\phi_{\max}}. \tag{58}$$

Therefore we obtain

$$|\mathrm{I}_3|_{2,\infty} \le \frac{1}{2}\lambda\sqrt{L}\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_\infty |\hat{g}_S|_\infty \le \frac{\sqrt{\phi_{\max}}}{2\phi_{\min}}\cdot\lambda\sqrt{s_0}L. \tag{59}$$

Combining (56), (57) and (59), we have, with probability at least $1-(pL)^{-c_1}$,

$$\begin{aligned}|\hat{b}_S - b_S^*|_{2,\infty} &= \max_{(j,k)\in S}|\hat{b}_{j,k,\cdot} - b_{j,k,\cdot}^*|_2 \\ &\le c_2\phi_{\min}^{-1}\frac{L\sqrt{s_0\log(pL)}}{\sqrt{n}} + B^2C(2\beta-1)^{-1}\phi_{\min}^{-1}s_0^{3/2}L^{3/2-\beta} + \frac{\sqrt{\phi_{\max}}}{2\phi_{\min}}\cdot\lambda\sqrt{s_0}L.\end{aligned} \tag{60}$$

By (37) and (38), it follows that, on an event $\mathcal{Z}_1$ with probability approaching 1,

$$\max_{(j,k)\in S}|\hat{b}_{j,k,\cdot} - b_{j,k,\cdot}^*|_2 \to 0.$$

Since $\max_{(j,k)\in S}|b_{j,k,\cdot}^*|_2 > 0$ and will not converge to 0 asymptotically, (51) holds on an event $\mathcal{Z}_1$ with probability approaching 1.

(ii). Proof of (52).

Since $\hat{b}_{S^c} = b^*_{S^c} = 0$, for all $(j,k) \in S^c$, (48) reduces to

$$\frac{2}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})(\Psi_{S_j}(X_{i-1})^\top \hat{b}_{S_j} - X_{i,j}) + \lambda \hat{g}_{jk} = 0.$$

It implies that

$$\hat{g}_{jk} = \frac{2}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})(\Psi_{S_j}(X_{i-1})^\top (b^*_{S_j} - \hat{b}_{S_j}) + \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(\epsilon_{ij} + r_{ij})\right).$$

By (54), we have

$$\begin{aligned}
\hat{g}_{jk} = & \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})\Psi_{S_j}(X_{i-1})^\top \hat{\Sigma}^{-1}_{S_j,S_j}\right)\hat{g}_{S_j} \\
& - \frac{2}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})\Psi_{S_j}(X_{i-1})^\top \hat{\Sigma}^{-1}_{S_j,S_j}\right)\frac{1}{n}\sum_{i=1}^{n}\Psi_{S_j}(X_{i-1})\epsilon_{ij} \\
& - \frac{2}{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})\Psi_{S_j}(X_{i-1})^\top \hat{\Sigma}^{-1}_{S_j,S_j}\right)\frac{1}{n}\sum_{i=1}^{n}\Psi_{S_j}(X_{i-1})r_{ij} \\
& + \frac{2}{\lambda}\cdot\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}\epsilon_{ij} + \frac{2}{\lambda}\cdot\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k,\cdot}r_{ij} \\
:= & \, \mathrm{II}_1 - \mathrm{II}_2 - \mathrm{II}_3 + \mathrm{II}_4 + \mathrm{II}_5.
\end{aligned}$$

Since for all $(j,k) \in S^c$,

$$\hat{g}_{jk}^\top \left(\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})\psi_{j,k,\cdot}(X_{i-1,k})^\top\right)^{-1} \hat{g}_{jk} \le \frac{1}{\phi_{\min}}|\hat{g}_{jk}|_2^2.$$

It suffices to show $\max_{(j,k)\in S^c}|\hat{g}_{jk}|_2 < \sqrt{\phi_{\min}}$. We now proceed to bound $\mathrm{II}_1, \mathrm{II}_2, \mathrm{II}_3, \mathrm{II}_4$ and II_5.

For II_1, by (36) and (58),

$$\begin{aligned}
|\mathrm{II}_1|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^{n} \psi_{j,k,\cdot}(X_{i-1,k})\Psi_{S_j}(X_{i-1})^\top \hat{\Sigma}^{-1}_{S_j,S_j}\right\|_2 |\hat{g}_{S_j}|_2 \\
&\le \sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0}\sqrt{\phi_{\max}} \\
&\le (1-\delta)\sqrt{\phi_{\min}}. \qquad (61)
\end{aligned}$$

For II_2, by Lemma 14, as $s_0 < n$, with probability at least $1-(nL)^{-c_3}$

$$\begin{aligned}|\mathrm{II}_2|_2 &\le \frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0 L}\left|\frac{1}{n}\sum_{i=1}^{n}\Psi_{S_j}(X_{i-1})\epsilon_{ij}\right|_\infty \\ &\le \frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0 L}\cdot c_4\sqrt{\frac{\log(nL)}{n}} \\ &= c_5\frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}}. \end{aligned} \tag{62}$$

For II_3, by (22) and Proposition 6, we have

$$|\mathrm{II}_3|_2 \le \frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0 L}\cdot B^2 C(2\beta-1)^{-1}s_0 L^{1/2-\beta} = c_6\frac{s_0 L^{1-\beta}}{\lambda}. \tag{63}$$

Similarly, for II_4, with probability at least $1-(nL)^{-c_7}$,

$$|\mathrm{II}_4|_2 \le c_8\frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}}. \tag{64}$$

For II_5,

$$|\mathrm{II}_5|_2 \le 2B^2 C(2\beta-1)^{-1}\frac{s_0 L^{1-\beta}}{\lambda} = c_9\frac{s_0 L^{1-\beta}}{\lambda}. \tag{65}$$

In view of (61), (62), (63), (64) and (65), for all $(j,k)\in S^c$, we can obtain, with probability at least $1-(nL)^{-c_3}-(nL)^{-c_7}$,

$$|\hat{g}_{jk}|_2 \le (1-\delta)\sqrt{\phi_{\min}} + (c_5+c_8)\frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}} + (c_6+c_9)\frac{s_0 L^{1-\beta}}{\lambda}. \tag{66}$$

By (38), it follows that, on an event $\mathcal{Z}_2$ with probability approaching 1,

$$|\hat{g}_{jk}|_2 \le (1-\delta)\sqrt{\phi_{\min}} + o(1).$$

Hence, (52) holds on an event $\mathcal{Z}_2$ with probability approaching 1. Then Theorem 8 follows. ■

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

Tarmo Äijö and Harri Lähdesmäki. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, 25(22):2937–2944, 2009.

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

Denis Bosq. Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics*, 24(1):59–70, 1993.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Donald L. Burkholder. Sharp inequalities for martingales and stochastic integrals. *Astérisque*, (157-158):75–94, 1988. ISSN 0303-1179. Colloque Paul Lévy sur les Processus Stochastiques (Palaiseau, 1987).

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Likai Chen and Wei Biao Wu. Stability and asymptotics for autoregressive processes. *Electronic Journal of Statistics*, 10(2):3723–3751, 2016.

Likai Chen and Wei Biao Wu. Concentration inequalities for empirical processes of linear time series. *The Journal of Machine Learning Research*, 18(1):8639–8684, 2018.

Rong Chen and Ruey S Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

Persi Diaconis and David Freedman. Iterated random functions. *SIAM review*, 41(1):45–76, 1999.

Randal Douc, Arnaud Guillin, and Eric Moulines. Bounds on regeneration times and limit theorems for subgeometric markov chains. In *Annales de l'IHP Probabilités et statistiques*, volume 44, pages 239–257, 2008.

Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.

Jianqing Fan, Yuan Liao, and Weichen Wang. Projected principal component analysis in factor models. *Annals of Statistics*, 44(1):219, 2016.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's lemma for markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.

Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, 2004.

Zhaoxing Gao, Yingying Ma, Hansheng Wang, and Qiwei Yao. Banded spatio-temporal autoregressions. *Journal of econometrics*, 208(1):211–230, 2019.

Satyajit Ghosh, Kshitij Khare, and George Michailidis. High-dimensional posterior consistency in bayesian vector autoregressive models. *Journal of the American Statistical Association*, 114(526):735–748, 2019.

Shaojun Guo, Yazhen Wang, and Qiwei Yao. High-dimensional and banded vector autoregressions. *Biometrika*, 103(4):889–903, 10 2016.

Eric C Hall, Garvesh Raskutti, and Rebecca M Willett. Learning high-dimensional generalized linear autoregressive models. *IEEE Transactions on Information Theory*, 65(4): 2401–2422, 2018.

Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150, 2015.

SF Jarner and RL Tweedie. Locally contracting iterated functions and stability of markov chains. *Journal of applied probability*, 38(2):494–507, 2001.

Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

Neil D Lawrence, Mark Girolami, Magnus Rattray, and Guido Sanguinetti. *Learning and inference in computational systems biology*. MIT press Cambridge, MA, 2010.

Sophie Lèbre. Inferring dynamic genetic networks with low order independencies. *Statistical applications in genetics and molecular biology*, 8(1):1–38, 2009.

Néhémy Lim, Florence dAlché Buc, Cédric Auliac, and George Michailidis. Operator-valued kernel-based vector autoregressive models for network inference. *Machine learning*, 99(3): 489–513, 2015.

Jiahe Lin and George Michailidis. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. *The Journal of Machine Learning Research*, 18 (1):4188–4236, 2017.

Yan Liu, Alexandru Niculescu-Mizil, Aurelie C Lozano, and Yong Lu. Learning temporal causal graphs for relational time-series analysis. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 687–694, 2010.

Johanna Mazur, Daniel Ritter, Gerhard Reinelt, and Lars Kaderali. Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. *BMC bioinformatics*, 10(1):448, 2009.

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.

Dharmendra S Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145, 1996.

Rebecca B Morton and Kenneth C Williams. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press, 2010.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*, 2013.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Emmanuel Rio. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1):146–163, 2009. ISSN 0894-9840.

Herbert Robbins. A remark on stirling's formula. *The American mathematical monthly*, 62 (1):26–29, 1955.

Michal Ronen, Revital Rosenberg, Boris I Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the national academy of sciences*, 99(16):10555–10560, 2002.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Xiaofeng Shao and Wei Biao Wu. Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics*, 35(4):1773–1801, 2007.

Chao Sima, Jianping Hua, and Sungwon Jung. Inference of gene regulatory networks using time-series data: a survey. *Current genomics*, 10(6):416–429, 2009.

Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Wei Biao Wu and Xiaofeng Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Danna Zhang. Robust estimation of the mean and covariance matrix for high dimensional time series. *Statistica Sinica*, 31(2):797–820, 2021.

Danna Zhang and Wei Biao Wu. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

Danna Zhang and Wei Biao Wu. Convergence of covariance and spectral density estimates for high dimensional locally stationary processes. *The Annals of Statistics*, To appear, 2020.

Hao Henry Zhou and Garvesh Raskutti. Non-parametric sparse additive auto-regressive network models. *IEEE Transactions on Information Theory*, 65(3):1473–1492, 2018.