Sparse Principal Component Analysis in Frequency Domain for Time Series

Junwei Lu[*], Yichen Chen[†], Xiuneng Zhu[‡], Fang Han[§], and Han Liu[¶]

Abstract

We consider the sparse principal component analysis of high dimensional time series and propose a dynamic component analysis to compress the time series to lower dimension subseries based on the spectral density matrix. The number of sample cross-covariance matrices we need for our spectral density matrix estimator is only logarithmic to the sample size, which reveals a novel phenomenon in comparison to the classical methods. A new chained sparse orthogonal pursuit algorithm combined with a cosine interpolation method is proposed to efficiently estimate the eigenvectors of the spectral density matrices. We also provide theoretical results on the error of low-dimension approximation and compare the results with the optimal compression. Our method and theoretical results are model-free and only rely on the dependency of the time series. The numerical results on both synthetic data and neural imaging dataset are provided to support the theoretical results.

1 Introduction

The exploration of the intrinsic dimension of high dimensional time series is vital in the study of econometrics, financial analysis, signal processing and neuroscience. Given a d-dimensional stationary time series $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$, we aim to find a sequence of low dimensional time series whose linear combination approximates $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$. If the observed time series is a sequence of independent data points, this problem becomes the high dimensional sparse principal component analysis (PCA) (Johnstone and Lu, 2009). It aims at estimating the sparse principal subspaces of the marginal covariance matrix of $\boldsymbol{X}_t$. Various sparse eigenvector estimators have been proposed in the literature. Witten et al. (2009); Zou et al. (2006) and Shen and Huang (2008) consider the penalized low rank approximation. Yuan and Zhang (2013); Ma et al. (2013) and Wang et al. (2014) propose truncated power method and thresholding QR iterative algorithm to compute the sparse eigenspaces. d'Aspremont et al. (2008); Vu et al. (2013) and Lei and Vu (2015) consider a convex relaxation and

[*]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: junweil@princeton.edu

[†]Computer Science Department, Princeton University, Princeton, NJ 08544, USA;e-mail: yichenc@princeton.edu

[‡]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: xiunengz@princeton.edu

[§]Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA; e-mail: fhan@jhsph.edu

[¶]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu

apply Fantope projected algorithm to get the eigenvectors. However, these works are restricted to the dataset with independent data points and leave the time series models untouched.

For the dependent time series dataset, there are two major tracks on the dimension reduction. One is reducing the parameter space in the time domain and the other in the frequency domain. The family of methods in the time domain mostly relies on specific multivariate time series models or structural assumptions. Stock and Watson (2002) study the dynamic factor model and estimate the factors by principal component analysis and similar ideas are also discussed in Stock and Watson (2005); Forni et al. (2005); Pan and Yao (2008) and Lam and Yao (2012) . Fan et al. (2013) consider the structural assumption that the covariance matrix of the factor model can be decomposed into a low rank and a sparse matrix and suggest a thresholding principal orthogonal complements method to estimate the factor structure. There are also many other dimension reduction methods for the factor model including the canonical component analysis (Box and Tiao, 1977), the independent component analysis (Back and Weigend, 1997), the scalar component model (Tiao and Tsay, 1989) and the dynamic orthogonal components analysis (Matteson and Tsay, 2011). Chang et al. (2014) seek for a generalized PCA to segment the time series into a contemporaneous linear combination of lower dimensional subseries. Besides the factor model, the vector autoregressive (VAR) model is considered by imposing sparsity assumptions on the parameters to reduce the dimension (Qiu et al., 2015; Guo et al., 2015).

Another track of the time series principal component analysis is in the frequency domain. The advantage of the PCA in the frequency domain over the time domain is that it can handle larger family of models. Brillinger (1969) shows that the spectral density matrix obtained from the Fourier transform of the cross-covariance matrices is crucial to the principal component analysis. Stoffer (1999) utilizes the spectral density matrix to detect the series with common spectral power at similar frequency. Jung et al. (2014) study the conditional independence graphical model of discrete Gaussian processes and propose regularized estimator based on the spectral density estimator.

We consider the dimension reduction for time series whose dimensions are much larger than the sample size. We estimate the sparse principal eigenvectors for the high dimensional spectral density matrices and build a time series compression procedure based on these eigenvectors. Our paper contributes into the area of time series PCA in three aspects. First, we propose a computationally efficient spectral density matrix estimator. The spectral density matrix has many applications in time series analysis (Martin, 2001; Manolakis et al., 2005). It plays a vital role in revealing the dependency properties of time series (Xiao and Wu, 2012; Basu and Michailidis, 2015). However, existing spectral density matrix estimators are computationally intractable when both the sample size and variable dimension are large. For example, the Blackman-Tukey estimator (Stoica and Moses, 1997) requires the number of cross-covariance matrix estimators involved in the calculation of Fourier transform to be of the same order of the sample size T. The same sample complexity is needed for the high dimensional scenario considered in Jung et al. (2014). This leads to the computation bottleneck of the implementation of dimension reduction for high dimensional time series. However, we show that only $O(\log T)$ cross-covariance matrices are enough to achieve the optimal rate for the spectral density matrix estimator and its eigenspace estimator. This result significantly improves the computation complexity of our algorithm. To obtain this result, we develop the theory for the concentration inequality for the sample cross-covariance matrix of dependent random variables. Second, we provide efficient procedures to obtain estimators with uniform statistical rates on the full spectrum of both time and frequency domain. To uniformly estimate the eigenvectors of the spectral density matrix on different frequencies, we consider a

discretization and interpolation strategy. The discretization step separates the frequency domain into grids and estimates the sparse eigenvectors on these frequencies. The estimation is based on a new multitask sparse PCA algorithm for complex valued Hermitian matrices, which is generalized from the work of Wang et al. (2014) for a single real matrix. The interpolation step considers a novel cosine interpolation procedure interlacing the eigenvectors of frequencies among the grids. Our theoretical results show that eigenvectors on $O(\sqrt{T})$ grids is enough to achieve a valid compressed time series. Third, our theoretical analysis is model-free and do not impose any specific time series model. The theoretical analysis of this paper only requires the α-mixing condition on the dependency of the time series, which makes our method applicable to a wide range of models. We provide a general theoretical result that the compression error of our method converges to the optimal one and characterize the convergence rate by the level of dependency.

1.1 Notation System

Without further notice, all the vectors and matrices will be assumed to take values in $\mathbb{C}$. In this paper, the notation $i = \sqrt{-1}$ is the imaginary unit. For a complex number $z = x + iy$, $\bar{z}$ is the complex conjugate of z and $|z| = \sqrt{x^2 + y^2}$ is the absolute value of z. The ℓ_p-norm of a complex value vector $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)^T \in \mathbb{C}^d$ is $\|\mathbf{v}\|_p = (\sum_{j=1}^d |\mathbf{v}_j|^p)^{1/p}$. We define the unit ball in $\mathbb{C}^d$ as $\mathbb{S}^{d-1}(\mathbb{C}) = \{\mathbf{v} \in \mathbb{C}^d | \|\mathbf{v}\|_2 = 1\}$. For a complex matrix $\mathbf{A}$, $\mathbf{A}^\dagger = \bar{\mathbf{A}}^T$ denotes the Hermitian conjugate of $\mathbf{A}$. $\mathbf{A}$ is called Hermitian if $\mathbf{A} = \mathbf{A}^\dagger$. For a matrix $\mathbf{A}$ of size $p \times q$, the (a, b)-norm for $a, b \in [1, \infty]$, denoted by $\|\mathbf{A}\|_{a,b}$, is defined as the ℓ_b norm of the vector consisting of all the ℓ_a norms of the rows of $\mathbf{A}$. Hence $\|\mathbf{A}\|_{2,0}$ is the number of nonzero rows of $\mathbf{A}$. The spectral norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_2 = \max\{|\mathbf{v}^\dagger \mathbf{A} \mathbf{v}| \,\big|\, \|\mathbf{v}\|_2 = 1\}$. For two square complex matrices of the same dimensions $\mathbf{A}$ and $\mathbf{B}$, the matrix inner product $\langle \mathbf{A}, \mathbf{B} \rangle$ is defined as $\mathrm{tr}(\mathbf{A}^\dagger \mathbf{B})$. The induced norm is called Frobenius norm and and is given by $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathbf{A}^\dagger \mathbf{A})}$. We use $\lambda_j(\mathbf{A})$ to denote the j-th largest eigenvalue of a Hermitian matrix $\mathbf{A}$, and use $\sigma_j(\mathbf{A})$ to denote the j-th largest singular value of $\mathbf{A}$.

1.2 Paper Organization

The remainder of this paper is organized as follows. In Section 2, we provide the preliminary results on the time series and spectral density matrix. In Section 3, we propose a computationally efficient estimator for spectral density matrix. In Section 4, we develop the Dynamic Component Analysis (DCA) algorithm to compress the high dimension time series based on the estimators in Section 3. Its statistical and computational performance is analyzed in Section 4.2. Section 5 is devoted to numerical studies. The technical proofs are listed in the Appendix.

2 Background

This section provides necessary background for the dimension reduction of time series. Before presenting the main part of this section, we first define several notations. For a d-dimensional, mean-zero and stationary time series $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$, define the cross-covariance matrices $\{\mathbf{R}_u\}_{u \in \mathbb{Z}}$ as $\mathbf{R}_u = \mathbb{E}[\boldsymbol{X}_u \boldsymbol{X}_0^\dagger]$, which captures the second-moment characteristics of the interdependencies between two random variables in the time series with lag u. In this paper, we assume that $\sum_{t=-\infty}^{\infty} |(\mathbf{R}_t)_{jk}| < \infty$

for each $1 \le j, k \le d$. The spectral density matrices are defined as

$$\mathbf{S}(\omega) = \sum_{t=-\infty}^{\infty} \mathbf{R}_t \exp(-i2\pi\omega t), \quad \omega \in [0, 1). \tag{2.1}$$

2.1 Optimality Characterization of PCA

We first quickly review the optimal compression characterization of PCA, and generalize it for time series. We assume that we know the population covariance matrix in this section.

Assume that we have a d-dimensional random variable $\boldsymbol{X}$ such that $\mathbb{E}[\|X\|_2^2] < \infty$, our objective is to linearly transform it into a q-dimensional random variable where $q < d$ by multiplying $\boldsymbol{X}$ by a $q \times d$ matrix $\mathbf{B}$, and then transform it back into dimension d by multiplying a $d \times q$ matrix $\mathbf{C}$. The expected error or information loss, due to this compression-recovery procedure, can be quantified as $\mathbb{E}\big[(\boldsymbol{X} - \mathbf{CB}\boldsymbol{X})^\dagger(\boldsymbol{X} - \mathbf{CB}\boldsymbol{X})\big]$. The matrices $\mathbf{B}$ and $\mathbf{C}$ minimizing $\mathbb{E}\big[(\boldsymbol{X} - \mathbf{CB}\boldsymbol{X})^\dagger(\boldsymbol{X} - \mathbf{CB}\boldsymbol{X})\big]$ are given by

$$\mathbf{B} = (\mathbf{v}_1, \ldots, \mathbf{v}_q)^T, \quad \mathbf{C} = \mathbf{B}^T,$$

where $\mathbf{v}_1, \cdots, \mathbf{v}_q$ are the leading q eigenvectors of population covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\dagger]$. Note that the q principal components $\{\mathbf{v}_i^T \boldsymbol{X}\}_{1 \le i \le q}$ are uncorrelated. This characterization is well-known in the literature and its proof can be found in, e.g., Jolliffe (2002).

A similar optimal compression problem can also be considered in the time series context, and accordingly, the linear transformations of data should be replaced by two linear filters. More precisely, assume that we have a stationary d-dimensional time series $\{\boldsymbol{X}_t\}_{t \in \mathbb{Z}}$ taking values in $\mathbb{C}$. Let

$$\mathbf{b}(t) : \mathbb{Z} \to \mathcal{M}_{q \times d}(\mathbb{C}), \quad \mathbf{c}(t) : \mathbb{Z} \to \mathcal{M}_{d \times q}(\mathbb{C}),$$

be two linear filters, where q is an integer smaller than d. Define the q-dimensional time series $\boldsymbol{\zeta}_t$ via the filter $\mathbf{b}(\cdot)$ as $\boldsymbol{\zeta}_t = \sum_{u \in \mathbb{Z}} \mathbf{b}(t-u)\boldsymbol{X}_u$, and we back out an approximate of $\boldsymbol{X}$ via $\mathbf{c}(\cdot)$ as

$$\boldsymbol{X}_t^* = \sum_{u \in \mathbb{Z}} \mathbf{c}(t-u)\boldsymbol{\zeta}_u. \tag{2.2}$$

Because of the stationarity assumption, the quality of this approximation can be measured via the expected squared recover error between $\boldsymbol{X}_t$ and $\boldsymbol{X}_t^*$ at a single time $\mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_t^*\|_2^2]$. We have the following theorem characterizing optimal filters $\mathbf{b}(\cdot)$ and $\mathbf{c}(\cdot)$ in the sense of minimizing the expected squared recover error.

Theorem 2.1 (Theorem 3.1, Brillinger (1969))**.** Let $\boldsymbol{X}_t^*$ be the compressed time series defined in (2.2). The $\mathbf{b}(\cdot)$ and $\mathbf{c}(\cdot)$ minimizing the expected squared recover error $\mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_t^*\|_2^2]$ is given by

$$\mathbf{b}(t) = \int_0^1 \mathbf{B}(\omega) \exp(i2\pi\omega t) d\omega, \quad \mathbf{c}(t) = \int_0^1 \mathbf{C}(\omega) \exp(i2\pi\omega t) d\omega,$$

where $\mathbf{B}(\omega) = (\mathbf{v}_1(\omega), \ldots, \mathbf{v}_q(\omega))^\dagger, \mathbf{C}(\omega) = \mathbf{B}^\dagger(\omega)$, and $\{\mathbf{v}_i(\omega)\}_{1 \le i \le q}$ are the q leading eigenvectors of $\mathbf{S}(\omega)$.

It can be readily checked that when $\boldsymbol{X}_t$ is a white noise, the spectral density matrices are all the same and equal to $\mathbf{R}_0$, so the result in Theorem 2.1 coincides with the usual PCA. Therefore, the variational interpretation of PCA is a special case of Theorem 2.1. The statistical counterpart of the above time series PCA has already been investigated in Brillinger (1969) in low dimensional regime, and we will explore the high dimensional regime in the following discussion.

3 Spectral Density Matrix Estimation

In this section, we propose a computationally efficient method to spectral density estimation.

Let $\{\boldsymbol{X}_t\}_{0\le t\le T}$ be the observations of the time series. Define the sample cross-covariance matrices $\{\widehat{\mathbf{R}}_t\}_{t\in\mathbb{Z}}$ as

$$\widehat{\mathbf{R}}_t = \frac{1}{N(t,T)+1}\sum_{k=0}^{N(t,T)} \boldsymbol{X}_{(k+1)t+k}\boldsymbol{X}_{kt+k}^{\dagger}, \quad 0\le t<T, \tag{3.1}$$

where $N(t,T)=\lfloor (T-t)/(t+1)\rfloor$. When $t<0$ we set $\widehat{\mathbf{R}}_t=\widehat{\mathbf{R}}_{-t}^{\dagger}$. Now that $\widehat{\mathbf{R}}_t$ is defined for all $t\in\mathbb{Z}$, the estimator for $\widehat{\mathbf{S}}(\omega)$ will be given by

$$\widehat{\mathbf{S}}(\omega) = \sum_{t=-C\lfloor \log T\rfloor}^{C\lfloor \log T\rfloor} \widehat{\mathbf{R}}_t \exp(-i2\pi\omega t). \tag{3.2}$$

where C is a constant which will be specified in Theorem 3.3. The cutoff at $C\lfloor \log T\rfloor$ is intimately related to the α-mixing condition where we will explain in details in Assumption 2, and will be later justified in the theoretical analysis in Theorem 3.3. This completes our estimator for spectral density estimators.

We now give the theoretical justification of our estimator and explain why $O(\log T)$ cross-covariance matrices are enough. The following two definitions are instrumental in measuring quantitatively how dependent the time series is across time.

Definition 3.1. Given two σ-algebras $\mathcal{A}$ and $\mathcal{B}$, the α-mixing coefficient between $\mathcal{A}$ and $\mathcal{B}$, denoted by $\alpha(\mathcal{A},\mathcal{B})$, is defined as

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} |\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)|.$$

The α-mixing coefficient between two σ-algebras in some sense measures the dependency between the two; when the two σ-algebras are independent their α-mixing coefficient is 0. We then proceed to define the α-mixing notion for a stationary time series.

Definition 3.2. Given a stationary time series $\boldsymbol{X}_t$, the α-mixing function at lag $u\ge 0$ is defined as

$$\alpha(u) = \alpha\left(\sigma(\boldsymbol{X}_t, t\le 0), \sigma(\boldsymbol{X}_t, t\ge u)\right).$$

Note that $\alpha(u)$ can be equivalently defined as $\alpha\left(\sigma(\boldsymbol{X}_t,t\le s),\sigma(\boldsymbol{X}_t,t\ge s+u)\right)$ for an arbitrary $s\in\mathbb{Z}$, because the time series is assumed to be stationary.

With the help of the above definitions we introduce four assumptions. We first introduce the sparisty assumption which is important in the approximation of the time series. Given a q-dimensional subspace $\mathcal{U}$, we suppose $\mathbf{u}_1,\dots,\mathbf{u}_q$ is a set of the orthonormal basis of $\mathcal{U}$. We say that the space $\mathcal{U}$ is s^*-sparse if the matrix $\mathbf{U}=(\mathbf{u}_1,\dots,\mathbf{u}_q)$ has at most s^* nonzero rows, i.e., $\|\mathbf{U}\|_{2,0}\le s^*$. Elementary algebra shows that this definition is well-defined because the number of non-zero rows doesn't change if we switch to another orthonormal basis of $\mathcal{U}$. The sparsity assumption is defined as follows.

Assumption 1. For all $\omega \in [0,1)$, the q-dimensional principal subspace of $\mathbf{S}(\omega)$ is s^*-sparse and these principal subspaces share the same support. Here, the principal subspace is the projection matrix spanned by the leading q eigenvectors. Moreover, the eigengaps of $\mathbf{S}(\omega)$'s satisfy that $\inf_{\omega\in[0,1)} \lambda_q(\mathbf{S}(\omega)) - \lambda_{q+1}(\mathbf{S}(\omega)) > \delta_q > 0$ for some universal constant δ_q.

Remark 3.1. This sparsity imposed on spectral density matrices is at first glance less straightforward compared to other sparsity assumptions in literature, and we would like to give some examples in order to shed more light on this assumption. For example, consider the factor model (Fan et al., 2013)

$$\boldsymbol{X}_t = \mathbf{A}\boldsymbol{F}_t + \boldsymbol{\epsilon}_t, \tag{3.3}$$

where $\boldsymbol{X}_t$ is d-dimensional and $\boldsymbol{F}_t$ is q-dimensional, and q is much smaller than d. $\boldsymbol{\epsilon}_t$ is a white noise with distribution $\mathcal{N}(0, \mathbf{I}_d)$. For the sake of identifiability, we assume that $\mathbf{A}$ has orthonormal column vectors, i.e. $\mathbf{A}^T\mathbf{A} = \mathbf{I}_q$. We also assume that the noise is independent of the factor process, and the factor process is uncorrelated with lags larger than h. A simple calculation reveals that, for all $\omega \in [0,1)$,

$$\mathbf{S}(\omega) = \mathbf{A}\left(\sum_{u=-h}^{h} \mathbb{E}[\boldsymbol{F}_u \boldsymbol{F}_0^\dagger] e^{i2\pi\omega u}\right)\mathbf{A}^T + \mathbf{I}_d = \mathbf{A}\boldsymbol{\Gamma}(\omega)\boldsymbol{\Lambda}(\omega)\boldsymbol{\Gamma}(\omega)^\dagger\mathbf{A}^T + \mathbf{I}_d, \tag{3.4}$$

and obviously a q-dimensional leading eigenspace of $\mathbf{S}(\omega)$ is generated by the column vectors of $\mathbf{A}\boldsymbol{\Gamma}(\omega)$. We hence see that if $\mathbf{A}$ is s^*-sparse in the sense that $\|\mathbf{A}\|_{2,0} \le s^*$, then for all $\omega \in [0,1)$ the leading q eigenvectors are jointly s^*-sparse. Moreover, the sparsity pattern is the same across all ω in $[0,1)$ for $\mathbf{S}(\omega)$.

Assumption 2. There exists positive constants c_1 and $\gamma_1 \ge 1$ such that for all lag $u \ge 1$, the α-mixing coefficient satisfies

$$\alpha_{\boldsymbol{X}}(u) \le e^{-c_1 u^{\gamma_1}}.$$

Assumption 3. There exist positive constants c_2 and γ_2 such that for all $\mathbf{v} \in \mathbb{S}^{d-1}(\mathbb{C})$ and all $\lambda \ge 0$, we have

$$\mathbb{P}(|\mathbf{v}^T\boldsymbol{X}_0| \ge \lambda) \le 2e^{-c_2\lambda^{\gamma_2}}.$$

Note that $\mathbb{S}^{d-1}(\mathbb{C})$ is the unit ball of dimension d and by strong stationarity of $\boldsymbol{X}$, the above inequality holds if we replace $\boldsymbol{X}_0$ by $\boldsymbol{X}_t$, $t \in \mathbb{Z}$.

Assumption 4. Define γ via $1/\gamma = 1/\gamma_1 + 2/\gamma_2$, where γ_1 and γ_2 are given in Assumption 2 and 3. We assume that $\gamma < 1$.

Remark 3.2. We would like to give some immediate comments on these assumptions. Assumption 2 means that the α-mixing function of the time series possesses an exponential decay, which gives the time series a "short memory" flavor. Assumption 2 is also assumed in Fan et al. (2013) for the factor model in (3.3). Assumption 3 is a bound on the marginal distribution, and apparently it is more general than the sub-Gaussian or sub-exponential model, with $\gamma_2 = 2$ corresponding to the sub-Gaussian case and $\gamma_2 = 1$ corresponding to the sub-exponential case. Assumption 4 is a technical assumption allowing us to apply Bernstein inequality presented in Merlevède et al. (2011). We would like to point out that Assumption 4 is not a stringent assumption, for example when marginal distribution is sub-Gaussian, this assumption yields no restriction on γ_1. The set of stationary time series distributions satisfying Assumptions 2 - 4 will be denoted by $\mathcal{M}(c_1, c_2, \gamma_1, \gamma_2)$.

We will work on the model $\mathcal{M}(c_1, c_2, \gamma_1, \gamma_2)$ exclusively in the remainder of the section. We first present the convergence rate of $\widehat{\mathbf{S}}(\omega)$ towards $\mathbf{S}(\omega)$, which is an interesting result by its own. Define the sparse operator norm of any matrix $\mathbf{\Sigma}$ as

$$\|\mathbf{\Sigma}\|_{\text{op},s^*} = \sup\left\{\mathbf{v}^\dagger \mathbf{\Sigma} \mathbf{v} | \|\mathbf{v}\|_2 \leq 1, \|\mathbf{v}\|_0 \leq s^*\right\}.$$

Actually, besides the estimator $\widehat{\mathbf{S}}(\omega)$ in (3.2), we consider the following general estimator

$$\widehat{\mathbf{S}}_M(\omega) = \sum_{t=-M(T)}^{M(T)} \widehat{\mathbf{R}}_t \exp(-i2\pi\omega t).$$

So $\widehat{\mathbf{S}}(\omega)$ is a special case of $\widehat{\mathbf{S}}_M(\omega)$ by choosing the cutoff value $M(T) = C\lfloor \log T \rfloor$. We can show in the following theorem the advantage of choosing such $M(T)$.

Theorem 3.3. Suppose Assumptions 1 – 4 are satisfied. If $M(T) \to \infty$, $M(T)/T^{1/3} \to 0$ and $(s^* \log d)^{2/\gamma - 1} M(T) \leq T \leq d$, $\widehat{\mathbf{S}}_M(\omega)$ is consistent in the following norms

$$\begin{aligned} \sup_{\omega \in [0,1)} \|\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\|_{\text{op},s^*} &= O_P\left(\exp(-c_0 M(T)) \bigvee M(T)^{3/2} \sqrt{\frac{s^* \log d}{T}} \right), \\ \sup_{\omega \in [0,1)} \|\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\|_{\infty,\infty} &= O_P\left(\exp(-c_0 M(T)) \bigvee M(T)^{3/2} \sqrt{\frac{\log d}{T}} \right). \end{aligned} \tag{3.5}$$

where c_0 is some constant only depending on $c_1, c_2, \gamma_1, \gamma_2$.

Remark 3.3. A straightforward conclusion from this theorem is that

$$\begin{aligned} \sup_{\omega \in [0,1)} \|\widehat{\mathbf{S}}(\omega) - \mathbf{S}(\omega)\|_{\text{op},s^*} &= O_P\big((\log T)^{3/2} \sqrt{s^* \log d / T}\big) \\ \sup_{\omega \in [0,1)} \|\widehat{\mathbf{S}}(\omega) - \mathbf{S}(\omega)\|_{\infty,\infty} &= O_P\big((\log T)^{3/2} \sqrt{\log d / T}\big) \end{aligned}$$

by setting $M(T) = \lfloor (2/c_0) \log T \rfloor$ in the theorem. Moreover, our estimator is computationally more efficient than the Blackman-Tukey estimator

$$\widetilde{\mathbf{S}}(\omega) = \sum_{t=-T}^{T} w_t \cdot \widehat{\mathbf{R}}_t \exp(-i2\pi\omega t),$$

in the sense that Blackman-Tukey estimator demands $O(T)$ cross-covariance matrices while our estimator only requires $O(\log T)$ cross-covariance matrices.

We shed more lights on the intuition of the statistical rate in (3.5). The statistical error of $\widehat{\mathbf{S}}(\omega)$ can be divided into two parts: the resolution error and the estimation error. The resolution error is caused by not using the whole spectrum in Fourier transform, and the estimation error is from cross-covariance matrix estimators. We can also interpret the resolution error as the bias and estimation error as the variance. Similar to the bias-variance trade-off, when the cutoff value is small, the resolution error tends to be large and the estimation error tends to be small. When the cutoff value is large, the resolution error tends to be small while the estimation error is large. Our theorem provides the scaling condition on the cutoff value when the estimator is consistent.

4 Dynamic Component Analysis

After estimating the spectral density matrix derived in the previous section, we conduct the Dynamic Component Analysis for subspace estimation. The algorithm we propose for DCA is a combination of CSOAP(Chained Sparse Orthogonal Pursuit) and cosine interpolation. In this section, we first describe our Dynamic Component Analysis (DCA) algorithm and then we show the theoretical results.

4.1 CSOAP Algorithm and Cosine Interpolation

For all frequencies ω, an estimator $\widehat{\mathbf{S}}(\omega)$ has been derived in the previous subsection. We then discretized the frequency space $[0, 1)$ into N equally-spaced points $\{0, 1/N, \cdots, (N-1)/N\}$, and we aim at calculating the principal subspaces for these N Hermitian matrices. As noted in Wang et al. (2014), an initialization with Alternate Direction Method of Multipliers (ADMM) and followed by a Sparse Orthogonal Pursuit (SOAP) can achieve an optimal statistical convergence rate to compute the principal subspace. This algorithm combined with the intuition that when N is large, adjacent $\widehat{\mathbf{S}}(k/N)$ are close to each other leads to the following CSOAP (Chained Sparse Orthogonal Pursuit) algorithm: An ADMM initialization is used for $\widehat{\mathbf{S}}(0)$, followed by a SOAP. The resulting principal subspace can then be plugged in as initialization for the SOAP for $\widehat{\mathbf{S}}(1/N)$, and this procedure propagates through $2/N, \cdots, (N-1)/N$.

We begin with the ADMM initialization step. This algorithm has been presented in detail in Wang et al. (2014). The central idea lies in the convex relaxation technique: if we define the complex q-dimensional fantope $\mathcal{F}^q \subset \mathbb{C}^{d\times d}$ as

$$\mathcal{F}^q = \{\mathbf{A} \text{ Hermitian} | \operatorname{tr}(\mathbf{A}) = q,\ 0 \preceq \mathbf{A} \preceq \mathbf{I}\},$$

then it can be proved that it is actually the convex hull of the set $\left\{\mathbf{V}\mathbf{V}^\dagger | \mathbf{V}^\dagger\mathbf{V} = \mathbf{I}\right\}$ where $\mathbf{V} \in \mathbb{C}^{d\times q}$. With this convex relaxation, the penalized optimization problem associated with the q-dimensional principal subspace of a Hermitian matrix $\mathbf{S}$ can be formulated as (see Vu et al. (2013))

$$\begin{aligned} &\text{Minimize} && -\langle \mathbf{S}, \mathbf{\Pi}\rangle + \rho\|\mathbf{\Pi}\|_{1,1} \\ &\text{Subject to} && \mathbf{\Pi} \in \mathcal{F}^q. \end{aligned} \tag{4.1}$$

We want to numerically solve this optimization problem for $\mathbf{S} = \widehat{\mathbf{S}}(0)$. As is discovered and analyzed in Vu et al. (2013) and Wang et al. (2014), the ADMM is especially suited for this kind of optimization problem, and it proceeds as follows. We first define the augmented Lagrangian associated to the problem as

$$L_{\rho,\beta}(\mathbf{\Pi}, \mathbf{\Phi}, \mathbf{\Theta}) = -\langle \mathbf{S}, \mathbf{\Pi}\rangle + \rho\|\mathbf{\Phi}\|_{1,1} - \langle \mathbf{\Theta}, \mathbf{\Pi} - \mathbf{\Phi}\rangle + \frac{\beta}{2}\|\mathbf{\Pi} - \mathbf{\Phi}\|_F^2,$$

where $\mathbf{S}$, $\mathbf{\Pi}$, $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are all Hermitian matrices, so the augmented Lagrangian always takes real values. The ADMM algorithm iteratively optimizes L with respect to $\mathbf{\Pi}$, then with respect to $\mathbf{\Phi}$, and updates the Lagrange multiplier $\mathbf{\Theta}$. As we can readily observe, the element-wise soft-thresholding algorithm is subtly different compared to its counterpart when we are dealing with real matrices. This is due to the definition of the complex $\|\cdot\|_{1,1}$-norm, which creates a coupling term between the matrix's real and imaginary parts. A detailed description for the ADMM algorithm can be found in Appendix A.

After the ADMM algorithm outputs $\overline{\mathbf{\Pi}}^{(R)}$, we compute its top q leading eigenvectors and denote it by $\mathbf{U}^{\text{init}} \in \mathbb{C}^{q\times d}$. It will be used later for the initialization to compute $\widehat{\mathbf{S}}(0)$. In order to do that, we first give the SOAP algorithm, which is the building block for computing the principal subspaces. The value of iteration number $\widetilde{R}$ and the truncation parameter $\widehat{s}$ is specified in Section

Algorithm 1 SOAP Algorithm

Function: $\widehat{\mathbf{U}} \leftarrow \texttt{SOAP}(\mathbf{S}, \widehat{\mathbf{U}}^{(0)})$

Input: $d \times d$ Hermitian matrix $\mathbf{S}$, $d \times q$ orthonormal columns $\widehat{\mathbf{U}}^{(0)}$, truncation parameter $\widehat{s}$

1: **For** $t = 0$ to $\widetilde{R} - 1$ **do**
2: $\quad \mathbf{V}^{(t+1)} \leftarrow \mathbf{S} \cdot \widehat{\mathbf{U}}^{(t)}$
3: $\quad \widetilde{\mathbf{V}}^{(t+1)}, \mathbf{R}_1^{(t+1)} \leftarrow \texttt{Thin\_QR}(\mathbf{V}^{(t+1)})$
4: $\quad \widetilde{\mathbf{U}}^{(t+1)} \leftarrow \texttt{Truncate}(\widetilde{\mathbf{V}}^{(t+1)}, \widehat{s})$
5: $\quad \widehat{\mathbf{U}}^{(t+1)}, \mathbf{R}_2^{(t+1)} \leftarrow \texttt{Thin\_QR}(\widetilde{\mathbf{U}}^{(t+1)})$
6: **End For**

Output: $\widehat{\mathbf{U}} \leftarrow \widehat{\mathbf{U}}^{(\widetilde{R})}$

4.2. $\texttt{Truncate}(\Sigma, s)$ is defined as follows. We sort the ℓ_2 norm of each row of Σ in the descending order. The corresponding first s rows are left unchanged and all the other rows are truncated to 0. The result is defined to be $\texttt{Truncate}(\Sigma, s)$. The definition of Thin QR in line 3 and 5 can be found in Golub and Van Loan (2012).

After giving the SOAP algorithm, we are ready to give the whole algorithm. First, the interval $[0, 1)$ will be discretized into grids $\{0, 1/N, \cdots, (N-1)/N\}$, and the SOAP algorithm will be run on $\widehat{\mathbf{S}}(0), \widehat{\mathbf{S}}(1/N), \cdots, \widehat{\mathbf{S}}((N-1)/N)$. For the spectral density matrix on each frequency, it is unnecessary to start from the ADMM algorithm. According to the smoothness of $S(\omega)$ with respect to ω, the eigenvectors of $\widehat{\mathbf{S}}(k/N)$ is a good initialization of the SOAP algorithm for $\widehat{\mathbf{S}}((k+1)/N)$. Therefore, we can chain up the computation from $\widehat{\mathbf{S}}(0)$ to $\widehat{\mathbf{S}}((N-1)/N)$. A detailed description can be found in Algorithm 2.

Algorithm 2 CSOAP Algorithm

Function: $\widehat{\mathbf{U}} \leftarrow \texttt{CSOAP}(\{\widehat{\mathbf{S}}(k/N)\}_{k=0}^{N-1})$

1: $\overline{\mathbf{\Pi}} \leftarrow \texttt{ADMM}(\widehat{\mathbf{S}}(0))$
2: Set the columns of $\mathbf{U}^{\text{init}}$ to be the top q leading eigenvectors of $\overline{\mathbf{\Pi}}$
3: $\widetilde{\mathbf{U}}^{(0)} \leftarrow \texttt{Truncate}(\mathbf{U}^{\text{init}}, \widehat{s}), \mathbf{U}^{\text{init}} \leftarrow \texttt{Thin\_QR}(\widetilde{\mathbf{U}}^{(0)})$
4: $\widehat{\mathbf{U}}(0) = \texttt{SOAP}(\widehat{\mathbf{S}}(0), \mathbf{U}^{\text{init}})$
5: **For** $k = 1$ to $N - 1$ **do**
6: $\quad \widehat{\mathbf{U}}(k/N) = \texttt{SOAP}(\widehat{\mathbf{S}}(k/N), \widehat{\mathbf{U}}((k-1)/N))$
7: **End For**

Output: $\{\widehat{\mathbf{U}}(k/N)\}_{k=0}^{N-1}$

The CSOAP algorithm yields $\{\widehat{\mathbf{U}}(k/N)\}_{k=0}^{N-1}$. In order to get the eigenvectors on the full spectrum of $\omega \in [0, 1)$, one possible solution is to add to the resolution by increasing N. However, this requires to run Algorithm 2 for more times and the computation cost is increased. There is also a theoretical reason why we cannot apply this naïve method. This is because the output eigenvectors $\widehat{\mathbf{U}}(\omega)$ from Algorithm 2 cannot be guaranteed to be smooth with respect to ω, which will affect

the accuracy of inverse Fourier transform and the convolution when we apply the filtering step in (2.2). To overcome this challenge, we apply the interpolation between $\widehat{\mathbf{U}}(k/N)$ and $\widehat{\mathbf{U}}((k+1)/N)$. We interpolate $\widehat{\mathbf{U}}(\omega)$ for $\omega \in [k/N, (k+1)/N)$ by using cosine function,

$$\widehat{\mathbf{U}}(\omega) = (\widehat{\mathbf{U}}(k/N)/2 - \widehat{\mathbf{U}}(k+1/N)/2)\cos(\pi N\omega - k/N) + \widehat{\mathbf{U}}(k/N)/2 + \widehat{\mathbf{U}}(k+1/N)/2. \quad (4.2)$$

Here we do not use the spline interpolation or other widely used methods because they violate the convergence conditions for the Fourier transform on $\widehat{\mathbf{U}}(\omega)$. To apply the Fourier transform, $\widehat{\mathbf{U}}(\omega)$ cannot fluctuate too much inside each interval $[k/N, (k+1)/N)$. By applying the cosine transform, for each $\omega \in [k/N, (k+1)/N)$, $\widehat{\mathbf{U}}_{st}(\omega)$ is bounded between the values of $\widehat{\mathbf{U}}_{st}(k/N)$ and $\widehat{\mathbf{U}}_{st}((k+1)/N)$ which guarantees that $\widehat{\mathbf{U}}(\omega)$ has nice smooth property.

The estimators of filters $\mathbf{b}$ and $\mathbf{c}$ used in the compression step in (2.2) are given by

$$\widehat{\mathbf{b}}(t) = \int_0^1 \widehat{\mathbf{U}}^\dagger(\omega)\exp(i2\pi\omega t)d\omega, \quad \widehat{\mathbf{c}}(t) = \int_0^1 \widehat{\mathbf{U}}(\omega)\exp(i2\pi\omega t)d\omega.$$

In practice, we cannot add infinite terms for the convolution step in (2.2). Therefore, the final estimator $\widetilde{\mathbf{b}}$ and $\widetilde{\mathbf{c}}$ is defined as

$$\widetilde{\mathbf{b}}(t) = \begin{cases} \widehat{\mathbf{b}}(t) & |t| \le N \\ 0 & |t| > N \end{cases} \text{ and } \widetilde{\mathbf{c}}(t) = \begin{cases} \widehat{\mathbf{c}}(t) & |t| \le N \\ 0 & |t| > N. \end{cases} \quad (4.3)$$

Here the truncation parameter N is the same as the number of frequency grids used in Algorithm 2. This is due to the duality between the Fourier transform and inverse Fourier transform. Therefore, the final compressed time series becomes

$$\boldsymbol{X}_t^{\widetilde{\mathbf{b}}} = \sum_{u\in\mathbb{Z}} \widetilde{\mathbf{c}}(t-u) \sum_{s\in\mathbb{Z}} \widetilde{\mathbf{b}}(u-s)\boldsymbol{X}_s. \quad (4.4)$$

Notice that the summation above is essentially a finite summation and we will show how to choose the truncation parameter N in the next section.

4.2 Theoretical Results

We proceed to present the main convergence rate result for the whole DCA algorithm. First of all, we list the following parameters required in order to implement the algorithm.

- ρ, the regularization parameter in ADMM algorithm.
- β, the penalization parameter in ADMM algorithm.
- $\widehat{s}$, the truncation parameter in SOAP algorithm.
- R, the number of steps taken in ADMM algorithm.
- $\widetilde{R}$, the number of steps taken in SOAP algorithm.
- N, the number of discretization points on frequency interval $[0, 1)$.

Our theoretical analysis gives a concrete guidance on how to choose the parameters above. We list the requirements in the following assumption. The constant C in the assumption is generic and independent to d, T and s.

Assumption 5 (Tuning parameters)**.** The parameters listed above need to follow the following scaling conditions.

- (**ADMM parameters**) The two parameters in ADMM algorithm satisfies

$$\rho = C(\log T)^{3/2}\sqrt{\frac{\log d}{T}} \quad \text{and} \quad \beta = \rho d/\sqrt{q}.$$

- (**Truncation parameter in SOAP**) Define the parameter

$$\eta = \max_{\omega\in[0,1)} \frac{3\lambda_{q+1}(\mathbf{S}(\omega)) + \lambda_q(\mathbf{S}(\omega))}{\lambda_{q+1}(\mathbf{S}(\omega)) + 3\lambda_q(\mathbf{S}(\omega))} < 1,$$

  The truncation parameter in SOAP algorithm satisfies $\widehat{s} = C\max\left\{4q/(\eta^{-1/2}-1)^{-2}, 1\right\} s^*$.

- (**Numbers of iterations**) Define $A = \min\{\sqrt{2\eta}/4, \sqrt{q\eta(1-\eta^{1/2})/2}\}$. The number of steps taken in ADMM algorithm and each of the SOAP algorithm satisfy

$$R \geq \left(\frac{qd^2(\log T)^3 \log d}{T}\right)^{1/2} \cdot \left(A - C(\log T)^{3/2}\sqrt{\frac{\log d}{T}}\right)^{-1} \text{ and}$$
$$\widetilde{R} \geq 4(\log(1/\eta))^{-1} \log\left(C(\eta/8)^{1/2}(\log T)^{3/2}\sqrt{\frac{s^* \log d}{T}}\right).$$

- (**Number of discretization**) The number of discretization points on $[0,1)$ satisfies

$$N \geq C(\log T)^{-3/2}\left(\frac{s^* \log d}{T}\right)^{-1/2}.$$

As described in the previous section, the CSOAP algorithm consists of one single ADMM in order to initiate the whole process and a sequence of SOAP steps. Denote $\mathbf{U}^*(\omega)$ as the matrix assembling the q-leading eigenvectors of $\mathbf{S}(\omega)$ and $\mathcal{U}^*(\omega)$ as the eigenspace spanned by $\mathbf{U}^*(\omega)$. Let $\mathbf{\Pi}_0^*(\omega) = \mathbf{U}^*(\omega)(\mathbf{U}^*(\omega))^\dagger$ and recall that $\overline{\mathbf{\Pi}}^{(R)}(\omega)$ is the output of ADMM algorithm after R steps. The following theorem analyze the convergence property of ADMM algorithm under the present setting.

Theorem 4.1. Assume that the time series are generated from $\mathcal{M}(c_1, c_2, \gamma_1, \gamma_2)$ and satisfy Assumption 1. Denote $\delta_{q0} = \lambda_q(\mathbf{S}(0)) - \lambda_{q+1}(\mathbf{S}(0)) > 0$. Under Assumption 5 and the scaling in Theorem 3.3, we have

$$\sup_{0\leq k\leq N-1} \|\mathbf{\Pi}_0^*(k/N) - \overline{\mathbf{\Pi}}^{(R)}(k/N)\|_{\mathrm{F}} = O_P\left(\frac{1}{\delta_{q0}}(\log T)^{3/2}\sqrt{\frac{\log d}{T}}\right). \tag{4.5}$$

We then proceed to the convergence analysis of SOAP algorithm. The space spanned by the eigenvectors $\widehat{\mathbf{U}}(\omega)$ in (4.2) is denoted as $\widehat{\mathcal{U}}(\omega)$. The projection matrix for $\widehat{\mathcal{U}}(\omega)$ is $\widehat{\mathbf{\Pi}}(\omega) = \widehat{\mathbf{U}}(\omega)\widehat{\mathbf{U}}(\omega)^\dagger$. We define the distance between $\widehat{\mathbf{U}}(\omega)$ and $\mathbf{U}^*(\omega)$ as

$$D(\widehat{\mathcal{U}}(\omega), \mathcal{U}^*(\omega)) = \|\widehat{\mathbf{\Pi}}(\omega) - \mathbf{\Pi}_0^*(\omega)\|_{\mathrm{F}}$$

Theorem 4.2. Let $\{\boldsymbol{X}_t\}_{0\le t\le T}$ be a time series generated from the model $\mathcal{M}(c_1, c_2, \gamma_1, \gamma_2)$. Under the same conditions as Theorem 4.1, we have

$$\sup_{\omega\in[0,1)} D(\widehat{\mathcal{U}}(\omega), \mathcal{U}^*(\omega)) = O_P\left((\log T)^{3/2}\sqrt{\frac{qs^*(q\vee\log d)}{T}}\right). \tag{4.6}$$

We then present the convergence property of the whole CSOAP algorithm. Compared with the usual rates derived in the literature for usual sparse PCA (Vu et al., 2013; Wang et al., 2014), an extra term of the order $(\log T)^{3/2}$ appears due to the fact that in order to estimate spectral density matrices, we have to sum up a certain number of error terms corresponding to estimated cross-covariance matrices.

Theorem 4.3. Let $\{\boldsymbol{X}_t\}_{0\le t\le T}$ be a time series generated from the model $\mathcal{M}(c_1, c_2, \gamma_1, \gamma_2)$ satisfying Assumption 1. Assume the eigengaps of $\mathbf{S}(\omega)$ satisfies $\inf_{\omega\in[0,1)} \lambda_i(\mathbf{S}(\omega)) - \lambda_{i+1}(\mathbf{S}(\omega)) > \delta > 0$ for all $i = 1, ..., q$. Let $\{\widehat{\mathbf{b}}(t)\}_{t\in\mathbb{Z}}$ be the estimators we obtain using the DCA algorithm that satisfying Assumption 5. If $c(s^*\log d)^{2/\gamma-1}\log T \le T \le c'd$ where c and c' are two constants, then we have

$$\left|\mathbb{E}\left[\|\boldsymbol{X}_t - \boldsymbol{X}_t^{\mathbf{b}}\|^2\right] - \mathbb{E}\left[\|\boldsymbol{X}_t - \boldsymbol{X}_t^{\widetilde{\mathbf{b}}}\|^2\right]\right| = O_P\left((\log T)^{3/2}q^{3/2}s^*\sqrt{\frac{\log d}{T}}\right). \tag{4.7}$$

This theorem guarantees that with high probability, by using the estimators we obtain from the DCA algorithm, the compression performance achieved will be very close to the optimal performance, and an upper bound is provided by the right-hand side of (4.7).

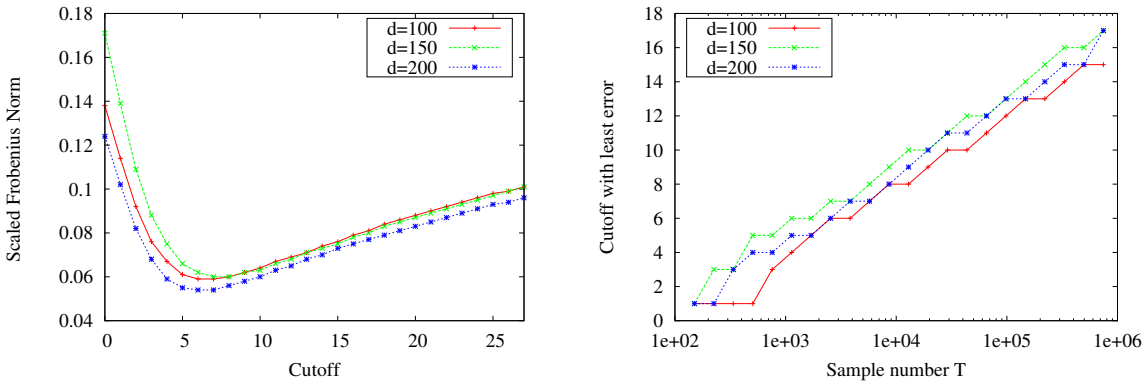
5 Simulation

In this section, we provide numerical results to verify the statistical accuracy of our dynamic component analysis (DCA). The whole procedure can be decomposed into three steps. In the first step, we estimate the spectral density matrices by applying Fourier transform on the cross-covariance matrices (3.2). In the second step, we apply our CSOAP algorithm on the estimated spectral density matrices to get their leading eigenvectors. In the last step, we use cosine interpolation and Fourier transform to transform the eigenvectors of estimated spectral density matrix back to the estimator $\widetilde{\mathbf{b}}$ and $\widetilde{\mathbf{c}}$ in (4.3). In the following, we evaluate the accuracy of these three steps for both the synthetic data and brain imaging dataset.

5.1 Synthetic Data

The model we consider is the factor model $\boldsymbol{X}_t = \mathbf{A}\boldsymbol{F}_t + \boldsymbol{\epsilon}_t$ where $\boldsymbol{X}_t$ is d-dimensional, $\boldsymbol{F}_t$ is q-dimensional and the white noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I}_d)$. $\mathbf{A}$ is a s^*-sparse $d\times q$ matrix where only the first s^* rows are non-zero and every entry in these s^* rows is independently generated from $\mathcal{N}(0,1)$. In all of the following synthetic simulations, we set $s^* = 10$ and $q = 5$. We let $\boldsymbol{F}_t = \mathbf{D}\boldsymbol{F}_{t-1} + \boldsymbol{\theta}_{t-1}$ where $\mathbf{D}$ is a $q\times q$ matrix and $\boldsymbol{\theta}_{t-1} \sim \mathcal{N}(0, \mathbf{I}_q)$. Note that this is the autoregressive model with order 1 (Johansen, 1995). We set $\mathbf{D} = 0.8\mathbf{I}$. Since the time series is stationary, the covariance matrix of $\boldsymbol{F}_t$ can be evaluated by solving the linear system $\mathrm{Cov}(\boldsymbol{F}_t) = \mathrm{Cov}(\boldsymbol{D}\boldsymbol{F}_t + \boldsymbol{\theta}_t)$ for all $t \ge 0$. We generate $T+1$ samples of the factor model $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_T$ with the initial vector $\boldsymbol{F}_0 \sim \mathcal{N}(0, \mathrm{Cov}(\boldsymbol{F}_0))$.

Spectral Density Matrices Estimation: In order to estimate the spectral density matrices $\mathbf{S}(\omega) = \sum_{t=-\infty}^{\infty}\mathbf{R}_t\exp(-i2\pi\omega t)$, we apply Fourier transform on the sample cross-covariance



(a) Statistical error v.s. cutoff (b) Statistical error minimizer v.s sample size

Figure 1: Spectral density matrices estimation

matrices $\widehat{\mathbf{R}}_t$. However, the definition of spectral density matrix requires us to add infinite cross-covariance matrices terms, which is unrealistic in real applications. We prove in Theorem 3.3 that cutting off at $t = \lfloor \log T \rfloor$ suffices it to estimate the spectral density matrices assuming that we have T samples. In the simulation, we set $T = 10^5$ and $d = 100, 150, 200$. To calculate the true spectral density matrices, we first obtain the cross-covariance matrix $\mathbf{R}_t = \mathbb{E}[\boldsymbol{F}_t \boldsymbol{F}_0^\dagger] = \mathbf{D}^t \operatorname{Cov}(\boldsymbol{F}_0)$. Then, we plug it into (3.4) to get the spectral density matrices. In Figure 1(a), we illustrate how the estimation error of spectral density matrix estimator varies with the cutoff value. We discretize the frequency space into 100 equally-spaced points and average the estimation errors $d^{-1}\|\widehat{\mathbf{S}}(\omega) - \mathbf{S}(\omega)\|_{\mathrm{F}}$ on these 100 equally-spaced ω's. We use this scaled value to represent the estimation error.

As we can see, the estimation error first decreases and starts to increase after some point. The reason is because getting a more accurate Fourier transform requires us to increase the cutoff value. However, this involves more sample cross-covariance estimators in (3.2) which brings additional estimation errors. It means that there is a trade-off between the error incurred by calculating Fourier transform and the error incurred by the estimation error. We demonstrate how the optimal cutoff value achieving the minimal estimation error changes with the sample size T in Figure 1(b). Note that the axis of sample size is logarithmic and the optimal cutoff value almost linearly increase. From the figure, it is clear that the cutoff value with least error has a logarithmic relationship with the sample size, showing that our setting of cutoff value is reasonable.

CSOAP Statistical Rate of Convergence: The core of CSOAP is to estimate the leading eigenvectors of $\widehat{\mathbf{S}}(\omega)$. One way to evaluate the accuracy of the leading eigenvectors is to compute the distance between the subspace spanned by them and the subspace spanned by the true leading eigenvectors (Wang et al., 2014). The distance here is the Frobenius norm between the projection matrices of the corresponding subspaces. Note that the leading eigenvectors for different ω's are different. Therefore, we discretize the frequency space $[0, 1)$ into 100 equally-spaced points and add together the distance at these points. We plot the results of three tests to show the relationship between the convergence rate and the statistical error. All our results are compared to the PCA algorithm, which simply calculate the eigenvectors of $\mathbf{S}(\omega)$ instead of using CSOAP algorithm.

We vary the size of the sample size T and use $\lfloor \log T \rfloor$ as the cutoff value. In Figure 2, we illustrate the relationship between the sample size and the subspace distance. We can see that our CSOAP algorithm performs better than PCA algorithm consistently. When the number of samples

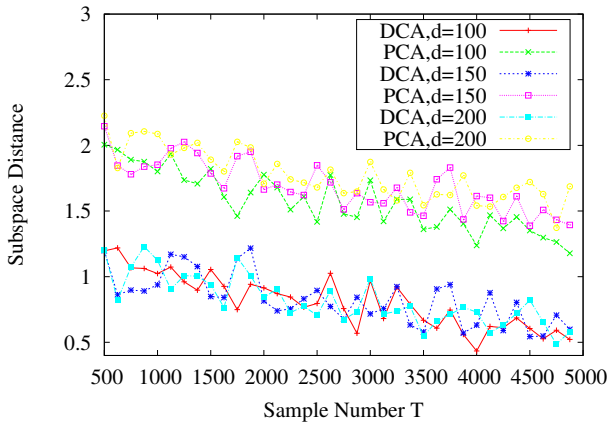


Figure 2: CSOAP Statistical Rate of Convergence

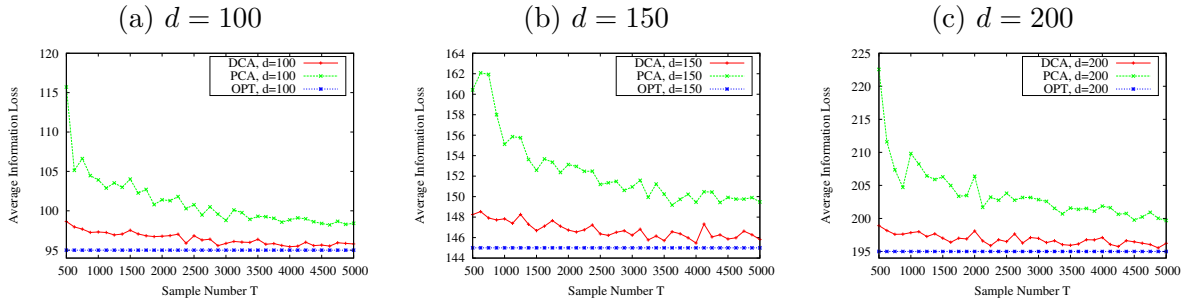


Figure 3: Prediction error in different settings

increases, our estimator performs better, which is in accordance with our convergence rate.

Prediction Accuracy: In this section, we show the prediction accuracy of our DCA algorithm. We conduct the simulation under $d = 100, 150, 200$ respectively. In each setting, we vary the number of samples from 500 to 5000. The samples are generated from the factor model illustrated before. We use the expectation of the information loss in (B.1) between the original sample and the compressed sample to evaluate the accuracy of the prediction. We compare our algorithm with the PCA algorithm and the optimal compression. To get the optimal compression, observe that we can compute the true spectral density matrices and get the compression by Theorem 2.1.

The results are shown in Figure 3. We can see from the figure that our algorithm works better than PCA algorithm. Also, when the dimension increases, the information loss decreases, which is in accordance with our convergence rate. Our results also show that the estimation error is close to the optimal value even if the sample size is relatively small.

5.2 Real Data

We consider the ADHD-200 Data (Biswal et al., 2010) as the real application of the time series PCA. The dataset consists of rs-fMRI images of 973 subjects. Some subjects are diagnosed with ADHD type 1,2 and 3 and some are healthy controls. We use the images of 478 healthy subjects and their age information. Note that we only use the data on 264 seed regions for analysis (Qiu et al., 2015). That is to say, every image is a 264×1 vector. Also, we take the median of the images for each subject to get the single image vector for that subject. Assume the images for healthy

subjects at the same age are similar, we can think of the images as a time series with regard to the age. Then, we apply our DCA algorithm to compress the images. We divide randomly the images into the training set and the testing set. The training set consists of 340 samples while the testing set consists of 138 samples.

Since the ages ranging from 7.1 years old to 21.82 are continuous, we first need to discretize the lags. In specific, we consider $T = 369$ discrete lags by transforming the age interval $[7.1, 21.82]$ to $[1, 369]$ in the sense that the t-th lag corresponds to the age $7.1 + 0.04(t-1)$. We apply the weighted average of the samples with different ages, where the weights are determined by Gaussian kernel. Therefore, our discretized time series $\boldsymbol{X}'_t$ for any $t = 1, \ldots, 369$ is

$$\boldsymbol{X}'_t = \frac{\sum_u K_\sigma(7.1 + 0.04(t-1), u)\boldsymbol{X}_u}{\sum_u K_\sigma(7.1 + 0.04(t-1), u)} \tag{5.1}$$

where $\boldsymbol{X}_u$ is the original sample at age u and K_σ is the Gaussian kernel where $K_\sigma(x, y) = \exp\big(-\|x-y\|^2/(2\sigma^2)\big)$. Here we choose $\sigma = 0.01$. By applying DCA algorithm to our new samples $\{\boldsymbol{X}'_t\}_{t=1}^{369}$, we plot the heatmaps of the spectral density matrices $\mathbf{S}(\omega)$ for $\omega = 0.2, 0.5$ and 0.9 in Figure 4. As the entries $\widehat{\mathbf{S}}(\omega)$ are complex numbers, we illustrate the absolute values of these entries. We can see from the figure that the values of a small proportion of rows are significantly larger than

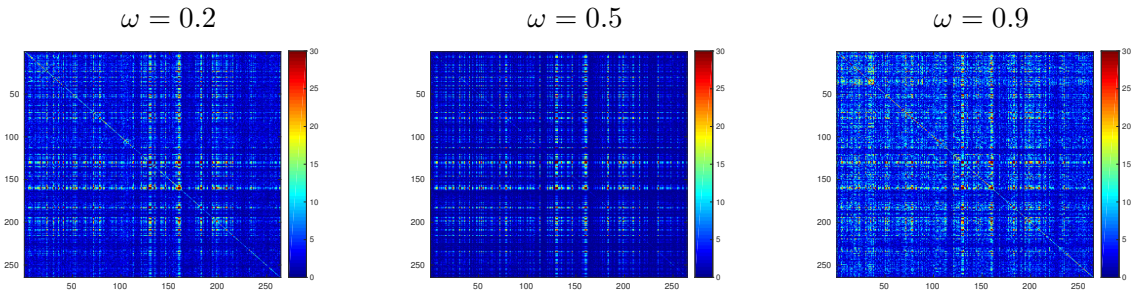


Figure 4: The spectral density matrix when $\omega = 0.2, 0.5$ and 0.9

the values of other rows. It is expected that the corresponding positions of the leading eigenvectors are also nonzero. We use CSOAP to estimate the leading eigenvectors of spectral density matrices $\mathbf{S}(0.2), \mathbf{S}(0.5)$ and $\mathbf{S}(0.9)$, which are all nonzero at the 132nd and 161st components.

Then we show that our compression is useful in predicting the image data given the subject's age. We use the previous setting to evaluate $\widehat{\mathbf{b}}(t)$ and $\widehat{\mathbf{c}}(t)$. Then we compress the converted samples using (2.2). Denote the data after compression to be $\boldsymbol{X}^*_t$. As we assume that the samples of the same age are similar, we apply the weighted average with Gaussian kernel for prediction. Given a test sample's age t_0, we predict the data $\boldsymbol{X}_{t_0}$ using two ways. One way is to use the compressed sample, in which

$$\widehat{\boldsymbol{X}}_{t_0} = \frac{\sum_{i=1}^{369} K_\sigma(7.1 + 0.04(i-1), t^*)\boldsymbol{X}^*_i}{\sum_{i=1}^{369} K_\sigma(7.1 + 0.04(i-1), t^*)} \tag{5.2}$$

Another way for prediction is to use the raw data, in which we simply replace the compressed time series $\{\boldsymbol{X}^*_i\}_{i=1}^{369}$ in (5.2) by the non-compressed data $\{\boldsymbol{X}'_i\}_{i=1}^{369}$ defined in (5.1). We then calculate the ℓ_2-norm of the difference between the value predicted and the true value dividing the dimension 264. The results are shown in Figure 5. Each point at age t represents the median of the error of the test examples whose age is in the interval $[t, t + 0.5)$. We can see from the figure that using

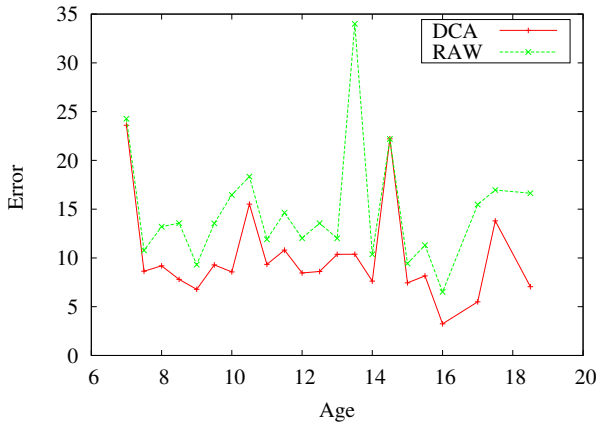


Figure 5: Prediction using compressed data and the raw data

Algorithm 3 ADMM Algorithm

Function: $\overline{\mathbf{\Pi}} \leftarrow \texttt{ADMM}(\widehat{\mathbf{S}}(0))$
Input: $d \times d$ Hermitian matrix $\widehat{\mathbf{S}}(0)$
Parameters: regularization parameter ρ, penalization parameter β

1: **For** $t = 0$ to R **do**
2: $\quad \mathbf{\Pi}^{(t+1)} \leftarrow \arg\min \left\{ L_{\rho,\beta}(\mathbf{\Pi}, \mathbf{\Phi}^{(t)}, \mathbf{\Theta}^{(t)}) | \mathbf{\Pi} \in \mathcal{F}^q \right\}$
3: $\quad \mathbf{\Phi}^{(t+1)} \leftarrow \arg\min \left\{ L_{\rho,\beta}(\mathbf{\Pi}^{(t+1)}, \mathbf{\Phi}, \mathbf{\Theta}^{(t)}) | \mathbf{\Phi} \in \mathbb{H} \right\}$
4: $\quad \mathbf{\Theta}^{(t+1)} \leftarrow \mathbf{\Theta}^{(t)} - \beta(\mathbf{\Pi}^{(t+1)} - \mathbf{\Phi}^{(t+1)})$
5: **End For**

Output: $\overline{\mathbf{\Pi}}^{(R)} = \frac{1}{R} \sum_{t=1}^{R} \mathbf{\Pi}^{(t)}$

compressed data gives us better results than using the raw data. It shows that our compression scheme is useful in extracting important information of the data.

A Algorithms for ADMM Initialization Step

In this section, we describe the detailed steps for the initialization steps for SOAP algorithm. The ADMM algorithm essentially solves the convex relaxation problems in (4.1).

The algorithm for the first step corresponding to the minimization with respect to $\mathbf{\Pi}$ is given in Algorithm 4. This step projects the matrices to the Fantope cone and it has a closed form solution through singular value decomposition.

Algorithm 4 Fantope Projection

Function: $\mathbf{\Pi}^{(t+1)} \leftarrow \texttt{Projection}_{\beta}(\mathbf{\Phi}^{(t)}, \mathbf{\Theta}^{(t)}, \mathbf{S})$
Spectral Decomposition: $\mathbf{Q}\Lambda^{(t)}\mathbf{Q}^{\dagger} \leftarrow \mathbf{\Phi}^{(t)} + \mathbf{\Theta}^{(t)}/\beta + \mathbf{S}/\beta$
Quadratic Programming: $\mathbf{v}' \leftarrow \arg\min \left\{ \|\mathbf{v} - \text{diag}(\Lambda^{(t)})\|_2^2 | \mathbf{v} \in \mathbb{R}^d, \ \sum \mathbf{v}_j = q, \ 0 \le \mathbf{v}_j \le 1 \right\}$
Output: $\mathbf{\Pi}^{(t+1)} = \mathbf{Q}\text{diag}(\mathbf{v}')\mathbf{Q}^{\dagger}$

The step corresponding to the minimization with respect to $\mathbf{\Phi}$ is given by Algorithm 5. This step is a soft-thresholding for the matrices.

Algorithm 5 Soft-thresholding

Function: $\mathbf{\Phi}^{(t+1)} \leftarrow \texttt{Soft\_Thresholding}_{\rho,\beta}(\mathbf{\Pi}^{(t+1)}, \mathbf{\Theta}^{(t)})$

Output: $\mathbf{\Phi}_{jk}^{(t+1)} = (\mathbf{\Pi}_{jk}^{(t+1)} - 1/\beta\mathbf{\Theta}_{jk}^{(t)})/|\mathbf{\Pi}_{jk}^{(t+1)} - 1/\beta\mathbf{\Theta}_{jk}^{(t)}| \times \max\left\{|(\mathbf{\Pi}_{jk}^{(t+1)} - 1/\beta\mathbf{\Theta}_{jk}^{(t)})| - \rho/\beta, 0\right\}$

B Proof for Theorem 2.1

To make this paper self-contained, we provide a brief proof of Theorem 2.1 and the detailed proof can be found in Brillinger (1969). Define the Fourier transform associated with $\mathbf{b}(\cdot)$ and $\mathbf{c}(\cdot)$ respectively as

$$\mathbf{B}(\omega) = \sum_{t\in\mathbb{Z}} e^{-i2\pi\omega t}\mathbf{b}(t), \quad \mathbf{C}(\omega) = \sum_{t\in\mathbb{Z}} e^{-i2\pi\omega t}\mathbf{c}(t),$$

then the squared recovery error $\mathbb{E}[\|\boldsymbol{X}_t - \boldsymbol{X}_t^*\|_2^2]$ as

$$\mathbb{E}\left[\mathrm{tr}((\boldsymbol{X}_t - \boldsymbol{X}_t^*)(\boldsymbol{X}_t - \boldsymbol{X}_t^*)^\dagger)\right] = \int_0^1 \mathrm{tr}\left((\mathbf{I} - \mathbf{C}(\omega)\mathbf{B}(\omega))\mathbf{S}(\omega)(\mathbf{I} - \mathbf{C}(\omega)\mathbf{B}(\omega))^\dagger\right) d\omega \tag{B.1}$$

$$= \int_0^1 \|\mathbf{S}^{1/2}(\omega) - \mathbf{C}(\omega)\mathbf{B}(\omega)\mathbf{S}^{1/2}(\omega)\|_F^2 d\omega, \tag{B.2}$$

where (B.1) is by the convolution property of the Fourier transform. Since the rank of $\mathbf{C}(\omega)\mathbf{B}(\omega)$is less or equal to q, when ω is fixed the minimization problem for the integrand is a low-rank approximation problem for $\mathbf{S}^{1/2}(\omega)$, and it is well-known that the solution is provided by the projection matrix of q-dimensional principal subspace of $\mathbf{S}(\omega)$. The conclusion of the theorem follows.

C Proofs of Main Results

In this section, we proof the main result on Theorem 3.3. The proofs of Theorem 4.1 and Theorem 4.3 are deferred to the Supplementary Material.

C.1 Proof for Theorem 3.3

In order to prove Theorem 3.3, we first define the complex valued sub-Gaussian random variable.

Definition C.1. A complex random variable $Z = X + iY$ is said sub-Gaussian (resp. sub-exponential) if $|Z|$ is sub-Gaussian (resp. sub-exponential). A complex random vector $\boldsymbol{Z} = \boldsymbol{X} + i\boldsymbol{Y}$ is called sub-Gaussian (resp. sub-exponential) if for all $\mathbf{v} \in \mathbb{C}^d$, $\mathbf{v}^\dagger\boldsymbol{Z}$ is sub-Gaussian (resp. sub-exponential).

We next describe lemmas on the concentration properties for the cross-covariance matrices. The following lemma shows the decaying rate of cross-covariance matrices.

Lemma C.2. There exist constants c_3 and c_4 which only depend on c_1, c_2 and γ_2, such that for all $t \in \mathbb{Z}$ and all $\omega \in [0, 1)$,

$$\|\mathbf{R}_t e^{-i2\pi\omega t} + \mathbf{R}_{-t} e^{i2\pi\omega t}\|_2 \le c_3 e^{-c_4 t^{\gamma_1}}.$$

We defer the proof to Section G.3 in the Supplementary Material.

Note that a direct consequence of Lemma C.2 is that the series $\{\mathbf{R}_t\}_{t\in\mathbb{Z}}$ is absolutely summable, and for all $\omega \in [0,1)$, $\mathbf{S}(\omega)$ is well-defined. Moreover, as a function of ω, the smoothness of $\mathbf{S}(\omega)$ is intimately related to the decay pattern of $\{\mathbf{R}_t\}_{t\in\mathbb{Z}}$.

Lemma C.3. There exists a constant c_5 only depending on c_1, c_2, γ_1 and γ_2 such that for all ω and $\widetilde{\omega}$ in $[0,1)$, we have $\|\mathbf{S}(\widetilde{\omega}) - \mathbf{S}(\omega)\|_2 \le c_5|\widetilde{\omega} - \omega|$.

We defer the proof to Section G.4 in the Supplementary Material.

The above lemma will be used to prove Theorem 4.3.

We now proceed to the estimation procedure for the covariance matrix function $\mathbf{R}_u$ and the spectral density matrix $\mathbf{S}(\omega)$. Recall the estimators $\widehat{\mathbf{R}}_t$ and $\widehat{\mathbf{S}}(\omega)$ in (3.1) and (3.2). At first glance of (3.2), the scaling of $\log T$ for the number of estimates of $\mathbf{R}_u$ we use seems counter-intuitive. We will provide extensive explanations and intuitions in the following proofs and remarks. We have the following proposition.

Proposition C.4. Recall that γ is defined via $1/\gamma = 1/\gamma_1 + 2/\gamma_2$ and $N(t,T)$ is the number of the samples we use in (3.1). There exist constant C_1, C_2, C_3, C_4 and V which only depend on c_1, c_2, γ_1 and γ_2 such that for all $t \in \mathbb{Z}$, $\mathbf{v} \in \mathbb{S}^{d-1}(\mathbb{C})$ and $\lambda > 1/(N(t,T)+1)$,

$$\mathbb{P}\left(|\mathbf{v}^\dagger\left(\widehat{\mathbf{R}}_t - \mathbf{R}_t\right)\mathbf{v}| \ge \lambda\right) \le (N(t,T)+1)\exp\left(-\frac{(N(t,T)+1)^\gamma\,\lambda^\gamma}{C_1}\right)$$
$$+\exp\left(-\frac{(N(t,T)+1)^2\,\lambda^2}{C_2(1+(N(t,T)+1)\,V)}\right) + \exp\left(-\frac{(N(t,T)+1)\,\lambda^2}{C_3}\exp\left(\frac{(N(t,T)+1)^{\gamma(1-\gamma)}\,\lambda^{\gamma(1-\gamma)}}{C_4\,(\log(N(t,T)+1)\,\lambda)^\gamma}\right)\right).$$

Proof. Define

$$\boldsymbol{Y}_{\mathbf{v}}^t(k) = \mathbf{v}^\dagger \boldsymbol{X}_{(k+1)t+k}\boldsymbol{X}_{kt+k}^\dagger \mathbf{v} - \mathbb{E}[\mathbf{v}^\dagger \boldsymbol{X}_{(k+1)t+k}\boldsymbol{X}_{kt+k}^\dagger \mathbf{v}],$$

which is a one-dimensional centered stationary process. Then by the definition of $\widehat{\mathbf{R}}_t$ we have

$$\mathbf{v}^\dagger\left(\widehat{\mathbf{R}}_t - \mathbf{R}_t\right)\mathbf{v} = \frac{1}{N(t,T)+1}\sum_{k=0}^{N(t,T)} \boldsymbol{Y}_{\mathbf{v}}^t(k).$$

We first note that by Lemma C.2, we have

$$|\mathbb{E}[\mathbf{v}^\dagger \boldsymbol{X}_{(k+1)t+k}\boldsymbol{X}_{kt+k}^\dagger \mathbf{v}]| \le \|\mathbf{R}_t\|_2 \le \frac{8\sqrt{2}}{\gamma_2},$$

where the upper bound doesn't depend on t or $\mathbf{v}$. We check readily that for all $n \ge 1$, we have $\alpha_{\boldsymbol{Y}_{\mathbf{v}}^t}(n) \le \alpha_{\boldsymbol{X}}(n(t+1)+1)$, so by Assumption 2, for all $u \ge 1$,

$$\alpha_{\boldsymbol{Y}_v^t}(u) \le e^{-c_1 u^{\gamma_1}}. \tag{C.1}$$

On the other hand, by Assumption 3 and Lemma G.1, there exist c_2' which only depends on c_2 such that for all $\mathbf{v} \in \mathbb{S}^{d-1}(\mathbb{C})$, $t \in \mathbb{Z}$, $k \in \{1,2,\cdots,N(t,T)\}$ and $\lambda \ge 0$, we have

$$\mathbb{P}(|\boldsymbol{Y}_{\mathbf{v}}^t(k) - \mathbb{E}[\boldsymbol{Y}_{\mathbf{v}}^t(k)]| \ge \lambda) \le 2e^{-c_2'\lambda^{\gamma_2/2}}. \tag{C.2}$$

Finally, in order to apply Theorem 1 in Merlevède et al. (2011), we first define

$$V_{\mathbf{v}}^t = \sup_{M\geq 1}\sup_{k>0} \operatorname{Var}(\varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(k))) + 2\sum_{j>k} |\operatorname{Cov}(\varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(k)), \varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(j)))|$$
$$= \sup_{M\geq 1} \operatorname{Var}(\varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(0))) + 2\sum_{j>0} |\operatorname{Cov}(\varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(0)), \varphi_M(\boldsymbol{Y}_{\mathbf{v}}^t(j)))|,$$

where $\varphi_M(x) = \max(\min(x, M), -M)$ and in the second equality we used the stationarity of $\boldsymbol{Y}_{\mathbf{v}}^t$. Combining (C.1) with (C.2), by applying Theorem 1 in Merlevède et al. (2011), we can readily find a positive constant V such that $V \geq V_{\mathbf{v}}^t$ for all $\mathbf{v} \in \mathbb{S}^{d-1}(\mathbb{C})$ and $t \in \mathbb{Z}$. That theorem also implies that there thus exist constants $C_1, \cdots, C_4$, depending only on c, γ_1 and γ_2, such that for all T and $\lambda > 1/(N(t,T)+1)$, we have the desired inequality. Note that the condition $\lambda > 1/(N(t,T)+1)$ guarantees that the term $(\log((N(t,T)+1)\lambda))^{1/2}$ is well-defined. □

With the preparation above, we can start proving Theorem 3.3.

Proof for Theorem 3.3. We simply write the number of cross-covariance $M(T)$ as M in the proof. We start with the ε-net argument. It can be readily proved that in order to cover $\mathbb{S}^{d-1}(\mathbb{C}) \cap \mathbb{B}_0(s^*)$ using balls of radisu $1/2$, we need only $\binom{d}{s^*}6^{s^*}$ points. The set consisting of these points will be denoted $\mathcal{N}_{1/2}$ and is called a $1/2$-net for $\mathbb{S}^{d-1}(\mathbb{C}) \cap \mathbb{B}_0(s^*)$. There exists a universal constant C such that for all Hermitian matrices $\mathbf{A}$,

$$\|\mathbf{A}\|_{\mathrm{op},s^*} \leq C \max_{\mathbf{v}\in\mathcal{N}_{1/2}} |\mathbf{v}^\dagger \mathbf{A} \mathbf{v}|. \tag{C.3}$$

We can thus replace $\|\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\|_{\mathrm{op},s^*}$ by $\max_{\mathbf{v}\in\mathcal{N}_{1/2}} |\mathbf{v}^\dagger(\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega))\mathbf{v}|$ at a cost of adding a constant factor. Assume that the cutoff value is M. Thus, for $\lambda \geq 0$, we have $\widehat{\mathbf{R}}_t = 0$ for $t > M$ or $t < M$. Then by union bound, we get

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\|_{\mathrm{op},s^*} \geq C\lambda\right) \leq \sum_{\mathbf{v}\in\mathcal{N}_{1/2}} \mathbb{P}\left(|\mathbf{v}^\dagger\left(\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\right)\mathbf{v}| \geq \lambda\right) \\
\leq &\binom{d}{s^*}6^{s^*}\mathbb{P}\left(\left|\sum_{t=-M}^{M} \mathbf{v}^\dagger\left(\widehat{\mathbf{R}}_t - \mathbf{R}_t\right)\mathbf{v}e^{-i2\pi\omega t}\right| + \left|\sum_{|t|>M} \mathbf{v}^\dagger\mathbf{R}_t\mathbf{v}e^{-i2\pi\omega t}\right| \geq \lambda\right) \\
\leq &\binom{d}{s^*}6^{s^*}\left(\mathbb{P}\left(\left|\sum_{t=-M}^{M} \mathbf{v}^\dagger\left(\widehat{\mathbf{R}}_t - \mathbf{R}_t\right)\mathbf{v}e^{-i2\pi\omega t}\right| \geq \frac{\lambda}{2}\right) + \mathbb{P}\left(\left|\sum_{|t|>M} \mathbf{v}^\dagger\mathbf{R}_t\mathbf{v}e^{-i2\pi\omega t}\right| \geq \frac{\lambda}{2}\right)\right),
\end{aligned} \tag{C.4}$$

where on the second line $\mathbf{v}$ is an arbitrary vector in $\mathbb{S}^{d-1}$. We will study the two terms seperately and start with the second term which corresponds to the population tail behaviour of the auto-covariance matrices. By applying Lemma C.2 and noticing that in Assumption 2 we assume $\gamma_1 \geq 1$, we have

$$\begin{aligned}
\left|\sum_{|t|>M} \mathbf{v}^\dagger\mathbf{R}_t\mathbf{v}e^{-i2\pi\omega t}\right| \leq &2\sum_{t>M} \|\mathbf{R}_t e^{-i2\pi\omega t} + \mathbf{R}_{-t}e^{i2\pi\omega t}\|_2 \\
\leq &2c_3\sum_{t>M} e^{-c_4 t^{\gamma_1}} \leq 2c_3 e^{-c_4 M}.
\end{aligned} \tag{C.5}$$

We then turn to the first term and distribute the mass equality among the terms:

$$\mathbb{P}\left(\sum_{t=-M}^{M}|\mathbf{v}^{\dagger}\left(\widehat{\mathbf{R}}_t-\mathbf{R}_t\right)\mathbf{v}|\geq\frac{\lambda}{2}\right)\leq\sum_{t=-M}^{M}\mathbb{P}\left(|\mathbf{v}^{\dagger}(\widehat{\mathbf{R}}_t-\mathbf{R}_t)\mathbf{v}|\geq\frac{\lambda}{2(2M+1)}\right). \tag{C.6}$$

The union bound combined with this observation in (C.4) yields

$$\sum_{t=-M}^{M}\binom{d}{s^*}6^{s^*}\mathbb{P}\left(|\mathbf{v}^{\dagger}(\widehat{\mathbf{R}}_t-\mathbf{R}_t)\mathbf{v}|\geq\frac{\lambda}{2(2M+1)}\right)\leq 2(\mathbf{I}+\mathbf{II}+\mathbf{III}), \tag{C.7}$$

where

$$\mathbf{I}=\binom{d}{s^*}6^{s^*}\sum_{t=0}^{M}(N(t,T)+1)\exp\left(-\frac{(N(t,T)+1)^{\gamma}\lambda^{\gamma}(4M+2)^{-\gamma}}{C_1}\right),$$

$$\mathbf{II}=\binom{d}{s^*}6^{s^*}\sum_{t=0}^{M}\exp\left(-\frac{(N(t,T)+1)^2\lambda^2(4M+2)^{-2}}{C_2(1+(N(t,T)+1)V)}\right),$$

$$\mathbf{III}=\binom{d}{s^*}6^{s^*}\sum_{t=0}^{M}\exp\left(-\frac{(N(t,T)+1)\,\lambda^2}{C_3}\exp\left(\frac{(N(t,T)+1)^{\gamma(1-\gamma)}\,\lambda^{\gamma(1-\gamma)}}{C_4\left(\log\left(N(t,T)+1\right)\lambda\right)^{\gamma}(4M+2)^{\lambda(1-\lambda)}}\right)\right).$$

We are going to analyze and provide concentration upper bounds for each of these three terms.

- For term $\mathbf{I}$, note that for all $t\in\{0,1,\cdots,M\}$ we have

$$\frac{T}{M}\leq N(t,T)+1\leq T+1, \tag{C.8}$$

by plugging these two bounds into the expression of $\mathbf{I}$ we find the following upper bound

$$\mathbf{I}\leq c_1(6d)^{s^*}TM\exp\left(-\frac{T^{\gamma}\lambda^{\gamma}}{\widetilde{C}_1M^{2\gamma}}\right),$$

where c_1 is a universal constant and $\widetilde{C}_1$ is a constant only depends on the c_1, c_2, γ_1 and γ_2. If we set the right hand side of the above inequality to be ε, we reversely solve $\lambda^{(\mathbf{I})}$ which gives

$$\lambda^{(\mathbf{I})}=\frac{(-\log\varepsilon+s^*\log d+\log(TM))^{1/\gamma}(M)^2}{T}.$$

- For term $\mathbf{II}$, again by using (C.8), we have

$$\mathbf{II}\leq c_2(6d)^{s^*}M\exp\left(-\frac{T\lambda^2}{\widetilde{C}_2M^3}\right),$$

where c_2 is a universal constant and $\widetilde{C}_2$ is a constant only depends on γ_1 and γ_2. Again by setting the right hand side to be ε and assume that $\log M\leq c\cdot s\log d$ where c is a universal constant, we solve for $\lambda^{(\mathbf{II})}$ which gives

$$\lambda^{(\mathbf{II})}\leq\sqrt{\frac{M^3(-\log\varepsilon+s^*\log d+\log M)}{T}}.$$

- For term **III**, we simply observe that the inner exponential term is greater or equal to 1. As a result, **II** $\geq$ **III** asymptotically. Therefore, we only need to consider the influence of **II**.

Putting all the three pieces together, since there exists a universal constant c such that for all T, s^* and d satisfying the scaling $M(s^* \log d)^{2/\gamma-1} \leq T \leq d$, we have $\max\{\lambda^{(\mathbf{I})}, \lambda^{(\mathbf{II})}, \lambda^{(\mathbf{III})}\} \leq c\lambda^{(\mathbf{II})}$. It means that the rate given by the first term of (C.4) is $\lambda^{(\mathbf{II})}$. Let $\varepsilon = 1/d^{s^*}$ and we can see that the rate is

$$CR_1 = M^{3/2}\sqrt{\frac{s^* \log d}{T}} \tag{C.9}$$

Observe that the rate given by the second term of (C.4) is specified in (C.5), which is $CR_2 = e^{-c_4 M}$. The final rate should be the bigger one of CR_1 and CR_2. We can see that with the increasement of M, CR_1 increases and CR_2 decreases. When $M = O_P(1)$, CR_2 does not converge and when $T^{1/3}/M = O_P(1)$, CR_1 does not converge, When $M \to \infty$ and $M/T^{1/3} \to 0$, the estimator converges under our scaling and the rate is $CR_1 \vee CR_2$, which completes our proof for the rate of $\|\cdot\|_{\text{op},s^*}$-norm. The proof for the convergence rate in $\|\cdot\|_{\infty,\infty}$-norm is the same as the previous proof. We only need to change the cut-off step in (C.4) into the following argument. Define the set $\mathcal{V} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, where $\mathbf{e}_j$ is the j-th canonical basis of $\mathbb{R}^d$.

$$\begin{aligned}
&\mathbb{P}\Big(\|\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\|_{\infty,\infty} \geq C\lambda\Big) \leq \sum_{\mathbf{u},\mathbf{v}\in\mathcal{V}} \mathbb{P}\Big(|\mathbf{u}^\dagger\Big(\widehat{\mathbf{S}}_M(\omega) - \mathbf{S}(\omega)\Big)\mathbf{v}| \geq \lambda\Big) \\
\leq& d^2 6^{s^*}\mathbb{P}\Big(\Big|\sum_{t=-M}^{M} \mathbf{u}^\dagger\Big(\widehat{\mathbf{R}}_t - \mathbf{R}_t\Big)\mathbf{v}e^{-i2\pi\omega t}\Big| + \Big|\sum_{|t|>M} \mathbf{u}^\dagger\mathbf{R}_t\mathbf{v}e^{-i2\pi\omega t}\Big| \geq \lambda\Big) \\
\leq& d^2 6^{s^*}\Big(\mathbb{P}\Big(\Big|\sum_{t=-M}^{M} \mathbf{u}^\dagger\Big(\widehat{\mathbf{R}}_t - \mathbf{R}_t\Big)\mathbf{v}e^{-i2\pi\omega t}\Big| \geq \frac{\lambda}{2}\Big) + \mathbb{P}\Big(\Big|\sum_{|t|>M} \mathbf{u}^\dagger\mathbf{R}_t\mathbf{v}e^{-i2\pi\omega t}\Big| \geq \frac{\lambda}{2}\Big)\Big).
\end{aligned} \tag{C.10}$$

The remaining part of the proof follows the same way. □

References

Back, A. D. and Weigend, A. S. (1997). A first application of independent component analysis to extracting structure from stock returns. *International journal of neural systems* **8** 473–484.

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

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S. and Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107** 4734–4739.

Box, G. E. and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika* **64** 355–365.

Brillinger, D. (1969). The canonical analysis of stationary time series. In *Multivariate Analysis - II* (P. R. Krishnaiah, ed.). Academic Press, 331–350.

Chang, J., Guo, B. and Yao, Q. (2014). Segmenting multiple time series by contemporaneous linear transformation. *arXiv preprint arXiv:1410.2323* .

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

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association* **100** 830–840.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, vol. 3. JHU Press.

Guo, S., Wang, Y. and Yao, Q. (2015). High dimensional and banded vector autoregressions. *arXiv preprint arXiv:1502.07831* .

Johansen, S. (1995). Likelihood-based inference in cointegrated vector autoregressive models. *OUP Catalogue* .

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

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Jung, A., Heckel, R., Bolcskei, H. and Hlawatsch, F. (2014). Compressive nonparametric graphical model selection for time series. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE.

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

Lei, J. and Vu, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *The Annals of Statistics* **43** 299–322.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* **41** 772–801.

Manolakis, D. G., Ingle, V. K. and Kogon, S. M. (2005). *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*, vol. 46. Artech House Norwood.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on* **9** 504–512.

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

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** 435–474.

Overton, M. L. and Womersley, R. S. (1992). On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications* **13** 41–45.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379.

Qiu, H., Han, F., Liu, H. and Caffo, B. (2015). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* to appear.

Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques* **29** 587–597.

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

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97** 1167–1179.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. Tech. rep., National Bureau of Economic Research.

Stoffer, D. S. (1999). Detecting common signals in multiple time series using the spectral envelope. *Journal of the American Statistical Association* **94** 1341–1356.

Stoica, P. and Moses, R. L. (1997). *Introduction to spectral analysis*, vol. 1. Prentice hall Upper Saddle River.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society. Series B (Methodological)* **51** 157–213.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*.

Wang, Z., Lu, H. and Liu, H. (2014). Nonconvex statistical optimization: Minimax-optimal sparse pca in polynomial time. *arXiv preprint arXiv:1408.5352* .

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

Xiao, H. and Wu, W. B. (2012). Covariance matrix estimation for stationary time series. *The Annals of Statistics* **40** 466–493.

Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *The Journal of Machine Learning Research* **14** 899–925.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics* **15** 265–286.

Supplementary material to

Sparse Principal Component Analysis in Frequency Domain for Time Series

Junwei Lu*, Yichen Chen†, Xiuneng Zhu‡, Fang Han§, and Han Liu¶

Abstract

This document contains the supplementary material to the paper "Sparse Principal Component Analysis in Frequency Domain for Time Series". In Section D, we prove Theorem 4.1. In Section E, we prove Theorem 4.2. In Section F, we prove Theorem 4.3. In Section G, we prove auxiliary lemmas for the proof of Theorem 3.3.

D Proof of Theorem 4.1

In this section we present the proof for Theorem 4.1. The proof is almost same as the one of Theorem 4.3 in Wang et al. (2014) except that in Theorem 4.1, all matrices take values in $\mathbb{C}$. Here we are dealing with Hermitian complex matrices instead of symmetric real matrices, but we expect all the important properties and results stay unchanged. We will go over some of them and point out the difference between these two settings.

First, we define the distance of two complex linear spaces. Assume that we have two subspaces $\mathcal{U}$ and $\mathcal{U}'$ which are both q-dimensional. We use $\mathbf{\Pi}$ and $\mathbf{\Pi}'$ to denote the canonical projections onto these two spaces which are Hermitian matrices. Then we define the distance between these two subspaces as

$$D(\mathcal{U},\mathcal{U}') = \|\mathbf{\Pi} - \mathbf{\Pi}'\|_{\mathrm{F}}.$$

We will use $\mathbf{u}_1, \cdots, \mathbf{u}_q$ to denote a set of orthonormal basis of $\mathcal{U}$ and $\mathbf{u}'_1, \cdots, \mathbf{u}'_q$ to denote that of $\mathcal{U}'$. Note that in contrast to the canonical projections corresponding to these subspaces, these orthonormal basis are not unique. However, if we use $\mathbf{U}$ to denote $(\mathbf{u}_1|\cdots|\mathbf{u}_q)$, then $\mathbf{\Pi} = \mathbf{U}\mathbf{U}^\dagger$. We also denote the orthogonal projection matrix $\mathbf{U}'^{\perp} = \mathbf{I} - \mathbf{U}'\mathbf{U}'^\dagger$.

Lemma D.1. We have the following identity on the distance of two subspaces:

$$D(\mathcal{U},\mathcal{U}') = \sqrt{2}\|\mathbf{U}^\dagger\mathbf{U}'^{\perp}\|_{\mathrm{F}} = \sqrt{2}\left(q - \|\mathbf{U}^\dagger\mathbf{U}'\|_{\mathrm{F}}^2\right)^{1/2} \le \sqrt{2q}.$$

Proof. For the first two equalities, we can easily check by calculation. By definition, we have

$$D(\mathcal{U},\mathcal{U}')^2 = \mathrm{tr}\left((\mathbf{U}\mathbf{U}^\dagger - \mathbf{U}'\mathbf{U}'^\dagger)(\mathbf{U}\mathbf{U}^\dagger - \mathbf{U}'\mathbf{U}'^\dagger)^\dagger\right) = 2q - 2\,\mathrm{tr}(\mathbf{U}^\dagger\mathbf{U}'\mathbf{U}'^\dagger\mathbf{U}) = 2q - 2\|\mathbf{U}^\dagger\mathbf{U}'\|_{\mathrm{F}}^2.$$

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: junweil@princeton.edu,

†Computer Science Department, Princeton University, Princeton, NJ 08544, USA;e-mail: yichenc@princeton.edu,

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: xiunengz@princeton.edu,

§Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA; e-mail: fhan@jhsph.edu,

¶Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: hanliu@princeton.edu.

Similarly we have,

$$2\|\mathbf{U}^\dagger\mathbf{U}'^\perp\|_{\mathrm{F}}^2 = 2\,\mathrm{tr}\left(\mathbf{U}^\dagger(\mathbf{I}-\mathbf{U}'\mathbf{U}'^\dagger)(\mathbf{I}-\mathbf{U}'\mathbf{U}'^\dagger)\mathbf{U}\right) = 2q - 2\,\mathrm{tr}(\mathbf{U}^\dagger\mathbf{U}'\mathbf{U}'^\dagger\mathbf{U}) = 2q - 2\|\mathbf{U}^\dagger\mathbf{U}'\|_{\mathrm{F}}^2.$$

The last inequality follows trivially from the positiveness of the Frobenius norm. □

The ADMM algorithm for sparse PCA problem is initially proposed in Vu et al. (2013) in the real number setting. We next check the curvature lemma first proposed there for real case.

Lemma D.2. Let $\mathbf{A} \in \mathbb{C}^{d\times d}$ be a Hermitian matrix and $\mathbf{E}$ be the projection matrix that projects vectors onto the principal q-dimensional subspace of $\mathbf{A}$. If $\delta_{\mathbf{A}} = \lambda_q(\mathbf{A}) - \lambda_{q+1}(\mathbf{A})$, then

$$\frac{\delta_{\mathbf{A}}}{2}\|\mathbf{E}-\mathbf{F}\|_{\mathrm{F}}^2 \le \langle \mathbf{A}, \mathbf{E}-\mathbf{F}\rangle,$$

where $\mathbf{F}$ satisfies $0 \preceq \mathbf{F} \preceq \mathbf{I}$ and $\mathrm{tr}(\mathbf{F}) = q$.

Proof. If $\mathbf{A}$ is not a positive definite matrix, there exists a constant $\rho > 0$ that $\mathbf{A} + \rho\mathbf{I} \succeq 0$. Since $\langle \mathbf{A}, \mathbf{E}-\mathbf{F}\rangle = \langle \mathbf{A}+\rho\mathbf{I}, \mathbf{E}-\mathbf{F}\rangle$ and $\mathbf{A}+\rho\mathbf{I}$ has the same eigengap as $\mathbf{A}$, we can just prove the case that $\mathbf{A}$ is positive definite. We suppose the Hermitian matrix $\mathbf{A} = \sum_{j=1}^d \lambda_j \mathbf{u}_j\mathbf{u}_j^\dagger$, where $\lambda_1 \ge \ldots \ge \lambda_d > 0$ are eigenvalues and $\{\mathbf{u}_j\}_{j=1}^d$ are the eigenvectors. So $\mathbf{E} = \sum_{j=1}^q \lambda_j \mathbf{u}_j\mathbf{u}_j^\dagger$. We first have

$$\frac{1}{2}\|\mathbf{E}-\mathbf{F}\|_{\mathrm{F}}^2 = \frac{1}{2}\Big(\|\mathbf{E}\|_{\mathrm{F}}^2 - 2\langle \mathbf{E}, \mathbf{F}\rangle + \|\mathbf{F}\|_{\mathrm{F}}^2\Big) \le \frac{1}{2}\Big(d - 2\langle \mathbf{E}, \mathbf{F}\rangle + \mathrm{tr}(\mathbf{R})\Big) = d - \langle \mathbf{E}, \mathbf{F}\rangle.$$

On the other side, we have

$$\begin{aligned}\langle \mathbf{A}, \mathbf{E}-\mathbf{F}\rangle &= \langle \mathbf{E}\mathbf{A}, \mathbf{I}-\mathbf{F}\rangle + \langle (\mathbf{I}-\mathbf{E})\mathbf{A}, \mathbf{F}\rangle = \langle \sum_{j=1}^q \lambda_j \mathbf{u}_j\mathbf{u}_j^\dagger, \mathbf{I}-\mathbf{F}\rangle + \langle \sum_{j=q+1}^d \lambda_j \mathbf{u}_j\mathbf{u}_j^\dagger, \mathbf{F}\rangle \\ &\ge \lambda_q\langle \mathbf{E}, \mathbf{I}-\mathbf{F}\rangle - \lambda_{q+1}\langle \mathbf{I}-\mathbf{E}, \mathbf{F}\rangle = \delta_{\mathbf{A}}(d - \langle \mathbf{E}, \mathbf{F}\rangle).\end{aligned}$$

Therefore, the lemma is proved. □

With the help of the above lemma, we can directly bound the distance between two principal subspaces as it is in Vu et al. (2013). The following $\sin\Theta$ theorem summarizes this fact and we omit the proof.

Lemma D.3. Let $\mathbf{A}$ and $\mathbf{B}$ be Hermitian matrices and $\mathcal{M}_{\mathbf{A}}$ and $\mathcal{M}_{\mathbf{B}}$ be their respective q-dimensional principal subspaces. If $\delta_{\mathbf{A},\mathbf{B}} = (\lambda_q(\mathbf{A}) - \lambda_{q+1}(\mathbf{A})) \wedge (\lambda_q(\mathbf{B}) - \lambda_{q+1}(\mathbf{B}))$, then

$$D(\mathcal{M}_{\mathbf{A}}, \mathcal{M}_{\mathbf{B}}) \le \frac{\sqrt{2}}{2\delta_{\mathbf{A},\mathbf{B}}}\|\mathbf{A}-\mathbf{B}\|_{\mathrm{F}}.$$

As explained in Vu et al. (2013), ADMM for sparse principal subspace problems relies on convex relaxation of the constraint. The following lemma is the complex counterpart.

Lemma D.4. The complex q-fantope $\mathcal{F}^q$ is the convex hull of the set $\{\mathbf{V}\mathbf{V}^\dagger | \mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_q\}$.

Proof. Actually we can prove that the latter is the set of all extremal points of $\mathcal{F}^q$. The proof is similar to Theorem 3 in Overton and Womersley (1992) in terms of the real case. First, we consider the simplex

$$\mathcal{S} = \Big\{(\lambda_1, \ldots, \lambda_d) | \sum_{j=1}^{d} \lambda_j = q, \lambda_j \geq 0, \forall 1 \leq j \leq d\Big\}.$$

It is easy to know that the extremal points set is

$$\mathcal{S}_0 = \Big\{(\lambda_1, \ldots, \lambda_d) | \#\{\lambda_j = 1\} = q, \#\{\lambda_j = 0\} = d - q\Big\}.$$

Therefore, the extremal points set of $\mathcal{F}^q$ is the subset of the set $\big\{\mathbf{V}\mathbf{V}^\dagger | \mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_q\big\}$. On the other hand, the compactness of $\mathcal{F}^q$ implies that the extremal points set is not empty. Therefore, there must exists at least one $\mathbf{V}_0\mathbf{V}_0^\dagger \in \big\{\mathbf{V}\mathbf{V}^\dagger | \mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_q\big\}$ being the extremal point. Finally, it is trivial to see that if $\mathbf{V}_0\mathbf{V}_0^\dagger$ is the extremal point, for any Hermitian matrix $\mathbf{U}$, the matrix $\mathbf{U}\mathbf{V}_0\mathbf{V}_0^\dagger\mathbf{U}^\dagger$ is also the extremal point. In conclusion, we prove that $\big\{\mathbf{V}\mathbf{V}^\dagger | \mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_q\big\}$ is the extremal points set of $\mathcal{F}^q$. □

Our final piece of lemma is the following statistical rate of convergence in infinity norm. Combined with all the deterministic results developed above we can prove Theorem 4.1.

Proof of Theorem 4.1. Define $\widetilde{\mathbf{\Theta}}^{(t)} = -\rho \cdot \mathrm{sign}(\overline{\mathbf{\Pi}}^{(t)} - \overline{\mathbf{\Phi}}^{(t)})$, recalling that $\overline{\mathbf{\Pi}}^{(t)}$ and $\overline{\mathbf{\Phi}}^{(t)}$ for $1 \leq t \leq R$ are defined in Algorithm A. In Equation (C.19) of Wang et al. (2014), as long as Lemmas D.1 - D.4 are satisfied, it can be proved that when $\rho \geq \|\widehat{\mathbf{S}}(0) - \mathbf{S}(0)\|_{\infty,\infty}$, we have for any $1 \leq t \leq R$,

$$\|\mathbf{\Pi}^* - \overline{\mathbf{\Pi}}^{(t)}\|_{\mathrm{F}} \leq \frac{4s^*\rho}{\lambda_q(\mathbf{S}(0)) - \lambda_{q+1}(\mathbf{S}(0))} + \frac{\sqrt{\beta}\|\mathbf{\Pi}^*\|_{\mathrm{F}} + \|\widetilde{\mathbf{\Theta}}^{(t)}\|_{\mathrm{F}}/\sqrt{\beta}}{\sqrt{\lambda_q(\mathbf{S}(0)) - \lambda_{q+1}(\mathbf{S}(0))}} \frac{1}{\sqrt{t}}, \tag{D.1}$$

where in the current context $\mathbf{\Pi}^*$ is the q-dimensional principal subspace of $\mathbf{S}(0)$ and $\overline{\mathbf{\Pi}}^{(t)}$ is the output we obtain at step t from ADMM algorithm associated to $\widehat{\mathbf{S}}(0)$. As a projection matrix on a q-dimensional subspace, $\|\mathbf{\Pi}^*\|_{\mathrm{F}} = \sqrt{q}$. On the other hand, by the definition of $\widetilde{\mathbf{\Theta}}^{(t)}$, $\|\widetilde{\mathbf{\Theta}}^{(t)}\|_{\mathrm{F}} \leq \rho d$. By Theorem 3.3, it suffices to choose $\rho = C(\log T)^{3/2}\sqrt{\log d/T}$ to guarantee (D.1) with probability at least $1 - 1/d$. Plug ρ into (D.1) and choose β to minimize it, we obtain the desired bound. □

E Proof of Theorem 4.2

Proof. Let $\mathcal{U}^{init}$ be the q-leading eigenspace of $\overline{\mathbf{\Pi}}^{(R)}$, since

$$R \geq \left(\frac{qd^2(\log T)^3 \log d}{T}\right)^{1/2} \cdot \left(A - C(\log T)^{3/2}\sqrt{\frac{\log d}{T}}\right)^{-1},$$

by Theorem 4.1, we have

$$D(\mathcal{U}^{init}, \mathcal{U}^*) \leq A = \min\left\{\sqrt{2\eta}/4, \sqrt{q\eta(1 - \eta^{1/2})/2}\right\}, \tag{E.1}$$

According to Theorem 4.2 in Wang et al. (2014), we can obtain

$$D(\mathcal{U}^{(R+t)},\mathcal{U}^*) \le \eta^{t/4} D(\mathcal{U}^{init},\mathcal{U}^*) + 8\sqrt{2}\eta^{-1/2}\frac{1-\eta^{1/4}}{1-\eta^{t/4}}\frac{\|\widehat{\mathbf{S}}(0)-\mathbf{S}(0)\|_{2,2\widehat{s}}}{\lambda_q(\mathbf{S}(0))-\lambda_{q+1}(\mathbf{S}(0))}.$$

Plugging in the concentration result on $\widehat{\mathbf{S}}(\omega)$ towards $\mathbf{S}(\omega)$, i.e., Theorem 3.3, we have the desired result since

$$\widetilde{R} \ge 4(\log(1/\eta))^{-1}\log\left(C(\eta/8)^{1/2}(\log T)^{3/2}\sqrt{\frac{s^*\log d}{T}}\right).$$

To analyze the CSOAP, we prove by induction. Suppose for $k-1$, $\widehat{\mathcal{U}}((k-1)/N)$ satisfies (4.6) that

$$D(\widehat{\mathcal{U}}((k-1)/N),\mathcal{U}^*((k-1)/N)) \le C(\log T)^{3/2}\sqrt{\frac{qs^*(q\vee\log d)}{T}}.$$

By Lemma C.3, we have

$$D(\mathcal{U}^*(k/N),\mathcal{U}^*((k-1)/N)) \le \sqrt{q}\|\mathbf{S}(\widetilde{\omega})-\mathbf{S}(\omega)\|_2 \le c_5\sqrt{q}/N$$

for $0 \le k \le N$. Therefore, since $N \ge C(\log T)^{-3/2}q^{-1}(s^*\log d/T)^{-1/2}$, we have

$$\begin{aligned}&D(\widehat{\mathcal{U}}((k-1)/N),\mathcal{U}^*(k/N))\\ &\le D(\widehat{\mathcal{U}}((k-1)/N),\mathcal{U}^*((k-1)/N)) + D(\widehat{\mathcal{U}}((k-1)/N),\mathcal{U}^*((k-1)/N)) \le A\end{aligned}$$

for sufficiently large T. Therefore, $\widehat{\mathcal{U}}((k-1)/N)$ satisfies the initialization condition in (E.1) and following the same proof as before, we have

$$D(\widehat{\mathcal{U}}(k/N),\mathcal{U}^*(k/N)) \le C(\log T)^{3/2}\sqrt{\frac{qs^*(q\vee\log d)}{T}},$$

which completes our proof.

□

F Proof of Theorem 4.3

Proof. We first note that

$$\begin{aligned}&\left|\mathbb{E}\left[\|\boldsymbol{X}_t-\boldsymbol{X}_t^{\mathbf{b}}\|_2^2\right]-\mathbb{E}\left[\|\boldsymbol{X}_t-\boldsymbol{X}_t^{\widetilde{\mathbf{b}}}\|_2^2\right]\right|\\ &\le \int_0^1 |\operatorname{tr}\left((\mathbf{I}-\boldsymbol{\Pi}(\omega))\mathbf{S}(\omega)(\mathbf{I}-\boldsymbol{\Pi}(\omega))^\dagger\right)-\operatorname{tr}\left((\mathbf{I}-\widetilde{\boldsymbol{\Pi}}(\omega))\mathbf{S}(\omega)(\mathbf{I}-\widetilde{\boldsymbol{\Pi}}(\omega))^\dagger\right)|d\omega\\ &= \int_0^1 |\operatorname{tr}\left((\mathbf{I}-\boldsymbol{\Pi}(\omega))\mathbf{S}(\omega)\right)-\operatorname{tr}\left((\mathbf{I}-2\widetilde{\boldsymbol{\Pi}}(\omega)+\widetilde{\boldsymbol{\Pi}}(\omega)\widetilde{\boldsymbol{\Pi}}(\omega))\mathbf{S}(\omega)\right)|d\omega\\ &\le \int_0^1 |\operatorname{tr}\left((\boldsymbol{\Pi}(\omega)-\widetilde{\boldsymbol{\Pi}}(\omega))\mathbf{S}(\omega)\right)|+|\operatorname{tr}\left((\widetilde{\boldsymbol{\Pi}}(\omega)(\mathbf{I}-\widetilde{\boldsymbol{\Pi}}(\omega))\mathbf{S}(\omega)\right)|d\omega,\end{aligned} \tag{F.1}$$

where $\boldsymbol{\Pi}(\omega)$ is the true principal subspace of $\mathbf{S}(\omega)$ and $\widetilde{\boldsymbol{\Pi}}(\omega)$ is the estimated one derived from the Fourier transform of $\widetilde{\mathbf{b}}(t)$ and $\widetilde{\mathbf{c}}(t)$ defined in (4.3). The first inequality is due to equation

(B.2) and we use the fact that $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ in the equality step. We then give the bound of $\mathrm{tr}\left((\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega))\mathbf{S}(\omega)\right)$. We use $\|\boldsymbol{A}\|_{F,s}$ to denote the s-sparse Frobenius norm, which is defined to be the maximal Frobenius norm taking arbitrary s rows and s columns of $\mathbf{A}$. As $\mathbf{\Pi}(\omega)$ and $\widetilde{\mathbf{\Pi}}(\omega)$ are $\widehat{s}$-sparse, the matrix $\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)$ is $2\widehat{s}$-sparse. This entails

$$\mathrm{tr}\left((\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega))\mathbf{S}(\omega)\right) \leq \|\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} \|\mathbf{S}(\omega)\|_{\mathrm{F},2\widehat{s}} \leq C\sqrt{qs^*}\|\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} \|\mathbf{S}(\omega)\|_{2,2\widehat{s}}, \tag{F.2}$$

where the last inequality is due to the assumption that $\widehat{s} \leq qs^*$. As $\mathbf{S}(\omega)$ is s^* sparse, we have $\|\mathbf{S}(\omega)\|_{2,2\widehat{s}} = \|\mathbf{S}(\omega)\|_2$. Combining that $\|\mathbf{S}(\omega)\|_2$ is bounded by a constant, we have

$$\begin{aligned}
&\int_0^1 |\,\mathrm{tr}\left((\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega))\mathbf{S}(\omega)\right)| \leq C\sqrt{qs^*}\sum_{k=0}^{N-1}\int_{k/N}^{(k+1)/N} \|\mathbf{\Pi}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} d\omega \\
&\leq C\sqrt{qs^*}\sum_{k=0}^{N-1}\int_{k/N}^{(k+1)/N} \|\mathbf{\Pi}(\omega) - \mathbf{\Pi}(\frac{k}{N})\|_{\mathrm{F}} + \|\mathbf{\Pi}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\frac{k}{N})\|_{\mathrm{F}} \\
&+ \|\widehat{\mathbf{\Pi}}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} + \|\widehat{\mathbf{\Pi}}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} d\omega.
\end{aligned} \tag{F.3}$$

Here, $\widehat{\mathbf{\Pi}}$ is the estimated principal subspace using the Fourier transform of $\widehat{\mathbf{b}}(t)$ and $\widehat{\mathbf{c}}(t)$ defined in Section 4.1. By using Davis-Kahan $\sin\theta$ theorem and Lemma C.3, we can bound the first term as

$$\|\mathbf{\Pi}(\omega) - \mathbf{\Pi}(\frac{k}{N})\|_{\mathrm{F}} \leq \sqrt{2}\|\sin\Theta(\mathbf{B}(\omega), \mathbf{B}(\frac{k}{N}))\|_F \leq c\sqrt{s}\|\mathbf{S}(\omega) - \mathbf{S}(\frac{k}{N})\|_2 \leq c\sqrt{s}|\omega - \frac{k}{N}| \leq \frac{c\sqrt{s}}{N}.$$

where $\mathbf{B}(\omega)$ are the columns of eigenvectors of $\mathbf{S}(\omega)$. By Theorem 4.2, we can bound the second term as

$$\|\mathbf{\Pi}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\frac{k}{N})\|_{\mathrm{F}} \leq (\log T)^{3/2} q\sqrt{\frac{s^* \log d}{T}}. \tag{F.4}$$

To bound the third term, we observe that

$$\|\widehat{\mathbf{\Pi}}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}}^2 = \|\widehat{\mathbf{B}}(\frac{k}{N})\widehat{\mathbf{B}}(\frac{k}{N})^\dagger - \widehat{\mathbf{B}}(\omega)\widehat{\mathbf{B}}(\omega)^\dagger\|_{\mathrm{F}}^2 \leq q\sum_{i=1}^{q}\|\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\|_{\mathrm{F}}^2, \tag{F.5}$$

where $\widehat{\mathbf{v}}_i(\omega)$ is the i-th column of $\widehat{\mathbf{B}}(\omega)$ and $\widehat{\mathbf{B}}(\omega)$ is the Fourier transform of $\widehat{\mathbf{b}}(t)$ and $\widehat{\mathbf{c}}(t)$ computed in Section 4.1. By some calculations and $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, we have for any $i = 1, \ldots, q$,

$$\begin{aligned}
&\|\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\left(\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\right)\left(\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\right)\right) \\
&= \mathrm{tr}\Big(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N}) \\
&\quad + \widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\omega) - \widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\omega)\Big).
\end{aligned}$$

We can therefore further get the equality that

$$\begin{aligned}
\|\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\|_{\mathrm{F}}^2 = \mathrm{tr}\Big(&\left(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N}) + \widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\omega)\right)\left(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)^\dagger\right)\widehat{\mathbf{v}}_i(\frac{k}{N}) \\
&+ \left(\widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\omega) + \widehat{\mathbf{v}}_i(\omega)^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N})\right)\left(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)^\dagger\right)\widehat{\mathbf{v}}_i(\omega)\Big).
\end{aligned}$$

Combining with the observation that $\|\widehat{\mathbf{v}}_i(\omega)\|_2$ are bounded by some constant and $\|\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)^\dagger\|_2 \leq \|\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\frac{k+1}{N})^\dagger\|_2$, we can bound $\left(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\frac{k}{N}) + \widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger\widehat{\mathbf{v}}_i(\omega)\right)\left(\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)^\dagger\right)\widehat{\mathbf{v}}_i(\frac{k}{N})$ by $c\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k+1}{N})\|_2$ for some constant c. As a result,

$$\|\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \widehat{\mathbf{v}}_i(\omega)\widehat{\mathbf{v}}_i(\omega)^\dagger\|_{\mathrm{F}}^2 \leq C\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k+1}{N})\|_2.$$

Substituting the above inequality into equation (F.5), we have

$$\|\widehat{\mathbf{\Pi}}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}}^2 \leq Cq\sum_{i=1}^{q}\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k+1}{N})\|_2. \tag{F.6}$$

We then need to bound $Cq\sum_{i=1}^{q}\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k+1}{N})\|_2$. We know that by triangle inequalities

$$\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \widehat{\mathbf{v}}_i(\frac{k+1}{N})\|_2 \leq \|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \mathbf{v}_i(\frac{k}{N})\|_2 + \|\mathbf{v}_i(\frac{k}{N}) - \mathbf{v}_i(\frac{k+1}{N})\|_2 + \|\widehat{\mathbf{v}}_i(\frac{k+1}{N}) - \mathbf{v}_i(\frac{k+1}{N})\|_2. \tag{F.7}$$

Note that

$$\|\widehat{\mathbf{v}}_i(\frac{k}{N}) - \mathbf{v}_i(\frac{k}{N})\|_2 \leq \|\widehat{\mathbf{v}}_i(\frac{k}{N})\widehat{\mathbf{v}}_i(\frac{k}{N})^\dagger - \mathbf{v}_i(\frac{k}{N})\mathbf{v}_i(\frac{k}{N})^\dagger\|_2 \leq C(\log T)^{3/2}\sqrt{\frac{s^*\log d}{T}}.$$

by Theorem 4.2. Then by the assumption that the eigengaps of $\mathbf{S}(\omega)$ satisfies $\inf_{\omega\in[0,1)}\lambda_i(\mathbf{S}(\omega)) - \lambda_{i+1}(\mathbf{S}(\omega)) > \delta > 0$ for all $i = 1, ..., q$, we can use Davis-Kahan $\sin\theta$ theorem and Lemma C.3 to obtain

$$\|\mathbf{v}_i(\frac{k}{N}) - \mathbf{v}_i(\frac{k+1}{N})\|_2 \leq \|\mathbf{v}_i(\frac{k}{N})\mathbf{v}_i(\frac{k}{N})^\dagger - \mathbf{v}_i(\frac{k+1}{N})\mathbf{v}_i(\frac{k+1}{N})^\dagger\|_2 \leq c\|\mathbf{S}(\omega) - \mathbf{S}(\frac{k}{N})\|_2 \leq \frac{c}{N}$$

So as long as N satisfies

$$N \geq C(\log T)^{-3/2}\left(\frac{s^*\log d}{T}\right)^{-1/2},$$

we can bound all the three terms in (F.7), which gives us

$$\|\widehat{\mathbf{\Pi}}(\frac{k}{N}) - \widehat{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}}^2 \leq C(\log T)^{3/2}q^2\sqrt{\frac{s^*\log d}{T}}, \tag{F.8}$$

by (F.6). Now, we are ready to bound the fourth term in (F.3). Note that

$$\|\widehat{\mathbf{\Pi}}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} = \sqrt{\sum_{i\in\mathcal{S}_\omega}\sum_{j\in\mathcal{S}_\omega}\left(\sum_{k=1}^{q}\widehat{\mathbf{B}}_{ik}(\omega)\widehat{\mathbf{B}}_{jk}(\omega) - \widetilde{\mathbf{B}}_{ik}(\omega)\widetilde{\mathbf{B}}_{jk}(\omega)\right)^2},$$

where $\mathcal{S}_\omega$ is the support of $\widehat{\mathbf{\Pi}}(\omega)$. Recall that $\widehat{\mathbf{B}}_{ik}(\omega)$ is continuously differentiable. So the partial sum $\widetilde{\mathbf{B}}_{ik}(\omega) = \sum_{t=-N}^{N}\widehat{\boldsymbol{b}}_{ik}(t)e^{-i2\pi\omega t}$ satisfies

$$\max_\omega|\widehat{\mathbf{B}}_{ik}(\omega) - \widetilde{\mathbf{B}}_{ik}(\omega)| \leq \frac{C}{\sqrt{N}},$$

by the property of the Fourier transform. Therefore, we have

$$\|\widehat{\mathbf{\Pi}}(\omega) - \widetilde{\mathbf{\Pi}}(\omega)\|_{\mathrm{F}} \leq C\frac{\widehat{sq}}{\sqrt{N}} \tag{F.9}$$

What remains is to bound the second term of (F.1). We have

$$|\operatorname{tr}\left((\widetilde{\mathbf{\Pi}}(\omega)(\mathbf{I}-\widetilde{\mathbf{\Pi}}(\omega))\mathbf{S}(\omega)\right)| \leq C\sqrt{qs^*}\|\widetilde{\mathbf{B}}(\omega)(\mathbf{I}-\widetilde{\mathbf{B}}(\omega)^T\widetilde{\mathbf{B}}(\omega))\widetilde{\mathbf{B}}(\omega)\|_{\mathrm{F}} \leq C\sqrt{qs^*}\|\mathbf{I}-\widetilde{\mathbf{B}}(\omega)^T\widetilde{\mathbf{B}}(\omega)\|_{\mathrm{F}}$$

where the first inequality use the same argument for (F.2). Similarly, we have $\|\mathbf{I}-\widetilde{\mathbf{B}}(\omega)^T\widetilde{\mathbf{B}}(\omega)\|_{\mathrm{F}} \leq \|\mathbf{I}-\widehat{\mathbf{B}}(\omega)^T\widehat{\mathbf{B}}(\omega)\|_{\mathrm{F}} + \|\mathbf{I}-\widetilde{\mathbf{B}}(\omega)^T\widetilde{\mathbf{B}}(\omega)\|_{\mathrm{F}}$. Note that the later term has been bounded before. To bound the first term, we just need to consider the difference caused by each entry and then to add them together, which is given by

$$\|\mathbf{I}-\widehat{\mathbf{B}}(\omega)^T\widehat{\mathbf{B}}(\omega)\|_{\mathrm{F}} \leq C(\log T)^{3/2}d\sqrt{\frac{s^*\log d}{T}}. \tag{F.10}$$

Combining all the bounds together, we can see that when N satisfies

$$N \geq C(\log T)^{-3/2}\left(\frac{s^*\log d}{T}\right)^{-1/2}, \tag{F.11}$$

we have

$$\left|\mathbb{E}\left[\|\boldsymbol{X}_t - \boldsymbol{X}_t^{\mathbf{b}}\|^2\right] - \mathbb{E}\left[\|\boldsymbol{X}_t - \boldsymbol{X}_t^{\widetilde{\mathbf{b}}}\|^2\right]\right| = O_P\left((\log T)^{3/2}q^{3/2}s^*\sqrt{\frac{\log d}{T}}\right), \tag{F.12}$$

which completes our proof.

□

G Auxiliary Lemmas

G.1 Complex Sub-Gaussian Random Variables

Notice that Definition C.1 is consistent with Assumption 3 in the sense that $\boldsymbol{X}_0$ is complex sub-Gaussian random variable under Assumption 3. The following lemma shows a property of the product of two sub-Gaussian random variables.

Lemma G.1. If X and Y are two complex sub-Gaussian random variables, then XY is a complex sub-exponential random variable.

Proof. We first treat the case where X and Y are real-valued centered random variables. If we use m to denote $\mathbb{E}[XY]$, we can write, for t large enough,

$$\mathbb{P}(|XY-m|>t) = \mathbb{P}(XY>m+t) + \mathbb{P}(XY<m-t) \leq 2\mathbb{P}(|XY|>t-m),$$

the inequality is true because $t > m$. Since X and Y are centered and sub-Gaussian, there exists $a, b > 0$ such that for all $t \geq 0$ we have $\mathbb{P}(|X|>t) \leq 2e^{-at^2}$ and $\mathbb{P}(|Y|>t) \leq 2e^{-bt^2}$. Without loss of generality, we assume that $a \leq b$, then for all t large enough, we have

$$\begin{aligned}\mathbb{P}(|XY-m|>t) &\leq 2(\mathbb{P}(|X|>\sqrt{t-m}) + \mathbb{P}(|Y|>\sqrt{t-m})) \\ &\leq 4e^{-a(t-m)} = 4e^{am}e^{-at}.\end{aligned}$$

There thus exists a constant $c > 0$ and $t_0 > 0$ such that for all $t > t_0$, $\mathbb{P}(|XY - m| > t) \leq e^{-ct}$, yielding the fact that XY is sub-exponential.

We now turn to the case where X and Y are not centered. In this case, XY can be written as $\bar{X}\bar{Y} + m_Y\bar{X} + m_X\bar{Y} + m_X m_Y$, where m_X and m_Y are means of X and Y, and $\bar{X} = X - m_X$, $\bar{Y} = Y - m_Y$. Now that $\bar{X}$ and $\bar{Y}$ are centered, we can apply the first part of the proof and the fact that finite linear combinations of sub-exponential random variables are sub-exponential to conclude that XY is sub-exponential. Finally, the case where X and Y are complex-valued follows directly from the definition of sub-Gaussianality and sub-exponentiality for complex-valued random variables.

□

G.2 Auxiliary Lemmas on Variance-Covariance and Spectral Density Matrices

We first state a complex version of Theorem 1 in Rio (1993) which provides us a tool to bound the covariance by the mixing coefficient of two random variables.

Lemma G.2. Let X and Y be two complex random variables such that $|X|$ and $|Y|$ are square-integrable. If we denote α to be the mixing coefficient between the σ-algebras generated by X and Y and $Q_{|X|}(\cdot)$, $Q_{|Y|}(\cdot)$ to be the quantile functions of $|X|$ and $|Y|$, where $Q_{|X|}(1-u) = \inf\{t : \mathbb{P}(|X| > t) \leq u)$. Assuming that $Q_{|X|}Q_{|Y|}$ is integrable on $[0,1]$, then

$$|\operatorname{Cov}(X,Y)| \leq 4\sqrt{2}\int_0^{2\alpha} Q_{|X|}(1-u)Q_{|Y|}(1-u)du.$$

Proof. We write $X = X_1 + X_2 i$ and $Y = Y_1 + Y_2 i$ where X_1, X_2, Y_1, Y_2 are real random variables. Because the covariance actually defines a Hermitian product on the space of random variables, we can write

$$\operatorname{Cov}(X,Y) = \operatorname{Cov}(X_1,Y_1) + \operatorname{Cov}(X_2,Y_2) + i(\operatorname{Cov}(X_1,Y_2) - \operatorname{Cov}(X_2,Y_1)).$$

In the following we use $c_{j,k}$ to denote $|\operatorname{Cov}(X_j, Y_k)|$. Now that X_j and X_k are real random variables. Let $\alpha(X_j, Y_k)$ be the α-mixing coefficient between X_j, Y_k and we can apply Theorem 1 in Rio (1993) and get

$$c_{j,k} \leq 2\int_0^{2\alpha(X_j,Y_k)} Q_{|X_j|}(1-u)Q_{|Y_k|}(1-u)du, \quad j = 1,2, \quad k = 1,2.$$

We then note that $\alpha(X_j, Y_k) \leq \alpha$, because the σ-algebra generated by X (resp. Y) contains that generated by X_j (resp. Y_k). We also have $Q_{|X_j|}(u) \leq Q_{|X|}(u)$ (resp. $Q_{|Y_k|}(u) \leq Q_{|Y|}(u)$) for any $u \in [0,1]$ because $|X_j| \leq |X|$ (resp. $|Y_k| \leq |Y|$). We thus have

$$c_{j,k} \leq 2\int_0^{2\alpha} Q_{|X|}(1-u)Q_{|Y|}(1-u)du, j = 1,2, \quad k = 1,2. \tag{G.1}$$

Lastly we write

$$|\operatorname{Cov}(X,Y)|^2 = (c_{1,1} + c_{2,2})^2 + (c_{1,2} - c_{2,1})^2 \leq 2(c_{1,1}^2 + c_{1,2}^2 + c_{2,1}^2 + c_{2,2}^2),$$

and plug (G.1) into the above equation we obtain the desired result. □

G.3 Proof of Lemma C.2

As we've already noted, $\mathbf{R}_t^\dagger = \mathbf{R}_{-t}$ so the matrix $\mathbf{R}_t e^{-i2\pi\omega t} + \mathbf{R}_{-t} e^{i2\pi\omega t}$ is Hermitian. In order to bound the spectral norm of this matrix we start by writing, for all $\mathbf{v} \in \mathbb{S}^{d-1}(\mathbb{C})$,

$$\begin{aligned} |\mathbf{v}^\dagger(\mathbf{R}_t e^{-i2\pi\omega t} + \mathbf{R}_{-t} e^{i2\pi\omega t})\mathbf{v}| =& |\mathbb{E}[\mathbf{v}^\dagger \boldsymbol{X}_t \boldsymbol{X}_0^\dagger \mathbf{v}] e^{-i2\pi\omega t} + \mathbb{E}[\mathbf{v}^\dagger \boldsymbol{X}_0 \boldsymbol{X}_t^\dagger \mathbf{v}] e^{i2\pi\omega t}| \\ \leq& |\operatorname{Cov}(\mathbf{v}^\dagger \boldsymbol{X}_0, \mathbf{v}^\dagger \boldsymbol{X}_t)| + |\operatorname{Cov}(\mathbf{v}^\dagger \boldsymbol{X}_t, \mathbf{v}^\dagger \boldsymbol{X}_0)|. \end{aligned} \tag{G.2}$$

Note that by Assumption 3, $F_{|\mathbf{v}^\dagger \boldsymbol{X}_u|}(\lambda) \geq 1 - 2e^{-\gamma_2 \lambda^2}$ where F stands for cumulative distribution function. This yields the following upper bound on the quantile function of $|\mathbf{v}^\dagger \boldsymbol{X}_u|$:

$$Q_{|\mathbf{v}^\dagger \boldsymbol{X}_u|}(x) \leq \left(-\frac{1}{c_2} \log \frac{1-x}{2}\right)^{1/\gamma_2}.$$

We hence have, by applying Lemma G.2 and Assumption 2,

$$|\operatorname{Cov}(\mathbf{v}^\dagger \boldsymbol{X}_0, \mathbf{v}^\dagger \boldsymbol{X}_t)| \leq 4\sqrt{2} \int_0^{2e^{-c_1 t^{\gamma_1}}} \left(-\frac{1}{c_1} \log \frac{x}{2}\right)^{2/\gamma_2} dx \leq 8\sqrt{2} e^{-c_1 t^{\gamma_1}} P(ct^{\gamma_1}), \tag{G.3}$$

where P is a polynomial of degree bigger than $2/\gamma_2$. Since we have the same bound on the second term of the right-hand side of (G.2), it suffices to choose c_3 large enough and c_4 small enough to obtain the desired result.

G.4 Proof of Lemma C.3

We first write, using the definition of $\mathbf{S}(\omega)$,

$$\|\mathbf{S}(\widetilde{\omega}) - \mathbf{S}(\omega)\|_2 \leq \sum_{t \geq 1} \|\mathbf{R}_t(e^{-i2\pi\widetilde{\omega}t} - e^{-i2\pi\omega t}) + \mathbf{R}_{-t}(e^{i2\pi\widetilde{\omega}t} - e^{i2\pi\omega t})\|_2.$$

By using the same technique as used in the proof of Lemma C.2, we have

$$\|\mathbf{R}_t(e^{-i2\pi\widetilde{\omega}t} - e^{-i2\pi\omega t}) + \mathbf{R}_{-t}(e^{i2\pi\widetilde{\omega}t} - e^{i2\pi\omega t})\|_2 \leq 2\pi|\widetilde{\omega} - \omega| c_3 t e^{-c_4 t^{\gamma_1}},$$

and by setting c_5 equal to $2\pi c_3 \sum_{t \geq 1} t e^{-c_4 t^{\gamma_1}}$ we obtain the desired inequality.