# Sparse Principal Component Analysis for High Dimensional Multivariate Time Series

**Zhaoran Wang**
Department of Electrical Engineering,
Princeton University

**Fang Han**
Department of Biostatistics,
Johns Hopkins University

**Han Liu**
Department of Operations Research
and Financial Engineering,
Princeton University

## Abstract

We study sparse principal component analysis (sparse PCA) for high dimensional multivariate vector autoregressive (VAR) time series. By treating the transition matrix as a nuisance parameter, we show that sparse PCA can be directly applied on analyzing multivariate time series as if the data are i.i.d. generated. Under a double asymptotic framework in which both the length of the sample period $T$ and dimensionality $d$ of the time series can increase (with possibly $d \gg T$), we provide explicit rates of convergence of the angle between the estimated and population leading eigenvectors of the time series covariance matrix. Our results suggest that the spectral norm of the transition matrix plays a pivotal role in determining the final rates of convergence. Implications of such a general result is further illustrated using concrete examples. The results of this paper have impacts on different applications, including financial time series, biomedical imaging, and social media, etc.

## 1 Introduction

This paper considers sparse principal component analysis for weakly stationary vector autoregressive (VAR) time series (In this paper, we only consider VAR(1) model, i.e., the model with lag 1. We extend our results to VAR(p) in the longer version of this paper): Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T \in \mathbb{R}^d$ be $T$ observations from a time series $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T$. Here we assume that each $\boldsymbol{X}_t \in \mathbb{R}^d$ is a $d$-dimensional random vector and follows a VAR model

$$\boldsymbol{X}_{t+1} = \mathbf{A}\boldsymbol{X}_t + \boldsymbol{Z}_t, \quad \text{for } t = 1, 2, \ldots, T-1, \tag{1.1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is called transition matrix and $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots \overset{\text{i.i.d}}{\sim} N_d(\mathbf{0}, \mathbf{\Psi})$ are independent colored Gaussian noise with covariance matrix $\mathbf{\Psi}$. Since the process is weakly stationary, we denote $\mathbf{\Sigma}$ to be the covariance matrix of the time series,

$$\mathbf{\Sigma} := \mathrm{Var}(\boldsymbol{X}_1) = \cdots = \mathrm{Var}(\boldsymbol{X}_T). \tag{1.2}$$

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ be the top $m$ leading eigenvectors of $\mathbf{\Sigma}$. We want to find $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_m$ which can estimate $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ accurately. In the special case where $\mathbf{A}$ is a zero matrix, this problem reduces to classical principal component analysis for i.i.d. Gaussian data. In more general settings where $\mathbf{A} \neq \mathbf{0}$, it is well known that to secure the weakly stationarity of the time series in (1.1), we must have the spectral norm of $\mathbf{A}$ (i.e., the largest singular value of $\mathbf{A}$) smaller than 1.

In this paper we consider high dimensional time series under a double asymptotic framework, i.e., we allow the time series dimension $d$ to scale with the length of the sample period $T$ with possibly $d \gg T$. Compared with the classical asymptotic framework for time series in which only $T$ increases while $d$ remains fixed, such a theoretical framework better reflects the challenge in many real-world applications. For example, in fMRI image processing, the machine collects $T$ scans of the human brain, each of which contains $d$ voxels. In a typical setting, the number of scans $T$ is around hundreds, while the number of voxels $d$ could be tens of thousands. In another application on modeling social media stream, e.g., twitter data, we simultaneously monitor the number of tweets for $d$ persons across $T$ time units (e.g. hours). In a typical setting, $T$ could be hundreds or thousands while $d$ could be millions. Other applications include low-frequency stock data in which $T$ represents the number of records of the closing price and $d$ represents the number of stocks in the market.

Such a double asymptotic framework, though more re-

Appearing in Proceedings of the $16^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. 

alistic, poses significant theoretical challenges. Even in a simplified setting where $\mathbf{A} = \mathbf{0}$, Johnstone and Lu (2009) show that the classical PCA is inconsistent under some conditions. In other words, letting the estimator $\widehat{\boldsymbol{u}}_1$ be the leading eigenvector of the sample covariance matrix, the angle between $\boldsymbol{u}_1$ and $\widehat{\boldsymbol{u}}_1$, denoted as $\angle(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1)$, doesn't converge to 0 as $T$ goes to infinity. To avoid such a curse of dimensionality, the population leading eigenvector $\boldsymbol{u}_1$ is in general assumed to be sparse. More specifically, let $s$ be the number of nonzero elements in $\boldsymbol{u}_1$, we assume $s \ll T$. With this sparsity assumption, different versions of sparse PCA have been proposed to handle i.i.d. Gaussian data (i.e., $\mathbf{A} = \mathbf{0}$): For example, greedy algorithms (d'Aspremont et al., 2008), lasso-type methods including SCoTLASS (Jolliffe et al., 2003), SPCA (Zou, 2006) and sPCA-rSVD (Shen and Huang, 2008), a number of power methods (Journée et al., 2010; Yuan, 2010; Ma, 2011), the biconvex algorithm PMD (Witten et al., 2009) and the semidefinite relaxation DSPCA (d'Aspremont et al., 2004). Sparse PCA has been widely used in finance (d'Aspremont et al., 2005), text mining (Zhang and El Ghaoui, 2011) and voting data analysis (Zhang et al., 2012).

One drawback for these existing sparse PCA theories is that they all assume the $T$ observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ are independently and identically distributed. Such an assumption is obviously violated in most real-world applications. For example, in the fMRI imaging application we described before, the scans from two adjacent time points are obviously correlated. In the applications for stock price or twitter data, the existence of non-dependence is easily justified. Though some work exists for low-dimensional PCA on dependent data (Skinner et al., 1986), no such result exists for high dimensional settings. There are some related results on dependent data analysis in high dimensions (Loh and Wainwright, 2011; Fan et al., 2012), they are mainly for other learning methods. For example, Loh and Wainwright (2011) study the high dimensional regression for Gaussian data with missing values and dependent data. Very recently, Fan et al. (2012) analyze the penalized least square estimators, taking a weakly dependence structure, called $\alpha$-mixing, of the noisy term into consideration.

In this paper, we study sparse PCA for weakly stationary VAR time series. By treating the transition matrix $\mathbf{A}$ as a nuisance parameter, we directly apply sparse PCA on the multivariate time series $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ to estimate the leading eigenvector $\boldsymbol{u}_1$ as if the data are i.i.d. generated. Let $\angle(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1)$ be the angle between $\widehat{\boldsymbol{u}}_1$ and $\boldsymbol{u}_1$, and $\lambda_k(\boldsymbol{\Sigma})$ be the $k$-th largest value of $\boldsymbol{\Sigma}$, we provide an explicit rate of convergence for $\sin\angle(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1)$, i.e., for some absolute constant $C$,

$$\left|\sin\angle\left(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1\right)\right| \leq \frac{C}{\lambda_1(\boldsymbol{\Sigma}) - \lambda_2(\boldsymbol{\Sigma})} \sqrt{\frac{s \log d}{T}\left(\frac{\|\boldsymbol{\Sigma}\|_2}{1 - \|\mathbf{A}\|_2}\right)},$$

with high probability.

Our result allows the quantities $\lambda_1(\boldsymbol{\Sigma})$, $\lambda_2(\boldsymbol{\Sigma})$ and $\|\mathbf{A}\|_2$ all scale with $d$. Since $\boldsymbol{\Sigma}$ is jointly determined by the structure of the transition matrix $\mathbf{A}$ and noise covariance matrix $\boldsymbol{\Psi}$, this result suggests that the interaction between the structure of $\mathbf{A}$ and $\boldsymbol{\Psi}$ plays a pivotal role in determining the final rate of convergence. We also discuss specific structures of $\mathbf{A}$ and $\boldsymbol{\Psi}$ to gain more insights. The results of this paper provide theoretical justifications for the popular practices in which sparse PCA is directly applied on high dimensional time series data for data visualization and feature selection. Examples areas include financial time series, biomedical imaging, and social media, etc.

The rest of the paper is organized as follows. In the next section, we briefly introduce sparse PCA and VAR time series model. In Section 3, we derive several useful results about the rate of convergence of sparse PCA. We prove the main theoretical result in Section 4. In section 5, we conduct numerical experiments on both simulated and real-world data to back up our theory.

## 2 Background

In this section, we briefly introduce the background of this paper. We start with notation. Let $\mathbf{A} = [\mathbf{A}_{jk}] \in \mathbb{R}^{d\times d}$ and $\boldsymbol{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$. For $0 \leq q \leq 1$, we define the vector $\ell_q$ norms as $\|\boldsymbol{v}\|_q := \left(\sum_{j=1}^d |v_j|^q\right)^{\frac{1}{q}}$. Specifically $\|\boldsymbol{v}\|_0 = \operatorname{card}\{\operatorname{supp}(\boldsymbol{v})\}$. We denote $\|\mathbf{A}\|_q$ to be the operator norm of matrix $\mathbf{A}$. In particular, $\|\mathbf{A}\|_2$ is the spectral norm. Specifically, for $q = 1$ or $q = \infty$, $\|\mathbf{A}\|_1 = \max_{1\leq j\leq d} \sum_{i=1}^d |\mathbf{A}_{ij}|$, and $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq d} \sum_{j=1}^d |\mathbf{A}_{ij}|$. We have $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty}$. Let $\lambda_j(\mathbf{A})$ be the $j$-th largest eigenvalue of $\mathbf{A}$. The $d$-dimension Euclidean unit sphere is $\mathbb{S}^{d-1} := \left\{\boldsymbol{v} \middle| \boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\|_2 = 1\right\}$. The $\ell_q$ ball with radius $R_q$ is $\mathbb{B}_q(R_q) := \left\{\boldsymbol{v} \in \mathbb{R}^d : \left\|\boldsymbol{v}\right\|_q \leq R_q\right\}$. For vector $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, define inner product as $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle := \boldsymbol{v}_1^T \boldsymbol{v}_2$. For matrix $\mathbf{A}_1, \mathbf{A}_2$, we define the inner product as $\langle \mathbf{A}_1, \mathbf{A}_2 \rangle := \operatorname{tr}\left(\mathbf{A}_1^T \mathbf{A}_2\right)$. For a set $\mathbb{K}$, $|\mathbb{K}|$ is its cardinality. We use $\mathbf{0}$ to denote the all-zero matrix.

### 2.1 Vector Autoregressive Time Series

The weakly stationary Vector Autoregressive (VAR) time series model linear dependencies between different movements. In particular, the model assumes the $T$ observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T$ are generated by the lag 1 autoregressive process:

$$\boldsymbol{X}_{t+1} = \mathbf{A}\boldsymbol{X}_t + \boldsymbol{Z}_t, \quad \text{for } t = 1, \ldots, T-1. \tag{2.1}$$

To secure the weakly stationary of the above process, the transition matrix $\mathbf{A}$ must have bounded spectral norm $\|\mathbf{A}\|_2 < 1$. We assume the Gaussian colored noise $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots \overset{\text{i.i.d}}{\sim} N_d(\mathbf{0}, \mathbf{\Psi})$. By assumption $\boldsymbol{Z}_t$ and $\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \ldots$ are independent. The stationary property indicates

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T + \mathbf{\Psi}. \tag{2.2}$$

For $j \geq k$, the covariance between $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ is

$$\mathrm{Cov}\big(\boldsymbol{X}_j, \boldsymbol{X}_k\big) = \underbrace{\mathbf{A}\cdots\mathbf{A}}_{j-k}\mathbf{\Sigma} := \mathbf{A}^{j-k}\mathbf{\Sigma}.$$

In the sequel we call $\mathbf{A}$ the transition matrix and $\mathbf{\Psi}$ the noise matrix. VAR model is widely used in the analysis of economic time series (Sims, 1980; Hatemi-J, 2004; Briiggemann and Liitkepohl, 2001), signal processing (de Waele and Broersen, 2003) and brain fMRI (Goebel et al., 2003; Roebroeck et al., 2005).

## 3 Sparse PCA for VAR Time Series

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ be the $T$ observations of a random vector $\boldsymbol{X} \in \mathbb{R}^d$. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ be the top $m$ eigenvectors of the covariance matrix $\mathbf{\Sigma}$. Let $\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_m$ be the top $m$ eigenvector of the sample covariance matrix $\mathbf{S}$. In low dimensions, PCA uses $\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_m$ to estimate $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$.

In the high dimensional settings where $d > T$, we assume that the leading eigenvector of $\mathbf{\Sigma}$ are under certain sparse constraints. In other words, we assume that $\boldsymbol{u}_1$ satisfies that $||\boldsymbol{u}||_0 \leq s$ and $||\boldsymbol{u}||_1 = 1$, i.e. $\boldsymbol{u}_1 \in \mathbb{S}^{d-1} \cap \mathbb{B}_0\big(s\big)$. In this way, we define the model $\mathcal{M}(s, \mathbf{\Sigma}, \lambda_1, \lambda_2)$ as follows:

$$\begin{aligned}\mathcal{M}(s, \mathbf{\Sigma}, \lambda_1, \lambda_2) := \{\boldsymbol{X} : \boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \mathbf{\Sigma}),\\ \lambda_1(\mathbf{\Sigma}) = \lambda_1, \lambda_2(\mathbf{\Sigma}) = \lambda_2, ||\boldsymbol{u}_1||_0 = s\},\end{aligned}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^d$. We define $\widehat{\boldsymbol{u}}_1$ to be the solution to the following optimization problem:

$$\begin{aligned}\widehat{\boldsymbol{u}}_1 := \arg\max_{\boldsymbol{v} \in \mathbb{R}^d} \boldsymbol{v}^T\mathbf{S}\boldsymbol{v},\\ \text{subject to} \quad \boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0\big(s\big).\end{aligned} \tag{3.1}$$

Here $\widehat{\boldsymbol{u}}_1$ is the global optimal estimator of $\boldsymbol{u}_1$. In this paper we only discuss about $\widehat{\boldsymbol{u}}_1$ corresponding to the leading eigenvalue $\lambda_1(\mathbf{\Sigma})$. Analysis for $\widehat{\boldsymbol{u}}_k$ corresponding to $\lambda_k(\mathbf{\Sigma})$ is discussed in the longer version of this paper.

## 4 Theoretical Properties

In this section we provide the theoretical properties of the sparse PCA estimator for VAR time series. In particular, we provide the nonasymptotic upper bound of the rate of convergence in parameter estimation under the VAR model. To our knowledge, this is the first work analyzing the theoretical performance of PCA for the dependent data in high dimensions.

### 4.1 Main Result

The main result states that under the VAR model, the estimator $\widehat{\boldsymbol{u}}_1$ obtained by (3.1) can approximate $\boldsymbol{u}_1$ in a parametric rate with respect to $(n, d, s)$, and the upper bound is also related to the transition matrix $\mathbf{A}$ and the noisy matrix $\mathbf{\Psi}$.

**Theorem 4.1.** *Provided that the random vector sequence $\{\boldsymbol{X}_t\}_{t=1}^T$ follows the VAR model described in (2.1) and $\boldsymbol{X}_t \in \mathcal{M}(s, \mathbf{\Sigma}, \lambda_1, \lambda_2)$ for $t = 1, \ldots, T$, the estimator $\widehat{\boldsymbol{u}}_1$ derived in (3.1) has the following property:*

$$\big|\sin\angle\big(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1\big)\big| = O_P\left(\frac{1}{\lambda_1 - \lambda_2}\sqrt{\frac{s\log d}{T}\left(\frac{\|\mathbf{\Sigma}\|_2}{1 - \|\mathbf{A}\|_2}\right)}\right).$$

*Here for any two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2 \in \mathbb{S}^{d-1}$, $|\sin(\boldsymbol{v}_1, \boldsymbol{v}_2)| := \sqrt{1 - (\boldsymbol{v}_1^T\boldsymbol{v}_2)^2}$.*

**Remark 4.2.** *The bound obtained in (4.1) depends on $\mathbf{\Sigma}, \mathbf{A}, \mathbf{\Psi}$, where $\mathbf{A}$ characterizes the data dependence degree. When both $||\mathbf{A}||_2$ and $||\mathbf{\Psi}||_2$ do not scale with $(n, d, s)$, this is the parametric optimal rate (Ma, 2011; Vu and Lei, 2012).*

### 4.2 Technical Proofs

To prove Theorem 4.1, first we need to prove several lemmas. The following Lemma connects $\sin\angle(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1)$ with $\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0(2s)} |\boldsymbol{v}^T(\mathbf{\Sigma} - \mathbf{S})\boldsymbol{v}|$. In the sequel, we assume that the assumptions in Theorem 4.1 hold.

**Lemma 4.3.** *$\boldsymbol{u}_1$ and $\widehat{\boldsymbol{u}}_1$ satisfy*

$$\sin\angle\big(\boldsymbol{u}_1, \widehat{\boldsymbol{u}}_1\big) \leq \frac{2}{\lambda_1 - \lambda_2}\sup_{\boldsymbol{v} \in \mathbb{S}^{d-1} \cap \mathbb{B}_0(2s)} |\boldsymbol{v}^T(\mathbf{\Sigma} - \mathbf{S})\boldsymbol{v}|. \tag{4.1}$$

*Proof.* Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $\mathbf{\Sigma}$. Let $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ be the corresponding eigenvectors. We have $\boldsymbol{u}_i^T\boldsymbol{u}_j = 0$ for $i \neq j$ and $\mathbf{\Sigma} = \sum_{j=1}^d \lambda_j\boldsymbol{u}_j\boldsymbol{u}_j^T$. Let $\mathbf{\Sigma} = \lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T + \mathbf{\Phi}_0$, where $\mathbf{\Phi}_0$ can represented by $\boldsymbol{u}_1$ and $\mathbf{\Sigma}$ as

$$\begin{aligned}\mathbf{\Phi}_0 &= \mathbf{\Sigma} - \lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T\\ &= \mathbf{\Sigma} - \lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T - \lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T + \lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T\\ &= \mathbf{\Sigma} - \boldsymbol{u}_1\boldsymbol{u}_1^T\mathbf{\Sigma} - \mathbf{\Sigma}\boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_1\big(\boldsymbol{u}_1^T\mathbf{\Sigma}\boldsymbol{u}_1\big)\boldsymbol{u}_1^T\\ &= \big(\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T\big)\mathbf{\Sigma}\big(\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T\big).\end{aligned}$$

For any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}&\big\langle\mathbf{\Sigma}, \boldsymbol{u}_1\boldsymbol{u}_1^T - \boldsymbol{u}\boldsymbol{u}^T\big\rangle\\ &= \big\langle\mathbf{\Sigma}, \boldsymbol{u}_1\boldsymbol{u}_1^T\big\rangle - \big\langle\lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T + \mathbf{\Phi}_0, \boldsymbol{u}\boldsymbol{u}^T\big\rangle\\ &= \lambda_1 - \lambda_1\big\langle\boldsymbol{u}_1, \boldsymbol{u}\big\rangle^2 - \big\langle\mathbf{\Phi}_0, \boldsymbol{u}\boldsymbol{u}^T\big\rangle\\ &= \lambda_1 - \lambda_1\big\langle\boldsymbol{u}_1, \boldsymbol{u}\big\rangle^2 - \boldsymbol{u}^T\big(\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T\big)\mathbf{\Sigma}\big(\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T\big)\boldsymbol{u}.\end{aligned}$$

Now we consider the unit vector

$$\boldsymbol{a} = (\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T)\boldsymbol{u}/\|(\mathbf{I}_d - \boldsymbol{u}_1\boldsymbol{u}_1^T)\boldsymbol{u}\|_2 \in \mathbb{R}^d.$$

It is easy to verify $\boldsymbol{a}$ is orthogonal to $\boldsymbol{u}_1$. Therefore, $\boldsymbol{a} \in \mathrm{span}\{\boldsymbol{u}_2, \ldots, \boldsymbol{u}_d\}$. Letting $a_j = \boldsymbol{a}^T\boldsymbol{u}_j, j = 2, \ldots, d$, we can get $\sum_{j=2}^d a_j^2 = 1$. Therefore we have

$$\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}=\boldsymbol{a}^T\Big(\lambda_1\boldsymbol{u}_1\boldsymbol{u}_1^T+\sum_{j=2}^{d}\lambda_j\boldsymbol{u}_j\boldsymbol{u}_j^T\Big)\boldsymbol{a}=\sum_{j=2}^{d}\lambda_j a_j^2\le\lambda_2,$$

which indicates

$$\boldsymbol{u}^T\big(\mathbf{I}_d-\boldsymbol{u}_1\boldsymbol{u}_1^T\big)\boldsymbol{\Sigma}\big(\mathbf{I}_d-\boldsymbol{u}_1\boldsymbol{u}_1^T\big)\boldsymbol{u}\le\lambda_2\big\|\big(\mathbf{I}_d-\boldsymbol{u}_1\boldsymbol{u}_1^T\big)\boldsymbol{u}\big\|_2^2.$$

Since $\boldsymbol{u}$ is on the unit sphere $\mathbb{S}^{d-1}$, we have $\|\boldsymbol{u}\|_2=1$. Therefore, $\big\|\big(\mathbf{I}_d-\boldsymbol{u}_1\boldsymbol{u}_1^T\big)\boldsymbol{u}\big\|_2^2=1-\big\langle\boldsymbol{u}_1,\boldsymbol{u}\big\rangle^2$. Now we obtain

$$\big\langle\boldsymbol{\Sigma},\boldsymbol{u}_1\boldsymbol{u}_1^T-\boldsymbol{u}\boldsymbol{u}^T\big\rangle\ge(\lambda_1-\lambda_2)\Big(1-\big\langle\boldsymbol{u}_1,\boldsymbol{u}\big\rangle^2\Big).\quad(4.2)$$

Since (4.2) holds for any $\boldsymbol{u}\in\mathbb{S}^{d-1}$, letting $\boldsymbol{u}=\widehat{\boldsymbol{u}}_1$ we have

$$\big\langle\boldsymbol{\Sigma},\boldsymbol{u}_1\boldsymbol{u}_1^T-\boldsymbol{u}\boldsymbol{u}^T\big\rangle\ge(\lambda_1-\lambda_2)\sin^2\angle\big(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1\big).$$

Since $\widehat{\boldsymbol{u}}_1$ is defined as $\widehat{\boldsymbol{u}}_1:=\arg\max_{\boldsymbol{v}\in\mathbb{R}^d}\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}$, we know

$$\widehat{\boldsymbol{u}}_1^T\mathbf{S}\widehat{\boldsymbol{u}}_1-\boldsymbol{u}_1^T\mathbf{S}\boldsymbol{u}_1\le0.$$

Therefore, $\big\langle\mathbf{S},\boldsymbol{u}_1{\boldsymbol{u}_1}^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T\big\rangle\le0$, we have

$$\begin{aligned}\sin^2\angle\big(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1\big)&\le\frac{1}{\lambda_1-\lambda_2}\big\langle\boldsymbol{\Sigma},\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T\big\rangle\\&\le\frac{1}{\lambda_1-\lambda_2}\big\langle\boldsymbol{\Sigma}-\mathbf{S},\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T\big\rangle.\end{aligned}$$

Let $\boldsymbol{\Pi}$ be the diagonal matrix with diagonal values being 1 if and only if the corresponding entries in $\boldsymbol{u}_1$ or $\widehat{\boldsymbol{u}}_1$ are zero. Therefore, we know there are at most $2s$ nonzero values in $\boldsymbol{\Pi}$. Then we have

$$\begin{aligned}&\sin^2\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\\&\le\frac{1}{\lambda_1-\lambda_2}\big\langle\boldsymbol{\Sigma}-\mathbf{S},\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T\big\rangle\\&=\frac{1}{\lambda_1-\lambda_2}\big\langle\boldsymbol{\Sigma}-\mathbf{S},\boldsymbol{\Pi}(\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T)\boldsymbol{\Pi}\big\rangle\\&=\frac{1}{\lambda_1-\lambda_2}\big\langle\boldsymbol{\Pi}(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{\Pi},\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T\big\rangle\\&\le\frac{1}{\lambda_1-\lambda_2}||\boldsymbol{\Pi}(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{\Pi}||_2||\boldsymbol{u}_1\boldsymbol{u}_1^T-\widehat{\boldsymbol{u}}_1\widehat{\boldsymbol{u}}_1^T||_S\\&=\frac{1}{\lambda_1-\lambda_2}||\boldsymbol{\Pi}(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{\Pi}||_2\cdot2|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)|.\end{aligned}$$

Here $||\cdot||_S$ denotes the sum of singular values. This implies

$$\begin{aligned}\big|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\big|&\le\frac{2}{\lambda_1-\lambda_2}||\boldsymbol{\Pi}(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{\Pi}||_2\\&\le\frac{2}{\lambda_1-\lambda_2}\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_0(2s)}|\boldsymbol{v}^T(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{v}|.\end{aligned}$$

This completes the proof. □

The next lemma comes from Ledoux and Talagrand (2011) and is informative in the proof of the main theorem.

**Lemma 4.4.** *Provided that* $x_1,x_2,\ldots,x_T\overset{\text{i.i.d}}{\sim}N(0,1)$. $\boldsymbol{X}=(x_1,x_2,\ldots,x_T)^T\in\mathbb{R}^T$ *is a random vector. Mapping* $f:\mathbb{R}^T\to\mathbb{R}$ *is Lipschitz, i.e., for any* $\boldsymbol{v}_1,\boldsymbol{v}_2\in\mathbb{R}^T$*:*

$$\exists L\ge0,\quad s.t.\ \big|f(\boldsymbol{v}_1)-f(\boldsymbol{v}_2)\big|\le L\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2,$$

*Then for any* $t>0$ *we have,*

$$\mathbb{P}\Big(\big|f(\boldsymbol{X})-\mathbb{E}f(\boldsymbol{X})\big|>t\Big)\le2\exp\left(-\frac{t^2}{2L^2}\right).\quad(4.3)$$

*Proof.* We refer to Ledoux and Talagrand (2011)'s proof of this lemma. □

The next lemma quantifies the difference between $\boldsymbol{\Sigma}$ and $\mathbf{S}$ for any fixed vector $\boldsymbol{v}\in\mathbb{R}^d$.

**Lemma 4.5.** *Letting* $\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_q\big(R_q\big)$*, we have*

$$\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}-\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|=O_P\left(\sqrt{\frac{1}{T}}\right).\quad(4.4)$$

*Proof.* Let $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T$ be the $T$ observations. $\boldsymbol{Y}=\Big(\boldsymbol{x}_1^T\boldsymbol{v},\boldsymbol{x}_2^T\boldsymbol{v},\ldots,\boldsymbol{x}_T^T\boldsymbol{v}\Big)^T\in\mathbb{R}^T$ is a zero-mean Gaussian random vector. We denote the covariance matrix of $\boldsymbol{Y}$ as $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$. $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ is symmetric and semi-definite. Thus $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ can be decomposed as $\boldsymbol{\Sigma}_{\boldsymbol{Y}}=\mathbf{Q}^T\mathbf{Q}$, where $\mathbf{Q}$ is a matrix with orthogonal columns. Let $\sigma=\|\mathbf{Q}\|_2=\sqrt{\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_2}$. According to the definition of $\boldsymbol{Y}$, we have

$$\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}=\boldsymbol{v}^T\left(\frac{\sum_{i=1}^T\boldsymbol{x}_i\boldsymbol{x}_i^T}{T}\right)\boldsymbol{v}=\frac{\boldsymbol{Y}^T\boldsymbol{Y}}{T}=\left(\frac{\|\boldsymbol{Y}\|_2}{\sqrt{T}}\right)^2,$$

$$\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}=\boldsymbol{v}^T\mathbb{E}\big(\boldsymbol{X}\boldsymbol{X}^T\big)\boldsymbol{v}=\mathbb{E}\left(\frac{\boldsymbol{Y}^T\boldsymbol{Y}}{T}\right)=\mathbb{E}\left(\frac{\|\boldsymbol{Y}\|_2}{\sqrt{T}}\right).$$

For convenience we define $\boldsymbol{W}:=\|\boldsymbol{Y}\|_2/\sqrt{T}$ and $f(\boldsymbol{v}):=\|\mathbf{Q}\boldsymbol{v}\|_2/\sqrt{T}$ for $\boldsymbol{v}\in\mathbb{R}^T$. Since we have $\boldsymbol{Y}=\mathbf{Q}\boldsymbol{V}$, $\boldsymbol{V}\in\mathbb{R}^T$ is a zero-mean Gaussian vector with covariance matrix $\mathbf{I}_T$. It can be verified that mapping $f:\mathbb{R}^T\to\mathbb{R}$ is Lipschitz with $L=\sigma/\sqrt{T}$. Using Lemma 4.4, we can get

$$\mathbb{P}\left(\big|\boldsymbol{W}-\mathbb{E}\big(\boldsymbol{W}\big)\big|\ge t\right)\le2\exp\left(-\frac{t^2T}{2\sigma^2}\right).\quad(4.5)$$

Since $\mathrm{Var}\big(\boldsymbol{W}\big)=\mathbb{E}\big(\boldsymbol{W}^2\big)-\mathbb{E}^2\big(\boldsymbol{W}\big)\ge0$, at the same time we have $\mathbb{E}\big(\boldsymbol{W}^2\big)\ge0,\mathbb{E}\big(\boldsymbol{W}\big)\ge0$, which implies $\sqrt{\mathbb{E}\big(\boldsymbol{W}^2\big)}-\mathbb{E}\big(\boldsymbol{W}\big)\ge0$. Thus we get

$$\left(\sqrt{\mathbb{E}\big(\boldsymbol{W}^2\big)}-\mathbb{E}\big(\boldsymbol{W}\big)\right)^2\le\mathbb{E}\left(\big|\boldsymbol{W}-\mathbb{E}\big(\boldsymbol{W}\big)\big|^2\right)=\frac{4\sigma^2}{T},$$

which indicates

$$\Big|\sqrt{\mathbb{E}\big(\boldsymbol{W}^2\big)}-\mathbb{E}\big(\boldsymbol{W}\big)\Big|\le\frac{2\sigma}{\sqrt{T}}.\quad(4.6)$$

Here (4.6) together with $\big|\boldsymbol{W}-\mathbb{E}\big(\boldsymbol{W}\big)\big|\le t$ implies $\Big|\boldsymbol{W}-\sqrt{\mathbb{E}\big(\boldsymbol{W}^2\big)}\Big|\le t+2\sigma/\sqrt{T}$. Therefore, according

to the definition of $\boldsymbol{Y}$ and $\boldsymbol{W}$,

$$\begin{aligned}
&\mathbb{P}\bigg(\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}-\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|\geq t+\frac{2\sigma}{\sqrt{T}}\bigg) \qquad (4.7)\\
=\;&\mathbb{P}\bigg(\Big|\boldsymbol{W}-\sqrt{\mathbb{E}(\boldsymbol{W}^2)}\Big|\geq t+\frac{2\sigma}{\sqrt{T}}\bigg)\\
\leq\;&\mathbb{P}\big(\big|\boldsymbol{W}-\mathbb{E}(\boldsymbol{W})\big|\geq t\big)\\
\leq\;&2\exp\bigg(-\frac{t^2T}{2\sigma^2}\bigg).
\end{aligned}$$

We reach the conclusion

$$\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}-\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|=O_p\bigg(\sqrt{\frac{1}{T}}\bigg).$$

This completes the proof. □

**Lemma 4.6.** *An $\epsilon$-net $\mathbb{N}_\epsilon$ of a sphere $\mathbb{S}^{n-1}$ (equipped with Euclidean distance) is a subset of $\mathbb{S}^{n-1}$ such that for any $\boldsymbol{v}\in\mathbb{S}^{n-1}$, there exists $\boldsymbol{u}\in\mathbb{N}_\epsilon$ subject to $\|\boldsymbol{u}-\boldsymbol{v}\|_2\leq\epsilon$. It is shown by Vershynin (2010) that for any $\epsilon>0$,*

$$\big|\mathbb{N}_\epsilon\big|\leq\bigg(1+\frac{2}{\epsilon}\bigg)^n. \qquad (4.8)$$

*Also, for matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$, the inequality below holds for any $\epsilon\in[0,1)$*

$$\max_{\boldsymbol{v}_1\in\mathbb{S}^{n-1}}|\boldsymbol{v}_1\mathbf{A}\boldsymbol{v}_1|\leq(1-2\epsilon)^{-1}\max_{\boldsymbol{v}_2\in\mathbb{N}_\epsilon}|\boldsymbol{v}_2\mathbf{A}\boldsymbol{v}_2|. \qquad (4.9)$$

*Proof.* Construct a maximal $\epsilon$-separated subset $\mathbb{N}_\epsilon$ of $\mathbb{S}^{n-1}$. Here an $\epsilon$-separated set is defined as the set whose arbitrary two different elements $\boldsymbol{x},\boldsymbol{y}$ are at least distance $\epsilon$ away, i.e. $\|\boldsymbol{x}-\boldsymbol{y}\|_2\geq\epsilon$. Since $\mathbb{N}_\epsilon$ is a maximal $\epsilon$-separated subset $\mathbb{N}_\epsilon$ of $\mathbb{S}^{n-1}$, there isn't any other $\epsilon$-separated $\mathbb{N}'_\epsilon$ of $\mathbb{S}^{n-1}$ such that $\mathbb{N}_\epsilon\subset\mathbb{N}'_\epsilon$.

We can prove that $\mathbb{N}_\epsilon$ is an $\epsilon$-net of $\mathbb{S}^{n-1}$ by contradiction. If we assume for $\boldsymbol{u}\in\mathbb{S}^{n-1}$, there is a point $\boldsymbol{v}\in\mathbb{N}_\epsilon$, $\|\boldsymbol{u}-\boldsymbol{v}\|_2>\epsilon$, then $\mathbb{N}'_\epsilon=\{\boldsymbol{v}\}\cup\mathbb{N}_\epsilon$ is a larger $\epsilon$-separated subset of $\mathbb{S}^{n-1}$ that contains $\mathbb{N}_\epsilon$, which contradicts with the fact that $\mathbb{N}_\epsilon$ is the maximal $\epsilon$-separated subset of $\mathbb{S}^{n-1}$.

Now we derive the bound of $|\mathbb{N}_\epsilon|$. We cover the neighborhood of every $\boldsymbol{x}_i\in\mathbb{N}_\epsilon$ with disjoint balls $\mathbb{B}_{\boldsymbol{y}i}=\big\{\boldsymbol{y}_i\big|\|\boldsymbol{y}_i-\boldsymbol{x}_i\|_2<\epsilon/2\big\}$. For any $\mathbb{B}_{\boldsymbol{y}i}\subset\mathbb{B}_0=\big\{\boldsymbol{y}\big|\|\boldsymbol{y}\|_2<1+\epsilon/2\big\}$ we have

$$|\mathbb{N}_\epsilon|\cdot|\mathbb{B}_{\boldsymbol{y}i}|=\sum_{i=1}^{|\mathbb{N}_\epsilon|}|\mathbb{B}_{\boldsymbol{y}i}|=\bigg|\bigcup_{i=1}^{|\mathbb{N}_\epsilon|}\mathbb{B}_{\boldsymbol{y}i}\bigg|\leq|\mathbb{B}_0|,$$

$$|\mathbb{N}_\epsilon|\leq\frac{|\mathbb{B}_0|}{|\mathbb{B}_{\boldsymbol{y}_i}|}=\frac{\big(1+\frac{\epsilon}{2}\big)^n}{\big(\frac{\epsilon}{2}\big)^n}=\bigg(1+\frac{2}{\epsilon}\bigg)^n.$$

Then we turn to prove (4.9). For $\boldsymbol{v}_1\in\mathbb{S}^{n-1}$ and $\boldsymbol{v}_2\in\mathbb{N}_\epsilon$, we know since $\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2\leq\epsilon$,

$$\begin{aligned}
|\boldsymbol{v}_1\mathbf{A}\boldsymbol{v}_1-\boldsymbol{v}_2\mathbf{A}\boldsymbol{v}_2|&\leq\|\mathbf{A}\|_2\|\boldsymbol{v}_1\|_2\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2\\
&\quad+\|\mathbf{A}\|_2\|\boldsymbol{v}_2\|_2\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2\leq2\epsilon\|\mathbf{A}\|_2
\end{aligned}$$

It follows that

$$|\boldsymbol{v}_2\mathbf{A}\boldsymbol{v}_2|\geq(1-2\epsilon)\|\mathbf{A}\|_2=(1-2\epsilon)\max_{\boldsymbol{v}_1\in\mathbb{S}^{d-1}}|\boldsymbol{v}_1\mathbf{A}\boldsymbol{v}_1|$$

Taking the maximum over $\boldsymbol{v}_2\in\mathbb{N}_\epsilon$, we completes the proof. □

Combining the above Lemmas, we can now proceed to the main proof.

*Proof of Theorem 4.1.* Using Lemma 4.5, it is easy to show

$$\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}=\sqrt{\mathbb{E}(\boldsymbol{W}^2)}=\sqrt{\frac{tr\big(\mathbf{Q}\mathbf{Q}^T\big)}{T}}=\frac{\|\mathbf{Q}\|_F}{\sqrt{T}}\leq\sigma.$$

Then using (4.7) we can get

$$\begin{aligned}
&\mathbb{P}\bigg(\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}+\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|\geq t+4\sigma\bigg)\\
\leq\;&\mathbb{P}\bigg(\Big|\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}+\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}\Big|\geq t+\frac{2\sigma}{\sqrt{T}}+2\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}\bigg)\\
\leq\;&\mathbb{P}\bigg(\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}-\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|\geq t+\frac{2\sigma}{\sqrt{T}}\bigg)\\
\leq\;&2\exp\bigg(-\frac{t^2T}{2\sigma^2}\bigg). \qquad (4.10)
\end{aligned}$$

We define events $\mathcal{E}_1$ and $\mathcal{E}_2$, combining (4.7) (with $t=t_1$) and (4.10) (with $t=t_2$),

$$\begin{aligned}
\mathcal{E}_1&:=\bigg\{\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}-\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|\geq t_1+2\frac{\sigma}{\sqrt{T}}\bigg\}\\
\mathcal{E}_2&:=\bigg\{\Big|\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}}+\sqrt{\boldsymbol{v}^T\mathbf{S}\boldsymbol{v}}\Big|\geq t_2+4\sigma\bigg\}.
\end{aligned}$$

We derive for any $t_1,t_2>0$,

$$\begin{aligned}
&\mathbb{P}\bigg(\Big|\boldsymbol{v}^T\big(\boldsymbol{\Sigma}-\mathbf{S}\big)\boldsymbol{v}\Big|\geq\bigg(t_1+2\frac{\sigma}{\sqrt{T}}\bigg)(t_2+4\sigma)\bigg)\\
&\leq\mathbb{P}(\mathcal{E}_1)+\mathbb{P}(\mathcal{E}_2)\\
&\leq2\exp\bigg(-\frac{t_1^2T}{2\sigma^2}\bigg)+2\exp\bigg(-\frac{t_2^2T}{2\sigma^2}\bigg). \qquad (4.11)
\end{aligned}$$

Now we turn to upper bound $\sup_{\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_0(2s)}|\boldsymbol{v}^T(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{v}|$ in Lemma 4.3. Assuming we have a fixed subset $\mathbb{K}\subset\{1,\ldots,d\}$, we define

$$\mathbb{B}_{\mathbb{K}}=\big\{\boldsymbol{v}\big|\text{for any } i\in\{1,\ldots,d\}\setminus\mathbb{K},v_i=0\big\}.$$

For any $t_1,t_2>0$, we define event $\mathcal{E}_{\mathbb{K}}$ and $\mathcal{E}_{\boldsymbol{v}}$ as

$$\begin{aligned}
\mathcal{E}_{\mathbb{K}}&:=\bigg\{\max_{\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_{\mathbb{K}}}\Big|\boldsymbol{v}^T\big(\boldsymbol{\Sigma}-\mathbf{S}\big)\boldsymbol{v}\Big|\geq2\bigg(t_1+2\frac{\sigma}{\sqrt{T}}\bigg)(t_2+4\sigma)\bigg\},\\
\mathcal{E}_{\boldsymbol{v}}&:=\bigg\{\Big|\boldsymbol{v}^T\big(\boldsymbol{\Sigma}-\mathbf{S}\big)\boldsymbol{v}\Big|\geq\bigg(t_1+2\frac{\sigma}{\sqrt{T}}\bigg)(t_2+4\sigma)\bigg\}.
\end{aligned}$$

According to Lemma 4.6, we define the $\frac{1}{4}$-net of $\mathbb{S}^{d-1}\cap\mathbb{B}_{\mathbb{K}}$ as $\mathbb{N}_{\mathbb{K}}$, then we can get

$$\mathcal{E}_{\mathbb{K}}\subset\bigcup_{\boldsymbol{v}\in\mathbb{N}_{\mathbb{K}}}\bigg\{\Big|\boldsymbol{v}^T\big(\boldsymbol{\Sigma}-\mathbf{S}\big)\boldsymbol{v}\Big|\geq\bigg(t_1+2\frac{\sigma}{\sqrt{T}}\bigg)(t_2+4\sigma)\bigg\},$$

Combining (4.11) and (4.8) and letting $|\mathbb{K}| = 2s$, we obtain

$$\mathbb{P}(\mathcal{E}_{\mathbb{K}}) \le \sum_{\boldsymbol{v}\in\mathbb{N}_{\mathbb{K}}} \mathbb{P}(\mathcal{E}_{\boldsymbol{v}}) = |\mathbb{N}_{\mathbb{K}}|\mathbb{P}(\mathcal{E}_{\boldsymbol{v}}) \le 9^{2s}\left(2\exp\left(-\frac{t_1^2T}{2\sigma^2}\right)+2\exp\left(-\frac{t_2^2T}{2\sigma^2}\right)\right).$$

Now we consider arbitrary subset $\mathbb{K}\subset\{1,\dots,d\}$ with cardinality $2s$. We define

$$\mathcal{E}'_{\mathbb{K}}:=\left\{\max_{\boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_0(2s)}\left|\boldsymbol{v}^T(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{v}\right|\ge 2\left(t_1+2\frac{\sigma}{\sqrt{T}}\right)(t_2+4\sigma)\right\}.$$

Then we have

$$\mathbb{P}(\mathcal{E}'_{\mathbb{K}}) \le \sum_{\mathbb{K}\subset\{1,\dots,d\}} \mathbb{P}(\mathcal{E}_{\mathbb{K}}) \le \binom{d}{2s}\mathbb{P}(\mathcal{E}_{\mathbb{K}}) \le 9^{2s}\binom{d}{2s}\left(2\exp\left(-\frac{t_1^2T}{2\sigma^2}\right)+2\exp\left(-\frac{t_2^2T}{2\sigma^2}\right)\right).$$

Using (4.1), we can get for any $t_1,t_2>0, \boldsymbol{v}\in\mathbb{S}^{d-1}\cap\mathbb{B}_0(2s)$, which implies

$$\mathbb{P}\left(\left|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\right|\ge\frac{4}{\lambda_1-\lambda_2}\left(t_1+2\frac{\sigma}{\sqrt{T}}\right)(t_2+4\sigma)\right) \le \mathbb{P}\left(\max\left|\boldsymbol{v}^T(\boldsymbol{\Sigma}-\mathbf{S})\boldsymbol{v}\right|\ge 2\left(t_1+2\frac{\sigma}{\sqrt{T}}\right)(t_2+4\sigma)\right) \le 9^{2s}\binom{d}{2s}\left(2\exp\left(-\frac{t_1^2T}{2\sigma^2}\right)+2\exp\left(-\frac{t_2^2T}{2\sigma^2}\right)\right).$$

Now we are going to derive the upper bound for $\sigma^2 = \left\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right\|_2$. We note that VAR model assumes $\|\mathbf{A}\|_2<1$. Thus $\sigma^2$ can be bounded as follows,

$$\sigma^2=\left\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right\|_2\le\left\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\right\|_\infty=\max_{i\in\{1,\dots,T\}}\sum_{j=1}^{T}\left|(\boldsymbol{\Sigma}_{\boldsymbol{Y}})_{ij}\right|. \quad (4.12)$$

We have $\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_2\le\sqrt{\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_1\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_\infty}=\|\boldsymbol{\Sigma}_{\boldsymbol{Y}}\|_\infty$, since $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ is symmetric. According to the definition (1.1) we have,

$$\boldsymbol{v}^T\operatorname{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j)\boldsymbol{v}=\boldsymbol{v}^T\operatorname{Cov}(\boldsymbol{x}_j,\boldsymbol{x}_i)\boldsymbol{v}=\boldsymbol{v}^T\mathbf{A}^{|i-j|}\boldsymbol{\Sigma}\boldsymbol{v},$$

In order to get $\max_{i\in\{1,\dots,T\}}\sum_{j=1}^{T}\left|(\boldsymbol{\Sigma}_{\boldsymbol{Y}})_{ij}\right|$ in (4.12), we first derive the upper bound of $\sum_{j=1}^{T}\left|(\boldsymbol{\Sigma}_{\boldsymbol{Y}})_{ij}\right|$.

$$\sum_{j=1}^{T}|(\boldsymbol{\Sigma}_{\boldsymbol{Y}})_{ij}|=\sum_{j\neq i}\left|\boldsymbol{v}^T\operatorname{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j)\boldsymbol{v}\right|+\left|\boldsymbol{v}^T\operatorname{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_i)\boldsymbol{v}\right| = \sum_{j\neq i}\left|\boldsymbol{v}^T\mathbf{A}^{|i-j|}\boldsymbol{\Sigma}\boldsymbol{v}\right|+\left|\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}\right| \le \sum_{j\neq i}\left\|\boldsymbol{\Sigma}\right\|_2\|\mathbf{A}\|_2^{|i-j|}+\left\|\boldsymbol{\Sigma}\right\|_2 \le \frac{2\left\|\boldsymbol{\Sigma}\right\|_2}{1-\|\mathbf{A}\|_2}.$$

Thus, letting $\delta^2=2\|\boldsymbol{\Sigma}\|_2/(1-\|\mathbf{A}\|_2)\ge\sigma^2$, we have for any $t_1,t_2>0$,

$$\mathbb{P}\left(\left|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1\right|\ge\frac{4}{\lambda_1-\lambda_2}\left(t_1+2\frac{\delta}{\sqrt{T}}\right)(t_2+4\delta)\right) \le 9^{2s}\binom{d}{2s}\left(2\exp\left(-\frac{t_1^2T}{2\delta^2}\right)+2\exp\left(-\frac{t_2^2T}{2\delta^2}\right)\right).$$

Letting $t_1=\sqrt{s\delta^2\log d/T}$ and $t_2$ be a constant, we reach the conclusion,

$$\left|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\right|=O_P\left(\frac{1}{\lambda_1-\lambda_2}\sqrt{\frac{s\log d}{T}\left(\frac{\|\boldsymbol{\Sigma}\|_2}{1-\|\mathbf{A}\|_2}\right)}\right).$$

This is the main result of the rate of convergence. □

## 5 Experiments

In this section we show some experimental results on both synthetic and real-world data to back up the theoretical results we obtain in last section. We use the truncated power method proposed by Yuan and Zhang (2011) to approximate the global estimator $\widehat{\boldsymbol{u}}_1$ obtained in (4.2).

### 5.1 Synthetic Data

In this section we experiment with sparse PCA on synthetic data. We show how $\mathbf{A}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ affect $\left|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\right|$. First we create $\mathbf{A}$ given $\|\mathbf{A}\|_2$. With $\lambda_1(\mathbf{A})$ and $\lambda_2(\mathbf{A})$, we generate

$$\boldsymbol{\Sigma}=(\lambda_1(\mathbf{A})-1)\boldsymbol{u}_1^T\boldsymbol{u}_1+(\lambda_2(\mathbf{A})-1)\boldsymbol{u}_2^T\boldsymbol{u}_2+\mathbf{I},$$

where $\|\boldsymbol{u}_1\|_0=s$, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ is orthogonal. According to the stationary property (2.2), the covariance matrix of the noise random vector $\boldsymbol{Z}_i$ is $\boldsymbol{\Psi}=\boldsymbol{\Sigma}-\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}$, where $\boldsymbol{\Psi}$ must be a positive semidefinite matrix. We generate $T=50$ data points according to the VAR time series model in (1.1). We illustrate how the scaling of $\|\mathbf{A}\|_2$ affects the accuracy of estimator $\widehat{\boldsymbol{u}}_1$. Set $\lambda_1(\boldsymbol{\Sigma})=10$, $\lambda_2(\boldsymbol{\Sigma})=5$, $s=20$, $d=200$ and $\|\mathbf{A}\|_2\in[0,0.9]$, we repeatedly experiment for 3000 times for each $\|\mathbf{A}\|_2$. Here $\boldsymbol{\Sigma}$ is fixed while $\mathbf{A}$ and $\boldsymbol{\Psi}$ are varying. The results are illustrated in Figure 1 by plotting the relevant part in the theoretical upper bound $\frac{||\boldsymbol{\Sigma}||_2}{1-||\mathbf{A}||_2}$ against the empirical error $|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)|$. As can be observed in Fig. 1, the empirical error increases when the spectral norm of transition matrix $\|\mathbf{A}\|_2$ increases. This makes sense because when $\|\mathbf{A}\|_2=0$, VAR model is reduced to the i.i.d. case, since $\boldsymbol{X}_{t+1}$ doesn't depend on $\boldsymbol{X}_t$. When $\|\mathbf{A}\|_2\to1$, the degree of dependency increases and sparse PCA loses its estimation accuracy, as quantified in Theorem 4.1. Table 1 shows the values and the standard deviations of $\left|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)\right|$ for each $\|\mathbf{A}\|_2$.

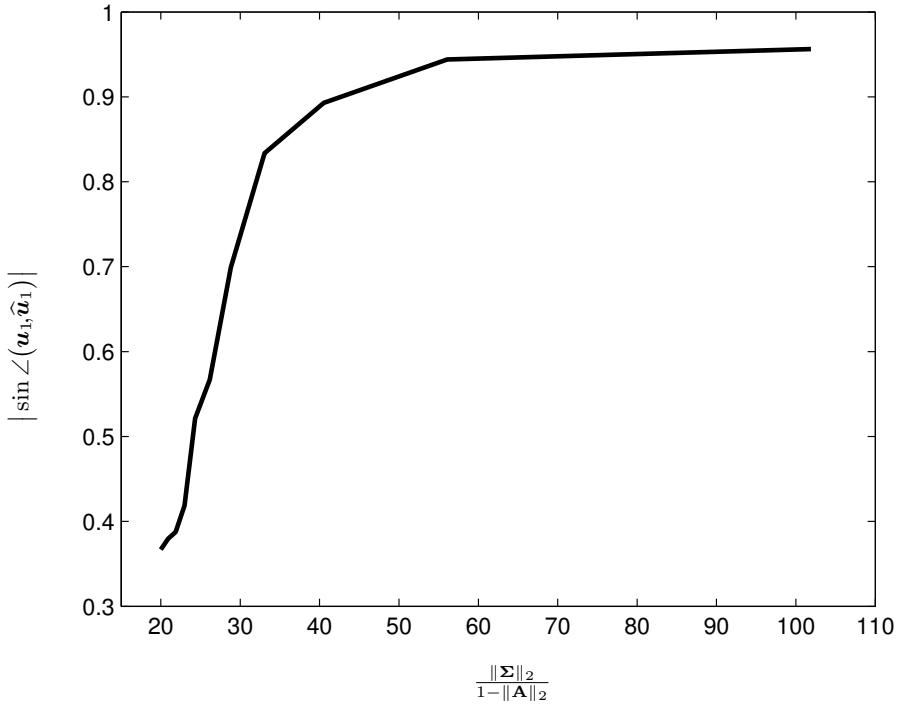


Figure 1: $\|\mathbf{A}\|_2$'s impact on the estimation accuracy $|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)|$. When dependency $\|\mathbf{A}\|_2$ increases toward 1, the upper bound $\frac{\|\boldsymbol{\Sigma}\|_2}{1-\|\mathbf{A}\|_2}$ increases, thus the estimator $\widehat{\boldsymbol{u}}_1$' accuracy decreases, i.e. $|\sin\angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)|$ increases towards 1.

Table 1: Corresponding values when $\|\mathbf{A}\|_2$ increases. Here $\kappa* = \frac{\|\boldsymbol{\Sigma}\|_2}{1-\|\mathbf{A}\|_2}$ and $\theta = \angle(\boldsymbol{u}_1,\widehat{\boldsymbol{u}}_1)$

| $\|\mathbf{A}\|_2$ | $\kappa^*$ | $\lvert\sin\theta\rvert$ | $\sigma(\lvert\sin\theta\rvert)$ |
|---|---|---|---|
| 0 | 20.0000 | 0.3669 | 0.1284 |
| 0.0900 | 20.9080 | 0.3795 | 0.2150 |
| 0.1800 | 21.8711 | 0.3876 | 0.1319 |
| 0.2700 | 22.9696 | 0.4186 | 0.1591 |
| 0.3600 | 24.3290 | 0.5214 | 0.2201 |
| 0.4500 | 26.1568 | 0.5671 | 0.2625 |
| 0.5400 | 28.8231 | 0.6989 | 0.1981 |
| 0.6300 | 33.0580 | 0.8336 | 0.1549 |
| 0.7200 | 40.5303 | 0.8930 | 0.1410 |
| 0.8100 | 56.0706 | 0.9441 | 0.0765 |
| 0.9000 | 101.9000 | 0.9563 | 0.1196 |

### 5.2 Equity Data

In this section we apply sparse PCA on daily closing prices of 452 stocks in the S&P 500 index between January 1, 2003 through January 1, 2008 from Yahoo! Finance (`finance.yahoo.com`). That is to say, we have altogether $T = 1257$ observations corresponding to the vector of closing prices on a trading day. We categorize the 452 stocks into 10 Global Industry Classification Standard (GICS) sectors, including `Consumer Discretionary` (70 stocks), `Consumer Staples` (35 stocks), `Energy` (37 stocks), `Financials` (74 stocks), `Health Care` (46 stocks), `Industrials` (59 stocks), `Information Technology` (64 stocks) `Telecommunications Services` (6 stocks), , `Materials` (29 stocks), and `Utilities` (32 stocks). We expect that the stocks from the same sector are likely to appear in the non-zero entries of the same principal component, since stocks from the same sector tend to be more correlated.

Let $\mathbf{P} = [\mathbf{P}_{t,j}]$ be the closing price of stock $j$ on day $t$. In this paper we are interested in the transformed data, where we calculate the log-ratio of the price at time $t$ to price at $t-1$:

$$\mathbf{X}_{t,j} = \log(\mathbf{P}_{t+1,j}/\mathbf{P}_{t,j}), \quad t = 1,\ldots,T-1.$$

It is obvious that there exists data dependence struc-

ture between $\mathbf{X}_{t_1,j}$ and $\mathbf{X}_{t_2,j}$ for any $t_1, t_2 \leq T-1$ and it accordingly raises concern for conducting classical sparse PCA algorithms on this dataset $\mathbf{X}$. However, the argument in this paper provides justifications to such a procedure. In particular, we conclude that the same parametric rate can be persisted if the operator norm of the transition matrix does not scale with $(n, d, s)$. Here we present labels of the first three estimated leading eigenvectors in Table 2. As can be observed, the sparse PCA algorithm tend to group stocks from the same sector into the same eigenvector.

**Acknowledgements**

This research was supported by NSF award IIS-1116730.

Table 2: Non-zero terms' sectors in the 1st, 2nd and 3rd eigenvectors obtained.

| $\widehat{\boldsymbol{u}}_1$ | $\widehat{\boldsymbol{u}}_2$ | $\widehat{\boldsymbol{u}}_3$ |
|---|---|---|
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Consumer Discretionary |
| Financials | Industrials | Financials |
| Financials | Industrials | Financials |
| Financials | Industrials | Financials |
| Financials | Industrials | Financials |
| Financials | Industrials | Financials |
| Financials | Industrials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Industrials |
| Financials | Materials | Information Technology |
| Financials | Materials | Information Technology |
| Financials | Materials | Materials |
| Financials | Materials | Materials |
| Financials | Materials | Materials |

## References


Briiggemann, R. and Liitkepohl, H. (2001). Lag selection in subset var models with an application to a us monetary system. *Econometric Studies: A Festschrift in Honour of Joachim Frohn, LIT-Verlag, Miinster* 107–28.

d'Aspremont, A., Bach, F. and Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research* **9** 1269–1294.

d'Aspremont, A., El Ghaoui, L., Jordan, M. and Lanckriet, G. (2004). *A direct formulation for sparse PCA using semidefinite programming*. Computer Science Division, University of California.

d'Aspremont, A., El Ghaoui, L., Jordan, M. and Lanckriet, G. (2005). Sparse pca with applications in finance .

de Waele, S. and Broersen, P. (2003). Order selection for vector autoregressive models. *Signal Processing, IEEE Transactions on* **51** 427–433.

Fan, J., Qi, L. and Tong, X. (2012). Penalized least squares estimation with weakly dependent data .

Goebel, R., Roebroeck, A., Kim, D. and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging* **21** 1251–1261.

Hatemi-J, A. (2004). Multivariate tests for autocorrelation in the stable and unstable var models. *Economic Modelling* **21** 661–683.

Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104** 682–693.

Jolliffe, I., Trendafilov, N. and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12** 531–547.

Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research* **11** 517–553.

Ledoux, M. and Talagrand, M. (2011). *Probability in Banach Spaces: isoperimetry and processes*, vol. 23. Springer.

Loh, P. and Wainwright, M. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)* .

Ma, Z. (2011). Sparse principal component analysis and iterative thresholding. *Arxiv preprint arXiv:1112.2432* .

Roebroeck, A., Formisano, E., Goebel, R. et al. (2005). Mapping directed influence over the brain using granger causality and fmri. *Neuroimage* **25** 230–242.

Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis* **99** 1015–1034.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.
URL `http://ideas.repec.org/a/ecm/emetrp/v48y1980i1p1-48.html`

Skinner, C., Holmes, D. and Smith, T. (1986). The effect of sample design on principal component analysis. *Journal of the American Statistical Association* **81** 789–798.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Vu, V. and Lei, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. *Arxiv preprint arXiv:1202.0786* .

Witten, D., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* **99** 2261–2286.

Yuan, X. and Zhang, T. (2011). Truncated power method for sparse eigenvalue problems. *Arxiv preprint arXiv:1112.2679* .

Zhang, Y., dAspremont, A. and Ghaoui, L. (2012). Sparse pca: Convex relaxations, algorithms and applications. *Handbook on Semidefinite, Conic and Polynomial Optimization* 915–940.

Zhang, Y. and El Ghaoui, L. (2011). Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems (NIPS)*.

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