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 $x_1, x_2, \ldots, x_T \in \mathbb{R}^d$ be $T$ observations from a time series $X_1, X_2, \ldots, X_T$. Here we assume that each $X_t \in \mathbb{R}^d$ is a $d$-dimensional random vector and follows a VAR model

$$X_{t+1} = AX_t + Z_t, \quad t = 1, 2, \ldots, T - 1,$$

where $A \in \mathbb{R}^{d \times d}$ is called transition matrix and $Z_1, Z_2, \ldots \sim_{i.i.d.} N_d(0, \Psi)$ are independent colored Gaussian noise with covariance matrix $\Psi$. Since the process is weakly stationary, we denote $\Sigma$ to be the covariance matrix of the time series,

$$\Sigma := \text{Var}(X_1) = \cdots = \text{Var}(X_T).$$

Let $u_1, \ldots, u_m$ be the top $m$ leading eigenvectors of $\Sigma$. We want to find $\hat{u}_1, \ldots, \hat{u}_m$ which can estimate $u_1, \ldots, u_m$ accurately. In the special case where $A$ is a zero matrix, this problem reduces to classical principal component analysis for i.i.d. Gaussian data. In more general settings where $A \neq 0$, it is well known that to secure the weakly stationarity of the time series in (1.1), we must have the spectral norm of $A$ (i.e., the largest singular value of $A$) 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-
alistic, poses significant theoretical challenges. Even in a simplified setting where $A = 0$, Johnstone and Lu (2009) show that the classical PCA is inconsistent under some conditions. In other words, letting the estimator $\hat{u}_1$ be the leading eigenvector of the sample covariance matrix, the angle between $u_1$ and $\hat{u}_1$, denoted as $\angle(u_1, \hat{u}_1)$, doesn’t converge to 0 as $T$ goes to infinity. To avoid such a curse of dimensionality, the population leading eigenvector $u_1$ is in general assumed to be sparse. More specifically, let $s$ be the number of nonzero elements in $u_1$, 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., $A = 0$). For example, greedy algorithms (d’Aspremont et al., 2008), lasso-type methods including SCuTLASS (Jolliffe et al., 2003), SPCA (Zou, 2006) and sPCA-rSVD (Shen and Huang, 2008), a number of power methods (Journé 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 $x_1, x_2, \ldots, x_T$ 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 we described before, the scans from two adjacent time points are obviously correlated. Though some work exists for low-dimensional PCA on dependent data (Skinner et al., 1986), no such result exists for high-dimensional PCA on dependent data (Skinner et al., 1986), no such result exists for low-dimensional PCA on dependent data. For example, in the fMRI imaging application we described before, the scans from two adjacent time points are obviously correlated. Though some work exists for low-dimensional PCA on dependent data (Skinner et al., 1986), no such result exists for high-dimensional PCA on dependent data. 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 dependent 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 $A$ as a nuisance parameter, we directly apply sparse PCA on the multivariate time series $x_1, \ldots, x_T$ to estimate the leading eigenvector $u_1$ as if the data are i.i.d. generated. Let $\angle(u_1, \hat{u}_1)$ be the angle between $u_1$ and $\hat{u}_1$, and $\lambda_1(\Sigma)$ be the $k$-th largest value of $\Sigma$, we provide an explicit rate of convergence for $\sin \angle(u_1, \hat{u}_1)$, i.e., for some absolute constant $C$,

$$|\sin \angle(u_1, \hat{u}_1)| \leq \frac{C}{\lambda_1(\Sigma)-\lambda_2(\Sigma)} \sqrt{\frac{s \log d}{T (1-\|A\|_2^2)}},$$

with high probability.

Our result allows the quantities $\lambda_1(\Sigma)$, $\lambda_2(\Sigma)$ and $\|A\|_2$ all scale with $d$. Since $\Sigma$ is jointly determined by the structure of the transition matrix $A$ and noise covariance matrix $\Psi$, this result suggests that the interaction between the structure of $A$ and $\Psi$ plays a pivotal role in determining the final rate of convergence. We also discuss specific structures of $A$ and $\Psi$ 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 results on dependent data analysis in high dimensions. We also discuss specific structures of $A$ and $\Psi$ 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.

2 Background

In this section, we briefly introduce the background of this paper. We start with notation. Let $A = [A_{jk}] \in \mathbb{R}^{d \times d}$ and $v = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$. For $0 \leq q \leq 1$, we define the vector $\ell_q$ norms as $\|v\|_q := (\sum_{j=1}^d |v_j|^q)^{\frac{1}{q}}$. Specifically $\|v\|_0 = \text{card}\{\text{supp}(v)\}$. We denote $\|A\|_q$ to be the operator norm of matrix $A$. In particular, $\|A\|_2$ is the spectral norm. Specifically, for $q = 1$ or $q = \infty$, $\|A\|_1 = \max_{1 \leq i \leq d} \sum_{j=1}^d |A_{ij}|$, and $\|A\|_\infty = \max_{1 \leq i \leq d} \sum_{j=1}^d |A_{ij}|$. We have $\|A\|_1 \leq \sqrt{\|A\|_1 \|A\|_\infty}$. Let $\lambda_j(A)$ be the $j$-th largest eigenvalue of $A$. The $d$-dimensional Euclidean unit sphere is $S^{d-1} := \{v \in \mathbb{R}^d, \|v\|_2 = 1\}$. The $\ell_q$ ball with radius $R_q$ is $B_q(R_q) := \{v \in \mathbb{R}^d, \|v\|_q \leq R_q\}$. For vector $v_1$ and $v_2$, define inner product as $\langle v_1, v_2 \rangle := v_1^T v_2$. For matrix $A_1, A_2$, we define the inner product as $\langle A_1, A_2 \rangle := \text{tr}(A_1^T A_2)$. For a set $K$, $|K|$ is its cardinality. We use $0$ 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 $x_1, x_2, \ldots, x_T$ are generated by the lag 1 autoregressive process:

$$x_{t+1} = AX_t + Z_t, \quad \text{for } t = 1, \ldots, T - 1.$$  

(2.1)
To secure the weakly stationary of the above process, the transition matrix $A$ must have bounded spectral norm $\|A\|_2 < 1$. We assume the Gaussian colored noise $Z_1, Z_2, \ldots \overset{i.i.d.}{\sim} N_d(0, \Psi)$. By assumption $Z_t$ and $X_t, X_{t-1}, \ldots$ are independent. The stationary property indicates

$$\Sigma = A\Sigma A^T + \Psi. \quad (2.2)$$

For $j \geq k$, the covariance between $X_j$ and $X_k$ is

$$\text{Cov}(X_j, X_k) = A \cdots A \Sigma := A^{j-k} \Sigma.$$ 

In the sequel we call $A$ the transition matrix and $\Psi$ the noise matrix. VAR model is widely used in the analysis of economic time series (Sims, 1980; Hatemi-J, 2004; Briiggemann and Liitkohl, 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 $x_1, x_2, \ldots, x_T$ be the $T$ observations of a random vector $X \in \mathbb{R}^d$. Let $u_1, \ldots, u_m$ be the top $m$ eigenvectors of the covariance matrix $\Sigma$. Let $\tilde{u}_1, \ldots, \tilde{u}_m$ be the top $m$ eigenvector of the sample covariance matrix $S$. In low dimensions, PCA uses $\tilde{u}_1, \ldots, \tilde{u}_m$ to estimate $u_1, \ldots, u_m$.

In the high dimensional settings where $d > T$, we assume that the leading eigenvector of $\Sigma$ are under certain sparse constraints. In other words, we assume that $u_1$ satisfies that $||u||_0 \leq s$ and $||u||_1 = 1$, i.e. $u_1 \in S^{d-1} \cap B_0(s)$. In this way, we define the model $\mathcal{M}(s, \Sigma, \lambda_1, \lambda_2)$ as follows:

$$\mathcal{M}(s, \Sigma, \lambda_1, \lambda_2) := \{ X : X \sim N_d(\mu, \Sigma), \lambda_1(\Sigma) = \lambda_1, \lambda_2(\Sigma) = \lambda_2, ||u||_0 = s \},$$

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

$$\tilde{u}_1 := \arg \max_{v \in \mathbb{R}^d} v^T S v,$$

subject to $v \in S^{d-1} \cap B_0(s). \quad (3.1)$

Here $\tilde{u}_1$ is the global optimal estimator of $u_1$. In this paper we only discuss about $\tilde{u}_1$ corresponding to the leading eigenvalue $\lambda_1(\Sigma)$. Analysis for $\tilde{u}_k$ corresponding to $\lambda_k(\Sigma)$ 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 nonsymmetric 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 $\tilde{u}_1$ obtained by (3.1) can approximate $u_1$ in a parametric rate with respect to $(n, d, s)$, and the upper bound is also related to the transition matrix $A$ and the noisy matrix $\Psi$.

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

$$|\sin \angle(u_1, \tilde{u}_1)| = O_P \left( \frac{1}{\lambda_1-\lambda_2} \sqrt{s \log d \left( \frac{\|\Sigma\|_2}{T} \right)} \right).$$

Here for any two vectors $v_1$ and $v_2 \in S^{d-1}$, $|\sin \angle(v_1, v_2)| := \sqrt{1 - (v_1^T v_2)^2}$.

**Remark 4.2.** The bound obtained in (4.1) depends on $\Sigma, A, \Psi$, where $A$ characterizes the data dependence degree. When both $||A||_2$ and $||\Psi||_2$ 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 $\sin \angle(u_1, \tilde{u}_1)$ with $\sup_{v \in S^{d-1} \cap B_0(2s)} |v^T (\Sigma - S) v|$. In the sequel, we assume that the assumptions in Theorem 4.1 hold.

**Lemma 4.3.** $u_1$ and $\tilde{u}_1$ satisfy

$$\sin \angle(u_1, \tilde{u}_1) \leq \frac{2}{\lambda_1-\lambda_2} \sup_{v \in S^{d-1} \cap B_0(2s)} |v^T (\Sigma - S) v|. \quad (4.1)$$

**Proof.** Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $\Sigma$. Let $u_1, u_2, \ldots, u_d$ be the corresponding eigenvectors. We have $u_i^T u_j = 0$ for $i \neq j$ and $\Sigma = \sum_{j=1}^d \lambda_j u_j u_j^T$. Let $\Phi_0$ be represented by $u_1$ and $\Sigma$ as

$$\Phi_0 = \Sigma - \lambda_1 u_1 u_1^T = \Sigma - \lambda_1 u_1 u_1^T - \lambda_1 u_1 u_1^T + \lambda_1 u_1 u_1^T = \Sigma - u_1 u_1^T - \sum_{j=2}^d \lambda_j u_j u_j^T + u_1 u_1^T = (I_d - u_1 u_1^T) \Sigma (I_d - u_1 u_1^T).$$

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

$$\langle \Sigma, u_1 u_1^T - uu^T \rangle = \langle \Sigma, u_1 u_1^T \rangle - \langle \lambda_1 u_1 u_1^T + \Phi_0, uu^T \rangle = \lambda_1 - \lambda_1 \langle u_1, u_1 \rangle^2 - \langle \Phi_0, uu^T \rangle = \lambda_1 - \lambda_1 \langle u_1, u_1 \rangle^2 - \langle u^T (I_d - u_1 u_1^T) \Sigma (I_d - u_1 u_1^T) u \rangle.$$ 

Now we consider the unit vector $a = (I_d - u_1 u_1^T) u / \| (I_d - u_1 u_1^T) u \|_2 \in \mathbb{R}^d$.

It is easy to verify $a$ is orthogonal to $u_1$. Therefore, $a \in \text{span}(u_2, \ldots, u_d)$. Letting $a_j = a^T u_j, j = 2, \ldots, d$, we can get $\sum_{j=2}^d a_j^2 = 1$. Therefore we have
\[ a^T \Sigma a = a^T \left( \lambda_1 u_1 u_1^T + \sum_{j=2}^{d} \lambda_j u_j u_j^T \right) a = \sum_{j=2}^{d} \lambda_j a_j^2 \leq \lambda_2, \]
which indicates
\[ u^T (I_d - u_1 u_1^T) \Sigma (I_d - u_1 u_1^T) u \leq \lambda_2 \left\| (I_d - u_1 u_1^T) u \right\|_2^2. \]
Since \( u \) is on the unit sphere \( S^{d-1} \), we have \( \|u\|_2 = 1 \). Therefore, \( \| (I_d - u_1 u_1^T) u \|_2^2 = 1 - \langle u_1, u \rangle^2 \). Now we obtain
\[ \langle \Sigma, u_1 u_1^T - uu^T \rangle \geq (\lambda_1 - \lambda_2) \left( 1 - \langle u_1, u \rangle^2 \right). \tag{4.2} \]
Since (4.2) holds for any \( u \in S^{d-1} \), letting \( u = \hat{u}_1 \) we have
\[ \langle \Sigma, u_1 u_1^T - uu^T \rangle \geq (\lambda_1 - \lambda_2) \sin^2 \angle (u_1, \hat{u}_1). \]
Since \( \hat{u}_1 \) is defined as \( \hat{u}_1 := \arg \max_{u \in \mathbb{R}^d} v^T Su \), we know
\[ \hat{u}_1^T S \hat{u}_1 - u_1^T S u_1 \leq 0. \]
Therefore, \( \langle S, u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \rangle \leq 0 \), we have
\[ \sin^2 \angle (u_1, \hat{u}_1) \leq \frac{1}{\lambda_1 - \lambda_2} \langle \Sigma, u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \rangle \leq \frac{1}{\lambda_1 - \lambda_2} \langle \Sigma - S, u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \rangle. \]
Let \( \Pi \) be the diagonal matrix with diagonal values being 1 if and only if the corresponding entries in \( u_1 \) or \( \hat{u}_1 \) are zero. Therefore, we know there are at most \( 2s \) nonzero values in \( \Pi \). Then we have
\[ \sin^2 \angle (u_1, \hat{u}_1) \leq \frac{1}{\lambda_1 - \lambda_2} \langle \Sigma - S, u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \rangle \leq \frac{1}{\lambda_1 - \lambda_2} \langle \Pi (\Sigma - S) \Pi, u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \rangle \leq \frac{1}{\lambda_1 - \lambda_2} \| \Pi (\Sigma - S) \Pi \|_2 \| u_1 u_1^T - \hat{u}_1 \hat{u}_1^T \|_S \leq \frac{1}{\lambda_1 - \lambda_2} \| \Pi (\Sigma - S) \Pi \|_2 \cdot 2 \sin \angle (u_1, \hat{u}_1). \]
Here \( \| \cdot \|_S \) denotes the sum of singular values. This implies
\[ \| \sin \angle (u_1, \hat{u}_1) \| \leq \frac{2}{\lambda_1 - \lambda_2} \| \Pi (\Sigma - S) \Pi \|_2 \leq \frac{2}{\lambda_1 - \lambda_2} \sup_{v \in S^{d-1} \cap \mathbb{B}_n(2s)} |v^T (\Sigma - S) v|. \]
This completes the proof. \qed

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

**Lemma 4.4.** Provided that \( x_1, x_2, \ldots, x_T \overset{i.i.d.}{\sim} N(0, 1) \). \( X = (x_1, x_2, \ldots, x_T)^T \in \mathbb{R}^T \) is a random vector. Mapping \( f : \mathbb{R}^T \to \mathbb{R} \) is Lipschitz, i.e., for any \( v_1, v_2 \in \mathbb{R}^T \):
\[ \exists L \geq 0, \quad \text{s.t.} \quad |f(v_1) - f(v_2)| \leq L \|v_1 - v_2\|_2, \]
Then for any \( t > 0 \) we have,
\[ \mathbb{P} \left( |f(X) - \mathbb{E} f(X)| > t \right) \leq 2 \exp \left( -\frac{t^2}{2L^2} \right). \tag{4.3} \]
**Proof.** We refer to Ledoux and Talagrand (2011)’s proof of this lemma. \qed

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

**Lemma 4.5.** Letting \( v \in S^{d-1} \cap \mathbb{B}_q(R_q) \), we have
\[ \left| \sqrt{v^T \Sigma v} - \sqrt{v^T S v} \right| = O_p \left( \frac{1}{T} \right). \tag{4.4} \]
**Proof.** Let \( x_1, x_2, \ldots, x_T \) be the \( T \) observations. \( Y = (x_1^T v, x_2^T v, \ldots, x_T^T v)^T \in \mathbb{R}^T \) is a zero-mean Gaussian random vector. We denote the covariance matrix of \( Y \) as \( \Sigma_Y \). \( \Sigma_Y \) is symmetric and semi-definite. Thus \( \Sigma_Y \) can be decomposed as \( \Sigma_Y = Q^T Q \), where \( Q \) is a matrix with orthogonal columns. Let \( \sigma = \|Q\|_2 = \|\Sigma_Y\|_2^{1/2} \). According to the definition of \( Y \), we have
\[ v^T S v = v^T \left( \sum_{i=1}^T x_i x_i^T \right) v = \frac{Y^T Y}{T} = \left( \frac{\|Y\|_2}{\sqrt{T}} \right)^2, \]
\[ v^T \Sigma v = v^T \mathbb{E} (XX^T) v = \mathbb{E} \left( \frac{Y^T Y}{T} \right) = \mathbb{E} \left( \frac{\|Y\|_2^2}{\sqrt{T}} \right). \]
For convenience we define \( W := \frac{\|Y\|_2^2}{\sqrt{T}} \) and \( f(v) := \|Qv\|_2^2 / \sqrt{T} \) for \( v \in \mathbb{R}^T \). Since we have \( Y = QV \), \( V \in \mathbb{R}^T \) is a zero-mean Gaussian vector with covariance matrix \( I_T \). 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
\[ \mathbb{P} \left( |W - \mathbb{E}(W)| \geq t \right) \leq 2 \exp \left( -\frac{t^2 T}{2 \sigma^2} \right). \tag{4.5} \]
Since \( \text{Var}(W) = \mathbb{E}(W^2) - \mathbb{E}^2(W) \geq 0 \), at the same time we have \( \mathbb{E}(W^2) \geq 0, \mathbb{E}(W) \geq 0 \), which implies \( \sqrt{\mathbb{E}(W^2)} - \mathbb{E}(W) \geq 0 \). Thus we get
\[ \left( \sqrt{\mathbb{E}(W^2)} - \mathbb{E}(W) \right)^2 \leq \mathbb{E} \left( \left| W - \mathbb{E}(W) \right|^2 \right) = \frac{4\sigma^2}{T}, \]
which indicates
\[ \left| \sqrt{\mathbb{E}(W^2)} - \mathbb{E}(W) \right| \leq \frac{2\sigma}{\sqrt{T}}. \tag{4.6} \]
Here (4.6) together with \( |W - \mathbb{E}(W)| \leq t \) implies \( |W - \sqrt{\mathbb{E}(W^2)}| \leq t + 2\sigma / \sqrt{T} \). Therefore, according
to the definition of $Y$ and $W$,
$$
P \left( \left| \sqrt{v^T \Sigma v} - \sqrt{v^T W^2} \right| \geq t + 2\sigma \sqrt{T} \right) = P \left( |W - \sqrt{E(W^2)}| \geq t + 2\sigma \sqrt{T} \right) \leq P \left( |W - E(W)| \geq t \right) \leq 2 \exp \left\{ -\frac{t^2 T}{2\sigma^2} \right\}.
$$
(4.7)

We reach the conclusion
$$
\left| \sqrt{v^T \Sigma v} - \sqrt{v^T W^2} \right| = O_p \left( \sqrt{\frac{1}{T}} \right).
$$

This completes the proof. □

**Lemma 4.6.** An $\epsilon$-net $N_\epsilon$ of a sphere $S^{n-1}$ (equipped with Euclidean distance) is a subset of $S^{n-1}$ such that for any $v \in S^{n-1}$, there exists $u \in N_\epsilon$, subject to $\|u - v\|_2 \leq \epsilon$. It is shown by Vershynin (2010) that for any $\epsilon > 0$,
$$
|N_\epsilon| \leq \left( 1 + \frac{2}{\epsilon} \right)^n.
$$
(4.8)

Also, for matrix $A \in \mathbb{R}^{n \times n}$, the inequality below holds for any $\epsilon \in (0, 1)$
$$
\max_{v_i \in S^{n-1}} |v_i^T A v_1| \leq (1 - 2\epsilon)^{-1} \max_{v_2 \in N_\epsilon} |v_2^T A v_2|.
$$
(4.9)

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

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

Now we derive the bound of $|N_\epsilon|$. We cover the neighborhood of every $x_i \in N_\epsilon$ with disjoint balls $B_{y_i} = \{y_i \mid \|y_i - x_i\|_2 < \epsilon/2\}$. For any $B_{y_i} \subset B_0 = \{y \mid \|y\|_2 < 1\}$ we have
$$
|N_\epsilon| \cdot |B_{y_i}| = \sum_{i=1}^{N_\epsilon} |B_{y_i}| = \left| \bigcup_{i=1}^{N_\epsilon} B_{y_i} \right| \leq |B_0|,
$$
$$
|N_\epsilon| \leq \frac{|B_0|}{|B_{y_i}|} = \left( \frac{1 + \frac{\epsilon}{2}}{\epsilon} \right)^n = \left( 1 + \frac{2}{\epsilon} \right)^n.
$$

Then we turn to prove (4.9). For $v_1 \in S^{n-1}$ and $v_2 \in N_\epsilon$, we know since $\|v_1 - v_2\|_2 \leq \epsilon$,
$$
|v_1^T A v_1 - v_2^T A v_2| \leq \|A\|_2 \|v_1\|_2 |v_1 - v_2|_2 + \|A\|_2 |v_2|_2 |v_1 - v_2|_2 \leq 2\epsilon \|A\|_2
$$

It follows that
$$
|v_2^T A v_2| \geq (1 - 2\epsilon) \|A\|_2 = (1 - 2\epsilon) \max_{v_1 \in S^{n-1}} |v_1^T A v_1|
$$

Taking the maximum over $v_2 \in N_\epsilon$, 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
$$
\sqrt{v^T \Sigma v} = \sqrt{E(W^2)} = \sqrt{\frac{\text{tr}(QQ^T)}{T}} \leq \frac{\|Q\|_F}{\sqrt{T}} \leq \sigma.
$$

Then using (4.7) we can get
$$
P \left( \left| \sqrt{v^T \Sigma v} + \sqrt{v^T W^2} \right| \geq t + 4\sigma \right) \leq P \left( \left| \sqrt{v^T \Sigma v} - \sqrt{v^T W^2} \right| \geq t + 2\sigma \right) \leq 2 \exp \left\{ -\frac{t^2 T}{2\sigma^2} \right\}.
$$
(4.10)

We define events $\mathcal{E}_1$ and $\mathcal{E}_2$, combining (4.7) (with $t = t_1$) and (4.10) (with $t = t_2$),
$$
\mathcal{E}_1 := \left\{ \left| \sqrt{v^T \Sigma v} - \sqrt{v^T W^2} \right| \geq t_1 + 2\sigma \sqrt{T} \right\},
$$
$$
\mathcal{E}_2 := \left\{ \left| \sqrt{v^T \Sigma v} + \sqrt{v^T W^2} \right| \geq t_2 + 4\sigma \right\}.
$$

We derive for any $t_1, t_2 > 0$,
$$
P \left( \|v^T (\Sigma - S)v\| \geq t_1 + 2\sigma \sqrt{T} \right) \leq \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) \leq 2 \exp \left\{ -\frac{t_1^2 T}{2\sigma^2} \right\} + 2 \exp \left\{ -\frac{t_2^2 T}{2\sigma^2} \right\}.
$$
(4.11)

Now we turn to upper bound $\sup_{v \in S^{d-1} \cap B_0(2\sigma)} |v^T (\Sigma - S)v|$ in Lemma 4.3. Assuming we have a fixed subset $K \subset \{1, \ldots, d\}$, define
$$
\mathbb{B}_K = \{v \mid \text{for any } i \in \{1, \ldots, d\} \setminus K, v_i = 0\}.
$$

For any $t_1, t_2 > 0$, we define event $\mathcal{E}_K$ and $\mathcal{E}_v$ as
$$
\mathcal{E}_K := \left\{ \max_{v \in S^{d-1} \cap B_0} |v^T (\Sigma - S)v| \geq 2 \left( t_1 + 2\sigma \sqrt{T} \right) (t_2 + 4\sigma) \right\},
$$
$$
\mathcal{E}_v := \left\{ |v^T (\Sigma - S)v| \geq 2 \left( t_1 + 2\sigma \sqrt{T} \right) (t_2 + 4\sigma) \right\}.
$$

According to Lemma 4.6, we define the $\frac{1}{4}$-net of $S^{d-1} \cap \mathbb{B}_K$ as $N_{K_\epsilon}$, then we can get
$$
\mathcal{E}_K \subset \bigcup_{v \in N_{K_\epsilon}} \left\{ |v^T (\Sigma - S)v| \geq 2 \left( t_1 + 2\sigma \sqrt{T} \right) (t_2 + 4\sigma) \right\},
$$
Combining (4.11) and (4.8) and letting $|K| = 2s$, we obtain
\[
\mathbb{P}(E_K) \leq \sum_{v \in N_K} \mathbb{P}(E_v) = |N_K| \mathbb{P}(E_v) \\
\leq g^{2s} \left( 2 \exp \left( - \frac{t_1^2 T}{2\sigma^2} \right) + 2 \exp \left( - \frac{t_2^2 T}{2\sigma^2} \right) \right).
\]

Now we consider arbitrary subset $K \subset \{1, \ldots, d\}$ with cardinality $2s$. We define
\[
E_K := \left\{ \max_{v \in \mathbb{S}^{d-1} \cap B_0(2s)} |v^T(S - S)v| \geq 2 \left( 1 + 2 \frac{\sigma}{\sqrt{T}} \right) (t_2 + 4\sigma) \right\}.
\]

Then we have
\[
\mathbb{P}(E_K) \leq \sum_{K \subset \{1, \ldots, d\}} \mathbb{P}(E_v) \leq \binom{d}{2s} \mathbb{P}(E_K) \\
\leq g^{2s} \left( \frac{d}{2s} \right) \left( 2 \exp \left( - \frac{t_1^2 T}{2\sigma^2} \right) + 2 \exp \left( - \frac{t_2^2 T}{2\sigma^2} \right) \right).
\]

Using (4.1), we can get for any $t_1, t_2 > 0, v \in \mathbb{S}^{d-1} \cap B_0(2s)$, which implies
\[
\mathbb{P} \left( \sin \angle(u_1, \tilde{u}_1) \right) \geq \frac{4}{\lambda_1 - \lambda_2} \left( t_1 + 2 \frac{\sigma}{\sqrt{T}} \right) (t_2 + 4\sigma) \\
\leq \mathbb{P} \left( \max_{v \in \mathbb{S}^{d-1}} |v^T(S - S)v| \geq 2 \left( 1 + 2 \frac{\sigma}{\sqrt{T}} \right) (t_2 + 4\sigma) \right) \\
\leq g^{2s} \left( \frac{d}{2s} \right) \left( 2 \exp \left( - \frac{t_1^2 T}{2\sigma^2} \right) + 2 \exp \left( - \frac{t_2^2 T}{2\sigma^2} \right) \right).
\]

Now we are going to derive the upper bound for $\sigma^2 = \|\Sigma_Y\|_2$. We note that VAR model assumes $\|A\|_2 < 1$. Thus $\sigma^2$ can be bounded as follows,
\[
\sigma^2 = \|\Sigma_Y\|_2 \leq \|\Sigma_Y\|_\infty = \max_{i \in \{1, \ldots, T\}} \sum_{j=1}^T |(\Sigma_Y)_{ij}| \quad (4.12)
\]

We have $\|\Sigma_Y\|_2 \leq \sqrt{\|\Sigma_Y\|_1 \|\Sigma_Y\|_\infty} = \|\Sigma_Y\|_\infty$, since $\Sigma_Y$ is symmetric. According to the definition (1.1) we have,
\[
v^T \text{Cov}(x_i, x_j)v = v^T \text{Cov}(x_j, x_i)v = v^T A^{i-j} \Sigma v,
\]

In order to get $\max_{i \in \{1, \ldots, T\}} \sum_{j=1}^T |(\Sigma_Y)_{ij}|$ in (4.12), we first derive the upper bound of $\sum_{j=1}^T |(\Sigma_Y)_{ij}|$
\[
\sum_{j=1}^T |(\Sigma_Y)_{ij}| = \sum_{j \neq i} |v^T \text{Cov}(x_i, x_j)v| + |v^T \text{Cov}(x_i, x_i)v| \\
= \sum_{j \neq i} |v^T A^{i-j} \Sigma v| + |v^T \Sigma v| \\
\leq \sum_{j \neq i} \|\Sigma\|_2 \|A\|_{2}^{i-j} + \|\Sigma\|_2 \\
\leq 2\|\Sigma\|_2 \left( 1 - \|\Sigma\|_2 \right).
\]

Thus, letting $\delta^2 = 2\|\Sigma\|_2 / (1 - \|\Sigma\|_2) \geq \sigma^2$, we have
\[
\mathbb{P} \left( |\sin \angle(u_1, \tilde{u}_1)| \geq \frac{4}{\lambda_1 - \lambda_2} \left( t_1 + 2 \frac{\delta}{\sqrt{T}} \right) (t_2 + 4\delta) \right) \\
\leq g^{2s} \left( \frac{d}{2s} \right) \left( 2 \exp \left( - \frac{t_1^2 T}{2\delta^2} \right) + 2 \exp \left( - \frac{t_2^2 T}{2\delta^2} \right) \right).
\]

Letting $t_1 = \sqrt{s\delta^2 \log d/T}$ and $t_2$ be a constant, we reach the conclusion,
\[
|\sin \angle(u_1, \tilde{u}_1)| = O_P \left( \frac{1}{\lambda_1 - \lambda_2} \sqrt{\frac{s \log d}{T}} \left( \|\Sigma\|_2 \right) \right).
\]

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 $\hat{u}_1$ obtained in (4.2).

5.1 Synthetic Data

In this section we experiment with sparse PCA on synthetic data. We show how $A$, $\Sigma$ and $\Psi$ affect $|\sin \angle(u_1, \tilde{u}_1)|$. First we create $A$ given $\|A\|_2$. With $\lambda_1(A)$ and $\lambda_2(A)$, we generate
\[
\Sigma = (\lambda_1(A) - 1)u_1^T u_1 + (\lambda_2(A) - 1)u_2^T u_2 + I,
\]

where $\|u_1\|_0 = s$, $u_1$ and $u_2$ is orthogonal. According to the stationary property (2.2), the covariance matrix of the noise random vector $\mathbf{Z}_t$ is $\Psi = \Sigma - A^T \Sigma A$, where $\Psi$ 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 $\|A\|_2$ affects the accuracy of estimator $\tilde{u}_1$. Set $\lambda_1(\Sigma) = 10$, $\lambda_2(\Sigma) = 5$, $s = 20$, $d = 200$ and $\|A\|_2 \in [0,0.9]$, we repeatedly experiment for 3000 times for each $\|A\|_2$. Here $\Sigma$ is fixed while $A$ and $\Psi$ are varying. The results are illustrated in Figure 1 by plotting the relevant part in the theoretical upper bound $\|\Sigma\|_2 / (1 - \|A\|_2)$ against the empirical error $|\sin \angle(u_1, \tilde{u}_1)|$. As can be observed in Fig. 1, the empirical error increases when the spectral norm of transition matrix $\|A\|_2$ increases. This makes sense because when $\|A\|_2 = 0$, VAR model is reduced to the i.i.d. case, since $X_{t+1}$ doesn’t depend on $X_t$. When $\|A\|_2 \to 1$, 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 $|\sin \angle(u_1, \tilde{u}_1)|$ for each $\|A\|_2$. 

\[53\]
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 $P = [P_{t,j}]$ 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 $t - 1$:

$$X_{t,j} = \log(P_{t+1,j}/P_{t,j}), \quad t = 1, \ldots, T - 1.$$  

It is obvious that there exists data dependence struc-
ture between $X_{t_1,j}$ and $X_{t_2,j}$ for any $t_1, t_2 \leq T - 1$ and it accordingly raises concern for conducting classical sparse PCA algorithms on this dataset $X$. 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.

<table>
<thead>
<tr>
<th>$\hat{u}_1$</th>
<th>$\hat{u}_2$</th>
<th>$\hat{u}_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Consumer Discretionary</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Financials</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Financials</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Financials</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Financials</td>
</tr>
<tr>
<td>Financials</td>
<td>Industrials</td>
<td>Financials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Industrials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Industrials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Industrials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Industrials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Industrials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Information Technology</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Information Technology</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Materials</td>
</tr>
<tr>
<td>Financials</td>
<td>Materials</td>
<td>Materials</td>
</tr>
</tbody>
</table>
References


