Estimation of High-dimensional Nonlinear Vector Autoregressive Models

Yuefeng Han, Likai Chen and Wei Biao Wu*

University of Notre Dame, Washington University in St. Louis and The University of Chicago

Abstract

High-dimensional vector autoregressive (VAR) models have numerous applications in fields such as econometrics, biology, climatology, among others. While prior research has mainly focused on linear VAR models, these approaches can be restrictive in practice. To address this, we introduce a high-dimensional non-parametric sparse additive model, providing a more flexible framework. Our method employs basis expansions to construct high-dimensional nonlinear VAR models. We derive convergence rates and model selection consistency for least squared estimators, considering dependence measures of the processes, error moment conditions, sparsity, and basis expansions. Our theory significantly extends prior linear VAR models by incorporating both non-Gaussianity and non-linearity. As a key contribution, we derive sharp Bernstein-type inequalities for tail probabilities in both non-sub-Gaussian linear and nonlinear VAR processes, which match the classical Bernstein inequality for independent random variables. Additionally, we present numerical experiments that support our theoretical findings and demonstrate the advantages of the nonlinear VAR model for a gene expression time series dataset.

Index Terms: Nonlinear vector autoregression, time series analysis, high-dimensional analysis, Bernstein inequality, non-parametric, sparsity, basis expansion, Lasso estimation, martingale

1 Introduction

The increasing variety of scientific applications has created a growing need for employing a large set of time series (variables) to model complex social and physical systems. This demand arises from various fields, including genomics (Sharon et al., 2013), neuroscience (Möller et al., 2001, Pereda et al., 2005, Kato et al., 2006), social networks (Aït-Sahalia et al., 2015), economics (Barigozzi and Hallin, 2017), environmental studies (Lichstein et al., 2002), and communication engineering (Baddour and Beaulieu, 2005). For example, economic policymakers rely on large-scale models of

*Yuefeng Han is Assistant Professor, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556. Email: yuefeng.han@nd.edu. Likai Chen is Assistant Professor, Department of Statistics and Data Science, Washington University in St. Louis, St. Louis, MO 63130. E-mail: likai.chen@wustl.edu. Weibiao Wu is Professor, Department of Statistics, The University of Chicago, Chicago, IL 60637. E-mail: wbwu@uchicago.edu. Han was supported in part by National Science Foundation grant DMS-2412578. Chen was supported in part by National Science Foundation grant EF-2222403 and DMS-2311251. Wu was supported in part by National Science Foundation grants DMS-2311249 and DMS-2027723.

economic indicators (Sims, 1980, Bernanke et al., 2005, Bańbura et al., 2010), as empirical evidence has shown that they improve forecasts and provide better estimates of how current economic shocks will propagate, which guides policy actions more effectively. Similarly, in genomics and neuroscience, the advent of high-throughput technologies has enabled researchers to collect measurements on hundreds of genes or brain regions (Shojaie and Michailidis, 2010, Seth et al., 2015), facilitating comprehensive modeling and deeper insights into biological mechanisms. In social sciences, many key variables are not directly observable but can be inferred through related time series variables, enabling a more nuanced understanding of policy decisions (Lin and Michailidis, 2020). Given the wide availability of high-dimensional time series data, understanding their underlying dynamic patterns is crucial for improving practical applications in these domains.

A widely used and informative model for capturing linear temporal dependencies between time series is the vector autoregression (VAR) model. Properties of VAR have been extensively studied in low-dimensional settings; see Lütkepohl (2005) for a comprehensive overview. Over the past decade, a growing body of literature has leveraged structured sparsity and regularized estimation frameworks to achieve consistent estimation of VAR parameters in high-dimensional settings. Basu and Michailidis (2015) investigated the theoretical properties of Lasso-penalized high-dimensional VAR models for Gaussian processes. Their result was extended to multi-block VAR models by Lin and Michailidis (2017) and to factor-augmented VAR models by Lin and Michailidis (2020). Guo et al. (2016) introduced a class of VAR models with banded coefficient matrices, which was further developed into spatio-temporal VAR models by Gao et al. (2019). Basu et al. (2019) explored high-dimensional VAR models involving low-rank and group-sparse components in network structures. Hall et al. (2018) studied regularized high-dimensional autoregressive generalized linear models, focusing on Bernoulli and Poisson distributions. Additionally, Ghosh et al. (2019, 2021) developed Bayesian VAR models and analyzed their posterior and strong selection consistency. For further related work, see Zheng and Raskutti (2019), Pandit et al. (2020), Wang et al. (2022), Wang and Tsay (2023), Chen et al. (2023), among others.

Although many mechanisms, such as regulatory processes in biology (cf. Sima et al. (2009) for a survey), involve nonlinear dynamics, research on high-dimensional time series models addressing such dynamics remains limited. Mazur et al. (2009) and Äijö and Lähdesmäki (2009) employed Bayesian learning to manage the stochasticity of biological data. Lim et al. (2015) introduced a family of VAR models using operator-valued kernels to identify nonlinear dynamic systems. Zhou and Raskutti (2018) proposed a framework for non-parametric autoregressive models within generalized linear models by utilizing reproducing kernel Hilbert spaces, analyzing the convex penalized sparse and smooth estimator. Shen et al. (2019) investigated nonlinear structural VAR models with application to brain networks. Additional applications can be found in Pereda et al. (2005), Balcilar et al. (2016), Yu et al. (2021), among others. Among these works, only Zhou and Raskutti (2018) provided theoretical guarantees, although their concentration inequalities are not sharp. In this paper, we extend the framework of sparse linear VAR models to sparse non-parametric nonlinear VAR models, with rigorous theoretical guarantees.

This paper has two primary objectives: (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

$$X_t = h^{(1)}(X_{t-1}) + h^{(2)}(X_{t-2}) + \ldots + h^{(d)}(X_{t-d}) + \epsilon_t, \tag{1}$$

where $\epsilon_t \in \mathbb{R}^p, t \in \mathbb{Z}$, are i.i.d. random vectors, $X_t = (X_t^{(1)}, \ldots, X_t^{(p)})^\top \in \mathbb{R}^p$, $h^{(j)} = (h_1^{(j)}, \ldots, h_p^{(j)})^\top$ and $h_k^{(j)} : \mathbb{R}^p \to \mathbb{R}$, $1 \leqslant j \leqslant d, 1 \leqslant k \leqslant p$, 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

$$X_t = h(X_{t-1}) + \epsilon_t. \tag{2}$$

Based on model (2), we shall develop sharp Bernstein-type inequalities. Establishing exponential-type tail probability inequalities for temporal dependent processes is a challenging problem. There has been some effort to derive concentration inequalities for non-i.i.d. processes. For example, generalizations of Bernstein's 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), Adamczak (2008), Lemańczyk (2021). Unfortunately, all these inequalities involve extra non-constant factors to account for weak dependence, and are not as sharp as the original Bernstein's inequality for independent random variables. Recently, Fan et al. (2021) 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 the one-dimensional ones. In this work, we establish Bernstein-type inequalities for nonlinear VAR processes. Our inequalities, up to some constants, 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 such sharp Bernstein-type inequalities for time series. Notably, we do not use the commonly employed "blocking" technique for sequences of dependent random variables (Hall et al., 2018), which allows us to avoid logarithmic factors. Our technical approach can be used to improve existing studies on high-dimensional VAR models, such as in Kock and Callot (2015), Jiang et al. (2023), Dahlhaus and Richter (2023), Wang and Tsay (2023).

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, $1 \leqslant j \leqslant p$, in model (2) can be expressed as:

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

where $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ 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 $\{(j,k) : h_{jk}(\cdot) \neq 0\}$ 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 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 (2010a), 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 (RKHS). 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. Recently, Düker and Waterbury (2025) developed an RKHS-based framework for nonlinear VAR processes and derived non-asymptotic probabilistic bounds.

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 speaking, p can be as large as e^{n^c} for some constant $0 < c < 1$ if ϵ_t 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 on 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 Sections 4 and 5, respectively. Proofs of theorems and technical lemmas are contained in Section 6.

We now introduce some notation. For a vector $x = (x_1, \ldots, x_p)^\top$, define $\|x\|_q = (|x_1|^q + \ldots + |x_p|^q)^{1/q}$, $q \geqslant 1$, $\|x\| = \|x\|_2$, $\|x\|_\infty = \max_{1 \leqslant j \leqslant p} |x_j|$, and $\mathrm{abs}(x) := (|x_1|, \ldots, |x_p|)^\top$. 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 $\|A\|_2 = \max_{\|x\|_2 \leqslant 1} \|Ax\|_2$ 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. 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 $|a_n| \leqslant C|b_n|$ (resp. $1/C \leqslant a_n/b_n \leqslant C$) holds for all sufficiently large n, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

Let $\epsilon_t, t \in \mathbb{Z}$, 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 $X_t = g(\ldots, \epsilon_{t-1}, \epsilon_t)$, where g is a measurable function, we define the coupled version $X_{t,\{k\}} = g(\ldots, \epsilon_{k-1}, \epsilon'_k, \epsilon_{k+1}, \ldots, \epsilon_t)$, which has the same distribution as X_t 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. random variables, 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 (2021) 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 Theorem 1, 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 the function h and the errors ϵ_t in model (2). Recall that $\|\cdot\|_\infty$ is the matrix infinity norm.

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

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

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

The above assumption requires componentwise Lipschitz condition for nonlinear VAR processes. 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, $X_t = \rho X_{t-1} + \epsilon_t$. Larger ρ suggests stronger dependence.

Remark 1 (Existence of stationary distribution)**.** For the sake of completeness, in this remark, we shall apply the theory in Chen and Wu (2016) and show the existence of stationary distribution. Construct a collection of backward series of random vectors $X_{(-n),t}$, for $t \geqslant -n$, as follows. For all $t \in \mathbb{Z}$, define $X_{(t),t} = 0$ and the recursion,

$$X_{(-n),t} = h\big(X_{(-n),t-1}\big) + \epsilon_t.$$

Let $X^{(j)}_{(-n),t}$ denote the j-th component of $X_{(-n),t}$. Then $X_{(-n),t} = (X^{(1)}_{(-n),t}, \ldots, X^{(p)}_{(-n),t})^\top$. Under Assumption 1, we have

$$\begin{aligned}\|X_{(-n+1),t} - X_{(-n),t}\|_\infty &\leqslant \max_{1\leqslant j\leqslant p}\sum_{k=1}^{p} H_{jk}|X^{(k)}_{(-n+1),t-1} - X^{(k)}_{(-n),t-1}| \\ &\leqslant \rho\|X_{(-n+1),t-1} - X_{(-n),t-1}\|_\infty \\ &\leqslant \rho^{t+n-1}\|X_{(-n+1),-n+1} - X_{(-n),-n+1}\|_\infty. \end{aligned} \tag{4}$$

Taking the L_q norm and defining $c_p = \|\max_{1\leqslant j\leqslant p}|h_j(0) + \epsilon^{(j)}_1|\|_q$, we obtain

$$\big\|\|X_{(-n+1),t} - X_{(-n),t}\|_\infty\big\|_q \leqslant \rho^{t+n-1}c_p.$$

Since $\rho < 1$, for fixed t, the sequence $X^{(j)}_{(-n),t}$ converges as $n \to \infty$ for any $1 \leqslant j \leqslant p$. Denote the limit by $Y^{(j)}_t$ and set $Y_t = (Y^{(1)}_t, \ldots, Y^{(p)}_t)^\top$. We now show that Y_t is the stationary solution of model (2). For any $\kappa > 0$, there exists an $N_0 \in \mathbb{N}$, such that for $n > N_0$ we have $\|Y_k - X_{(-n),k}\|_\infty < \kappa$. Then, using the Lipschitz condition on h,

$$\max_{1\leqslant j\leqslant p} |h_j(Y_{t-1}) + \epsilon^{(j)}_t - Y^{(j)}_t| \leqslant \max_{1\leqslant j\leqslant p} |h_j(Y_{t-1}) - h_j(X_{(-n),t-1})| + \max_{1\leqslant j\leqslant p} |X^{(j)}_{(-n),t} - Y^{(j)}_t| \leqslant 2\kappa.$$

Since κ is arbitrary, this implies $Y_t = h(Y_{t-1}) + \epsilon_t$ almost surely, so Y_t indeed satisfies the VAR recursion and is stationary.

■

Remark 2. Denote H^m as the result of multiplying the matrix H by itself m times. We can weaken Assumption 1 by requiring only that there exists an integer $m \geqslant 1$ such that $\|H^m\|_\infty \leqslant \rho < 1$. In other words, we allow $\|H\|_\infty \geqslant 1$ as long as repeated application of H eventually satisfies this condition. All of our subsequent results remain valid under this relaxed assumption. To see this, we consider the previous argument for the existence of a stationary distribution. Repeatedly applying the first inequality in (4), we have

$$\begin{aligned}\|X_{(-n+1),t} - X_{(-n),t}\|_\infty &\leqslant \max_{1\leqslant j\leqslant p} \sum_{k=1}^{p} (H^m)_{jk} |X^{(k)}_{(-n+1),t-m} - X^{(k)}_{(-n),t-m}| \\ &\leqslant \rho \|X_{(-n+1),t-m} - X_{(-n),t-m}\|_\infty. \end{aligned} \tag{5}$$

Iterating (5), we conclude that $\|X_{(-n+1),t} - X_{(-n),t}\|_\infty \lesssim \rho^{\lfloor (n+t)/m \rfloor}$, where $\lfloor x \rfloor$ is the largest integer less or equal to x. Then for fixed t, we have that $X_{(-n),t}$ converges as $n \to \infty$. Similar adaptations apply throughout the paper under this relaxed condition instead of Assumption 1. ■

Assumption 2. For i.i.d. random vectors $\epsilon_t = (\epsilon^{(1)}_t, \ldots, \epsilon^{(p)}_t)^\top \in \mathbb{R}^p$, $t \in \mathbb{Z}$, assume one of the following holds:

(i) (finite moment) $\mu_q := \max_{1\leqslant j\leqslant p} (\mathbb{E}|\epsilon^{(j)}_t|^q)^{1/q} < \infty$ for some $q \geqslant 2$.

(ii) (exponential moment) $\mu_e := \max_{1\leqslant j\leqslant p} \mathbb{E}\big(\exp(c_0|\epsilon^{(j)}_t|)\big)$, for some $c_0 > 0$.

Assumption 3. Let function $g : \mathbb{R}^p \to \mathbb{R}$, be Lipschitz continuous with $|g(x) - g(y)| \leqslant \sum_{j=1}^p G_j |x_j - y_j|$, for any $x = (x_1, \ldots, x_p)^\top$, $y = (y_1, \ldots, y_p)^\top \in \mathbb{R}^p$, where G_j are Lipschitz coefficients. Denote $G = (G_1, \ldots, G_p)^\top$ and $\tau := \|G\|_1 = \sum_{j=1}^p G_j$.

The following theorem presents a Bernstein-type inequality for bounded Lipschitz continuous functions, under both the finite moment condition and the exponential moment condition of the error vectors ϵ_t, respectively.

Theorem 1. *Consider the VAR process defined in (2), where the function h satisfies Assumption 1. Let g be any function satisfying Assumption 3 with $\tau = \|G\|_1$. Then:*

(i) If Assumption 2 (i) holds and g is bounded with $\|g\|_\infty = \sup_x |g(x)| \leqslant M$, then for all $z \geqslant 0$,

$$\mathbb{P}\left(\left| \sum_{t=1}^{n} (g(X_t) - \mathbb{E}g(X_t)) \right| \geqslant z \right) \leqslant 2\exp\left\{ -\frac{z^2}{c_1 \tau^2 n + c_2 \tau M z} \right\}, \tag{6}$$

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

(ii) If Assumption 2 (ii) holds, then for all $z \geqslant 0$,

$$\mathbb{P}\left(\left|\sum_{t=1}^{n}(g(X_t)-\mathbb{E}g(X_t))\right| \geqslant z\right) \leqslant 2\exp\left\{-\frac{z^2}{c_3\tau^2 n + c_4\tau z}\right\}, \tag{7}$$

where c_3 and c_4 are positive constants depending only on ρ and μ_e.

Theorem 1(i) addresses the finite moment case for the error vectors ϵ_t (cf. Assumption 2 (i)). If the error vectors $\epsilon_t, t \in \mathbb{Z}$, satisfy stronger moment condition than merely having a finite q-th moment, we can expect a stronger inequality than (6). Indeed, when ϵ_t has subexponenial tail (Assumption 2 (ii)), we obtain an improved Bernstein-type inequality in (7). Different from Theorem 1 (i), in Theorem 1(ii), function g can be unbounded.

Remark 3. Based on the proof of Theorem 1(i), 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 (6) and obtain the following Hoeffding type inequality. ■

Corollary 1. *Consider the VAR process defined in (2), where the function h satisfies Assumption 1. Let g be any function satisfying Assumption 3. Suppose Assumption 2 (i) or 2 (ii) holds. If g is bounded with $\|g\|_\infty \leqslant 1$, then we have*

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{n}\big(g(X_t)-\mathbb{E}g(X_t)\big)\Big| \geqslant z\Big) \leqslant 2e^{-c_1 z^2/(\tau^2 n)}, \tag{8}$$

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

Remark 4. Note that up to a multiplicative constant, our Bernstein-type inequality (6) coincides with classical Bernstein's inequality for i.i.d. random variables. Thus one can expect sharper convergence rates for estimators of nonlinear VAR processes (2). We remark that the 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 $|X_t| \leqslant M$, Zhang (2021) provided the following Bernstein-type inequality,

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

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,

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{n}\big(X_t-\mathbb{E}X_t\big)\Big| \geqslant z\Big) \leqslant \exp\left\{-\frac{c_1 z^2}{nM^2 + M\log(n)\log\log(n) z}\right\},$$

where $c_1 > 0$ and $|X_t| \leqslant M$. Both involve an unpleasant $\log(n)$-type multiplicative 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 Sketch. The proof of Theorem 1 is quite involved. The key steps involve employing a martingale decomposition and deriving a sharp bound for the martingale differences. To be more specific, without loss of generality, assume $\|G\|_1 = 1$ with G defined in Assumption 3. Recall that $\mathcal{F}_k = (\ldots, \epsilon_{k-1}, \epsilon_k)$ and the projection operator $P_k(\cdot) = \mathbb{E}(\cdot|\mathcal{F}_k) - \mathbb{E}(\cdot|\mathcal{F}_{k-1})$, for $k \in \mathbb{Z}$. The summation can then be decomposed into a sum of martingale differences:

$$S_n(g) := \sum_{t=1}^{n} \big(g(X_t) - \mathbb{E}g(X_t)\big) = \sum_{k \leqslant n} \xi_k, \quad \text{where } \xi_k = P_k(S_n(g)).$$

For $X_t = \mathcal{G}(\ldots, \epsilon_{t-1}, \epsilon_t)$, where $\mathcal{G}$ is some measurable function, following Wu (2005), we define the coupled version

$$X_{t,\{k\}} = \mathcal{G}(\ldots, \epsilon_{k-1}, \epsilon'_k, \epsilon_{k+1}, \ldots, \epsilon_t).$$

For $x = (x_1, \ldots, x_p)^\top$, write $\mathrm{abs}(x) = (|x_1|, \ldots, |x_p|)^\top$. Since the mapping h is componentwise Lipschitz continuous, by induction, we have $\mathrm{abs}(X_t - X_{t,\{k\}}) \leqslant H^{t-k}\mathrm{abs}(\epsilon_k - \epsilon'_k)$. Hence

$$\begin{aligned} \big|P_k(g(X_t))\big| &= \big|\mathbb{E}(g(X_t) - g(X_{t,\{k\}})|\mathcal{F}_k)\big| \\ &\leqslant \mathbb{E}\left(G^\top \mathrm{abs}(X_t - X_{t,\{k\}})\big|\mathcal{F}_k\right) \\ &\leqslant \mathbb{E}\left(G^\top H^{t-k}\mathrm{abs}(\epsilon_k - \epsilon'_k)\big|\mathcal{F}_k\right). \end{aligned} \tag{9}$$

Since the function $g(\cdot)$ is bounded by M, it follows that $|P_k(g(X_t))| \leqslant 2M$. Therefore, combining this with (9), we obtain

$$|\xi_k| \leqslant \sum_{t=1}^{n} |P_k(g(X_t))| \leqslant \sum_{t=k\vee 1}^{n} \min\Big\{v_{t-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon'_k)\big|\mathcal{F}_k), 2M\Big\}, \quad \text{with} \quad \|v_t\|_1 \leqslant \rho^t. \tag{10}$$

Since $\|v_t\|_1$ decays exponentially fast, for all sufficiently large t, one shall expect the first term $v_{t-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon'_k)\big|\mathcal{F}_k)$ to be small. Then by carefully leveraging between the two terms as detailed in Lemma 2, we obtain that

$$\mathbb{E}(e^{|\xi_k|h}) < \infty$$

for any $h \leqslant h^*$ some constant $h^* > 0$. Since ξ_k's are martingale differences,

$$\begin{aligned} \mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) &= 1 + \mathbb{E}(e^{\xi_k h} - \xi_k h - 1|\mathcal{F}_{k-1}) \\ &\leqslant 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big]h^2, \end{aligned} \tag{11}$$

where the conditional expectation in the last line can be shown to be bounded for any $h \leqslant h^*$ with the bound denoted by c. Hence

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \leqslant 1 + ch^2. \tag{12}$$

The above applies for $-n \leqslant k \leqslant n$. For $k < -n$, we can show that those terms are negligible. The desired result then follows by Markov's inequality and recursively using (12) for $-n \leqslant k \leqslant n$.

It should be emphasized that our Bernstein-type concentration inequalities are sharp, and does not contain any annoying extra logarithmic terms. These inequalities are useful for handling non-Gaussian VAR problems.

3 Sparse additive nonlinear VAR models

In this section, we study sparse additive nonlinear VAR models. We first introduce a population-level optimization problem and then derive a sample-based algorithm through basis expansion. Our theoretical analysis builds on the technical tools developed in previous section.

3.1 The model

Assume that we are provided with observed time series data $X_1, \ldots, X_n \in \mathbb{R}^p$, which are sampled from a dynamical system involving p variables. Our primary goal is to infer the direct influence that each variable j exerts on every other variable k (with $k \neq j$, $1 \leqslant k \leqslant p$). For instance, in the case of linear VAR models, the evolution of the system is often characterized by $X_t = GX_{t-1} + \epsilon_t$, where G is a $p \times p$ coefficient matrix, and ϵ_t represents noise. In our study, we assume that a first-order stationary model provides a sufficient approximation of the temporal dependencies within the system. Accordingly, we recall the nonlinear VAR framework in (2),

$$X_t = h(X_{t-1}) + \epsilon_t,$$

where the function h can capture potentially complex, nonlinear dynamics.

In this section, we propose a new class of high-dimensional, sparse, additive non-parametric VAR models. Here, each component h_j of the function h is assumed to decompose additively in terms of the individual components of the state vector $x \in \mathbb{R}^p$. Specifically, we posit that for each variable j

$$h_j(x) = \sum_{k=1}^{p} h_{jk}(x_k), \tag{13}$$

where each function $h_{jk} : \mathbb{R} \to \mathbb{R}$ captures the individual contribution of the k-th variable to the dynamics of the j-th variable.

Let Π denote the joint distribution of the vector X_t, and let Π_k denote the marginal distribution of the k-th component $X_t^{(k)}$ for each $1 \leqslant k \leqslant p$. For practical purposes, we define the $L_2(\Pi_k)$-norm of the function h_{jk} as

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

This definition is particularly relevant because it allows us to accommodate functions h_{jk} that might not be Lebesgue integrable over the entire real line; instead, the integrability is considered relative to the distribution Π_k.

The classical nonlinear ridge regression is defined as

$$\frac{1}{n} \sum_{t=1}^{n} \|X_t - h(X_{t-1})\|_2^2 + \lambda \sum_{j=1}^{p} \sum_{k=1}^{p} \|h_{jk}\|_{\Pi_k,2}^2,$$

where the norms measure the overall discrepancy and the smoothness penalty on each component. To encourage sparsity in high-dimensional settings, we replace the squared norm $\|h_{jk}\|_{\Pi_k,2}^2$ with the

norm $\|h_{jk}\|_{\Pi_k,2}$ itself. This substitution leads to a population-level penalized least squares estimator defined by the optimization problem

$$(\hat{h}_{jk}, 1 \leqslant j,k \leqslant p) := \underset{h_{jk}\in\mathcal{I}_k, 1\leqslant j,k\leqslant p}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{t=1}^{n} \|X_t - h(X_{t-1})\|_2^2 + \lambda \sum_{j=1}^{p}\sum_{k=1}^{p} \|h_{jk}\|_{\Pi_k,2} \right\}. \tag{14}$$

Here, h is decomposed as in (13) into a sum of univariate functions, and $\mathcal{I}_k$ is an appropriate function class for the k-th component. In practice, the norm $\|h_{jk}\|_{\Pi_k,2}$ can be estimated empirically by $(n^{-1}\sum_{t=1}^{n} h_{jk}^2(X_{t-1}^{(k)}))^{1/2}$.

By decomposing h_j into additive components, our framework enhances interpretability and computational efficiency in high-dimensional settings. The imposed sparsity helps to pinpoint which variables have a direct influence on the dynamics. This model is an extension of the sparse additive models developed for the i.i.d. case (Ravikumar et al., 2009) and is especially relevant when the system exhibits nonlinear structure that traditional linear models fail to capture, while still preserving a structure that is amenable to rigorous analysis and estimation.

For each $k \in \{1,\ldots,p\}$, let $\mathcal{H}_k$ denote the Hilbert subspace $L_2(\Pi_k)$ consisting of measurable functions $f(\cdot)$ satisfying $\mathbb{E}f(X_t^{(k)}) = 0$ and the norm $\|f\| = (\mathbb{E}f^2(X_t^{(k)}))^{1/2} < \infty$. The inner product on $\mathcal{H}_k$ is defined as

$$\langle f, g\rangle = \mathbb{E}\big(f(X_t^{(k)})g(X_t^{(k)})\big).$$

We denote by $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots \oplus H_p$ the Hilbert space of functions of $(x_1,\ldots,x_p)$ that admit an additive representation $m(x) = \sum_k f_k(x_k)$ with each $f_k \in \mathcal{H}_k, k = 1,\ldots,p$.

We now impose the following assumption on our basis expansion.

Assumption 4 (Basis function)**.** Assume that the functions $h_{jk}(x)$ in (13) have compact support for all $1 \leqslant j,k \leqslant p$, that is, $|h_{jk}(x)| = 0$ for any $|x| > c_0$, for some constant $c_0 > 0$. Moreover assume $h_{jk} \in \mathcal{I}_k$ where

$$\mathcal{I}_k = \left\{ h_{jk}(\cdot) \in \mathcal{H}_k : h_{jk}(\cdot) = \sum_{l=1}^{\infty} b_{jk}^{(l)*}\psi_{k,l}(\cdot), \quad \sum_{l=1}^{\infty}(b_{jk}^{(l)*})^2 l^{2\beta} \leqslant C^2 \right\},$$

where $(\psi_{k,l}(\cdot) : l = 1,2,\ldots)$ is a uniformly bounded orthonormal basis on $[-c_0,c_0]$, that is $|\psi_{k,l}(x)| \leqslant B$, for some $0 < B, C < \infty$ and $\beta \geqslant 1$.

For example, we can choose the Fourier basis functions to satisfy Assumption 4. In standard nonparametric regression such as Ravikumar et al. (2009), covariates are often assumed to be bounded (i.e., to have compact support). Similarly, in our nonlinear VAR framework we assume $h_{jk}(x)$ in (13) have compact support for mathematical convenience and tractability; see also Raskutti et al. (2012), Zhou and Raskutti (2018). Many of our results can be extended to the case of unbounded $h_{jk}(x)$ via truncation arguments with proper tail decay conditions. We omit such arguments for the sake of presentation simplicity. For example, the Fourier basis satisfies this assumption. This assumption implies that the tail of the expansion satisfies $\sum_{l=L+1}^{\infty}(b_{jl}^{(l)*})^2 \leqslant C^2L^{-2\beta}$, which corresponds to the functional class condition of Ravikumar et al. (2009) and is a standard requirement in basis expansion methods. The parameter β captures the level of smoothness, effectively linking our function class to a function space. Although one could allow β to vary adaptively with k, we confine ourselves to a common smoothness level in this work.

Let $L = L_n$ be a truncation parameter, and let $h_{jk}^{(L)}$ be the approximation of h_{jk} defined by

$$h_{jk}^{(L)}(\cdot) = \sum_{l=1}^{L} b_{jk}^{(l)*} \psi_{k,l}(\cdot). \tag{15}$$

In this formulation, $h_{jk}^{(L)}$ is interpreted as the projection of h_{jk} onto the truncated set of basis functions $\{\psi_{k,1}, \ldots, \psi_{k,L}\}$. Then, for $1 \leqslant j, k \leqslant p$, the model can be written as

$$X_t^{(j)} = \sum_{k=1}^{p} h_{jk}^{(L)}(X_{t-1}^{(k)}) + r_t^{(j)} + \epsilon_t^{(j)}, \text{ where } r_t^{(j)} = \sum_{k=1}^{p} [h_{jk}(X_{t-1}^{(k)}) - h_{jk}^{(L)}(X_{t-1}^{(k)})] \tag{16}$$

is the reminder term and captures the bias introduced by truncating the basis expansion.

We now define the oracle coefficients on the population level for the basis expansion and the design matrix. For any $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, set vectors

$$\begin{aligned}
b_{j,k}^* &= (b_{j,k}^{(1)*}, \ldots, b_{j,k}^{(L)*})^\top, \\
b_j^* &= (b_{j,1}^{*\top}, \ldots, b_{j,p}^{*\top})^\top, \\
b^* &= (b_1^{*\top}, \ldots, b_p^{*\top})^\top, \\
\psi_k(x_k) &= (\psi_{k,1}(x_k), \ldots, \psi_{k,l}(x_k))^\top, \\
\psi(x) &= (\psi_1^\top(x_1), \ldots, \psi_p^\top(x_p))^\top.
\end{aligned} \tag{17}$$

Let $r_t = (r_t^{(1)}, \ldots, r_t^{(p)})^\top$. With these definitions, the model can be rewritten in a compact form as

$$X_t := \Psi(X_{t-1})^\top b^* + r_t + \epsilon_t, \tag{18}$$

where

$$\Psi(X_{t-1}) = \begin{pmatrix}
\psi(X_{t-1}) & 0 & 0 & \cdots & 0 \\
0 & \psi(X_{t-1}) & 0 & \cdots & 0 \\
0 & 0 & \psi(X_{t-1}) & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \psi(X_{t-1})
\end{pmatrix} \in \mathbb{R}^{p \times p^2 L}.$$

Consequently, the solution to our optimization problem (14) can be approximately estimated by solving

$$\hat{b} := \operatorname*{argmin}_b \left\{ \frac{1}{n} \sum_{t=1}^{n} \|X_t - \Psi(X_{t-1})^\top b\|_2^2 + \lambda \sum_{j=1}^{p} \sum_{k=1}^{p} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(\sum_{l=1}^{L} \psi_{k,l}(X_{t-1}^{(k)}) b_{j,k}^{(l)} \right)^2} \right\}. \tag{19}$$

This formulation can be viewed as a functional version of the group lasso, and the standard convexity arguments guarantee the existence of a minimizer.

Compared with the approach in Lim et al. (2015), which employs operator-valued reproducing kernels for VAR models, our formulation offers a key advantage: it decouples the smoothness and sparsity components. This separation allows us to employ a block coordinate descent algorithm (cf. Ravikumar et al. (2009)) to efficiently construct the estimator. In the following section, we leverage the technical tools developed in Section 2 to establish the theoretical properties of our ℓ_1-regularized estimator, under the assumption that the particular smoother in (19) is used.

3.2 Asymptotic properties

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

Assumption 5. There exist constants $\phi_U, \phi_L > 0$, so that

$$\lambda_{\min}\left\{\mathbb{E}\psi(X_{t-1})\psi(X_{t-1})^\top\right\} \geqslant \phi_L, \tag{20}$$

and

$$\max_{1\leqslant k\leqslant p} \lambda_{\max}\left\{\mathbb{E}\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top\right\} \leqslant \phi_U. \tag{21}$$

Condition (20) in Assumption 5 is similar to the smallest population eigenvalue conditions commonly used in high-dimensional statistics (Raskutti et al., 2011, van de Geer et al., 2014). In addition, it parallels the population minimum eigenvalue condition in Assumption 4 of Chen and Christensen (2015) and Assumption S.3 of Belloni et al. (2019) for sieve basis expansion functions. If the marginal density of $X_t^{(k)}$ satisfies $0 < f_{\min} \leqslant f_k(x) \leqslant f_{\max} < \infty$ for $1 \leqslant k \leqslant p$ and almost all $x \in [-c_0, c_0]$, then

$$\mathbb{E}u^\top\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top u = \int_{-c_0}^{c_0} (u^\top\psi_k(x))^2 f_k(x)\,\mathrm{d}x \leqslant f_{\max}\,\|u\|_2^2,$$
$$\mathbb{E}u^\top\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top u = \int_{-c_0}^{c_0} (u^\top\psi_k(x))^2 f_k(x)\,\mathrm{d}x \geqslant f_{\min}\,\|u\|_2^2.$$

This verifies condition (21) in Assumption 5. In the following Proposition 1, we use concentration inequalities to establish the sample version of Assumption 5.

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

(i). Assume that (20) holds and that for some constant $c_1 > 0$ *does not rely on* p, L, *such that for all* $u \in \mathbb{R}^{pL}$,

$$\mathbb{E}(u^\top\psi(X_t)\psi(X_t)^\top u)^2 \leqslant c_1\big(u^\top\mathbb{E}(\psi(X_t)\psi(X_t)^\top)u\big)^2. \tag{22}$$

Then, with probability at least $1 - p^{-c_2} - pe^{-c_3 n/\log(n)}$, *for all* $u \in \mathbb{R}^{pL}$ *with* $\|u\|_2 = 1$,

$$\frac{1}{n}\sum_{t=1}^n u^\top\psi(X_t)\psi(X_t)^\top u \geqslant \frac{\phi_L}{2} - \frac{1}{n} - c_4\frac{\log(n)\log(pL)\cdot\|u\|_1^2}{n}, \tag{23}$$

where $c_2, c_3, c_4 > 0$ *are constants independent of* n, p, L.

(ii). Assume that (21) holds. Then, with probability at least $1 - p^{-c_5} - e^{-c_6 n/\log(n)}$, *for all* $u \in \mathbb{R}^L$ *with* $\|u\|_2 = 1$,

$$\max_{1\leqslant k\leqslant p}\frac{1}{n}\sum_{t=1}^n u^\top\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top u \leqslant \phi_U + c_7 L\sqrt{\frac{\log(n)(\log p + \log L)}{n}}, \tag{24}$$

where $c_5, c_6, c_7 > 0$ *are constants independent of* n, p, L.

Remark 5. Condition (22) is the L_2-L_4 norm equivalence condition for $\psi(X_t)$; see Mendelson and Zhivotovskiy (2020). Let $\xi = w^\top \psi(X_t)$. Then it becomes $\mathbb{E}(\xi^4) \leqslant c_1 \big(\mathbb{E}(\xi^2)\big)^2$, implying that the kurtosis of ξ is bounded. The L_2-L_4 norm equivalence plays an important role in random matrix theory and it holds in various settings, such as sub-Gaussian random vectors. See Mendelson and Zhivotovskiy (2020) for more details and more examples.

In addition, to ensure (23), following Oliveira (2016), condition (22) can be relaxed by letting u be sparse vectors satisfying $\|u\|_0 \leqslant n$. ■

Assumption 6. Let $S := \{(j,k) : h_{jk}(\cdot) \not\equiv 0, 1 \leqslant j,k \leqslant p\}$ and $S_j := \{k : h_{jk}(\cdot) \not\equiv 0, 1 \leqslant k \leqslant p\}$, $1 \leqslant j \leqslant p$. Assume that nonzero indices

$$s_0 := \max_{1 \leqslant j \leqslant p} \sum_{k=1}^{p} \mathbf{1}_{\{h_{jk} \not\equiv 0\}} = \max_{1 \leqslant j \leqslant p} \mathrm{Card}(S_j) = o(p) \text{ and } s := \sum_{j=1}^{p} \sum_{k=1}^{p} \mathbf{1}_{\{h_{jk} \not\equiv 0\}} = \mathrm{Card}(S) = o(p^2).$$

Assumption 6 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. To achieve convergence rates without an additional factor of p, as is typically desired in high-dimensional settings, global boundedness of the quantities in Assumption 6 is usually required, as in Koltchinskii and Yuan (2010b). However, Raskutti et al. (2012) finds an elaborate way to circumvent this requirement when studying sparse additive models with RKHS components.

The following Proposition 2 establishes an upper bound on the remainder term $\|r_t\|_\infty$ as a function of the smoothness level β, the number of basis functions L, and the sparsity level s_0. Moreover, the quantity $\frac{1}{n}\sum_{t=1}^{n}[h_{jk}(X_{t-1}^{(k)}) - h_{jk}^{(L)}(X_{t-1}^{(k)})]^2$ serves as a measure of the L_2 bias between h_{jk} and its orthogonal projection onto the finite-dimensional subspace spanned by the chosen basis functions.

Proposition 2. *Under Assumptions 4 and 6, we have*

$$\|r_t\|_\infty = \max_{1 \leqslant j \leqslant p} \left| \sum_{k=1}^{p} [h_{jk}(X_{t-1}^{(k)}) - h_{jk}^{(L)}(X_{t-1}^{(k)})] \right| \leqslant BC(2\beta-1)^{-1} s_0 L^{1/2-\beta},$$

$$\max_{1 \leqslant j,k \leqslant p} \frac{1}{n} \sum_{t=1}^{n} \left[h_{jk}(X_{t-1}^{(k)}) - h_{jk}^{(L)}(X_{t-1}^{(k)}) \right]^2 \leqslant B^2 C^2 (2\beta-1)^{-2} L^{1-2\beta}.$$

Formally, we have the following asymptotic properties for the ℓ_1 regularized estimators. Theorem 2 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 2. *Suppose Assumptions 1, 2(ii), 4, 5 and 6 hold. Let $\hat{b}$ be the corresponding LASSO solution given in the optimization problem (19). Consider the estimator*

$$\hat{h}_{jk}(x) = \sum_{l=1}^{L} \psi_{k,l}(x) \hat{b}_{j,k}^{(l)}, \qquad 1 \leqslant j,k \leqslant p. \tag{25}$$

Suppose that condition (22) holds. Assume that

$$\lambda \geqslant c_2 \left(\sqrt{\frac{L \log(pL)}{n}} + s_0 L^{1-\beta} \right), \tag{26}$$

for some $c_2 > 0$. Also suppose that

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

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

$$\|\hat{b} - b^*\|_2 \leqslant c_4 \sqrt{s}\lambda, \tag{27}$$

$$\sum_{j=1}^{p}\sum_{k=1}^{p} \|\hat{h}_{jk} - h_{jk}\|_2^2 \leqslant c_5 s\lambda^2 + c_5 s L^{-2\beta}, \tag{28}$$

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{j,k=1}^{p} (\hat{h}_{jk}(X_{t-1}^{(k)}) - h_{jk}(X_{t-1}^{(k)}))^2 \leqslant c_6 s\lambda^2 + c_6 s L^{1-2\beta}, \tag{29}$$

where $c_4, c_5, c_6 > 0$ are constants depending on ρ and μ_e.

Observe that since $s \leqslant s_0 p$, the bounds in (27), (28) and (29) imply that

$$\max_{1\leqslant j\leqslant p} \|\hat{b}_j - b_j^*\|_2 \leqslant c_4\sqrt{s_0}\lambda,$$

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

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{p} (\hat{h}_{jk}(X_{t-1}^{(k)}) - h_{jk}(X_{t-1}^{(k)}))^2 \leqslant c_6 s_0\lambda^2 + c_6 s_0 L^{1-2\beta},$$

where b^* and b_j^* are defined in (17) and (18). The quantity ρ measures the dependence strength of the processes, and the constant μ_e encodes the moment condition. Theorem 2 shows that, provided Assumptions 1 and 2(ii) hold with $\rho \leqslant \rho_0 < 1$ and ρ_0 is a constant, neither the dependence strength nor the moment constant μ_e affects these convergence rates. The second terms in (28) and (29) quantify the bias due to truncating the basis expansion. Moreover, Theorem 2 implies that if the noise ϵ_t has finite exponential moments, then we may allow the dimension p to grow as fast as e^{n^c} for some constant $0 < c < 1$; the exponent c depends on the chosen truncation level L of basis expansion.

It is instructive to compare the two terms in the tuning requirement λ from (26). In the case with relative low dimension $\log(p) \lesssim s_0^2 n L^{1-2\beta}$ and low basis number $L \lesssim s_0^{2/(2\beta-1)}(n/\log n)^{1/(2\beta-1)}$, the basis-expansion bias term $s_0 L^{1-\beta}$ 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 $L \gtrsim s_0^{2/(2\beta-1)}(n/\log n)^{1/(2\beta-1)}$, the stochastic term $(n^{-1}L\log(pL))^{1/2}$ becomes the leading factor.

Remark 6. The convergence rates of the penalized estimators in (28) and (29) contain two sources of bias: (a) the first from the penalty λ, and (b) the second from the truncation parameter L (which depends on the smoothness of the function space, β). ■

Remark 7 (Use of Bernstein-type Inequalities)**.** Bernstein-type inequalities play a crucial role in the theoretical analysis of high-dimensional methods with regularization. Define the loss function

$$F(b) = \frac{1}{n}\sum_{t=1}^{n} \|X_t - \Psi(X_{t-1})^\top b\|_2^2 + \lambda \sum_{j,k=1}^{p} \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\psi_k(X_{t-1}^{(k)})^\top b_{j,k})^2},$$

and define

$$\Sigma_k = \frac{1}{n}\sum_{t=1}^{n} \psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top \quad \text{and} \quad J_n = \frac{1}{n}\sum_{t=1}^{n} \Psi(X_{t-1})\Psi(X_{t-1})^\top.$$

Following the standard proof technique for regularized estimators (Negahban et al., 2012), we compare $F(\hat{b})$ to $F(b^*)$, where $\hat{b}$ minimizes $F(b)$, to obtain

$$0 \geqslant F(\hat{b}) - F(b^*) = -2\nabla_n^\top(\hat{b} - b^*) + (\hat{b} - b^*)^\top J_n(\hat{b} - b^*) + \lambda \sum_{j,k=1}^{p} \big(\|\Sigma_k^{1/2}\hat{b}_{j,k}\|_2 - \|\Sigma_k^{1/2}b^*_{j,k}\|_2\big),$$

where ∇_n is the gradient of the least squares loss, defined in (56) below. In our analysis, Theorem 1 is not applied verbatim in the proof of Theorem 2, but its underlying arguments and closely related concentration inequalities are used. First, we establish a high probability bound on $|\nabla_n|_{2,\infty}$, where $|\cdot|_{2,\alpha}$ is defined in (52) below. In particular, Lemma 6 requires an exponential-type tail probability bound for $\frac{1}{n}\sum_{t=1}^{n} g(X_{t-1})\epsilon_t^{(j)}$ analogous to the bound in Theorem 1. Next, we need a high probability bound for the quadratic term $(\hat{b}-b^*)^\top J_n(\hat{b}-b^*)$. Obtaining this bound also relies on Bernstein-type inequalities, as generalized in Lemma 5. However, because of temporal dependence, the quantities in Lemmas 5 and 6 involve quadratic forms or noise terms rather than simple Lipschitz functions g of X_t as in Assumption 3, so Theorem 1 cannot be applied directly. We therefore adapt its technical arguments to establish a corresponding exponential-type tail probability bound and then use those bounds to prove Theorem 2.

■

Remark 8. Our framework in Theorem 2 is quite general: it accommodates a broad class of nonlinear VAR processes whose innovations need not be sub-Gaussian. By contrast, Han et al. (2015) and Basu and Michailidis (2015) focus on linear VAR models with i.i.d. Gaussian errors, estimating the transition matrix. Like those linear VAR analyses, we also allow the ambient dimension p to vastly exceed the sample size n.

A crucial distinction arises in the tuning parameter condition (26). The second term on the right, originating from the bias in truncating the basis expansion, enters the gradient of the loss and must be retained when verifying restricted strong convexity (Negahban et al., 2012). Consequently, the truncation level L influences both the choice of λ and the estimator's convergence rate.

In the fully nonlinear setting, one typically requires $L \to \infty$, so the first term $\sqrt{L\log(pL)/n}$ in λ's bound exceeds the familiar $\sqrt{\log(p)/n}$ rate for linear VARs (Basu and Michailidis, 2015). This inflation can be viewed as the statistical "cost of nonlinearity". However, in special cases where each h_{jk} admits an exact (or arbitrarily precise) finite dimensional basis representation, the bias term $s_0 L^{1-\beta}$ in (26) vanishes and the first term collapses to $\sqrt{\log(p)/n}$. Under those circumstances, our nonlinear estimator attains the same tuning and convergence rates as its linear counterpart.

■

Next, we turn to model-selection consistency. In place of Assumptions 5, we present an alternative condition that directly targets the support of each component. To simplify the notation, let $\Psi_{S_j}(X_t) = (\psi_k(X_t^{(k)})^\top, k \in S_j)$ be the truncated feature vector in $\mathbb{R}^{L\cdot\mathrm{Card}(S_j)}$, where ψ_k is defined

in (17). We then assemble these vectors into the block-diagonal matrix

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

Assumption 7. There are some constants $\phi_{\max}, \phi_{\min} > 0, 0 < \delta \leqslant 1$, so that

$$\min_{1\leqslant j\leqslant p} \lambda_{\min}\Big\{\mathbb{E}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\} \geqslant \phi_{\min} > 0, \tag{30}$$

$$\max_{1\leqslant j\leqslant p} \lambda_{\max}\Big\{\mathbb{E}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\} \leqslant \phi_{\max} < \infty, \tag{31}$$

and

$$\max_{1\leqslant j\leqslant p} \Big\|\Big(\mathbb{E}\Psi_{S_j^c}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big)\Big(\mathbb{E}\Psi_{S_j}(X_{l-1})\Psi_{S_j}(X_{l-1})^\top\Big)^{-1}\Big\|_{2,\infty} \leqslant \sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}, \tag{32}$$

where the induced matrix $(2,\infty)$-norm is defined as $\|A\|_{2,\infty} = \max_{1\leqslant j\leqslant m_1}\sqrt{\sum_{k=1}^{m_2} A_{jk}^2}$ for $A \in \mathbb{R}^{m_1\times m_2}$.

This assumption corresponds to the condition of Ravikumar et al. (2009, 2010). Similar to Assumption 5, (30) and (31) are also standard, and are commonly imposed for high-dimensional regression analysis. Besides, (32) relates to the incoherence condition, see e.g. Wainwright (2009), Ravikumar et al. (2010). In the following proposition, we establish a sample version of Assumption 7.

Proposition 3. *Suppose Assumptions 1, 2(ii), 6 and 7 hold. Assume* $\sup_x |\psi_{k,l}(x)| \leqslant B$ *for any* $1 \leqslant k \leqslant p, 1 \leqslant l \leqslant L$. *Assume that (37) holds and that for some constant* $c_1 > 0$ *does not rely on* p, s_0, L, *such that for all* $u \in \mathbb{R}^{\mathrm{Card}(S_j)L}$,

$$\mathbb{E}(u^\top\Psi_{S_j}(X_t)\Psi_{S_j}(X_t)^\top u)^2 \leqslant c_1\big(u^\top\mathbb{E}(\Psi_{S_j}(X_t)\Psi_{S_j}(X_t)^\top)u\big)^2. \tag{33}$$

Then, with probability approaching one (as $n, p \to \infty$*), we have*

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

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

and

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

In Theorem 3, we show that, under certain conditions, our method recovers the sparsity pattern asymptotically. Recall $S = \{(j,k) : h_{jk}(\cdot) \not\equiv 0, 1 \leqslant j,k \leqslant p\}$. Then $S = \{(j,k) : b^*_{j,k} \neq 0, 1 \leqslant j,k \leqslant p\}$. Let $\hat{S}_n := \{(j,k) : \hat{b}_{j,k} \neq 0, 1 \leqslant j,k \leqslant p\}$.

Theorem 3. *Suppose Assumptions 1, 2(ii), 4, 6 and 7 hold. Let $\hat{b}$ be the corresponding LASSO solution given in the optimization problem (19). Let $\beta > 3/2$. Suppose that condition (33) holds. Assume that*

$$\frac{s_0 L^2 \cdot \log(pL)}{n} + s_0 L^{1-2\beta/3} \to 0, \tag{37}$$

and

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

Then the solution $\hat{b}$ to problem (19) is unique and satisfies $\hat{S}_n = S$, with probability approaching one (as $n, p \to \infty$).

In a p-dimensional vector time series, the pattern of direct influences among variables can be represented by a binary adjacency matrix $A = (a_{jk}) \in \{0,1\}^{p\times p}$, where

$$a_{jk} = \begin{cases} 1, & \text{if variable } k \text{ directly influences variable } j, \\ 0, & \text{otherwise.} \end{cases}$$

In a linear VAR model $X_t = G\,X_{t-1} + \epsilon_t$, this network structure is typically inferred from the nonzero entries of the transition matrix G, which is often assumed to be sparse (Hall et al., 2018). A theory-free principle was advocated in Sims (1980) for inferring economic relations between variables of linear VARs.

In our nonlinear VAR framework, each component function h_{jk} quantifies the influence of k on j. Moreover, the group lasso formulation in (19) yields a sparse estimate $\hat{b}$, so that many blocks $\hat{b}_{j,k}$ are exactly zero. We therefore define the estimated adjacency matrix $\hat{A} = (\hat{a}_{jk})$ by

$$\hat{a}_{jk} = \begin{cases} 1, & \text{if } \hat{b}_{j,k} \neq 0, \\ 0, & \text{if } \hat{b}_{j,k} = 0. \end{cases}$$

Since $\hat{A}$ need not be symmetric, it encodes a directed graph. Our Theorem 3 then guarantees model selection consistency for $\hat{A}$, ensuring that the true influence network is recovered with high probability. We demonstrate the proposed network estimation method on real data in Section 5.

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, $1 \leqslant j \leqslant p$, 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 $1 \leqslant j \leqslant p$, we randomly select 5 functions h_{jk} $(1 \leqslant k \leqslant p)$ to be the above nonzero functions. The rest $p - 5$ functions of h_{jk} $(1 \leqslant k \leqslant p)$ 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 procedure 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

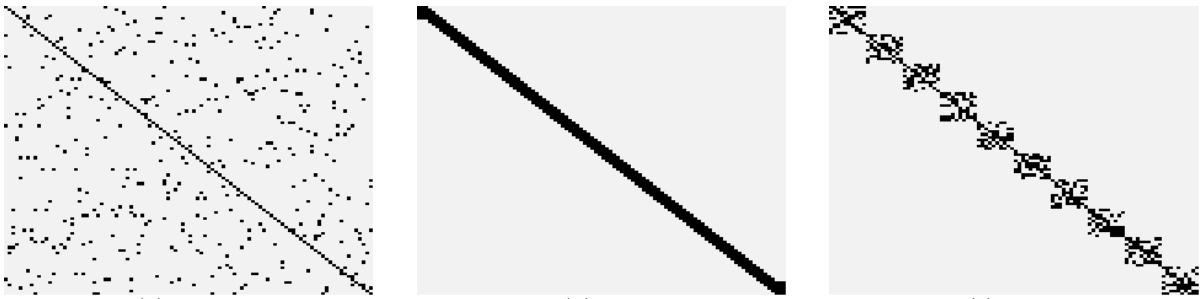
$$\text{TPR} = \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

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 ($n \geqslant 100$), 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.

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,



(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

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 and the number of basis function L 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.

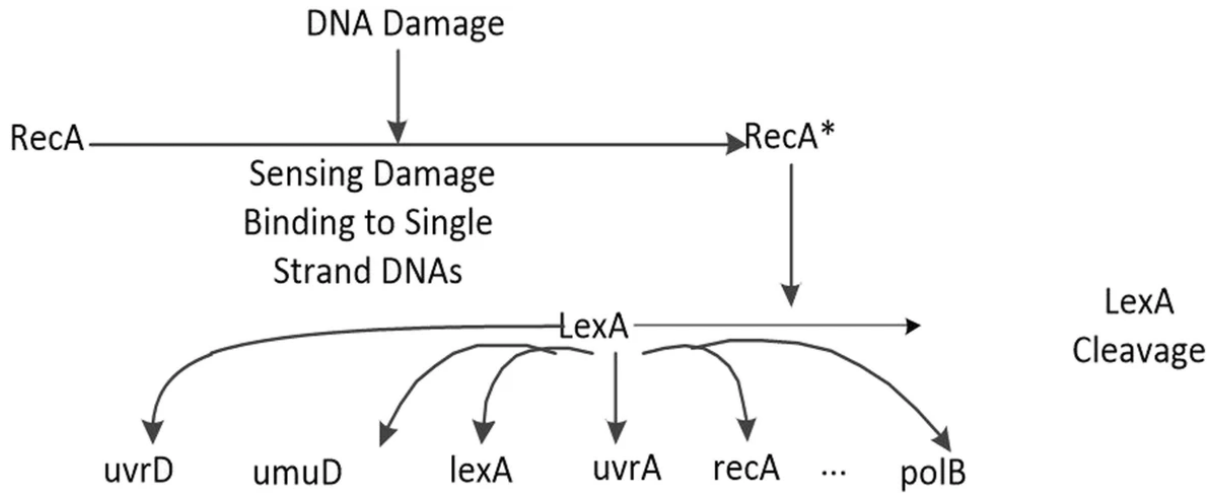


Figure 2: The bacterial SOS DNA repair system

6 Proofs

Write random variable $\boldsymbol{\xi} \in \mathcal{L}^m$, $m \geqslant 1$, if the m-norm $\|\boldsymbol{\xi}\|_m := (\mathbb{E}|\boldsymbol{\xi}|^m)^{1/m} < \infty$. Denote $\|\boldsymbol{\xi}\| := \|\boldsymbol{\xi}\|_2$. 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 $(\epsilon_k)_{k\in\mathbb{Z}}$, so that ϵ_i, ϵ'_j, $i, j \in \mathbb{Z}$ are i.i.d. For any $X_t = \mathcal{G}(\ldots, \epsilon_{t-1}, \epsilon_t)$, where $\mathcal{G}$ is a measurable function, we define the coupled version $X_{t,\{k\}} = \mathcal{G}(\ldots, \epsilon_{k-1}, \epsilon'_k, \epsilon_{k+1}, \ldots, \epsilon_t)$. If $k > i$, then $X_{t,\{k\}} = X_t$.

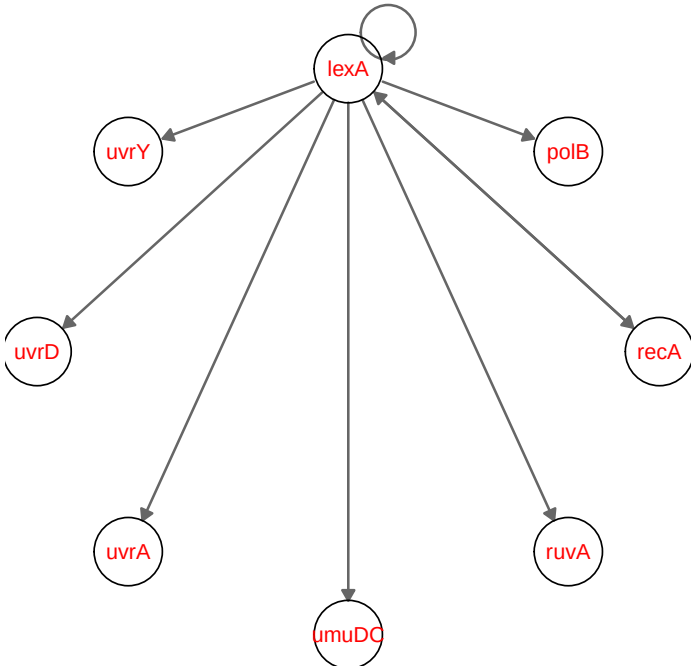


Figure 3: The target SOS DNA repair network

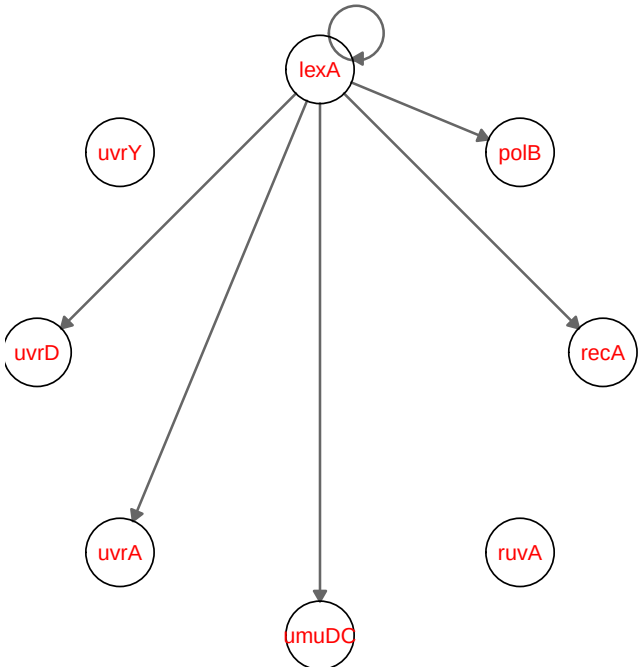


Figure 4: Reconstruction of SOS DNA repair network by nonlinear VAR model

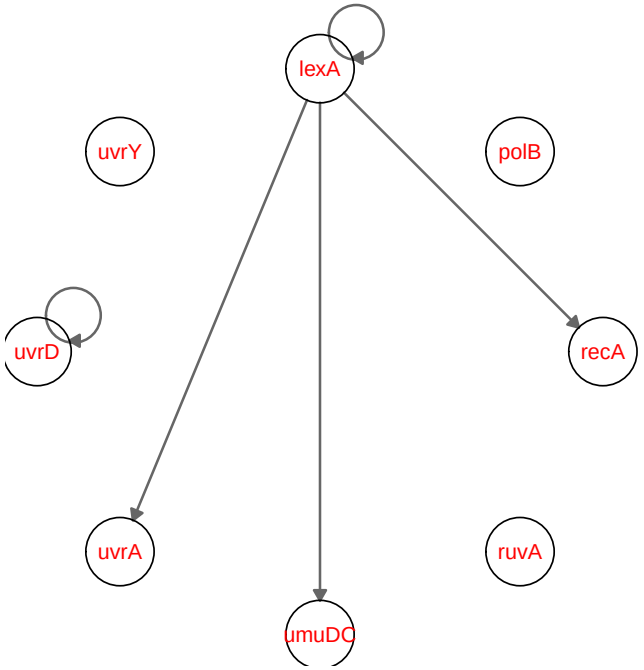


Figure 5: Reconstruction of SOS DNA repair network by linear VAR model

6.1 Proofs of Theorems in Section 2

Lemma 1 (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*

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

Lemma 2. *Let $\epsilon \in \mathbb{R}^p$ be a random vector with non-negative entries, satisfying Assumption 2(i) with $\mu_q < \infty$ for some $q \geqslant 2$. For non-negative vectors $v_t \in \mathbb{R}^p$, $i \geqslant 0$, assume $\|v_t\|_1 \leqslant \rho^t$ where $\rho < 1$. Denote*

$$X = \sum_{t=0}^{\infty} \min\left\{v_t^\top \epsilon, M\right\}.$$

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

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

Proof. Note that we have the decomposition

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

For I_1 part, we have for any $m \geqslant 1$,

$$\mathbb{E}|\mathrm{I}_1|^m \leqslant M^m \left(\sum_{t=0}^{\infty} \left\| \mathbf{1}_{\{v_t^\top \epsilon \geqslant M\}} \right\|_m \right)^m = M^m \Big(\sum_{t=0}^{\infty} \mathbb{P}(v_t^\top \epsilon \geqslant M)^{1/m} \Big)^m. \tag{39}$$

By Markov's inequality,

$$\mathbb{P}(v_t^\top \epsilon \geqslant M) \leqslant \|v_t^\top \epsilon\|_2^2 / M^2 \leqslant \rho^{2i} \mu_2^2 / M^2.$$

Applying above into (39), we further have

$$\mathbb{E}|\mathrm{I}_1|^m \leqslant M^m \Big(\mu_2^{2/m} M^{-2/m} \sum_{t=0}^{\infty} \rho^{2i/m} \Big)^m \leqslant \mu_2^2 (1-\rho^{2/m})^{-m} M^{m-2}.$$

Since for any $m \geqslant 1$, we have

$$1 - \rho^{2/m} \geqslant (1-\rho^2)/m. \tag{40}$$

We further obtain

$$\mathbb{E}|\mathrm{I}_1|^m \leqslant \mu_2^2 ((1-\rho^2)/m)^{-m} M^{m-2}.$$

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

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

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

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

where the last inequality is by (40). Therefore

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

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

$$\mathbb{E}e^{c_0 X/M} - 1 - \mathbb{E}(c_0 X/M) = \sum_{m \geqslant 2} \frac{\mathbb{E}((c_0 X/M)^m)}{m!} \leqslant \sum_{m \geqslant 2} \frac{\mathbb{E}((c_{1,M}\mathrm{I}_1)^m)}{m!} + \sum_{m \geqslant 2} \frac{\mathbb{E}((c_{1,M}\mathrm{I}_2)^m)}{m!} \leqslant \mu_2^2 M^{-2}.$$

□

Proof of Theorem 1. Part (i). Without loss of generality, assume $\|G\|_1 = 1$. For $X_t = \mathcal{G}(\ldots, \epsilon_{t-1}, \epsilon_t)$, where $\mathcal{G}$ is some measurable function, we define the coupled version

$$X_{t,\{k\}} = \mathcal{G}(\ldots, \epsilon_{k-1}, \epsilon_k', \epsilon_{k+1}, \ldots, \epsilon_t).$$

Recall that $\mathrm{abs}(x) = (|x_1|, \ldots, |x_p|)^\top$ for $x = (x_1, \ldots, x_p)^\top$. By Assumption 1, for $k \leqslant i-1$, we have

$$\mathrm{abs}(X_t - X_{t,\{k\}}) \leqslant H\mathrm{abs}(X_{t-1} - X_{t-1,\{k\}}),$$

and for $k = i$, $\mathrm{abs}(X_t - X_{t,\{k\}}) = \epsilon_k - \epsilon_k'$. Hence by induction, we obtain

$$\mathrm{abs}(X_t - X_{t,\{k\}}) \leqslant H^{t-k}\mathrm{abs}(\epsilon_k - \epsilon_k').$$

Since the function g is Lipschitz continuous, combined with the above inequality, we have

$$\begin{aligned} |P_k g(X_t)| &= \big|\mathbb{E}(g(X_t) - g(X_{t,\{k\}})|\mathcal{F}_k)\big| \\ &\leqslant \mathbb{E}\left(G^\top \mathrm{abs}(X_t - X_{t,\{k\}})\Big|\mathcal{F}_k\right) \\ &\leqslant \mathbb{E}\left(G^\top H^{t-k}\mathrm{abs}(\epsilon_k - \epsilon_k')\Big|\mathcal{F}_k\right). \end{aligned} \tag{41}$$

Let $S_n(g) = \sum_{t=1}^n \big(g(X_t) - \mathbb{E}g(X_t)\big)$. For $k \leqslant n$, denote $\xi_k = P_k(S_n(g))$. Then

$$S_n(g) = \sum_{k \leqslant n} \xi_k.$$

The tail probability can be decomposed into two parts

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

In the following, we will first bound ξ_k and then address I_1 and I_2 separately. The first part I_1 is the leading term, while the second part I_2 is relatively small. By Assumption 1 and $\|G\|_1 \leqslant 1$, we have

$$\|H^{t-k\top}G\|_1 \leqslant \|H\|_\infty^{t-k}\|G\|_1 \leqslant \rho^{t-k}.$$

Denote $v_t = H^{t\top}G$. Since $|g|_\infty \leqslant M$, we have $|P_k g(X_t)| \leqslant 2M$. Thus by (41),

$$|\xi_k| \leqslant \sum_{t=1}^n |P_k g(X_t)| \leqslant \sum_{t=k\vee 1}^n \min\Big\{v_{t-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')|\mathcal{F}_k), 2M\Big\}, \quad \text{with } \|v_t\|_1 \leqslant \rho^t. \tag{42}$$

For part I_1, let $h^* := -\rho^2(\log\rho)/(4eM)$. By Lemma 2 and (42) for any $0 < h \leqslant h^*$, $\mathbb{E}(e^{|\xi_k|h}) < \infty$. Note that $\mathbb{E}(\xi_k|\mathcal{F}_{k-1}) = 0$. Then

$$\begin{aligned} \mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) &= 1 + \mathbb{E}(e^{\xi_k h} - \xi_k h - 1|\mathcal{F}_{k-1}) \\ &\leqslant 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big]h^2, \end{aligned} \tag{43}$$

in view of $e^x - x \leqslant e^{|x|} - |x|$ 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 2, we have

$$\begin{aligned} \mathbb{E}\Big[\frac{e^{|\xi_k|h} - |\xi_k|h - 1}{h^2}\Big|\mathcal{F}_{k-1}\Big] &\leqslant \mathbb{E}\Big[\frac{e^{|\xi_k|h^*} - |\xi_k|h^* - 1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big] \\ &\leqslant (h^*)^{-2}\mu_2^2(2M)^{-2} \\ &\leqslant c_3, \end{aligned} \tag{44}$$

where $c_3 = 4e^2(-\rho^2\log\rho)^{-2}\mu_2^2$. Hence for any $h \leqslant h^*$, by (43) and (44),

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \leqslant 1 + c_3 h^2. \tag{45}$$

By Markov's inequality we have $\mathrm{I}_1 \leqslant e^{-zh}\mathbb{E}[\exp(\sum_{-n<k\leqslant n}\xi_k h)]$. Let $h = \min\{z(4c_3 n)^{-1}, h^*\}$, then by recursively applying (45),

$$\begin{aligned}
\mathrm{I}_1 \leqslant & e^{-zh}\mathbb{E}\Big(e^{\sum_{k=-n+1}^{n-1}\xi_k h}\mathbb{E}(e^{\xi_n h}|\mathcal{F}_{n-1})\Big) \\
\leqslant & e^{-zh}(1+c_3h^2)^{2n} \\
\leqslant & \exp\big(-zh + 2nc_3h^2\big) \\
\leqslant & \exp\Big\{-\frac{z^2}{8c_3 n + c_4 M z}\Big\},
\end{aligned} \tag{46}$$

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

For I_2, by (42), $\|\xi_k\|_q \leqslant \sum_{t=1}^n \rho^{t-k}\mu_q \leqslant \rho^{1-k}(1-\rho)^{-1}\mu_q$, for $k \leqslant 0$. Then by Lemma 1,

$$\begin{aligned}
\mathrm{I}_2 &\leqslant z^{-q}\Big((q-1)\sum_{k\leqslant -n}\|\xi_k\|_q^2\Big)^{q/2} \\
&\leqslant (q-1)^{q/2}z^{-q}\Big(\sum_{k\leqslant -n}\|\xi_k\|_q^2\Big)^{q/2} \\
&\leqslant c_5\rho^{qn}/z^q = c_5 e^{-qn\log(\rho^{-1})}/z^q,
\end{aligned} \tag{47}$$

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 $z \leqslant 2Mn$.

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

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

Denote $v_t = H^{t\top}G$ and $\omega_k = \sum_{t=1\vee k}^n v_{t-k}$. Since (41) still holds, we have

$$|\xi_k| \leqslant \sum_{t=k\vee 1}^n v_{t-k}^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')|\mathcal{F}_k) = \omega_k^\top \mathbb{E}(\mathrm{abs}(\epsilon_k - \epsilon_k')|\mathcal{F}_k). \tag{48}$$

For I_2, $k \leqslant -n$, $\|w_k\|_1 \leqslant \rho^{1-k}/(1-\rho)$. Let $h^* := c_0(1-\rho)/\rho$. By (43) and (44), for any $0 \leqslant h \leqslant h^*$,

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) \leqslant 1 + \mathbb{E}\Big[\frac{e^{|\xi_k|h^*} - |\xi_k|h^* - 1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big]h^2 \leqslant 1 + \frac{\mathbb{E}(e^{|\xi_k|h^*} - 1|\mathcal{F}_{k-1})}{h^{*2}}h^2. \tag{49}$$

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

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

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

$$\begin{aligned}\mathbb{E}(e^{|\xi_k|h^*}-1|\mathcal{F}_{k-1}) &\leqslant \int_0^\infty e^{t\rho^{-k}}\rho^{-k}\mathbb{P}(c_0 u_k^\top \mathrm{abs}(\epsilon_k-\epsilon_k')\geqslant t)\mathrm{d}t \\ &\leqslant \rho^{-k}\int_0^\infty e^{-t(1-\rho^{-k})}\mu_e^2\mathrm{d}t \leqslant \rho^{-k}(1-\rho)^{-1}\mu_e^2. \end{aligned} \tag{50}$$

Since $1+x\leqslant e^x$, by (49) and (50),

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1})\leqslant 1+\rho^{-k}(1-\rho)^{-1}\mu_e^2(h^*)^{-2}h^2\leqslant e^{c_3\rho^{-k}h^2}, \tag{51}$$

where $c_3=\mu_e^2(1-\rho)^{-3}\rho^2 c_0^{-2}$. Recursively applying (51), we can obtain

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

where $c_4=c_3/(1-\rho)$. Similar to (46), we can bound the I_1 part and we complete the proof. □

6.2 Proofs of Theorems in Section 3

By (17), for vector $b=(b_{j,k})_{1\leqslant j,k\leqslant p}$ and $b_{j,k}\in\mathbb{R}^L$, define the $(2,\alpha)$ group structure norm

$$|b|_{2,\alpha}:=||b_{j,k}|_2|_\alpha=\left(\sum_{j=1}^p\sum_{k=1}^p\left(\sum_{l=1}^L(b_{j,k}^{(l)})^2\right)^{\alpha/2}\right)^{1/\alpha}, \tag{52}$$

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

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

Proof of Proposition 2. Note that since basis functions are orthonormal, $\|h_{jk}\|_2=(\sum_{l=1}^\infty(b_{jl}^{(l)*})^2)^{1/2}$. Since basis functions are bounded by B, by Assumption 4, we have

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

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

$$\|r_t\|_\infty\leqslant\sum_{k=1}^p\|h_{jk}-h_{jk}^{(L)}\|_\infty\leqslant BC(2\beta-1)^{-1}s_0L^{1/2-\beta}.$$

Furthermore, we have

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\left[h_{jk}(X_{t-1}^{(k)})-h_{jk}^{(L)}(X_{t-1}^{(k)})\right]^2 &= \frac{1}{n}\sum_{t=1}^{n}\left[\sum_{l\geqslant L+1}\psi_{k,l}(X_{t-1}^{(k)})b_{jk}^{(l)*}\right]^2\\ &\leqslant B^2\left[\sum_{l\geqslant L+1}b_{jk}^{(l)*}\right]^2\\ &\leqslant B^2C^2(2\beta-1)^{-2}L^{1-2\beta}.\end{aligned}$$

Then we obtain the desired result. □

Proof of Proposition 1. We first prove part (i). By (20), we have, for any $u\in\mathbb{R}^{pL}$ with $\|u\|_2=1$,

$$\mathbb{E}u^\top\psi(X_t)\psi(X_t)^\top u\geqslant\phi_L.$$

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

$$\frac{1}{n}\sum_{t=1}^{n}u^\top\mathbb{E}\big(\psi(X_t)\psi(X_t)^\top|\mathcal{F}_{t-m+1}^n\big)u\geqslant\frac{1}{2}u^\top\mathbb{E}\psi(X_t)\psi(X_t)^\top u-\frac{c_2\log(n)\log(pL)}{n}\|u\|_1^2.$$

Note that $L=o(n)$. Let $z=1$ in Lemma 4, we can obtain, with probability at least $1-mp^{-c_1}/12-2mpLe^{-3n/(10m)}-e^{-c_3n}$, for any $u\in\mathbb{R}^{pL}$,

$$\frac{1}{n}\sum_{t=1}^{n}u^\top\big(\psi(X_t)\psi(X_t)^\top\big)u\geqslant\frac{1}{2}u^\top\mathbb{E}\psi(X_t)\psi(X_t)^\top u-\frac{c_2\log(n)\log(pL)}{n}\|u\|_1^2-\frac{1}{n}\|u\|_2^2.$$

Then (23) follows.

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

$$\mathbb{P}\left(\Big|\frac{1}{N}\sum_{t\in\mathcal{N}}\mathbb{E}\Big((\psi_k(X_t^{(k)})\psi_k(X_t^{(k)})^\top)_{l_1,l_2}|\mathcal{F}_{t-m+1}^n\Big)-(\Omega_k)_{l_1,l_2}\Big|\geqslant z\right)\leqslant 2\exp\{-c_3Nz^2\}.$$

Therefore with probability at least $1-2L^2\exp\{-c_3Nz^2\}$, for any $u\in\mathbb{R}^L$ with $\|u\|_2=1$,

$$\Big|\frac{1}{N}\sum_{t\in\mathcal{N}}\mathbb{E}\Big(u^\top\psi_k(X_t^{(k)})\psi_k(X_t^{(k)})^\top u|\mathcal{F}_{t-m+1}^n\Big)-u^\top\Omega_k u\Big|\leqslant Lz.$$

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

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big(u^\top\psi_k(X_t^{(k)})\psi_k(X_t^{(k)})^\top u|\mathcal{F}_{t-m+1}^n\Big)\leqslant\phi_U+c_5L\sqrt{\frac{\log(p)+\log(L)}{N}}.$$

Then (24) follows by combining above and Lemma 4 with $z=1$ and $m=4(-\log\rho)^{-1}\log(n)$. □

Lemma 3. *For $m = o(n)$, denote $N = [(n-1)/m]$ and $\mathcal{N} = \{1, m+1, 2m+1, \ldots, (N-1)m+1\}$. Consider the VAR process (2), suppose Assumptions 1 and 2(ii) hold. Assume that there exists a constant $c > 0$, such that for all $u \in \mathbb{R}^{pL}$, $\mathbb{E}[(u^\top \psi(X_t)\psi(X_t)^\top u)^2] \leqslant c\big(u^\top \mathbb{E}(\psi(X_t)\psi(X_t)^\top)u\big)^2$. Let $N \geqslant C\log(pL)$, where $C > 0$ is a sufficiently large constant. Then, we have, with probability at least $1 - p^{-c_1}/12 - 2pLe^{-3N/10}$,*

$$\forall u \in \mathbb{R}^{pL}, \frac{1}{N}\sum_{t\in\mathcal{N}} u^\top \mathbb{E}\big(\psi(X_t)\psi(X_t)^\top | \mathcal{F}_{t-m+1}^n\big)u \geqslant \frac{1}{2}u^\top \mathbb{E}\psi(X_t)\psi(X_t)^\top u - \frac{c_2\log(pL)}{N}\|u\|_1^2,$$

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

Proof. Recall for any $1 \leqslant k \leqslant p, 1 \leqslant l \leqslant L$, $\sup_x |\psi_{k,l}(x)| \leqslant B$, some $B \geqslant 1$, and $\mathcal{F}_k^n = \{\epsilon_k, \ldots, \epsilon_n\}$. Denote $\Sigma = \mathbb{E}(\psi(X_t)\psi(X_t)^\top)$ and

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

Let $\tilde{\Sigma}_{\text{diag}}$ be the diagonal of $\tilde{\Sigma}_N$. Note that $\mathbb{E}\big(\psi(X_t)\psi(X_t)^\top | \mathcal{F}_{t-m+1}^n\big) = \mathbb{E}\big(\psi(X_t)\psi(X_t)^\top | \mathcal{F}_{t-m+1}^t\big)$ are independent for all $t \in \mathcal{N}$. By Jensen's inequality,

$$\mathbb{E}\Big[\big(\mathbb{E}\big(u^\top\psi(X_t)\psi(X_t)^\top u | \mathcal{F}_{t-m+1}^n\big)\big)^2\Big] \leqslant \mathbb{E}[(u^\top\psi(X_t)\psi(X_t)^\top u)^2] \leqslant c\big(u^\top\mathbb{E}(\psi(X_t)\psi(X_t)^\top)u\big)^2.$$

Then, employing similar arguments as in the proof of Lemmas 5.1 and 5.2 in Oliveira (2016), we can obtain, for $N \geqslant 1568c(c_3+1)\log(pL)$ and $c_3 > 0$,

$$\mathbb{P}\left(\forall u \in \mathbb{R}^{pL}, u^\top\tilde{\Sigma}_N u \geqslant \frac{1}{2}u^\top\Sigma u - \frac{1568c(c_3+1)\log(pL)}{N}\left|\tilde{\Sigma}_{\text{diag}}^{1/2}u\right|_1^2\right) \geqslant 1 - \frac{1}{12}p^{-c_3}. \tag{53}$$

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

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{t\in\mathcal{N}}(\psi_{k,l}(X_t^{(k)})^2 - \mathbb{E}[\psi_{k,l}(X_t^{(k)})^2 | \mathcal{F}_{t-m+1}^n])\right| \geqslant z\right) \leqslant 2\exp\left(-\frac{Nz^2}{2B^4 + 4B^2z/3}\right).$$

Hence, we have

$$\mathbb{P}\left(\max_{1\leqslant k\leqslant p, 1\leqslant l\leqslant L}\left|\frac{1}{N}\sum_{t\in\mathcal{N}}\psi_{k,l}(X_t^{(k)})^2\right| \geqslant 2B^2\right) \leqslant 2pL\exp\left(-10N/3\right).$$

Combining the above inequality with (53), it follows that, with probability at least $1 - p^{-c_3}/12 - 2pLe^{-3N/10}$, for any $u \in \mathbb{R}^{pL}$,

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

□

Lemma 4. *(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. We have*

$$\mathbb{P}\left(\sup_{\|u\|_2=1,\ \|u\|_1^2=s_0L}\left|\sum_{t=1}^n u^\top\big[\psi(X_t)\psi(X_t)^\top - \mathbb{E}\big(\psi(X_t)\psi(X_t)^\top | \mathcal{F}_{t-m+1}^n\big)\big]u\right| \geqslant z\right) \leqslant s_0^2L^2e^{-cn},$$

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 $\mathbb{E}_{t-m+1}(\cdot) = (\cdot) - \mathbb{E}(\cdot|\mathcal{F}^n_{t-m+1})$, then we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\|u\|_2=1,\ \|u\|_1^2=s_0L}\Big|u^\top\sum_{t=1}^n\mathbb{E}_{t-m+1}\big(\psi(X_t)\psi(X_t)^\top\big)u\Big|\geqslant z\right)\\
\leqslant&\mathbb{P}\left(\sup_{\|u\|_2=1,\ \|u\|_1^2=s_0L}\|u\|_1^2\max_{1\leqslant k_1,k_2\leqslant pL}\Big|\sum_{t=1}^n\mathbb{E}_{t-m+1}\big((\psi(X_t)\psi(X_t)^\top)_{k_1,k_2}\big)\mathbf{1}_{u_{k_1},u_{k_2}\neq 0}\Big|\geqslant z\right)\\
\leqslant&s_0^2L^2\max_{1\leqslant k_1,k_2\leqslant pL}\mathbb{P}\left(\Big|\sum_{t=1}^n\mathbb{E}_{t-m+1}\big((\psi(X_t)\psi(X_t)^\top)_{k_1,k_2}\big)\Big|\geqslant z/(s_0L)\right).
\end{aligned}$$

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

$$\phi_1,\phi_2\in\{f:\mathbb{R}^p\to\mathbb{R}|f(x)=\psi_{k,l}(x_k)\text{ for some }1\leqslant k\leqslant p, 1\leqslant l\leqslant L\}$$

such that $(\psi(X_t)\psi(X_t)^\top)_{k_1,k_2}=\phi_1(X_t)\phi_2(X_t)$. Since function $\psi_{k,l}$ satisfies conditions in Lemma 5, we complete the proof. □

Lemma 5. *Consider the VAR process (2), suppose Assumptions 1 and 2(ii) hold. Assume functions $\phi_1,\phi_2:\mathbb{R}^p\to\mathbb{R}$ are both bounded with $|\phi_t|_\infty\leqslant B$, $i=1,2$. For any $x,y\in\mathbb{R}^p$, assume $|\phi_t(x)-\phi_t(y)|\leqslant\beta^\top\mathrm{abs}(x-y)=\sum_{j=1}^p\beta_j|x_j-y_j|$, where $\|\beta\|_1\leqslant 1$. Then we have*

$$\mathbb{P}\Big(\Big|\sum_{t=1}^n\big[\phi_1(X_t)\phi_2(X_t)-\mathbb{E}\big(\phi_1(X_t)\phi_2(X_t)|\mathcal{F}^n_{t-m+1}\big)\big]\Big|\geqslant z\Big)\leqslant e^{-c\min\{n,z\rho^{-m},z^2\rho^{-2m}/n\}},\tag{54}$$

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

Proof. Recall $\mathcal{F}^n_k=\{\epsilon_k,\dots,\epsilon_n\}$. Denote

$$S_n=\sum_{t=1}^n\big[\phi_1(X_t)\phi_2(X_t)-\mathbb{E}\big(\phi_1(X_t)\phi_2(X_t)|\mathcal{F}^n_{t-m+1}\big)\big]\quad\text{and}\quad\xi_k=\mathbb{E}(S_n|\mathcal{F}^n_{k-1})-\mathbb{E}(S_n|\mathcal{F}^n_k).$$

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

$$\begin{aligned}
|\xi_k|\leqslant&\sum_{t=(k+m-1)\vee 1}^n\mathbb{E}\Big(|\phi_1(X_{t,\{k\}})-\phi_1(X_t)||\phi_2(X_t)|\big|\mathcal{F}^n_k\Big)\\
&+\sum_{t=(k+m-1)\vee 1}^n\mathbb{E}\Big(|\phi_1(X_{t,\{k\}})||\phi_2(X_{t,\{k\}})-\phi_2(X_t)|\big|\mathcal{F}^n_k\Big)=:\xi_{1k}+\xi_{2k}.
\end{aligned}\tag{55}$$

Since $|\phi_1(X_{t,\{k\}})-\phi_1(X_t)|\leqslant\beta^\top H^{t-k}\mathrm{abs}(\epsilon'_k-\epsilon_k)$ and $|\phi_1|_\infty\leqslant B$, we have

$$\xi_{1k}\leqslant\sum_{t=(k+m-1)\vee 1}^n B\cdot\mathbb{E}\Big(\beta^\top H^{t-k}\mathrm{abs}(\epsilon'_k-\epsilon_k)\big|\mathcal{F}^n_k\Big).$$

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

$$|\xi_k| \leqslant \mathbb{E}\big(\omega_k^\top \mathrm{abs}(\epsilon_k' - \epsilon_k)|\mathcal{F}_k^n\big), \text{ where } \omega_k^\top = 2B\beta^\top \sum_{t=(k+m-1)\vee 1}^{n} H^{t-k}.$$

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

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

Thus by Markov's inequality

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

For $k > -n$, let $h^* = (2B)^{-1}(1-\rho)\rho c_0$ and $\xi_k' = \xi_k/\rho^m$ Then $\mathbb{E}\exp(h^*|\xi_k'|) \leqslant 2\mu_e < \infty$. By (43), (44) and (45), we have for any $h \leqslant h^*$,

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

where $c_1 = 2\mu_e h^{*-2}$. Similar as (46), we have

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

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

Remark 9. The proof of Lemma 5 follows a similar approach to that of Theorem 1. ■

Proof of Theorem 2. Let

$$F(b) = \frac{1}{n}\sum_{t=1}^{n} \|X_t - \Psi(X_{t-1})^\top b\|_2^2 + \lambda \sum_{j,k=1}^{p} \sqrt{\frac{1}{n}\sum_{t=1}^{n} (\psi_k(X_{t-1}^{(k)})^\top b_{j,k})^2}.$$

Define

$$\nabla_n = \frac{1}{n}\sum_{t=1}^{n} \Psi(X_{t-1})(X_t - \Psi(X_{t-1})^\top b^*). \tag{56}$$

Recall the definition of $|\cdot|_{2,\alpha}$ in (52). Then

$$\begin{aligned} |\nabla_n|_{2,\infty} &= \big| \frac{1}{n}\sum_{t=1}^{n} \Psi(X_{t-1})(\epsilon_t + r_t) \big|_{2,\infty} \\ &\leqslant \frac{1}{n}\sum_{t=1}^{n} L^{1/2} \big\|\Psi(X_{t-1})\big\|_\infty \|r_t\|_\infty + \big| \frac{1}{n}\sum_{t=1}^{n} \Psi(X_{t-1})\epsilon_t \big|_{2,\infty} \\ &:= \mathrm{I}_1 + \mathrm{I}_2. \end{aligned} \tag{57}$$

For I_1 part, by (18) and Proposition 2, we have $\|\Psi(X_{t-1})\|_\infty \leqslant B$ and thus $\mathrm{I}_1 \leqslant B^2C(2\beta-1)^{-1}s_0L^{1-\beta}$. For I_2 part, by Lemma 6, with probability at least $1-(pL)^{-c'}$, $\mathrm{I}_2 \leqslant c\sqrt{L\log(pL)/n}$, for some constants $c, c' > 0$.

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

$$\lambda \geqslant (12/\phi_L)\big(c\sqrt{L\log(pL)/n} + B^2C(2\beta-1)^{-1}s_0L^{1-\beta}\big) \geqslant 12|\nabla_n|_{2,\infty}/\phi_L.$$

Let

$$\widetilde{\phi}_L = \frac{\phi_L}{2} - \frac{1}{n} - \frac{c_4(s_0L)\log(n)\log(pL)}{n},$$

and

$$\widetilde{\phi}_U = \phi_U + c_7L\sqrt{\frac{\log(n)\log(pL)}{n}},$$

where $\|u\|_1 = s_0L$ in Proposition 1, and c_4, c_7 are the constants in (23) and (24). Then, for $n \geqslant c_3(s_0L)\log(n)\log(pL) + c_3L^2\log(n)\log(pL)$ with sufficient large constant $c_3 > 0$, we have

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

Denote

$$\Sigma_k = \frac{1}{n}\sum_{t=1}^{n}\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top \quad \text{and} \quad J_n = \frac{1}{n}\sum_{t=1}^{n}\Psi(X_{t-1})\Psi(X_{t-1})^\top.$$

Hence, by Assumption 5 and Proposition 1, with probability approaching one, we have

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

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

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

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

Furthermore,

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

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

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

which also implies (28).

Moreover,

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\sum_{j,k=1}^{p}(\hat{h}_{jk}(X_{t-1}^{(k)})-h_{jk}^{(L)}(X_{t-1}^{(k)}))^2 &= (\hat{b}-b^*)^\top J_n(\hat{b}-b^*)\\ &\lesssim \sum_{j,k=1}^{p}(\hat{b}-b^*)^2=\|\hat{b}-b^*\|_2^2\\ &\lesssim s\lambda^2.\end{aligned}$$

By Proposition 2, we can obtain (29). □

Lemma 6. *For function $g:\mathbb{R}^p\to\mathbb{R}$, assume $|g|_\infty\leqslant B$. Under Assumption 2(ii), we have*

$$\mathbb{P}\Big(\big|\frac{1}{n}\sum_{t=1}^{n}g(X_{t-1})\epsilon_t^{(j)}\big|\geqslant z\Big)\leqslant\begin{cases}2\exp\Big(-\frac{nz^2}{4c_1}\Big), & \text{if } z\leqslant 2c_0c_1B^{-1},\\ 2\exp\big(-c_0nz/(2B)\big), & \text{if } z>2c_0c_1B^{-1},\end{cases}\tag{58}$$

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

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

$$\begin{aligned}\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1}) &= 1+\mathbb{E}(e^{\xi_k h}-\xi_k h-1|\mathcal{F}_{k-1})\\ &\leqslant 1+\mathbb{E}\Big[\frac{e^{|\xi_k|h}-|\xi_k|h-1}{h^2}\Big|\mathcal{F}_{k-1}\Big]h^2.\end{aligned}\tag{59}$$

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

$$\mathbb{E}\Big[\frac{e^{|\xi_k|h}-|\xi_k|h-1}{h^2}\Big|\mathcal{F}_{k-1}\Big]\leqslant\mathbb{E}\Big[\frac{e^{|\xi_k|h^*}-|\xi_k|h^*-1}{h^{*2}}\Big|\mathcal{F}_{k-1}\Big]\leqslant\frac{\mathbb{E}(e^{Bh^*|\epsilon_t^{(j)}|})}{h^{*2}}\leqslant c_1,\tag{60}$$

where $c_1=\mu_e B^2c_0^{-2}$. Combining (59) and (60), we can obtain

$$\mathbb{E}(e^{\xi_k h}|\mathcal{F}_{k-1})\leqslant 1+c_1h^2.$$

Then, by recursively applying the above inequality, we have

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\sum_{t=1}^{n}\xi_t \geqslant z\Big) &\leqslant e^{-nzh}\mathbb{E}\Big(e^{\sum_{t=1}^{n-1}\xi_t h}\mathbb{E}(e^{\xi_n h}|\mathcal{F}_{n-1})\Big)\\ &\leqslant e^{-nzh}(1+c_1h^2)^n\\ &\leqslant \exp\big(-nzh+nc_1h^2\big).\end{aligned}$$

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

$$\mathbb{P}\Big(\frac{1}{n}\sum_{t=1}^{n}\xi_t \geqslant z\Big) \leqslant \exp\Big(-\frac{nz^2}{4c_1}\Big)\mathbf{1}_{\{h^*\geqslant z/(2c_1)\}} + \exp\big(-c_0nz/(2B)\big)\mathbf{1}_{\{h^*<z/(2c_1)\}}.$$

Similar argument can be applied to $\mathbb{P}(n^{-1}\sum_{t=1}^{n}\xi_t \leqslant -z)$ and the desired result follows. □

Remark 10. The proof of Lemma 6 follows a similar approach to that of Theorem 1. ■

Proof of Proposition 3. Note that

$$\begin{aligned}\lambda_{\min}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_S(X_{t-1})\Psi_S(X_{t-1})^\top\Big\} &= \min_{1\leqslant j\leqslant p}\lambda_{\min}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\},\\ \lambda_{\max}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_S(X_{t-1})\Psi_S(X_{t-1})^\top\Big\} &= \max_{1\leqslant j\leqslant p}\lambda_{\max}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\}.\end{aligned}$$

Then, under (37), applying (33) and similar arguments in the proofs of Proposition 1, we have, in an event Ω_1 with probability approaching one (as $n,p\to\infty$),

$$\begin{aligned}\min_{1\leqslant j\leqslant p}\lambda_{\min}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\} &\geqslant (1+o(1))\phi_{\min}>0,\\ \max_{1\leqslant j\leqslant p}\lambda_{\max}\Big\{\frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top\Big\} &\leqslant (1+o(1))\phi_{\max}>0.\end{aligned}$$

Thus, in the event Ω_1, (34) and (35) hold.

Define

$$\begin{aligned}\widehat{Q}_{S_j,S_j} &= \frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top,\\ \widehat{Q}_{S_j^c,S_j} &= \frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j^c}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top,\\ Q_{S_j,S_j} &= \mathbb{E}\Psi_{S_j}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top,\\ Q_{S_j^c,S_j} &= \mathbb{E}\Psi_{S_j^c}(X_{t-1})\Psi_{S_j}(X_{t-1})^\top.\end{aligned}$$

Then, similar to Ravikumar et al. (2010), we decompose the sample matrix as follows

$$\begin{aligned}\widehat{Q}_{S_j^c,S_j}\widehat{Q}^{-1}_{S_j,S_j} &= Q_{S_j^c,S_j}(\widehat{Q}^{-1}_{S_j,S_j}-Q^{-1}_{S_j,S_j}) + (\widehat{Q}_{S_j^c,S_j}-Q_{S_j^c,S_j})Q^{-1}_{S_j,S_j}\\ &\quad + (\widehat{Q}_{S_j^c,S_j}-Q_{S_j^c,S_j})(\widehat{Q}^{-1}_{S_j,S_j}-Q^{-1}_{S_j,S_j}) + Q_{S_j^c,S_j}Q^{-1}_{S_j,S_j}\\ &= \mathrm{I}_1+\mathrm{I}_2+\mathrm{I}_3+\mathrm{I}_4.\end{aligned}$$

Similar to the proofs of Proposition 1, Lemma 4 and Lemma 5, we can show in an event Ω_2 with probability approaching one (as $n, p \to \infty$),

$$\|\widehat{Q}_{S_j^c,S_j} - Q_{S_j^c,S_j}\|_{2,\infty} = o(1).$$

Based on the properties of the induced matrix norms ($\|\cdot\|_{2,2} = \|\cdot\|_2$), we have in the event Ω_1,

$$\begin{aligned}
\|I_1\|_{2,\infty} &\leqslant \|Q_{S_j^c,S_j} Q_{S_j,S_j}^{-1}\|_{2,\infty} \|Q_{S_j,S_j}(\widehat{Q}_{S_j,S_j}^{-1} - Q_{S_j,S_j}^{-1})\|_{2,2} \\
&\leqslant \|Q_{S_j^c,S_j} Q_{S_j,S_j}^{-1}\|_{2,\infty} \|Q_{S_j,S_j}\|_2 \|\widehat{Q}_{S_j,S_j}^{-1} - Q_{S_j,S_j}^{-1}\|_2 \\
&\leqslant o(1) \|Q_{S_j^c,S_j} Q_{S_j,S_j}^{-1}\|_{2,\infty}.
\end{aligned}$$

Similarly, in the event $\Omega_1 \cap \Omega_2$

$$\begin{aligned}
\|I_2\|_{2,\infty} &\leqslant \|\widehat{Q}_{S_j^c,S_j} - Q_{S_j^c,S_j}\|_{2,\infty} \|Q_{S_j,S_j}^{-1}\|_{2,2} = o(1), \\
\|I_3\|_{2,\infty} &\leqslant \|\widehat{Q}_{S_j^c,S_j} - Q_{S_j^c,S_j}\|_{2,\infty} \|\widehat{Q}_{S_j,S_j}^{-1} - Q_{S_j,S_j}^{-1}\|_{2,2} = o(1).
\end{aligned}$$

It follows that in the event $\Omega_1 \cap \Omega_2$,

$$\|\widehat{Q}_{S_j^c,S_j} \widehat{Q}_{S_j,S_j}^{-1}\|_{2,\infty} \leqslant (1 + o(1)) \|Q_{S_j^c,S_j} Q_{S_j,S_j}^{-1}\|_{2,\infty} + o(1).$$

Thus, in the event $\Omega_1 \cap \Omega_2$ with probability approaching one (as $n, p \to \infty$), (34), (35) and (36) hold. □

Proof of Theorem 3. Let $b_S = (b_{j,k}, (j,k) \in S) \in \mathbb{R}^{sL}$, and

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

Denote

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

and

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

By Assumption 7 and Proposition 3, (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 $\hat{b} \in \mathbb{R}^{p^2 L}$ is an optimum of the objective function in (14) if and only if there is a subgradient $\hat{g} \in \partial\Omega(\hat{b})$, such that

$$\frac{2}{n} \sum_{t=1}^{n} \Psi(X_{t-1})(\Psi(X_{t-1})^\top \hat{b} - X_t) + \lambda \hat{g} = 0. \tag{61}$$

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

$$g_{jk} = \frac{\frac{1}{n}\sum_{t=1}^{n} \psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top b_{j,k}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\psi_k(X_{t-1}^{(k)})^\top b_{j,k})^2}}, \tag{62}$$

$$g_{jk}^\top \left(\frac{1}{n}\sum_{t=1}^{n} \psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top\right)^{-1} g_{jk} \leqslant 1. \tag{63}$$

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

$$\hat{b}_{j,k} \neq 0, \quad \text{for } (j,k) \in S, \tag{64}$$

$$\hat{g}_{jk}^\top \left(\frac{1}{n}\sum_{t=1}^{n} \psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top\right)^{-1} \hat{g}_{jk} < 1, \quad \text{for } (j,k) \in S^c, \tag{65}$$

hold with probability approaching 1.

(i). Proof of (64).

Since $\hat{b}_{S^c} = b^*_{S^c} = 0$, (61) reduces to

$$\frac{2}{n}\sum_{t=1}^{n} \Psi_S(X_{t-1})(\Psi_S(X_{t-1})^\top \hat{b}_S - X_t) + \lambda \hat{g}_S = 0. \tag{66}$$

It implies that

$$\hat{b}_S - b^*_S = \hat{\Sigma}_{S,S}^{-1} \cdot \frac{1}{n}\sum_{t=1}^{n} \Psi_S(X_{t-1})\epsilon_t + \hat{\Sigma}_{S,S}^{-1} \cdot \frac{1}{n}\sum_{t=1}^{n} \Psi_S(X_{t-1})r_t - \frac{\lambda}{2}\hat{\Sigma}_{S,S}^{-1} \cdot \hat{g}_S := \mathrm{I}_1 + \mathrm{I}_2 - \mathrm{I}_3. \tag{67}$$

We now proceed to bound $\mathrm{I}_1, \mathrm{I}_2$ and I_3. Recall the definition of $|\cdot|_{2,\alpha}$ in (52). Also recall that $\|A\|_\infty$ is the matrix ∞ norm of $A = (a_{ij})_{n\times m}$ with $\|A\|_\infty = \max_{1\leqslant i\leqslant n}\sum_{j=1}^{m}|a_{ij}|$.

For I_1, we have

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

By Lemma 6, with probability at least $1 - (pL)^{-c_1}$,

$$\left\|\frac{1}{n}\sum_{t=1}^{n} \Psi_S(X_{t-1})\epsilon_t\right\|_\infty \leqslant c_2\sqrt{\frac{\log(pL)}{n}}. \tag{68}$$

Note that

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

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

$$|\mathrm{I}_1|_{2,\infty} \leqslant c_2\sqrt{L}\cdot\frac{\sqrt{s_0 L}}{\phi_{\min}}\cdot\sqrt{\frac{\log(pL)}{n}} = c_2\phi_{\min}^{-1}\frac{L\sqrt{s_0\log(pL)}}{\sqrt{n}}. \tag{69}$$

For I_2, by (18) and Proposition 2, we have

$$|\mathrm{I}_2|_{2,\infty} \leqslant \sqrt{L}\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_\infty \|\Psi_S(X_{t-1})\|_\infty \|r_t\|_\infty \leqslant cB^2C(2\beta-1)^{-1}\phi_{\min}^{-1}s_0^{3/2}L^{3/2-\beta}. \tag{70}$$

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

$$\frac{1}{(1+o(1))\phi_{\max}}\|\hat{g}_{jk}\|_2^2 \leqslant \hat{g}_{jk}^\top\left(\frac{1}{n}\sum_{t=1}^n \psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top\right)^{-1}\hat{g}_{jk} \leqslant 1.$$

It follows that

$$\|\hat{g}_S\|_\infty = \max_{(j,k)\in S}\|\hat{g}_{jk}\|_\infty \leqslant \max_{(j,k)\in S}\|\hat{g}_{jk}\|_2 \leqslant \sqrt{(1+o(1))\phi_{\max}}. \tag{71}$$

Therefore we obtain

$$|\mathrm{I}_3|_{2,\infty} \leqslant \frac{1}{2}\lambda\sqrt{L}\left\|\hat{\Sigma}_{S,S}^{-1}\right\|_\infty\|\hat{g}_S\|_\infty \leqslant \frac{\sqrt{(1+o(1))\phi_{\max}}}{2\phi_{\min}}\cdot\lambda\sqrt{s_0}L. \tag{72}$$

Combining (69), (70) and (72), we have, with probability at least $1-(pL)^{-c_1}$,

$$\begin{aligned}|\hat{b}_S - b_S^*|_{2,\infty} &= \max_{(j,k)\in S}\|\hat{b}_{j,k}-b_{j,k}^*\|_2 \\ &\leqslant c_2\phi_{\min}^{-1}\frac{L\sqrt{s_0\log(pL)}}{\sqrt{n}} + cB^2C(2\beta-1)^{-1}\phi_{\min}^{-1}s_0^{3/2}L^{3/2-\beta} + \frac{\sqrt{(1+o(1))\phi_{\max}}}{2\phi_{\min}}\cdot\lambda\sqrt{s_0}L.\end{aligned} \tag{73}$$

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

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

Since $\max_{(j,k)\in S}\|b_{j,k}^*\|_2 > 0$ and will not converge to 0 asymptotically, (64) holds on an event $\mathcal{Z}_1$ with probability approaching 1.

(ii). Proof of (65).

Since $\hat{b}_{S^c} = b_{S^c}^* = 0$, for all $(j,k)\in S^c$, (61) reduces to

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

It implies that

$$\hat{g}_{jk} = \frac{2}{\lambda}\left(\frac{1}{n}\sum_{t=1}^n\psi_k(X_{t-1}^{(k)})(\Psi_{S_j}(X_{t-1})^\top(b_{S_j}^*-\hat{b}_{S_j}) + \frac{1}{n}\sum_{t=1}^n\psi_k(\epsilon_t^{(j)}+r_t^{(j)})\right).$$

By (67), we have

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

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

$$\hat{g}_{jk}^\top\left(\frac{1}{n}\sum_{t=1}^{n}\psi_k(X_{t-1}^{(k)})\psi_k(X_{t-1}^{(k)})^\top\right)^{-1}\hat{g}_{jk}\leqslant\frac{1}{\phi_{\min}}\left\|\hat{g}_{jk}\right\|_2^2.$$

It suffices to show $\max_{(j,k)\in S^c}\|\hat{g}_{jk}\|_2<\sqrt{(1+o(1))\phi_{\min}}$. 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 (71),

$$\begin{aligned}\|\mathrm{II}_1\|_2 &\leqslant \left\|\frac{1}{n}\sum_{t=1}^{n}\psi_k(X_{t-1}^{(k)})\Psi_{S_j}(X_{t-1})^\top\hat{\Sigma}^{-1}_{S_j,S_j}\right\|_2\|\hat{g}_{S_j}\|_2\\ &\leqslant (1+o(1))\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0}\sqrt{\phi_{\max}}\\ &\leqslant (1+o(1))(1-\delta)\sqrt{\phi_{\min}}.\end{aligned}\tag{74}$$

For II_2, by Lemma 6, as $s_0<n$, with probability at least $1-(nL)^{-c_3}$

$$\begin{aligned}\|\mathrm{II}_2\|_2 &\leqslant \frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0L}\left\|\frac{1}{n}\sum_{t=1}^{n}\Psi_{S_j}(X_{t-1})\epsilon_t^{(j)}\right\|_\infty\\ &\leqslant \frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0L}\cdot c_4\sqrt{\frac{\log(nL)}{n}}\\ &= c_5\frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}}.\end{aligned}\tag{75}$$

For II_3, by (18) and Proposition 2, we have

$$\|\mathrm{II}_3\|_2\leqslant\frac{2}{\lambda}\cdot\sqrt{\frac{\phi_{\min}}{\phi_{\max}}}\cdot\frac{1-\delta}{\sqrt{s_0}}\cdot\sqrt{s_0L}\cdot B^2C(2\beta-1)^{-1}s_0L^{1/2-\beta}=c_6\frac{s_0L^{1-\beta}}{\lambda}.\tag{76}$$

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

$$\|\mathrm{II}_4\|_2 \leqslant c_8 \frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}}. \tag{77}$$

For II_5,

$$\|\mathrm{II}_5\|_2 \leqslant 2B^2C(2\beta-1)^{-1}\frac{s_0L^{1-\beta}}{\lambda} = c_9\frac{s_0L^{1-\beta}}{\lambda}. \tag{78}$$

In view of (74), (75), (76), (77) and (78), for all $(j,k) \in S^c$, we can obtain, with probability at least $1-(nL)^{-c_3}-(nL)^{-c_7}$,

$$\|\hat{g}_{jk}\|_2 \leqslant (1+o(1))(1-\delta)\sqrt{\phi_{\min}} + (c_5+c_8)\frac{1}{\lambda}\sqrt{\frac{L\log(nL)}{n}} + (c_6+c_9)\frac{s_0L^{1-\beta}}{\lambda}. \tag{79}$$

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

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

Hence, (65) holds on an event $\mathcal{Z}_2$ with probability approaching 1. Then Theorem 3 follows.

□

References

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

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

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. J. (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606.

Baddour, K. E. and Beaulieu, N. C. (2005). Autoregressive modeling for fading channel simulation. *IEEE Transactions on Wireless Communications*, 4(4):1650–1662.

Balcilar, M., Thompson, K., Gupta, R., and Van Eyden, R. (2016). Testing the asymmetric effects of financial conditions in South Africa: A nonlinear vector autoregression approach. *Journal of International Financial Markets, Institutions and Money*, 43:30–43.

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

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 66(3):581–605.

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

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120(1):387–422.

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

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

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

Chen, E. Y., Fan, J., and Zhu, X. (2023). Community network auto-regression for high-dimensional time series. *Journal of Econometrics*, 235(2):1239–1256.

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

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

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

Chen, X. and Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465.

Dahlhaus, R. and Richter, S. (2023). Adaptation for nonparametric estimators of locally stationary processes. *Econometric Theory*, 39(6):1123–1153.

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review*, 41(1):45–76.

Douc, R., Guillin, A., and Moulines, E. (2008). Bounds on regeneration times and limit theorems for subgeometric Markov chains. In *Annales de l'IHP Probabilités et statistiques*, volume 44, pages 239–257.

Düker, M.-C. and Waterbury, A. (2025). Kernel estimation for nonlinear dynamics. *arXiv preprint arXiv:2502.18634*.

Fan, J., Jiang, B., and Sun, Q. (2021). Hoeffding's inequality for general Markov chains and its applications to statistical learning. *The Journal of Machine Learning Research*, 22(139):1–35.

Fan, J. and Yao, Q. (2008). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Gao, Z., Ma, Y., Wang, H., and Yao, Q. (2019). Banded spatio-temporal autoregressions. *Journal of Econometrics*, 208(1):211–230.

Ghosh, S., Khare, K., and Michailidis, G. (2019). High-dimensional posterior consistency in Bayesian vector autoregressive models. *Journal of the American Statistical Association*, 114(526):735–748.

Ghosh, S., Khare, K., and Michailidis, G. (2021). Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *The Annals of Statistics*, 49(3):1267–1299.

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

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

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

Jarner, S. and Tweedie, R. (2001). Locally contracting iterated functions and stability of Markov chains. *Journal of Applied Probability*, 38(2):494–507.

Jiang, B., Li, J., and Yao, Q. (2023). Autoregressive networks. *The Journal of Machine Learning Research*, 24(227):1–69.

Jiang, B., Sun, Q., and Fan, J. (2018). Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*.

Kato, H., Taniguchi, M., and Honda, M. (2006). Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Transactions on Signal Processing*, 54(9):3414–3425.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

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

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

Lemańczyk, M. (2021). General Bernstein-like inequality for additive functionals of Markov chains. *Journal of Theoretical Probability*, 34(3):1426–1454.

Lichstein, J. W., Simons, T. R., Shriner, S. A., and Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, 72(3):445–463.

Lim, N., dAlché Buc, F., Auliac, C., and Michailidis, G. (2015). Operator-valued kernel-based vector autoregressive models for network inference. *Machine Learning*, 99(3):489–513.

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

Lin, J. and Michailidis, G. (2020). Regularized estimation of high-dimensional factor-augmented vector autoregressive (FAVAR) models. *The Journal of Machine Learning Research*, 21(117):1–51.

Lütkepohl, H. (2005). New introduction to multiple time series analysis. *NY: Springer*.

Mazur, J., Ritter, D., Reinelt, G., and Kaderali, L. (2009). Reconstructing nonlinear dynamic models of gene regulation using stochastic sampling. *BMC Bioinformatics*, 10(1):448.

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

Mendelson, S. and Zhivotovskiy, N. (2020). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648 – 1664.

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

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

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

Möller, E., Schack, B., Arnold, M., and Witte, H. (2001). Instantaneous multivariate eeg coherence analysis by means of adaptive high-dimensional autoregressive models. *Journal of Neuroscience Methods*, 105(2):143–158.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Oliveira, R. I. (2016). The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166:1175–1194.

Pandit, P., Sahraee-Ardakan, M., Amini, A. A., Rangan, S., and Fletcher, A. K. (2020). Generalized autoregressive linear models for discrete high-dimensional data. *IEEE Journal on Selected Areas in Information Theory*, 1(3):884–896.

Pereda, E., Quiroga, R. Q., and Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2):1–37.

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

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13(1):389–427.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287 – 1319.

Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, 22(1):146–163.

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29.

Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. (2002). 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.

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

Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297.

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

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research*, 23(1):111–120.

Shen, Y., Giannakis, G. B., and Baingana, B. (2019). Nonlinear structural vector autoregressive models with application to directed brain networks. *IEEE Transactions on Signal Processing*, 67(20):5325–5339.

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

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

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202.

Wainwright, M. J. (2009). 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.

Wang, D. and Tsay, R. S. (2023). Rate-optimal robust estimation of high-dimensional vector autoregressive models. *The Annals of Statistics*, 51(2):846–877.

Wang, D., Zheng, Y., Lian, H., and Li, G. (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117(539):1338–1356.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

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

Yu, P.-N., Liu, C. Y., Heck, C. N., Berger, T. W., and Song, D. (2021). A sparse multiscale nonlinear autoregressive model for seizure prediction. *Journal of Neural Engineering*, 18(2):026012.

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

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

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

Zhang, D. and Wu, W. B. (2021). Convergence of covariance and spectral density estimates for high dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233 – 254.

Zheng, L. and Raskutti, G. (2019). Testing for high-dimensional network parameters in auto-regressive models. *Electronic Journal of Statistics*, 13(2):4977 – 5043.

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