HIGH DIMENSIONAL DATA ANALYSIS WITH DEPENDENCY AND UNDER LIMITED MEMORY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yiming Sun

December 2019

HIGH DIMENSIONAL DATA ANALYSIS WITH DEPENDENCY AND UNDER LIMITED MEMORY

Yiming Sun, Ph.D.

Cornell University 2019

Several methods for high dimensional analysis are proposed in this thesis under the condition that there are data dependency and limited memory. The first part of the work proposes a model free method for building networks for time series data when data are dependent from weakly stationary time series. We develop a thresholding based on methods to estimate multivariate spectral density under weakly sparsity assumption for high dimensional time series. Our theoretical analysis ensures that consistent estimations of spectral density matrix of a p-dimensional time series using n samples are possible under high-dimensional regime $\log p/n \to 0$ as long as the true spectral density is approximately sparse. A key technical component of our analysis is a new concentration inequality of average periodogram around its expectation, which is of independent interest. Our estimation consistency results complement existing results for shrinkage based estimators of multivariate spectral density, which require no assumption on sparsity but only ensure consistent estimation in a regime $p^2/n \to 0$. In addition, our proposed thresholding based estimators perform consistent and automatic edge selection when coherence networks among the components of a multivariate time series are learned. We demonstrate the advantages of our estimators using simulation studies and a real data application on functional connectivity analysis with fMRI data.

We further show that with a simple modification in the classic estimator, we can build a rigorous theory for adaptive thresholding in estimating multivariate spectral

density for Gaussian process. This adaptive estimator can capture the heterogeneity across different positions in spectral density matrix at a better convergence rate in comparison to the hard thresholding estimator.

The second part delves into compressing/analyzing high dimensional data with limited memory. We fixate on developing a streaming algorithm for Tucker Decomposition, generalization of singular value decomposition. The method applies a randomized linear map to the tensor to obtain a *sketch* that captures the important directions within each mode as well as the interactions among the modes. The sketch can be extracted from streaming or distributed data or with a single pass over the tensor which uses storage proportional to the degrees of freedom in the output Tucker approximation. Although the algorithm can exploit another view to compute a superior approximation, it does not require a second pass over the tensor. In conclusion, the paper provides a rigorous theoretical guarantee on elimination of the approximation error. Extensive numerical experiments show that the algorithm produces useful results that improve the state of the art for streaming Tucker decomposition.

Along the development of one-pass Tucke decomposition, we propose a memory efficient random mapping which we call Tensor random projection. We further study its theoretical property in application to several areas like random projection, sketching algorithms for fast computation for tensor regression.

BIOGRAPHICAL SKETCH

Yiming Sun was born in Yancheng Jiangsu, China. His general interest lies in interplay of machine learning and statistics. He is generally interested in any innovative things and enjoys discussing with any person who can provide him with different perspective. Yiming entered Cornell for PhD in statistics in 2016. After that he never stopped his steps to acquire the ability to implement end to end machine learning product while understanding statistical guarantees behind it.

To my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my sincere gratitude towards my PhD advisers, Sumanta Basu and Madeleine Udell for their guidance, time and commitment to my development towards an independent researcher. In particular, professor Udell's enthusiasm toward almost all novelties and energetic working style are an inspiration to me, while her ability to address numerous issues also leaves me a deep impression. I deeply appreciate the precious assistance my professor Udell offered who has quickly ushered me into new fields and provided me with insightful suggestions. It is professor Basu that imparts the necessity of cooperation with people of different backgrounds like economics and biology to me. I appreciate the precious assistance professor Basu offered who has taught me how to conduct research independently.

In addition, professor Michael Nussbaum and Yudong Chen always provide innovative and rigorous mathematical insights. I really appreciate that during the winter time, professor Nussbaum went through many results in spectral density estimation with me purely out of fervent devotion to researches. Every time talking with professor Yudong, I can learn many refreshing updates ranging from optimization, high dimensional statistics under various settings.

My heartfelt tribute would also be payed to the talented and intellectual people who have enlightened me during my PhD pursuit. I acquired the down to earth work ethics from Congzheng Song and Zhengqi Li, two PhDs in computer science who are my project partners. Ding Ma, who has already become an assistant professor, shared lots of her stories to assist me in facing uncertainties and anxiety in life. Lots of my old friends like Chengcheng Liu, Kuangyan Song always provides me with support.

During my several industrial interns, I was fortunate enough to be mentored by Olivia Simpson, Jessica Stauth, David Sargent etc.. What I appreciate the most is their

genuine collaborative and caring working style besides numerous expertise that they have embedded in me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LARGE SPECTRAL DENSITY MATRIX ESTIMATION BY THRESHOLDING

1.1 Introduction

Multivariate spectral density estimation is an important problem in time series and signal processing, with applications in many scientific disciplines including economics [Granger, 1969] and neuroscience [Bowyer, 2016]. Spectral density of a stationary multivariate time series is the frequency domain analogue of covariance and is based on the Fourier transform of autocovariance function. It aggregates information on linear association, both contemporaneous and across different lags, among the components of a multivariate time series. So it can be used to provide a richer description of cross-sectional dependency than Pearson correlation, which only accounts for contemporaneous association among the time series components.

In particular, multivariate spectral density and coherence (frequency domain analogue of correlation) are routinely used in neuroscience as metrics of functional connectivity among brain regions using time series of neurophysiological signals (e.g., fMRI, EEG and MEG) and to construct networks of interactions in a data-driven fashion [Bowyer, 2016]. These connectivity networks, where each node corresponds to a brain region and edge weights correspond to strengths of coherence between regions, are often used to study differential brain connectivity patterns in patients suffering from neurological disorders. More recently, coherence metrics have also been used to construct similarity measures when clustering high-dimensional time series of brain signals [Euan et al., 2016]. With advances in data collection and storage technologies, it is now

feasible to analyze time series data on a large number of brain regions. For instance, the freeSurfer brain atlas used in this paper summarizes voxel level data to $p = 86$ brain regions. Consequently, there is an increasing interest among neuroscientists in constructing coherence networks among a large number of brain regions in a principled manner from temporally dependent samples of small to moderate size ($n \ll p^2$). For instance, we use only $n = 200$ samples for our fMRI data analysis in this paper.

This recent interest in learning the cross-sectional dependence from spectral density matrix at different frequencies is complementary to developments in classical time series and signal processing literature, which focused more on studying the *shape* of spectral density function in a low-dimensional asymptotic regime (p fixed, $n \to \infty$) [Brillinger, 1981; Brockwell and Davis, 2013]. In another line of work, Dahlhaus and Eichler [2003]; Dahlhaus et al. [1997]; Eichler [2007] investigated in depth the issues of inference with coherence and testing of marginal independence between components of multivariate time series using integrated spectral density. Finer and uniform convergence rates of smoothed periodograms were more recently provided by Wu and Zaffaroni [2015]. However, as the dimension of the time series increases, so does the estimation risk of smoothed periodograms. This was first pointed out by Böhm and von Sachs [2009], who showed that shrinking smoothed periodogram towards a simpler structure can reduce risk and make the estimates better-conditioned for studying inverse spectral density matrix. The authors also proved consistency of their estimates under a double-asymptotic regime $p \to \infty, n \to \infty, p^2/n \to 0$. In a series of papers, Böhm and Von Sachs [2008]; Fiecas and Ombao [2016]; Fiecas and von Sachs [2014] have made significant progress in this direction by providing a wide variety of shrinkage methods with attractive theoretical and empirical properties.

In this work, we make two additions to this research direction of learning large spe-

tral density matrices. First, we propose a family of *sparsity regularized estimators* of spectral density matrix based on thresholding averaged periodograms. Our proposed estimators have the added advantage of performing automatic edge selection and providing sparse, interpretable networks among the component time series. Second, we develop a non-asymptotic theory for estimation of spectral density and coherence that explicitly connects estimation error bounds to a notion of approximate sparsity of the true spectrum. As a consequence, our theory shows that consistent estimation is possible in a high-dimensional regime $\log p/n \to 0$ as long as the underlying structure is approximately sparse.

Our proposal is motivated by recent developments in covariance matrix estimation literature, where several thresholding based strategies [Bickel and Levina, 2008; Cai et al., 2016; Cai and Liu, 2011; Rothman et al., 2009] have shown to provide good theoretical and empirical properties compared to the shrinkage based estimators proposed in Ledoit and Wolf [2004]. The thresholding techniques developed in this literature serve as promising candidates for high-dimensional spectral density estimation as well. However, their implementation and theoretical analysis require addressing additional technical challenges. From an implementation consideration, choice of threshold in covariance matrix estimation for i.i.d. data is carried out using multiple sample-splitting [Bickel and Levina, 2008] which is not feasible when the data have a temporal ordering. On the theoretical side, non-asymptotic analysis of periodograms averaged across nearby frequencies requires understanding concentration behavior of a sum of random matrices that are *neither independent nor identically distributed*. Unlike sample covariance estimation with i.i.d. data, the lack of identical distribution results in smoothing bias well-known in nonparametric density estimation. In addition, the additional temporal dependence complicates deriving finite sample deviation of averaged

periodogram from its expectation.

We make three technical contributions in this paper to address the above challenges. First, we select thresholding parameters using a frequency-domain sample-splitting scheme based on the heuristic of approximate independence of periodograms at different Fourier frequencies. Second, we provide upper bounds on the finite sample bias of averaged periodograms and provide insight into how it is affected by temporal dependence in data for some commonly used families of time series. Finally, we develop a non-asymptotic upper bound on the deviation of averaged periodogram using a Hanson-Wright type inequality for complex quadratic forms of temporally dependent random vectors. Building upon these technical ingredients, our main theoretical results include (i) consistency of thresholded averaged periodograms in operator and scaled Frobenius norms in a high-dimensional regime under a weak sparsity assumption on true spectrum, and (ii) sparsistency results ensuring selection of marginally correlated pairs of time series in a coherence network with high probability. Our analysis framework accommodates Gaussian time series, and linear processes with subGaussian or generalized subexponential errors, or errors with finite fourth moments. The rates of convergence of thresholded estimators change with the nature of tail distribution of errors.

We demonstrate the merits of our proposed methods using extensive numerical experiments and a real data application on constructing functional connectivity networks from fMRI data. Our numerical experiments show that thresholding methods achieve estimation accuracy comparable with the shrinkage method, while simultaneously performing automatic coherence selection. In particular, a lasso and an adaptive lasso based thresholding strategy show promising performance across different simulation settings. In the real data application, these two methods were able to extract sparse, in-

terpretable networks that nicely captured known biological patterns in brain networks and distinguished different brain regions from each other.

The rest of the paper is organized as follows. In section 2.2, we formally introduce our problem, provide a brief review of shrinkage estimators, and describe our proposed thresholding methods. In section 3.5, we derive non-asymptotic upper bounds on our proposed spectral density estimates for Gaussian time series. In section 1.4 we extend the results for Gaussian time series to general linear processes with different non-Gaussian noise distributions. In section 4.4, we conduct simulation studies to assess the finite sample properties of our proposed estimators. Section 1.6 contains an empirical application of our proposed method to a functional connectivity analysis with real fMRI data. We defer the proofs of all of our technical results to the Appendix.

Notation. Throughout this paper, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of integers, real numbers and complex numbers, respectively. We use $|c|$ to denote the modulus of a complex number and the absolute value of a real number. We use $\|v\|$ to denote ℓ_2-norm of a vector v. For a matrix A, $\|A\|_1$, $\|A\|_\infty$, $\|A\|$ and $\|A\|_F$ will denote maximum complex modulus column sum norm, maximum complex modulus row sum norm, spectral norm $\sqrt{\Lambda_{\max}(A^\dagger A)}$ and Frobenius norm $\sqrt{\text{tr}(A^\dagger A)}$, respectively, where $A^\dagger$ is conjugate transpose of A. We also let $\lambda_{\max}(A)$ denote the spectral radius of a $n \times n$ matrix A, i.e., $\lambda_{\max}(A) = \max(|\lambda_1|, \cdots |\lambda_n|)$, where λ_i are the eigenvalues of matrix A. If A is symmetric or Hermitian, we denote its maximum and minimum eigenvalues by $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$. We use e_i to denote the i^{th} unit vector in $\mathbb{R}^p$, for $i = 1, 2, \ldots, p$. For vectors $v_i \in \mathbb{R}^p, i = 1, \ldots, n$, we use $[v_1 : \ldots : v_n]$ to denote the $p \times n$ matrix formed by horizontally stacking these column vectors v_i, and $[v_1^\top; \ldots; v_n^\top]$ to denote the $n \times p$ matrix by vertically stacking row vectors $v_i^\top$. Let $vec(A)$ represent the vector got from vectorization of a matrix A by stacking all its columns. We use $rk(A)$ to

denote the rank of a matrix A. For a complex vector $v \in \mathbb{C}^p$ and any $q > 0$, we define $\|v\|_q := (\sum_{i=1}^p |v_i|^q)^{1/q}$. We use $\|v\|_0$ to denote the number of non-zero elements in v. Note that when $0 \leqslant q < 1$, it is not really a norm since triangle inequality does not hold, but we keep the notation of a norm for convenience . Then we define the induced matrix norm, $\|A\|_{\alpha,\beta} = \sup_{x \neq 0} \|Ax\|_\alpha / \|x\|_\beta$, for any $\alpha > 0, \beta > 0$. We will also use $\|A\|_\alpha$ to denote the induced norm $\|A\|_{\alpha,\alpha}$ for any $\alpha > 0$ and any complex matrix $A \in \mathbb{C}^{p \times p}$. Also, to be succinct, we use $\|A\|_{\max} := \max_{r,s} |A_{rs}|$. Throughout the paper, we write $A \gtrsim B$ if there exists a universal constant $c > 0$, not depending on model dimension or any model parameters, such that $A \geqslant cB$. We use $A \asymp B$ to denote $A \gtrsim B$ and $B \gtrsim A$.

1.2 Background and Methods

Consider a p-dimensional weakly stationary real-valued time series $X_t = (X_{t1}, \ldots, X_{tp})^\top$, $t \in \mathbb{Z}$. Let $\mathcal{X} = [X_1 : \ldots : X_n]^\top$ be the *data matrix* containing n consecutive observations from the time series $\{X_t\}$ in its rows. We assume $\mathbb{E}X_t = 0,\ t = 1, \ldots, n$ for ease of exposition. In practice, multivariate time series are often de-meaned before performing correlation based analysis. Weak stationarity implies that $\text{Cov}(X_t, X_{t-\ell}) = \mathbb{E}X_t X_{t-\ell}^\top$ only depends on ℓ, so we can define autocovariance as function of the lag ℓ, viz., $\Gamma(\ell) = \text{Cov}(X_t, X_{t-\ell})$. Spectral density aggregates information of autocovariance of different lag orders ℓ at a specific frequency $\omega \in [-\pi, \pi]$ as

$$f(\omega) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\omega}. \tag{1.2.1}$$

Note that the autocovariance functions of different lags can be recovered from the spectral density using the transformation $\Gamma(\ell) = \int_{-\pi}^{\pi} f(\omega) e^{i\ell\omega} d\omega$, for any $\ell \in \mathbb{Z}$.

For the matrix-valued spectral density function f over $[-\pi, \pi]$, we define, for $q \geqslant 0$,

$$|||f|||_q = \operatorname*{ess\,sup}_{\omega \in [-\pi,\pi]} \|f(\omega)\|_q.$$

Following Basu and Michailidis [2015], we will also use $|||f||| := |||f|||_2 = \operatorname{ess\,sup}_{\omega\in[-\pi,\pi]} \|f(\omega)\|$ as a measure of stability of the time series X_t. Larger values of $|||f|||$ are associated with processes having stronger temporal and cross-sectional dependence and less stability. Since every coordinate of the spectral density matrix is calculated using at most two components of the p-dimensional time series X_t and $f(\omega)$ is non-negative definite, a smaller measure of stability, viz. $\max_{1\leqslant r\leqslant p} \operatorname{ess\,sup}_\omega \|f_{rr}(\omega)\|$ can be also used in our error bound analysis instead, although we present our results in terms of $|||f|||$ for ease of exposition.

In many applications, in particular functional connectivity analyses in neuroscience, it is of interest to estimate standardized spectral density or coherence matrix, an analogue of correlation in the frequency domain, defined as

$$g_{rs}(\omega) = \frac{f_{rs}(\omega)}{\sqrt{f_{rr}(\omega) f_{ss}(\omega)}}, \tag{1.2.2}$$

assuming $f_{rr}(\omega) \neq 0$ for all $1 \leqslant r \leqslant p$.

1.2.1 Background: Periodogram Smoothing and Shrinkage

The classical estimate of spectral density is based on the periodogram [Brockwell and Davis, 2013; Rosenblatt, 1985] defined as

$$I(\omega) = \sum_{|\ell|<n} \hat{\Gamma}(\ell) e^{-i\ell\omega}, \tag{1.2.3}$$

where $\hat{\Gamma}(\ell) = n^{-1} \sum_{t=\ell+1}^{n} X_t X_{t-\ell}^\top$ for $\ell \geqslant 0$, and $\hat{\Gamma}(\ell) = n^{-1} \sum_{t=1}^{n+\ell} X_t X_{t-\ell}^\top$ for $\ell < 0$. Note the connection between periodogram and discrete Fourier transformation (DFT)

Input: j, m, N, periodograms at Fourier frequency $\{I(\omega_k)\}_{k\in F_n}$, finite grid of thresholds $\mathcal{L}$

for $\lambda \in \mathcal{L}$ **do**

  for $\nu \leftarrow 1$ ***to*** N **do**

    Randomly divide $\{j-m, \ldots, j, \ldots, j+m\}$ into two subsets J_1 and J_2 such that $||J_1| - |J_2|| \leqslant 1$ and for any $k \in F_n$, $k \in J_1$ iff $-k \in J_1$

    $\hat{f}_{1,\nu}(\omega_j) \leftarrow \sum_{k\in J_1} I(\omega_k), \quad \hat{f}_{2,\nu}(\omega_j) \leftarrow \sum_{k\in J_2} I(\omega_k)$

    $\hat{R}_\nu(\omega_j, \lambda) \leftarrow \left\| T_\lambda(\hat{f}_{1,\nu}(\omega_j)) - \hat{f}_{2,\nu}(\omega_j) \right\|_F^2$

  end

  $\hat{R}(\omega_j, \lambda) \leftarrow \sum_{\nu=1}^{N} \hat{R}_\nu(\omega_j, \lambda)/N$

end

Output: $\hat{\lambda}_j := \hat{\lambda}(\omega_j) = \operatorname{argmin}_{\lambda\in\mathcal{L}} \hat{R}(\omega_j, \lambda)$

Algorithm 1: Threshold Selection by Frequency Domain Sample-splitting

$d(\omega) = \mathcal{X}^\top(C(\omega) - iS(\omega))$, where

$$\begin{aligned} C(\omega) &= \frac{1}{\sqrt{n}}(1, \cos\omega, \ldots, \cos(n-1)\omega)^\top, \\ S(\omega) &= \frac{1}{\sqrt{n}}(1, \sin\omega, \ldots, \sin(n-1)\omega)^\top. \end{aligned} \tag{1.2.4}$$

We can rewrite $I(\omega)$ as $d(\omega)d(\omega)^\dagger$. In classical asymptotic analysis of time series (p fixed, $n \to \infty$), it is known that $\frac{1}{2\pi}I(\omega)$ is asymptotically unbiased for $f(\omega)$ but not consistent due to non-diminishing variance. For instance, for i.i.d Gaussian white noise $X_t \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2 I)$, the variance of $I(\omega)$ is of the order σ^4 [Proposition 10.3.2, Brockwell and Davis [2013]]. To achieve consistency, it is common to resort to smoothing periodograms over nearby frequencies. In this paper, we focus on the simplest form of smoothing, viz. averaging, of periodograms

$$\hat{f}(\omega; m) = \frac{1}{2\pi(2m+1)} \sum_{|k|\leqslant m} I(\omega + \omega_k), \tag{1.2.5}$$

where $\omega_k = 2\pi k/n,\ k \in F_n$, the set of Fourier frequencies. To be precise, F_n denotes the set $\left\{-[\frac{n-1}{2}], \ldots, [\frac{n}{2}]\right\}$ where $[x]$ is the integer part of x. F_n contains exactly the same frequencies used to calculate discrete Fourier transformation. It is common to

evaluate the periodogram at these Fourier frequencies, in which case the smoothing periodogram in (2.1.5) becomes

$$\hat{f}(\omega_j; m) = \frac{1}{2\pi(2m+1)} \sum_{|k| \leqslant m} I(\omega_{j+k}). \tag{1.2.6}$$

Note that even though the values of $j + k$ can fall outside F_n, it is enough to evaluate periodograms at Fourier frequencies F_n since $I(\omega)$ is 2π-periodic in ω. Theorem 10.4.1 in Brockwell and Davis [2013] shows that if $m = o(\sqrt{n})$, (2.1.6) is a consistent estimator. As in general nonparametric function estimation, one can replace the weights $1/(2m+1)$ in (2.1.6) by a more general kernel function. For more details, we refer the readers to Brockwell and Davis [2013]. To make notations simpler, in this paper we will omit the subscript m and use $\hat{f}(\omega_j)$ whenever m is clear from the context.

This nonparametric smoothing method can be unstable for high-dimensional multivariate spectral density estimation since smoothed periodograms start to become ill-conditioned. Generalizing shrinkage estimation strategy for high-dimensional covariance matrix [Ledoit and Wolf, 2004], Böhm and von Sachs [2009] proposed shrinking averaged periodogram to estimate spectral density in high-dimension. The idea of shrinkage method is to reduce condition numbers for smoothed periodograms. In particular, the authors changed the estimation target to $f^0(\omega) = \mathbb{E}\hat{f}(\omega)$ and argued that $f^0(\omega)$ is close enough to $f(\omega)$ asymptotically. Subsequently, they considered a Hilbert space for square complex random matrices with inner product defined as $\mathbb{E}\langle A, B\rangle$ where A, B are two matrices and

$$\langle A, B\rangle = \frac{1}{p}\text{tr}(A^\dagger B).$$

In this Hilbert space and with the fact that $\hat{f}(\omega)$ is an unbiased estimator for $f^0(\omega)$, Böhm and von Sachs [2009] applied the projection argument similar to Ledoit and Wolf [2004] to build the shrinkage estimator for $f^0(\omega)$. To this end, the authors first

projected $f^0(\omega)$ on the space spanned by the identity matrix as $\mu(\omega)I_p$, where I_p is the identity matrix and $\mu(\omega) = \frac{1}{p}\text{tr}(f(\omega))$. Then the shrinkage estimator is defined as the minimizer of the convex program

$$\hat{f}^\star(\omega) = \underset{\tilde{f}(\omega)\in S(\omega)}{\text{argmin}} \frac{1}{p}\|f^0(\omega) - \tilde{f}(\omega)\|_F^2,$$

where

$$S(\omega) = \rho(\omega)\mu(\omega)I_p + (1-\rho(\omega))\hat{f}(\omega), \quad 0 \leqslant \rho(\omega) \leqslant 1.$$

Böhm and von Sachs [2009] derived an explicit formula $\rho(\omega) = \alpha^2(\omega)/\delta^2(\omega)$, where

$$\alpha^2(\omega) = \frac{1}{p}\|f^0(\omega) - \mu(\omega)I_p\|_F^2, \;\; \beta^2(\omega) = \frac{1}{p}\|f^0(\omega) - \hat{f}(\omega)\|_F^2,$$

and $\delta^2(\omega) = \alpha^2(\omega) + \beta^2(\omega)$. Then they plugged in estimators of $\alpha(\omega), \beta(\omega), \delta(\omega)$ into the above formula to get the final data-driven estimator of spectral density.

1.2.2 Method: Thresholding Averaged Periodogram

In this section, we present our proposed thresholding estimators. We restrict our methodology description and theoretical development on the finite grid of Fourier frequencies for convenience, although all our theoretical results hold for any arbitrary frequency $\omega \in [-\pi, \pi]$. We briefly explain why all theoretical developments still hold for thresholding on smoothed periodograms at a general frequency defined in (2.1.5). The key property we used to develop error bound analysis for thresholding estimators is orthogonality of $d(\omega_j), j \in F_n$. For general frequency ω, we can show that $d(\omega + \omega_j), j = -m, \cdots, 0, \cdots, m$, are also orthogonal to each other. Based on this property, we could follow all arguments for Fourier frequencies to achieve the same theoretical results.

We propose *hard thresholding* of averaged periodograms, i.e.,

$$T_\lambda(\hat{f}_{rs}(\omega_j)) = \begin{cases} \hat{f}_{rs}(\omega_j) & \text{if } |\hat{f}_{rs}(\omega_j)| \geqslant \lambda \\ 0 & \text{if } |\hat{f}_{rs}(\omega_j)| < \lambda, \end{cases} \tag{1.2.7}$$

where $\lambda > 0$ is a threshold chosen by the user, and can potentially be a frequency dependent number λ_j. $T_\lambda(\cdot)$ is a thresholding operator on spectral density, $T_\lambda(\hat{f}_{rs}(\omega_j))$ represents the $(r,s)^{th}$ element of the thresholded matrix, where $1 \leqslant r, s \leqslant p$. For notational convenience, we will often use $\hat{f}_{\lambda,rs}(\cdot)$ instead of $T_\lambda(\hat{f}_{rs}(\cdot))$.

Following Rothman et al. [2009], we also propose a variety of *generalized thresholding operators* $S_\lambda(.)$ that combine the benefits of shrinkage and thresholding. In particular, we consider elememt-wise shrinkage operator $S_\lambda(.)$ satisfying the following three conditions for any $z \in \mathbb{C}$:

(1) $|S_\lambda(z)| \leqslant |z|$,

(2) $S_\lambda(z) = 0$ if $|z| \leqslant \lambda$,

(3) $|S_\lambda(z) - z| \leqslant \lambda$.

Similar to hard thresholding $T_\lambda(.)$, we apply this operator to individual elements of averaged periodogram. It turns out conditions (1)-(3) are satisfied by a number of thresholding and shrinkage procedures. In particular, the hard thresholding operator $T_\lambda(.)$ satisfies these conditions. In addition, generalizing Rothman et al. [2009] to the case of complex variables, we propose a soft thresholding (lasso) operator

$$S^s_\lambda(z) = \frac{z}{|z|}\left(|z| - \lambda\right)_+, \quad z \in \mathbb{C},$$

and adaptive lasso operator

$$S^{\text{AL}}_\lambda = \frac{z}{|z|}\left(|z| - \lambda^{(\eta+1)}|z|^{-\eta}\right)_+, \quad z \in \mathbb{C}.$$

Our proposed hard and soft thresholding procedures require selection of two tuning parameters: (i) smoothing span m and (ii) level of threshold λ. In Section 3.5, we provide a detailed discussion of the theoretical choices of these parameters that ensure consistent estimation in high-dimensional regime. In the next subsection 1.2.3, we discuss how to choose these two parameters in a data-driven fashion. The adaptive lasso based soft thresholding method has a third tuning parameter η. In our numerical and real data analyses, we set $\eta = 2$ following the suggestion of Rothman et al. [2009], although a more general sample-splitting based choice along the line of Algorithm 2 can be adopted in practice.

When the thresholded spectral density matrices are sparse, they can be used to construct networks to visualize and analyze marginal dependence relationships among the component time series. However, just like thresholded covariance matrix estimators, thresholding individual entries does not necessarily ensure that the thresholded spectral density matrix estimate is positive definite. Our operator norm consistency results in Section 3.5 implies that as long as the true spectral density is positive definite and the sample size is large enough, the thresholded estimate is positive definite with high probability. However, in finite sample, this is a limitation since the estimates cannot be directly used to calculate inverse spectral density and partial coherence. On the other hand, regularization is required to calculate inverse spectral density in high-dimension, and a more principled approach along the line of graphical lasso can be used to directly regularize entries of the inverse spectral density [Jung, 2015; Jung et al., 2015]. We expect that the key concentration inequalities developed in our analysis will be useful in the estimation of inverse spectral density as well.

1.2.3 Choice of Tuning Parameters

At any Fourier frequency ω_j, we need to choose two tuning parameters for our method - (i) the smoothing span $2m+1$, and (ii) the threshold level λ. In this work, we select a single smoothing span for all the frequencies, but choose the threshold level separately for each frequency.

The smoothing span plays the role of "effective sample size" in estimating $f(\omega_j)$. Recall that

$$\hat{f}(\omega_j; m) = \frac{1}{2\pi(2m+1)} \sum_{|k| \leqslant m} I(\omega_{j+k}).$$

In classical asymptotics (p fixed, $n \to \infty$) and the Kolmogorov asymptotics ($p \to \infty, n \to \infty, p^2/n \to 0$) [Böhm and von Sachs, 2009; Brockwell and Davis, 2013], it is shown that for $\hat{f}(\omega_j; m)$ to be consistent, m (depending on n) must go to infinity and $m/n \to 0$ as $n \to \infty$. Our non-asymptotic analysis in Section 3.5 suggests that $m/[n\Omega_n(f)] \to 0$, where $\Omega_n(f)$, defined as $\max_{r,s} \sum_{\ell=-n}^{n} |\ell||\Gamma_{rs}(\ell)|$ is a measure of temporal dependence in the time series. For our numerical and real data applications, we choose m in the order of $\sqrt{n}$, with smaller values of m for processes with stronger temporal dependence and larger $\Omega_n(f)$. A more data-driven approach along the line of Ombao et al. [2001] and Fiecas and von Sachs [2014] can be designed with suitable modification to account for high-dimensionality, although we do not pursue this direction in this work.

The second tuning parameter is the threshold value. Unlike the shrinkage estimators of spectral density matrices, finding asymptotically optimal plug-in estimators for threshold level is challenging due to the non-smooth nature of thresholding operators. For covariance estimation from i.i.d. data using thresholding, a sample-splitting method proposed in Bickel and Levina [2008] or its variants are normally employed.

In this method, the entire sample is split into two sub-samples, and the Frobenius norm difference between thresholded estimation in one sub-sample and regular sample covariance in the other sub-sample is compared for different levels of threshold. The entire exercise is repeated N times and the level of threshold minimizing the average Frobenius norm difference is selected as the threshold.

This approach is not directly amenable to spectral density estimation since for any two given sub-sample sizes, only $N = 1$ split is possible maintaining the temporal ordering. However, the periodograms at different positive Fourier frequencies $\omega_j \in F_n, \omega_j \geqslant 0$, are asymptotically independent. This suggests an analogous sample-splitting algorithm can be designed in the frequency domain. With this heuristic, we propose the following algorithm.

For each frequency $j \in \{1, \cdots, [n/2]\}$, we randomly split the periodograms in $\{j - m, \cdots, j + m\}$ into two sub-samples J_1, J_2 of size m_1 and m_2, where $|m_1 - m_2| \leqslant 1$. Since $I(\omega_{-k}) = I(\omega_k)$, we keep $I(\omega_k)$ and $I(-\omega_k)$ in the same sub-sample. Then, for every λ on a finite grid of possible threshold choices $\mathcal{L}$, we calculate the squared Frobenius norm of the difference between thresholded averaged periodogram on J_1, viz., $\hat{f}_1(\omega_j)$, and averaged periodogram $\hat{f}_2(\omega_j)$ on J_2. This exercise is repeated N times and the threshold $\lambda \in \mathcal{L}$ minimizing squared Frobenius norm is selected as $\hat{\lambda}_j$ for frequency ω_j. A complete description is provided in Algorithm 2.

1.3 Theoretical Properties

In this section, we analyze asymptotic properties of thresholded averaged periodograms under high-dimensional regime. In particular, we derive non-asymptotic upper bound

on the estimation error under operator and Frobenius norms and relate them to a notion of weak sparsity of the spectral density matrices. A key technical ingredient of our analysis is a concentration inequality of complex quadratic forms of temporally dependent Gaussian random vectors. In section 1.4 we extend these results to linear processes with more general noise distributions, including subGaussian and subexponential families.

In contrast with classical asymptotic framework where p is fixed and $n \to \infty$, a non-asymptotic analysis for high-dimensional time series requires careful quantification of the convergence rates, in particular how they are affected by cross-sectional and temporal dependence inherent in the time series. Therefore, before proceeding with the main theoretical results, we describe parameters of the multivariate time series X_t that appears in our estimation error bounds.

Weak Sparsity of Spectral Density: In order to make meaningful estimation in a high-dimensional regime, we focus on a class of spectral density matrices with suitable low-dimensional structure of *weak sparsity* measured by $|||f|||_q$ for some $0 \leqslant q < 1$. Matrices with small $|||f|||_0$ are *exactly sparse*, while small $|||f|||_q$ correspond to matrices within a small ℓ_q ball in $\mathbb{C}^{p\times p}$. Weak sparsity of regression coefficients and covariance matrices have been proposed earlier in van de Geer [2016] and Bickel and Levina [2008] respectively. Weakly sparse covariance matrices have been applied to climate studies according to Cai et al. [2016] and gene expression array analysis, as mentioned in Cai and Zhou [2012].

Although the induced norm defined in notation section does not satisfy triangle inequality for $0 \leqslant q < 1$, $\|A\|_q^q$ satisfies the triangle inequality leading to

$$\max_{s=1}^{p} \sum_{s=1}^{p} |f_{rs}(\omega)|^q = \|f(\omega)\|_q^q \leqslant |||f|||_q^q,$$

where $|||f|||_q = \operatorname{ess\,sup}_{\omega\in[-\pi,\pi]} \|f(\omega)\|_q$ as defined before. We provide a proof of this

statement in lemma 1.C.1. Since spectral density $f(\omega)$ is a Hermitian matrix, $|||f|||_q^q$ also measures the row weak sparsity. This weakly sparse class covers a variety of sparse patterns as shown in Bickel and Levina [2008].

Strength of Temporal and Cross-sectional Dependence: The decay rates of the strengths of cross- and autocorrelation between components of X_t capture the strength of temporal and cross-sectional dependence in data, which in turn relates to the effective sample size and appear in our error bounds. For meaningful estimation, we restrict ourselves to the class of short-range dependent time series X_t with the following summability assumption on its underlying autocovariance function $\Gamma(\ell)$:

Assumption 1.3.1. $\sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\|_{max} < \infty$.

Under this assumption, we will present our bounds in terms of three quantities. The first one is $|||f|||$ defined before, and will be used to assess the *concentration of averaged periodogram around its expectation*. Note that $|||f|||$ is finite since

$$\|f(\omega)\| = \left\| \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\omega\ell} \right\| \leqslant \sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\| \leqslant \sum_{\ell=-\infty}^{\infty} p\|\Gamma(\ell)\|_{\max}. \tag{1.3.1}$$

The other two quantities that capture the strength of temporal and contemporaneous dependence in the multivariate time series $\{X_t\}_{t\in\mathbb{Z}}$ are

$$\Omega_n(f) = \max_{1\leqslant r,s\leqslant p} \sum_{\ell=-n}^{n} |\ell||\Gamma_{rs}(\ell)|, \quad L_n(f) = \max_{1\leqslant r,s\leqslant p} \sum_{|\ell|>n} |\Gamma_{rs}(\ell)|. \tag{1.3.2}$$

Together, these two quantities help assess how the *bias of averaged periodogram* depends on the the degree of decay of the autocovariance function with increasing lag order ℓ. Under Assumption 2.1.1, both of these quantities are finite. In Proposition 1.3.4, we show how these quantities grow for some common classes of multivariate time series.

1.3.1 Estimation Consistency: Stable Gaussian Time Series

We start with a key technical ingredient of our analysis, a Hanson-Wright type inequality [Rudelson and Vershynin, 2013] for quadratic forms of random vectors generated by a multivariate Gaussian time series. This result generalizes Proposition 2.4 in Basu and Michailidis [2015] by allowing an arbitrary matrix A in the quadratic form. In Section 1.4, we extend this inequality to accommodate more general non-Gaussian time series.

Our modified Hanson-Wright inequality is crucial for understanding the concentration behaviour of averaged periodograms around the true spectral density $\left|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)\right|$, for a fixed coordinate (r, s) of the $p \times p$ spectral density matrix. This deviation is required for selecting threshold λ that ensures consistency in high-dimension. Unlike high-dimensional covariance estimation problem where sample covariance is an unbiased estimator of population covariance, the averaged periodogram at frequency ω_j is a biased estimator of $f(\omega_j)$. This requires developing upper bounds on both the "bias" and "variance" terms in the deviation of $\hat{f}_{rs}$ around f_{rs}:

$$\left|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)\right| \leqslant \left|\mathbb{E}\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)\right| + \left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right|.$$

Note that while the first term above is indeed capturing bias of $\hat{f}_{rs}(\omega_j)$, the second term is not technically "variance" since this is the centered version of $\hat{f}_{rs}(\omega_j)$ and not its L_2 norm. Nevertheless, we continue to use the term 'variance' in this context since it captures the fluctuation of $\hat{f}_{rs}(\omega_j)$ around its expectation. The upper bounds on bias and variance terms are obtained in Propositions 1.3.3 and 1.3.5, respectively. Finally, in Proposition 2.3.7 we extend the deviation bound on a single (r, s) to all p^2 elements of $f(\omega_j)$ and provide a non-asymptotic upper bound on the estimation error of the hard-thresholded averaged periodogram.

Lemma 1.3.2. *Suppose* $\mathcal{X}_{n\times p} = [X_1 : \ldots : X_n]^\top$ *is a data matrix from a stable*

Gaussian time series X_t satisfying Assumption 2.1.1. Then there exists a universal constant $c > 0$ such that for any $\eta > 0$ and any $p \times p$ real matrix A,

$$\mathbb{P}\left(\left|vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top) - \mathbb{E}\left[vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top)\right]\right| > 2\pi\eta |||f|||\right)$$
$$\leqslant 2\exp\left[-c\min\left\{\frac{\eta}{\|A\|}, \frac{\eta^2}{rk(A)\|A\|^2}\right\}\right].$$

For Gaussian $\mathcal{X}$, the above lemma generalizes Hanson-Wright inequality by allowing dependence among the entries of $\mathcal{X}$, and controlling the effect of dependence in the tail bound using $|||f|||$, $\|A\|$ and rk(A). As will be evident from our analysis, this simple generalization will be immensely useful for studying concentration behaviour of averaged periodogram around the true spectral density in appropriate norms. Note that we replace $\|A\|_F^2$ in standard Hanson-Wright inequality by a larger quantity $rk(A)\|A\|^2$, which makes the presentation easier in the asymptotic regime of our interest. In a lower dimensional regime, it is possible to get sharper rate using $\|A\|_F^2$ and $\int_{[-\pi,\pi]} \|f(\omega)\|^2 d\omega$ instead of $|||f|||$, as discussed in Basu and Michailidis [2015].

Bound on Bias Term: In low-dimensional asymptotic regime (p fixed, $n \to \infty$) the bias term is asymptotically negligible. In the double-asymptotic analysis of [Böhm and von Sachs, 2009] as well, the authors claim the bias of the estimator i.e., $|\mathbb{E}\hat{f}(\omega_j) - f(\omega)| = o(\frac{m}{n})$ which is negligible. In our non-asymptotic analysis, we need to derive an upper bound for this bias term in terms of $\{\Gamma(\ell)\}_{\ell\in\mathbb{Z}}$, since the choice of threshold λ depends crucially on this. The following proposition establishes such an upper bound in terms of the temporal dependence present in the multivariate time series X_t.

Proposition 1.3.3. *For any coordinate (r, s) with $1 \leqslant r, s \leqslant p$ and any Fourier frequency ω_j, $j \in F_n$, the estimation bias of averaged periodogram with a smoothing span*

$2m + 1$ *satisfies*

$$\left|\mathbb{E}\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)\right| \leqslant \frac{m + 1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f).$$

A consequence of this proposition is that it shows $m/[n/(\Omega_n(f)] \to 0$ is sufficient to ensure bias vanishes asymptotically. In particular, for two p-dimensional time series and same sample size n, it suggests choosing a smaller m for the series with stronger temporal dependence (larger $\Omega_n(f)$) since the effective sample size after accounting for dependence ($n/\Omega_n(f)$) is smaller.

We defer its proof to Appendix A. The upper bound on the bias depends on two terms: $\Omega_n(f)$ and $L_n(f)$. In previous works Böhm and Von Sachs [2008]; Böhm and von Sachs [2009], authors argue that this upper bound on bias is of the order $\mathcal{O}(m/n)$. But since we focus on non-asymptotic analysis, these two terms $\Omega_n(f)$ and $L_n(f)$ appear in the choices of our two tuning parameters: threshold λ and the smoothing span $2m + 1$. To ensure we choose these parameters appropriately so that the bias vanishes asymptotically under a high-dimensional regime, it is important to understand how the above quantities grow with sample size n. Our next proposition provides some upper bounds on these quantities under three different conditions. The first one is assuming a geometric decay rate on $\|\Gamma(\ell)\|_{\max}$, second one is about ρ-mixing condition (equivalent to strongly mixing for stationary Gaussian processes [Bradley, 2005]) and VAR processes. Before that, we briefly review definition of ρ mixing for condition 2 in Proposition 1.3.4 and VAR process for condition 3 in Proposition 1.3.4.

Bradley [2005] provides a good summary of various mixing conditions. Here we introduce the definition for ρ mixing: for two σ-algebras $\mathcal{A}$ and $\mathcal{B}$, we define

$$\rho(\mathcal{A}, \mathcal{B}) = \sup |\text{Corr}(f, g)|, \quad f \in L^2(\mathcal{A}), g \in L^2(\mathcal{B}),$$

where f, g are two measurable functions with respect to σ-algebras $\mathcal{A}$ and $\mathcal{B}$ respectively. For stationary multivariate time series X_t, we define the ρ-mixing coefficient for gap ℓ as

$$\rho(\ell) = \rho(\sigma(X_t, t \leqslant 0), \sigma(X_t, t \geqslant \ell)). \tag{1.3.3}$$

The two characteristics $\|\Gamma(\ell)\|_{\max}$ and $\rho(\ell)$ are usually easy to describe for finite order VMA and $\mathrm{VAR}(1)$ model. For $\mathrm{VAR(d)}$ with $d > 1$, however, it is more complicated. It is well known that we can rewrite a VAR(d) model

$$X_t = \sum_{\ell=1}^{d} A_\ell X_{t-\ell} + \varepsilon_t,$$

as a VAR(1) model $\tilde{X}_t = \tilde{A}_1 \tilde{X}_{t-1} + \tilde{\varepsilon}_t$, where

$$\tilde{X}_t = \begin{bmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-d+1} \end{bmatrix}_{dp\times 1} \quad \tilde{A}_1 = \begin{bmatrix} A_1 & A_2 & \cdots & A_{d-1} & A_d \\ I_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_p & \mathbf{0} \end{bmatrix}_{dp\times dp} \quad \tilde{\varepsilon}_t = \begin{bmatrix} \varepsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}_{dp\times 1}.$$

The sufficient and necessary condition for X_t being stationary is that $\lambda_{\max}(\tilde{A}_1) < 1$. As we will discuss later that first two conditions in Proposition 1.3.4 could be achieved by assuming coefficients has operator norm less than 1 for VAR(1) model. But for VAR(d) with $d > 1$, it is known that $\|\tilde{A}_1\| \geqslant 1$ [Basu and Michailidis, 2015]. So we cannot directly verify the geometric decay conditions 1 and 2 in Proposition 1.3.4. But we can still get some compact bound by assuming $\tilde{A}_1$ is diagonalizable. Note that the assumption of diagonalizability is not stringent since we can add a sufficiently small perturbation to the entries of $\tilde{A}_1$ so that its eigenvalues are distinct and we still have $\lambda_{\max}(\tilde{A}_1) < 1$. We make this statement precise in Lemma 1.C.2 in the Appendix.

Proposition 1.3.4. *Consider a weakly stationary, centered time series X_t.*

1 *Suppose* X_t *satisfies* $\|\Gamma(\ell)\|_{\max} \leqslant \sigma_X \rho_X^{|\ell|}$ *for all* $\ell \in \mathbb{Z}$ *for some* $\sigma_X > 0$ *and* $\rho_X \in (0,1)$. *Then*

$$\Omega_n \leqslant 2\sigma_X \rho_X \left[\frac{1-(n+1)\rho_X^n + n\rho_X^{n+1}}{(1-\rho_X)^2}\right], \quad L_n \leqslant \frac{2\sigma_X \rho_X^{n+1}}{1-\rho_X}.$$

2 *Suppose* X_t *satisfies* $\rho(\ell) \leqslant \sigma_X \rho_X^{|\ell|}$ *where* $\rho(\ell)$ *is the* ρ*-mixing coefficient defined in* (1.3.3). *Then*

$$\Omega_n \leqslant 2\|\Gamma(0)\|_{\max}\sigma_X \rho_X \left[\frac{1-(n+1)\rho_X^n + n\rho_X^{n+1}}{(1-\rho_X)^2}\right], \quad L_n \leqslant \frac{2\sigma_X \|\Gamma(0)\|_{\max}\rho_X^{n+1}}{1-\rho_X}.$$

3 *Suppose* X_t *is a stable VAR(d) process* $X_t = \sum_{\ell=1}^d A_d X_{t-d} + \varepsilon_t$, *where* $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2 I)$. *Set* $\tilde{A}_1$ *as in* (1.3.4)*, and assume* $\tilde{A}_1$ *is diagonalizable with an eigen-decomposition* $\tilde{A}_1 = SDS^{-1}$. *Then*

$$\Omega_n \leqslant 2\kappa^2 \frac{\lambda_{\max}(\tilde{A}_1)(1 + n\lambda_{\max}^{n+1}(\tilde{A}_1) - (n+1)\lambda_{\max}(\tilde{A}_1))}{(1-\lambda_{\max}(\tilde{A}_1))^2(1-\lambda_{\max}^2(\tilde{A}_1))},$$
$$L_n \leqslant 2\kappa^2 \frac{\lambda_{\max}^{n+1}(\tilde{A}_1)}{(1-\lambda_{\max}(\tilde{A}_1))(1-\lambda_{\max}^2(\tilde{A}_1))},$$

where $\kappa = \|S\|\|S^{-1}\|$.

Remark. *These bounds show that for a large class of stationary processes* X_t, $\Omega_n(f)/n \to 0$ *and* $L_n(f) \to 0$ *as* $n \to \infty$. *This implies it is possible to choose a large smoothing span* $m \to \infty$ *(required for asymptotically vanishing variance) that also ensures bias vanishing at a rate* $O(m\Omega_n(f)/n)$.

Bound on Variance term: Unlike the bias term, the variance term $|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)|$ is non-deterministic, so we need to establish high probability upper bound on this quantity. Compared to analogous bounds derived in covariance estimation for i.i.d. [Bickel and Levina, 2008] or time series [Shu and Nan, 2014] data, concentration of sample average of periodograms over nearby frequencies requires additional care since the

summands are neither independent nor identically distributed to each other. However, the following proposition shows that the deviation bounds are the same order as i.i.d. data modulo a *price of dependence* captured by $|||f|||$. From a purely technical perspective, this Proposition forms the core of all our subsequent theoretical developments, and we believe this deviation bound will potentially be useful in other problems involving high-dimensional spectral density, e.g., estimation of partial coherence using graphical lasso type algorithms [Jung et al., 2015].

Proposition 1.3.5. *There exist universal positive constants c_1, c_2 such that for any $\eta > 0$,*

$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right| \geqslant |||f|||\eta\right) \leqslant c_1 \exp\left[-c_2(2m+1)\min\{\eta, \eta^2\}\right]. \quad (1.3.4)$$

A complete proof is provided in Appendix 1.A. It is worth noting that the effective sample size in this bound is $(2m + 1)$, a function of the smoothing span. The proof proceeds by separating the real and imaginary parts of $\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)$ into two quadratic forms involving random vectors $\{X_t\}_{t=1}^n$, subsequently applying Lemma 1.3.2 to each part and deriving upper bounds on the spectral norm and ranks of the resulting A matrices.

With the aforementioned bounds on bias and variance parts, we are now ready to present our main result that provides non-asymptotic upper bounds on the estimation error of the high-dimensional thresholded averaged periodogram in operator norm and Frobenius norm for Gaussian time series. The proof adapts techniques of Bickel and Levina [2008] and Rothman et al. [2009] to combine the individual, entry-wise bounds on bias and variance terms across all the entries of the high-dimensional matrix.

Proposition 1.3.6. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series satisfying Assumption 2.1.1, and consider a single Fourier*

frequency $\omega_j \in [-\pi, \pi]$. *Assume* $n \gtrsim \Omega_n(f)|||f|||^2 \log p$. *Then for any* m *satisfying* $m \lesssim n/\Omega_n(f)$ *and* $m \gtrsim |||f|||^2 \log p$, *and any* $R > 0$, *there exist universal constants* $c_1, c_2 > 0$ *such that choosing a threshold*

$$\lambda = 2R|||f|||\sqrt{\frac{\log p}{m}} + 2\left[\frac{m + 1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right], \tag{1.3.5}$$

the estimation error of thresholded averaged periodogram satisfies

$$\mathbb{P}\left(\left\|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\right\| \geqslant 7|||f|||_q^q \lambda^{(1-q)}\right) \leqslant c_1 \exp\left[-(c_2R^2 - 2)\log p\right].$$

Similarly, there exist universal positive constants c_1, c_2 *such that for any* $R > 0$, *with the same choice of threshold in* (2.3.20), *we have*

$$\mathbb{P}\left(\frac{1}{p}\left\|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\right\|_F^2 \geqslant 13|||f|||_q^q \lambda^{2-q}\right) \leqslant c_1 \exp\left[-(c_2R^2 - 2)\log p\right].$$

Remark. *The estimation errors of our thresholded averaged periodogram in both operator norm and Frobenius norm depend on three factors: (i) the weak sparsity level of the true spectral density matrix* $|||f|||_q$; *(ii) measure of stability of the process* $|||f|||$ *to control variance of our estimate; (iii) rate of decay of autocovariances* Ω_n *and* L_n *to control bias of our estimates. For any process satisfying* $\Omega_n/n \to 0$ *faster than* $1/|||f|||^2 \log p$, *it is possible to find a sequence of smoothing span* m *such that* $\lambda \to 0$ *as* $n \to \infty$. *The two appears in the threshold is only for an easy writing for technical proof.*

The above result is non-asymptotic in nature, and our choice of threshold includes an upper bound on the bias. This is in contrast with existing works in the regime $p^2/n \to 0$, where this bias term is asymptotically negligible. Our choices of tuning parameters m and λ then ensure that both bias and variance decrease as n, p grow, which is necessary for meaningful estimation, i.e.,

$$\max\left\{R|||f|||\sqrt{\frac{\log p}{m}}, \frac{m}{n}\Omega_n(f)\right\} = o(1). \tag{1.3.6}$$

Generalized Thresholding of Averaged Periodogram: Building up on the bounds on bias and variance terms of the individual entries of averaged periodogram, we are now ready to present our results for the generalized thresholding operator $S_\lambda(.)$. Suppose we have a generalized thresholding operator $S_\lambda(.)$ satisfying conditions (1)-(3) in Section 2.2. The following proposition generalizes our previous estimation guarantees of hard thresholding to this more generalized family of estimates that includes lasso and adaptive lasso thresholds.

Proposition 1.3.7. *Suppose $S_\lambda(.)$ satisfies conditions (1) - (3) above. Then, for any Fourier frequency $\omega_j, j \in F_n$, and the same choices of tuning parameters m and λ as in Proposition 2.3.7, there exist universal constants $c_i > 0$ such that*

$$\mathbb{P}\left(\|S_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\| > 7|||f|||_q^q \lambda^{(1-q)}\right) \leqslant c_1 \exp\left[-(c_2 R^2 - 2)\log p\right].$$

As pointed out in Rothman et al. [2009], the key is to build concentration inequality for each element of $\hat{f}(\omega_j) - f(\omega_j)$ which is provided by proof in Proposition 1.3.3 and 1.3.5. After building the concentration inequality, all the proof left is exactly same as in Proposition 2.3.7 and Rothman et al. [2009]. We omit this proof for sake of brevity.

Sparsistency of Thresholded Averaged Periodograms: A key motivation for using thresholded averaged periodogram for estimating high-dimensional spectral density matrix is the automatic selection of marginal independence graph among the p times series. Our next result provides a support recovery guarantee at each frequency, justifying usage of these estimates to build weighted networks for downstream functional connectivity analysis in neuroscience problems (see Section 1.6). In particular, the results show that with an appropriate choice of threshold, the support of estimated spectral

density matrix is contained in the true support of $f(\omega)$ with high probability. In addition, if the spectral density is exactly sparse and minimum strength of cross-spectral density is sufficiently large, the entire support is recovered with high probability. For general weakly sparse spectral densities, our proposed thresholding procedures can still recover the strong connections with high probability.

Proposition 1.3.8. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series satisfying Assumption 2.1.1, and consider a single Fourier frequency $\omega_j, j \in F_n$. Assume $n \gtrsim \Omega_n(f)|||f|||^2 \log p$. Then for any m satisfying $m \lesssim n/\Omega_n(f)$ and $m \gtrsim |||f|||^2 \log p$, and any $R > 0$, if we set threshold value λ as* (2.3.20) *, then there exists universal constant c_1, c_2 s.t.*

$$\mathbb{P}\left(\exists\, r, s : T_\lambda(\hat{f}_{rs}(\omega_j)) \neq 0, f_{rs}(\omega_j) = 0\right) \leqslant c_1 \exp[-(c_2 R^2 - 2)\log p].$$

Define $\mathcal{S}(\gamma) = \{(r, s) : |f_{rs}(\omega_j)| \geqslant \gamma\lambda\}$ with some $\gamma > 3/2$, then

$$\mathbb{P}\left(\exists\, (r, s) \in \mathcal{S}(\gamma) : T_\lambda(\hat{f}_{rs}(\omega_j)) = 0, f_{rs}(\omega_j) \neq 0\right) \leqslant c_1 \exp[-(c_2(\gamma-1)^2 R^2 - 2)\log p].$$

Remark. *The first probabilistic bound claims that probability of false positive selection goes to zero if $\lambda = o(1)$ with R large enough and the second probabilistic bound claims that we could recover the signal with strength larger than the threshold we choose ($\gamma > 3/2$).*

Coherence Matrix Estimation: Our next proposition provides an error bound for each element of this plug-in estimator of coherence matrix defined in (1.2.2),

$$\hat{g}_{rs}(\omega_j) = \frac{\hat{f}_{rs}(\omega_j)}{\sqrt{\hat{f}_{rr}(\omega_j)\hat{f}_{ss}(\omega_j)}}.$$

Note that $\hat{f}_{rr} \neq 0$ ($\hat{f}_{rr}$ is a real number) almost surely for Gaussian time series X_t. The sparsistency results can be generalized along the line of Proposition 1.3.8 to ensure coherence graph selection consistency.

Proposition 1.3.9. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series X_t satisfying Assumption 2.1.1, and $\tau := \min_{r=1}^{p} f_{rr}(\omega_j) > 0$. Consider a single Fourier frequency $\omega_j, j \in F_n$. Assume $n \gtrsim \Omega_n(f)|||f|||^2 \log p$. Then for any m satisfying $m \lesssim n/\Omega_n(f)$ and $m \gtrsim |||f|||^2 \log p$ and λ as in* (2.3.20)*, there exist universal positive constants c_1, c_2 such that for any $R > 0$,*

$$\mathbb{P}\left(\exists\, r, s : |T_{2\lambda/\tau}(\hat{g}_{rs}(\omega_j))| > 0, g_{rs}(\omega_j) = 0\right) \leqslant c_1 \exp[-(c_2 R^2 - 2)\log p].$$

Define $\mathcal{S}(\gamma) := \{(r, s) : |g_{rs}(\omega_j)| \geqslant \gamma\lambda/\tau\}$ with some $\gamma > 3/2$. Then we have

$$\mathbb{P}\left(\exists\, (r, s) \in \mathcal{S}(\gamma) : T_{2\lambda/\tau}(\hat{g}_{rs}(\omega_j)) = 0, |g_{rs}(\omega_j)| > 0\right) \leqslant c_1 \exp[-(c_2(\gamma-1)^2 R^2 - 2)\log p].$$

1.4 Spectral Density Estimation of Linear Processes

In this section, we extend the estimation consistency results of our thresholding based spectral density estimators beyond Gaussian time series. The proof of the Hanson-Wright type inequality for temporally dependent data in Lemma 1.3.2 crucially relies on the fact that uncorrelated Gaussian random variables are also independent with each other. This does not apply for non-Gaussian time series in general. However, we show in this section that for some linear processes with error tail heavier than Gaussian distribution, it is possible to derive similar concentration inequalities. Using these concentration inequalities, we then extend the theoretical results of previous section to a larger class of non-Gaussian linear time series.

We focus on linear processes with absolutely summable MA(∞) coefficients:

$$X_t = \sum_{\ell=0}^{\infty} B_\ell \varepsilon_{t-\ell}, \tag{1.4.1}$$

where $B_\ell \in \mathbb{R}^{p\times p}$ and $\varepsilon_t \in \mathbb{R}^p$ have i.i.d. centered distribution with possibly heavier tails than Gaussian. Rosenblatt [1985] shows that stationarity of X_t is ensured under element-wise absolute summability of MA coefficients

$$\sum_{\ell=0}^{\infty} |B_{\ell,(r,s)}| < \infty \tag{1.4.2}$$

for any r, s, $1 \leqslant r, s \leqslant p$. Under this condition, the autocovariance $\Gamma(\ell) = \sum_{t=0}^{\infty} B_t B_{t+\ell}^\top$ is well-defined for every $\ell \in \mathbb{Z}$, and Assumption 2.1.1 holds. A proof is given in Lemma 1.C.7 for completeness.

We assume that each component ε_{tr}, $1 \leqslant r \leqslant p$, of the random vector ε_t is from one of the following three types of distributions.

(C1) sub-Gaussian: there exists some $\sigma > 0$ such that for all $\eta > 0$, $\mathbb{P}[|\varepsilon_{tr}| > \eta] \leqslant 2\exp\left(-\frac{\eta^2}{2\sigma^2}\right)$;

(C2) generalized sub-exponential with parameter $\alpha > 0$: there exist positive constants a, b such that for all $\eta > 0$, $\mathbb{P}[|\varepsilon_{tr}| \geqslant \eta^\alpha] \leqslant a\exp(-b\eta)$ [Erdős et al., 2012];

(C3) ε_{tr} has finite 4^{th} moment: $\mathbb{E}\varepsilon_{tr}^4 \leqslant K < \infty$.

Remark. *ε_{tr} has generalized sub-exponential distribution defined in Erdős et al. [2012], which is more general than the usual definition of sub-exponential used in the literature with $\alpha = 1$. In some recent works [Faradonbeh et al., 2018; Wong and Tewari, 2017], such distributions were also referred to as sub-Weibull distributions.*

Next we establish concentration inequalities similar to Lemma 1.3.2 for linear processes where the distribution of each coordinate of noise terms comes from one of the families C1, C2 and C3.

For i.i.d. data, existing works have generalized Hanson-Wright type inequality for distributions in C1 and C2 [Erdős et al., 2012; Rudelson and Vershynin, 2013]. We can use Markov inequality to get an upper bound for C3 as well. We summarize these results in the following lemma. Its proof is defered to Appendix 1.B.

Lemma 1.4.1. *Consider a random vector $\varepsilon \in \mathbb{R}^p$ with i.i.d. coordinates following one of the three distributions C1 - C3, and a deterministic $p \times p$ matrix A. For simplicity, let us assume A is a real matrix, and $\mathbb{E}\varepsilon_r = 0$ and $\mathbb{E}\varepsilon_r^2 = 1$ for every r, $1 \leqslant r \leqslant p$. Then*

$$\mathbb{P}\left(|\varepsilon^\top A\varepsilon - \mathbb{E}\varepsilon^\top A\varepsilon| \geqslant \eta\right) \leqslant \mathcal{T}_j(\eta, A),$$

where $\mathcal{T}_j(\eta, A), j = 1, 2, 3$, are tail decay functions for the three families, given by

$$\begin{aligned}
\mathcal{T}_1(\eta, A) &= 2\exp\left[-c\min\left\{\frac{\eta}{\|A\|}, \frac{\eta^2}{rk(A)\|A\|^2}\right\}\right], \\
\mathcal{T}_2(\eta, A) &= c_1\exp\left[-c_2\left(\frac{\eta}{\sqrt{rk(A)}\|A\|}\right)^{\frac{1}{2+2\alpha}}\right], \\
\mathcal{T}_3(\eta, A) &= \frac{c_3 rk(A)\|A\|^2}{\eta^2}.
\end{aligned}$$

Here c only depends on σ in C1, c_1, c_2 only depend on a, b in C2 and c_3 only depends on K in C3, and none of them depends on the MA coefficients B_ℓ, $\ell \geqslant 0$.

Now we extend these three inequalities by replacing ε with n random variables of the form $X_t = \sum_{\ell \geqslant 0} B_\ell \varepsilon_{t-\ell}$. The main technical difficulty stems from handling the sum of infinitely many terms ε_t. We apply a truncation argument to overcome this.

Proposition 1.4.2. *Suppose $\mathcal{X} = [X_1 : X_2 : \ldots : X_n]^\top$ is a data matrix with n consecutive observations from a stationary linear process $\{X_t\}$ in* (1.4.2) *with each coordinate of ε_t is i.i.d. from one of the families C1, C2 or C3, and consider a deterministic $np \times np$*

matrix A. Then

$$\mathbb{P}\left(\left|vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top) - \mathbb{E}\left[\, vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top)\right]\right| > 2\pi\eta|||f|||\right) \leqslant \mathcal{T}_j(\eta, A),$$

where $\mathcal{T}_j(\eta, A), j = 1, 2, 3$, are tail decay functions for the three families, as defined in Lemma 1.4.1.

Remark. *The main difference between the concentration inequalities in Lemma 1.4.1 and Proposition 1.4.2 is that $|||f|||$ appears in the right side of the inequality. As pointed by Basu and Michailidis [2015], $|||f|||$ can be viewed as a "price of dependence" present in time series data. For instance, if $B_\ell = 0$ for all $\ell > 0$, $B_0 = I$, and Var$(\varepsilon_{tr}) = 1$ for all r, t, we have $|||f||| = \frac{1}{2\pi}$ which coincides with the result in Lemma 1.4.1 applied to a np-dimensional random vector.*

This result generalizes the Hanson-Wright type concentration inequality in Lemma 1.3.2 to the case of three non-Gaussian families with potentially heavier tails. After building concentration inequalities for these three cases, we could bound the variance term as Proposition 1.3.5 which we listed as following Proposition. The proof follows the same line as the proof of Proposition 1.3.5, by replacing Gaussian Hanson-Wright type inequality with those in Proposition 1.4.2. We omit this for sake of brevity.

Proposition 1.4.3. *Suppose $\mathcal{X} = [X_1 : X_2 : \ldots : X_n]^\top$ is a data matrix with n consecutive observations from a stationary linear process $\{X_t\}$ in* (1.4.2) *, each coordinate of ε_t is i.i.d. from one of the families C1, C2 or C3. Then there exist general constants $c_i > 0$ (depending only on the error distribution but not on the coefficients B_ℓ of the linear process) such that for any r, s, $1 \leqslant r, s \leqslant p$, and any Fourier frequency $\omega_j \in F_n$, we have*

$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right| \geqslant |||f|||\eta\right) \leqslant \mathcal{B}_k(\eta, m), \tag{1.4.3}$$

where $\mathcal{B}_k$, $k = 1, 2, 3$, are defined as

$$\begin{aligned}
\mathcal{B}_1(\eta, m) &= c_1 \exp\left[-c_2 \min\{\eta, \eta^2\}\right], \\
\mathcal{B}_2(\eta, m) &= c_3 \exp(\left[-c_4 \left(\sqrt{m}\eta\right)^{\frac{1}{2+2\alpha}}\right]), \\
\mathcal{B}_3(\eta, m) &= \frac{c_5}{m\eta^2}.
\end{aligned}$$

After showing the bound for variance term for linear process, we can derive estimation consistency of hard-thresholding estimators similar to Proposition 2.3.7 for linear processes with any of the three different types of noise distributions.

Proposition 1.4.4. *Suppose $\{X_t\}$ is a linear process defined in* (1.4.1)*, with ε_t from one of the three distributions C1, C2 and C3, and consider a Fourier frequency $\omega_j \in F_n$. Assume $n \gtrsim \Omega_n(f)\mathcal{N}_k$, where $\mathcal{N}_1 = |||f|||^2 \log p$, $\mathcal{N}_2 = |||f|||^2 (\log p)^{4+4\alpha}$, and $\mathcal{N}_3 = p^2$ for the three families C1, C2 and C3. Then for any m satisfying $m \lesssim n/\Omega_n(f)$ and $m \gtrsim |||f|||^2 \mathcal{N}_k$, and any $R > 0$, if we choose threshold for the three different distributions as*

(C1) $\lambda = 2R|||f|||\sqrt{\frac{\log p}{m}} + 2\left[\frac{m+1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right]$,

(C2) $\lambda = 2|||f|||\frac{(R\log p)^{2+2\alpha}}{\sqrt{m}} + 2\left[\frac{m+1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right]$,

(C3) $\lambda = 2|||f|||\frac{p^{1+R}}{\sqrt{m}} + 2\left[\frac{m+1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right]$,

then

$$\mathbb{P}\left(\|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\| > 7|||f|||_q^q \lambda^{(1-q)}\right) \leqslant \mathcal{B}_k,$$

where the tail probability $\mathcal{B}_k$ are given as

$$\begin{aligned}
\mathcal{B}_1 &= c_1 \exp\left[-(c_2 R^2 - 2)\log p\right], \\
\mathcal{B}_2 &= c_3 \exp\left[-(c_4 R - 2)\log p\right], \\
\mathcal{B}_3 &= c_5 \exp\left[-2R\log p\right],
\end{aligned} \tag{1.4.4}$$

where $c_i > 0$ are some general constants depending only on the error distribution but not on the coefficients B_ℓ of the linear process.

The proof follows the same line as the proof of Proposition 2.3.7, by replacing Gaussian variance bound in Proposition 1.3.5 with Proposition 1.4.3. We omit this for sake of brevity.

Remark. *The heavier is the tail of the noise distribution, the wider bandwidth of periodogram averaging ($2m + 1$ in our notation) is required for consistent estimation. For generalized sub-exponential, we can ensure consistency in high-dimensional regime $p = O(n^\alpha), \alpha > 1$, while if we only assume existence of fourth moment, we will require $p = o(\sqrt{n})$ for consistency.*

1.5 Simulation Studies

We assess the finite sample properties of our proposed spectral density estimators through numerical experiments on simulated data sets. To this end, we compare the performance of smoothed periodogram, shrinkage estimator from Böhm and von Sachs [2009], hard thresholding, soft thresholding (lasso) and adaptive lasso thresholding. In particular, we simulate data from vector moving average (VMA) and autoregressive (VAR) processes with block-diagonal transition matrices and evaluate estimation and model selection performance of these methods for different values of n and p. Overall, the results demonstrate that thresholding methods provide substantial improvements in estimation accuracy over smoothed periodograms and shrinkage methods when p is large and the true spectral density is approximately sparse. In addition, thresholding methods accurately recovers the edges in coherence networks, as measured by their

precision, recall and area under receiver operating characteristic (ROC) curves.

Generative models: We consider VAR(1) models $X_t = AX_{t-1} + \varepsilon_t$ of three different dimensions: $p = 12, 48, 96$. Each element in ε_t is independent and identically distributed as $\mathcal{N}(0, 1)$, and the transition matrix A is composed of 3×3 block matrices on the diagonal. Each block matrix A^0 has 0.5 on the diagonal and 0.9 on the first upper off-diagonal. We also consider VMA(1) models $X_t = B\varepsilon_{t-1} + \varepsilon_t$ of the same dimensions as the VAR models. These transition matrix structures are adopted from Fiecas and von Sachs [2014], where a data-driven shrinkage method was shown to improve upon smoothed periodograms in high-dimensional settings. For each model, we generate $n = 100, 200, 400, 600$ consecutive observations from the multivariate time series.

The transition matrix A of VAR is a block diagonal composed of identical blocks consisting of a 3×3 upper triangular matrix A^0. Similarly, the VMA transition matrix B is a block diagonal matrix composed of identical 3×3 upper triangular matrix B^0.

$$A^0 = B^0 = \begin{bmatrix} 0.5 & 0.9 & 0 \\ 0 & 0.5 & 0.9 \\ 0 & 0 & 0.5 \end{bmatrix}. \tag{1.5.1}$$

The estimated spectral density matrices are compared to the true spectral densities. For stable, invertible VARMA(1,1) processes $X_t = AX_{t-1} + \varepsilon_t + B\varepsilon_{t-1}$, true spectral densities take the form

$$f(\omega) = \frac{1}{2\pi}(\mathcal{A}^{-1}(e^{-i\omega}))\mathcal{B}(e^{-i\omega})\Sigma_\varepsilon\mathcal{B}^\dagger(e^{-i\omega})(\mathcal{A}^{-1}(e^{-i\omega}))^\dagger,$$

where $\mathcal{A}(z) = I_p - Az$ and $\mathcal{B}(z) = I_p + Bz$.

Performance Metrics: We compare the estimation performances of different estimators of $f(\omega_j)$ using Relative Mean Integrated Squared Error (RMISE) in Frobenius norm, defined as

$$RMISE(\hat{f}) := \frac{\sum_{j\in F_n} \|\hat{f}(\omega_j) - f(\omega_j)\|_F^2}{\sum_{j\in F_n} \|f(\omega_j)\|_F^2}.$$

In order to capture how well the three thresholding methods recover the non-zero coordinates in a spectral density matrix under exactly sparse generative VMA and VAR models, we also record their precision, recall and F1 measures over all Fourier frequencies

$$\begin{aligned}
\text{precision}(\omega_j) &= \frac{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0, \ |f_{rs}(\omega_j)| \neq 0\}}{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0\}} \\
\text{recall}(\omega_j) &= \frac{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0, \ |f_{rs}(\omega_j)| \neq 0\}}{\#\{(r,s) : \ |f_{rs}(\omega_j)| \neq 0\}} \\
\text{F1}(\omega_j) &= 2 \times (\text{precision}(\omega_j) \cdot \text{recall}(\omega_j))/(\text{precision}(\omega_j) + \text{recall}(\omega_j)).
\end{aligned}$$

We calculate each of the three criteria averaged across all Fourier frequencies $j \in F_n$. All the experiments are replicated 50 times, and mean and standard deviation of the performance metrics are reported.

We also evaluate the accuracy of thresholding methods in selecting the graph $G = \{(r,s) \in V \times V : \hat{f}_{rs}(\omega_j) \neq 0 \text{ for some } \omega_j \in F_n\}$. For this purpose, we use averaged absolute coherence (across all Fourier frequencies) to construct a single $p \times p$ weighted adjacency matrix $\hat{G}$, and then measure its accuracy in selecting edges of the true graph G.

Tuning parameter selection: For each of the three thresholding methods, we use the sample-splitting algorithm 1 with $N = 1$ to determine the value of threshold for individual frequencies. We choose a grid $\mathcal{L}$ of equispaced values between the minimum and maximum moduli of off-diagonal entries in smoothed periodogram. Based on the

theoretical considerations in Section 3.5, the smoothing spans for VMA models are chosen by setting $m = \sqrt{n}$. Since $\Omega_n(f)$ is larger for VAR than VMA models considered here, a smaller smoothing span is chosen by setting $m = 2/3\sqrt{n}$. The results are qualitatively similar in our sensitivity analysis with different values of m of this order.

Results: The RMISE of smoothed (averaged) periodograms, shrinkage and thrsholding methods are reported in Table **??**. The results show that both shrinkage and thresholding outperform smoothed periodogram, and the improvement is more prominent for larger p. Further, thresholding procedures show some improvement over shrinkage methods in these approximately sparse data generative models. Amongst the three thresholding methods, lasso and adaptive lasso tend to have lower error than hard thresholding in most settings.

Precision, recall and F1 scores of the three thresholding methods are reported in Appendix 1.D. In most of the simulation settings, the methods have high precision but low recall, indicating higher true negative in general. This matches with our theoretical predictions for weakly sparse spectral densities in Proposition 1.3.8. The F1 scores are in the range of $50 - 60\%$ in most simulation settings. As in the RMISE results, lasso and adaptive lasso thresholds perform significantly better than hard thresholding in most simulation settings.

The ROC curves for the three thresholding methods in selecting coherence graph of a VAR(1) model with $p = 48$ and $n \in \{100, 200, 400, 600\}$ are provided in Figure 1.1. Consistent with the frequency-specific precision and recall results, lasso and adaptive lasso thresholding methods perform better than hard thresholding.

Overall, our numerical experiments confirm that thresholding procedures can be successfully used to estimate large spectral density matrices with same order of accu-

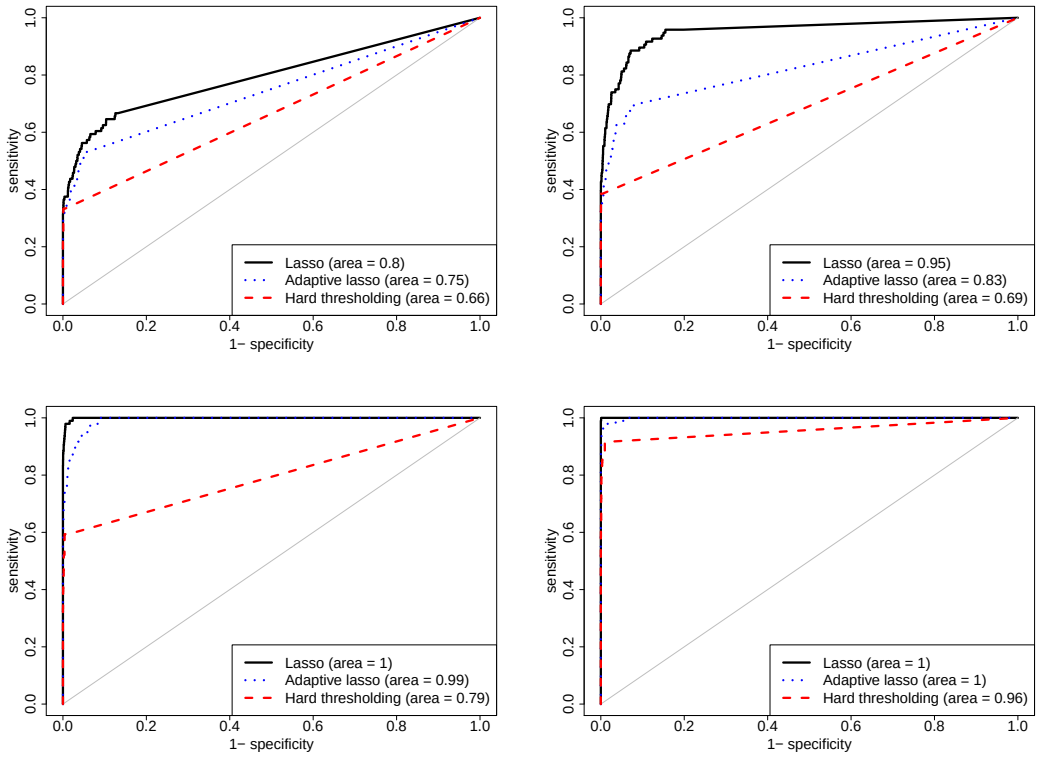


Figure 1.1: Receiver Operating Characeristic (ROC) curves of hard thresholding, lasso and adaptive lasso for recovering coherence network of a $p = 96$ dimensional VAR(1) model using $n = 100$ (top left), $n = 200$ (top right), $n = 400$ (bottom left) and $n = 600$ (bottom right) time series observations.

racy as shrinkage methods, and with an additional advantage of performing automatic edge selection in coherence networks.

1.6 Functional Connectivity Analysis with fMRI Data

We demonstrate the advantage of thresholding based spectral density estimators for visualization and interpretation in functional connectivity analysis among different brain regions of a human subject using resting state fMRI data. This data is part of a study involving 51 subjects (29.6 ± 8.6 years of age, 35 males) that suffered from mild trau-

	Smoothed	Shrinkage	Hard Threshold	Lasso	Adaptive Lasso
VMA					
p = 12					
n = 100	43.21(6.86)	22.15(1.77)	27.43(2.11)	22.89(2.05)	25.54(1.91)
n = 200	29.95(2.93)	17.67(1.01)	20.5(1.45)	16.18(1.33)	18.74(1.32)
n = 400	21.11(1.75)	14.24(0.67)	12.39(1.52)	10.84(1.01)	11.33(1.27)
n = 600	17.28(1.39)	12.58(0.59)	8.73(1.16)	8.84(0.71)	8.45(0.92)
p = 24					
n = 100	80.49(7.63)	26.28(1.62)	29.86(1.21)	26.36(1.5)	28.38(1.31)
n = 200	59.79(4.7)	22.92(0.81)	25.62(0.86)	19.29(1.35)	22.09(1.21)
n = 400	41.83(1.98)	19.54(0.45)	17.16(1.26)	13.0(0.99)	13.71(1.25)
n = 600	35.86(1.6)	17.83(0.36)	12.27(1.01)	10.36(0.65)	9.89(0.84)
p = 48					
n = 100	162.79(9.94)	29.58(1.24)	30.62(0.91)	28.78(0.83)	29.9(0.8)
n = 200	119.58(4.21)	27.0(0.57)	28.29(0.37)	22.68(0.76)	25.48(0.72)
n = 400	83.48(2.67)	24.09(0.37)	22.35(0.65)	15.86(0.54)	17.21(0.74)
n = 600	69.83(1.77)	22.58(0.28)	16.95(0.81)	12.88(0.48)	12.73(0.7)
p = 96					
n = 100	324.57(14.7)	32.34(1.15)	30.3(0.46)	29.71(0.43)	30.11(0.44)
n = 200	235.78(7.75)	29.58(0.67)	28.83(0.28)	25.28(0.43)	27.31(0.38)
n = 400	167.89(4.28)	27.44(0.37)	25.67(0.33)	18.58(0.5)	20.34(0.55)
n = 600	139.4(2.02)	26.26(0.24)	21.25(0.48)	15.35(0.37)	15.72(0.51)
VAR					
p = 12					
n = 100	39.11(10.1)	37.49(5.27)	41.09(6.36)	38.46(5.25)	41.81(5.18)
n = 200	28.06(8.4)	25.2(4.15)	30.52(5.83)	27.6(4.19)	30.69(5.21)
n = 400	17.31(4.63)	16.51(2.93)	19.37(3.74)	16.84(2.61)	19.5(3.41)
n = 600	25.0(5.86)	19.23(3.95)	23.07(4.62)	18.55(2.85)	21.65(3.92)
p = 24					
n = 100	73.83(15.52)	49.25(4.16)	49.18(4.78)	44.64(3.8)	47.59(3.54)
n = 200	54.77(9.83)	36.84(2.97)	40.95(3.46)	34.29(3.52)	38.46(3.47)
n = 400	35.53(6.01)	27.34(2.05)	27.43(2.86)	22.32(1.76)	25.05(2.67)
n = 600	28.53(2.24)	21.82(0.74)	17.25(0.97)	15.17(0.11)	16.64(0.98)
p = 48					
n = 100	131.88(20.49)	61.75(4.11)	49.3(3.35)	47.12(1.89)	48.1(2.25)
n = 200	99.46(12.68)	48.3(2.17)	44.24(1.53)	39.31(1.77)	42.63(1.41)
n = 400	69.19(7.07)	38.38(1.44)	35.52(1.55)	26.69(1.35)	30.5(1.75)
n = 600	53.08(1.38)	32.58(0.4)	25.23(0.41)	20.38(0.3)	21.16(0.52)
p = 96					
n = 100	259.85(31.63)	75.46(5.47)	48.6(1.69)	47.96(1.41)	48.15(1.59)
n = 200	200.12(16.87)	59.45(1.88)	45.18(1.23)	43.34(1.2)	44.63(1.06)
n = 400	135.52(8.76)	50.08(1.25)	41.41(0.7)	32.53(1.11)	37.13(1.14)
n = 600	97.13(1.32)	42.62(0.08)	31.65(0.45)	24.6(0.34)	24.56(0.69)

Table 1.1: Relative Mean Integrated Squared Error (RMISE, in %) of smoothed periodogram, shrinkage towards a diagonal target and three different thresholding methods - hard thresholding, lasso and adaptive lasso. Results are averaged over 20 replicates. Standard deviations (also in %) are reported in parentheses.

matic brain injury (TBI). Magnetic resonance imaging (MRI) data and neuropsychological data were collected at 1 week, 1 month, 6 months and 12 months post-injury. TBI is defined as Glasgow Coma Scale of 13-15 at injury, loss of consciousness less than 30 minutes and post-traumatic amnesia less than 24 hours. More details are available in Kuceyeski et al. [2018].

A 3T GE Signa EXCITE scanner was used to acquire the MRIs, which included structural scans (FSPGR T1, $1 \times 1 \times 1$ mm^3 voxels) and resting-state functional magnetic resonance imaging (fMRI) (7 min, $3.4 \times 3.4 \times 4.0$ mm^3 voxels, 2 sec sampling rate). The MRIs were processed by parcellating the gray matter into $p = 86$ anatomical regions of interest (ROIs) using the semi-automated FreeSurfer software [**?**]. Cortical and subcortical parcellations and the fMRI time series data were then used in the construction of coherence based functional connectivity (FC) networks. The adjacency matrix of FC network captures the similarity of the neuronal activation over time between pairs of ROIs.

We calculated coherence matrices at frequency 0 using adaptive lasso thresholding (with $\eta = 2$) and shrinkage of averaged periodograms. The smoothing span was chosen by setting $m = \sqrt{n}$, and the tuning parameters in our sample-splitting algorithm were selected as in our simulation studies.

Results: In Figure 1.2, we show an example of the FC coherence network for a particular TBI patient using our proposed adaptive lasso thresholding (top left) and the same patient's FC network estimated using the shrinkage method (top right) of Böhm and von Sachs [2009] that does not perform automatic coherence selection. One of the many issues with using fMRI data is the spurious functional connections that arise from the method's abundant noise (due to instrumentation and physiology). It is often

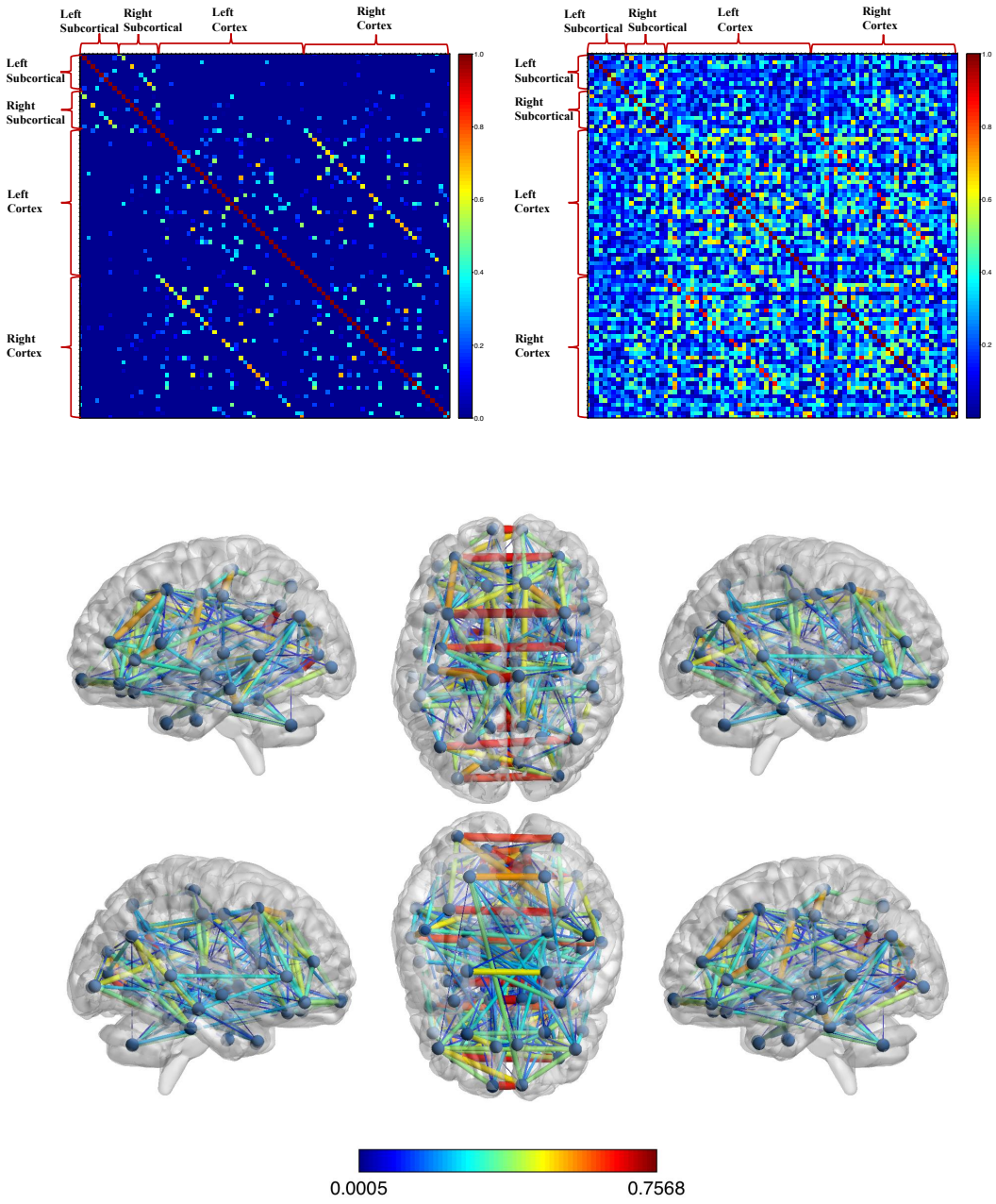


Figure 1.2: [top]: Heat maps of absolute coherence matrices (at frequency 0) obtained from spectral density estimated using [top left] adaptive lasso thresholding and [top right] a shrinkage method. [bottom]: Absolute coherence network among brain regions obtained using adaptive lasso and visualized using BrainNet Viewer. The coherence network estimated by adaptive lasso retains known biological patterns, including presence of bilateral homologues, i.e. strong connectivity between same ROIs in the left and right parts of brain.

preferable in a clinical context to filter out this noise, but it is not currently done in a universally accepted and statistically principled way. As shown in the top panel of Figure 1.2, the coherence matrix estimated by adaptive lasso thresholding obviously is more sparse in nature compared to the one from shrinkage method, while maintaining known physiological connections. For example, we see strong FC in the bilateral homologues (the same ROI in the left versus right hemisphere), which are known to have strong functional connections [Zuo et al., 2010]. This is even more readily apparent in the bottom panel of Figure 1.2 where we see strong connections between the same ROI in the left and right sides of the brain. Other than the bilateral homologues, the left and right precuneus, isthmus cingulate, lingual gyrus and pericalcarine have prominent connections to many regions (see Figures 1.3 and 1.4 in Appendix 1.D). The precuneus, which plays a role in visual, sensorimotor, and attentional information processing, is central to resting-state (task negative) fMRI networks detected using correlation analysis [**?**]. Additionally, the isthmus cingulate, part of the posterior cingulate cortex, is known to be highly functionally connected to many regions across the brain at rest [**?**]. In addition, we see a stronger FC between the left and right homologues in the subcortical ROIs (upper left corner) than between subcortical and cortical ROIs. It is interesting to note that while some of these connections are also strong in the shrinkage based coherence matrix estimate, it is not easy to separate them from other moderately strong coherences between brain regions.

1.7 Discussion

We proposed hard thresholding and generalized thresholding of averaged periodogram for estimation of high-dimensional spectral density matrices of stable Gaussian time

series and linear processes with errors having potentially heavier tails than Gaussian. Under high-dimensional regime $\log p/n \to 0$, we established consistency of the above estimation procedures when the true spectral densities are weakly sparse. At the core of our technical results lie concentration inequalities of complex quadratic forms of temporally dependent, high-dimensional random vectors, which were used to derive finite sample deviation of averaged periodograms around their expectation. These results are of independent interest and are potentially useful in other problems involving high-dimensional spectral density. In our next steps, we plan to extend the theoretical analyses to more general adaptive thresholding methods [Cai and Liu, 2011], which will explicitly account for heterogeneity in the strengths of cross-spectral association across different pairs of time series and different frequency bands. We also plan to develop estimation and inference procedures for high-dimensional partial coherence at different frequencies.

Another direction of potential interest is to develop thresholding strategies that incorporate information on different brain regions and prior biological knowledge on brain networks. Dynamic functional connectivity of brain networks is known to play important roles behind progression of neurodegenerative diseases. A common approach to build such networks is using coherence measures of Fourier or wavelet transform of multi-channel fMRI/EEG/MEG signals and thresholding small entries of zero. Selection of threshold level that represents heterogeneous modular structure of human brain has been a topic of active research [Bordier et al., 2017]. We expect that more sophisticated thresholding methods, building up on universal and adaptive thresholds and incorporating prior neuroscientific knowledge, will be potentially useful in data-driven discovery of scientifically and clinically relevant connectivity patterns in human brain.

1.A Appendix: Proofs for Gaussian Time Series

1.A.1 Proof of Lemma 1.3.2

Proof. We can write $vec(\mathcal{X}^\top) \overset{d}{=} \Sigma^{1/2} Z$, where Σ is the covariance matrix of the np-dimensional random vector $vec(\mathcal{X}^\top)$ and $Z \sim N(0, I)$. Then using Hanson-Wright inequality [Theorem 1.1, Rudelson and Vershynin [2013]] and the fact that the sub-Gaussian norm of Z is 1, we conclude that there exists a universal constant $c > 0$ satisfying

$$\begin{aligned}
&\mathbb{P}\left(\left|vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top) - \mathbb{E}\left[vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top)\right]\right| > 2\pi\eta|||f|||\right) \\
&= \mathbb{P}\left(\left|Z^\top \Sigma^{1/2} A \Sigma^{1/2} Z - \mathbb{E}\left[Z^\top \Sigma^{1/2} A \Sigma^{1/2} Z\right]\right| > 2\pi\eta|||f|||\right) \\
&\leqslant 2\exp\left[-c\min\left\{\frac{2\pi\eta|||f|||}{\|\Sigma^{1/2}A\Sigma^{1/2}\|}, \frac{4\pi^2\eta^2|||f|||^2}{\|\Sigma^{1/2}A\Sigma^{1/2}\|_F^2}\right\}\right].
\end{aligned} \tag{1.A.1}$$

Using Lemma 1.C.5, $\|\Sigma^{1/2}A\Sigma^{1/2}\| \leqslant \|\Sigma\|\|A\| \leqslant 2\pi|||f|||\|A\|$. It follows from Golub and Van Loan [2012],

$$\begin{aligned}
\|\Sigma^{1/2}A\Sigma^{1/2}\|_F &\leqslant \sqrt{\text{rk}(\Sigma^{1/2}A\Sigma^{1/2})}\|\Sigma^{1/2}A\Sigma^{1/2}\| \\
&\leqslant \sqrt{\text{rk}(A)}\|\Sigma^{1/2}A\Sigma^{1/2}\| \leqslant 2\pi\sqrt{\text{rk}(A)}\|A\|||f|||.
\end{aligned}$$

Then plugging in the bound for $\|\Sigma^{1/2}A\Sigma^{1/2}\|$ and $\|\Sigma^{1/2}A\Sigma^{1/2}\|_F$ into (1.A.1) completes the proof. □

1.A.2 Proof of Proposition 1.3.3

Proof. It suffices to show that for any two unit vectors e_r, e_s,

$$\left|e_r^\top \left[\mathbb{E}\hat{f}(\omega_j) - f(\omega_j)\right] e_s\right| \leqslant \frac{m}{n}\Omega_n(f) + \frac{1}{2\pi}\left(\frac{\Omega_n(f)}{n} + L_n(f)\right).$$

Since

$$\hat{f}(\omega_j) = \frac{1}{2\pi(2m+1)} \sum_{\ell=-m}^{m} I(\omega_{j+\ell}),$$

we have

$$\begin{aligned}\left|e_r^\top \left[\mathbb{E}\hat{f}(\omega_j) - f(\omega_j)\right] e_s\right| \leqslant \frac{1}{2\pi(2m+1)} \sum_{\ell=-m}^{m} \left|e_r^\top \left[\mathbb{E}I(\omega_{j+\ell}) - \mathbb{E}I(\omega_j)\right] e_s\right| \\ + \left|e_r^\top \left[\frac{1}{2\pi}\mathbb{E}I(\omega_j) - f(\omega_j)\right] e_s\right|.\end{aligned} \tag{1.A.2}$$

By definition of $I(\omega_j)$ in (2.1.3), we have $\mathbb{E}I(\omega_j) = \sum_{|k|\leqslant n} \Gamma(k)\frac{(n-|k|)}{n}e^{-ik\omega_j}$. Therefore, the second term above takes the form

$$\begin{aligned}\left|\frac{1}{2\pi}e_r^\top \left[\mathbb{E}I(\omega_j) - 2\pi f(\omega_j)\right] e_s\right| &= \frac{1}{2\pi}\left|\sum_{|k|\leqslant n}\frac{|k|}{n}\Gamma_{rs}(k)e^{-ik\omega_j} + \sum_{|k|>n}\Gamma_{rs}(k)e^{-ik\omega_j}\right| \\ &\leqslant \frac{1}{2\pi}\left[\sum_{|k|\leqslant n}\frac{|k|}{n}|\Gamma_{rs}(k)| + \sum_{|k|>n}|\Gamma_{rs}(k)|\right] \\ &= \frac{1}{2\pi}\left(\frac{\Omega_n(f)}{n} + L_n(f)\right).\end{aligned} \tag{1.A.3}$$

For the first term, note that $|e^{ix} - e^{iy}| \leqslant |x-y|$ and $|\omega_j - \omega_{j+\ell}| = 2\pi\frac{|\ell|}{n}$. This implies

$$\begin{aligned}\left|\frac{1}{2\pi}e_r^\top \left[\mathbb{E}I(\omega_j) - \mathbb{E}I(\omega_{j+\ell})\right] e_s\right| &= \frac{1}{2\pi}\left|\sum_{|k|\leqslant n}\left(1 - \frac{|k|}{n}\right)|\Gamma_{rs}(k)|(e^{-ik\omega_j} - e^{-ik\omega_{j+\ell}})\right| \\ &\leqslant \frac{1}{2\pi}\sum_{|k|\leqslant n}|\Gamma_{rs}(k)||k||\omega_j - \omega_{j+\ell}| = |\ell|\Omega_n(f)/n.\end{aligned} \tag{1.A.4}$$

Plugging in (1.A.3) and (1.A.4) into (1.A.2),

$$\begin{aligned}\left|e_r^\top \left[\mathbb{E}\hat{f}(\omega_j) - f(\omega_j)\right] e_s\right| &\leqslant \frac{1}{2\pi}\left(\frac{\Omega_n(f)}{n} + L_n(f)\right) + \left(\frac{\sum_{|\ell|\leqslant m}|\ell|}{2m+1}\right)\frac{\Omega_n(f)}{n} \\ &\leqslant \frac{m}{n}\Omega_n(f) + \frac{1}{2\pi}\left(\frac{\Omega_n(f)}{n} + L_n(f)\right).\end{aligned}$$

□

1.A.3 Proof for Proposition 1.3.4

Proof. We will prove the proposition one by one for its three conditions. Proof for all three conditions uses the simple fact that, for $0 < x < 1$

$$\sum_{\ell=1}^{n} \ell x^\ell = \frac{x(1 + nx^{n+1} - (n+1)x^n)}{(1-x)^2}. \tag{1.A.5}$$

Condition 1: Directly plug the bound on $\|\Gamma(\ell)\|_{\max}$ and with $|\Gamma_{r,s}(\ell)| \leqslant \|\Gamma(\ell)\|_{max}$, we have

$$\Omega_n \leqslant 2\sum_{\ell=1}^{n} \ell\|\Gamma(\ell)\|_{\max} \leqslant 2\sigma_X \sum_{\ell=1}^{n} \ell\rho_X^\ell = \frac{2\sigma_X\rho_X(1 + n\rho_X^{n+1} - (n+1)\rho_X^n)}{(1-\rho_X)^2}.$$

For L_n,

$$L_n \leqslant 2\sum_{\ell>n} \|\Gamma_\ell\|_{\max} \leqslant 2\sigma_X \sum_{\ell>n} \rho_X^\ell = \frac{2\sigma_X\rho_X^{n+1}}{1-\rho_X}.$$

Condition 2: Note that condition of geometrically decaying ρ-mixing coefficient leads to condition 1 as

$$\begin{aligned}|\Gamma_{rs}(\ell)| &= \left|\frac{\mathbb{E}e_r' X_\ell X_0^\top e_s}{\sqrt{|\Gamma_{rr}(0)||\Gamma_{ss}(0)|}}\right| \sqrt{|\Gamma_{rr}(0)||\Gamma_{ss}(0)|} \\ &\leqslant \|\Gamma(0)\|_{\max}\sigma_X\rho_X^{|\ell|}.\end{aligned}$$

Then follow the argument for condition 1, we finish the proof.

Condition 3: Let $\tilde{\Omega}_n$, $\tilde{L}_n$ be the Ω_n, L_n defzined before for time series $\tilde{X}_t$ and $\tilde{\Gamma}(\ell)$ be the auto-covariance for $\tilde{X}$. We first show that $\tilde{\Omega}_n$, $\tilde{L}_n$ are upper bounds for Ω_n, L_n. Then we present upper bounds for $\tilde{\Omega}_n$ and $\tilde{L}_n$ although it may lose the tightness in controlling of growth rates of these two. To see this, we partition $\tilde{\Gamma}(\ell)$ into blocks as follows.

$$\tilde{\Gamma}(\ell) = \begin{bmatrix} \Gamma(\ell) & \Gamma(\ell+1) & \cdots & \Gamma(\ell+d-1) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(\ell-d+1) & \Gamma(\ell-d) & \cdots & \Gamma(\ell) \end{bmatrix}.$$

Since $\Gamma(\ell)$ appears as diagonal block of $\tilde{\Gamma}(\ell)$, based on the definition of Ω_n and L_n, we can claim that $\tilde{\Omega}_n$, $\tilde{L}_n$ are upper bounds for Ω_n and L_n respectively. Next, we focus on gettting upper bound for $\tilde{\Omega}_n$ and $\tilde{L}_n$. Consider the infinite moving average representation of $\tilde{X}_t$

$$\tilde{X}_t = \sum_{\ell=0}^{\infty} \tilde{B}_\ell \tilde{\varepsilon}_{t-\ell},$$

where $\tilde{B}_\ell = (\tilde{A}_1)^\ell$ and autocovariance becomes

$$\tilde{\Gamma}(\ell) = \sum_{t=0}^{\infty} \tilde{B}_{t+\ell} \begin{bmatrix} I_p & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{bmatrix} \tilde{B}_t^\top.$$

Since $\|\tilde{A}^\ell\| = \|SD^\ell S^{-1}\| = \kappa \lambda_{\max}^\ell(\tilde{A}_1)$,

$$\begin{aligned} \|\tilde{\Gamma}(\ell)\| &\leqslant \sum_{t=0}^{\infty} \|\tilde{B}_{t+\ell}\| \|\tilde{B}_t\| \\ &\leqslant \kappa^2 \lambda_{\max}^\ell(\tilde{A}_1) \sum_{k=0}^{\infty} \lambda_{\max}^2(\tilde{A}_1) = \kappa^2 \frac{\lambda_{\max}^\ell(\tilde{A}_1)}{1-\lambda_{\max}^2(\tilde{A}_1)}. \end{aligned}$$

Then noticing $\|\tilde{\Gamma}(\ell)\|_{\max} \leqslant \|\tilde{\Gamma}(\ell)\|$, using (1.A.5)

$$\begin{aligned}\tilde{\Omega}_n &\leqslant 2\sum_{\ell=1}^{n}|\ell|\|\tilde{\Gamma}(\ell)\| \\ &\leqslant 2\kappa^2\sum_{\ell=1}^{n}\frac{\ell\lambda_{\max}^{\ell}(\tilde{A}_1)}{(1-\lambda_{\max}(\tilde{A}_1))(1-\lambda_{\max}^2(\tilde{A}_1))} = 2\kappa^2\frac{\lambda_{\max}(\tilde{A}_1)(1+n\lambda_{\max}^{n+1}(\tilde{A}_1)-(n+1)\lambda_{\max}(\tilde{A}_1))}{(1-\lambda_{\max}(\tilde{A}_1))^2(1-\lambda_{\max}^2(\tilde{A}_1))}.\end{aligned}$$

For $\tilde{L}_n$,

$$\tilde{L}_n \leqslant 2\sum_{\ell>n}\|\tilde{\Gamma}(\ell)\| = 2\kappa^2\sum_{\ell>n}\frac{\lambda_{\max}^{\ell}(\tilde{A}_1)}{1-\lambda_{\max}^2(\tilde{A}_1)} = 2\kappa^2\frac{\lambda_{\max}^{n+1}(\tilde{A})}{(1-\lambda_{\max}(\tilde{A}_1))(1-\lambda_{\max}^2(\tilde{A}_1))}.$$

□

1.A.4 Proof of Proposition 1.3.5

Proof. We focus on bounding the tail probability of the variance term

$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j)-\mathbb{E}\hat{f}_{rs}(\omega_j)\right| \geqslant |||f|||\eta\right).$$

First, note that $\mathbb{P}\left(\left|\hat{f}_{rs}(\omega)-\mathbb{E}\hat{f}_{rs}(\omega)\right| \geqslant |||f|||\eta\right)$ is at most

$$\mathbb{P}\left(\left|\mathbf{Re}\left(\hat{f}_{rs}(\omega)-\mathbb{E}\hat{f}_{rs}(\omega)\right)\right| \geqslant \frac{|||f|||\eta}{2}\right)+\mathbb{P}\left(\left|\mathbf{Im}\left(\hat{f}_{rs}(\omega)-\mathbb{E}\hat{f}_{rs}(\omega)\right)\right| \geqslant \frac{|||f|||\eta}{2}\right),$$

so it is sufficient to derive upper bounds for the real and imaginary parts separately.

The main idea of our proof is to express the real and imaginary parts of $\hat{f}(\omega_j)$ as quadratic forms in $vec(\mathcal{X}^\top)$, and apply Lemma 1.3.2 on each part. First, we express the periodogram $I(\omega_j)$ in terms of trigonometric series. As pointed out before, $I(\omega_j)$ defined in (2.1.3) can be written as

$$\begin{aligned}I(\omega_j) =& \left(\mathcal{X}^\top C_j - i\mathcal{X}^\top S_j\right)\left(\mathcal{X}^\top C_j - i\mathcal{X}^\top S_j\right)^\dagger \\ =& \mathcal{X}^\top(C_jC_j^\top + S_jS_j^\top)\mathcal{X} + i\mathcal{X}^\top(C_jS_j^\top - S_jC_j^\top)\mathcal{X}.\end{aligned} \tag{1.A.6}$$

Note that in the univariate case ($p = 1$) the imaginary part becomes zero and we only need to bound the real part. However, for multivariate case, we need to understand the concentration behaviour of both parts.

Concentration Inequality for Real Part: We claim that there exists a universal constant $c > 0$ s.t. for any two unit vectors u and v,

$$\begin{aligned}&\mathbb{P}\left(\left|u^T\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)v\right| \geqslant |||f|||\eta/2\right)\\ \leqslant& 6\exp\left(-c\min\left\{(2m+1)\eta^2, (2m+1)\eta\right\}\right).\end{aligned}$$

We notice for any symmetric matrix A and unit vectors u and v:

$$2|u^\top Av| \leqslant |u^\top Au| + |v^\top Av| + |(u+v)^\top A(u+v)|. \tag{1.A.7}$$

Now, $\mathbf{Re}\left(\hat{f}(\omega_j)\right) = \mathcal{X}^\top \sum_{|\ell|\leqslant m}(C_{j+\ell}C_{j+\ell}^\top + S_{j+\ell}S_{j+\ell}^\top)\mathcal{X}$ is a symmetric matrix, and so is $\mathbb{E}\left[\mathbf{Re}\left(\hat{f}(\omega_j)\right)\right]$. Thus, $\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)$ is a symmetric matrix. Then applying (1.A.6), we get

$$\begin{aligned}&\mathbb{P}\left(\left|u^\top\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)v\right| \geqslant 1/2|||f|||\eta\right)\\ &\leqslant \mathbb{P}\left(\left|u^\top\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)u\right| \geqslant 1/4|||f|||\eta\right)\\ &+ \mathbb{P}\left(\left|v^\top\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)v\right| \geqslant 1/4|||f|||\eta\right)\\ &+ \mathbb{P}\left(\left|(u+v)^\top\mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right)(u+v)\right| \geqslant 1/2|||f|||\eta\right).\end{aligned} \tag{1.A.8}$$

Next, we note that

$$\mathbf{Re}(\hat{f}(\omega_j)) = \frac{1}{2\pi(2m+1)}\|Q_j\mathcal{X}\|^2,$$

where

$$Q_j := \begin{bmatrix} C_{j-m}^\top \\ S_{j-m}^\top \\ \vdots \\ C_j^\top \\ S_j^\top \\ \vdots \\ C_{j+m}^\top \\ S_{j+m}^\top \end{bmatrix}_{(4m+2)\times n}.$$

Then for any unit vector v,

$$\left|v^\top \mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right) v\right| = \frac{1}{2\pi(2m+1)} \left|v^\top \mathcal{X}^\top Q_j^\top Q_j \mathcal{X} v - \mathbb{E} v^\top \mathcal{X}^\top Q_j^\top Q_j \mathcal{X} v\right|.$$

Let $Y_t = v^\top X_t$, and let $\mathcal{Y} = [Y_1 : \ldots : Y_n]^\top$ be a data matrix with n consecutive observations. Using Lemma 1.C.6, $|||f_Y||| \leqslant \|v\|^2 |||f||| = |||f|||$. Now note that $\text{rk}(Q_j^\top Q_j) \leqslant 4m+2$, and $\|Q_j\| \leqslant \|Q_{F_n}\| = 1$, where Q_{F_n} expands the rows of Q_j to include all the Fourier frequencies (see Lemma 1.C.4 for definition). Since all the rows of Q_j are partially selected from those in Q_{F_n} and Lemma 1.C.4 states that $\|Q_{F_n}\| = 1$, using this bound and applying Lemma 1.3.2, we get

$$\begin{aligned}
&\mathbb{P}\left(\left|v^\top \mathbf{Re}\left(\hat{f}(\omega_j) - \mathbb{E}\hat{f}(\omega_j)\right) v\right| \geqslant 1/4 |||f||| \eta\right) \\
&\leqslant P\left(\frac{1}{2\pi}\left|\mathcal{Y}^\top Q_j^\top Q_j \mathcal{Y} - \mathbb{E}\mathcal{Y}^\top Q_j^\top Q_j \mathcal{Y}\right| \geqslant 1/4 |||f_Y|||(2m+1)\eta\right) \\
&\leqslant 2\exp\left[-c_1 \min\left\{\frac{(2m+1)\eta}{\|Q_j\|^2}, \frac{(2m+1)^2\eta^2}{\text{rk}(Q_j)\|Q_j\|^4}\right\}\right] \\
&\leqslant 2\exp\left[-c_1 \min\left\{(2m+1)\eta, \frac{(2m+1)^2\eta^2}{(4m+2)}\right\}\right] \\
&\leqslant 2\exp\left[-c \min\left((2m+1)\eta^2, (2m+1)\eta\right)\right],
\end{aligned} \tag{1.A.9}$$

where c_1, c are universal constants not depending on n, p or any other model parameters. We can write

$$\begin{aligned}&\mathbb{P}\left(|(u+v)^\top \mathbf{Re}\left(\hat{f}(\omega_j)-\mathbb{E}\hat{f}(\omega_j)\right)(u+v)| \geqslant 1/2|||f|||\eta\right)\\ =&\mathbb{P}\left(\left|\frac{(u+v)}{\sqrt{2}}^\top \mathbf{Re}\left(\hat{f}(\omega_j)-\mathbb{E}\hat{f}(\omega_j)\right)\frac{(u+v)}{\sqrt{2}}\right| \geqslant 1/4|||f|||\eta\right),\end{aligned}$$

with $\frac{u+v}{\sqrt{2}}$ as a unit vector. Thus three terms appearing in right hand side of inequality (1.A.8) can all be bounded by (1.A.9), which completes our proof. Note that when u, v are canonical vectors e_r, e_s respectively, then $(u+v)$ has at most two non-zero entries. Further, since f is non-negative definite, the quantity $|||f_Y|||$ can be upper bounded by a smaller quantity $\max_{1\leqslant r\leqslant p}|||f_r|||$, where f_r denotes the spectral density of the r^{th} component of X_t.

Concentration Inequality for Imaginary Part: We claim that there exists a universal positive constant c such that for any two unit vectors u and v and any $\eta > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|u^\top \mathbf{Im}\left(\hat{f}(\omega_j)-\mathbb{E}\hat{f}(\omega_j)\right)v\right| \geqslant 1/2|||f|||\eta\right)\\ &\leqslant 4\exp\left[-c\min\left\{(2m+1)\eta^2,(2m+1)\eta\right\}\right].\end{aligned}$$

To prove this claim, note that (1.A.6) implies

$$\mathbf{Im}\left(\hat{f}(\omega_j)\right)=\mathcal{X}^\top\sum_{|\ell|\leqslant m}(C_{j+\ell}S_{j+\ell}^\top - S_{j+\ell}C_{j+\ell}^\top)\mathcal{X}.$$

Therefore, for any $\eta > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|u^\top \mathbf{Im}\left(\hat{f}(\omega_j)-\mathbb{E}\hat{f}(\omega_j)\right)v\right| \geqslant 2|||f|||\eta\right)\\ &\leqslant \mathbb{P}\left(\frac{1}{2\pi(2m+1)}\left|u^\top\mathcal{X}^\top\sum_{|\ell|\leqslant m}(S_{j+\ell}C_{j+\ell}^\top)\mathcal{X}v-\mathbb{E}\left[u^\top\mathcal{X}^\top\sum_{|\ell|\leqslant m}(S_{j+\ell}C_{j+\ell}^\top)\mathcal{X}v\right]\right| \geqslant |||f|||\eta\right)\\ &+\mathbb{P}\left(\frac{1}{2\pi(2m+1)}\left|u^\top\mathcal{X}^\top\sum_{|\ell|\leqslant m}(C_{j+\ell}S_{j+\ell}^\top)\mathcal{X}v-\mathbb{E}\left[u^\top\mathcal{X}^\top\sum_{|\ell|\leqslant m}(C_{j+\ell}S_{j+\ell}^\top)\mathcal{X}v\right]\right| \geqslant |||f|||\eta\right).\end{aligned} \tag{1.A.10}$$

It takes the same technique to get upper bound for two parts in the right hand side of inequality (1.A.10). So we will only show the proof for getting upper bound for the first part.

Let $Y_t = [v^\top; u^\top]X_t$ be a 2-dimensional time series. It follows from Lemma 1.C.6 that $|||f_Y||| \leqslant \|[v^\top; u^\top]\|^2 |||f||| = 2|||f|||$.

Define

$$P_j = \begin{bmatrix} M_j & 0 \\ 0 & N_j \end{bmatrix}_{(4m+2)\times 2n}, \tag{1.A.11}$$

where

$$M_j = \begin{bmatrix} S_{j-m}^\top \\ \vdots \\ S_j^\top \\ \vdots \\ S_{j+m}^\top \end{bmatrix}_{(2m+1)\times n} \qquad N_j = \begin{bmatrix} C_{j-m}^\top \\ \vdots \\ C_j^\top \\ \vdots \\ C_{j+m}^\top \end{bmatrix}_{(2m+1)\times n}.$$

We can express the first part in (1.A.10) as

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{2\pi(2m+1)}\left|u^\top \mathcal{X}^\top \sum_{|\ell|\leqslant m}(S_{j+\ell}C_{j+\ell}^\top)\mathcal{X}v - \mathbb{E}\left[u^\top \mathcal{X}^\top \sum_{|\ell|\leqslant m}(S_{j+\ell}C_{j+\ell}^\top)\mathcal{X}v\right]\right| \geqslant 1/2|||f|||\eta\right) \\ &= \mathbb{P}\left(\frac{1}{2\pi(2m+1)}\left|vec(\mathcal{Y}^\top)^\top P_j^\top M P_j vec(\mathcal{Y}^\top) - \mathbb{E}\left[vec(\mathcal{Y}^\top)^\top P_j^\top M P_j vec(\mathcal{Y}^\top)\right]\right| \geqslant 1/2|||f|||\eta\right), \end{aligned} \tag{1.A.12}$$

where

$$M = \begin{bmatrix} 0_{2m+1,2m+1} & I_{2m+1,2m+1} \\ 0_{2m+1,2m+1} & 0_{2m+1,2m+1} \end{bmatrix}.$$

Since M_j and N_j are both composed with rows from Q_{F_n}, $\|M_j\| \leqslant \|Q_{F_n}\| = 1$ and $\|N_j\| \leqslant \|Q_{F_n}\| = 1$. Furthermore, as block-wise diagonal matrix, $\|P_j\| = \max\{\|M_j\|, \|N_j\|\} =$

1. Now $\|P_j^\top M P_j\| \leqslant \|P_j\|^2 \|M\| \leqslant 1$ and $\mathrm{rk}(P_j^\top M P_j) \leqslant \mathrm{rk}(M) = 2m+1$. Since $|||f_Y||| \leqslant 2|||f|||$, we can apply lemma 1.3.2 to show that the probability in (1.A.12) is at most

$$\begin{aligned}
&2\exp\left[-c\min\left\{\frac{(2m+1)\eta}{\|P_j M P_j^\top\|}, \frac{(2m+1)^2\eta^2}{\mathrm{rk}(P_j)\|P_j M P_j^\top\|}\right\}\right] \\
&\leqslant 2\exp\left[-c\min\left\{(2m+1)\eta, \frac{(2m+1)^2\eta^2}{(4m+2)}\right\}\right] \\
&\leqslant 2\exp\left[-c\min\left\{(2m+1)\eta^2, (2m+1)\eta\right\}\right],
\end{aligned}$$

where c is an universal constant.

Combining bounds for real and imaginary parts, and plugging these two bounds in (1.A.6), we can show that there exist universal positive constants c_1, c_2 such that for any $\eta > 0$,

$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right| \geqslant |||f|||\eta\right) \leqslant c_1 \exp\left[-c_2(2m+1)\min\{\eta, \eta^2\}\right]. \quad (1.A.13)$$

□

1.A.5 Proof of Proposition 2.3.7

Proof. For any Hermitian matrix M Golub and Van Loan [2012], we have

$$\|M\| \leqslant \sqrt{\|M\|_1 \|M\|_\infty} = \|M\|_1. \quad (1.A.14)$$

Since both $f(\omega_j)$ and $\hat{f}_\lambda(\omega_j)$ are Hermitian, we can bound spectral norm of estimation error matrix with its maximum absolute column sum norm, i.e.

$$\|\hat{f}_\lambda(\omega_j) - f(\omega_j)\| \leqslant \|\hat{f}_\lambda(\omega_j) - f(\omega_j)\|_1. \quad (1.A.15)$$

Following the proof technique of Theorem 1 in Bickel and Levina [2008], the first step is to bound probability of event

$$A_0 = \left\{ \max_{1 \leqslant r,s \leqslant p} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \geqslant \lambda/2 \right\}.$$

Our goal is to prove that there exist universal constants $c_1,\ c_2$ such that for any $r, s \in \{1, \cdots, p\}$,

$$\mathbb{P}\left(\left| \hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j) \right| \geqslant \frac{\lambda}{2} \right) \leqslant c_1 \exp\left[-c_2 \min\left\{ (2m+1)\eta^2, (2m+1)\eta \right\} \right],$$

where

$$\lambda = 2\left[R|||f|||\sqrt{\log p/m} + \frac{m + 1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f) \right].$$

Then, with union bound, we could get probability bound for $\mathcal{A}_0$.

To accomplish this, we first divide the error into two terms along the line of a bias-variance decomposition.

$$\left| \hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j) \right| \leqslant \left| \mathbb{E}\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j) \right| + \left| \hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j) \right|.$$

Proposition 1.3.3 provides an upper bound on the bias term

$$\left| \mathbb{E}\hat{f}_{rs}(\omega_j) - f_{rs}(\omega) \right| \leqslant \frac{m + 1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)..$$

This bound in bias shows that

$$\mathbb{P}\left(\left| \hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j) \right| \geqslant \lambda/2 \right) \leqslant \mathbb{P}\left(\left| \hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j) \right| \geqslant R|||f|||\sqrt{\frac{\log p}{m}} \right). \tag{1.A.16}$$

Next, proposition 1.3.5 shows that there exists general constants c_1, c_2 s.t. such that for any $\eta > 0$,

$$\mathbb{P}\left(\left| \hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j) \right| \geqslant |||f|||\eta \right) \leqslant c_1 \exp\left[-c_2(2m+1)\min\{\eta, \eta^2\} \right].$$

We set $\eta = R\sqrt{\frac{\log p}{m}}$. Combined with (1.A.16), and noting that we are working in the regime $m \gtrsim \log p$, we conclude $\eta^2 = R^2\frac{\log p}{m} \leqslant \eta = R\sqrt{\frac{\log p}{m}}$. This implies

$$P(A_0) = \mathbb{P}(\max_{r,s} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \geqslant \lambda/2) \leqslant c_1 p^2 \exp\left[-c_2(2m+1)R^2\frac{\log p}{m}\right].$$

This concentration playes essential role in the proof of Theorem 1 as equation (12) in Bickel and Levina [2008]. Theorem 1 in Bickel and Levina [2008] provides the techniques to complete the asymptotic analysis, while here we do some modification to achieve non-asymptotic analysis.

L_2 norm bound: We separate our target into two terms

$$\|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\| \leqslant \|T_\lambda(f(\omega_j)) - f(\omega_j)\| + \|T_\lambda(f(\omega_j)) - T_\lambda(\hat{f}(\omega_j))\|$$

The first term can be bounded by its L_1 norm

$$\begin{aligned} &\|T_\lambda(f(\omega_j)) - f(\omega_j)\| \leqslant \|T_\lambda(f(\omega_j)) - f(\omega_j)\|_1 \\ &\leqslant \max_{r=1}^{p} \sum_{s=1}^{p} |f_{rs}(\omega_j)| \mathbb{1}(|f_{rs}(\omega_j)| < \lambda) \leqslant \lambda^{1-q} |||f|||_q^q, \end{aligned} \tag{1.A.17}$$

for any $0 \leqslant q < 1$.

Then we can upper bound the second term in (1.A.17) by three terms as follows:

$$\begin{aligned} &\|T_\lambda(f(\omega_j)) - T_\lambda(\hat{f}(\omega_j))\| \\ &\leqslant \max_{r=1}^{p} \sum_{s=1}^{p} |\hat{f}_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \leqslant \lambda) \\ &+ \max_{r=1}^{p} \sum_{s=1}^{p} |f_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \leqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\ &+ \max_{r=1}^{p} \sum_{s=1}^{p} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III} \end{aligned}$$

Define three events:

$$A_1 = \left\{ \mathrm{I} \geqslant 3|||\mathrm{f}|||_{\mathrm{q}}^{\mathrm{q}} \lambda^{(1-\mathrm{q})} \right\}$$
$$A_2 = \left\{ \mathrm{II} \geqslant 2|||\mathrm{f}|||_{\mathrm{q}}^{\mathrm{q}} \lambda^{(1-\mathrm{q})} \right\}$$
$$A_3 = \left\{ \mathrm{III} \geqslant |||\mathrm{f}|||_{\mathrm{q}}^{\mathrm{q}} \lambda^{(1-\mathrm{q})} \right\}$$

We will show that on $A_0^{\complement}$, none of these three events can happen, i.e.,

$$A_1 \cup A_2 \cup A_3 \subset A_0.$$

To this end, note that on $A_0^{\complement}$,

$$\begin{aligned} \mathrm{III} &\leqslant \max_r \left| \hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j) \right| \sum_{s=1}^{p} \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) \\ &\leqslant \lambda \sum_{s=1}^{p} \frac{|f_{rs}(\omega_j)|^q}{\lambda^q} \leqslant |||f|||_q^q \lambda^{1-q}. \end{aligned}$$

Here we use the fact that on event $A_0^{\complement}$, $|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \leqslant \frac{\lambda}{2} < \lambda$. Similarly, on $A_0^{\complement}$,

$$\begin{aligned} \mathrm{II} &\leqslant \max_{r=1}^{p} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \sum_{s=1}^{p} \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) + |\hat{f}_{rs}(\omega_j)| \sum_{s=1}^{p} \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \leqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\ &\leqslant \max_{r=1}^{p} \left[\lambda \sum_{s=1}^{p} \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) + \lambda \sum_{s=1}^{p} \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) \right] \leqslant 2|||f|||_q^q \lambda^{1-q}, \end{aligned}$$

where the last inequality follows from the same argument as in (1.A.18). Next, we focus on A_1.

$$\begin{aligned} \mathrm{I} &= \max_{\mathrm{r}=1}^{\mathrm{p}} \sum_{\mathrm{s}=1}^{\mathrm{p}} |\hat{\mathrm{f}}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \mathbb{1}(|\hat{\mathrm{f}}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \geqslant \lambda, |\mathrm{f}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \leqslant \lambda) \\ &\leqslant \max_{r=1}^{p} \sum_{s=1}^{p} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \leqslant \lambda) \\ &+ \max_{r=1}^{p} \sum_{s=1}^{p} |f_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \leqslant \lambda) \\ &= \mathrm{IV} + \mathrm{V}. \end{aligned}$$

A similar argument as above can show that

$$\mathrm{V} \leqslant |||\mathrm{f}|||_{\mathrm{q}}^{\mathrm{q}} \lambda^{1-\mathrm{q}}.$$

For IV, on A_0^{c},

$$\begin{aligned}
\mathrm{IV} &= \max_{\mathrm{r}=1}^{\mathrm{p}} \sum_{\mathrm{s}=1}^{\mathrm{p}} |\hat{\mathrm{f}}_{\mathrm{rs}}(\omega_{\mathrm{j}}) - \mathrm{f}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \mathbb{1}(|\hat{\mathrm{f}}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \geqslant \lambda, |\mathrm{f}_{\mathrm{rs}}(\omega_{\mathrm{j}})| \leqslant \lambda) \\
&= \max_{r=1}^{p} \sum_{s=1}^{p} |\hat{f}_{rs}(\omega) - f_{rs}(\omega_j)| \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda,\ \lambda/2 < |f_{rs}(\omega_j)| \leqslant \lambda) \\
&\leqslant \max_{r=1}^{p} \sum_{s=1}^{p} \lambda \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda/2) \leqslant \max_{r} \sum_{s=1}^{p} \lambda \sum_{s=1}^{p} \frac{|f_{rs}(\omega_j)|^q}{(\lambda/2)^q} \\
&\leqslant 2\lambda^{1-q} |\!|\!| f |\!|\!|_q^q.
\end{aligned}$$

Combining these two parts, we have $\mathrm{I} \leqslant 3\lambda^{1-\mathrm{q}} |\!|\!| \mathrm{f} |\!|\!|_{\mathrm{q}}^{\mathrm{q}}$. Also, since

$$\left\{ \|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\| \geqslant 7\lambda^{1-q} |\!|\!| f |\!|\!|_q^q \right\} \subset A_1 \cup A_2 \cup A_3 \subset A_0,$$

we have

$$\mathbb{P}(\|\hat{f}(\omega_j) - f(\omega_j)\| \geqslant 7\lambda^{1-q} |\!|\!| f |\!|\!|_q^q\}) \leqslant \mathbb{P}(A_0) \leqslant c_1 p^2 \exp\left[-c_2(2m+1)\min\{\eta, \eta^2\}\right].$$

Proof of upper bound on Frobenius norm: Like the proof for operator norm, we decompose the error term as

$$\|T_\lambda(\hat{f})(\omega_j) - f(\omega_j)\|_F^2 \leqslant \|T_\lambda(f(\omega_j)) - f(\omega_j)\|_F^2 + \|T_\lambda(f(\omega_j)) - T_\lambda(\hat{f}(\omega_j))\|_F^2.$$

The same argument for opertator norm then ensures that on A_0^c

$$\begin{aligned}
\|T_\lambda(f) - f\|_F^2 &= \sum_{r,s} |f_{rs}(\omega_j)|^2 \mathbb{1}(|f_{rs}(\omega_j)| \leqslant \lambda) \\
&\leqslant \sum_{r,s} \lambda^{2-q} |f_{rs}(\omega_j)|^q \leqslant \lambda^{2-q} |\!|\!| f |\!|\!|_q^2.
\end{aligned}$$

As before, we decompose the second term in the next step as follows:

$$
\begin{aligned}
&\|T_\lambda(f(\omega_j)) - T_\lambda(\hat{f}(\omega_j))\|_F^2 \\
&\leqslant \sum_{r,s} |\hat{f}_{rs}(\omega_j)|^2 \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \leqslant \lambda) \\
&+ \sum_{r,s} |f_{rs}(\omega_j)|^2 \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \leqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\
&+ \sum_{r,s} |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)|^2 \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \geqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III},
\end{aligned}
$$

and we define following events:

$$
\begin{aligned}
A_1 &= \left\{ \mathrm{I} \geqslant 7\mathrm{p} |\!|\!| \mathrm{f} |\!|\!|_{\mathrm{q}}^{\mathrm{q}} \lambda^{2-\mathrm{q}} \right\} \\
A_2 &= \left\{ \mathrm{II} \geqslant 4\mathrm{p} |\!|\!| \mathrm{f} |\!|\!|_{\mathrm{q}}^{\mathrm{q}} \lambda^{2-\mathrm{q}} \right\} \\
A_3 &= \left\{ \mathrm{III} \geqslant \mathrm{p} |\!|\!| \mathrm{f} |\!|\!|_{\mathrm{q}}^{\mathrm{q}} \lambda^{2-\mathrm{q}} \right\}.
\end{aligned}
$$

We will show that $A_1 \cup A_2 \cup A_3 \subset A_0$ by showing on $A_0^{\complement}$, none of these three events can happen. $\mathrm{III} \leqslant \mathrm{p}\lambda^{2-\mathrm{q}} |\!|\!| \mathrm{f} |\!|\!|_{\mathrm{q}}^{\mathrm{q}}$ is obvious with same techniques before. For II, on A_0,

$$
\begin{aligned}
\mathrm{II} \leqslant & \left[|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)|^2 + |\hat{f}_{rs}(\omega_j)|^2 + 2|\hat{f}_{rs}(\omega_j)||\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \right] \\
& \mathbb{1}(|\hat{f}_{rs}(\omega_j)| \leqslant \lambda, |f_{rs}(\omega_j)| \geqslant \lambda) \\
& \leqslant \sum_{r,s} \lambda^2 \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) + \lambda^2 \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) + 2\lambda^2 \mathbb{1}(|f_{rs}(\omega_j)| \geqslant \lambda) \\
& \leqslant 4p\lambda^{2-q} |\!|\!| f |\!|\!|_q^q.
\end{aligned}
$$

For I, on A_0, we have

$$
\begin{aligned}
\mathrm{I} \leqslant & \sum_{r,s} \left[|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)|^2 + |f_{rs}(\omega_j)|^2 + 2|f_{rs}(\omega_j)||\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \right] \\
& \mathbb{1}(|f_{rs}(\omega_j)| \leqslant \lambda, |\hat{f}_{rs}(\omega_j)| \geqslant \lambda) \\
& = \mathrm{V} + \mathrm{VI} + \mathrm{VII}.
\end{aligned}
$$

Note that on $A_0^{\complement}$, $\mathbb{1}(|f_{rs}(\omega_j)| \leqslant \lambda, |\hat{f}_{rs}(\omega_j)| \geqslant \lambda) = \mathbb{1}(\lambda/2 < |f_{rs}(\omega_j)| \leqslant \lambda, |\hat{f}_{rs}(\omega_j)| \geqslant \lambda)$. Using this, we can show that

$$\mathrm{V} = \lambda^2 \sum_{\mathrm{r,s}} \mathbb{1}(|\mathrm{f_{rs}}(\omega_\mathrm{j})| \geqslant \lambda/2) =\leqslant \lambda^2 \sum_{\mathrm{r,s}} \left(\frac{|\mathrm{f_{rs}}(\omega_\mathrm{j})|^\mathrm{q}}{(\lambda/2)^\mathrm{q}} \right) \leqslant 2\mathrm{p}\lambda^{2-\mathrm{q}} |||\mathrm{f}|||_\mathrm{q}^\mathrm{q}$$

$$\mathrm{VI} \leqslant \sum_{\mathrm{r,s}} |\mathrm{f_{rs}}(\omega_\mathrm{j})|^2 1(|\mathrm{f_{rs}}(\omega_\mathrm{j})| \leqslant \lambda) \leqslant \sum_{\mathrm{r,s}} \left(\frac{\lambda}{|\mathrm{f_{rs}}(\omega_\mathrm{j})|} \right)^{2-\mathrm{q}} |\mathrm{f_{rs}}(\omega_\mathrm{j})|^2 = \mathrm{p}|||\mathrm{f}|||_\mathrm{q}^\mathrm{q} \lambda^{2-\mathrm{q}}$$

$$\mathrm{VII} \leqslant 2\lambda^2 \sum_{\mathrm{r,s}} 1(|\mathrm{f_{rs}}(\omega_\mathrm{j})| \geqslant \lambda/2) =\leqslant 2\lambda^2 \sum_{\mathrm{r,s}} \left(\frac{|\mathrm{f_{rs}}(\omega_\mathrm{j})|^\mathrm{q}}{(\lambda/2)^\mathrm{q}} \right) \leqslant 4\mathrm{p}\lambda^{2-\mathrm{q}} |||\mathrm{f}|||_\mathrm{q}^\mathrm{q}.$$

Thus, we have shown that $\mathrm{I} \leqslant 7\mathrm{p}\lambda^{2-\mathrm{q}} |||\mathrm{f}|||_\mathrm{q}^\mathrm{q}$. Putting all these pieces together, we obtain

$$\left\{ \|T_\lambda(\hat{f}(\omega_j)) - f(\omega_j)\|_F^2 \geqslant 13p\lambda^{2-q} |||f|||_q^q \right\} \subset A_1 \cup A_2 \cup A_3 \subset A_0,$$

which completes the proof. □

1.A.6 Proof of Proposition 1.3.8

Proof. In order to prove the first bound, we note that

$$\begin{aligned}
&\mathbb{P}\left(\exists\, r, s : |T_\lambda(\hat{f}_{rs}(\omega_j))| > 0, f_{rs}(\omega_j) = 0 \right) \\
&\leqslant \mathbb{P}\left(\exists\, r, s : |T_\lambda(\hat{f}_{rs}(\omega_j)) - f_{rs}(\omega_j)| > \lambda \right) \\
&\leqslant p^2 c_1 \exp[-c_2 R^2 \log p].
\end{aligned}$$

where the last inequality comes from proposition 2.3.7.

Now we turn to the second part. Since $\mathcal{S}(\gamma) = \{(r,s) : |f_{rs}(\omega_j)| \geqslant \gamma\lambda\}$ with some $\gamma > 1$.

$$\begin{aligned}
&\mathbb{P}\left(\exists\, (r,s) \in \mathcal{S}(\gamma) : T_\lambda(\hat{f}_{rs}(\omega_j)) = 0, |f_{rs}(\omega_j)| > 0 \right) \\
&\mathbb{P}\left(\exists\, (r,s) \in S(\gamma), |\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| > (\gamma - 1)\lambda \right) \\
&\leqslant p^2 c_1 \exp[-c_2 (\gamma - 1)^2 R^2 \log p].
\end{aligned}$$

The last inequality comes from the following decomposition

$$(\gamma-1)\lambda = 2(\gamma-1)R|||f|||\sqrt{\frac{\log p}{m}} + 2(\gamma-1)\left[\frac{m+1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right],$$

where the second part serves as an upper bound for bias because $\gamma > 1.5$. □

1.A.7 Proof of Proposition 1.3.9

We first build the concentration bound for error terms under asymptotic region stated in the proposition 1.3.9, i.e., there exist universal positive constants c_1, c_2 s.t.

$$\mathbb{P}\left(\max_{r,s}|\hat{g}_{rs}(\omega_j) - g_{rs}(\omega_j)| \geqslant \frac{2\lambda}{\tau}\right) \leqslant c_1 p^2 \exp[-c_2 R\log p]. \tag{1.A.18}$$

Define the events

$$A_0 = \left\{\max_{r,s}|\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)| \geqslant \lambda\right\}$$

and

$$A_1 = \left\{\max_{r,s}|\hat{g}_{rs}(\omega_j) - g_{rs}(\omega_j)| \geqslant 2\lambda/\tau\right\}.$$

We will show that $A_1 \subset A_0$. Since

$$|\hat{g}_{rs}(\omega_j) - g_{rs}(\omega_j)| \leqslant |\hat{g}_{rs}(\omega_j) - \tilde{g}_{rs}(\omega_j)| + |\tilde{g}_{rs}(\omega_j) - g_{rs}(\omega_j)|$$

with $\tilde{g}_{rs}(\omega_j) = \frac{\hat{f}_{rs}(\omega_j)}{\sqrt{f_{rr}(\omega_j)f_{ss}(\omega_j)}}$, it suffices to show that for any r, s,

$$\{|\tilde{g}_{rs}(\omega_j) - g_{rs}(\omega_j)| \geqslant \lambda/\tau\} \subset A_0$$

$$\{|\hat{g}_{rs}(\omega_j) - \tilde{g}_{rs}(\omega_j)| \geqslant \lambda/\tau\} \subset A_0$$

For the first inclusion, note that with $|f_{rr}(\omega_j)| \geqslant \tau$ for $1 \leqslant r \leqslant p$,

$$\begin{aligned}
&\{|g_{rs}(\omega_j) - \tilde{g}_{rs}(\omega_j)| \geqslant \lambda/\tau\} \\
&= \left\{\left|\frac{\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)}{\sqrt{f_{rr}(\omega_j)f_{ss}(\omega_j)}}\right| \geqslant \lambda/\tau\right\} \\
&\subset \left\{\left|\frac{\hat{f}_{rs}(\omega_j) - f_{rs}(\omega_j)}{\tau}\right| \geqslant \lambda/\tau\right\} = A_0.
\end{aligned}$$

Similarly, for the second one,

$$\begin{aligned}
&\{|\hat{g}_{rs}(\omega_j) - \tilde{g}_{rs}(\omega_j)| \geqslant \lambda/\tau\} \\
&= \left\{ |\hat{g}_{rs}(\omega_j)| \left| \sqrt{\frac{\hat{f}_{rr}(\omega_j)\hat{f}_{ss}(\omega_j)}{f_{rr}(\omega_j)f_{ss}(\omega_j)}} - 1 \right| \geqslant \lambda/\tau \right\}.
\end{aligned}$$

Since the averaged periodogram ($\hat{f}(\omega_j)$) is positive semi-definite with positive diagonal elements(almost surely), we have $|\hat{g}_{rs}(\omega_j)| \leqslant 1$. This implies that the above event is a subset of

$$\left\{ \left| \sqrt{\frac{\hat{f}_{rr}(\omega_j)\hat{f}_{ss}(\omega_j)}{f_{rr}(\omega_j)f_{ss}(\omega_j)}} - 1 \right| \geqslant \lambda/\tau \right\}.$$

For all $1 \leqslant r \leqslant p$,

$$\begin{aligned}
&\left\{ \left| \frac{\hat{f}_{rr}(\omega_j)}{f_{rr}(\omega_j)} - 1 \right| \geqslant \frac{\lambda}{\tau} \right\} \\
&= \left\{ \left| \frac{f_{rr}(\omega_j) - f_{rr}(\omega_j)}{f_{rr}(\omega_j)} \right| \geqslant \frac{\lambda}{\tau} \right\} \\
&\subset \{|f_{rr}(\omega_j) - f_{rr}(\omega_j)| \geqslant \lambda\} = A_0.
\end{aligned}$$

This indicates that

$$\left\{ \max_{r=1}^{p} \left| \frac{\hat{f}_{rr}(\omega_j)}{f_{rr}(\omega_j)} - 1 \right| \geqslant \frac{\lambda}{\tau} \right\} \subset A_0.$$

Noticing on the event $A_0^{\complement}$, for all $1 \leqslant r \leqslant p$,

$$\left\{ 1 - \frac{\lambda}{\tau} \leqslant \left| \frac{\hat{f}_{rr}(\omega_j)}{f_{rr}(\omega_j)} \right| \leqslant 1 + \frac{\lambda}{\tau} \right\},$$

with $\lambda/\tau < 1$ (since $\lambda = o(1)$),

$$1 - \frac{\lambda}{\tau} \leqslant \sqrt{\frac{\hat{f}_{rr}(\omega_j)\hat{f}_{ss}(\omega_j)}{f_{rr}(\omega_j)f_{ss}(\omega_j)}} \leqslant 1 + \frac{\lambda}{\tau}, \tag{1.A.19}$$

indicating

$$\left\{ \left| \sqrt{\frac{\hat{f}_{rr}(\omega_j)\hat{f}_{ss}(\omega_j)}{f_{rr}(\omega_j)f_{ss}(\omega_j)}} - 1 \right| \geqslant \lambda/\tau \right\} \subset A_0.$$

This in turn implies

$$\{|\hat{g}_{rs}(\omega_j) - \tilde{g}_{rs}(\omega_j)| \geqslant \lambda/\tau\} \subset A_0.$$

Combining two inclusion relations, we can claim that

$$\left\{|\hat{g}_{rs}(\omega_j) - g_{rs}(\omega_j)| \geqslant \frac{2\lambda}{\tau}\right\} \subset A_0,$$

which completes building the concentration inequality for event A_1 since proposition 1.3.8 presents the concentration inequality for event A_0. Then following the argument in proof of proposition 1.3.8, we could complete the proof.

1.B Appendix: Proofs for Linear Processes

1.B.1 Proof for Lemma 1.4.1

Proof. Proof for sub-Gaussian case is given by Rudelson and Vershynin [2013] and proof for the sub-exponential case is given by Lemma 8.3 in Erdős et al. [2012]. We will show the proof for case (3) based on Markov inequality. We will show tail bound for both diagonal part and non-diagonal part for any $\eta > 0$ one by one. For diagonal part, let $y_i = \varepsilon_{ii}^2 - 1$. Then $\mathbb{E}y_i = 0$ and $\mathbb{E}y_i^2 = \mathbb{E}\varepsilon_{ii}^4 - 2\mathbb{E}\varepsilon_{ii}^2 + 1 \leqslant K - 1 < K$. Therefore, noticing $\mathbb{E}\varepsilon^\top A\varepsilon = \text{tr}(A)$ under this setting,

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{i=1}^n \varepsilon_{ii}^2 A_{ii} - \text{tr}(A)\right| \geqslant \eta\right] &= \mathbb{P}\left[\left|\sum_{i=1}^n y_i A_{ii}\right| \geqslant \eta\right] \\ &\leqslant \frac{\mathbb{E}(\sum_{i=1}^n y_i A_{ii})^2}{\eta^2} \leqslant \frac{K\sum_{i=1}^n A_{ii}^2}{\eta^2},\end{aligned}$$

where the second last inequality follows from $\mathbb{E}y_i y_j = 0$. For the non-diagonal part, note that

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{1\leqslant i\neq j\leqslant n} A_{ij}\varepsilon_i\varepsilon_j\right| \geqslant \eta\right] &\leqslant \frac{\mathbb{E}\left|\sum_{1\leqslant i\neq j\leqslant n} A_{ij}\varepsilon_i\varepsilon_j\right|^2}{\eta^2} \\ &= \frac{\sum_{1\leqslant i\neq j\leqslant n} A_{ij}^2(\mathbb{E}\varepsilon_1^2)^2}{\eta^2} + \frac{\sum_{1\leqslant i\neq j\leqslant n} A_{ij}A_{ji}(\mathbb{E}\varepsilon_1^2)^2}{\eta^2} \\ &\leqslant \frac{2\sum_{1\leqslant i\neq j\leqslant n} A_{ij}^2}{\eta^2}.\end{aligned}$$

Here the second line holds since $\mathbb{E}\varepsilon_i\varepsilon_j\varepsilon_p\varepsilon_q \neq 0$ iff $i = p, j = q$ or $i = q, j = p$ and the third line comes from the simple fact that $A_{ij}A_{ji} \leqslant \frac{1}{2}(A_{ij}^2 + A_{ji}^2)$.

Then plugging $\frac{\eta}{2}$ into above two parts, we get

$$\begin{aligned}&\mathbb{P}\left[|\varepsilon^\top A\varepsilon - \mathbb{E}\varepsilon^\top A\varepsilon| \geqslant \eta\right] \\ \leqslant&\mathbb{P}\left[\left|\sum_{i=1}^n \varepsilon_{ii}^2 A_{ii} - \mathrm{tr}(A)\right| \geqslant \eta/2\right] + \mathbb{P}\left[\left|\sum_{1\leqslant i\neq j\leqslant n} A_{ij}\varepsilon_i\varepsilon_j\right| \geqslant \eta/2\right] \\ \leqslant&\max\{4K, 8\}\frac{\|A\|_F^2}{\eta^2},\end{aligned}$$

where we can set $c_3 = \max\{4K, 8\}$ and use the fact $\|A\|_F^2 \leqslant \mathrm{rk}(A)\|A\|^2$ to complete our proof. □

1.B.2 Proof of Proposition 1.4.2

Proof. The proofs of the above inequalities for these three cases follow a common structure. We work with fixed values of n and p, and construct a limiting argument as $L \to \infty$. In the first step, we apply inequality in Lemma 1.4.1 to the truncated process $X_{(L),t} = \sum_{\ell=0}^L B_\ell\varepsilon_{t-\ell}$, for some $L > 0$. Then we show that this inequality holds in the limit $L \to \infty$. For the sake of brevity, we only present the proof for sub-Gaussian case here.

Let $\mathcal{X}_{(L)}$ be a $n \times p$ data matrix with n consecutive observations from process $\{X_{(L),t}\}_{t\in\mathbb{Z}}$. We can write $vec(\mathcal{X}_{(L)}^\top) = \Pi_L E_n$ where

$$\Pi_L = \begin{bmatrix} 0 & 0 & \dots & 0 & B_0 & B_1 & \dots & B_{L-1} & B_L \\ 0 & 0 & \dots & B_0 & B_1 & B_2 & \dots & B_L & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ B_0 & B_1 & \dots & \dots & \dots & \dots & B_L & 0 & 0 \end{bmatrix}$$

and $E_n = (\varepsilon_n^\top, \dots, \varepsilon_{1-L}^\top)^\top$. Without loss of generality, we assume $L > n$ in our representation of Π_L and E_n. It follows from Lemma 1.C.5 that $\|\text{Cov}(vec(\mathcal{X}_{(L)}^\top), vec(\mathcal{X}_{(L)}^\top))\| = \|\Pi_L \Pi_L^\top\| \leqslant \|\!|\!| f_{(L)} |\!|\!|$, where $f_{(L)}(\omega)$ is the spectral density of $X_{(L),t}$. Then using the same technique as in the proof of Lemma 1.3.2 and inequality for sub-Gaussian i.i.d. case introduced in Lemma 1.4.1, we get

$$\begin{aligned} &\mathbb{P}\left(\left|vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top) - \mathbb{E}\left[\, vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top)\right]\right| > 2\pi\eta \|\!|\!| f_{(L)} |\!|\!|\right) \\ &\leqslant 2\exp\left[-c\min\left\{\frac{\eta}{\|A\|}, \frac{\eta^2}{\text{rk}(A)\|A\|^2}\right\}\right]. \end{aligned} \tag{1.B.1}$$

Next we note that by Lemma 1.C.8, for any fixed n, p, $vec(\mathcal{X}_{(L)}^\top) \xrightarrow{L_2} vec(\mathcal{X}^\top)$ as $L \to \infty$. Since L_2 convergenece implies convergence in probability, by continuous mapping theorem, we have

$$vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top) \xrightarrow{\mathbb{P}} vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top) \tag{1.B.2}$$

as $L \to \infty$. The L_2-norm convergence also ensures L_1-norm convergence, which implies

$$\mathbb{E}\left[\, vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top)\right] \to \mathbb{E}\left[\, vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top)\right]. \tag{1.B.3}$$

A detailed derivation is outlined in the remarks after Lemma 1.C.8. Together with Lemma 1.C.9, we obtain $2\pi\eta\|\!|\!| f_{(L)} |\!|\!| \to 2\pi\eta\|\!|\!| f |\!|\!|$. Putting pieces together, we have

$$vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top) - \mathbb{E}\left[\, vec(\mathcal{X}_{(L)}^\top)^\top A\, vec(\mathcal{X}_{(L)}^\top)\right] - 2\pi\eta\|\!|\!| f_{(L)} |\!|\!|$$

converges in probability, and hence in distribution, to

$$vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top) - \mathbb{E}\left[\, vec(\mathcal{X}^\top)^\top A\, vec(\mathcal{X}^\top)\right] - 2\pi\eta|||f|||.$$

Thus, if we take $L \to \infty$ from both sides in (1.B.1), we obtain the final bound. □

1.C Appendix: Additional Proofs of Technical Results

Lemma 1.C.1. *For any matrix $A \in \mathbb{C}^{p\times p}$ and $0 \leqslant q < 1$, define $\|A\|_q := \max_{\|x\|_q=1} \|Ax\|_q$, where q norm for vector is defined as $\|x\|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$ for any vector x of length p(Again, it is indeed a norm iff $q \geqslant 1$). Then*

$$\max_{s=1}^{p} \sum_{r=1}^{p} |A_{rs}|^q = \|A\|_q^q.$$

Proof. First, for two vectors $v_1, v_2 \in \mathbb{C}^p$, $\|v_1 + v_2\|_q^q \leqslant \|v_1\|_q^q + \|v_2\|_q^q$ for $0 \leqslant q < 1$, since for scalars $x, y \in \mathbb{C}$, $|x+y|^q \leqslant |x|^q + |y|^q$. Then let A_i be the i^{th} column of A. Based on the definition of $\|A\|_q$, we have

$$\begin{aligned}
\|A\|_q^q &= \max_{\|x\|_q=1} \|\sum_{i=1}^{p} A_i x_i\|_q^q \\
&\leqslant \max_{\|x\|_q=1} \sum_{i=1}^{p} \|A_i x_i\|_q^q = \sum_{i=1}^{p} |x_i|^q \|A_i\|_q^q \\
&\leqslant (\max_{i=1}^{p} \|A_i\|_q^q) \sum_{i=1}^{p} \|x_i\|^q = \max_{i=1}^{p} \|A_i\|_q^q.
\end{aligned}$$

Noticing if we set x above as the indicator vector e_r, where $r = \mathrm{argmax}_i \|A_i\|^q$, the equality holds, we finish the proof. □

Lemma 1.C.2. *For any matrix $A \in \mathbb{R}^{p\times p}$, and a positive constant ϵ, we could find a matrix E such that $A + E$ has distinct eigenvalues and $\|E\| \leqslant \epsilon$.*

Proof. Consider the Schur decomposition([Golub and Van Loan, 2012]) of A as $A = QUQ^\dagger$ where Q is an unitary matrix and U is an upper triangular matrix. Construct a diagonal matrix D with each element less than ϵ and make $U_{i,i} + D_{i,i}$ distinct. Set $E = QDQ^\dagger$, we have $A + QDQ^\dagger = Q(E + D)Q^\dagger$ with eigenvalues as $U_{i,i} + D_{i,i}, i = 1, \dots p$ which are distinct. By setting $E = QDQ^\dagger$ and noticing $\|E\| = \|QDQ^\dagger\| = \|D\| \leqslant \epsilon$ we complete the proof. □

Remark. *$\lambda_{max}(A)$ is continuous mapping from the set of $p \times p$ complex matrices to the set of real numbers. Thus, we can always find perturbation $\|E\|$ small enough to guarantee $\|A + E\| < 1$. To quantify this, we can apply the result from Bhatia et al. [1990] for perturbation bound on potentially non-symmetric matrices*

$$|\lambda_{\max}(A + E) - \lambda_{\max}(A)| \leqslant 12\|A\|^{1-1/p}\|E\|^{1/p}. \tag{1.C.1}$$

Lemma 1.C.3. *For any j, k in F_n, the inner product between C_j and S_k can only have the following forms:*

(a) $C_j^\top S_k = 0$

(b)

$$C_j^\top C_k = 0 \text{ if } |j| \neq |k|; \quad C_j^\top C_j = \begin{cases} 1 & \text{if } j \in \{0, \frac{n}{2}\} \\ \frac{1}{2} & \text{otherwise} \end{cases}; \quad C_j^\top C_{-j} = \begin{cases} 1 & \text{if } j = 0 \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

(c)

$$S_j^\top S_k = 0 \text{ if } |j| \neq |k|; \quad S_j^\top S_j = \begin{cases} 0 & \text{if } j \in \{0, \frac{n}{2}\} \\ \frac{1}{2} & \text{otherwise} \end{cases}; \quad S_j^\top S_{-j} = \begin{cases} 0 & \text{if } j = 0 \\ -\frac{1}{2} & \text{otherwise.} \end{cases}$$

Proof. We first state Lagrange's trigonometric identities:

$$\sum_{\ell=1}^{n}\cos(\ell\theta) = \begin{cases} n & \theta = 2k\pi \text{ for some integer } k \\ -\frac{1}{2} + \frac{\sin\left(n+\frac{1}{2}\right)\theta}{2\sin\frac{\theta}{2}} & \text{otherwise} \end{cases} \tag{1.C.2}$$

and

$$\sum_{\ell=1}^{n}\sin(\ell\theta) = \begin{cases} 0 & \theta = 2k\pi \text{ for some integer } k \\ \frac{\cos(\frac{1}{2}\theta)}{2\sin(\frac{1}{2}\theta)} - \frac{\cos\left(n+\frac{1}{2}\right)\theta}{2\sin\frac{\theta}{2}} & \text{otherwise} \end{cases} \tag{1.C.3}$$

Now we consider a special case where we set $\theta = \omega_j = \frac{2j\pi}{n}, j \in \mathbb{Z}$. Here we relax $j \in F_n$ to all integers. After this relaxation, we can write $\omega_j + \omega_k = \omega_{j+k}$ and $\omega_j - \omega_k = \omega_{j-k}$. Using (1.C.2) and (1.C.3), for any ω_j, $j \in \mathbb{Z}$, and fixed n, we have the following identities

$$\sum_{\ell=1}^{n}\cos(\ell\omega_j) = \begin{cases} n & \text{if } j \equiv 0 \pmod{n} \\ 0 & \text{otherwise} \end{cases}, \tag{1.C.4}$$

$$\sum_{\ell=1}^{n}\sin(\ell\omega_j) = 0. \tag{1.C.5}$$

Now we prove (a), (b) and (c).

(a) For any j and k in F_n, (1.C.5) implies

$$\begin{aligned} C_j^\top S_k &= \frac{1}{2n}\sum_{\ell=1}^{n}[\sin(\ell(\omega_j + \omega_k)) - \sin(\ell(\omega_k - \omega_j))] \\ &= \frac{1}{2n}\left[\sum_{\ell=1}^{n}\sin(\ell\omega_{j+k}) - \sum_{\ell=1}^{n}\sin(\ell\omega_{k-j})\right] = 0 \end{aligned} \tag{1.C.6}$$

(b) For any $j, k \in F_n$,

$$C_j^\top C_k = \frac{1}{2n}\left(\sum_{\ell=1}^{n}\cos(\ell\omega_{j+k}) + \sum_{\ell=1}^{n}\cos(\ell\omega_{j-k})\right) \tag{1.C.7}$$

For the case $j = k$ or $j = -k$, we have

$$C_j^\top C_k = \frac{1}{2n}\left(\sum_{\ell=1}^{n}\cos(\ell\omega_{2k}) + \sum_{\ell=1}^{n}\cos(\ell\omega_0)\right). \tag{1.C.8}$$

(1.C.2) implies that if $j \in \left\{0, \frac{n}{2}\right\}$, (1.C.8) is 1. In other cases, $0 < 2k < n$, (1.C.4) implies that $\sum_{\ell=1}^{n}\cos(\ell\omega_{2k}) = 0$ which further shows that the right hand side in (1.C.8) is $1/2$.

For the other cases, since $-n < j + k < n$ and $-n < j - k < n$, $j + k \not\equiv 0 \pmod{n}$ and $j - k \not\equiv 0 \pmod{n}$, the right hand side in equation (1.C.7) becomes 0.

(c) for any $j, k \in F_n$,

$$C_j^\top C_k + S_j^\top S_k = \frac{1}{n}\sum_{\ell=1}^{n}\cos(\ell\omega_j)\cos(\ell\omega_k) + \sin(\ell\omega_j)\sin(\ell\omega_k) = \frac{1}{n}\sum_{\ell=1}^{n}\cos(\ell\omega_{j-k}) \tag{1.C.9}$$

If $k = j$, the right hand side in (1.C.9) is 1 and in other cases, the right hand side is 0. Then plugging in the value of $C_j^\top C_k$ listed in case (a), we complete our proof for case (c).

□

Lemma 1.C.4. $\|Q_{F_n}\| = 1$ *where*

$$Q_{F_n} = \begin{bmatrix} C^\top_{-[\frac{n-1}{2}]} \\ S^\top_{-[\frac{n-1}{2}]} \\ \vdots \\ C^\top_{[\frac{n}{2}]} \\ S^\top_{[\frac{n}{2}]} \end{bmatrix} \tag{1.C.10}$$

and each $C_j, S_j, j \in F_n$ *follow the definition in* (2.1.4)

Proof. Since row permutation does not change the L_2 norm of a matrix, we can stack rows in Q_{F_n} such that S_j, C_j, S_{-j}, C_{-j} appear adjacently, if there exists such a pair $\{j, -j\}$. Then $\|Q_{F_n}\| = \|Q_{F_n}^\top\| = \sqrt{\lambda_{\max}(Q_{F_n}Q_{F_n}^\top)}$. Lemma 1.C.3 implies that $Q_{F_n}Q_{F_n}^\top$ can only be block-wise diagonal with three possible blocks:

$$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \quad B_3 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

Here B_1 corresponds to the block formed with C_0, S_0, B_2 corresponds to the block formed of $C_j, S_j, C_{-j}, S_{-j}, j \neq 0$ and B_3 corresponds to the block formed of single j: C_j, S_j. It can be checked that $\|B_i\| \leqslant 1$ for $i = 1, 2, 3$. It follows that $\|Q_{F_n}\| = \sqrt{\lambda_{\max}(Q_{F_n}Q_{F_n}^\top)} \leqslant \max_{i=1}^3 \|B_i\| = 1$, completing our proof. □

Lemma 1.C.5.

$$\|Cov(vec(\mathcal{X}^\top), vec(\mathcal{X}^\top))\| \leqslant 2\pi |||f|||,$$

where $|||f||| = \operatorname{ess\,sup}_{\omega\in[-\pi,\pi]} \|f(\omega)\|$.

Proof. The proof follows from Proposition 2.3 in Basu and Michailidis [2015]. □

Lemma 1.C.6. *For any matrix $A_{p\times m}$, the time series $Y_t = A^\top X_t$ satisfies*

$$|||f_Y||| \leqslant \|A\|^2 |||f|||.$$

Proof. The autocovariance function of the time series Y_t can be written as

$$\Gamma_Y(\ell) = \text{Cov}(A^\top X_t, A^\top X_{t-\ell}) = A^\top \Gamma_X(\ell) A, \tag{1.C.11}$$

which immediately leads to

$$f_Y(\omega) = \sum_{\ell=-\infty}^{\infty} A^\top \Gamma_X(\ell) A e^{-i\omega\ell} = A^\top f(\omega) A. \tag{1.C.12}$$

Thus for any $\omega \in [-\pi, \pi]$, $\|f_Y(\omega)\| \leqslant \|A\|^2 |||f|||$. Taking supremum over ω on the left side completes the proof. □

Lemma 1.C.7. *For stationary linear processes $\Gamma(\ell)$ is well defined, and Assumption 2.1.1 holds.*

Proof. Since $(\sum_{i=1}^n |a_i|)^2 \geqslant \sum_{i=1}^n a_i^2$, we have

$$\sum_{\ell=0}^{\infty} \|B_\ell\|_F \leqslant \sum_{\ell=0}^{\infty} \sum_{1 \leqslant i,j \leqslant p} |B_{\ell,(i,j)}| < \infty.$$

Then by equivalence of norms, it follows that

$$\sum_{\ell=0}^{\infty} \|B_\ell\| < \infty. \tag{1.C.13}$$

Noticing for $h > 0$, $\Gamma(h) = \Gamma^\top(-h)$, we have $\|\Gamma(h)\| = \|\Gamma(-h)\|$ for $h \geqslant 0$. Therefore,

$$\begin{aligned} \sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\| &\leqslant 2 \sum_{\ell=0}^{\infty} \|\Gamma(\ell)\| = 2 \sum_{\ell=0}^{\infty} \| \sum_{t=0}^{\infty} B_{t+\ell} B_t^\top \| \\ &< 2 \sum_{\ell=0}^{\infty} \sum_{t=0}^{\infty} \|B_{\ell+t}\| \|B_\ell^\top\| = 2 \sum_{t_1=0}^{\infty} \sum_{t_2=0}^{\infty} \|B_{t_1}\| \|B_{t_2}\| = 2 \left[\sum_{t=0}^{\infty} \|B_t\| \right]^2 < \infty. \end{aligned} \tag{1.C.14}$$

□

Lemma 1.C.8.

$$\lim_{L \to \infty} \mathbb{E} \left[\|vec(\mathcal{X}_{(L)}^\top) - vec(\mathcal{X}^\top)\|^2 \right] = 0,$$

where $\mathcal{X}_{n \times p} = [X_1 : \ldots : X_n]^\top$ is a $n \times p$ data matrix with n consecutive observations from a stationary linear process defined in (1.4.1)*.*

Proof. Since

$$\|vec(\mathcal{X}_{(L)}^\top) - vec(\mathcal{X}^\top)\|^2 = \sum_{t=1}^{n} \|X_t - X_{(L),t}\|^2,$$

it suffices to show that $\lim_{L\to\infty}\mathbb{E}\left[\|X_{(L),t}-X_t\|^2\right]=0$ for any given $t\in\{1,\dots,n\}$. It follows that

$$\|X_{(L),t}-X_t\|^2=\sum_{\ell_1=L+1}^{\infty}\sum_{\ell_2=L+1}^{\infty}\varepsilon_{t-\ell_1}^{\top}B_{\ell_1}^{\top}B_{\ell_2}\varepsilon_{t-\ell_2}\leqslant\sum_{\ell_1=0}^{\infty}\sum_{\ell_2=0}^{\infty}\|B_{\ell_1}\|\|B_{\ell_2}\|\|\varepsilon_{t-\ell_1}\|\|\varepsilon_{t-\ell_1}\|. \tag{1.C.15}$$

Since each coordinate of ε_t has finite second moment (1 to be precise), we let $\mathbb{E}\|\varepsilon_{t-\ell_1}\|=c_\varepsilon<\infty$. Then the expected value of right part in (1.C.15) is

$$c_\varepsilon^2\sum_{\ell_1=0}^{\infty}\sum_{\ell_2=0}^{\infty}\|B_{\ell_1}\|\|B_{\ell_2}\|=c_\varepsilon^2(\sum_{\ell=0}^{\infty}\|B_\ell\|)^2<\infty,$$

where the last inequality was established in the proof of lemma 1.C.7. Then we apply dominated convergence theorem to show that

$$\begin{aligned}\mathbb{E}\left[\|X_{(L),t}-X_t\|^2\right]&=\sum_{\ell_1=L+1}^{\infty}\sum_{\ell_2=L+1}^{\infty}\mathbb{E}\left[\varepsilon_{t-\ell_1}^{\top}B_{\ell_1}^{\top}B_{\ell_2}\varepsilon_{t-\ell_2}\right]\\&=\sum_{\ell=L+1}^{\infty}\mathbb{E}\left[\varepsilon_{t-\ell}^{\top}B_{\ell}^{\top}B_{\ell}\varepsilon_{t-\ell}\right]\leqslant c_\varepsilon^2(\sum_{\ell=L+1}^{\infty}\|B_\ell\|)^2,\end{aligned}$$

because $\sum_{\ell=0}^{\infty}\|B_\ell\|<\infty$, above goes to zero when $L\to\infty$. □

Remark. *The above convergence argument immediately implies several useful results,*

1 $vec(\mathcal{X}_{(L)}^{\top})\xrightarrow{\mathbb{P}}vec(\mathcal{X}^{\top})$

2 *For any real matrix* $A_{np\times np}$,

$$\lim_{L\to\infty}\mathbb{E}\left[vec(\mathcal{X}_{(L)}^{\top})^{\top}A\,vec(\mathcal{X}_{(L)}^{\top})\right]=\mathbb{E}\left[vec(\mathcal{X}^{\top})^{\top}A\,vec(\mathcal{X}^{\top})\right]$$

This is because

$$\begin{aligned}&\left|\mathbb{E}\left[vec(\mathcal{X}_{(L)}^{\top})^{\top}A\,vec(\mathcal{X}_{(L)}^{\top})\right]-\mathbb{E}\left[vec(\mathcal{X}^{\top})^{\top}A\,vec(\mathcal{X}^{\top})\right]\right|\\&\leqslant\left|\mathbb{E}\left[vec(\mathcal{X}_{(L)}^{\top})^{\top}A\left(vec(\mathcal{X}_{(L)}^{\top})-vec(\mathcal{X}^{\top})\right)\right]\right|+\left|\mathbb{E}\left[\left(vec(\mathcal{X}_{(L)}^{\top})-vec(\mathcal{X}^{\top})\right)^{\top}A\,vec(\mathcal{X}^{\top})\right]\right|.\end{aligned} \tag{1.C.16}$$

Applying Cauchy-Schwarz inequality to the first part in second line of (1.C.16)*, we get*

$$\begin{aligned}&\left|\mathbb{E}\left[vec(\mathcal{X}_{(L)}^{\top})^{\top}A\left(vec(\mathcal{X}_{(L)}^{\top})-vec(\mathcal{X}^{\top})\right)\right]\right|^{2}\\&\leqslant\|A\|\mathbb{E}\left[\|vec(\mathcal{X}_{(L)}^{\top})\|^{2}\right]\mathbb{E}\left[\left\|\left(vec(\mathcal{X}_{(L)}^{\top})-vec(\mathcal{X}^{\top})\right)\right\|^{2}\right].\end{aligned}$$

In addition, from Lemma 1.C.8, we have $\mathbb{E}\left[\|vec(\mathcal{X}_{(L)}^{\top})\|^{2}\right]\rightarrow\mathbb{E}\left[\|vec(\mathcal{X}^{\top})\|^{2}\right]$ *and*

$\mathbb{E}\left[\|\left(vec(\mathcal{X}_{(L)}^{\top})-vec(\mathcal{X}^{\top})\right)\|^{2}\right]\rightarrow 0$. *This implies that the first part in second line of* (1.C.16) *converges to zero when* L *goes to infinity. A similar argument ensures that the second part in second line of* (1.C.16) *goes to zero as well, completing our proof.*

Lemma 1.C.9. $\lim_{L\rightarrow\infty}\left|\left|\left|f_{(L)}\right|\right|\right|=|||f|||$.

Proof. Let $\Gamma_{(L)}(h)$ and $f_{(L)}(\omega)$ be the autocovariance function and spectral density of the truncated process $X_{(L),t}$. We list expressions for $\Gamma_{(L)}(h)$ and $\Gamma(h)$ in order to make a comparison later where we focus on the case $h>0$ (as pointed before, $\Gamma(h)=\Gamma^{\top}(-h)$ for $h>0$)

$$\Gamma(h)=\mathbb{E}X_{t}X_{t-h}^{\top}=\mathbb{E}\left(\sum_{\ell=0}^{\infty}B_{\ell}\varepsilon_{t-\ell}\right)\left(\sum_{\ell=0}^{\infty}B_{\ell}\varepsilon_{t-h-\ell}\right)^{\top}=\sum_{\ell=0}^{\infty}B_{\ell+h}B_{\ell}^{\top},$$

and

$$\Gamma_{L}(h)=\mathbb{E}X_{t}X_{t-h}^{\top}=\mathbb{E}\left(\sum_{\ell=0}^{L}B_{\ell}\varepsilon_{t-\ell}\right)\left(\sum_{\ell=0}^{L}B_{\ell}\varepsilon_{t-h-\ell}\right)^{\top}=\sum_{\ell=0}^{L-h}B_{\ell+h}B_{\ell}^{\top},$$

which indicates $\Gamma_{L}(h)=0$ if $h>L$.

Now we show that $\|\Gamma_{(L)}(h)-\Gamma(h)\|$ goes to zero with $L\rightarrow\infty$. Since we consider L goes to inftty, we assume $L>|h|$. Without losing generality, for any given positive

integer h,

$$\begin{aligned}\lim_{L\to\infty}\|\Gamma_{(L)}(h)-\Gamma(h)\| &= \lim_{L\to\infty}\left\|\sum_{\ell=L-h+1}^{\infty}B_{\ell+h}B_{\ell}^{\top}\right\| \\ &\leqslant \lim_{L\to\infty}\sum_{\ell=L-h+1}^{\infty}\|B_{\ell}\|\|B_{\ell+h}\| \\ &\leqslant \lim_{L\to\infty}\left(\sum_{\ell=0}^{\infty}\|B_{\ell}\|\right)\left(\sum_{\ell=L+1}^{\infty}\|B_{\ell}\|\right)=0.\end{aligned}$$

The last equality comes from the fact that $\sum_{\ell=0}^{\infty}\|B_{\ell}\|<\infty$. Considering the relation that $\Gamma(h)=\Gamma^{\top}(-h)$, for $h<0$, following also holds:

$$\lim_{L\to\infty}\|\Gamma_L(h)-\Gamma(h)\|=0. \tag{1.C.17}$$

Based on the expression of $\Gamma(h)$ and $\Gamma_L(h)$, we have

$$\max\left\{\sum_{h=-\infty}^{\infty}\|\Gamma(h)\|,\sum_{h=-\infty}^{\infty}\|\Gamma_{(L)}(h)\|\right\}\leqslant 2(\sum_{\ell=0}^{\infty}\|B_{\ell}\|)^2<\infty,$$

which in turn implies

$$\sum_{h=-\infty}^{\infty}\|\Gamma_{(L)}(h)-\Gamma(h)\|\leqslant 2(\sum_{\ell=0}^{\infty}\|B_{\ell}\|)^2<\infty. \tag{1.C.18}$$

Therefore, by dominant convergence theorem,

$$\begin{aligned}&\lim_{L\to\infty}\operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]}\|f_{(L)}(\omega)-f(\omega)\|\leqslant\lim_{L\to\infty}\sum_{h=-\infty}^{\infty}\|\Gamma_{(L)}(h)-\Gamma(h)\| \\ &=\sum_{h=-\infty}^{\infty}\lim_{L\to\infty}\|\Gamma_{(L)}(h)-\Gamma(h)\|=0,\end{aligned}$$

Finally

$$\begin{aligned}\lim_{L\to\infty}\left|\|\!|f_{(L)}|\!\|-\|\!|f|\!\|\right| &= \lim_{L\to\infty}\left|\operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]}\|f_{(L)}(\omega)\|-\operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]}\|f(\omega)\|\right| \\ &\leqslant \lim_{L\to\infty}\operatorname*{ess\,sup}_{\omega\in[-\pi,\pi]}\|f_{(L)}(\omega)-f(\omega)\|=0,\end{aligned}$$

which completes the proof. □

1.D Appendix: Additional Table and Graphs

This section contains a table on precision, recall and F1 measures of the three different types of thresholding methods in selecting the non-zero entries of the spectral density matrices of VMA and VAR models of different dimension using different sample sizes. The simulation settings are described in Section 4.4.

We also present enlarged images of the adjacency matrices of coherence networks obtained using adaptive lasso thresholding and shrinkage methods on the real data analysis in Section 1.6. These images contain names of brain regions so that interesting strong connectivity patterns between regions can be identified easily.

		Hard Thresholding			Lasso			Adaptive Lasso		
VMA		precision	recall	F1	precision	recall	F1	precision	recall	F1
p = 12										
	n = 100	93.07(2.79)	55.95(2.58)	68.8(1.93)	80.1(4.87)	75.97(4.19)	72.7(2.54)	93.07(2.79)	61.04(3.62)	70.55(2.07)
	n = 200	92.31(2.45)	68.04(2.84)	77.65(2.1)	75.53(4.11)	91.07(2.34)	79.25(2.54)	92.31(2.45)	75.92(2.88)	80.35(1.92)
	n = 400	91.07(2.15)	83.72(1.96)	88.4(1.51)	70.37(3.39)	98.16(0.68)	79.83(2.52)	91.07(2.15)	91.32(1.61)	89.67(1.6)
	n = 600	91.89(1.82)	90.02(1.59)	92.75(1.06)	69.73(3.74)	99.4(0.3)	80.13(2.77)	91.89(1.82)	95.68(1.02)	92.83(1.28)
p = 24										
	n = 100	97.93(0.98)	45.58(0.94)	62.33(0.81)	88.59(3.35)	63.0(3.6)	70.4(1.89)	97.93(0.98)	50.07(2.04)	65.36(1.44)
	n = 200	96.12(1.36)	52.79(1.6)	68.13(1.2)	80.36(4.06)	84.67(2.81)	79.53(2.22)	96.12(1.36)	66.14(2.77)	76.51(1.72)
	n = 400	94.62(1.0)	71.14(2.26)	81.57(1.53)	74.21(2.63)	96.7(0.88)	82.47(1.74)	94.62(1.0)	86.86(1.81)	89.47(1.1)
	n = 600	94.56(1.1)	81.6(1.73)	88.68(1.11)	71.44(2.66)	99.0(0.34)	81.79(1.9)	94.56(1.1)	93.81(0.95)	93.64(0.79)
p = 48										
	n = 100	99.42(0.37)	43.4(0.23)	60.5(0.22)	94.44(1.85)	52.88(1.96)	66.51(1.14)	99.42(0.37)	45.31(0.75)	62.08(0.61)
	n = 200	98.57(0.51)	45.51(0.56)	62.4(0.48)	87.23(2.05)	75.2(2.38)	78.7(1.31)	98.57(0.51)	55.86(1.86)	70.44(1.33)
	n = 400	96.99(0.6)	58.26(1.35)	72.75(1.03)	79.44(2.01)	93.35(0.9)	84.74(1.17)	96.99(0.6)	79.2(1.65)	86.21(1.09)
	n = 600	96.87(0.45)	70.21(1.46)	81.53(1.02)	77.86(1.56)	97.52(0.48)	85.89(1.01)	96.87(0.45)	89.23(1.1)	92.38(0.61)
p = 96										
	n = 100	99.9(0.08)	42.95(0.09)	60.09(0.09)	98.48(0.57)	46.43(0.85)	62.87(0.68)	99.9(0.08)	43.5(0.3)	60.59(0.28)
	n = 200	99.58(0.17)	43.47(0.22)	60.58(0.21)	93.6(0.91)	64.54(1.62)	75.21(1.02)	99.58(0.17)	49.24(0.99)	65.61(0.83)
	n = 400	98.67(0.25)	49.43(0.7)	65.85(0.6)	86.03(1.31)	87.97(1.14)	86.08(0.64)	98.67(0.25)	70.45(1.18)	81.42(0.78)
	n = 600	98.18(0.24)	59.63(0.96)	74.05(0.73)	82.83(1.08)	94.85(0.53)	87.92(0.57)	98.18(0.24)	82.83(0.92)	89.29(0.54)
VAR		precision	recall	F1	precision	recall	F1	precision	recall	F1
p = 12										
	n = 100	90.08(2.96)	45.23(2.11)	57.52(1.41)	79.5(4.16)	60.65(4.06)	61.26(1.93)	90.08(2.96)	48.83(2.96)	58.42(1.74)
	n = 200	88.99(2.48)	51.84(2.15)	62.68(1.25)	75.61(3.8)	73.26(3.52)	68.11(2.0)	88.99(2.48)	57.57(2.7)	64.86(1.83)
	n = 400	88.99(1.96)	61.3(2.08)	71.14(1.49)	72.33(2.58)	86.34(2.19)	74.52(1.59)	88.99(1.96)	70.4(2.42)	74.7(1.69)
	n = 600	86.88(0.11)	69.07(0.09)	76.14(0.41)	66.34(0.41)	93.65(0.53)	74.74(0.4)	86.88(0.11)	79.31(0.18)	79.61(0.05)
p = 24										
	n = 100	94.85(1.6)	36.88(0.81)	52.6(0.55)	85.36(3.08)	47.49(2.25)	56.34(1.05)	94.85(1.6)	39.38(1.1)	53.8(0.76)
	n = 200	94.43(1.39)	39.78(0.85)	55.71(0.74)	81.51(2.38)	59.89(2.32)	64.32(1.43)	94.43(1.39)	45.61(1.43)	59.44(1.08)
	n = 400	92.9(1.21)	48.0(1.0)	63.48(0.86)	75.01(2.24)	77.5(1.81)	72.69(1.26)	92.9(1.21)	59.46(1.57)	70.39(1.19)
	n = 600	92.74(0.46)	58.03(0.27)	72.01(0.18)	71.01(1.52)	88.3(0.12)	76.35(0.86)	92.74(0.46)	71.38(0.25)	79.11(0.05)
p = 48										
	n = 100	97.63(0.95)	34.28(0.3)	50.8(0.23)	91.73(2.08)	39.63(1.17)	53.21(0.73)	97.63(0.95)	35.47(0.57)	51.44(0.39)
	n = 200	97.72(0.51)	35.35(0.33)	52.02(0.31)	88.42(1.47)	48.61(1.34)	60.42(0.92)	97.72(0.51)	39.06(0.66)	55.16(0.55)
	n = 400	96.34(0.59)	40.86(0.48)	57.54(0.47)	80.55(1.42)	66.78(1.11)	70.84(0.81)	96.34(0.59)	50.95(0.69)	65.76(0.63)
	n = 600	95.71(0.12)	48.49(0.72)	64.74(0.77)	77.08(1.29)	79.01(0.31)	76.29(0.73)	95.71(0.12)	61.51(0.52)	74.1(0.44)
p = 96										
	n = 100	99.02(0.52)	33.59(0.11)	50.21(0.08)	96.16(1.12)	35.66(0.52)	51.24(0.36)	99.02(0.52)	34.02(0.25)	50.51(0.19)
	n = 200	99.15(0.31)	33.84(0.12)	50.54(0.12)	94.04(1.06)	41.37(0.77)	56.44(0.58)	99.15(0.31)	35.67(0.34)	52.29(0.31)
	n = 400	98.25(0.29)	36.62(0.26)	53.44(0.26)	87.0(1.18)	57.22(0.96)	67.82(0.59)	98.25(0.29)	44.75(0.62)	61.02(0.58)
	n = 600	97.74(0.12)	43.05(0.19)	59.87(0.18)	82.79(0.08)	69.24(0.59)	74.33(0.38)	97.74(0.12)	55.26(0.48)	70.27(0.33)

Table 1.2: Precision, Recall, F1 Score (in %) of three different thresholding methods: hard threshold, lasso and adaptive lasso.

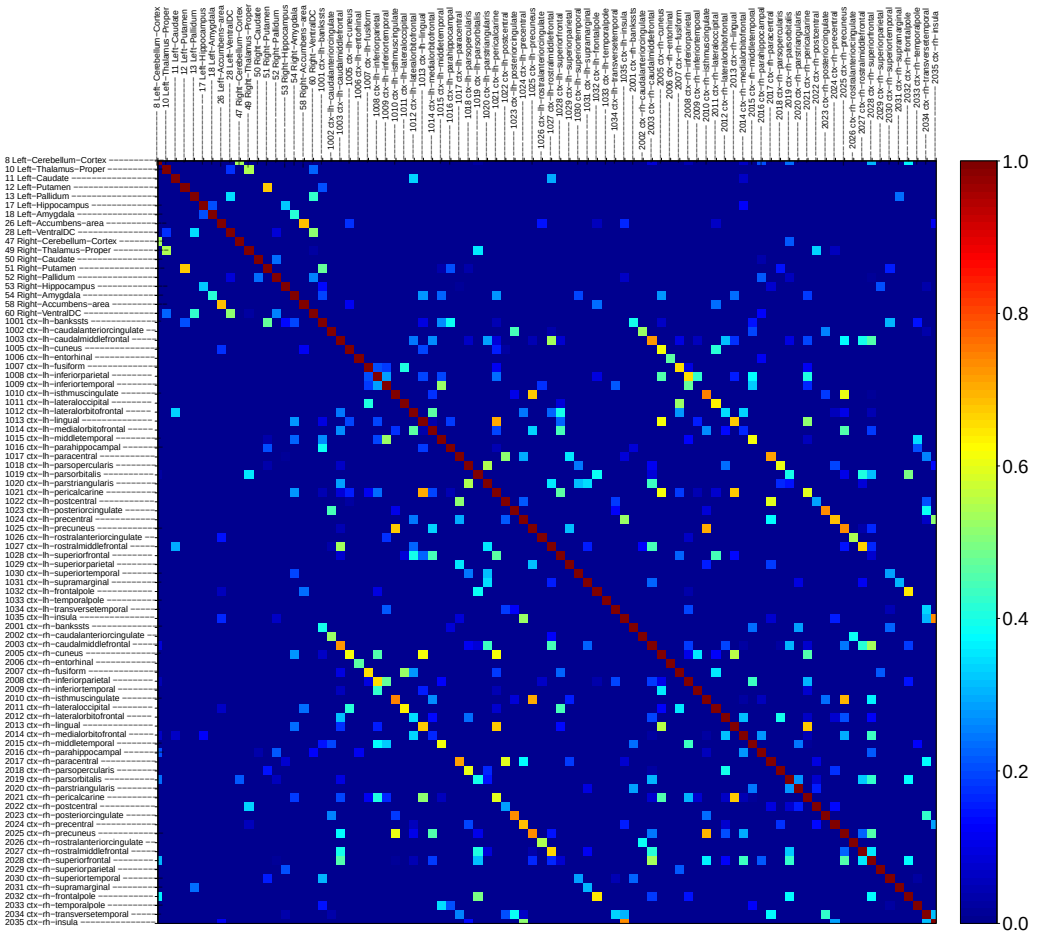


Figure 1.3: Heat map of absolute coherence matrix (at frequency 0) estimated using adaptive lasso thresholding of averaged periodogram.

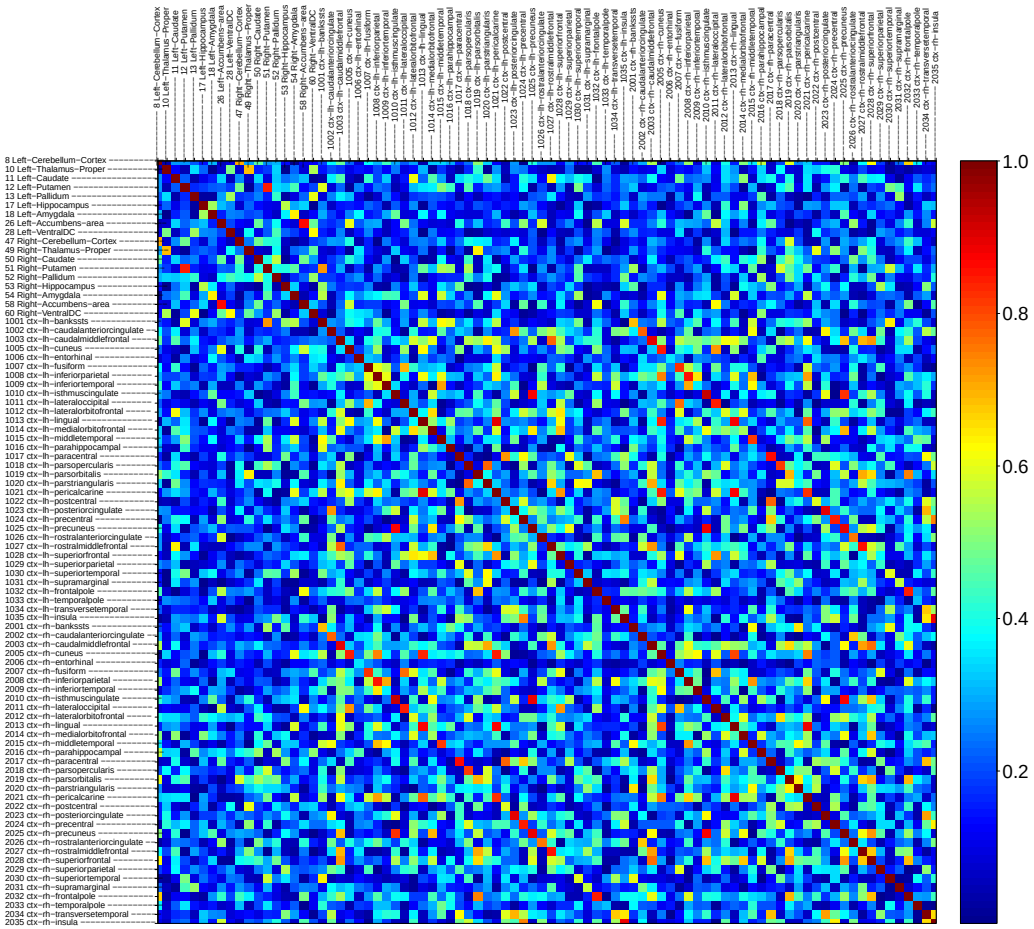


Figure 1.4: Heat map of absolute coherence matrix (at frequency 0) estimated using diagonal shrinkage of averaged periodogram.

CHAPTER 2

LARGE SPECTRAL DENSITY MATRIX ESTIMATION FOR GAUSSIAN PROCESS BY ADAPTIVE THRESHOLDING

2.1 Introduction

Consider a p-dimensional real-valued time series $X_t = (X_{t1}, \ldots, X_{tp})^\top,\ t \in \mathbb{Z}$. It is called weakly stationary if $\mathbb{E}X_t = \mathbb{E}X_s$ and autocovariance $\Gamma(\ell) = \text{Cov}(X_t, X_{t-\ell})$ only depends on the lag ℓ. If for any finite sequence $X_{t_1}, \cdots, X_{t_n})$ and any integer τ, $(X_{t_1}, \cdots, X_{t_n})$ has the same joint distribution with $(X_{t_1+\tau}, \cdots, X_{t_n+\tau})$, we say the process is strongly stationary. If the joint distribution is multivariate Gaussian, we call this process as Gaussian process where weak stationarity is equivalent to strong stationarity.

For completeness of this chapter, we restate the definition for spectral density for stationary time series in first chapter here. We assume $\mathbb{E}X_t = 0,\ t = 1, \ldots, n$ for ease of exposition. In practice, multivariate time series are often de-meaned before performing correlation based analysis. Strong/Weak stationarity for Gaussian process implies that $\Gamma(\ell) = \text{Cov}(X_t, X_{t-\ell}) = \mathbb{E}X_t X_{t-\ell}^\top$ only depends on ℓ. Spectral density aggregates information of autocovariance of different lag orders ℓ at a specific frequency $\omega \in [-\pi, \pi]$ as

$$f(\omega) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\omega}. \tag{2.1.1}$$

Note that the autocovariance functions of different lags can be recovered from the spectral density using the transformation $\Gamma(\ell) = \int_{-\pi}^{\pi} f(\omega) e^{i\ell\omega} d\omega$, for any $\ell \in \mathbb{Z}$.

Spectral density matrix in Gaussian process is a generalization to covariance matrix

for Guassian distribution. For multi-variate Gaussian variable x, position (r, s) for covariance matrix: $\Sigma_{r,s}$ is zero iff variable x_r, x_s are independent with each other. For p dimensional Gaussian process , this equivalent certificate for marginal independence becomes $f_{rs}(\omega) = 0$ for all frequency $\omega \in [-\pi, \pi)$ iff time series X_r is independent of X_s.

Recently, researchers have made progresses in developing methods with theoretical support in high-dimension for spectral density estimation for high dimensional time series data, assuming weak sparsity in spectral density, i.e. the L_1 norm of the spectral density matrix is small. For example, Fiecas et al. [2018]; Sun et al. [2018b] follow a similar path to build theory: first build concentration inequality for smoothing periodogram, then utilize weak sparsity property to demonstrate consistency in estimation. The difference mainly lays in the way of building concentration inequality part, or in other words, the way of measuring the dependency in the data. Fiecas et al. [2018] directly uses the magnitude of autocovariance as the measure while Sun et al. [2018b] follows the way of measuring dependency from frequency domain perspective as Basu and Michailidis [2015]. Those thresholding schemes are adaptive to heterogeneity in frequencies, not to variability of the individual entries. In other words, in terms of fixed frequency, they do universal thresholding for the spectral density matrix. In the next subsection, we explain why we need adaptive thresholding for estimating spectral matrix for each fixed frequency.

2.1.1 Why Adaptive Thresholding?

As pointed by Cai and Liu [2011], universal thresholding was firstly proposed by Donoho and Johnstone [1994] for estimating sparse normal mean vectors in the context

of wavelet function estimation where noise is homoscedastic. Although in many cases, the universal thresholding can achieve good theoretical property for heteroscedastic problem within certain weakly sparse space by setting the thresholding value proportional to an upper bound of the standard deviation of the noise. For example, consider following sparse multivariate Gaussian mean estimation. Suppose that we have n observation of p-variate Gaussian distribution as follows.

$$y_i \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \sigma_p^2 \end{bmatrix}\right),$$

Each position in the Gaussian vector has different unknown variance σ_j^2. But if we assume σ_j are uniformly bounded i.e., $\max_j \sigma_j \leqslant B$ for some positive number B, let $\bar{y}_j = \frac{1}{n}\sum_{i=1}^n y_{ij}$, and we have

$$\mathbb{P}(|\bar{y}_j - \mu_j| \geqslant \eta) \leqslant 2\exp(-n\eta^2/2\sigma_j^2). \tag{2.1.2}$$

Combine these with techniques introduced in Bickel and Levina [2008], if we choose threshold to be at order as $B\sqrt{\frac{\log p}{n}}$, we can have asymptotic consistency in estimation of μ if

$$\mu \in \left\{ \mu \in \mathbb{R}^p, \sum_{j=1}^p |\mu_j|^q \leqslant c(p) \right\},$$

for some $0 \leqslant q < 1$ and $c(p)$ is the measure for weak sparsity. The key assumption is that $\sigma_j s$ are uniformly bounded. But it is apparent, if σ_j variates too much or is not bounded, the argument in Bickel and Levina [2008] will not work and the accuracy in estimation will get hurt. So people resort to adaptive thresholding: set the thresholding value to be proportional to an estimator of σ_j , say sample standard deviation $\sqrt{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2/(n-1)}$.

Now we go to a more relevant case. Consider a number of i.i.d normal distributed y_i with

$$y_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_{p\times p}),$$

if we want to estimate each position r, s of Σ: Σ_{rs}, the estimation problem can be treated as estimating the expectation of $(y_i y_i^\top)_{rs}$. Then maximum likelihood estimator (MLE) is simply the sample average. Suppose the covariance matrix has a weakly sparse structure i.e., a relatively small L_1 norm, Bickel and Levina [2008] proposes to apply universal thresholding and they assume diagonal elements for Σ are uniformly bounded. Compared to above sparse mean estimation for multi-variate Gaussian, now it is like the case where we estimate the expectation of a p^2 random vector $(y_j y_j^\top)_{rs}$, and $\sigma_{rs}^2 = \mathrm{Var}[(y_i y_i^\top)_{rs}] = \Sigma_{rr}\Sigma_{ss} + \Sigma_{rs}^2 \leqslant 2\Sigma_{rr}\Sigma_{ss}$. Now it becomes clear why Bickel and Levina [2008] requires an upper bound for Σ_{rr}: $2\max_{r=1}^p \Sigma_{rr}^2$ is an upper bound for variance of the target. To conquer the shortcomings we mentioned above about universal thresholding, Cai and Liu [2011] proposed adaptive thresholding estimator by setting thresholding value proportional to sample standard deviation of sequence $[(y_1 y_1^\top)_{rs}, \cdots, (y_n y_n^\top)_{rs}]$ for n observations.

Now let us go to our case. As we will introduce later, the spectral density $f(\omega_j)$ can be taken as expectation of $d_\infty(\omega_j) d_\infty^\top(\omega_j)$ where $d_\infty(\omega_j)$ has the distribution same as limiting distribution of discrete Fourier coefficient. Then it is almost same as covariance setting although it is complex matrix not real anymore. Sun et al. [2018b] proposes hard thresholding in the modulus of the estimator. With Cauchy-Schwarz inequality,

$$\mathbb{E}|((d(\omega_j)_\infty d(\omega_j)_\infty^\top)_{rs}|^2 \leqslant \mathbb{E} d_{\infty,ss}^2(\omega_j) d_{\infty,rr}^2(\omega_j) \leqslant \operatorname*{ess\,sup}_\omega \|f(\omega)\|.$$

Although Sun et al. [2018b] allows $\operatorname{ess\,sup}_\omega \|f(\omega)\|$ grow with n, p, this upper bound appears in the thresholding value, and its growth rate is required to be controlled. As

discussed before, it is better to replace this upper bound with the estimated variance. We list our contributions to solve those problems mentioned.

- we clearly define the adaptive thresholding estimation problem, and show what variance should our estimator be adaptive to.
- we propose a modified version of periodogram which assists us to develop the theory for consistent estimation of adaptive thresholding estimator under high dimensional setting.
- non-asymptotic bound analysis is provided for our adaptive thresholding estimator which relaxes the constraint in operator norm of spectral density for universal thresholding proposed by Sun et al. [2018b] and achieves better error bound.

2.1.2 Periodogram Smoothing

Let $\mathcal{X} = [X_1 : \ldots : X_n]^\top$ be the *data matrix* containing n consecutive observations from the time series $\{X_t\}$ in its rows. The classical estimate of spectral density is based on the periodogram [Brockwell and Davis, 2013; Rosenblatt, 1985] defined as

$$I(\omega) = \sum_{|\ell|<n} \hat{\Gamma}(\ell) e^{-i\ell\omega}, \tag{2.1.3}$$

where $\hat{\Gamma}(\ell) = n^{-1} \sum_{t=\ell+1}^{n} X_t X_{t-\ell}^\top$ for $\ell \geqslant 0$, and $\hat{\Gamma}(\ell) = n^{-1} \sum_{t=1}^{n+\ell} X_t X_{t-\ell}^\top$ for $\ell < 0$. It is noticed that the periodogram can be written as outer product of Discrete Fourier Transformation(DFT):

$$I(\omega) = d(\omega) d^\dagger(\omega),$$

where $d(\omega)$ is defined as $d(\omega) = \mathcal{X}^\top(C(\omega) - iS(\omega))$, where

$$\begin{aligned} C(\omega) &= \frac{1}{\sqrt{n}}(1, \cos\omega, \ldots, \cos(n-1)\omega)^\top, \\ S(\omega) &= \frac{1}{\sqrt{n}}(1, \sin\omega, \ldots, \sin(n-1)\omega)^\top. \end{aligned} \tag{2.1.4}$$

This leads to fast computation with fast Fourier transformation. For brevity, we let $c_j = C(\omega_j)$ and $s_j = S(\omega_j)$.

It is common to resort to smoothing periodograms over nearby frequencies to achieve asymptotic consistency. The simplest smoothing is

$$\hat{f}(\omega; m) = \frac{1}{2\pi(2m+1)} \sum_{|k| \leqslant m} I(\omega + \omega_k), \tag{2.1.5}$$

where $\omega_k = 2\pi k/n,\ k \in F_n$, the set of Fourier frequencies. To be precise, F_n denotes the set $\{-[\frac{n-1}{2}], \ldots, [\frac{n}{2}]\}$ where $[x]$ is the integer part of x. F_n contains exactly the same frequencies used to calculate discrete Fourier transformation. It is common to evaluate the periodogram at these Fourier frequencies, in which case the smoothing periodogram in (2.1.5) becomes

$$\hat{f}(\omega_j; m) = \frac{1}{2\pi(2m+1)} \sum_{|k| \leqslant m} I(\omega_{j+k}). \tag{2.1.6}$$

Note that even though the values of $j + k$ can fall outside F_n, it is enough to evaluate periodograms at Fourier frequencies F_n since $I(\omega)$ is 2π-periodic in ω. The effectiveness of reducing variance via smoothing lies in the fact that $d(\omega_j), d(\omega_k)$ are asymptotically independent if $k \notin \{-j, j\}$. Since our theory development is based on the asymptotically independence, we change the smoothing set a little bit to make sure that $\{-j, j\}$ would not appear at the same time in the index set. Also, we exclude frequency $0, \frac{\pi}{2}, -\pi$ to avoid degenerate limiting distribution. Then we can present the smoothing

periodogram estimator as

$$g(\omega_j; m) = \frac{1}{m} \sum_{k \in \mathcal{B}_j^m} I(\omega_k). \tag{2.1.7}$$

where $\mathcal{B}_j^m$ is a set containing all indices nearest to j excluding $0, [n/2]$ and all possible pairs $\{j, -j\}$. Assuming $m < n/2$, in fact the expression for $\mathcal{B}_j^m$ is quite simple

$$\mathcal{B}_j^m = \begin{cases} \{j, j+1, \cdots, j+m\} & j > 0 \\ \{j, j-1, \cdots, j-m\} & j < 0. \end{cases} \tag{2.1.8}$$

as pointed before, we use 2π-periodic in ω if index falls out of F_n. For simplicity, we ignore the m in subscription in $\mathcal{B}_j^m$ and we always let it be $\mathcal{B}_j$. We introduce set $\mathcal{B}_j$ mainly for the purpose of ensuring all $H(\omega_j)$ are asymptotically independents with similar distribution. We exclude pairs like $\{j, -j\}$ and for simplicity, we do not consider frequency $\omega_0, \omega_{[n/2]}$ since they have degenerate limiting distribution. The theory can be easily extended to both cases though. In the next section, we will answer the question: what kind of variance should our thresholding estimator adapt to.

2.1.3 What Variance should thresholding Value be Adaptive to?

In order to derive adaptive thresholding for spectral density estimation, the first question is which variance should the estimator be adaptive to? Different from i.i.d. Gaussian case, to describe the variance of spectral density, we need to consider the asymptotic distribution of discrete Fourier transformation. We listed Theorem 4.4.1 in Brillinger [2001] as the following lemma

Assumption 2.1.1. $\sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\| < \infty$.

Lemma 2.1.2. *Suppose $\mathcal{X}_{n\times p} = [X_1 : \ldots : X_n]^\top$ is a data matrix from a strongly stationary Gaussian time series X_t, and assumption 2.1.1 is satisfied, we have for all $j \in F_n$ with $\omega_j \neq 0$ or π,*

$$d(\omega_j) = \begin{bmatrix} \mathbf{Re}(d(\omega_j)) \\ \mathbf{Im}(d(\omega_j)) \end{bmatrix} = \begin{bmatrix} \mathcal{X}^\top c_j \\ \mathcal{X}^\top s_j \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{2}\begin{bmatrix} \mathbf{Re}(f(\omega_j) & -\mathbf{Im}(f(\omega_j)) \\ \mathbf{Im}(f(\omega_j)) & \mathbf{Re}(f(\omega_j)) \end{bmatrix}\right). \tag{2.1.9}$$

For $\omega_j = 0$ or $\omega_j = \pi$,

$$\mathcal{X}^\top c_j \xrightarrow{d} \mathcal{N}\left(0, f(\omega_j)\right). \tag{2.1.10}$$

For $k \notin \{j, -j\}$, $\begin{bmatrix} \mathcal{X}^\top c_j \\ \mathcal{X}^\top s_j \end{bmatrix}$ is asymptotically independent of $\begin{bmatrix} \mathcal{X}^\top c_k \\ \mathcal{X}^\top s_k \end{bmatrix}$. Here the convergence is convergence in distribution.

This lemma indicates that if we restrict our focus to $\mathcal{B}_j$, all $I(\omega_k), k \in \mathcal{B}_j^m$ are asymptotically independently distributed and within a neighborhood, their asymptotic distributions are very similar to each other, in other words, behave like i.i.d. data. This explains why smoothing periodogram can effectively reduce the variance. What is more, we can treat the spectral density as the expectation of limiting distribution. Let $d_\infty(\omega_j)$ be the variable whose distribution is the limiting distribution of $d(\omega_j)$, then

$$f(\omega_j) = \mathbb{E}(d_\infty(\omega_j)d_\infty^\dagger(\omega_j)).$$

Now it is clear that the thresholding estimator should adapt to:

$$\mathrm{Var}(\mathbf{Re}(d_\infty(\omega_j)d_\infty^\dagger(\omega_j))_{rs}); \mathrm{Var}(\mathbf{Im}(d_\infty(\omega_j)d_\infty^\dagger(\omega_j))_{rs})$$

for position (r, s)'s real and imaginary part respectively. Asymptotically, the formation of the problem is almost the same as what is proposed by Cai and Liu [2011]: we use

sample average of data to estimate its expectation and try to let our estimator be adaptive to its variances. Rearranging those expressions for Gaussian forth moments in appendix, we can express the variances for real and imaginary part of $(d_\infty(\omega_j)d_\infty(\omega_j)^\dagger)_{rs}$ as

$$\begin{aligned}&\mathrm{Var}(\mathbf{Re}(d_\infty(\omega_j)d_\infty^\dagger(\omega_j))_{rs})\\ =\ &\tfrac{1}{2}\left[f_{rr}(\omega_j)f_{ss}(\omega_j)+\mathbf{Re}(f_{rs}(\omega_j))^2-\mathbf{Im}(f_{rs}(\omega_j))^2\right]\end{aligned} \tag{2.1.11}$$

and

$$\begin{aligned}&\mathrm{Var}(\mathbf{Im}(d_\infty(\omega_j)d_\infty^\dagger(\omega_j))_{rs})\\ =\ &\tfrac{1}{2}\left[f_{rr}(\omega_j)f_{ss}(\omega_j)+\mathbf{Im}(f(\omega_j)_{rs})^2-\mathbf{Re}(f_{rs}(\omega_j))^2\right].\end{aligned} \tag{2.1.12}$$

As mentioned above, the key ingredient in the theoretical development of Cai and Liu [2011] is that variances of position r, s of $y_1y_1^\top$ for p variate Gaussian variable y_1, is in the same order of $\Sigma_{rr}\Sigma_{ss}$. However, variances in (2.1.11) and (2.1.12) do not meet this condition. We provide a counter example in Appendix 2.E to demonstrate this phenomenon. We not only utilize the framework provided by Cai and Liu [2011], but also try to avoid diminishing variances in the estimator, which may cause instability in the theory.

Notation. Throughout this paper, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of integers, real numbers and complex numbers, respectively. We use $\mathbf{Re}(c)$, $\mathbf{Im}(c)$ to present the real and imaginary part of complex number c respectively and $|c|$ to denote its modulus(absolute value for real number). We use $\|v\|$ to denote ℓ_2-norm of a vector v. For a matrix A, $\|A\|_1$, $\|A\|_\infty$, $\|A\|$ and $\|A\|_F$ will denote maximum complex modulus column sum norm, maximum complex modulus row sum norm, spectral norm $\sqrt{\Lambda_{\max}(A^\dagger A)}$ and Frobenius norm $\sqrt{\mathrm{tr}(A^\dagger A)}$, respectively, where $A^\dagger$ is the conjugate transpose of A. We use e_i to denote the i^{th} unit vector in $\mathbb{R}^p$, for $i = 1, 2, \ldots, p$. Also, we let E_p be the set containing $e_1, \cdots, e_p$. For vectors $v_i \in \mathbb{R}^p, i = 1, \ldots, n$, we use $[v_1 : \ldots : v_n]$ to denote the $p \times n$ matrix formed by horizontally stacking these column vectors v_i, and $[v_1^\top; \ldots; v_n^\top]$

to denote the $n \times p$ matrix by vertically stacking row vectors $v_i^\top$. Let $vec(A)$ represent the vector obtained from vectorization of a matrix A by stacking all its columns. We use $rk(A)$ to denote the rank of a matrix A. For a complex vector $v \in \mathbb{C}^p$ and any $q > 0$, we define $\|v\|_q := (\sum_{i=1}^p |v_i|^q)^{1/q}$. We use $\|v\|_0$ to denote the number of non-zero elements in v. Note that when $0 \leqslant q < 1$, it is not really a norm, for triangle inequality does not hold, but we keep the notation of a norm for convenience. Then we define the induced matrix norm, $\|A\|_{\alpha,\beta} = \sup_{x \neq 0} \|Ax\|_\alpha / \|x\|_\beta$, for any $\alpha > 0, \beta > 0$. We will also use $\|A\|_\alpha$ to denote the induced norm $\|A\|_{\alpha,\alpha}$ for any $\alpha > 0$ and any complex matrix $A \in \mathbb{C}^{p \times p}$. Also, to be succinct, we use $\|A\|_{\max} := \max_{r,s} |A_{rs}|$. Throughout the paper, we write $A \gtrsim B$ if there exists a universal constant $c > 0$, not depending on model dimension or any model parameters, such that $A \geqslant cB$. We use $A \asymp B$ to denote $A \gtrsim B$ and $B \gtrsim A$.

2.2 Background and Methods

2.2.1 Modified Periodogram and Its Smoothing Estimator

In this section, we show that with only a small but effective modification in periodogram, we can still preserve asymptotic unbiasedness while letting the order of the variance for each entry (r, s) still at $f_{rr}(\omega_j) f_{ss}(\omega_j)$.

$$\begin{aligned} f(\omega_j) &= \mathbb{E}[d_\infty(\omega_j) d_\infty^\top(\omega_j)] \\ &= \mathbb{E}[\mathbf{Re}(d_\infty(\omega_j))\mathbf{Re}(d_\infty^\top(\omega_j)) + \mathbf{Im}(d_\infty(\omega_j))\mathbf{Im}(d_\infty^\top(\omega_j))] \\ &+ i\mathbb{E}[\mathbf{Im}(d_\infty(\omega_j))\mathbf{Re}(d_\infty^\top(\omega_j)) - \mathbf{Re}(d_\infty(\omega_j))\mathbf{Im}(d_\infty^\top(\omega_j))] \end{aligned} \tag{2.2.1}$$

Lemma 2.1.2 claims that $\mathbf{Re}(d_\infty(\omega_j))$ and $\mathbf{Im}(d_\infty(\omega_j))$ share the same marginal distribution, and $\mathbb{E}[\mathbf{Im}(d_\infty(\omega_j))\mathbf{Re}(d_\infty^\top(\omega_j)) = -\mathbb{E}\mathbf{Re}(d_\infty(\omega_j))\mathbf{Im}(d_\infty^\top(\omega_j))$. Then we have for real and imaginary parts of spectral density,

$$\begin{aligned}\mathbf{Re}(f(\omega_j)) &= 2\mathbb{E}\mathbf{Re}(d_\infty(\omega_j))\mathbf{Re}(d_\infty(\omega_j))^\top \\ \mathbf{Im}(f(\omega_j)) &= 2\mathbb{E}\mathbf{Im}(d_\infty(\omega_j))\mathbf{Re}(d_\infty(\omega_j))^\top\end{aligned} \tag{2.2.2}$$

where $d_\infty(\omega_j)$'s distribution is the same as the limiting distribution of $d(\omega_j)$. Therefore, we could use

$$H(\omega_j) = 2\mathbf{Re}(d(\omega_j))\mathbf{Re}(d(\omega_j))^\top + 2\mathbf{Im}(d(\omega_j))\mathbf{Re}(d(\omega_j))^\top i \tag{2.2.3}$$

as our modified periodogram, with its limiting version of

$$H_\infty(\omega_j) = 2\mathbf{Re}(d_\infty(\omega_j))\mathbf{Re}(d_\infty(\omega_j))^\top + 2\mathbf{Im}(d_\infty(\omega_j))\mathbf{Re}(d_\infty(\omega_j))^\top i \tag{2.2.4}$$

whose expectation is exactly $f(\omega_j)$. Then the variance for real and imaginary parts of $H_{\infty,rs}(\omega_j)$ can be calculated with fourth moments of Gaussian distribution:

$$\mathrm{Var}(\mathbf{Re}(H_{\infty,rs}(\omega_j))) = [f_{rr}(\omega_j)f_{ss}(\omega_j) + \mathbf{Re}(f_{rs}(\omega_j))^2] \tag{2.2.5}$$

and

$$\mathrm{Var}(\mathbf{Im}(H_{\infty,rs}(\omega_j))) = [f_{rr}(\omega_j)f_{ss}(\omega_j) + \mathbf{Im}(f_{rs}(\omega_j))^2]\,. \tag{2.2.6}$$

Both of them are at the same order of $f_{rr}(\omega_j)f_{ss}(\omega_j)$. More specifically,

$$\begin{aligned}f_{rr}(\omega_j)f_{ss}(\omega_j) \leqslant \mathrm{Var}(\mathbf{Re}(H_{\infty,rs})) \leqslant 2f_{rr}(\omega_j)f_{ss}(\omega_j) \\ f_{rr}(\omega_j)f_{ss}(\omega_j) \leqslant \mathrm{Var}(\mathbf{Im}(H_{\infty,rs})) \leqslant 2f_{rr}(\omega_j)f_{ss}(\omega_j)\end{aligned} \tag{2.2.7}$$

In the following, we will show how to use the modified periodogram to perform adaptive thresholding. The modified smoothing periodogram can be written as

$$g(\omega_j) = \frac{1}{m}\sum_{k\in\mathcal{B}_j} H(\omega_k). \tag{2.2.8}$$

2.2.2 Method: Adaptive Thresholding

Firstly we introduce definition of generalized thresholding proposed by Rothman et al. [2009]: consider a thresholding operator $S_\lambda(\cdot)$ that integrates the benefits of shrinkage and thresholding:

(1) $|S_\lambda(z)| \leqslant |z|$,

(2) $S_\lambda(z) = 0$ if $|z| \leqslant \lambda$,

(3) $|S_\lambda(z) - z| \leqslant \lambda$.

Rothman et al. [2009] show that this estimator can recover the covariance matrix assuming weak sparsity. Cai and Liu [2011] further let the thresholding value be proportional to the estimator of standard deviation.

Estimation of Variance: As pointed out above, all $H(\omega_k), k \in \mathcal{B}_j$ behave like i.i.d. variables, thus we propose to use sample variance to estimate variances at position (r, s) of $\mathbf{Re}(H(\omega_j))$ and $\mathbf{Im}(H(\omega_j))$. We let $\theta^{(r)}_{j,rs}$ represent $\mathrm{Var}(\mathbf{Re}(H_{\infty,rs}(\omega_j)))$ and $\theta^{(i)}_{j,rs}$ represent $\mathrm{Var}(\mathbf{Im}(H_{\infty,rs}(\omega_j)))$ respectively. Then we can write the estimator $\hat{\theta}^{(r)}_{j,rs}$, $\hat{\theta}^{(i)}_{j,rs}$ as

$$\begin{aligned}
\hat{\theta}^{(r)}_{j,rs} &= \frac{1}{m-1} \sum_{q \in \mathcal{B}_j} \left[\mathbf{Re}(H_{rs}(\omega_q)) - \frac{1}{m} \sum_{k \in \mathcal{B}_j} \mathbf{Re}(H_{rs}(\omega_k)) \right]^2 \\
\hat{\theta}^{(i)}_{j,rs} &= \frac{1}{m-1} \sum_{q \in \mathcal{B}_j} \left[\mathbf{Im}(H_{rs}(\omega_q)) - \frac{1}{m} \sum_{k \in \mathcal{B}_j} \mathbf{Im}(H_{rs}(\omega_k)) \right]^2 .
\end{aligned} \tag{2.2.9}$$

Adaptive Estimator With the variance estimator, the adaptive thresholding value for the real and imaginary parts at frequency ω_j at position (r, s), which we call $\lambda^{(r)}_{rs}$ and

Input: j, m, N, modified periodograms at Fourier frequency $H(\omega_j)$, an estimator of variance of real part $\hat{\theta}_j^{(r)}$, finite grid of thresholds $\mathcal{L}$

for $\lambda \in \mathcal{L}$ **do**

  for $\nu \leftarrow 1$ ***to*** N **do**

    Randomly divide $\mathcal{B}_j$ into two subsets J_1 and J_2 such that $||J_1| - |J_2|| \leqslant 1$

    $g_{1,\nu}(\omega_j) \leftarrow \frac{1}{|J_1|}\sum_{k\in J_1} \mathbf{Re}(H(\omega_k)), \quad g_{2,\nu}(\omega_j) \leftarrow \frac{1}{|J_2|}\sum_{k\in J_2} H(\omega_k)$

    $\hat{R}_\nu(\omega_j, \lambda) \leftarrow \left\| T_{\hat{\theta}_j^{(r)}\lambda}(g_{1,\nu}(\omega_j)) - g_{2,\nu}(\omega_j) \right\|_F^2$

  end

  $\hat{R}(\omega_j, \lambda) \leftarrow \sum_{\nu=1}^{N} \hat{R}_\nu(\omega_j, \lambda)/N$

end

Output: $\hat{\lambda}_j := \hat{\lambda}(\omega_j) = \mathrm{argmin}_{\lambda\in\mathcal{L}} \hat{R}(\omega_j, \lambda)$

Algorithm 2: Threshold Selection by Frequency Domain Sample-splitting for Real Part

$\lambda_{rs}^{(i)}$ respectively, will look like

$$\begin{aligned} \lambda_{rs}^{(r)} &= \sqrt{\hat{\theta}_{j,rs}^{(r)}}\lambda^{(r)} \\ \lambda_{rs}^{(i)} &= \sqrt{\hat{\theta}_{j,rs}^{(i)}}\lambda^{(i)}. \end{aligned} \tag{2.2.10}$$

Here $\lambda^{(r)}, \lambda^{(i)}$ are the same across all positions for spectral density at frequency ω_j for real and imaginary part respectively. We delete j from notation for brevity.

Let $\hat{\theta}_j^{(r)}$ be the matrix of size $p \times p$, with each element equal to $\hat{\theta}_{j,rs}^{(r)}$ and we define $\hat{\theta}_j^{(i)}$ in a similar way. For the real part, we can define following adaptive thresholding operator:

$$T_{\hat{\theta}_j^{(r)}\lambda}(M) = \tilde{M} \tag{2.2.11}$$

where

$$\tilde{M}(r,s) = \begin{cases} M(r,s) & \text{if } |M(r,s)| \leqslant \sqrt{\hat{\theta}_{j,rs}^{(r)}}\lambda \\ 0 & \text{else} \end{cases} \tag{2.2.12}$$

We present the way to choose $\lambda_j^{(r)}$ for the real part in Algorithm 2, which also applies to choosing the threshold value for the imaginary part.

2.3 Theoretical Properties

In this section, we analyze asymptotic properties of adaptive thresholding average modified periodograms. First we need to present the bound for bias in the average modified periodogram and variance estimator. Then we build concentration inequality for estimators of both of them.

2.3.1 Bounding the Bias

Although the modified periodograms $H(\omega_k), k \in \mathcal{B}_j$ are asymptotically independent, and behave like i.i.d., we need to quantify the bias from two sources. One is from the gap between the finite sample to the limiting distribution; the other is from the fact that the limiting distributions are not exactly identical. In this section, we list the results of bounding the bias for both smoothing modified periodograms and variance estimator.

Our process requires some conditions in the minimal value of diagonal of spectral density.

Assumption 2.3.1. $\min_{r=1}^{p} f_{rr}(\omega_j) \geqslant \phi_0 > 0$.

For simplicity we give a deterministic lower bound for diagonal elements in spectral density.

Bias for Smoothing Modified Periodogram Similar to Sun et al. [2018b], two quantities are needed that capture the strength of temporal and contemporaneous dependence

in the multivariate time series $\{X_t\}_{t\in\mathbb{Z}}$, which are

$$\Omega_n = \max_{1\leqslant r,s\leqslant p} \sum_{\ell=-n}^{n} |\ell||\Gamma_{rs}(\ell)|, \quad L_n = \max_{1\leqslant r,s\leqslant p} \sum_{|\ell|>n} |\Gamma_{rs}(\ell)|. \tag{2.3.1}$$

Lemma 2.3.2. *For any $j \in F_n$, $1 \leqslant r, s \leqslant p$,*

$$\max\left\{\left|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbf{Re}(f_{rs}(\omega_j))\right|, \left|\mathbb{E}[\mathbf{Im}(g_{rs}(\omega_j))] - \mathbf{Im}(f_{rs}(\omega_j))\right|\right\} \leqslant B_f, \tag{2.3.2}$$

where $B_f = \frac{m}{n}\Omega_n + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n\right) + \frac{\Omega_n}{2\pi n}$.

The technique for proof starts with the triangular inequality,

$$\begin{aligned} &\left|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbf{Re}(f_{rs}(\omega_j))\right| \\ &\leqslant \left|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbb{E}[\mathbf{Re}(\hat{f}_{rs}(\omega_j))]\right| + \left|\mathbb{E}[\mathbf{Re}(\hat{f}_{rs}(\omega_j))] - \mathbf{Re}(f_{rs}(\omega_j))\right|, \end{aligned} \tag{2.3.3}$$

where the bound for the second part is already shown in Sun et al. [2018b], and the first part of above inequality we can handle properties of toeplitz matrices. We defer the detailed proof to Appendix.

Bias for Variance Estimation

Lemma 2.3.3.

$$\max\left\{\left|\mathbb{E}\hat{\theta}^{(r)}_{j,rs} - \theta^{(r)}_{j,rs}\right|, \left|\mathbb{E}\hat{\theta}^{(i)}_{j,rs} - \theta^{(i)}_{j,rs}\right|\right\} \leqslant B_\theta, \tag{2.3.4}$$

where $B_\theta = 2\max(f_{rr}(\omega_j), f_{ss}(\omega_j))(\delta_1 + \delta_2) + \delta_1^2 + \delta_2^2 + \frac{\Omega_n^2}{\pi^2 n^2}$, and

$$\begin{aligned} \delta_1 &= \frac{\Omega_n}{2n\pi} + \frac{\sqrt{2}}{2\pi}\frac{m\Omega_n}{n} + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n\right) \\ \delta_2 &= \frac{\Omega_n}{2n\pi} + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n\right). \end{aligned} \tag{2.3.5}$$

Remark. *Assuming Ω_n, f_{rr} is bounded, then the bias in variance estimation has the same order of bias for estimator $g_{rs}(\omega_j)$: $\mathcal{O}(m/n)$.*

The proof is deferred to Appendix. Based on the bound for variance estimation, we here present another useful result. Since our intuition is that asymptotically modified periodograms are behaving like i.i.d. within $\mathcal{B}_j$, we expect that

$$\begin{aligned} \mathrm{Var}(\mathbf{Re}(g_{rs}(\omega_j))) &\approx \frac{1}{m}\theta_{j,rs}^{(r)} \\ \mathrm{Var}(\mathbf{Im}(g_{rs}(\omega_j))) &\approx \frac{1}{m}\theta_{j,rs}^{(i)}. \end{aligned} \tag{2.3.6}$$

The following lemma justifies and quantifies the above intuition.

Lemma 2.3.4. *For* $1 \leqslant r, s \leqslant p$*,*

$$\begin{aligned} &\frac{1}{m}(1 - B_\delta) \leqslant \min\left\{\left|\frac{\mathrm{Var}(\mathbf{Re}(g_{rs}(\omega_j)))}{\theta_{j,rs}^{(r)}}\right|, \left|\frac{\mathrm{Var}(\mathbf{Im}(g_{rs}(\omega_j)))}{\theta_{j,rs}^{(i)}}\right|\right\} \\ &\leqslant \max\left\{\left|\frac{\mathrm{Var}(\mathbf{Re}(g_{rs}(\omega_j)))}{\theta_{j,rs}^{(r)}}\right|, \left|\frac{\mathrm{Var}(\mathbf{Im}(g_{rs}(\omega_j)))}{\theta_{j,rs}^{(i)}}\right|\right\} \leqslant \frac{1}{m}(1 + B_\delta), \end{aligned} \tag{2.3.7}$$

where $B_\delta = 4\delta_1/\phi_0 + 3\delta_1^2/\phi_0^2$*,* δ_1, δ_2 *follow the definition in Lemma 2.3.3.*

2.3.2 Deviation Bound

In the previous section, we have shown the upper bound for bias for $g_{rs}(\omega_j)$, $\hat{\theta}_{j,rs}^{(r)}$ and $\hat{\theta}_{j,rs}^{(i)}$. In this section, we will build the non-asymptotic analysis for those estimators. The non-asymptotic analysis will lead to the main theory for consistency of our adaptive thresholding estimators.

Lemma 2.3.5. *For any positive number* $\eta > 0$*, there exit constants* c_1, c_2 *s.t.*

$$\begin{aligned} \mathbb{P}\left(\left|\hat{\theta}_{j,rs}^{(r)} - \theta_{j,rs}^{(r)}\right| \geqslant B_\theta + \eta\right) &\leqslant c_1 \exp(-c_2 m \min\left(\eta, \eta^2\right)), \\ \mathbb{P}\left(\left|\hat{\theta}_{j,rs}^{(i)} - \theta_{j,rs}^{(i)}\right| \geqslant B_\theta + \eta\right) &\leqslant c_1 \exp(-c_2 m \min\left(\eta, \eta^2\right)). \end{aligned} \tag{2.3.8}$$

The next lemma constitutes the key element for theoretical development. It is a non-asymptotic analysis.

Lemma 2.3.6. *For $j \in F_n$, and $\omega_j \notin \{0, -\pi\}$, $1 \leqslant r, s \leqslant p$, assuming $B_\delta \leqslant 3$ and $B_\theta/\phi_0^2 < 1/4$, given $\eta \leqslant 1$, there exist universal positive constants c_1, c_2, s.t.*

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \geqslant \frac{B_f}{\phi_0} + \eta\right) \leqslant c_1 \exp(-c_2 \min(\eta^2 m, \eta\sqrt{m})), \\
&\mathbb{P}\left(\left|\frac{\mathbf{Im}(g_{rs}(\omega_j)) - \mathbf{Im}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(i)}}}\right| \geqslant \frac{B_f}{\phi_0} + \eta\right) \leqslant c_1 \exp(-c_2 \min(\eta^2 m, \eta\sqrt{m})).
\end{aligned} \tag{2.3.9}$$

Remark. *The conditions like $B_\delta \leqslant 3$ and $B_\theta/\phi_0^2 < 1/4$ are set for convenience of proof. In the later main results section, asymptotically, they both go to zero.*

Proof. We will focus on the proof for the real part. First we handle the bias. Define event $\mathcal{A}_j$ and $\mathcal{B}_j$ as

$$\begin{aligned}
\mathcal{A}_j &= \{\hat{\theta}_{j,rs}^{(r)} \leqslant (4 - B_\theta/\phi_0^2)\theta_{j,rs}^{(r)}\}, \\
\mathcal{B}_j &= \{\hat{\theta}_{j,rs}^{(r)} \geqslant (1/4 + B_\theta/\phi_0^2)\theta_{j,rs}^{(r)}\}.
\end{aligned} \tag{2.3.10}$$

Given lemma 2.3.5 and the fact that $\theta_{j,rs}^{(r)}$ is at the same order of $f_{rr}(\omega_j)f_{ss}(\omega_j)$:

$$f_{rr}(\omega_j)f_{ss}(\omega_j) \leqslant \theta_{j,rs}^{(r)} \leqslant 2f_{rr}(\omega_j)f_{ss}(\omega_j). \tag{2.3.11}$$

We can show that there exist universal constants c_1, c_2 only depending on ϕ_0

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_j^c) &\leqslant \mathbb{P}\left(|\hat{\theta}_{j,rs}^{(r)} - \theta_{j,rs}^{(r)}| \geqslant (3 + B_\theta/\phi_0^2)f_{rr}(\omega_j)f_{ss}(\omega_j)\right) \\
&\leqslant c_1 \exp\{-c_2 m\}.
\end{aligned} \tag{2.3.12}$$

Similarly we can obtain the same bound for $\mathbb{P}(\mathcal{B}_j^c)$. From now on we restrict our analysis to the event $\mathcal{A}_j$ and $\mathcal{B}_j$.

By triangular inequality,

$$\begin{aligned}
&\left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \\
&\leqslant \left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbb{E}\mathbf{Re}(g_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \\
&+ \left|\frac{\mathbb{E}\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right|.
\end{aligned} \tag{2.3.13}$$

Noticing on event $\mathcal{B}_j$,

$$\left|\frac{\mathbb{E}\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \leqslant 2B_f/\phi_0, \tag{2.3.14}$$

which indicates that

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \geqslant \frac{B_f}{\phi_0} + \eta\right) \\
&\leqslant \mathbb{P}\left(\left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbb{E}\mathbf{Re}(g_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \geqslant \eta\right).
\end{aligned} \tag{2.3.15}$$

On event $\mathcal{A}_j$,

$$\sqrt{\frac{\hat{\theta}_{j,rs}^{(r)}}{\theta_{j,rs}^{(r)}}} \leqslant 2.$$

We write the left part of the result for the real part as

$$\begin{aligned}
&\left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}}\right| \\
&= \left|\frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\mathrm{Var}(\mathbf{Re}(g_{rs}(\omega_j)))}}\right| \\
&\times \left|\sqrt{\frac{\mathrm{Var}(g_{rs}(\omega_j))}{\theta_{rs}^{(r)}}}\right| \times \left|\sqrt{\frac{\theta_{rs}^{(r)}}{\hat{\theta}_{rs}^{(r)}}}\right|
\end{aligned} \tag{2.3.16}$$

Since we assume $B_\delta \leqslant 3$,

$$\left| \frac{\sqrt{\mathrm{Var}(g_{rs}(\omega_j))}}{\sqrt{\theta_{rs}^{(r)}}} \right| \leqslant \sqrt{(1+B_\delta)m} \leqslant 2\sqrt{m}.$$

Therefore, on $\mathcal{A}_j$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left| \frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\hat{\theta}_{j,rs}^{(r)}}} \right| \geqslant \frac{B_f}{\phi_0} + \eta \right) \\
&\leqslant \mathbb{P}\left(\left| \frac{\mathbf{Re}(g_{rs}(\omega_j)) - \mathbb{E}\mathbf{Re}(f_{rs}(\omega_j))}{\sqrt{\mathrm{Var}(\mathbf{Re}(g_{rs}(\omega_j)))}} \right| \geqslant 4\eta\sqrt{m} \right) \\
&\leqslant c_1 \exp(-c_2 \min(\eta^2 m, \eta\sqrt{m})),
\end{aligned} \tag{2.3.17}$$

where the last inequality comes from lemma 2.C.1 and the fact that we can write $\mathbf{Re}(f_{rs}(\omega_j))$ as quadratic function of Gaussian random variables as shown in Sun et al. [2018b]. □

2.3.3 Main Results

2.3.4 Sparse Class

In order to analyze the effectiveness of this estimator like consistency in L_2 norm, we require the following sparse class which is inspired by Bickel and Levina [2008]: for frequency ω_j,

$$\mathcal{U}(q, c_0(p), \omega) = \left\{ f(\omega) : \sum_{s=1}^{p} |f_{rs}(\omega)|^q \leqslant c_0(p) \text{ for all } r \right\}. \tag{2.3.18}$$

Sun et al. [2018b] shows that a generalized thresholding estimator can achieve L_2 and Frobenius norm estimation consistency. For adaptive thresholding, we follow the anal-

ogy of Cai and Liu [2011], defining the following sparse class:

$$\mathcal{U}^a(q, c_0(p), \omega) = \left\{ f(\omega) : \max_{r=1}^{p} \sum_{s=1}^{p} (f_{rr}(\omega) f_{ss}(\omega))^{(1-q)/2} |f_{rs}(\omega)|^q \leqslant c_0(p) \right\}. \tag{2.3.19}$$

There are detailed discussions by Cai and Liu [2011] on how $\mathcal{U}^a$ is compared to $\mathcal{U}$. Assuming $\max_{r=1}^{n} f_{rr}(\omega_j) \leqslant M$, is bounded, $\mathcal{U}(q, c_0(p), \omega) \subset \mathcal{U}^a(q, c_0(p), \omega)$. Although Sun et al. [2018b] lets the $\max_{r=1}^{p} f_{rr}(\omega_j)$ grow with dimension but the rate for growth must be controlled in order to make sure that the thresholding value go to zero asymptotically. While in this paper, the adaptive thresholding estimator does not put any constraint on the growth rate for this statistic, but instead, we need a lower bound for $\min_{r=1}^{p} f_{rr}(\omega_j)$ or control the decay rate to zero. But this lower bound only occurs in the bias term.

2.3.5 Consistency Under Weak Sparsity

Proposition 2.3.7. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series satisfying Assumption 2.1.1, and consider a single Fourier frequency $\omega_j \in [-\pi, \pi)$ and $\omega_j \neq 0$. For any m satisfying $m \lesssim n/\Omega_n(f)$ and $m \gtrsim c_0^2(p) \log p$, and a large enough $R > 0$, there exist universal constants $c_1, c_2 > 0$ such that choosing a threshold*

$$\lambda = R c_0(p) \sqrt{\frac{\log p}{m}} + 2B_f/\phi_0, \tag{2.3.20}$$

where $B_f = \frac{m}{n}\Omega_n(f_X) + \frac{1}{2\pi}\left(\frac{\Omega_n(f_X)}{n} + L_n(f_X)\right) + \frac{\Omega_n}{2\pi n}$, assuming $f(\omega_j) \in \mathcal{U}^a(q, c_0(p), \omega)$, the estimation error of adaptive thresholding average modified periodogram satisfies

$$\mathbb{P}\left(\left\| T_\lambda(\hat{f}(\omega_j)) - f(\omega_j) \right\| \geqslant 7\lambda^{(1-q)/2} \right) \leqslant c_1 \exp\left[-(c_2 R^2 - 2) \log p \right].$$

Proof. The logic for proof is almost the same as proof for consistency for operator norm appears in Sun et al. [2018b]. Using the concentration inequality introduced by Lemma 2.3.6 with the weakly sparse class coefficient, we can almost use exactly the same argument to finish the proof. So we omit the the details here. For those type of proof, the difficulty mainly lies in building the concentration inequality as Lemma 2.3.6. □

Remark. *Although B_f seems to have a complicated form, in many linear processes, as shown in Sun et al. [2018b], Ω_n is uniformly bounded. Also, if we set $m = \mathcal{O}(\sqrt{n})$, then $\log p/m \to 0$ if $p = \mathcal{O}(n^\delta)$ for any positive delta which is the same as modern high dimensional setting. Then it is not hard to find a sequence of thresholding value λ go to zero, as $n \to \infty$ and $p \to \infty$,*

Conclusion

We propose to do adaptive thresholding for weakly sparse spectral density estimation. In order to conquer some technical difficulties, we have proposed a new modified periodogram, which has the same order in bias compared with the classic periodogram, while providing great convenience in theoretical development. Our adaptive thresholding relaxes the constraint in maximum operator norm and achieves better convergence rate in theory.

2.A Proof for Bias Bounding

2.A.1 Proof for Lemma 2.3.2

Proof. We will show proof for real part only and with same argument, we can finish the proof for imaginary part. By triangular inequality,

$$\begin{aligned}&|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbf{Re}(f_{rs}(\omega_j))| \\ &\leqslant \left|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbb{E}[\mathbf{Re}(\hat{f}_{rs}(\omega_j))]\right| + \left|\mathbb{E}[\mathbf{Re}(\hat{f}_{rs}(\omega_j))] - \mathbf{Re}(f_{rs}(\omega_j))\right|,\end{aligned} \tag{2.A.1}$$

where $\hat{f}(\omega_j)$ is the usual smoothed estimator.

It is not hard to see that

$$\left|\mathbb{E}[\mathbf{Re}(g_{rs}(\omega_j))] - \mathbb{E}[\mathbf{Re}(\hat{f}_{rs}(\omega_j))]\right| \leqslant \max_{k\in\mathcal{B}_j} \left|e_r^\top (H(\omega_k) - I(\omega_k))e_s\right|. \tag{2.A.2}$$

Now for any unit vector u, v,

$$\begin{aligned}&\mathbb{E}\left[u^\top \mathbf{Re}(H(\omega_j))v - u^\top \mathbf{Re}(I(\omega_j))v\right] = \frac{1}{2\pi}\mathbb{E}\left[u^\top(\mathcal{X}^\top c_j c_j^\top \mathcal{X} - \mathcal{X}^\top s_j s_j^\top \mathcal{X})v\right] \\ &= \frac{1}{2\pi}\mathbb{E}\left[c_j^\top \mathcal{X} v u^\top \mathcal{X}^\top c_j - s_j^\top \mathcal{X} v u^\top \mathcal{X}^\top s_j\right].\end{aligned} \tag{2.A.3}$$

Let $\Sigma^{v,u} = \mathrm{Cov}(\mathcal{X}v, \mathcal{X}u)$, then $\Sigma^{v,u}$ is a toeplitz matrix with element as

$$\Sigma^{v,u}_{k,q} = u^\top \Gamma(q-k)v. \tag{2.A.4}$$

Then by first part of lemma 2.D.2, we know $\left|c_j^\top \Sigma^{v,u} c_j - s_j^\top \Sigma^{v,u} s_j\right| \leqslant \frac{\Omega(\Sigma^{v,u})}{2\pi n}$. Besides, $\Omega(\Sigma^{v,u}) \leqslant \Omega_n(f_{\mathcal{X}})$ since $\left|u^\top \Gamma(q-k)v\right| \leqslant \|\Gamma(q-k)\|$. Hence,we could bound (2.A.3) with $\frac{1}{2\pi}\Omega_n$.

Besides, Lemma A.4 in Sun et al. [2018b] shows that the second term in (2.A.1) is bounded by $\frac{m}{n}\Omega_n + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n\right)$. Combining these two bounds, we complete the proof. □

2.A.2 Proof for Lemma 2.3.3

Proof. We will only present the proof for building deviation bound for $\hat{\theta}^{(r)}_{j,rs}$ and the same argument applies to achieve deviation bound for $\hat{\theta}^{(i)}_{j,rs}$. For $j \in F_n$, let y_j(we omit r, s from notation for simplicity) be the vector of length m, composed by $\mathbf{Re}(H(\omega_k)_{rs}), k \in \mathcal{B}_j$. Then

$$\begin{aligned}\mathbb{E}\hat{\theta}^{(r)}_{j,rs} &= \frac{1}{m-1}\mathbb{E}\left\{y_j^\top\left[I-\frac{1}{m}11^\top\right]y_j\right\} = \frac{1}{m-1}\mathrm{Tr}\left(\left[\mathrm{I}-\frac{1}{\mathrm{m}}11^\top\right]\mathrm{D}\right)\\ &= \frac{1}{m-1}\left\{\sum_{q\in B^j_m}\left(1-\frac{1}{m}\right)D(q,q)+\frac{1}{m}\sum_{q_1\neq q_2,\in B^j_m}D(q_1,q_2)\right\},\end{aligned} \tag{2.A.5}$$

where

$$D(q_1,q_2) = \mathrm{Cov}(\mathbf{Re}(H_{rs}(\omega_{q_1})),\mathbf{Re}(H_{rs}(\omega_{q_2}))), q_1,q_2 \in \mathcal{B}_j. \tag{2.A.6}$$

If $D(q,q)$ share the same value as θ^r_{rs} and $D(q_1,q_2) = 0$ then $\hat{\theta}^{(r)}_{j,rs}$ is an unbiased estimator. So in this proof we bound $|D(q,q)-\theta^r_{rs}|$ and $|D(q_1,q_2)|$.

Bound $|D(q,q)-\theta^{(r)}_{j,rs}|$:

We first bound $|D(q,q)-\theta^{(r)}_{j,rs}|, q \in \mathcal{B}_j$. For e_r, e_s, applying results of Gaussian fourth moments in section 2.C,

$$\begin{aligned}&|D(q,q)-\theta^{(r)}_{j,rs}|\\ &= \left|4\mathrm{Var}(\langle\mathcal{X}^\top c_q,e_r\rangle\langle\mathcal{X}^\top c_q,e_s\rangle)-\mathrm{Var}(e_r^\top\mathbf{Re}(H_\infty(\omega_j))e_s)\right|\\ &\leqslant 4|\mathrm{Var}(\langle\mathcal{X}^\top c_q,e_r\rangle)\mathrm{Var}\langle\mathcal{X}^\top c_q,e_s\rangle)-\mathrm{Var}(\langle\mathbf{Re}(d_{\infty,j}),e_r\rangle)\mathrm{Var}\langle\mathbf{Re}(d_{\infty,j}),e_s\rangle)|\\ &+4|\mathbb{E}^2(\langle\mathcal{X}^\top c_q,e_r\rangle\langle\mathcal{X}^\top c_q,e_s\rangle))-\mathbb{E}^2(\langle\mathbf{Re}(d_{\infty,j}),e_r\rangle\langle\mathbf{Re}(d_{\infty,j}),e_s\rangle))|\end{aligned} \tag{2.A.7}$$

As shown before, $2\text{Var}(\langle \mathbf{Re}(d_{\infty,j}), e_r\rangle) = f_{rr}(\omega_j)$, then

$$
\begin{aligned}
&|\text{Var}(\langle \mathcal{X}^\top c_q, e_r\rangle) - \text{Var}(\langle \mathbf{Re}(d_{\infty,j}), e_r\rangle)| = \frac{1}{2}\left|e_r^\top \mathbb{E}[\mathbf{Re}(H(\omega_q))]e_r - e_r^\top \mathbf{Re}(f(\omega_j))e_r\right| \\
&\leqslant \frac{1}{2}|u^\top \mathbb{E}[\mathbf{Re}(H(\omega_q))]e_r - e_r^\top \mathbb{E}[\mathbf{Re}(I(\omega_q))]e_r| + \frac{1}{2}|e_r^\top \mathbb{E}[\mathbf{Re}(I(\omega_q))]e_r - e_r^\top \mathbf{Re}\mathbb{E}[(I(\omega_j))]e_r| \\
&+ \frac{1}{2}|e_r^\top \mathbb{E}[I(\omega_j)]e_r - e_r^\top f(\omega_j)e_r| \\
&\leqslant \frac{\Omega_n}{4n\pi} + \frac{\sqrt{2}}{4\pi}\frac{|j-k|\Omega_n}{n} + \frac{1}{4\pi}\left(\frac{\Omega_n}{n} + L_n\right)
\end{aligned}
\tag{2.A.8}
$$

where the first part of last line is from same technique used in lemma 2.3.2 and the second and third parts are from Sun et al. [2018b]. Therefore,

$$
\begin{aligned}
&4|\text{Var}(\langle \mathcal{X}^\top c_q, e_r\rangle)\text{Var}(\langle \mathcal{X}^\top c_q, e_s\rangle) - \text{Var}(\langle \mathbf{Re}(d_{\infty,j}), e_r\rangle)\text{Var}(\langle \mathbf{Re}(d_{\infty,j}), e_s\rangle)| \\
&\leqslant |4\text{Var}(\langle \mathcal{X}^\top c_q, e_r\rangle)\text{Var}(\langle \mathcal{X}^\top c_q, e_s\rangle) - 2\text{Var}(\langle \mathcal{X}^\top c_q, e_r\rangle)2\text{Var}(\langle d^r_{\infty,j}, e_s\rangle)| \\
&+ |2\text{Var}(\langle \mathcal{X}^\top c_q, e_r\rangle)2\text{Var}(\langle d^r_{\infty,j}, e_s\rangle) - 2\text{Var}(\langle d^r_{\infty,j}, e_r\rangle)2\text{Var}(\langle d^r_{\infty,j}, e_s\rangle)| \\
&\leqslant 2f_{rr}(\omega_j)\delta_1 + \delta_1^2
\end{aligned}
\tag{2.A.9}
$$

where $\delta_1 = \frac{\Omega_n}{2n\pi} + \frac{\sqrt{2}}{2\pi}\frac{m\Omega_n}{n} + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n\right)$. Here we assume $f_{rr}(\omega_j) \geqslant f_{ss}(\omega_j)$ without loss of generality. Now we bound last line in (2.A.7), noticing

$$
2\mathbb{E}[\langle \mathcal{X}^\top c_q, e_r\rangle\langle \mathcal{X}^\top c_q, e_s\rangle] = e_r^\top f(\omega_q)e_s
$$

,

$$
\begin{aligned}
&4|\mathbb{E}^2(\langle \mathcal{X}^\top c_q, e_r\rangle\langle \mathcal{X}^\top c_q, e_s\rangle)) - \mathbb{E}^2(\langle d^r_{\infty,j}, e_r\rangle\langle d^r_{\infty,j}, e_s\rangle))| \\
&\leqslant |e_r^\top \mathbb{E}[\mathbf{Re}(H(\omega_q)]e_s + e_r^\top \mathbf{Re}(f(\omega_j))e_s||\mathbb{E}(e_r^\top \mathbf{Re}(H(\omega_q))e_s - e_r^\top \mathbf{Re}(f(\omega_j))e_s| \\
&\leqslant 2f_{rr}(\omega_j)\delta_2 + \delta_2^2,
\end{aligned}
\tag{2.A.10}
$$

where $\delta_2 = \frac{\Omega_n}{2n\pi} + \frac{1}{2\pi}\left(\frac{\Omega_n}{n} + L_n(f_X)\right)$. Here we use the fact that for $u, v \in E_p$,

$$
\begin{aligned}
&|\mathbb{E}[2(\langle d^r_{\infty,j}, e_r\rangle\langle d^r_{\infty,j}, v\rangle)]| = |e_r^\top f(\omega_j)e_s| \\
&\leqslant \sqrt{(e_r^\top f(\omega_j)e_r)(e_s^\top f(\omega_j)e_s)} \leqslant f_{rr}(\omega_j),
\end{aligned}
\tag{2.A.11}
$$

and same techniques to get bound as (2.A.8).

Bound $|D(q_1, q_2)|$:

For any unit vector u, v, applying summary of Gaussian fourth moments 2.C.1,

$$\begin{aligned} &4|\text{Cov}\left(\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_1}, v\rangle, \langle \mathcal{X}^\top c_{q_2}, u\rangle\langle \mathcal{X}^\top c_{q_2}, v\rangle\right)| \\ &= 4\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right]\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, v\rangle\langle \mathcal{X}^\top c_{q_2}, v\rangle\right] \\ &+ 4\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, v\rangle\right]\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, v\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right] \end{aligned} \tag{2.A.12}$$

We will show each of these four terms well bounded. For $\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right]$,

$$\begin{aligned} &|\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right]| = |\mathbb{E}c_{q_1}^\top \Sigma^{u,u} c_{q_1}| \\ &\leqslant \frac{\Omega(\Sigma^{u,u})}{2n\pi} \leqslant \frac{\Omega_n}{2n\pi}, \end{aligned} \tag{2.A.13}$$

where $\Sigma_{rs}^{u,u} = u^\top \Gamma(s-r)u$ and the last inequality comes from the same argument we used in lemma 2.3.2. With almost same argument, we could show

$$\begin{aligned} &\max\left(\left|\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right]\right|, \left|\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, v\rangle\langle \mathcal{X}^\top c_{q_2}, v\rangle\right]\right| \right. \\ &\left.\left|\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, u\rangle\langle \mathcal{X}^\top c_{q_2}, v\rangle\right]\right|, \left|\mathbb{E}\left[\langle \mathcal{X}^\top c_{q_1}, v\rangle\langle \mathcal{X}^\top c_{q_2}, u\rangle\right]\right|\right) \leqslant \frac{\Omega_n}{2\pi n} \end{aligned} \tag{2.A.14}$$

Taking $u = e_{q_1}, v = e_{q_2}$, we show that

$$D(q_1, q_2) \leqslant \frac{\Omega_n^2}{\pi^2 n^2}. \tag{2.A.15}$$

Above all, $\left|\mathbb{E}\hat{\theta}_{j,(r,s)}^{(r)} - \text{Var}(\mathbf{Re}(H_{r,s}(\omega_j)))\right| \leqslant 2\max(f_{rr}(\omega_j), f_{ss}(\omega_j))(\delta_1 + \delta_2) + \delta_1^2 + \delta_2^2 + \frac{\Omega_n^2}{\pi^2 n^2}$. Then by same argument, we could get the proof for imaginary part. $\square$

2.A.3 Proof for Lemma 2.3.4

Proof. We will only provide the proof for real part.

$$\begin{aligned}\mathrm{Var}(g_{rs}(\omega_j)) &= \mathrm{Var}\left(\frac{1}{m}\sum_{q\in\mathcal{B}_j} H_{rs}(\omega_q)\right)\\ &= \sum_{q\in\mathcal{B}_j}\frac{1}{m^2}(\mathrm{Var}(H_{rs}(\omega_q)))\\ &+ \frac{1}{m^2}\sum_{q_1\neq q_2\in\mathcal{B}_j}\mathrm{Cov}[H_{rs}(\omega_{q_1}), H_{rs}(\omega_{q_2})].\end{aligned} \tag{2.A.16}$$

From proof in lemma 2.3.3, noticing $\delta_1 \geqslant \delta_2$,

$$\left|\mathrm{Var}(H_{rs}(\omega_q)) - \theta^{(r)}_{j,rs}\right| \leqslant 4\max\{f_{rr}(\omega_j), f_{ss}(\omega_j)\}\delta_1 + 2\delta_1^2,$$

and

$$\mathrm{Cov}(H_{rs}(\omega_{q_1}), H_{rs}(\omega_{q_2})) \leqslant \frac{\Omega_n^2}{\pi^2 m^2}.$$

Combining these two and the fact that $\theta^r_{j,rs} \geqslant f_{rr}(\omega_j)f_{ss}(\omega_j)$ and assumption $\min_{r=1}^p f_{rr}(\omega_j) \geqslant \phi_0$ we can get the result. □

2.B Proof for Deviation Bound

2.B.1 Proof for Lemma 2.3.5

From lemma 2.3.3, we have

$$\begin{aligned}&\mathbb{P}\left(|\hat{\theta}^{(r)}_{j,rs} - \theta^{(r)}_{j,rs}| \geqslant \eta + B_\theta\right)\\ &\leqslant \mathbb{P}\left(|\hat{\theta}^{(r)}_{j,rs} - \mathbb{E}[\hat{\theta}^{(r)}_{j,rs}]| \geqslant \eta\right)\end{aligned} \tag{2.B.1}$$

we could write $\hat{\theta}^{(r)}_{j,rs} = y_j' D y_j$ where y_j is vector of length m composed of $\mathrm{Re}(\mathrm{H_{rs}}(\omega_{\mathrm{q}}))), \mathrm{q} \in \mathcal{B}_{\mathrm{j}}$ and $D = \frac{1}{m-1}(I - \frac{1}{m}11^\top)$. $I - \frac{1}{m}11^\top$ is a symmetric projection matrix and we know for any symmetric projection matrix P, $\|P\| = 1$ and $\|P\|_F^2 = \mathrm{rk}(P)$. Thus, $\|D\| = 1/(m-1)$ and $\|D\|_F = \sqrt{1/(m-1)}$. Then we could apply Hanson-Wright inequality [Rudelson and Vershynin, 2013] to (2.B.1), there exists universal constants c_1, c_2,

$$\begin{aligned} &\mathbb{P}\left(|\hat{\theta}^{(r)}_{j,rs} - \mathbb{E}[\hat{\theta}^{(r)}_{j,rs}]| \geqslant \eta\right) \leqslant \exp\left(-c\min\left(\frac{\eta}{\|D\|}, \frac{\eta^2}{\|D\|_F^2}\right)\right) \\ &= c_1 \exp(-c_2 m \min\left(\eta, \eta^2\right)). \end{aligned} \tag{2.B.2}$$

2.C Technical Lemmas

2.C.1 Fourth Moments of Multivariate Normal Distribution

Let $X \in \mathbb{R}^p$, $X \sim \mathcal{N}(0, \Sigma)$ and $\sigma_{rs} = \Sigma_{r,s}$, we have

$$\begin{aligned} &\mathbb{E}X_r^4 = 3\sigma_{rr}^2 \\ &\mathbb{E}X_r^3 X_s = 3\sigma_{rr}\sigma_{rs} \\ &\mathbb{E}X_r^2 X_s^2 = \sigma_{rr}\sigma_{ss} + 2\sigma_{rs}^2 \\ &\mathrm{Var}(X_r X_s) = \mathbb{E}X_r^2 X_s^2 - (\mathbb{E}X_r X_s)^2 = \sigma_{rr}\sigma_{ss} + \sigma_{rs}^2 \\ &\mathbb{E}X_r^2 X_s X_k = \sigma_{rr}\sigma_{sk} + 2\sigma_{rs}\sigma_{rk} \\ &\mathbb{E}X_r X_s X_k X_q = \sigma_{rs}\sigma_{kq} + \sigma_{rk}\sigma_{sq} + \sigma_{rq}\sigma_{sk} \\ &\mathrm{Cov}(X_r X_s, X_k X_q) = \sigma_{rk}\sigma_{sq} + \sigma_{rq}\sigma_{sk}. \end{aligned} \tag{2.C.1}$$

Lemma 2.C.1. *Let y be a Gaussian random vector, $y \sim \mathcal{N}(0, \Sigma)$, given a symmetric real matrix A, there exit positive constants c_1, c_2, s.t.*

$$\mathbb{P}\left(\left|y^\top A y - \mathbb{E}y^\top A y\right| \geqslant \frac{1}{\sqrt{2}}\sqrt{\mathit{Var}(y^\top A y)}\eta\right) \leqslant \exp(-c\eta^2). \tag{2.C.2}$$

Proof. Since $\text{Var}(y^\top Ay) = \text{Var}(z^\top \Sigma^{1/2} A \Sigma^{1/2} z) = 2\|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2$ [Rencher and Schaalje, 2008], where $z \sim \mathcal{N}(0, I)$.

$$\begin{aligned} &\mathbb{P}\left(\left|y^\top Ay - \mathbb{E}y^\top Ay\right| \geqslant \frac{1}{\sqrt{2}}\sqrt{\text{Var}(y^\top Ay)}\eta\right) \\ &= \mathbb{P}\left(\left|z^\top \Sigma^{1/2} A \Sigma^{1/2} z - \mathbb{E}z^\top \Sigma^{1/2} A \Sigma^{1/2} z\right| \geqslant \|\Sigma^{1/2} A \Sigma^{1/2}\|_F \eta\right). \end{aligned} \tag{2.C.3}$$

Then applying Hanson Wright inequality [Rudelson and Vershynin, 2013] into above tail probability, there exit positive contants c_1, c_2,

$$\begin{aligned} &\mathbb{P}\left(\left|y^\top Ay - \mathbb{E}y^\top Ay\right| \geqslant \frac{1}{\sqrt{2}}\sqrt{\text{Var}(y^\top Ay)}\eta\right) \\ &\leqslant c_1 \exp\left(-c_2 \min\left(\eta^2, \eta\frac{\|\Sigma^{1/2} A \Sigma^{1/2}\|_F}{\|\Sigma^{1/2} A \Sigma^{1/2}\|}\right)\right) \\ &= c_1 \exp(-c_2 \min(\eta^2, \eta)). \end{aligned} \tag{2.C.4}$$

Here, the last equality comes from the fact that $\|\cdot\|_F \geqslant \|\cdot\|$. □

2.D Technical Results for Toeplitz Matrixz

Lemma 2.D.1.

$$\begin{aligned} \sum_{k=1}^{n} \cos(a + kx) &= \frac{\sin((1/2 + n)x + a) - \sin(x/2 + a)}{2\sin(x/2)} \\ \sum_{k=1}^{n} \sin(a + kx) &= \frac{\cos(x/2 + a) - \cos((1/2 + n)x + a)}{2\sin(x/2)} \end{aligned} \tag{2.D.1}$$

Proof. Proof is from element math which can be found in Zygmund [2002]. □

Lemma 2.D.2. *Let M be a toeplitz matrix and $M_{p,q} = g(q - p)$ for some function $g(\cdot)$. Now we claim following results with definition of definition of* $\cos, \sin$ *sequence in 2.1.4, for $j \neq k$ and $\omega_j \neq 0, \pi$*

$$\max\{|c_j^\top M c_j - s_j^\top M s_j|, |s_j^\top M c_j + c_j^\top M s_j|\} \leqslant \frac{\Omega(M)}{2\pi n}.$$

and

$$\max\left\{|c_j^\top M s_k|, |c_j^\top M c_k|, |s_j^\top M s_k|, |s_j^\top M c_k|\right\} \leqslant \frac{\Omega(M)}{2\pi n},$$

where linear operator Ω on toeplitz matrix is defined as

$$\Omega(M) = \sum_{\ell=-(n-1)}^{(n-1)} |\ell||g(\ell)|. \tag{2.D.2}$$

Remark. *Lemma 2.1.2 claims $X^\top c_j$ and $X^\top s_j$ have same limiting marginal distribution, and the first part of this lemma, in fact could be used to quantify this similarity in finite sample through quantification of the difference in two covariance matrix which we will show later. The second part could be used to quantify the rate of being asymptotically independence across different frequency claimed by lemma 2.1.2.*

There is many literature showing similar results in the second part, but usually it points out vanishing rate for any frequency between $[-\pi, \pi]$ in an asymptotic sense which is different from my theory. None of them provides an explicit finite sampling bound at discrete Fourier frequencies and.

Proof. We will only prove $|c_j^\top M c_j - s_j^\top M s_j| \leqslant \frac{\Omega(M)}{2\pi n}$, then all the others could be

proven with same techniques.

$$
\begin{aligned}
&|c_j^\top M c_j - s_j^\top M s_j| \\
&= \frac{1}{2\pi n}\left|\sum_{\ell=-(n-1)}^{(n-1)}\sum_{p=0}^{(n-1)-|\ell|} g(\ell)\left[\cos(p\omega_j)\cos((p+\ell)\omega_j) - \sin(p\omega_j)\sin((p+\ell)\omega_j)\right]\right| \\
&\leqslant \frac{1}{2\pi n}\left|\sum_{\ell=-(n-1)}^{n-1}\sum_{p=0}^{n-1} g(\ell)\left[\cos(p\omega_j)\cos((p+\ell)\omega_j) - \sin(p\omega_j)\sin((p+\ell)\omega_j)\right]\right| \\
&+ \frac{1}{2\pi n}\left|\sum_{\ell=-(n-1)}^{(n-1)}\sum_{p=(n-1)-|\ell|}^{n-1} g(\ell)\left[\cos(p\omega_j)\cos((p+\ell)\omega_j) - \sin(p\omega_j)\sin((p+\ell)\omega_j)\right]\right| \\
&\leqslant \frac{1}{2\pi n}\left|\sum_{\ell=-(n-1)}^{n-1}\sum_{p=0}^{n-1} g(\ell)\cos(2p\omega_j + \ell\omega_j)\right| + \frac{1}{2\pi n}\sum_{\ell=-(n-1)}^{(n-1)} |\ell||g(\ell)|.
\end{aligned}
\tag{2.D.3}
$$

By setting $a = (\ell-2)\omega_j$, $x = 2\omega_j$ in lemma 2.D.1 and noticing $2n\omega_j = 4j\pi$, we get $\sum_{p=0}^{n-1} g(\ell)\cos(2p\omega_j + \ell\omega_j) = 0$ for any ℓ. Therefore, (2.D.3) leads to

$$
|c_j^\top M c_j - s_j^\top M s_j| \leqslant \frac{1}{2\pi n}\sum_{\ell=-(n-1)}^{(n-1)} |\ell||g(\ell)| = \frac{\Omega(M)}{2\pi n}. \tag{2.D.4}
$$

With similar techniques, we could prove the second part. □

2.E An Example Explaining Why We Modify the Periodogram

In this session, we present a stationary time series which makes variance shown in (**??**) could be small as possible, which damages the argument in Cai and Liu [2011]. In our counter example, we first present the the spectral which makes order of its variance as small as possible.

Consider a Vector autoregression model with length 2 whose spectral density has

the form $f(\omega) = \begin{bmatrix} 1 & i\gamma(\omega) \\ -i\gamma(\omega) & 1 \end{bmatrix}$, where $\gamma(\omega) = \gamma_1(\omega) * \delta_1(\omega)$, $*$ denote the convolution, $\delta_1 \in C^3$ is an approximate function of δ-function and

$$\gamma_1(\omega) = \begin{cases} 1 & \text{if } \omega > \frac{\pi}{2} \\ \frac{2}{\pi}\omega & \text{if } |\omega| \leqslant \frac{\pi}{2} \\ -1 & \text{if } \omega < -\frac{\pi}{2} \end{cases}$$

We know that $\gamma(\omega) = \gamma_1(w) * \delta_1(\omega)$ and for any ϵ, there exists δ_1 such that $|\gamma(\omega) - \gamma_1(\omega)| \leqslant \epsilon$. Also, $\gamma(\omega) \in C^3$ is an odd function.

We say $f(\omega)$ can be a spectral density matrix since it satisfies that $\sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\| < \infty$. In fact for the diagonal entries, $\Gamma(\ell)_{ii} = 0, \ell \neq 0$ and $\Gamma(0)_{ii} = 1$. For the off diagonal entries, firstly they are all real numbers less than 1 and $\sum_{\ell=-\infty}^{\infty} |\Gamma_{12}(\ell)| \leqslant \sum_{\ell=1}^{\infty} \frac{M}{\ell^3} < \infty$ due to the property of inverse Fourier transform of C^3.

Therefore, given this $f(\omega)$, since $\gamma(\omega)$ can be close to 1 arbitrarily, we find the variance of the original estimator is not the same order as $f_{11}(\omega_j) f_{22}(\omega_j)$.

CHAPTER 3

LOW-RANK TUCKER APPROXIMATION OF A TENSOR FROM STREAMING DATA

3.1 Introduction

Large-scale datasets with natural tensor (multidimensional array) structure arise in a wide variety of applications including computer vision [Vasilescu and Terzopoulos, 2002], neuroscience [Cichocki, 2013], scientific simulation [Austin et al., 2016], sensor networks [Sun et al., 2008], and data mining [Kolda and Sun, 2008]. In many cases, these tensors are too large to manipulate, to transmit, or even to store in a single machine. Luckily, tensors often exhibit a low-rank structure, and can be approximated by a low-rank tensor factorization, such as CANDECOMP/PARAFAC (CP), tensor train, or Tucker factorization [Kolda and Bader, 2009]. These factorizations reduce the storage costs by exposing the latent structure. Sufficiently low rank tensors can be compressed by several orders of magnitude with negligible loss. However, computing these factorizations can require substantial computational resources. Indeed, one particular challenge is that these large tensors may not fit in main memory on our computer.

In this paper, we develop a new algorithm to compute a low-rank Tucker approximation for a tensor from streaming data, using storage proportional to the degrees of freedom in the output Tucker approximation. The algorithm forms a linear sketch of the tensor, and operates on the sketch to compute a low-rank Tucker approximation. Importantly, the main computational work is all performed on a small tensor, of size proportional to the core tensor of the Tucker factorization. We derive detailed probabilistic error bounds on the quality of the approximation in terms of the tail energy of

any matricization of the target tensor.

This algorithm is useful in at least three concrete problem settings:

1 **Streaming:** Data from the tensor is generated sequentially. At each time stamp, we may observe a low dimensional slice, an individual entry, or an additive update to the tensor (the so-called "turnstile" model [Muthukrishnan et al., 2005]). For example, each slice of the tensor may represent a subsequent time step in a simulation, or sensor measurements at a particular time. In the streaming setting, the complete tensor is not stored; indeed, it may be much larger than available computing resources.

   Our algorithm can approximate tensors revealed via streaming updates by sketching the updates and storing the sketch. Linearity of the sketch guarantees that sketching commutes with slice, entrywise, or additive updates. Our method forms an approximation of the tensor only after all the data has been observed, rather than approximating the tensor-observed-so-far at any time. This protocol allows for offline data analysis, including many scientific applications. Conversely, this protocol is not suitable for real-time monitoring.

2 **Limited memory:** Data describing the tensor is stored on a hard disk of a computer with much smaller RAM. This setting reduces to the streaming setting by streaming the data from disk.

3 **Distributed:** Data describing the tensor may be stored on many different machines. Communicating data between these machines may be costly due to low network bandwidth or high latency. Our algorithm can approximate tensors stored in a distributed computing enviroment by sketching the data on each slave machine and transmitting the sketch to a master, which computes the sum

of the sketches. Linearity of the sketch guarantees that the sum of the sketches is the sketch of the full tensor.

In the streaming setting, the tensor is not stored, so we require an algorithm that can compute an approximation from a single pass over the data. In constrast, multiple passes over the data are possible in the memory-limited or distributed settings.

This paper presents algorithms for all these settings, among other contributions:

- We present a new method to form a linear sketch of an unknown tensor. This sketch captures both the principal subspaces of the tensor along each mode, and the action of the tensor that links these subspaces. This tensor sketch can be formed from any dimension reduction map. Those sketches themselves are useful in many application even without forming Tucker decomposition like in video clustering.
- We develop a practical one-pass algorithm to compute a low rank Tucker approximation from streaming data. The algorithm sketches the tensor and then recovers a low rank Tucker approximation from this sketch.
- We propose a two-pass algorithm that improves on the one-pass method. Both the one-pass and two-pass methods are appropriate in a limited memory or distributed data setting.
- We develop provable probabilistic guarantees on the performance of both the one-pass and two-pass algorithms when the tensor sketch is composed of Gaussian dimension reduction maps.
- We exhibit several random maps that can be used to sketch the tensor. Compared to the Gaussian map, others are cheaper to store, easier to apply, and deliver sim-

ilar performance experimentally in tensor approximation error. In particular, we demonstrate the effective performance of a row-product random matrix, which we call the Tensor Random Projection (TRP), which uses exceedingly low storage.

- We perform a comprehensive simulation study with synthetic data, and consider applications to several real datasets, to demonstrate the practical performance of our method. Our methods reduce approximation error compared to the only existing one-pass Tucker approximation algorithm [Malik and Becker, 2018] by more than an order of magnitude given the same storage budget.
- We have developed and released an open-source package in python that implements our algorithms.

3.2 Background and Related Work

3.2.1 Notation

Our paper follows the notation of [Kolda and Bader, 2009]. We denote *scalar*, *vector*, *matrix*, and *tensor* variables respectively by lowercase letters (x), boldface lowercase letters ($\mathbf{x}$), boldface capital letters ($\mathbf{X}$), and boldface Euler script letters ($\boldsymbol{\mathcal{X}}$). For two vectors $\mathbf{x}$ and $\mathbf{y}$, we write $\mathbf{x} \succ \mathbf{y}$ if $\mathbf{x}$ is greater than $\mathbf{y}$ elementwise and $\mathbf{x} \prec \mathbf{y}$ as the opposite.

Define $[N] := \{1, \ldots, N\}$. For a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, we denote its i^{th} row, j^{th} column, and $(i,j)^{th}$ element as $\mathbf{X}(i,.)$, $\mathbf{X}(.,j)$, and $\mathbf{X}(i,j)$, respectively, for $i \in [m]$, $j \in [n]$. We use $\mathbf{X}^{\dagger} \in \mathbb{R}^{n \times m}$ to denote the *Moore–Penrose pseudoinverse* of the matrix

$\mathbf{X} \in \mathbb{R}^{m \times n}$. In particular, $\mathbf{X}^\dagger = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^T$ if $m \geqslant n$ and $\mathbf{X}$ has full column rank; $\mathbf{X}^\dagger = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$, if $m < n$ and $\mathbf{X}$ has full row rank.

Tail energy

To state our results, we will need a tensor equivalent for the decay in the spectrum of a matrix. For each unfolding $\mathbf{X}^{(n)}$, define the *ρth tail energy*

$$(\tau_\rho^{(n)})^2 := \sum_{k>\rho}^{\min(I_n, I_{(-n)})} \sigma_k^2(\mathbf{X}^{(n)}),$$

where $\sigma_k(\mathbf{X}^{(n)})$ is the kth largest singular value of $\mathbf{X}^{(n)}$.

Kronecker and Khatri-Rao product

For two matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$, we define the *Kronecker product* $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{IK \times JL}$ as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}(1,1)\mathbf{B} & \cdots & \mathbf{A}(1,J)\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}(I,1)\mathbf{B} & \cdots & \mathbf{A}(I,J)\mathbf{B} \end{bmatrix}. \tag{3.2.1}$$

For $J = L$, we define the *Khatri-Rao product* as $\mathbf{A} \odot \mathbf{B}$, i.e. the "matching column-wise" Kronecker product. The resulting matrix of size $(IJ) \times K$ is defined as

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{A}(\cdot,1) \otimes \mathbf{B}(\cdot,1) \cdots \mathbf{A}(\cdot,K) \otimes \mathbf{B}(\cdot,K)]$$

Tensor basics

For a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, its *mode* or *order* is the number N of dimensions. If $I = I_1 = \cdots I_N$, we denote $\mathbb{R}^{I_1 \times \cdots \times I_N}$ as $\mathbb{R}^{I^N}$. The inner product of two tensors $\mathcal{X}, \mathcal{Y}$

is defined as $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1 \ldots i_N} \mathcal{Y}_{i_1 \ldots i_N}$. The *Frobenius norm* of $\mathcal{X}$ is $\|\mathcal{X}\|_F = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$.

Tensor unfoldings

Let $\bar{I} = \Pi_{j=1}^{N} I_j$ and $I_{(-n)} = \Pi_{j \neq n} I_j$, and let $\mathbf{vec}(\mathcal{X})$ denote the vectorization of $\mathcal{X}$. The *mode-n unfolding* of $\mathcal{X}$ is the matrix $\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times I_{(-n)}}$. The inner product for tensors matches that of any mode-n unfolding:

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \langle \mathbf{X}^{(n)}, \mathbf{Y}^{(n)} \rangle = \mathrm{Tr}((\mathbf{X}^{(\mathrm{n})})^\top \mathbf{Y}^{(\mathrm{n})}). \tag{3.2.2}$$

Tensor rank

The *mode-n rank* is the rank of the mode-n unfolding. We say the *rank* of $\mathcal{X}$ is $\mathbf{r}(\mathcal{X}) = (r_1, \ldots, r_N)$ if its *mode-n rank* is r_n for each $n \in [N]$. This notion of rank corresponds to the size of the core tensor in a Tucker factorization of $\mathcal{X}$. A *superdiagonal* tensor generalizes a diagonal matrix: all entries are zero except for the entries whose indices in each dimension are equal.

Tensor contractions

Write $\mathcal{G} = \mathcal{X} \times_n \mathbf{U}$ for the *mode-n (matrix) product* of $\mathcal{X}$ with $\mathbf{U} \in \mathbb{R}^{J \times I_n}$. That is, $\mathcal{G} = \mathcal{X} \times_n \mathbf{U} \iff \mathbf{G}^{(n)} = \mathbf{U}\mathbf{X}^{(n)}$. The tensor $\mathcal{G}$ has dimension $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N$. Mode products with respect to different modes commute: for $\mathbf{U} \in \mathbb{R}^{J_1 \times I_n}$, $\mathbf{V} \in \mathbb{R}^{J_2 \times I_m}$

$$\mathcal{X} \times_n \mathbf{U} \times_m \mathbf{V} = \mathcal{X} \times_m \mathbf{V} \times_n \mathbf{U} \quad \text{if} \quad n \neq m.$$

Mode products along the same mode simplify: for $\mathcal{A} \in \mathbb{R}^{J_1 \times I_n}$, $\mathcal{B} \in \mathbb{R}^{J_2 \times J_1}$,

$$\mathcal{X} \times_n \mathbf{A} \times_n \mathbf{B} = \mathcal{X} \times_n (\mathbf{BA}).$$

3.2.2 Tucker Approximation

Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and target rank $\mathbf{r} = (r_1, \ldots, r_N)$, the idea of Tucker approximation is finding a *core tensor* $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$ and matrices with orthogonal columns $\mathbf{U}_n \in \mathbb{R}^{I_n \times r_n}$ for $n \in [N]$, called *factor matrices*, so that

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{U}_1 \times \cdots \times_N \mathbf{U}_N.$$

For brevity, we define $[\![\mathcal{G}; \mathbf{U}_1, \cdots, \mathbf{U}_N]\!] = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \cdots \times_N \mathbf{U}_N$. Any best rank-$\mathbf{r}$ Tucker approximation is of the form $[\![\mathcal{G}^\star; \mathbf{U}_1^\star, \cdots, \mathbf{U}_N^\star]\!]$, where $\mathcal{G}^\star, \mathbf{U}_n^\star$ solve the problem

$$\begin{aligned} &\text{minimize} && \|\mathcal{X} - \mathcal{G} \times_1 \times \ldots \mathbf{U}_{n+1} \times_N \mathbf{U}_N\|_F^2 \\ &\text{subject to} && \mathbf{U}_n^\top \mathbf{U}_n = \mathbf{I}. \end{aligned} \tag{3.2.3}$$

The problem 3.2.3 is a challenging nonconvex optimization problem. Moreover, the solution is not unique [Kolda and Bader, 2009]. We use the notation $[\![\mathcal{X}]\!]_\mathbf{r}$ to represent a best rank-$\mathbf{r}$ Tucker approximation of the tensor $\mathcal{X}$, which in general we cannot compute.

HOSVD

The standard approach to computing a rank $\mathbf{r} = (r_1, \ldots, r_N)$ Tucker approximation for a tensor $\mathcal{X}$ begins with the higher order singular value decomposition (HOSVD) [De Lathauwer et al., 2000; Tucker, 1966], (Algorithm 1):

Algorithm 1 Higher order singular value decomposition (HOSVD) [De Lathauwer et al., 2000; Tucker, 1966]

Given: tensor $\mathcal{X}$, target rank $\mathbf{r} = (r_1, \ldots, r_N)$

1 *Factors.* For $n \in [N]$, compute the top r_n left singular vectors $\mathbf{U}_n$ of $\mathbf{X}^{(n)}$.

2 *Core.* Contract these with $\mathcal{X}$ to form the core

$$\mathcal{G} = \mathcal{X} \times_1 \mathbf{U}_1^T \cdots \times_N \mathbf{U}_N^T.$$

Return: Tucker approximation $\mathcal{X}_{\text{HOSVD}} = [\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!]$

The HOSVD can be computed in two passes over the tensor [Battaglino et al., 2019; Zhou et al., 2014]. We describe this method briefly here, and in more detail in the next section. In the first pass, sketch each matricization $\mathbf{X}^{(n)}$, $n \in [N]$, and use randomized linear algebra (e.g., the randomized range finder of [Halko et al., 2011]) to (approximately) recover its range $\mathbf{U}_n$. To form the core $\mathcal{X} \times_1 \mathbf{U}_1^T \cdots \times_N \mathbf{U}_N^T$ requires a second pass over $\mathcal{X}$, since the factor matrices $\mathbf{U}_n$ depend on $\mathcal{X}$. The main algorithmic contribution of this paper is to develop a method to approximate both the factor matrices and the core in just one pass over $\mathcal{X}$.

ST-HOSVD

[Vannieuwenhoven et al., 2012] proposes a sequentially truncated higher-order singular value decomposition which enjoys the same approximation compared to HOSVD, but requires less operators. The idea, is to update the low rank approximation of the target tensor $\mathcal{X}$ when we get factor matrix in Algorithm 1 for the following factor extraction. Readers can go to [Vannieuwenhoven et al., 2012] for more details.

HOOI

The higher order orthogonal iteration (HOOI) [De Lathauwer et al., 2000] (2) improves on the resulting Tucker factorization by repeatedly minimizing the objective of 3.2.3 over the the core and the factor matrices. Notice the core update (3.2.5) admits the

Algorithm 2 Higher order orthogonal iteration (HOOI) [De Lathauwer et al., 2000]

Given: tensor $\mathcal{X}$, target rank $\mathbf{r} = (r_1, \ldots, r_N)$
Initialize: compute $\mathcal{X} \approx [\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!]$ using HOSVD
Repeat:

1 *Factors.* For each $n \in [N]$,

$$\mathbf{U}_n \leftarrow \underset{\mathbf{U_n}}{\operatorname{argmin}} \|[\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!] - \mathcal{X}\|_F^2 , \tag{3.2.4}$$

2 *Core.*

$$\mathcal{G} \leftarrow \underset{\mathcal{G}}{\operatorname{argmin}} \|[\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!] - \mathcal{X}\|_F^2 . \tag{3.2.5}$$

Return: Tucker approximation $\mathcal{X}_{\text{HOOI}} = [\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!]$

closed form solution $\mathcal{G} \leftarrow \mathcal{X} \times_1 \mathbf{U}_1^\top \cdots \times_N \mathbf{U}_N^\top$, which motivates the second step of HOSVD.

3.2.3 Previous Work

The only previous work on streaming Tucker approximation is [Malik and Becker, 2018], which develops a streaming method called Tucker TensorSketch (T.-TS) [Malik and Becker, 2018, Algorithm 2]. T.-TS improves on HOOI by sketching the data matrix in the least squares problems. However, the success of the approach depends on the quality of the initial core and factor matrices, and the alternating least squares algorithm takes several iterations to converge.

In contrast, our work is motivated by HOSVD (not HOOI), and requires no initialization or iteration. We treat the tensor as a *multilinear* operator. The sketch identifies a low-dimensional subspace *for each input argument* that captures the action of the operator. The reconstruction produces a low-Tucker-rank multilinear operator with the same action on this low-dimensional tensor product space. This linear algebraic view allows us to develop the first guarantees on approximation error for this class of problems[1]. Moreover, we show in numerical experiments that our algorithm achieves a better approximation of the original tensor given the same memory resources.

More generally, there is a large literature on randomized algorithms for matrix factorizations and for solving optimization problems; see e.g. the review articles [Halko et al., 2011; Woodruff et al., 2014]. In particular, our method is strongly motivated by the recent papers [Tropp et al., 2018, 2019b], which provide methods for one-pass matrix approximation. The novelty of this paper is in our design of a core sketch (and reconstruction) for the Tucker decomposition, together with provable performance guarantees. The proof requires a careful accounting of the errors resulting from the factor sketches and from the core sketch. The structure of the Tucker sketch guarantees that these errors are independent.

Many researchers have used randomized algorithms to compute tensor decompositions. For example, [Battaglino et al., 2018; Wang et al., 2015] apply sketching techniques to the CP decomposition, while [Tsourakakis, 2010] suggests sparsifying the tensor. Several papers aim to make Tucker decomposition efficient in the limited-memory or distributed settings [Austin et al., 2016; Baskaran et al., 2012; Battaglino

[1]The guarantees in [Malik and Becker, 2018] hold only when a new sketch is applied for each subsequent least squares solve; the resulting algorithm cannot be used in a streaming setting. In contrast, the practical streaming method T.-TS fixes the sketch for each mode, and so has no known guarantees. Interestingly, experiments in [Malik and Becker, 2018] show that the method achieves lower error using a fixed sketch (with no guarantees) than using fresh sketches at each iteration.

et al., 2019; Kaya and Uçar, 2016; Li et al., 2015; Zhou et al., 2014].

3.3 Dimension Reduction Maps

In this section, we first introduce some commonly used randomized dimension reduction maps together with some mathematical background, and explain how to calculate and update sketches.

3.3.1 Dimension Reduction Map

Dimension reduction maps (DRMs) take a collection of high dimensional objects to a lower dimensional space while preserving certain geometric properties Oymak and Tropp [2015]. For example, we may wish to preserve the pairwise distances between vectors, or to preserve the column space of matrices. We call the output of a DRM on an object x a *sketch* of x.

Some common DRMs include matrices with i.i.d. Gaussian entries or i.i.d. ± 1 entries. The Scrambled Subsampled Randomized Fourier Transform (SSRFT) Woolfe et al. [2008] and sparse random projections Achlioptas [2003]; Li et al. [2006] can achieve similar performance with fewer computational and storage requirements; see 3.F for details.

Our theoretical bounds rely on properties of the Gaussian DRM. However, our numerical experiments indicate that many other DRMs yield qualitatively similar results; see, e.g., Figure 3.1, Figure 3.10 and Figure 3.9) in Figure 4.B.

3.3.2 Tensor Random Projection

Here we present a strategy for reducing the storage of the random map that makes use of the tensor random projection (TRP), and extremely low storage structured dimension reduction map proposed in Sun et al. [2018a]. The *tensor random projection (TRP)* $\mathbf{\Omega}$: $\prod_{n=1}^{N} I_n \to \mathbb{R}^k$ is defined as the iterated Khatri-Rao product of DRMS $\mathbf{A}_n \in \mathbb{R}^{I_n \times k}$, $n \in [N]$:

$$\mathbf{\Omega} = \mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N. \tag{3.3.1}$$

Each $\mathbf{A}_n \in \mathbb{R}^{I_n \times k}$ can be a Gaussian map, sign random projection, SSRFT, etc. The number of constituent maps N and their dimensions I_n for $n \in [N]$ are parameters of the TRP, and control the quality of the map; see Sun et al. [2018a] for details. The TRP map is a row-product random matrix, which behaves like a Gaussian map in many respects Rudelson [2012]. Our experimental results confirm this behavior.

Supposing each I_n is the same for $n \in [N]$, the TRP can be formed (and stored) using only kNI random variables, while standard dimension reduction maps use randomness (and storage) that grows as I^N when applied to a generic (dense) tensor. 3.1 compares the computational and storage costs for different DRMs.

	Storage Cost	Computation Cost
Gaussian	kI^N	kI^N
Sparse	μkI^N	μkI^N
SSRFT	I^N	$I^N \log(k)$
TRP	kNI	kI^N

Table 3.1: Performance of Different Dimension Reduction Maps: We compare the storage cost and the computational cost of applying a DRM mapping $\mathbb{R}^{I^N}$ to $\mathbb{R}^k$ to a dense tensor in $\mathbb{R}^{I^N}$. Here μ is the sparse factor for sparse random projection. The TRP considered here is composed of Gaussian DRMs.

We do not need to explicitly form or store the TRP map $\mathbf{\Omega}$. Instead, we can store its

constituent DRMs $\mathbf{A}_1, \ldots, \mathbf{A}_N$ and compute the action of the map on the matricized tensor using the definition of the TRP. The additional computation required is minimal and empirically incurs almost no performance loss.

3.4 Algorithms for Tucker approximation

In this section, we present our proposed tensor sketch and algorithms for one- and two-pass Tucker approximation, and discuss the computational complexity and storage cost of these methods for both sparse and dense input tensors. We present guarantees for these methods in 3.5.

3.4.1 Tensor compression via sketching

The Tucker sketch Our Tucker sketch generalizes the matrix sketch of Tropp et al. [2018] to higher order tensors. To compute a Tucker sketch for tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with sketch size parameters $\mathbf{k}$ and $\mathbf{s}$, draw independent, random DRMs

$$\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_N \quad \text{and} \quad \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_N, \tag{3.4.1}$$

with $\boldsymbol{\Omega}_n \in \mathbb{R}^{I_{(-n)} \times k_n}$ and $\boldsymbol{\Phi}_n \in \mathbb{R}^{I_n \times s_n}$ for $n \in [N]$. Use these DRMs to compute

$$\begin{aligned} \mathbf{V}_n &= \mathbf{X}^{(n)} \boldsymbol{\Omega}_n && \in \mathbb{R}^{I_n \times k_n}, \quad n \in [N], \\ \mathcal{H} &= \mathcal{X} \times_1 \boldsymbol{\Phi}_1^\top \cdots \times_N \boldsymbol{\Phi}_N^\top && \in \mathbb{R}^{s_1 \times \cdots \times s_N}. \end{aligned}$$

The *factor sketch* $\mathbf{V}_n$ captures the span of the mode-n fibers of $\mathcal{X}$ for each $n \in [N]$, while the *core sketch* $\mathcal{H}$ contains information about the interaction between different modes. See 3 for pseudocode.

To produce a rank $\mathbf{r} = \{r_1, \ldots, r_N\}$ Tucker approximation of $\mathcal{X}$, choose sketch size parameters $\mathbf{k} = (k_1, \ldots, k_N) \geq \mathbf{r}$ and $\mathbf{s} = (s_1, \ldots, s_N) \preceq \mathbf{k}$. (Vector inequalities hold elementwise.) Our approximation guarantees depend closely on the parameters $\mathbf{k}$ and $\mathbf{s}$. As a rule of thumb, we suggest selecting $\mathbf{s} = 2\mathbf{k} + 1$, as the theory requires $\mathbf{s} > 2\mathbf{k}$, and choosing $\mathbf{k}$ as large as possible given storage limitations.

The sketches $\mathbf{V}_n$ and $\mathcal{H}$ are linear functions of the original tensor, so we can compute the sketches in a single pass over the tensor $\mathcal{X}$. Linearity enables easy computation of the sketch even in the streaming model (7) or distributed model (8). Storing the sketches requires memory $\sum_{n=1}^{N} I_n \cdot k_n + \Pi_{i=1}^{N} s_n$: much less than the full tensor.

Algorithm 3 Tucker Sketch

Given: RDRM (a function that generates a random DRM)

1: **function** TUCKERSKETCH($\mathcal{X}; \mathbf{k}, \mathbf{s}$)
2: Form DRMs $\mathbf{\Omega}_n = \text{RDRM}(I_{(-n)}, k_n)$ and $\mathbf{\Phi}_n = \text{RDRM}(I_n, s_n)$, $n \in [N]$
3: Compute factor sketches $\mathbf{V}_n \leftarrow \mathbf{X}^{(n)}\mathbf{\Omega}_n$, $n \in [N]$
4: Compute core sketch $\mathcal{H} \leftarrow \mathcal{X} \times_1 \mathbf{\Phi}_1^\top \times \cdots \times_N \mathbf{\Phi}_N^\top$
5: **return** $(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N, \{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n \in [N]})$
6: **end function**

Remark. *The DRMs $\mathbf{\Omega}_n \in \mathbb{R}^{I_{(-n)} \times k_n}$ are large—much larger than the size of the Tucker factorization we seek! Even using a low memory mapping such as the SSRFT and sparse random map, the storage cost required grows as $\mathcal{O}(I_{(-n)})$. However, we do not need to store these matrices. Instead, we can generate (and regenerate) them as needed using a (stored) random seed.*[2]

Remark. *Alternatively, the TRP (3.3.2) can be used to limit the storage of $\mathbf{\Omega_n}$ required. The Khatri-Rao structure in the sketch need not match the structure in the matricized tensor. However, we can take advantage of the structure of our problem to reduce*

[2]Our theory assumes the DRMs are random, whereas our experiments use pseudorandom numbers. In fact, for many pseudorandom number generators it is NP hard to determine whether the output is random or pseudorandom Arora and Barak [2009]. In particular, we expect both to perform similarly for tensor approximation.

storage even further. We generate DRMs $\mathbf{A}_n \in \mathbb{R}^{I_n \times k}$ *for* $n \in [N]$ *and define* $\mathbf{\Omega}_n = \mathbf{A}_1 \odot \cdots \mathbf{A}_{n-1} \odot \mathbf{A}_{n+1} \odot \cdots \odot \mathbf{A}_N$ *for each* $n \in [N]$*. Hence we need not store the maps* $\mathbf{\Omega}_n$*, but only the small matrices* $\mathbf{A}_n$*. The storage required is thereby reduced from* $\mathcal{O}(N(\prod_{n=1}^{N} I_n)k)$ *to* $\mathcal{O}((\sum_{n=1}^{N} I_n)k)$*, while the approximation error is essentially unchanged. We use this method in our experiments.*

3.4.2 Low-Rank Approximation

Now we explain how to construct a Tucker decomposition of $\mathcal{X}$ with target Tucker rank $\mathbf{k}$ from the factor and core sketches.

We first present a simple two-pass algorithm, 4, that uses only the factor sketches by projecting the unfolded matrix of original tensor $\mathcal{X}$ to the column space of each factor sketch. To project to the column space of each factor matrix, we calculate the QR decomposition of each factor sketch:

$$\mathbf{V}_n = \mathbf{Q}_n \mathbf{R}_n \quad \text{for } n \in [N], \tag{3.4.2}$$

where $\mathbf{Q}_n \in \mathbb{R}^{I_n \times k_n}$ has orthonormal columns and $\mathbf{R}_n \in \mathbb{R}^{k_n \times k_n}$ is upper triangular. Consider the tensor approximation

$$\tilde{\mathcal{X}} = \mathcal{X} \times_1 \mathbf{Q}_1 \mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N \mathbf{Q}_N^\top. \tag{3.4.3}$$

This approximation admits the guarantees stated in 3.5.1. Using the commutativity of the mode product between different modes, we can rewrite $\tilde{\mathcal{X}}$ as

$$\tilde{\mathcal{X}} = \underbrace{\left[\mathcal{X} \times \mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N^\top\right]}_{\mathcal{W}_2} \times_1 \mathbf{Q}_1 \times_2 \cdots \times_N \mathbf{Q}_N = [\![\mathcal{W}_2; \mathbf{Q}_1, \ldots, \mathbf{Q}_N]\!], \tag{3.4.4}$$

which gives an explicit Tucker approximation $\tilde{\mathcal{X}}$ of our original tensor. The core approximation $\mathcal{W}_2 \in \mathbb{R}^{k_1 \times \cdots \times k_N}$ is much smaller than the original tensor $\mathcal{X}$. To compute

this approximation, we need access to $\mathcal{X}$ twice: once to compute $\mathbf{Q}_1, \ldots, \mathbf{Q}_N$, and again to apply them to $\mathcal{X}$ in order to form $\mathcal{W}_2$.

Algorithm 4 Two Pass Sketch and Low Rank Recovery

Given: tensor $\mathcal{X}$, sketch parameters $\mathbf{k}$ and $\mathbf{s} \geqslant \mathbf{k}$

1 *Sketch.* $\left(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N, \{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n\in[N]}\right) = \textsc{TuckerSketch}\left(\mathcal{X}; \mathbf{k}, \mathbf{s}\right)$

2 *Recover factor matrices.* For $n \in [N]$, $(\mathbf{Q}_n, \sim) \leftarrow \mathrm{QR}(\mathbf{V}_\mathrm{n})$

3 *Recover core.* $\mathcal{W}_2 \leftarrow \mathcal{X} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N$

Return: Tucker approximation $\hat{\mathcal{X}}_2 = [\![\mathcal{W}_2; \mathbf{Q}_1, \ldots, \mathbf{Q}_N]\!]$ with rank $\preceq \mathbf{k}$

One-Pass Approximation To develop a one-pass method, we must use the core sketch $\mathcal{H}$ — the compression of $\mathcal{X}$ using the random projections $\mathbf{\Phi}_n$ – to approximate $\mathcal{W}_2$ — the compression of $\mathcal{X}$ using random projections $\mathbf{Q}_n$. To develop intuition, consider the following calculation: if the factor matrix approximations $\mathbf{Q}_n$ capture the range of $\mathcal{X}$ well, then projection onto their ranges in each mode approximately preserves the action of $\mathcal{X}$:

$$\mathcal{X} \approx \mathcal{X} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \times \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top$$

Recall that for tensor $\mathcal{A}$, and matrix $\mathbf{B}$ and $\mathbf{C}$ with compatible sizes, $\mathcal{A} \times_n (\mathbf{BC}) = (\mathcal{A} \times_n \mathbf{C}) \times_n \mathbf{B}$. Use this rule to collect terms to recognize the two pass core approximation $\mathcal{W}_2$:

$$\mathcal{X} \approx \left(\mathcal{X} \times_1 \mathbf{Q}_1^\top \times \cdots \times_N \mathbf{Q}_N^\top\right) \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N = \mathcal{W}_2 \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N$$

Now contract both sides of this approximate equality with the DRMs $\mathbf{\Phi}_n$ and recognize the core sketch $\mathcal{H}$:

$$\mathcal{H} := \mathcal{X} \times_1 \mathbf{\Phi}_1^\top \cdots \times_N \mathbf{\Phi}_N^\top \approx \mathcal{W}_2 \times_1 \mathbf{\Phi}_1^\top\mathbf{Q}_1 \times \cdots \times_N \mathbf{\Phi}_N^\top\mathbf{Q}_N.$$

We have chosen $\mathbf{s} > \mathbf{k}$ so each $\mathbf{\Phi}_n^\top \mathbf{Q}_n$ has a left inverse with high probability. Hence we can solve the approximate equality for $\mathcal{W}_2$:

$$\mathcal{W}_2 \approx \mathcal{H} \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger =: \mathcal{W}_1.$$

The right hand side of the approximation defines the one pass core approximation $\mathcal{W}_1$. 3.B.2 controls the error in this approximation.

5 summarizes the resulting one-pass algorithm. One (streaming) pass over the tensor can be used to sketch the tensor; to recover the tensor, we only access the sketches. 3.5.2 (below) bounds the overall quality of the approximation.

Algorithm 5 One Pass Sketch and Low Rank Recovery

Given: tensor $\mathcal{X}$, sketch parameters $\mathbf{k}$ and $\mathbf{s} \geqslant \mathbf{k}$

1 *Sketch.* $\left(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N, \{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n\in[N]}\right) = \textsc{TuckerSketch}\left(\mathcal{X}; \mathbf{k}, \mathbf{s}\right)$

2 *Recover factor matrices.* For $n \in [N]$, $(\mathbf{Q}_n, \sim) \leftarrow \mathrm{QR}(\mathbf{V}_\mathrm{n})$

3 *Recover core.* $\mathcal{W}_1 \leftarrow \mathcal{H} \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger$

Return: Tucker approximation $\hat{\mathcal{X}}_1 = [\![\mathcal{W}_1; \mathbf{Q}_1, \ldots, \mathbf{Q}_N]\!]$ with rank $\leqslant \mathbf{k}$

The time and storage cost of Algorithm 5 is given by Table 3.2. The time and storage complexity of these methods compare favorably to the only previous method for streaming Tucker approximation Malik and Becker [2018],

	Stage	Time Cost	Storage Cost
5 (One Pass)	Sketching	$\mathcal{O}(((1-(s/I)^N)/(1-(s/I)) + Nk)I^N)$	$kNI + s^N$
	Recovery	$\mathcal{O}((k^2 s^N (1-(k/s)^N))/(1-k/s) + k^2 NI)$	
	Total	$\mathcal{O}(((s(1-(s/I)^N))/(1-s/I) + Nk)I^N)$	

Table 3.2: Computational Complexity of 5 on tensor $\mathcal{X} \in \mathbb{R}^{I\times\cdots\times I}$ with parameters (k, s), using a TRP composed of Gaussian DRMs inside the Tucker sketch. By far the majority of the time is spent sketching the tensor $\mathcal{X}$.

3.4.3 Fixed-Rank Approximation

Algorithm 4 and algorithm 5 produce a two-pass and one-pass rank-$\mathbf{k}$ tensor approximation respectively. It is often valuable to truncate this approximation to a user-specified target rank $\mathbf{r} \leqslant \mathbf{k}$ [Tropp et al., 2019b, Figure 4].

Our fixed rank approximation method is motivated by the following lemma:

Lemma 3.4.1. *Let $\mathcal{W} \in \mathbb{R}^{k_1 \times \cdots \times k_N}$ be a tensor, and let $\mathbf{Q}_n \in \mathbb{R}^{I_n \times k_n}$ be orthogonal matrices with $k_n \geqslant r_n$ for $n \in [N]$. Then*

$$[\![\mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N]\!]_{\mathbf{r}} = [\![\mathcal{W}]\!]_{\mathbf{r}} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N.$$

(This lemma does not necessarily hold if the best rank-r Tucker approximation $[\![\cdot]\!]$ is replaced by the output of any concrete algorithm such as HOSVD or HOOI.) The proof of 3.4.1 appears in 3.C.

Motivated by this lemma, to produce a fixed rank $\mathbf{r}$ approximation of $\mathcal{X}$, we compress the core tensor approximation from 4 or 5 to rank $\mathbf{r}$. This compression is cheap because the core approximation $\mathcal{W} \in \mathbb{R}^{k_1 \times \cdots \times k_N}$ is small. We present this method (using HOOI as the the compression algorithm) as 6. Other compression algorithms can be used to trade off the quality of approximation with the difficulty of running the algorithm. Reasonable choices include the sequentially-truncated HOSVD (ST-HOSVD) Vannieuwenhoven et al. [2012] or TTHRESH Ballester-Ripoll et al. [2019]. Both HOSVD and ST-HOSVD are psedual

Algorithm 6 Fixed rank approximation

Given: Tucker approximation $[\![\mathcal{W}; \mathbf{Q}_1, \cdots, \mathbf{Q}_N]\!]$ of tensor $\mathcal{X}$, rank target $\mathbf{r}$

1 *Approximate core with fixed rank.* $\mathcal{G}, \mathbf{U}_1, \cdots, \mathbf{U}_N \leftarrow \text{HOOI}(\mathcal{W}, \mathbf{r})$

2 *Compute factor matrices.* For $n \in [N]$, $\mathbf{P}_n \leftarrow \mathbf{Q}_n \mathbf{U}_n$

Return: Tucker approximation $\hat{\mathcal{X}}_{\mathbf{r}} = [\![\mathcal{G}; \mathbf{P}_1, \ldots, \mathbf{P}_N]\!]$ with rank $\leqslant \mathbf{r}$

3.5 Guarantees

In this section, we present probabilistic guarantees on the preceding algorithms. We show that approximation error for the one-pass algorithm is the sum of the error from the two-pass algorithm and the error resulting from the core approximation. Proofs for the three theorems in this section can be found in the corresponding subsections of 3.A.

3.5.1 Low rank approximation

Theorem 3.5.1 guarantees the performance of the two pass method algorithm 4.

Theorem 3.5.1. *Sketch the tensor $\mathcal{X}$ using a Tucker sketch with parameters $\mathbf{k}$ using DRMs with i.i.d. Gaussian $\mathcal{N}(0,1)$ entries. Then the approximation $\hat{\mathcal{X}}_2$ computed with the two pass method 4 satisfies*

$$\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2 \leqslant \min_{1 \leqslant \rho_n < k_n - 1} \sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1}\right) (\tau_{\rho_n}^{(n)})^2.$$

The two pass method does not use the core sketch, so this result does not depend on $\mathbf{s}$.

Theorem 3.5.2 guarantees the performance of one pass method 5.

Theorem 3.5.2. *Sketch the tensor $\mathcal{X}$ using a Tucker sketch with parameters $\mathbf{k}$ and $\mathbf{s} > 2\mathbf{k}$ using DRMs with i.i.d. Gaussian $\mathcal{N}(0,1)$ entries. Then the approximation $\hat{\mathcal{X}}_1$ computed with the one pass method 5 satisfies the bound*

$$\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}_1\|_F^2 \leqslant (1+\Delta) \min_{1 \leqslant \rho_n < k_n - 1} \sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1}\right) (\tau_{\rho_n}^{(n)})^2,$$

where $\Delta := \max_{n=1}^{N} k_n/(s_n - k_n - 1)$.

The theorem shows that the method works best for tensors whose unfoldings exhibit spectral decay. As a simple consequence of this result, we see that the two pass method with $\mathbf{k} > \mathbf{r} + 1$ perfectly recovers a tensor with exact Tucker rank $\mathbf{r}$, since in that case $\tau_{r_n}^{(n)} = 0$ for each $n \in [N]$. However, this theorem states a stronger bound: the method exploits decay in the spectrum, wherever (in the first k_n singular values of each mode n unfolding) it occurs.

We see that the additional error due to sketching the core is a multiplicative factor Δ more than the error due to sketching the factor matrices. This factor Δ decreases as the size of the core sketch $\mathbf{s}$ increases.

Remark. *Both HOSVD, ST-HOSVD achieves so called pseudo optimal with parameter $\sqrt{N}$. For example, for ST-HOSVD,*

$$\|[\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{ST-k}}\|_F \leqslant \sqrt{N}\|[\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{k}}\|_F \leqslant \sqrt{\sum_{n=1}^{N} (\tau_{k_n}^{(n)})^2}, \tag{3.5.1}$$

where $[\![\mathcal{X}]\!]_{\mathbf{k}}$ is the optimal Tucker rank $\mathbf{k}$ approximation. Thus, the low rank approximation for two pass algorithm is pseduo optimal with factor $\sqrt{2N}$ while for one pass algorithm, if we choose $\mathbf{s} > 2\mathbf{k} + 1 (\Delta \leqslant 2)$, it is also pseduo optimal with factor $2\sqrt{N}$.

Theorem 3.5.2 also offers guidance on how to select the sketch size parameters $\mathbf{s}$ and $\mathbf{k}$. In particular, suppose that the mode-n unfolding has a good rank r_n approxi-

mation for each mode n. Then the choices $k_n = 2r_n + 1$ and $s_n = 2k_n + 1$ ensure that

$$\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 \leqslant 4 \sum_{n=1}^{N} (\tau_{r_n}^{(n)})^2.$$

More generally, as k_n/r_n and s_n/k_n increase, the leading constant in the approximation error tends to one.

3.5.2 Fixed rank approximation

We now present a conditional analysis of the fixed rank approximation method given a low rank approximation. Recall that $[\![\cdot]\!]_{\mathbf{r}}$ returns a best rank-$\mathbf{r}$ Tucker approximation.

Theorem 3.5.3. *Suppose* $\hat{\mathcal{X}} = [\![\mathcal{W}; \mathbf{Q}_1, \cdots, \mathbf{Q}_N]\!]$ *approximates the target tensor* $\mathcal{X}$*, and let* $\hat{\mathcal{X}}_{\mathbf{r}}$ *denote some rank* $\mathbf{r}$ *approximation to* $\hat{\mathcal{X}}$ *from some procedure like ST-HOSVD, Then for output any fix rank* $\mathbf{r}$

$$\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}_{\mathbf{r}}\|_F \leqslant \|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F + 2\sqrt{\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2}.$$

The second term on the right-hand side of 3.5.3 is controlled by 3.5.1 and 3.5.2. Hence we can combine these results to provide guarantees for fixed rank approximation with either the two pass or one pass algorithms.

The resulting bound shows that the best rank-$\mathbf{r}$ approximation of the output from the one or two pass algorithms is comparable in quality to a true best rank-$\mathbf{r}$ approximation of the input tensor. An important insight is that the sketch size parameters $\mathbf{s}$ and $\mathbf{k}$ that guarantee a good low rank approximation also guarantee a good fixed rank approximation: the error due to sketching depends only on the sketch size parameters $\mathbf{k}$ and $\mathbf{s}$, and not on the target rank $\mathbf{r}$.

In practice, one would truncate the rank of the approximation using HOOI (6), rather than the best rank $\mathbf{r}$ approximation. Guarantees for resulting algorithm are beyond the scope of this paper, since there are no strong guarantees on the performance of HOOI; however, it is widely believed to produce an approximation that is usually quite close to the best rank $\mathbf{r}$ approximation.

3.5.3 Proof sketch

To bound the approximation error of the algorithms presented in the main body of this paper, we first develop several structural results showing an additive decomposition of the error. First, the total error is the sum of the error due to sketching and the error due to fixed rank approximation. Second, the sketching error is the sum of the error due to the factor matrix approximations and to the core approximation. Third, the error due to the factor matrix approximations is the sum of the error in each the modes, as the errors due to each mode are mutually orthogonal. This finishes the approximation error bound for the two pass algorithm, 3.5.1. As for the error due to the core approximation, we rewrite the approximation error in the core tensor as a sum over each mode of errors that are mutually orthogonal. Indeed, these errors have the same form as the errors due to the factor matrix approximations, scaled down by a factor $\Delta(k, s)$ that depends on the sketch sizes $\mathbf{k}$ and $\mathbf{s}$. This argument shows the error due to the core approximation is at most a factor $\Delta(k, s)$ times the error due to the factor matrix approximation.

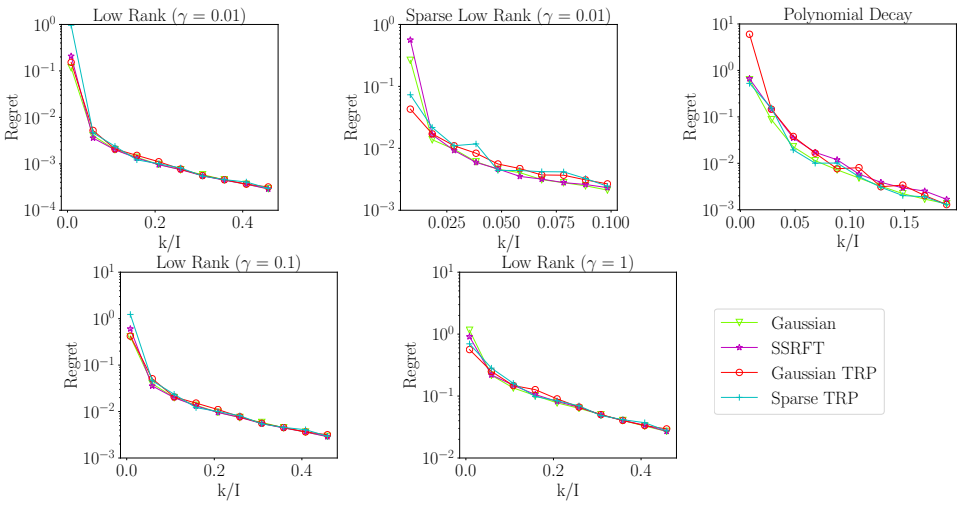


Figure 3.1: *Different DRMs perform similarly.* We approximate 3D synthetic tensors (see 3.6.1) with $I = 600$, using our one-pass algorithm with $r = 5$ and varying k ($s = 2k+1$), using a variety of DRMs in the Tucker sketch: Gaussian, SSRFT, Gaussian TRP, or Sparse TRP.

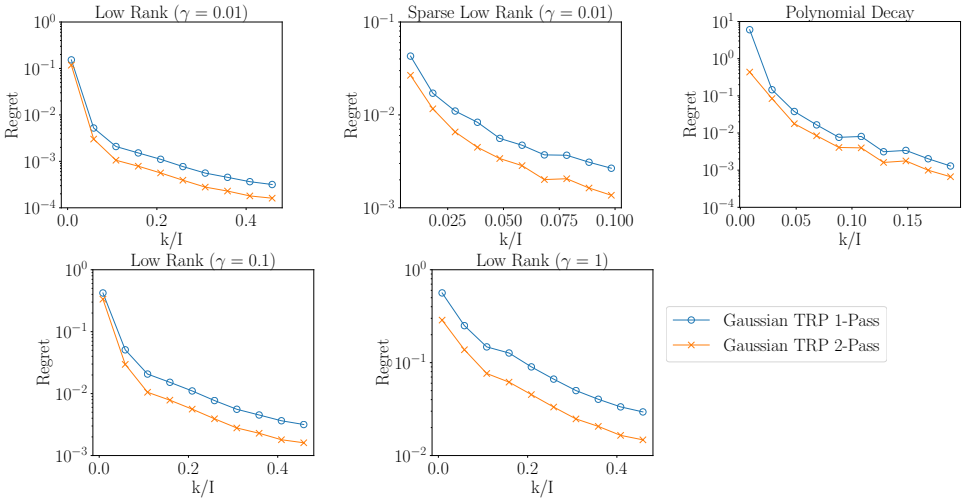


Figure 3.2: *Two-pass improves on one-pass.* We approximate 3D synthetic tensors (see 3.6.1) with $I = 600$, using our one-pass and two-pass algorithms with $r = 5$ and varying k ($s = 2k + 1$), using the Gaussian TRP in the Tucker sketch.

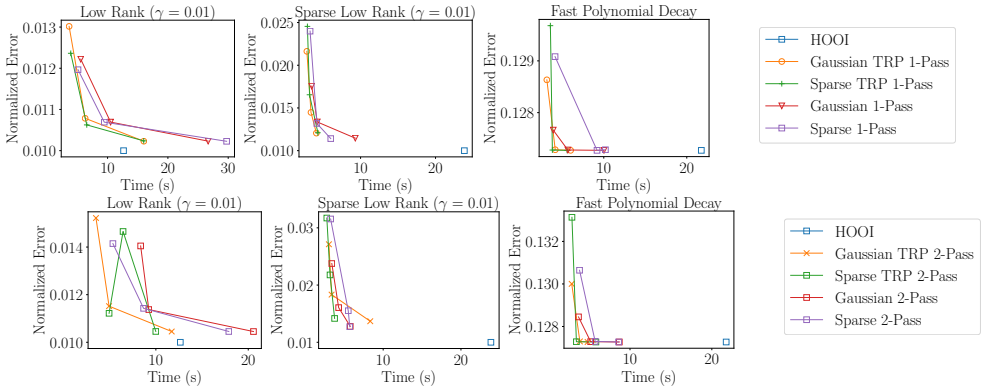


Figure 3.3: *Faster approximations.* We approximate 3D synthetic tensors with $I = 600$ generated as described in 3.6.1, using HOOI and our one-pass and two-pass algorithms with $r = 5$ for a few different k ($s = 2k + 1$).

3.6 Numerical Experiments

In this section, we study the performance of our method. We compare the performance of the method using various different DRMs, including TRP. We also compare our method with the algorithm proposed by Malik and Becker [2018] to show that for the same storage budget, our method produces better approximations. Our two-pass algorithm outperforms the one-pass version, as expected. (Contrast this to Malik and Becker [2018], where the multi-pass method performs less well than the one-pass version.)

We evaluate the experimental results using two metrics:

$$\text{normalized error:} \qquad \|\mathcal{X} - \hat{\mathcal{X}}\|_F / \|\mathcal{X}\|_F$$

$$\text{regret:} \qquad \left(\|\mathcal{X} - \hat{\mathcal{X}}\|_F - \|\mathcal{X} - \mathcal{X}_{\text{HOOI}}\|_F\right) / \|\mathcal{X}\|_F.$$

The normalized error measures the fraction of the energy in $\mathcal{X}$ captured by the approximation. The regret measures the increase in normalized error due to using the approximation $\hat{\mathcal{X}}$ rather than using $\mathcal{X}_{\text{HOOI}}$. The relative error measures the decrease in performance relative to HOOI. The normalized error of a rank $\mathbf{r}$ Tucker approximation $\hat{X}$ is always positive when $\mathcal{X}$ has a larger rank. In general, we find our proposed

methods approaches the performance of HOOI for large enough storage budgets.

We ran all experiments on a server with 128 Intel® Xeon® E7-4850 v4 2.10GHz CPU cores and 1056GB memory. The code for our method is available at an anonymous Github repository `https://github.com/tensorsketch/tensorsketch`.

3.6.1 Synthetic experiments

All synthetic experiments use an input tensor with equal side lengths I. We consider three different data generation schemes:

- *Low rank + noise.* Generate a core tensor $\mathcal{C} \in \mathbb{R}^{r^N}$ with entries drawn from $\mathrm{Unif}([0,1])$. Independently generate N orthogonal factor matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N \in \mathbb{R}^{r \times I}$. Define $\mathcal{X}^\natural = \mathcal{C} \times_1 \mathbf{A}_1 \cdots \times_N \mathbf{A}_N$ and the noise parameter $\gamma > 0$. Generate an input tensor as $\mathcal{X} = \mathcal{X}^\natural + (\gamma \|\mathcal{X}^\natural\|_F / I^{N/2})\boldsymbol{\epsilon}$ where the noise $\boldsymbol{\epsilon}$ has i.i.d. $\mathcal{N}(0,1)$ entries.
- *Sparse low rank + noise.* We construct the input tensor $\mathcal{X}$ as above (Low Rank + Noise), but with sparse factor matrices $\mathbf{A}_n$: If δ_n is the sparsity (proportion of non-zero elements) of $\mathbf{A}_n$, then the sparsity of the true signal $\mathcal{X}^\natural$ is $\prod_{n=1}^N \delta_n$. We use $\delta_n = 0.2$ unless otherwise specified.
- *Polynomial decay.* We construct the input tensor $\mathcal{X}$ as

$$\mathcal{X} = \textbf{superdiag}(1, \ldots, 1, 2^{-t}, 3^{-t}, \ldots, (I-r)^{-t}).$$

  The first r entries are 1. Recall **superdiag** converts a vector to N dimensional superdiagonal tensor. Our experiments use $t = 1$ (geometric decay).

Different dimension reduction maps perform similarly

Our first experiment investigates the performance of our one-pass fixed-rank algorithm as the sketch size (and hence, required storage) varies, for several types of dimension reductions maps, including Gaussian, SSRFT, Gaussian TRP, and Sparse TRP. We generate synthetic data as described above with $\mathbf{r} = (5, 5, 5), I = 600$. 3.1 shows the rank-$\mathbf{r}$ approximation error as a function of the compression factor k/I. (Results for other input tensors are presented as 3.9 and 3.10 in 4.B.) We see that the log relative error for our one-pass algorithm converges to that of HOOI as k increases for all input tensors. In the low rank case, the convergence rate is lower for higher noise levels. In general, the performance for different maps are approximately the same, although our theory only pertains to the Gaussian map.

We evaluate the run time for HOOI and our two algorithms with several different DRMs in 3.3. We can see that the one-pass algorithm is always slightly faster than the two-pass algorithm. The TRP generally provides a modest speedup in addition to the memory advantage. Both our one-pass and two-pass algorithms achieve nearly the accuracy of HOOI, and are usually much faster.

A second pass reduces error

The second experiment compares our two-pass and one-pass algorithm. The design is similar to the first experiment. 3.2 shows that the two-pass algorithm typically outperforms the one-pass algorithm, especially in the high-noise, sparse, or rank-decay case. Both converge at the same asymptotic rate. (Results for other input tensors are available in the supplement.)

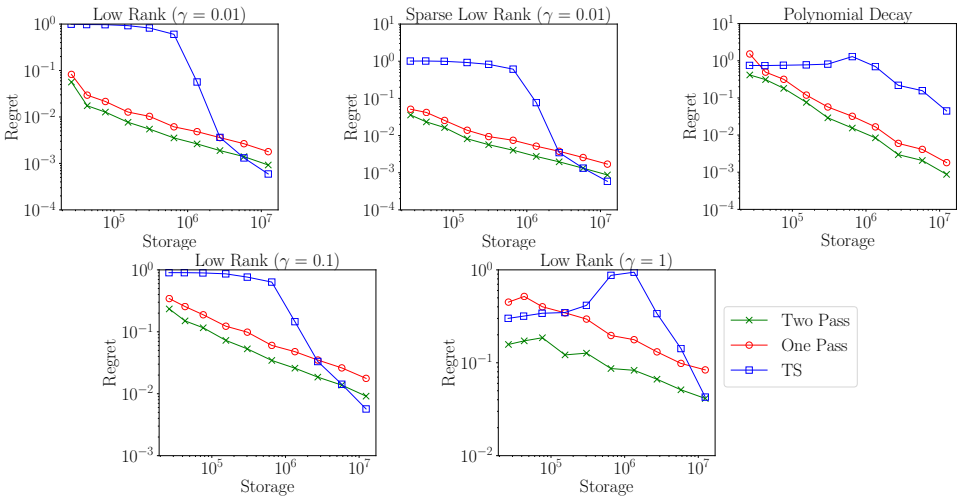


Figure 3.4: *Approximations improve with more memory: synthetic data.* We approximate 3D synthetic tensors (see 3.6.1) with $I = 300$, using T.-TS and our one-pass and two-pass algorithms with the Gaussian TRP to produce approximations with equal ranks $r = 10$. Notice every marker on the plot corresponds to a 2700× compression!

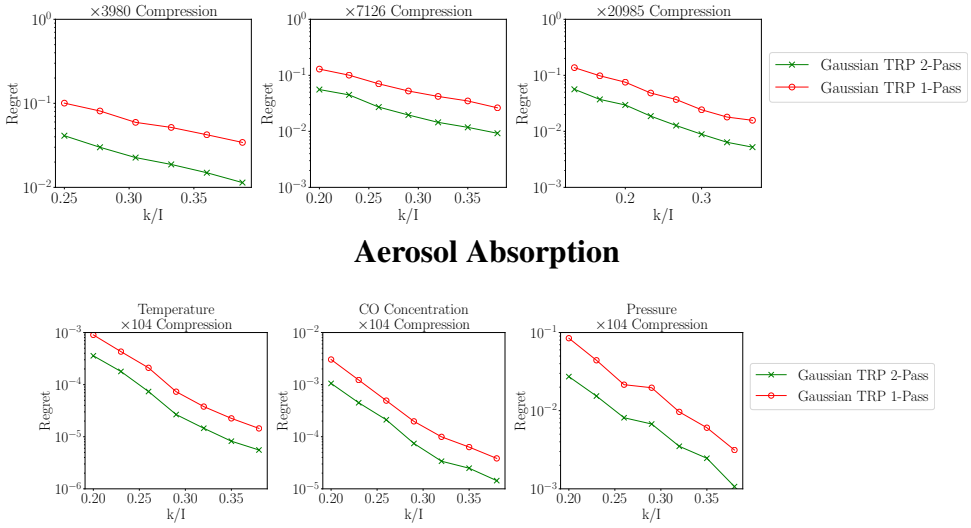
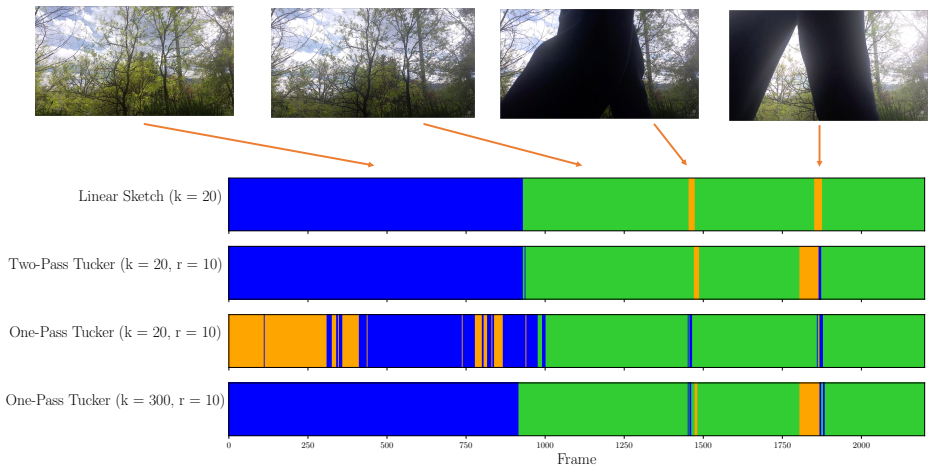


Figure 3.5: *Approximations improves with more memory: real data.* We approximate aerosol absorption and combustion data using our one-pass and two-pass algorithms with the Gaussian TRP. We compare three target ranks ($r/I = 0.125, 0.1, 0.067$) for the former, and use the same target rank ($r/I = 0.1$) for each measured quantity in the combustion dataset. Notice $r/I = 0.1$ gives a hundred-fold compression!



Video Scene Classification

Figure 3.6: *Video Scene Classification* ($2200 \times 1080 \times 1980$): We classify frames from the video data from Malik and Becker [2018] (collected as a third order tensor with size $2200 \times 1080 \times 1980$) using K-means with K=3 on vectors computed using four different methods. $s = 2k + 1$ throughout. 1) The linear sketch along the time dimension (Row 1). 2-3) the Tucker factor along the time dimension, computed via our two-pass (Row 2) and one-pass (Row 3) algorithms. 4) The Tucker factor along the time dimension, computed via our one-pass (Row 4) algorithm

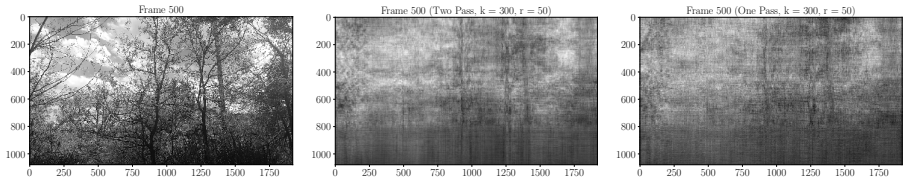


Figure 3.7: *Visualizing Video Recovery:* Original frame (left); approximation by two-pass sketch (middle); approximation by one-pass sketch (right).

Improvement on state-of-the-art

The third experiment compares the performance of our two-pass and one-pass algorithms and Tucker TensorSketch (T.–TS), as described in Malik and Becker [2018], the only extant one-pass algorithm. For a fair comparison, we allocate the same storage budget to each algorithm and compare the relative error of the resulting fixed-rank ap-

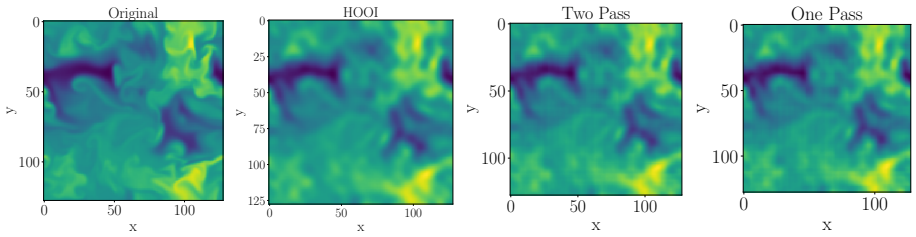


Figure 3.8: *Visualizing Combustion Simulation:* All four figures show a slice of the temperature data along the first dimension. The approximation uses $\mathbf{r} = (281, 25, 25)$, $\mathbf{k} = (562, 50, 50)$, $\mathbf{s} = (1125, 101, 101)$, with the Gaussian TRP in the Tucker sketch.

proximations. We approximate synthetic 3D tensors with side length $I = 300$ with Tucker rank $r = 10$. We use the suggested parameter settings for each algorithm: $k = 2r$ and $s = 2k + 1$ for our methods; $K = 10$ for T.–TS. Our one-pass algorithm (with the Gaussian TRP) uses $((2k + 1)^N + kIN)$ storage, whereas T.-TS uses $(Kr^{2N} + Kr^{2N-2})$ storage (see supplement).

Figure 3.4 shows that our algorithms generally perform as well as T.–TS, and dramatically outperforms for small storage budgets. For example, our method achieves 1/50, 1/50, 1/7, and 1/4 the relative error of T.–TS for low rank and sparse low rank ($\gamma = 0.01$), low rank ($\gamma = 0.1$), and polynomial-decay input tensors, respectively. For the low rank ($\gamma = 1$) tensor, the performance of T.–TS is not even monotone as the storage budget increases! The performance of T.–TS is comparable with that of the algorithms presented in this paper only when the storage budget is large.

Remark. *The paper Malik and Becker [2018] proposes a multi-pass method, Tucker Tensor-Times-Matrix-TensorSketch (TTMTS) that is dominated by the one-pass method Tucker TensorSketch(TS) in all numerical experiments; hence we compare only with T.-TS.*

3.6.2 Applications

We also apply our method to datasets drawn from three application domains: climate, combustion, and video.

- *Climate data.* We consider global climate simulation datasets from the Community Earth System Model (CESM) Community Atmosphere Model (CAM) 5.0 Hurrell et al. [2013]; Kay et al. [2015]. The dataset on aerosol absorption has four dimensions: times, altitudes, longitudes, and latitudes ($240 \times 30 \times 192 \times 288$). The data on net radiative flux at surface and dust aerosol burden have three dimensions: times, longitudes, and latitudes ($1200 \times 192 \times 288$). Each of these quantitives has a strong impact on the absorption of solar radiation and on cloud formation.
- *Combustion data.* We consider combustion simulation data from Lapointe et al. [2015]. The data consists of three measured quantities — pressure, CO concentration, and temperature — each observed on a $1408 \times 128 \times 128$ spatial grid.
- *Video data.* We use our streaming method to cluster frames of a video, as in Malik and Becker [2018]. Here, a low frame rate camera is mounted in a fixed position as people walk by. A 3D tensor is constructed with each video frames as a slice. The video consists of 2493 frames, each of size 1080 by 1980. As a tensor, stored as a `numpy.array`, the video data is 41.4 GB in total.

Data compression

We show that our proposed algorithms are able to successfully compress climate and combustion data even when the full data does not fit in memory. Since the Tucker rank

of the original tensor is unknown, we perform experiments for three different target ranks. In this experiment, we hope to understand the effect of different choices of storage budget k to achieve the same compression ratio. We define the compression ratio as the ratio in size between the original input tensor and the output Tucker factors, i.e. $\frac{\prod_{i=1}^{N} I_i}{\sum_{i=1}^{N} r_i I_i + \prod_{i=1}^{N} r_i}$. As in our experiments on simulated data, 3.5 shows that the two-pass algorithm outperforms the one-pass algorithm as expected. However, as the storage budget k increases, both methods converge to the performance of HOOI. The rate of convergence is faster for smaller target ranks. Performance of our algorithms on the combustion simulation is qualitatively similar, but converges faster to the performance of HOOI. 3.8 visualizes the recovery of the temperature data in combustion simulation for a slice along the first dimension. We could observe that the recovery for both two-pass and one-pass algorithm approximate the recovery from HOOI. 3.13 in 3.H shows similar results on another dataset.

Video scene classification

We show how to use our single pass method to classify scenes in the video data described above. The goal is to identify frames in which people appear. We remove the first 100 frames and last 193 frames where the camera setup happened, as in Malik and Becker [2018]. We stream over the tensor and sketch it using parameters $k = 300, s = 601$. Finally, we compute a fixed-rank approximation with $\mathbf{r} = (10, 10, 10)$ and $(20, 20, 20)$. We apply K-means clustering to the resulting 10 or 20 dimensional vectors corresponding to each of the remaining 2200 frames.

We experimented with clustering vectors found in three ways: from the two-pass or one-pass Tucker approximation, or directly from the factor sketch.

When matching the video frames with the classification result, we can see that the background light is relatively dark at the beginning, thus classified into `Class` 0. After a change in the backgroun light, most other frames of the video are classified into `Class` 1. When a person passes by the camera, the frames are classified into `Class` 2. Right after the person passed by, the frames are classified into `Class` 0, the brighter background scene, due to the light adjustment.

Our classification results (using the linear sketch or approximation) are similar to those in Malik and Becker [2018] while using only $1/500$ as much storage; the one pass approximation requires more storage (but still less than Malik and Becker [2018]) to achieve similar performance. In particular, using the sketch itself, rather than the Tucker approximation, to summarize the data enables very efficient video scene classification.

On the other hand, to reconstruct the original video frames we require much larger $\mathbf{k}$ and $\mathbf{r}$: the video is not very low rank along the spatial dimensions. 3.7 shows that even with $\mathbf{s} = 601, 601, 601, \mathbf{k} = (300, 300, 300), \mathbf{r} = (50, 50, 50)$, the recovered frame is very noisy.

3.7 Conclusion

This paper proposes a practical one-pass algorithm to compute the Tucker decomposition of a tensor with provable guarantees. Our algorithm uses a dimension reduction map to summarize the data in a linear sketch. This sketch can be efficiently stored or transmitted, which enables applications in modern large-scale setting, including streaming and distributed data storage, or computational environment with limited memory. We give the first comprehensive error analysis for one-pass Tucker decomposition algo-

rithm in 3.5.2 and 3.5.3 for the tensor approximation with error growing only linearly with the order N of the tensor. In practice, our algorithm significantly outperforms the current state-of-the-art one-pass Tucker decomposition algorithm Malik and Becker [2018] in the limited memory setting and noisy setting with much greater stability. Also, our one-pass achieves the same performance in Malik and Becker [2018]'s suggested setting when the memory requirement is much higher than our suggested setting. Our algorithm is available as an open-source python package.

3.A Proof of Main Results

3.A.1 Error bound for the two pass approximation Algorithm 4

Proof of Theorem 3.5.1. Suppose $\hat{\mathcal{X}}_2$ is the low-rank approximation from 4. Use the definition of the mode-n product to see

$$\begin{aligned}\hat{\mathcal{X}}_2 &= \left[\mathcal{X} \times_1 \mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N^\top\right] \times_1 \mathbf{Q}_1 \times_1 \cdots \times_N \mathbf{Q}_N \\ &= \mathcal{X} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top.\end{aligned}$$

Although it seems that we sequentially project tensor $\mathcal{X}$ to column space spanned with $\mathbf{Q}_n$, but since mode product is exchangeable, in fact $\hat{\mathcal{X}}_2$ is the projection to space $\{\mathcal{X} : \mathcal{X}^{(n)} \in \mathbf{col}(\mathbf{Q}_n)\}$. This is a generalization of projection matrix where **?** has a very detailed explanation and it is referred as multi-linear orthogonal projection. Following exact techniques in Theorem 5.1 in Vannieuwenhoven et al. [2012] by sequentially applying Pythagorean theory sequentially we can show that

$$\|\hat{\mathcal{X}}_2 - \mathcal{X}\|_F^2 \leqslant \sum_{n=1}^{N} \left\|(\mathbf{I} - \mathbf{Q}_n\mathbf{Q}_n^\top)\mathbf{X}^{(n)}\right\|_F^2. \tag{3.A.1}$$

Then taking expectation on $\mathbf{Q}_n$, and applying Lemma 3.D.2 we complete the proof.

□

3.A.2 Error bound for the one pass approximation Algorithm 5

Proof of Theorem 3.5.2. We show the approximation error can be decomposed as the error due to the factor matrix approximations and the error due to the core approximation. Let $\hat{\mathcal{X}}_1$ be the one pass approximation from 5, and let

$$\hat{\mathcal{X}}_2 = \mathcal{X} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top, \tag{3.A.2}$$

be the two pass approximation from 4. The difference in one-pass and two-pass approximation is in the core:

$$\hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2 = (\mathcal{W} - \mathcal{X} \times_1 \mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_n^\top) \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N.$$

Thus $\hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2$ is in the space defined above: $\{\mathcal{X} : \mathcal{X}^{(n)} \in \mathbf{col}(\mathbf{Q}_n)\}$ while as pointed before $\hat{\mathcal{X}}_2 - \mathcal{X}$ is perpendicular to that space. Therefore,

$$\langle \hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2, \hat{\mathcal{X}}_2 - \mathcal{X} \rangle = 0. \tag{3.A.3}$$

Now we use the (expectation of) the Pythagorean theorem to bound the expected error of the one pass approximation:

$$\mathbb{E}\|\hat{\mathcal{X}}_1 - \mathcal{X}\|_F^2 = \mathbb{E}\|\hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2\|_F^2 + \mathbb{E}\|\hat{\mathcal{X}}_2 - \mathcal{X}\|_F^2. \tag{3.A.4}$$

Consider the first term which is due to core approximation. Based in the definition of $\hat{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$ we can see that

$$\begin{aligned}\|\hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2\|_F^2 &= \|(\mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top) \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N\|_F^2 \\ &= \|(\mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top)\|_F^2,\end{aligned}$$

where we use the invariance of the Frobenius norm under orthonormal transformations to get the second line. Now using 3.B.2 to bound for the error due to the core approximation as

$$\mathbb{E}\|\hat{\mathcal{X}}_1 - \hat{\mathcal{X}}_2\|_F^2 \leqslant \Delta \left[\sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1} \right) (\tau_{\rho_n}^{(n)})^2 \right].$$

Finally, as shown in proof for 3.5.1 to bound the error due to the factor matrix approximations (the second term in (3.A.4)) as

$$\mathbb{E}\|\hat{\mathcal{X}}_2 - \mathcal{X}\|_F^2 \leqslant \left[\sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1} \right) (\tau_{\rho_n}^{(n)})^2 \right].$$

Summing these two bounds finishes the proof. □

3.A.3 Error bound for the fixed rank approximation Algorithm 6

Proof of Theorem 3.5.3. Our argument follows the proof of [Tropp et al., 2017, Proposition 6.1]:

$$\begin{aligned}\|\mathcal{X} - [\![\hat{\mathcal{X}}]\!]_{\mathbf{r}}\|_F &\leqslant \|\mathcal{X} - \hat{\mathcal{X}}\|_F + \|\hat{\mathcal{X}} - [\![\hat{\mathcal{X}}]\!]_{\mathbf{r}}\|_F \\ &\leqslant \|\mathcal{X} - \hat{\mathcal{X}}\|_F + \|\hat{\mathcal{X}} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F \\ &\leqslant \|\mathcal{X} - \hat{\mathcal{X}}\|_F + \|\hat{\mathcal{X}} - \mathcal{X} + \mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F \\ &\leqslant 2\|\mathcal{X} - \hat{\mathcal{X}}\|_F + \|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F.\end{aligned}$$

The first and the third line are the triangle inequality, and the second line follows from the definition of the best rank-r approximation. Take the expectation of $\|\mathcal{X} - \hat{\mathcal{X}}\|_F$ and use Jensen's inequality $\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}\|_F \leqslant \sqrt{\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}\|_F^2}$ to finish the proof. □

3.B Probabilistic Analysis of Core Sketch Error

This section contains the most technical part of our proof. We provide a probabilistic error bound for the difference between the two pass core approximation $\mathcal{W}_2$ from Algorithm 4 and the one pass core approximation $\mathcal{W}_1$ from Algorithm 5.

Introduce for each $n \in [N]$ the orthonormal matrix $\mathbf{Q}_n^{\perp}$ that forms a basis for the subspace orthogonal to $\mathbf{Q}_n$, so that $\mathbf{Q}_n^{\perp}(\mathbf{Q}_n^{\perp})^{\top} = \mathbf{I} - \mathbf{Q}_n\mathbf{Q}_n^{\top}$. Next, define

$$\boldsymbol{\Phi}_n^{Q} = \boldsymbol{\Phi}_n^{\top}\mathbf{Q}_n, \quad \boldsymbol{\Phi}_n^{Q^{\perp}} = \boldsymbol{\Phi}_n^{\top}\mathbf{Q}_n^{\perp}. \tag{3.B.1}$$

Recall that the DRMs $\boldsymbol{\Phi}_n$ are i.i.d. Gaussian. Hence conditional on $\mathbf{Q}_n$, $\boldsymbol{\Phi}_n^{Q}$ and $\boldsymbol{\Phi}_n^{Q^{\perp}}$ are independent.

3.B.1 Decomposition of Core Approximation Error

In this section, we characterize the difference between the one and two pass core approximations $\mathcal{W}_1 - \mathcal{W}_2 = \mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top$.

Lemma 3.B.1. *Suppose that $\mathbf{\Phi}_n$ has full column rank for each $n \in [N]$. Then*

$$\mathcal{W}_1 - \mathcal{W}_2 = \mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top = \sum_{(i_1,\ldots,i_N)\in\{0,1\}^N, \sum_{j=1}^N i_j \geqslant 1} \mathcal{Y}_{i_1 \ldots i_N},$$

where

$$\begin{aligned} \mathcal{Y}_{i_1\ldots i_N} = \mathcal{X} \times_1 &\left(\mathbf{1}_{i_1=0}\mathbf{Q}_1^\top + \mathbf{1}_{i_1=1}(\mathbf{\Phi}_1^{Q_1})^\dagger \mathbf{\Phi}_1^{Q_1^\perp}(\mathbf{Q}_1^\perp)^\top\right) \\ &\times_2 \cdots \times_N \left(\mathbf{1}_{i_N=0}\mathbf{Q}_N^\top + \mathbf{1}_{i_1=1}(\mathbf{\Phi}_N^{Q_N})^\dagger \mathbf{\Phi}_N^{Q_N^\perp}(\mathbf{Q}_N^\perp)^\top\right). \end{aligned} \tag{3.B.2}$$

Proof. Let $\mathcal{H}$ be the core sketch from 3. Write $\mathcal{W}_1$ as

$$\begin{aligned} \mathcal{W}_1 &= \mathcal{H} \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \\ &= (\mathcal{X} - \hat{\mathcal{X}}_2) \times_1 \mathbf{\Phi}_1^\top \times_2 \cdots \times_N \mathbf{\Phi}_N^\top \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \\ &+ \hat{\mathcal{X}}_2 \times_1 \mathbf{\Phi}_1^\top \times_2 \cdots \times_N \mathbf{\Phi}_N^\top \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger. \end{aligned}$$

Using the fact that $(\mathbf{\Phi}_n^\top \mathbf{Q}_n)^\dagger (\mathbf{\Phi}_n^\top \mathbf{Q}_n) = \mathbf{I}$, we can simplify the second term as

$$\begin{aligned} &\tilde{\mathcal{X}} \times_1 \mathbf{\Phi}_1^\top \times_2 \cdots \times_N \mathbf{\Phi}_N^\top \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \\ &= \mathcal{X} \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \mathbf{\Phi}_1^\top \mathbf{Q}_1 \mathbf{Q}_1^\top \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \mathbf{\Phi}_N^\top \mathbf{Q}_N \mathbf{Q}_N^\top \\ &= \mathcal{X} \times_1 \mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N^\top, \end{aligned}$$

which is exactly the two pass core approximation $\mathcal{W}_2$. Therefore

$$\mathcal{W}_1 - \mathcal{W}_2 = (\mathcal{X} - \tilde{\mathcal{X}}) \times_1 \mathbf{\Phi}_1^\top \times_2 \cdots \times_N \mathbf{\Phi}_N^\top \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger.$$

We continue to simplify this difference:

$$
\begin{aligned}
(\mathcal{X} - \tilde{\mathcal{X}}) &\times_1 \mathbf{\Phi}_1^\top \times_2 \cdots \times_N \mathbf{\Phi}_N^\top \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \\
&= (\mathcal{X} - \tilde{\mathcal{X}}) \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \mathbf{\Phi}_1^\top \times_2 \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \mathbf{\Phi}_N^\top \\
&= (\mathcal{X} - \tilde{\mathcal{X}}) \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \mathbf{\Phi}_1^\top (\mathbf{Q}_1 \mathbf{Q}_1^\top + \mathbf{Q}_1^\perp (\mathbf{Q}_1^\perp)^\top) \dots \\
&\times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger \mathbf{\Phi}_N^\top (\mathbf{Q}_N \mathbf{Q}_N^\top + \mathbf{Q}_N^\perp (\mathbf{Q}_N^\perp)^\top) \\
&= (\mathcal{X} - \tilde{\mathcal{X}}) \times_1 (\mathbf{Q}_1^\top + (\mathbf{\Phi}_1^Q)^\dagger \mathbf{\Phi}_1^{Q^\perp} (\mathbf{Q}_1^\perp)^\top) \times_2 \dots \\
&\times_N (\mathbf{Q}_N^\top + (\mathbf{\Phi}_N^{Q_N})^\dagger \mathbf{\Phi}_N^{Q_N^\perp} (\mathbf{Q}_N^\perp)^\top).
\end{aligned}
\tag{3.B.3}
$$

Many terms in this sum are zero. We use the following two facts:

1 $(\mathcal{X} - \tilde{\mathcal{X}}) \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top = 0.$

2 For each $n \in [N]$, $\tilde{\mathcal{X}} \times_n (\mathbf{\Phi}_n^{Q_n})^\dagger \mathbf{\Phi}_n^{Q_n^\perp} (\mathbf{Q}_n^\perp)^\top = 0.$

Here 0 means a tensor with all zero elements. These facts can be obtained from the exchange rule of the mode product and the orthogonality between $\mathbf{Q}_n^\perp$ and $\mathbf{Q}_n$. Using these two facts, we find that only the terms $\mathcal{Y}_{i_1 \dots i_N}$ (defined in (3.B.2)) remain in the expression. Therefore, to complete the proof, we write (3.B.3) as

$$
\sum_{(i_1,\dots,i_N) \in \{0,1\}^N, \sum n=1^N i_n \neq 0} \mathcal{Y}_{i_1 \dots i_N}.
$$

□

3.B.2 Probabilistic Core Error Bound

In this section, we derive a probabilistic error bound based on the core error decomposition from Lemma 3.B.1.

Lemma 3.B.2. *Sketch the tensor $\mathcal{X}$ using a Tucker sketch with parameters $\mathbf{k}$ and $\mathbf{s} > 2\mathbf{k}$ with i.i.d. Gaussian $\mathcal{N}(0,1)$ DRMs. Define $\Delta = \max_{n=1}^{N} \frac{k_n}{s_n - k_n - 1}$. Then for any natural numbers $1 \leqslant \rho < \mathbf{k} - 1$,*

$$\mathbb{E}\|\mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top\|_F^2 \leqslant \Delta \left[\sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1} \right) (\tau_{\rho_n}^{(n)})^2 \right].$$

Proof. It suffices to show

$$\mathbb{E}\left[\|\mathcal{W}_1 - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top\|_F^2 \mid \mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_N \right] \leqslant \Delta \|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2. \tag{3.B.4}$$

Then take the expectation with respect to $\mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_N$ and apply results in 3.5.1 to bound $\|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2$ to finish the proof. To show 3.B.4, we will use the fact that the core DRMs $\{\mathbf{\Omega_n}\}_{n\in[N]}$ are independent of the factor matrix DRMs $\{\mathbf{\Phi_n}\}_{n\in[N]}$, and that the randomness in each factor matrix approximation $\mathbf{Q}_n$ comes solely from $\mathbf{\Omega}_n$.

For $i \in \{0,1\}^N$, define $\mathcal{B}_{i_1 \ldots i_N} =$

$$\mathcal{X} \times_1 (\mathbf{1}_{i_1=0}\mathbf{Q}_1\mathbf{Q}_1^\top + \mathbf{1}_{i_1=1}\mathbf{Q}_1^\perp(\mathbf{Q}_1^\perp)^\top) \cdots \times_N (\mathbf{1}_{i_N=0}\mathbf{Q}_N\mathbf{Q}_N^\top + \mathbf{1}_{i_N=1}\mathbf{Q}_N^\perp(\mathbf{Q}_N^\perp)^\top).$$

3.B.1 decomposes the core error as the sum of $\mathcal{Y}_{i_1 \cdots i_n}$ where $\sum_{n=1}^{N} i_n \geqslant 1$. Applying 3.D.1 and using the orthogonal invariance of the Frobenius norm, we observe

$$\mathbb{E}\left[\|\mathcal{Y}_{i_1 \ldots i_N}\|_F^2 \mid \mathbf{\Omega}_1 \cdots \mathbf{\Omega}_N \right] = \left(\prod_{n=1}^{N} \Delta_n^{i_n} \right) \|\mathcal{B}_{i_1 \ldots i_N}\|_F^2 \leqslant \Delta \|\mathcal{B}_{i_1 \ldots i_N}\|_F^2$$

when $\sum_{n=1}^{N} i_n \geqslant 1$, where $\Delta_n = \frac{k_n}{s_n - k_n - 1} < 1$ and $\Delta = \max_{n=1}^{N} \Delta_n$.

Suppose $\mathbf{q}_1, \mathbf{q}_2 \in \{0,1\}^N$ are index (binary) vectors of length N. For different indices $\mathbf{q}_1$ and $\mathbf{q}_2$, there exists some $1 \leqslant r \leqslant N$ such that their r-th element is different. Without loss of generality, assume $\mathbf{q}_1(r) = 0$ and $\mathbf{q}_2(r) = 1$ to see

$$\langle \mathcal{B}_{q_1}, \mathcal{B}_{q_2} \rangle = \langle \ldots \mathbf{Q}_r^\top \mathbf{Q}_r^\perp \ldots \rangle = 0. \tag{3.B.5}$$

Similarly we can show that the inner product between $\mathcal{Y}_{q_1}$ and $\mathcal{Y}_{q_2}$ is zero with different $\mathbf{q}_1, \mathbf{q}_2$. Noticing that $\mathcal{B}_{0,\dots,0} = \hat{\mathcal{X}}_2$, we have

$$\|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2 = \left\| \sum_{(i_1,\dots,i_N)\in\{0,1\}^N, \sum_{n=1}^N i_n \geqslant 1} \mathcal{B}_{i_1\dots i_N} \right\|_F^2 = \sum_{\substack{(i_1,\dots,i_N)\in\{0,1\}^N, \\ \sum_{n=1}^N i_n \geqslant 1}} \|\mathcal{B}_{i_1\dots i_N}\|_F^2.$$

Putting all these together and using the Pythagorean theorem, to show 3.B.4:

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{W} - \mathcal{X} \times_1 \mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N^\top\|_F^2 \mid \mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_N\right] \\
&= \sum_{(i_1,\dots,i_N)\in\{0,1\}^N, \sum_{n=1}^N i_n \geqslant 1} \mathbb{E}\left[\|\mathcal{Y}_{i_1\dots i_N}\|_F^2 \mid \mathbf{\Omega_1}, \dots, \mathbf{\Omega}_N\right] \\
&\leqslant \Delta \left(\sum_{(i_1,\dots,i_N)\in\{0,1\}^N, \sum_{n=1}^N i_n \geqslant 1} \|\mathcal{B}_{i_1\dots i_N}\|_F^2 \right) = \Delta \|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2.
\end{aligned}$$

□

3.C Proof of fixed rank approximation lemma

Proof of 3.4.1. The target tensor to be approximated is $\mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N$ is apparently in the space $\{\mathcal{X} : \mathcal{X}^{(n)} \in \mathbf{col}(\mathbf{Q}_n)\}$. For any approximation $\hat{\mathcal{X}}$, we can project it into this space as

$$\hat{\mathcal{X}} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \times_2 \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top$$

and by Pythagorean theory,

$$\begin{aligned}
&\|\hat{\mathcal{X}} - \mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N\|_F \leqslant \|\hat{\mathcal{X}} - \hat{\mathcal{X}} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top\|_F^2 \\
&+ \|\hat{\mathcal{X}} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top - \mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N\|_F^2,
\end{aligned} \tag{3.C.1}$$

which indicates that the optimal Tucker decomposition resides in the space $\{\mathcal{X} : \mathcal{X}^{(n)} \in \mathbf{col}(\mathbf{Q}_n)\}$. Suppose $[\![\mathcal{W}; \mathbf{V}_1, \cdots, \mathbf{V}_N]\!]$ is the optimal solution to the problem, since its

unfolding is in the space spanned by $\mathbf{Q}_n$, each $\mathbf{V}_n$ can be written as $\mathbf{Q}_n\mathbf{U}_n$ for some orthogonal matrix $\mathbf{U}_n \in \mathbb{R}^{k_n \times r_n}$. Then, noticing orthogonal transformation does not change Frobenius norm,

$$\begin{aligned}
&\|\mathcal{W} \times_1 \mathbf{Q}_1 \times \cdots \times_N \mathbf{Q}_N - \mathcal{G} \times_1 \mathbf{Q}_1\mathbf{U}_1 \times \cdots \times_N \mathbf{Q}_N\mathbf{U}_N\|_F \\
&= \|\mathcal{W} - \mathcal{G} \times_1 \mathbf{U}_1 \times \cdots \times_N \mathbf{U}_N\|_F \geqslant \|\mathcal{W} - [\![\mathcal{W}]\!]_{\mathbf{r}}\|_F \\
&= \|\mathcal{W} \times_1 \mathbf{Q}_1 \times \cdots \times_N \mathbf{Q}_N - [\![\mathcal{W}]\!]_{\mathbf{r}} \times_1 \mathbf{Q}_1 \times \cdots \times_N \mathbf{Q}_N\|_F.
\end{aligned} \tag{3.C.2}$$

This finishes the proof. □

3.D Technical Lemmas

3.D.1 Random projections of matrices

Proofs for lemmas in this section appear in [Halko et al., 2011, chapters 9 and 10].

Lemma 3.D.1. *Assume that $t > q$. Let $\mathbf{G}_1 \in \mathbb{R}^{t\times q}$ and $\mathbf{G}_2 \in \mathbb{R}^{t\times p}$ be independent standard normal matrices. For any matrix $\mathbf{B}$ with conforming dimensions,*

$$\mathbb{E}\|\mathbf{G}_1^\dagger\mathbf{G}_2\mathbf{B}\|_F^2 = \frac{q}{t-q-1}\|\mathbf{B}\|_F^2.$$

Lemma 3.D.2. *Suppose that $\mathbf{A}$ is a real $m \times n$ matrtix with singular value $\sigma_1 \geqslant \sigma_2 \geqslant \cdots$, choose a target rank $k \geqslant 2$ and an oversampling parameter $p \geqslant 2$, where $k + p \leqslant \min\{m, n\}$. Draw an $n \times (k + p)$ standard Guassian matrix $\mathbf{\Omega}$, and construct the sample matrix $\mathbf{Y} = \mathbf{A\Omega}$, then the expectation of approximation error is*

$$\mathbb{E}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Y}})\mathbf{A}\|_F^2 \leqslant \left(1 + \frac{k}{p-1}\right)\left(\sum_{j>k}\sigma_j^2\right).$$

3.E More Algorithms

This section provides detailed implementations for a linear sketch appropriate to a streaming setting (Algorithm 7) or a distributed setting (8).

Algorithm 7 Linear Update to Sketches

function SKETCHLINEARUPDATE($\mathcal{F}, \mathbf{V}_1, \ldots, \mathbf{V}_N, \mathcal{H}; \theta_1, \theta_2$)
For $n = 1, \ldots, N$
    $\mathbf{V}_n \leftarrow \theta_1 \mathbf{V}_n + \theta_2 \mathbf{F}^{(n)} \mathbf{\Omega}_n$
    $\mathcal{H} \leftarrow \theta_1 \mathcal{H} + \theta_2 \mathcal{F} \times_1 \mathbf{\Phi}_1 \times \cdots \times_N \mathbf{\Phi}_N$
    return $(\mathbf{V}_1, \ldots, \mathbf{V}_N, \mathcal{H})$
end function

Algorithm 8 Sketching in Distributed Setting

Require: $\mathcal{X}_i$ is the part of the tensor $\mathcal{X}$ at local machine i and $\mathcal{X} = \sum_{i=1}^{m} \mathcal{X}_i$.
function COMPUTESKETCHDISTRIBUTED($\mathcal{X}_1, \ldots, \mathcal{X}_m$)
    Send the same random generating environment to every local machine.
    Generate the same DRM at each local machine.
For $i = 1 \ldots m$
    $(\mathbf{V}_1^{(i)}, \cdots, \mathbf{V}_n^{(i)}, \mathcal{H}^{(i)}) \leftarrow$ ComputeSketch($\mathcal{X}_i$)
For $j = 1 \ldots n$
    $\mathbf{V}_j \leftarrow \sum_{i=1}^{m} \mathbf{V}_j^{(i)}$
    $\mathcal{H} \leftarrow \sum_{i=1}^{m} \mathcal{H}^{(i)}$
    return $(\mathbf{V}_1, \ldots, \mathbf{V}_n, \mathcal{H})$
end function

3.F Scrambled Subsampled Randomized Fourier Transform

In order to reduce the cost of storing the test matrices, in particular, $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_N$, we can use the Scrambled Subsampled Randomized Fourier Transform (SSRFT). To reduce the dimension of a matrix, $\mathbf{X} \in \mathbb{R}^{m \times n}$, along either the row or the column to size k, we

define the SSRFT map $\boldsymbol{\Xi}$ as:

$$\boldsymbol{\Xi} = \begin{cases} \mathbf{R}\mathbf{F}^\top \boldsymbol{\Pi}\mathbf{F}\boldsymbol{\Pi}^\top \in \mathbb{F}^{k\times m} & \text{(Row linear transform)} \\ (\bar{\mathbf{R}}\bar{\mathbf{F}}^\top \bar{\boldsymbol{\Pi}}\bar{\mathbf{F}}\bar{\boldsymbol{\Pi}}^\top)^\top \in \mathbb{F}^{n\times k} & \text{(Column linear transform)}, \end{cases}$$

where $\boldsymbol{\Pi}, \boldsymbol{\Pi}' \in \mathbb{R}^{m\times m}, \bar{\boldsymbol{\Pi}}, \bar{\boldsymbol{\Pi}}' \in \mathbb{R}^{n\times n}$ are signed permutation matrices. That is, the matrix has exactly one non-zero entry, 1 or -1 with equal probability, in each row and column. $\mathbf{F} \in \mathbb{F}^{m\times m}, \mathbf{F} \in \mathbb{F}^{n\times n}$ denote the discrete cosine transform ($\mathbb{F} = \mathbb{R}$) or the discrete fourier transform ($\mathbb{F} = \mathbb{C}$). The matrix $\mathbf{R}, \bar{\mathbf{R}}$ is the restriction to k coordinates chosen uniformly at random.

In practice, we implement the SSRFT as in Algorithm 9. It takes only $\mathcal{O}(m)$ or $\mathcal{O}(n)$ bits to store $\boldsymbol{\Xi}$, compared to $\mathcal{O}(km)$ or $\mathcal{O}(kn)$ for Gaussian or uniform random map. The cost of applying $\boldsymbol{\Xi}$ to a vector is $\mathcal{O}(n \log n)$ or $\mathcal{O}(m \log m)$ arithmetic operations for fast Fourier transform and $\mathcal{O}(n \log k)$ or $\mathcal{O}(m \log k)$ for fast cosine transform. Though in practice, SSRFT behaves similarly to the Gaussian random map, its analysis is less comprehensive Ailon and Chazelle [2009]; Boutsidis and Gittens [2013]; Tropp [2011] than the Gaussian case.

Algorithm 9 Scrambled Subsampled Randomized Fourier Transform (Row Linear Transform)

Require: $\mathbf{X} \in \mathbb{R}^{m\times n}, \mathcal{F} = \mathbb{R}$, **randperm** creates a random permutaion vector, and **randsign** creates a random sign vector. **dct** denotes the discrete cosine transform.

1: **function** SSRFT($\mathbf{X}$)
2: **coords** ← **randperm**(m,k)
3: $\mathbf{perm}_j$ ← **randperm**(m) for $j = 1, 2$
4: $\mathbf{sgn}_j$ ← **randsign**(m) for $j = 1, 2$
5: $\mathbf{X}$ ← **dct**$(\mathbf{sgn}_1 \cdot \mathbf{X}[\mathbf{perm}_1, :])$ ▷ elementwise product
6: $\mathbf{X}$ ← **dct**$(\mathbf{sgn}_2 \cdot \mathbf{X}[\mathbf{perm}_2, :])$
7: **return** $\mathbf{X}[\mathbf{coords}, :]$
8: **end function**

3.G TensorSketch

Many authors have developed methods to perform dimension reduction efficiently. In particular Diao et al. [2017] proposed a method called tensor sketching aiming to solve least square problem with design matrix has kroneck product structure. Malik and Becker [2018] applied this technique to their one pass Tucker decomposition. Here we review the definition of tensor sketch and how it be applied in Malik and Becker [2018].

CountSketch Cormode and Hadjieleftheriou [2008] proposed the CountSketch method. A comprehensive theoretical analysis in the context of low-rank approximation problems appears in Clarkson and Woodruff [2017]. To compute the sketch $\mathbf{X}\mathbf{\Omega} \in \mathbb{R}^{d\times k}$ for $\mathbf{X} \in \mathbb{R}^{m\times d}$, CountSketch defines $\mathbf{\Omega} = \mathbf{D}\mathbf{\Phi}$, where

1 $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with each diagonal entry equal to $(-1, 1)$ with probability $(1/2, 1/2)$.

2 $\mathbf{\Phi} \in \mathbb{R}^{d\times k}$ is the matrix form of a Hashing function.

In total, these two matrices have $2d$ non-zero entries in total, thus requiring much less storage than the standard kd entries. Furthermore, these two matrices can act as an operator on each column of $\mathbf{X}$ and require only $\mathcal{O}(kd)$ operations.

TensorSketch Malik and Becker [2018] proposes to use the countsketch inside the HOOI method for Tucker decomposition. They apply sketching method solve least square problem appearing in (3.2.4) and (3.2.5) in 2. They use J_1, J_2 to denote the reduced dimension. Using a standard random map, it will need J_1-by-$I_{(-n)}$ random matrix for 3.2.4 and a J_2-by-$\prod_{n=1}^{N} I_n$ random matrix to compute 3.2.5.

But as shown in Malik and Becker [2018], these two stages can be expressed as

$$\text{For } n = 1, \dots, N, \text{update } \mathbf{U}^{(n)} = \underset{\mathbf{U} \in \mathbb{R}^{I_n \times R_n}}{\arg\min} \left\| \left(\bigotimes_{\substack{i=N \\ i \neq n}}^{1} \mathbf{U}^{(i)} \right) \mathbf{G}_{(n)}^{\top} \mathbf{U}^{\top} - \mathbf{Y}_{(n)}^{\top} \right\|_F^2. \quad (3.G.1)$$

$$\text{Update } \mathcal{G} = \underset{\mathcal{Z} \in \mathbb{R}^{R_1 \times \cdots \times R_N}}{\arg\min} \left\| \left(\bigotimes_{i=N}^{1} \mathbf{U}^{(i)} \right) \mathbf{vec}\, \mathcal{Z} - \mathbf{vec}\, \mathcal{Y} \right\|_2^2, \quad (3.G.2)$$

where $\mathcal{Y}$ is the original data. $\forall i \in [n]$, $\mathbf{U}_i$ is the factor matrix, and $\mathcal{G}$ is the core tensor. $R_1, \dots, R_N$ denote the rank of the data.

As what shown in Cormode and Hadjieleftheriou [2008], Malik and Becker [2018] proposes to apply tensorSketch to the Kronecker product structure of the input matrix in the sketch construction, i.e. $\otimes_{\substack{i=1 \\ i \neq n}}^{N} \mathbf{U}_i$ in 3.G.1 and $\otimes_{i=1}^{N} \mathbf{U}_i$ in 3.G.2. TensorSketch method combines the CountSketch of each factor matrix via the Khatri-Rao product and Fast Fourier Transform. Consider sketching $\otimes_{i=1}^{N} \mathbf{U}_i$ in 3.G.2. TensorSketch is defined as

$$\mathbf{\Omega X} = \text{FFT}^{-1}\left(\odot_{n=1}^{N} \left(\text{FFT}\big(\text{CountSketch}^{(n)}(\mathbf{U}^{(n)})\big)^{\top} \right)^{\top} \right) \quad (3.G.3)$$

By only storing $\text{CountSketch}^{(1)}, \dots, \text{CountSketch}^{(N)}$, TensorSketch only requires $2\sum_{i=1}^{N} I_n$ storage. Therefore, the storage cost of the sketch is dominated by the sketch size, $NR^{n-1}J_1 + J_2R^n \approx NKR^{2n-2} + KR^{2n}$, when $J_1 = KR^{n-1}, J_2 = KR^n$.

3.H More Numerics

This section provides more numerical results on simulated datasets in 3.9, 3.10, 3.11, and 3.12.

We also provide more numerical results on real datasets in Figure 3.13.

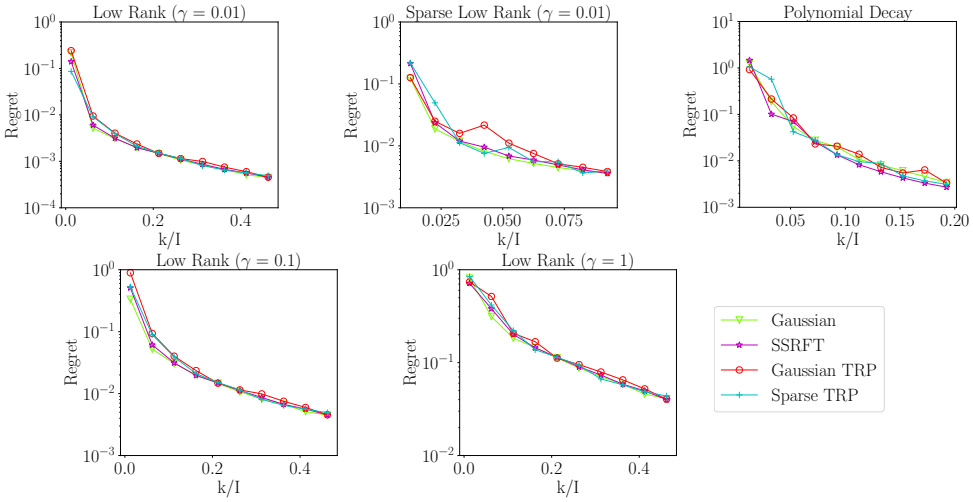


Figure 3.9: We approximate 3D synthetic tensors (see 3.6.1) with $I = 400$, using our one-pass algorithm with $r = 5$ and varying k ($s = 2k + 1$), using a variety of DRMs in the Tucker sketch: Gaussian, SSRFT, Gaussian TRP, or Sparse TRP.

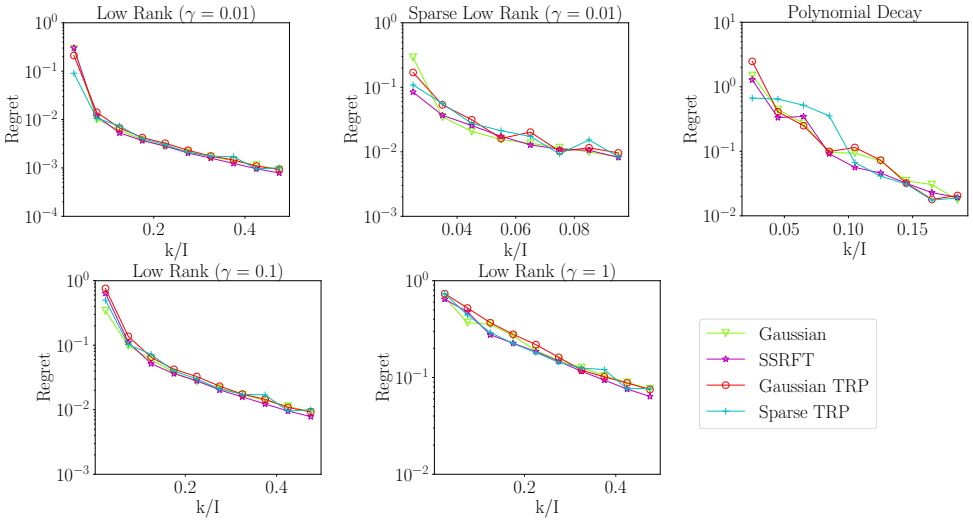


Figure 3.10: We approximate 3D synthetic tensors (see 3.6.1) with $I = 200$, using our one-pass algorithm with $r = 5$ and varying k ($s = 2k + 1$), using a variety of DRMs in the Tucker sketch: Gaussian, SSRFT, Gaussian TRP, or Sparse TRP.

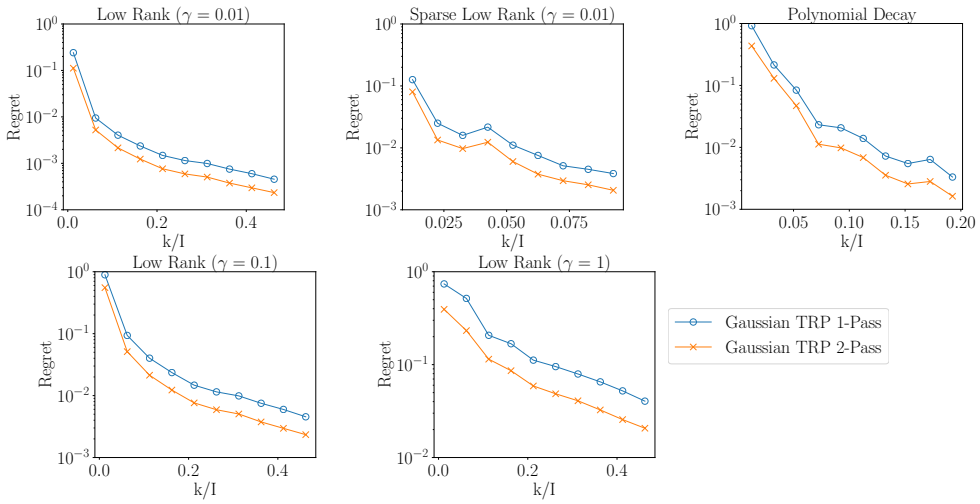


Figure 3.11: We approximate 3D synthetic tensors (see 3.6.1) with $I = 400$, using our one-pass and two-pass algorithms with $r = 5$ and varying k ($s = 2k + 1$), using the Gaussian TRP in the Tucker sketch.

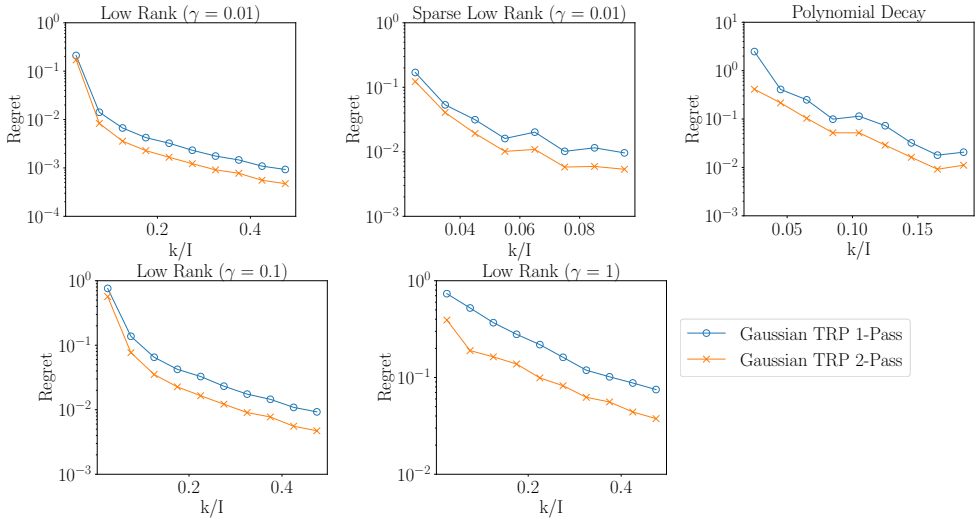
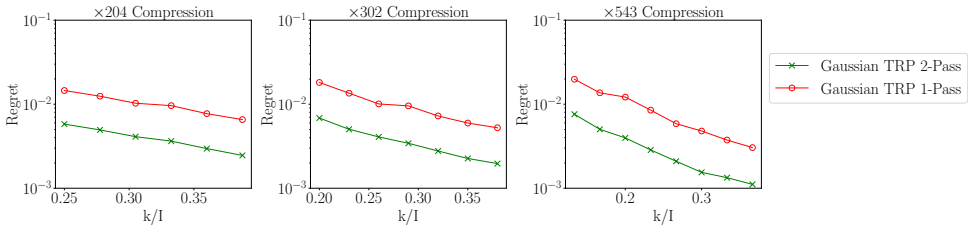
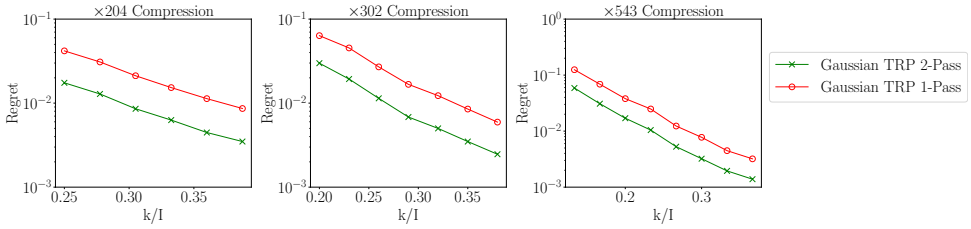


Figure 3.12: We approximate 3D synthetic tensors (see 3.6.1) with $I = 200$, using our one-pass and two-pass algorithms with $r = 5$ and varying k ($s = 2k + 1$), using the Gaussian TRP in the Tucker sketch.



Net Radiative Flux at Surface



Dust Aerosol Burden

Figure 3.13: We approximate the net radiative flux and dust aerosol burden data using our one-pass and two-pass algorithms using Gaussian TRP. We compare the performance under different ranks ($r/I = 0.125, 0.2, 0.067$). The dataset comes from the CESM CAM. The dust aerosol burden measures the amount of aerosol contributed by the dust. The net radiative flux determines the energy received by the earth surface through radiation.

CHAPTER 4

TENSOR RANDOM PROJECTION FOR LOW MEMORY DIMENSION REDUCTION

4.1 Introduction

Linear random projection is operating a random matrix onto the data which could be either high dimension vector or matrix to reduce the dimension while preserving the useful information residing in the data. A linear random projection can be represented as a random matrix $\mathbf{\Omega} \in \mathbb{R}^{d\times k}$, operating on a vector $\mathbf{x} \in \mathbb{R}^d$ or a matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ to reduce the dimension:

$$\begin{aligned} &\mathbf{x} \in \mathbb{R}^n \to \mathbf{\Omega}^\top \mathbf{x} \in \mathbb{R}^k \\ &\mathbf{X} \in \mathbb{R}^{m\times d} \to \mathbf{X}\mathbf{\Omega} \in \mathbb{R}^{m\times k}. \end{aligned} \tag{4.1.1}$$

Random projection application to vector has a very long history which enables a broad range of modern applications from bio-informatics, informational retrieval to computer vision like [Allen-Zhu et al., 2014; Bingham and Mannila, 2001; Buhler and Tompa, 2002; Fradkin and Madigan, 2003; Halko et al., 2011; Jegou et al., 2008; Wang et al., 2012; Wright et al., 2009]. In the context of large-scale relational databases, these maps enable applications like information retrieval [Papadimitriou et al., 2000], similarity search [Kaski, 1998; Sahin et al., 2005], and privacy preserving distributed data mining [Liu et al., 2006]. Later, along with the fast development in randomized algorithm, linear random projections are widely employed in constructing fast randomized algorithms in fields like matrix and tensor decomposition [Tropp et al., 2017; Woolfe et al., 2008], optimization [Yurtsever et al., 2017], streaming data compression [Sun et al., 2019; Tropp et al., 2019a]. The much smaller matrix after random projection in

the second line in (4.1.1) is named *sketch*. The term 'sketch' describes the fact that the matrix after random projection captures most of the action of the original matrix.

The effectiveness of random projection is measured by whether the information inside data is well preserved in the low dimensional embedding after random projection. For the case where we operates random matrix $\mathbf{\Omega}$ onto vector $\mathbf{x}$, we require

$$\|\mathbf{\Omega}^\top \mathbf{x}_1 - \mathbf{\Omega}^\top \mathbf{x}_2\| \approx \|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{4.1.2}$$

For the case, where we operating $\mathbf{\Omega}$ onto matrix $\mathbf{X}$, it requires that

$$\|\mathbf{Q}\mathbf{Q}^\top \mathbf{X} - \mathbf{X}\| \text{ is small,} \tag{4.1.3}$$

where $\mathbf{Q}$ is the ortho-normal matrix got from QR factorization from $\mathbf{X\Omega}$. For the case where $\mathbf{\Omega}$ has i.i.d. elements, there are many literature in how these two properties are preserved. We list two well known results in literature for those two cases separately.

Lemma 4.1.1 (Arriaga and Vempala [2006])**.** *Let $\mathbf{x} \in \mathbb{R}^d$, assume that the entries in $\mathbf{\Omega} \in \mathbb{R}^{d\times k}$ are sampled independently from $\mathcal{N}(0,1)$. Then*

$$\mathbb{P}\left((1-\epsilon)\|\mathbf{x}\|^2 \leqslant \left\|\frac{1}{\sqrt{k}}\mathbf{\Omega}^\top \mathbf{x}\right\| \leqslant (1+\epsilon)\|\mathbf{x}\|^2\right) \leqslant 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}. \tag{4.1.4}$$

Lemma 4.1.2 (Halko et al. [2011])**.** *Let $\mathbf{X} \in \mathbb{R}^{m\times d}$, assume that the entries in $\mathbf{\Omega} \in \mathbb{R}^{d\times(k+p)}$ are sampled independently from $\mathcal{N}(0,1)$. Then let $\mathbf{Q}$ be the orthonormal matrix from QR factorization $\mathbf{X\Omega} = \mathbf{QR}$, then*

$$\|\mathbf{X} - \mathbf{Q}\mathbf{Q}^\top \mathbf{X}\|_F \leqslant \left(1 + \frac{k}{p-1}\right)^{1/2} \left(\sum_{j>k} \sigma_j^2\right)^{1/2}. \tag{4.1.5}$$

Memory Efficient Random Projection Sparse random maps for low memory dimension reduction were first proposed by [Achlioptas, 2003], and further work has improved the memory requirements and guarantees of these methods [Ailon and Chazelle,

2006; Bourgain et al., 2015; Li et al., 2006]. Usually they propose the random projection to be sparse. But under modern 'big data' setting, their cost in storage/memory cost is still to big to be practical. Consider

Most closely related to our work is Rudelson's foundational study [Rudelson, 2012], which considers how the spectral and geometric properties of the random maps we use in this paper resemble a random map with iid entries, and shows that their largest and smallest singular values are of the same order. These results have been widely used to obtain guarantees for algorithmic privacy, but not for random projection. Battaglino et al. [Battaglino et al., 2018] use random projections of Khatri-Rao products to develop a randomized least squares algorithm for tensor factorization; in contrast, our method uses the (full) Khatri-Rao product to enable random projection. Sparse random projections to solve least squares problems were also explored in [Wang et al., 2015] and [Woodruff et al., 2014]. To our knowledge, this paper is the first to consider using the Khatri-Rao product for low memory random projection. However, if the dimension of vectors before reduction (here, the size of the lexicon) is too big, the storage cost of the random map is not negligible. Furthermore, even generating the pseudo-random numbers used to produce the random projection is expensive [Matsumoto and Nishimura, 1998].

To reduce the storage burden, we propose a novel use of the row-product random matrices in random projection, and call it the *Tensor Random Projection* (TRP), formed as the Khatri-Rao product of a list of smaller dimension reduction maps. We show this map is an approximate isometry, with tunable accuracy, and hence can serve as a useful dimension reduction primitive. Furthermore, the storage required to compress d dimension vectors scales as $\sqrt[N]{d}$ where N is the number of smaller maps used to form the TRP. We also develop a reduced variance version of the TRP that allows separate

control of the dimension of the range and the quality of the isometry.

4.1.1 Notation

We denote *scalar*, *vector*, and *matrix* variables, respectively, by lowercase letters (x), boldface lowercase letters ($\mathbf{x}$), and boldface capital letters ($\mathbf{X}$). Let $[N] = \{1, \ldots, N\}$. For matrix $\mathbf{X}$, we denote its i^{th} row, j^{th} column, and the $(i,j)^{th}$ element as $\mathbf{X}_{(i,.)}, \mathbf{X}_{(.,j)}, \mathbf{X}_{(i,j)}$. The *Kronecker product* of two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{p \times q}$, denoted as $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{mp \times nq}$, is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A_{11}\mathbf{B} & \cdots & A_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ A_{m1}\mathbf{B} & \cdots & A_{mn}\mathbf{B} \end{bmatrix}.$$

We let $\mathbf{X} \odot \mathbf{Y}$ denotes the *Khatri-Rao product*, $\mathbf{A} \in \mathbb{R}^{I \times K}, \mathbf{B} \in \mathbb{R}^{J \times K}$, i.e. the "matching column-wise" Kronecker product. The resulting matrix of size $(IJ) \times K$ is given by:

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{A}_{(1,.)} \otimes \mathbf{B}_{(1,.)}, \ldots, \mathbf{A}_{(K,.)} \otimes \mathbf{B}_{(K,.)}]. \tag{4.1.6}$$

4.2 Tensor Random Projection

We seek a random projection map to embed a collection of vectors $\mathcal{X} \subseteq \mathbb{R}^d$ into $\mathbb{R}^k$ with $k \ll d$. Let us take $d = \prod_{n=1}^{N} d_n$, motivated by the problem of compressing (the vectorization of) an order N tensor with dimensions $d_1, \ldots, d_N$. Conventional random projections use $O(kd)$ random variables. Generating so many random numbers

is costly; and storing them can be costly when d is large. Is so much randomness truly necessary for a random projection map?

To reduce randomness and storage requirements, we propose the *tensor random projection* (TRP):

$$f_{\text{TRP}}(\mathbf{x}) := (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)^\top \mathbf{x}, \tag{4.2.1}$$

where each $\mathbf{A}_i \in \mathbb{R}^{d_i \times k}$, for $i \in [N]$, can be an arbitrary RP map and $\mathbf{A} := (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)^\top$. We call N the *order* of the TRP. We show in this paper that the TRP is an expected isometry, has vanishing variance, and supports database-friendly operations.

The TRP requires only $k \sum_{i=1}^{N} d_i$ random variables (or $k \sqrt[N]{d}$ by choosing each d_i to be equal), rather than the kd random variables needed by conventional methods. Hence the TRP is database friendly: it significantly reduces storage costs and randomness requirements compared to its constituent DRMs.

In large scale database settings, where computational efficiency is critical and queries of vector elements are costly, practitioners often use sparse RPs. Let δ be the proportion of non-zero elements in the RP map. To achieve a δ-sparse RP, a common construction is the scaled sign random map: each element is distributed as $(-1/\sqrt{\delta}, 0, 1/\sqrt{\delta})$ with probability $(\delta/2, 1-\delta, \delta/2)$. Achlioptas [2003] proposed $\delta = 1/3$, while Li et al. [2006] further suggests a sparser scheme with $\delta = 1/\sqrt{d}$ that he calls the *Very Sparse* RP.

To further reduce memory requirements of random projection, we can form a TRP whose constituent submatrices are generated each with sparsity factor δ, which leads to a δ^N-sparse TRP. Under sparse setting, it is a $(1/3)^N$ sparse TRP while under very sparse setting, it is a $1/\sqrt{d}$ sparse TRP. Both TRPs can be applied to a vector using very few queries to vector elements and no multiplications. Below, we show both sparse and

very sparse TRP are low-variance approximate isometry empirically.

4.3 Main Theory

In this section, we discuss the properties of tensor random projection with application to length preservation and column space preservation.

4.3.1 Bias and Variance

In this section, we will show the TRP and TRP_T are expected isometries with vanishing variance. We provide a rate for the decrease in variance with k. We also prove a non-asymptotic concentration bound on the quality of the isometry when $N = 2$. We begin by showing the TRP is an approximate isometry.

Theorem 4.3.1. *Fix* $\mathbf{x} \in \mathbb{R}^{\prod_{n=1}^{N} d_n}$. *Form a TRP and* TRP_T *of order* N *with range* k *composed of independent matrices with independent columns whose entries are mean zero, variance one, and within each column every pair of elements has covariance zero. Then*

$$\mathbb{E}\|\text{TRP}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2 \qquad \textit{and} \qquad \mathbb{E}\|\text{TRP}_T(\mathbf{x})\|^2 = \|\mathbf{x}\|^2.$$

Interestingly, Theorem 4.3.1 does not require elements of $\mathbf{A}_n$ to be iid. Now we present an explicit form for the variance of the isometry.

Theorem 4.3.2. *Fix* $\mathbf{x} \in \mathbb{R}^{\prod_{n=1}^{N} d_n}$. *Form a* TRP *and* TRP_T *of order* N *with range* k *independent matrices whose entries are i.i.d. with mean zero, variance one, and fourth*

moment Δ. *Then*

$$\begin{aligned}
\textit{Var}(\|\text{TRP}(\mathbf{x})\|^2) &= \frac{1}{k}(\Delta^N - 3)\|\mathbf{x}\|_4^4 + \frac{2}{k}\|\mathbf{x}\|_2^4 \\
\textit{Var}(\|\text{TRP}_T(\mathbf{x})\|^2) &= \frac{1}{Tk}(\Delta^N - 3)\|\mathbf{x}\|_4^4 + \frac{2}{k}\|\mathbf{x}\|_2^4.
\end{aligned}$$

We can see the variance increases with N. In the $N = 1$ Gaussian case, this formula shows a variance of $2/k\|\mathbf{x}\|_2^4$, which agrees with the classic result. Notice the TRP_T only reduces the first term in the variance bound: as $T \to \infty$, the variance converges to that of a Gaussian random map.

Next, since TRP_T is a linear operator, treat $\mathbf{x} - \mathbf{y}$ as a vector, with above argument, we have the following lemma for pair-wise distance. Proof is omitted for the sake of brevity.

Corollary 4.3.3. *Fix* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\prod_{n=1}^N d_n}$. *Form a* TRP_T *of order* N *with range* k *independent matrices whose entries are i.i.d with mean zero, variance one, and fourth moment* Δ. *We have*

$$\begin{aligned}
&\mathbb{E}(\|\text{TRP}_T(\mathbf{x}) - \text{TRP}_T(\mathbf{y})\|^2) = \|\mathbf{x} - \mathbf{y}\|^2, \\
&\textit{Var}(\|\text{TRP}_T(\mathbf{x}) - \text{TRP}_T(\mathbf{y})\|^2) = \frac{1}{Tk}(\Delta^N - 3)\|\mathbf{x} - \mathbf{y}\|_4^4 + \frac{2}{k}\|\mathbf{x} - \mathbf{y}\|_2^4.
\end{aligned} \tag{4.3.1}$$

For completeness, we also present the analysis for bias and variance for inner product.

Lemma 4.3.4. *Fix* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\prod_{n=1}^N d_n}$. *For* TRP *and* TRP_T *of order* N *with range* k *independent matrices whose entries are i.i.d with mean zero, variance one, and fourth*

moment Δ, we have

$$\begin{aligned}
\mathbb{E}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle) &= \mathbb{E}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle) = \langle \mathbf{x}, \mathbf{y}\rangle \\
\textit{Var}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle) &= \frac{1}{k}[(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2]. \\
\textit{Var}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle) &= \frac{1}{kT}(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + (\frac{2}{k} - \frac{1}{kT})\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \frac{1}{kT}\langle \mathbf{x}, \mathbf{y}\rangle^2.
\end{aligned} \tag{4.3.2}$$

We can see as $T \to \infty$, $\text{Var}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle) \to \frac{2}{k}\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2$, same as the variance in the Gaussian Random map case.

Remark. *If we further assume each entry of $\mathbf{x}, \mathbf{y}$ to be a random variable with their second and fourth moment bounded by constants. We can see as $d \to \infty$, $\|\mathbf{x}\|_2^4$, $\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2$, $\langle \mathbf{x}, \mathbf{y}\rangle^2$ are $\mathcal{O}(d^2)$, and $\|\mathbf{x}\|_4^4$, $\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2$ are $\mathcal{O}(d)$ respectively. Thus, $\frac{2}{k}\|\mathbf{x}\|_2^4$, i.e. the term same as in the Gaussian RP, dominates Var($\|\text{TRP}(\mathbf{x})\|^2$), and $\frac{1}{k}(\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2)$ dominates Var($\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle$).*

4.3.2 Asymptotic Behavior

4.3.3 Finite Sample Bound?

Finally we show a non-asymptotic concentration bound for $N = 2$. We leave the parallel result for $N \geqslant 3$ open for future exploration.

Proposition 4.3.5. *Fix $\mathbf{x} \in \mathbb{R}^{d_1 d_2}$ with sub-Gaussian norm φ_2. Form a* TRP*(T) of order 2 with range k composed of two independent matrices whose entries are drawn i.i.d. from a sub-Gaussian distribution with mean zero and variance one. Then there exists a*

constant C depending on φ_2 and a universal constant c_1 so that

$$\mathbb{P}\left(\left|\|f_{\text{TRP}}(\mathbf{x})\|^2 - \|\mathbf{x}\|_2^2\right| \geqslant \epsilon\|x\|^2\right) \leqslant C\exp\left[-c_1\left(\sqrt{k}\epsilon\right)^{1/4}\right],$$

Here φ_2 is the sub-Gaussian norm defined in 4.D.1 in Appendix 4.A. 4.3.5 shows that for a TRP to form an ϵ-JL DRM with substantial probability on a dataset with n points, our method requires $k = \mathcal{O}(\epsilon^{-2}\log^8 n)$ while conventional random projections require $k = \mathcal{O}(\epsilon^{-2}\log n)$. Numerical experiments suggest this bound is pessimistic.

4.3.4 Column Space Preservation

(4.1.5) in Lemma 4.1.2 shows that the random projection preserve the information in column space of a matrix well and provide the error bound compared with the tail energy. It is hard to derive similar result for general matrix with tensor random projection. But if the matrix is in form of kroneck product, we can get a similar result based as following proposition:

Proposition 4.3.6. *Let $\mathbf{X}_n \in \mathbb{R}^{m_i\times d_n}$ be a series of matrix and $\mathbf{\Omega}_n \in \mathbb{R}^{d_n\times(k+p_n)}$ with eact element sampled from standard Gaussian distribution, let $\tau_n(k) = \sum_{j>k}\sigma_j^2((x))$ be the tail energy for $\mathbf{X}_i$. Let $\mathbf{Q} \in \mathbb{R}^{d\times k}$ be the orthonormal matrix from QR factorization:*

$$\mathbf{Q}, - = \mathrm{QR}[(\mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_\mathrm{N})(\mathbf{\Omega}_1 \odot \cdots \odot \mathbf{\Omega}_\mathrm{N})]$$

we have

$$\begin{aligned}&\mathbb{E}\|(\mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_N) - \mathbf{Q}\mathbf{Q}^\top(\mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_N)\|_F^2 \\ &\leqslant \prod_{i=1}^{N}\left(1 + \frac{k}{p_n - 1}\right)\tau_n^2(k).\end{aligned} \tag{4.3.3}$$

Proof. [Schäcke, 2013] has a detailed descriptions on properties for kronecker product. Here we mainly use the association rule between kronecker product and khatri rao product, which claims for $\mathbf{A} \in \mathbb{R}^{m_n \times d_n}$ and $\mathbf{B} \in \mathbb{R}^{d_n \times k}$

$$\begin{aligned}&(\mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_N)(\mathbf{B_1} \odot \cdots \odot \mathbf{B_N}) \\ &= (\mathbf{A}_1\mathbf{B_1} \odot \cdots \odot \mathbf{A}_N\mathbf{B}_N).\end{aligned} \tag{4.3.4}$$

Also, for orthogonal matrix $\mathbf{U}_n, n = 1, \cdots, N$, ($\mathbf{U}_n^\top \mathbf{U}_n = \mathbf{I}$), $\mathbf{U}_1 \odot$

Using this association rule,

$$\begin{aligned}&(\mathbf{X}_1 \otimes \cdots \otimes \mathbf{X}_N)(\mathbf{\Omega_1} \odot \cdots \odot \mathbf{\Omega_N}) \\ &= (\mathbf{X}_1\mathbf{\Omega_1} \odot \cdots \odot \mathbf{X}_N\mathbf{\Omega_N}).\end{aligned} \tag{4.3.5}$$

Let $\mathbf{Q}_n \in \mathbb{R}^{d_n \times k}$ be the orthonormal matrix from QR factorization: $\mathbf{Q}_i, - \leftarrow \mathrm{QR}(\mathbf{X_i}\mathbf{\Omega_i})$. The key observation is that $\mathbf{Q} = \mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_N$ then we finish the proof with some association rule for kronecker product in Schäcke [2013] and result in Lemma 4.1.2. Also this proposition, indicate in practice, we should sketch each small matrix $\mathbf{X}_n$ then combine them together which is equivalent to do sketch the whole matrix after kronecker product.

□

4.4 Experiment

In this section, we compare the quality of the isometry of conventional RPs, TRP, and TRP(5), for Gaussian, Sparse Achlioptas [2003], and Very Sparse random maps Li et al. [2006] on both synthetic data and MNIST data. We also use TRP and TRP(5) to

compute pairwise cosine similarity (Table 4.1 and Appendix 4.B) and to sketch matrices and tensors (Appendix 4.5), although the theory still remains open.

Our first experiment evaluates the quality of the isometry for maps $\mathbb{R}^d \to \mathbb{R}^k$. We generate $n = 10$ independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of sizes $d = 2500, 10000, 40000$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We consider the following three RPs: 1. Gaussian RP; 2. Sparse RP Achlioptas [2003]; 3. Very Sparse RP Li et al. [2006]. For each, we compare the performance of RP, TRP, and TRP(5) with order 2 and $d_1 = d_2$. We evaluate the methods by repeatedly generating a RP and computing the reduced vector, and plot the ratio of the pairwise distance $\frac{1}{n(n-1)} \sum_{n \geqslant i \neq j \geqslant 1} \frac{\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|_2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}$ and the average standard deviation for different k averaged over 100 replications. In the MNIST example, we choose the first $n = 50$ vectors of size $d = 784$, normalize them, and perform the same experiment. Figure 4.1 shows results on simulated ($d = 2500$) and MNIST data for the Gaussian and Very Sparse RP. See 4.B for additional experiments.

These experiments show that to preserve pairwise distance and cosine similarity, TRP performs nearly as well as RP for all three types of maps. With only five replicates, TRP(5) reduces the variance significantly in real data while not much in the simulation setting. The difference in accuracy between methods diminishes as k increases. When $d = d_1 d_2 = 40000$, the storage for TRP(5) is still $\frac{1}{20}$ of the Gaussian RP. The variance reduction is effective especially in sparse and very sparse setting.

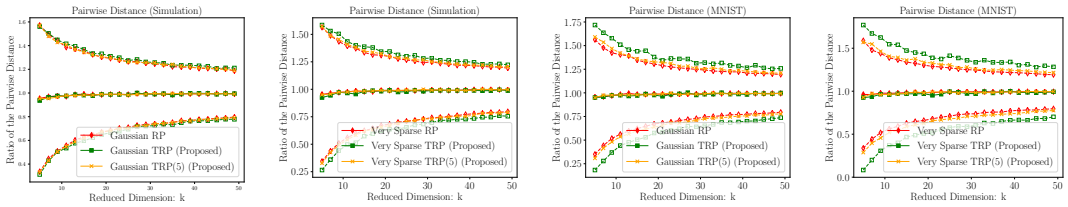


Figure 4.1: Isometry quality for simulated and MNIST data. The left two plots show results for Gaussian and Very Sparse RP, TRP, TRP(5) respectively applied to $n = 20$ standard normal data vectors in $\mathbb{R}^{2500}$. The right two plots show the same for 50 MNIST image vectors in $\mathbb{R}^{784}$. The dashed line shows the error two standard deviations from the average ratio.

	Gaussian	Sparse	Very Sparse
RP	0.1198 (0.0147)	0.1198 (0.0150)	0.1189 (0.0108)
TRP	0.1540 (0.0290)	0.1609 (0.0335)	0.1662 (0.0307)
TRP(5)	0.1262 (0.0166)	0.1264 (0.0194)	0.1276 (0.0164)

Table 4.1: RMSE for the estimate of the pairwise inner product of the MNIST data, where standard error is in the parentheses.

4.5 Application: Sketching

Beyond random projection, our novel TRP also has an important application in sketching. Sketching is an important technique to accelerate expensive computations with widespread applications, such as regression, low-rank approximation, and graph sparsification, etc. [Halko et al., 2011; Woodruff et al., 2014] The core idea behind sketching is to compress a large dataset, typically a matrix or tensor, into a smaller one by multiplying a random matrix. In this section, we will mainly focus on the low-rank matrix approximation problem. Consider a matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ with rank r, we want to find the best rank-r approximation with the minimal amount of time. The most common method is the randomized singular value decomposition (SVD), whose underlying idea is sketching.

First, we compute the linear sketch $\mathbf{Z} \in \mathbb{R}^{m \times k}$ by $\mathbf{Z} = \mathbf{X}\mathbf{\Omega}$, where $\mathbf{\Omega} \in \mathbb{R}^{d \times r}$ is the random map. Then we compute the QR decomposition of $\mathbf{X}\mathbf{\Omega}$ by $\mathbf{Q}\mathbf{R} = \mathbf{Z}$, where $\mathbf{Q} \in \mathbb{R}^{m \times k}, \mathbf{R} \in \mathbb{R}^{r \times r}$. At the end, we project $\mathbf{X}$ onto the column space of $\mathbf{Q}$, and obtain the approximation $\hat{\mathbf{X}} = \mathbf{Q}\mathbf{Q}^\top\mathbf{X}$.

With our TRP, we can significantly reduce the storage of the random map, while achieving similar rate of convergence as demonstrated in Figure 4.2. With further variance reduction by taking the geometric-median over multiple runs, our TRP with variance reduction can achieve even better performance. The detailed implementation is given in Algorithm 10. And we will delay the theoretical analysis of this method for future works.

Algorithm 10 Tensor Sketching with Variance Reduction

Require: $\mathbf{X} \in \mathbb{R}^{m \times d}$, where $d = \prod_{i=1}^{N} d_n$ and RMAP is a user-specified function that generates a random dimension reduction map. T is the number of runs for variance reduction averaging.

function $\mathrm{SSVR}(\mathbf{X}, \{d_n\}, k, T, \mathrm{RMAP})$
For $t = 1 \dots T$
For $i = 1 \dots N$
$\mathbf{\Omega}_i^{(t)} = \mathrm{RMAP}(\mathrm{d_i}, \mathrm{k})$
    $\mathbf{\Omega}^{(t)} = \mathbf{\Omega}_1^{(t)} \odot \cdots \odot \mathbf{\Omega}_N^{(t)}$
    $(\mathbf{Q}^{(t)}, \sim) = \mathrm{QR}(\mathbf{X}\mathbf{\Omega}^{(\mathrm{t})})$
    $\hat{\mathbf{X}}^{(t)} = \mathbf{Q}^{(t)}\mathbf{Q}^{(t)T}\mathbf{X}$
    $\hat{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{X}}^{(t)}$
    return $\mathbf{G}$
end function

Furthermore, the extension of TRP to tensor data is also natural. To be specific, the n^{th} unfolding of a large tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, denoted as $\mathbf{X}^{(n)}$, has dimension $I_n \times I_{(-n)}$, where $I_{(-n)} = \prod_{i \neq n, i \in [N]} I_i$. To construct a sketch for the unfolding, we need to create a random matrix of size $I_{(-n)} \times k$. Then, our TRP becomes a natural choice to avoid the otherwise extremely expensive storage cost. For many popular

tensor approximation algorithms, it is even necessary to perform sketching for every dimension of the tensor [De Lathauwer et al., 2000; Wang et al., 2015]. In the simulation section, we perform experiments for the unfolding of the higher-order order tensor with our structured sketching algorithms (Figure 4.2). For more details in tensor algebra, please refer to [Kolda and Bader, 2009].

Experimental Setup In sketching problems, considering a N-D tensor $\mathcal{X} \in \mathbb{R}^{I^N}$ with equal length along all dimensions, we want to compare the performance of the low rank approximation with different maps for its first unfolding $\mathbf{X}^{(1)} \in \mathbb{R}^{I \times I^{N-1}}$.

We construct the tensor $\mathcal{X}$ in the following way. Generate a core tensor $\mathcal{C} \in \mathbb{R}^{r^N}$, with each entry $\mathrm{Unif}([0,1])$. Independently generate N orthogonal arm matrices by first creating $\mathbf{A}_1, \ldots, \mathbf{A}_N \in \mathbb{R}^{r \times I}$ and then computing the arm matrices by $(\mathbf{Q}_n, \sim) = \mathrm{QR}(\mathbf{A}_\mathrm{n})$, for $1 \leqslant n \leqslant N$.

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N + \sqrt{\frac{0.01 \cdot \|\mathcal{X}^\natural\|_F^2}{I^N}} \mathcal{N}(0,1).$$

Then, we construct the mode-1 unfolding of $\mathbf{X} = \mathbf{X}^{(1)}$, which has a rank smaller than or equal to r.

In our simulation, we consider the scenarios of 2-D (900×900), 3-D ($400 \times 400 \times 400$), 4-D ($100 \times 100 \times 100 \times 100$) tensor data, with corresponding mode-1 unfolding of size 900×900, 400×160000, 100×1000000 respectively and $r = 5$. In each scenario, we compare the performance for Gaussian RP, TRP, and TRP_5 maps with varying k from 5 to 25. The TRP map in these scenarios has 2, 4, 6 components of size $30 \times k$, $20 \times k$, $10 \times k$ respectively. And the number of runs variance reduction averaging is $T = 5$. In the end, we evaluate the performance by generating the random matrix 100 times and compute the relative error $\frac{\|\mathbf{X} - \hat{\mathbf{X}}\|}{\|\mathbf{X}\|}$, and constructing a 95% confidence interval

for it.

Result From Figure 4.2, we can observe that the relative error decreases as k increases as expected for all dimension reduction maps. The difference of the performance between the Khatri-Rao map and Gaussian map is small when $N = 2$, but increases when N increases, whereas the Khatri-Rao variance reduced method is particularly effective producing strictly better performance than the other two.

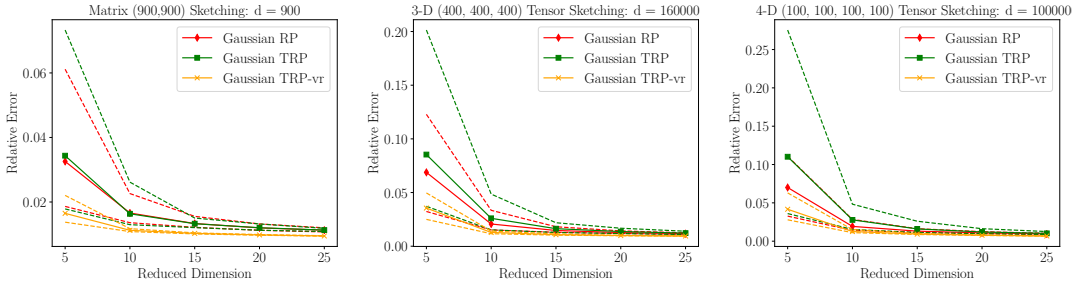


Figure 4.2: Relative Error for the low-rank tensor unfolding approximation: *we compare the relative errors for low-rank tensor approximation with different input size: 2-D (*900×900*), 3-D (*$400 \times 400 \times 400$*), 4-D (*$100 \times 100 \times 100 \times 100$*). In each setting, we compare the performance of Gaussian RP, TRP, and* TRP_5*. The dashed line stands for the 95% confidence interval.*

4.6 Conclusion

The TRP is a novel dimension reduction map composed of smaller DRMs. Compared to its constituent DRMs, it significantly reduces the requirements for randomness and for storage. Numerically, the variance-reduced TRP(5) method with only five replicates achieves accuracy comparable to the conventional RPs for $1/20$ of the original storage. We prove the TRP and TRP(T) are expected isometries with vanishing variance, and provide a non-asymptotic error bound for the order 2 TRP.

For the future work, we will provide a general non-asymptotic bound for the higher order TRP and develop the theory relevant for the application of the TRP in sketching low-rank approximation, given its practical effectiveness (shown in Appendix 4.5).

4.A Proof for Bias and Variance Analysis

Before presenting the proof for the main theory, we first define some new notations. Since these notations will only be used in technical proofs, we do not include them in the main body.

Extra Notations for Technical Proofs

For a vector $\mathbf{x}$ with length $\prod_{n=1}^{N} d_n$, for simplicity, we introduce the multi-index for it: let $\mathbf{x}_{r_1,\cdots,r_N}, \forall r_n \in [d_n]$, represent the $(1 + \sum_{n=1}^{N}(r_n - 1)s_n)^{th}$ element, where $s_n = \prod_{n+1}^{N} d_n$ for $n < N$, and 1 for $n = N$. For vector $\mathbf{r}_1$, $\mathbf{r}_2$, we say $\mathbf{r}_1 = \mathbf{r}_2$ if and only if all their elements are the same.

Also, we let $\mathbf{vec}(\mathbf{A})$ be the vectorization operator for any matrix $\mathbf{A} \in \mathbb{R}^{d\times k}$, which stacks all columns of matrix $\mathbf{A}$ and returns a vector of length kd, $[\mathbf{A}(\cdot, 1); \cdots ; \mathbf{A}(\cdot, k);]$. Here we use semi-colon to denote the vertical stack of vectors $\mathbf{x}$ and $\mathbf{y}$ as $[\mathbf{x}; \mathbf{y}]$. As comparison, we use comma to mean stack row vector horizontally like $[\mathbf{x}^\top, \mathbf{y}^\top]$.

Proof for Theorem 4.3.1

Proof. We first give a sufficient condition for general random matrix to let (4.3.1) be held, then we show that Khatri-Rao map with condition in Theorem 4.3.1 satisfies these two general sufficient conditions.

Consider a general random matrix $\mathbf{A} \in \mathbb{R}^{k\times d}$ and $\mathbf{x} \in \mathbb{R}^d$. we claim if $\mathbb{E}\mathbf{A}^2(r,s) = 1, \forall r, s$ and $\mathbb{E}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2) = 0, \forall r \in [k], s_1 \neq s_2 \in [d]$, then $\mathbb{E}\|\frac{1}{\sqrt{k}}\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$, when $\mathbf{y} = \mathbf{A}\mathbf{x}$. To see why, it suffices to show that $\mathbb{E}y_r^2 = \|x\|_2^2$.

$$\begin{aligned}\mathbb{E}y_r^2 &= \mathbb{E}\sum_{s_1=1}^{d}\sum_{s_2=1}^{d}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2)x_{s_1}x_{s_2}\\ &= \sum_{s=1}^{d}\mathbf{A}^2(r,s)x_s^2 = \|\mathbf{x}\|_2^2,\end{aligned}$$

where the first equation in the second line comes from the fact that $\mathbb{E}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2) = 0$ for $s_1 \neq s_2$ and the second equation in the second line comes from that $\mathbb{E}\mathbf{A}^2(r,s) = 1$.

Then, we will prove Theorem 4.3.1 by induction. We first show that for two matrices $\mathbf{B}_1 \in \mathbb{R}^{d_1\times k}, \mathbf{B}_2 \in \mathbb{R}^{d_2\times k}$ whose entries satisfy the two conditions in Theorem 4.3.1: $\mathbb{E}\mathbf{B}_n^2(r,s) = 1$ and $\mathbb{E}[\mathbf{B}_n(r_1,s)\mathbf{B}_n(r_2,s)] = 0$ for $n = 1,2, s \in [d], r, r_1 \neq r_2 \in [d_n]$, we have $\mathbf{A} = (\mathbf{B}_1 \odot \mathbf{B}_2)^\top$ satifies the two sufficient conditions stated previously. It suffices to restrict our focus to the first row of $\mathbf{\Omega}$ and we apply the multi-index to it. For any $1 \leqslant r_1 \leqslant d_1, 1 \leqslant r_2 \leqslant d_2$,

$$\begin{aligned}&\mathbb{E}\mathbf{A}_1^2(k_1,k_2) = \mathbb{E}\mathbf{B}_1^2(k_1,1)\mathbf{B}_2^2(k_2,1)\\ &= \mathbb{E}\mathbf{B}_1^2(k_1,1)\mathbb{E}\mathbf{B}_2^2(k_1,1) = 1. \text{ (independence bewteen } \mathbf{B}_i, i = 1,2)\end{aligned}$$

To avoid confusion in notation, we argue that $\mathbf{A}(1,\cdot)$ is the first row vector of $\mathbf{A}$ of size d_1d_2, and we apply the multi-index to it. Also, for two different elements in the first row of $\mathbf{A}$: $\mathbf{A}_1(k_1,k_2)\mathbf{A}_1(s_1,s_2)$ at least one of $k_1 \neq s_1$, $k_2 \neq s_2$ hold. Without losing

generality, assuming $k_1 \neq s_1$,

$$\begin{aligned}
&\mathbb{E}\mathbf{A}_1(k_1,k_2)\mathbf{A}_1(s_1,s_2) = \mathbb{E}\mathbf{B}_1(k_1,1)\mathbf{B}_2(k_2,1)\mathbf{B}_1(s_1,1)\mathbf{B}_2(s_2,1) \\
&= \mathbb{E}\mathbb{E}\left[\mathbf{B}_1(k_1,1)\mathbf{B}_1(k_2,1)\mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1) \mid \mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1)\right] \\
&= \mathbb{E}\mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1)\mathbb{E}\left[\mathbf{B}_1(k_1,1)\mathbf{B}_1(s_1,1)\right] = 0,
\end{aligned}$$

where we use the fact that entries within/across B_i are independent with each other and have zero expectation.

Notice that two conditions for $\mathbf{A} = (\mathbf{B}_1 \odot \mathbf{B}_2)^\top$ directly show that $\mathbf{B}_1 \odot \mathbf{B}_2$ satisfies two conditions in Theorem 4.3.1, we could use a standard mathematical induction argument to finish the proof for TRP. For TRP(T),

$$\begin{aligned}
\mathbb{E}\|\text{TRP}_T(\mathbf{x})\|_2^2 &= \frac{1}{T}\mathbb{E}\|\sum_{t=1}^{T}\text{TRP}^{(t)}(\mathbf{x})\|_2^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\text{TRP}^{(t)}(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^2,
\end{aligned}$$

where in the second line we use the fact that each $\text{TRP}^{(t)}$ is independent with each other. □

Next we introduce a lemma which shows that by bounding the deviation for the norm square of each vector, we could also bound the deviation for inner product. Although it is commonly known in any random projection literature, for completeness, we still list the lemma with proof here.

Proof for Theorem 4.3.2

Proof. Let $\mathbf{y} = \mathbf{A}\mathbf{x}$. We know from Theorem 4.3.1 that $\mathbb{E}\|\text{TRP}(\mathbf{x})\|_2^2 = \frac{1}{k}\mathbb{E}\|\mathbf{A}\mathbf{x}\|^2 = \|\mathbf{x}\|_2^2$. Notice

$$\mathbb{E}(\|\text{TRP}_T(\mathbf{x})\|_2^2) = \|\mathbf{x}\|_2^2,$$

and $\mathbb{E}y_1^2 = \|x\|_2^2$ as shown in the poof of Lemma 4.3.1. It is easy to see that

$$\mathbb{E}\|\mathbf{y}\|_2^4 = \sum_{i=1}^{k}\mathbb{E}y_i^4 + \sum_{i\neq j}\mathbb{E}y_i^2y_j^2.$$

Again, as shown in Theorem 4.3.1, $\mathbb{E}y_i^2y_j^2 = \mathbb{E}y_i^2\mathbb{E}y_j^2 = \|\mathbf{x}\|^4$. To find $\mathbb{E}\|\mathbf{y}\|_2^4$, it suffices to find $\mathbb{E}y_1^4$ by noticing that y_i are iid random variables. Let Ω be the set containing all corresponding multi-index vector for $\{1, \cdots, \prod_{n=1}^{N} d_n\}$.

$$y_1^4 = \left[\sum_{\mathbf{r}\in\Omega}\mathbf{A}(1,\mathbf{r})x_{\mathbf{r}}\right]^4 = \sum_{\mathbf{r}\in\Omega}\mathbf{A}^4(1,\mathbf{r})x_{\mathbf{r}}^4 + 3\sum_{\mathbf{r}_1\neq\mathbf{r}_2\in\Omega}\mathbf{A}^2(1,\mathbf{r}_1)x_{\mathbf{r}_1}^2\mathbf{A}^2(1,\mathbf{r}_2)x_{\mathbf{r}_2}^2$$
$$+ 6\sum_{\mathbf{r}_1\neq\mathbf{r}_2\neq\mathbf{r}_3\in\Omega}\mathbf{A}^2(1,\mathbf{r}_1)x_{\mathbf{r}_1}\mathbf{A}(1,\mathbf{r}_2)x_{\mathbf{r}_2}\mathbf{A}(1,\mathbf{r}_3)x_{\mathbf{r}_3} + 4\sum_{\mathbf{r}_1\neq\mathbf{r}_2\in\Omega}\mathbf{A}^3(1,\mathbf{r}_1)x_{\mathbf{r}_1}^3\mathbf{A}(1,\mathbf{r}_2)x_{\mathbf{r}_2}$$
$$+ \sum_{\mathbf{r}_1\neq\mathbf{r}_2\neq\mathbf{r}_3\neq\mathbf{r}_4\in\Omega}\mathbf{A}(1,\mathbf{r}_1)x_{\mathbf{r}_1}\mathbf{A}(1,\mathbf{r}_2)x_{\mathbf{r}_2}\mathbf{A}(1,\mathbf{r}_3)x_{\mathbf{r}_3}\mathbf{A}(1,\mathbf{r}_4)x_{\mathbf{r}_4}.$$

It is not hard to see that except for the first line, the expectation of second and third line is zero.

$$\mathbb{E}\mathbf{A}^4(1,\mathbf{r}) = \mathbb{E}\mathbf{A}_1^4(1,r_1)\cdots\mathbf{A}_N^4(1,r_N) = \Delta^N.$$

Also with proof in Theorem 4.3.1,

$$\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbf{A}^2(1,\mathbf{r}_2) = \mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_2) = 1.$$

Combining these two together, we have

$$\begin{aligned}\mathbb{E}\|\text{TRP}(\mathbf{x})\|^4 &= \frac{1}{k^2}\left[k(\Delta^N-3)\|\mathbf{x}\|_4^4 + 3k\|\mathbf{x}\|_2^4 + (k-1)k\|\mathbf{x}\|_2^4\right]\\ &= \frac{1}{k}\left[(\Delta^N-3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right] + \|\mathbf{x}\|_2^4.\end{aligned} \tag{4.A.1}$$

Therefore,

$$\text{Var}(\|\text{TRP}(\mathbf{x})\|_2^2) = \mathbb{E}\|\text{TRP}(\mathbf{x})\|_2^4 - (\mathbb{E}\|\text{TRP}(\mathbf{x})\|_2^2)^2 = \frac{1}{k}\left[(\Delta^N - 3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right].$$

Now we switch to see how much variance could be reduced by the variance reduction method. With Theorem 4.3.1, we already know that $\mathbb{E}\|\text{TRP}_T(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^2$. The rest is to calculate $\mathbb{E}\|\text{TRP}_T(\mathbf{x})\|_2^4$ out.

$$\|\text{TRP}_T(\mathbf{x})\|_2^4 = \frac{1}{T^2}\left[\sum_{t=1}^{T}\|\text{TRP}^{(t)}(\mathbf{x})\|_2^2 + \sum_{t_1\neq t_2}\langle\text{TRP}^{(t_1)}(\mathbf{x}), \text{TRP}^{(t_2)}(\mathbf{x})\rangle\right]^2$$

$$= \frac{1}{T^2}\left[\sum_{t=1}^{T}\|\text{TRP}^{(t)}(\mathbf{x})\|_2^4 + \sum_{t_1\neq t_2}\|\text{TRP}^{(t_1)}(\mathbf{x})\|_2^2\|\text{TRP}^{(t_2)}(\mathbf{x})\|_2^2 + 2\sum_{t_1\neq t_2}\langle\text{TRP}^{(t_1)}(\mathbf{x}), \text{TRP}^{(t_2)}(\mathbf{x})\rangle^2 + \text{rest}\right].$$

It is not hard to show that $\mathbb{E}(\text{rest}) = 0$. Following the definition of $\mathbf{y}$,

$$\mathbb{E}\|\text{TRP}^{(t_1)}(\mathbf{x})\|_2^2\|\text{TRP}^{(t_2)}(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^4,$$

and

$$\begin{aligned}
&\mathbb{E}\langle\text{TRP}^{(t_1)}(\mathbf{x}), \text{TRP}^{(t_2)}(\mathbf{x})\rangle^2 \\
&= \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^{k} y_i^{(t_1)} y_i^{(t_2)}\right]^2 \\
&= \frac{1}{k}\mathbb{E}(y_1^{(t_1)})^2\mathbb{E}(y_1^{(t_2)})^2 = \frac{1}{k}\|\mathbf{x}\|_2^4.
\end{aligned}$$

Combining all these together, we could show that

$$\begin{aligned}
\text{Var}(\|\text{TRP}_T(\mathbf{x})\|_2^2) &= \mathbb{E}\|\text{TRP}_T(\mathbf{x})\|_2^4 - (\mathbb{E}\|\text{TRP}_T(\mathbf{x})\|_2^2)^2 \\
&= \frac{1}{T^2}\left[\frac{T}{k}\left[(\Delta^N - 3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right]\right. \\
&\left.+T\|\mathbf{x}\|_2^4 + T(T-1)\|\mathbf{x}\|_2^4 + \frac{2T(T-1)}{k}\|\mathbf{x}\|_2^4\right] - \|\mathbf{x}\|_2^4 \\
&= \frac{1}{Tk}(\Delta^N - 3)\|\mathbf{x}\|_4^4 + \frac{2}{k}\|\mathbf{x}\|_2^4.
\end{aligned}$$

□

Proof for Lemma 4.3.4

Proof. First, we show the unbiasedness of the inner product estimation:

$$\mathbb{E}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y}) \rangle) = [\|\text{TRP}(\mathbf{x}) + \text{TRP}(\mathbf{y})\|_2^2 - \|\text{TRP}(\mathbf{x})\|_2^2 - \|\text{TRP}(\mathbf{y})\|_2^2]/2 = \langle \mathbf{x}, \mathbf{y} \rangle.$$

The equation above follows from Thm 4.3.1, the unbiasedness of norm estimation. We can apply the similar idea to get $\mathbb{E}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y}) \rangle) = \langle \mathbf{x}, \mathbf{y} \rangle$.

Now, let $\mathbf{u} = \mathbf{A}\mathbf{x}$, $\mathbf{v} = \mathbf{A}\mathbf{y}$. Then,

$$\begin{aligned}
(u_1 v_1)^2 &= [\sum_{\mathbf{r}\in\Omega} \mathbf{A}(1,\mathbf{r}) x_{\mathbf{r}}]^2 [\sum_{\mathbf{r}\in\Omega} \mathbf{A}(1,\mathbf{r}) y_{\mathbf{r}}]^2 \\
&= \sum_{\mathbf{r}} \mathbf{A}(1,\mathbf{r})^4 x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + \sum_{\mathbf{r}_1 \neq \mathbf{r}_2} \mathbf{A}(1,\mathbf{r}_1)^2 \mathbf{A}(1,\mathbf{r}_2)^2 x_{\mathbf{r}_1}^2 y_{\mathbf{r}_2}^2 \\
&\quad + 2 \sum_{\mathbf{r}_1 \neq \mathbf{r}_2} \mathbf{A}(1,\mathbf{r}_1)^2 \mathbf{A}(1,\mathbf{r}_2)^2 x_{\mathbf{r}_1} x_{\mathbf{r}_2} y_{\mathbf{r}_1} y_{\mathbf{r}_2} + \text{rest.},
\end{aligned}$$

Since $\mathbb{E}\mathbf{A}(1,\mathbf{r}) = 0,\ \forall \mathbf{r}$, $\mathbb{E}(\text{rest}) = 0$. Also with proof in Thm 4.3.1,

$$\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbf{A}^2(1,\mathbf{r}_2) = \mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_2) = 1.$$

And,

$$\mathbb{E}\mathbf{A}^4(1,\mathbf{r}) = \mathbb{E}\mathbf{A}_1^4(1,r_1)\cdots\mathbf{A}_N^4(1,r_N) = \Delta^N.$$

Then, similar to (4.A.1), we can obtain:

$$\begin{aligned}\mathbb{E}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle)^2 &= \frac{1}{k^2}\mathbb{E}[\sum_{i,j} u_i v_i u_j v_j]^2 = \frac{1}{k^2}\mathbb{E}[\sum_{i}(u_i v_i)^2] + \frac{1}{k^2}\mathbb{E}[\sum_{i\neq j}(u_i v_i u_j v_j)] \\ &= \frac{1}{k}\mathbb{E}(u_1 v_1)^2 + \frac{k(k-1)}{k^2}\langle \mathbf{x}, \mathbf{y}\rangle^2 \\ &= \frac{1}{k}[(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2] + \langle \mathbf{x}, \mathbf{y}\rangle^2.\end{aligned} \tag{4.A.2}$$

Then, with the unbiasedness of TRP map, we get

$$\begin{aligned}\text{Var}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle) &= \mathbb{E}(\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle^2) - (\mathbb{E}\langle \text{TRP}(\mathbf{x}), \text{TRP}(\mathbf{y})\rangle)^2 \\ &= \frac{1}{k}[(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2].\end{aligned}$$

Now, we proceed to find the variance for the inner product estimation with TRP_T. Since $\text{Var}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle) = \mathbb{E}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle^2) - (\mathbb{E}\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle)^2$, we first compute:

$$\begin{aligned}\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle^2 &= \frac{1}{T^2}\langle \sum_{i=1}^{T} \text{TRP}^{(i)}(\mathbf{x}), \sum_{j=1}^{T} \text{TRP}^{(j)}(\mathbf{y})\rangle^2 \\ &= \frac{1}{T^2}\sum_{i=1}^{T}\langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(i)}(\mathbf{y})\rangle^2 \\ &+ \frac{1}{T^2}\sum_{i\neq j}(\langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(i)}(\mathbf{y})\rangle)(\langle \text{TRP}^{(j)}(\mathbf{x}), \text{TRP}^{(j)}(\mathbf{y})\rangle) \\ &+ \frac{2}{T^2}\sum_{i\neq j}\langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(j)}(\mathbf{y})\rangle^2 + \text{rest}.\end{aligned}$$

Following the definition of the TRP map, we can see:

$$\begin{aligned}&\mathbb{E}\langle \text{TRP}^{(t_1)}(\mathbf{x}), \text{TRP}^{(t_2)}(\mathbf{y})\rangle^2 \\ &= \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^{k} u_i^{(t_1)} v_i^{(t_2)}\right]\left[\sum_{i=1}^{k} u_i^{(t_1)} v_i^{(t_2)}\right] \\ &= \frac{1}{k}\mathbb{E}[u_1^{(t_1)} u_1^{(t_1)}]\mathbb{E}[v_1^{(t_2)} v_1^{(t_2)}] = \frac{1}{k}\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2.\end{aligned}$$

First, $\mathbb{E}(\text{rest}) = 0$. Then, combining all the above results, we obtain:

$$
\begin{aligned}
\text{Var}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle) &= \mathbb{E}(\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle^2) - (\mathbb{E}\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle)^2 \\
&= \mathbb{E}(\langle \frac{1}{T^2}\sum_i \text{TRP}^{(i)}(\mathbf{x}), \frac{1}{T^2}\sum_j \text{TRP}^{(j)}(\mathbf{y})\rangle^2) - (\mathbb{E}\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle)^2 \\
&= \frac{1}{T^2}\mathbb{E}(\sum_i \langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(i)}(\mathbf{y})\rangle^2 \\
&+ \sum_{i\neq j}\langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(i)}(\mathbf{y})\rangle\langle \text{TRP}^{(j)}(\mathbf{x}), \text{TRP}^{(j)}(\mathbf{y})\rangle \\
&+ 2\sum_{i\neq j}\langle \text{TRP}^{(i)}(\mathbf{x}), \text{TRP}^{(j)}(\mathbf{y})\rangle^2) - (\mathbb{E}\langle \text{TRP}_T(\mathbf{x}), \text{TRP}_T(\mathbf{y})\rangle)^2 \\
&= \frac{1}{T^2}\left[\frac{T}{k}[(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2] + T\langle \mathbf{x}, \mathbf{y}\rangle^2\right. \\
&\left. + \frac{2T(T-1)}{k}\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + T(T-1)\langle \mathbf{x}, \mathbf{y}\rangle^2\right] - \langle \mathbf{x}, \mathbf{y}\rangle^2 \\
&= \frac{1}{kT}(\Delta^N - 3)\sum_{\mathbf{r}} x_{\mathbf{r}}^2 y_{\mathbf{r}}^2 + (\frac{2}{k} - \frac{1}{kT})\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + \frac{1}{kT}\langle \mathbf{x}, \mathbf{y}\rangle^2.
\end{aligned}
$$

□

4.B More Simulation Results

Pairwise Distance Estimation In Figure 4.3, 4.4, 4.5, we compare the performance of Gaussian, Sparse, Very Sparse random maps on the pairwise distance estimation problem with $d = 2500, 10000, 40000, N = 2$. Additionally, we compare their performance for $d = 125000, N = 3$ in Figure 4.6.

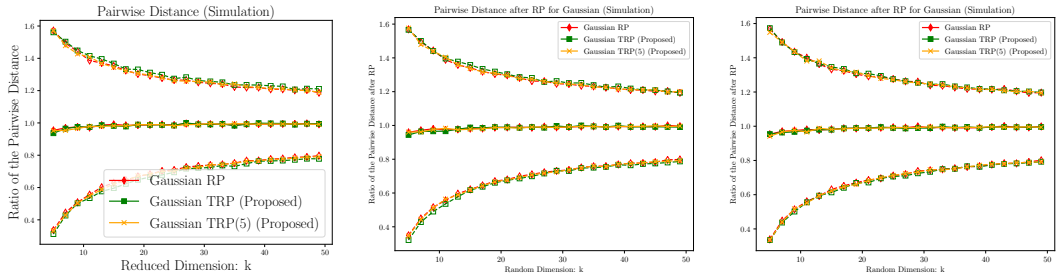


Figure 4.3: Average ratio of the pairwise distance for simulation data using Gaussian RP: *The plots correspond to the simulation for Gaussian RP, TRP,* TRP_5 *respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

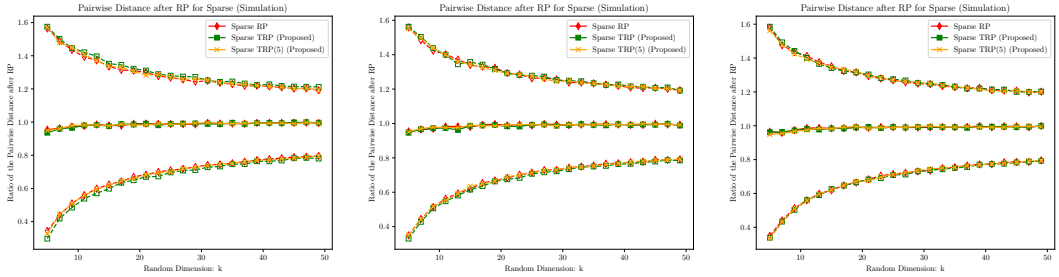


Figure 4.4: Average ratio of the pairwise distance for simulation data using Sparse RP: *The plots correspond to the simulation for Sparse RP, TRP,* TRP_5 *respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

Pairwise Cosine Similarity Estimation The second experiment is to estimate the pairwise cosine similarity, i.e. $\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$ for $\mathbf{x}_i, \mathbf{x}_j$. We use both the simulation data ($d = 10000$) and the MNIST data ($d = 784, n = 60000$). We experiment with Gaussian,

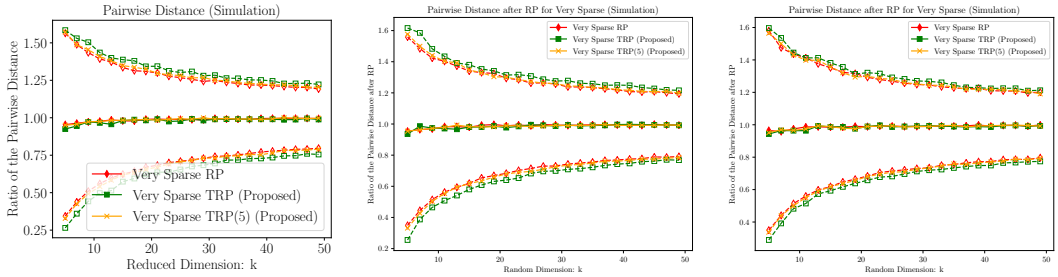


Figure 4.5: Average ratio of the pairwise distance for simulation data using Very Sparse RP: *The plots correspond to the simulation for Very Sparse RP, TRP,* TRP_5 *respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

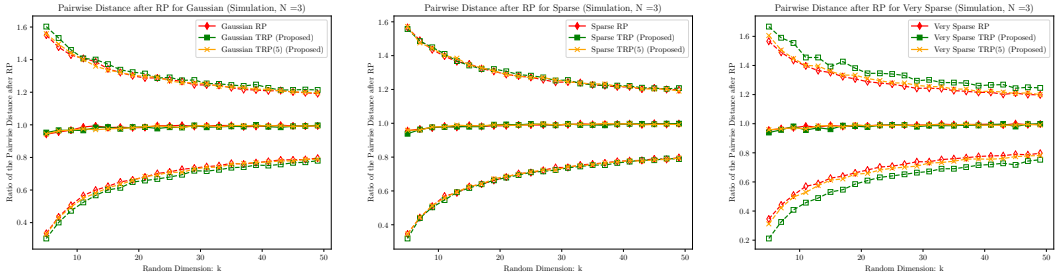


Figure 4.6: Average ratio of the pairwise distance for simulation data using: *The plots correspond to the simulation for Gaussian, Sparase, Very Sparse RP, TRP,* TRP_5 *respectively with* $n = 20, d = d_1 d_2 d_3 = 50 \times 50 \times 50 = 125000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

Sparse, Very Sparse RP, TRP, and TRP_5 with the same setting as above ($k = 50$). We evaluate the performance by the average root mean square error (RMSE). The results is given in Table 4.1, 4.2.

	Gaussian	Sparse	Very Sparse
RP	0.1409 (0.0015)	0.1407 (0.0013)	0.1412 (0.0014)
TRP	0.1431 (0.0016)	0.1431 (0.0015)	0.1520 (0.0033)
TRP_5	0.1412 (0.0012)	0.1411 (0.0015)	0.1427 (0.0014)

Table 4.2: RMSE for the estimate of the pairwise inner product of the simulation data ($d = 10000, k = 50, n = 100$), where standard error is in the parentheses.

4.C Appendix: Finite Sample Bound

Definition 4.C.1. *A random variable x is said to satisfy the generalized-sub-exponential moment condition with constant α, if for general positive integer k, there exists a general constant C(not depending on k), s.t.*

$$\mathbb{E}|x|^k \leqslant (Ck)^{k\alpha} \tag{4.C.1}$$

Proof for Proposition 4.3.5

Proof. From now on, with losing generality, we will assume $\|x\| = 1$. Let

$$\mathbf{y} = \frac{1}{\sqrt{k}}(\mathbf{A}_1 \odot \mathbf{A}_2)^\top \mathbf{x},$$

Lemma **??** asserts that $\mathbb{E}\|\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$ (conditions in lemma **??** naturally hold for iid random variables in our setting). The key observation is that $y_i, i \in [k]$ is quadratic form of elements of $\mathbf{A}_i, i = 1, 2$. Then as quadratic form of sub-Gaussian variables, y_i are identically independently distributed generalized sub-exponential random variable. Then we could use Hanson-Wright inequality to determine the constants in moments condition 4.C.1 which shall present tighter bound compared to directly citing results of linear combination of sub-exponential random variable defined in (4.C.1)

We aim to write y_i as a quadratic form of $\mathbf{z}_i := [\mathbf{vec}(\mathbf{A}_1(\cdot, i)); \mathbf{vec}(\mathbf{A}_2(\cdot, i))]$. Also, for convenience, we partition $\mathbf{x}$ into d_1 sub-vectors with equal length d_2 i.e., $\mathbf{x} = [\mathbf{x}_1; \cdots; \mathbf{x}_{d_1}]$. To make it clear, we consider writing y_1 as quadratic form of $\mathbf{z}_1$ first.

$$y_1 = \langle [\mathbf{A}_1(1,1)\mathbf{A}_2(\cdot,1); \cdots; \mathbf{A}_1(d_1,1)\mathbf{A}_2(\cdot,1)], [\mathbf{x}_1; \cdots; \mathbf{x}_{d_1}] \rangle$$

which indicates that we could write

$$y_1 = \mathbf{z}_1^\top \mathbf{M} \mathbf{z}_1,$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{D} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_{d_1}^\top. \end{bmatrix}$$

It is easy to see that $\|\mathbf{M}\| \leqslant \|\mathbf{D}\| \leqslant \|\mathbf{D}\|_F = \|\mathbf{M}\|_F = 1$ by assuming $\|\mathbf{x}\| = 1$. Then applying the Hanson Wright inequality in Lemma 4.D.1, we could have for any positive number η, there exists a general constant c_1 s.t.

$$\begin{aligned} \mathbb{P}(|y_i| \geqslant \eta) &\leqslant 2 \exp\left[-c_1 \min\left\{ -\frac{\eta}{\varphi_2^2 \|M\|}, \frac{\eta^2}{\varphi_2^4 \|M\|_F^2} \right\} \right] \\ &\leqslant 2 \exp\left[-c_1 \min\left\{ -\frac{\eta}{\varphi_2^2}, \frac{\eta^2}{\varphi_2^4} \right\} \right]. \end{aligned}$$

Then by Lemma 4.D.2, we could find a constant C depending on sub-Gaussian norm and general constant c_1 s.t.

$$\mathbb{E}|y_i|^k \leqslant (Ck)^k,$$

where in fact we could give the explicit form of C as

$$C = 1 + \frac{c_1}{\min\left\{\varphi_2^2, \varphi_2^4\right\}}. \tag{4.C.2}$$

Notice y_i has mean zero and variance 1 (assuming $\|x\| = 1$), then apply Lemma 4.D.3, we could assert that there exists a general constant c_2

$$\mathbb{P}\left(\left| \frac{1}{k} \mathbf{y}^\top \mathbf{I}_{k,k} \mathbf{y} - 1 \right| \geqslant \epsilon \right) \leqslant C \exp\left(-c_2 \left[\sqrt{k} \epsilon \right]^{1/4} \right),$$

where C is defined in (4.C.2) and we use the fact $\alpha = 1$ in our case which is defined in moments condition.

□

Lemma 4.C.1. *For a linear mapping from* $\mathbb{R}^d \to \mathbb{R}^k$*:* $f(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{\Omega}\mathbf{x}$*,*

$$\mathbb{P}(|\langle f(\mathbf{x}), f(\mathbf{y})\rangle - \langle \mathbf{x}, \mathbf{y}\rangle| \geqslant \epsilon|\langle \mathbf{x}, \mathbf{y}\rangle|) \leqslant 2 \sup_{\mathbf{x}\in\mathbb{R}^d} \mathbb{P}(|\|f(\mathbf{x})\|^2 - \|\mathbf{x}\|^2| \geqslant \epsilon\|\mathbf{x}\|_2^2).$$

Proof. Since f is a linear mapping, we have

$$4f(\mathbf{x})f(\mathbf{y}) = \|f(\mathbf{x}+\mathbf{y})\|_2^2 - \|f(\mathbf{x}-\mathbf{y})\|_2^2.$$

Consider the event

$$\mathcal{A}_1 = \left\{|\|f(\mathbf{x}+\mathbf{y})\|_2^2 - \|\mathbf{x}+\mathbf{y}\|_2^2| \geqslant \epsilon\|\mathbf{x}+\mathbf{y}\|_2^2\right\}$$
$$\mathcal{A}_2 = \left\{|\|f(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2| \geqslant \epsilon\|\mathbf{x}-\mathbf{y}\|_2^2\right\}$$

On the event $\mathcal{A}_1^\complement \cap \mathcal{A}_2^\complement$,

$$4f(\mathbf{x})f(\mathbf{y}) \geqslant (1-\epsilon)(\mathbf{x}+\mathbf{y})^2 - (1+\epsilon)(\mathbf{x}-\mathbf{y})^2 = 4\langle \mathbf{x}, \mathbf{y}\rangle - 2\epsilon(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2),$$

noticing $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \geqslant 2\langle \mathbf{x}, \mathbf{y}\rangle$, and by similar argument on the other side of the inequality, we could claim that

$$\{|\langle f(\mathbf{x}), f(\mathbf{y})\rangle - \langle \mathbf{x}, \mathbf{y}\rangle| \geqslant \epsilon|\langle \mathbf{x}, \mathbf{y}\rangle|\} \subseteq \mathcal{A}_1 \cup \mathcal{A}_2.$$

Then we finish the proof by simply applying an union bound of two events. □

Remark. *The key element of classic random projections is the dimension-free bound. Similarly, according to Prop. 4.3.5, our TRP has a norm preservation bound independent of the particular vector* $\mathbf{x}$ *and dimension* d *and thus a dimension-free inner product preservation bound according to Lemma 4.C.3.*

4.D Technical Lemmas

In this section, we list some technical lemmas we use in this paper. All of them are about tail probability of sub-Gaussian or generalized sub-exponential variables.

Definition 4.D.1. *A random variable x is called sub-Gaussian if $\mathbb{E}|x|^p = \mathcal{O}(p^{p/2})$ when $p \to \infty$. With this, we define sub-Gaussian norm for x (less than infinity) as*

$$\|x\|_{\varphi_2} = \sup_{p \geqslant 1} p^{-1/2} (\mathbb{E}|x|^p)^{1/p}. \tag{4.D.1}$$

Note that for Bernoulli random variable, i.e., $\{-1, 1\}$ with prob. $\{\frac{1}{2}, \frac{1}{2}\}$, $\varphi_2 = 1$; any bounded random variable with absolute value less than $M > 0$ has $\varphi_2 \leqslant M$. For standard Gaussian random variable, $\varphi_2 = 1$.

Lemma 4.D.1. *(Hanson-Wright Inequality) Let $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfies $\mathbb{E}\mathbf{x}_i = 0$ and $\varphi_2(x_1) \leqslant K$. Let A be an $n \times n$ matrix. Then, for every $\eta \geqslant 0$, there exists a general constant c s.t.*

$$\mathbb{P}\left(|\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^\top \mathbf{A}\mathbf{x}| \geqslant \eta\right) \leqslant 2\exp\left[-c\min\left\{\frac{\eta}{K^2\|A\|}, \frac{\eta^2}{\|A\|_F^2 K^4}\right\}\right].$$

Proof. Please refer to Rudelson and Vershynin [2013] □

Lemma 4.D.2. *Let $\mathbf{x}$ be a random variable whose tail probability satisfies for every $\eta \geqslant 0$, there exists a constant c_1 s.t.*

$$\mathbb{P}\left(|x| \geqslant \eta\right) \leqslant 2\exp\left[-c_1 \min\left(\eta, \eta^2\right)\right].$$

Then for any $k \geqslant 1$, x satisfies generalized sub-exponential moment condition 4.C.1 with $\alpha = 1$, i.e.,

$$\mathbb{E}|x|^k \leqslant (Ck)^k,$$

where $C = 1 + \frac{1}{c_1}$.

Proof.

$$\begin{aligned}\mathbb{E}|x|^k &= \int_0^1 kx^{k-1}2\exp[-c_1x^2]dx + \int_1^\infty kx^{k-1}2\exp[-c_1x]dx \\ &\leqslant 1 + + \frac{1}{c_1^k}\int_0^\infty ky^{k-1}2\exp[-y]dy \\ &= 1 + \frac{1}{c_1^k}k\Gamma(k-1) \leqslant \left[1 + \frac{1}{c_1^k}\right]k^k.\end{aligned} \tag{4.D.2}$$

Noticing $\left[1 + \frac{1}{c_1^k}\right]^{1/k} \leqslant 1 + \frac{1}{c_1}$, we finish the proof. □

Lemma 4.D.3. *For a random vector* $\mathbf{x}$ *with each element independent and identically distributed with mean zero and variance 1, suppose each element of* $\mathbf{x}$ *satisfies generalized sub-exponential moment condition as in* (4.D.2)*, that there exists a general constant* C *s.t.* $\mathbb{E}|x_1|^k \leqslant (Ck)^{\alpha k}$*. Then for any matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$*, there exists a general constant* c_1

$$\mathbb{P}\left(\left|\mathbf{x}^\top\mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^\top\mathbf{A}\mathbf{x}\right| \geqslant \eta\right) \leqslant C\exp\left(-c_1\left[\frac{\eta}{\|A\|_F}\right]^{1/(2(1+\alpha))}\right).$$

Proof. The proof is directly from Lemma 8.3 in Buhler and Tompa [2002] and we change the statement on generalized sub-exponential R.V. directly to the statement on the moment condition. □

BIBLIOGRAPHY

Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of Computing*, pages 557–563. ACM, 2006.

Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

Zeyuan Allen-Zhu, Rati Gelashvili, Silvio Micali, and Nir Shavit. Sparse sign-consistent johnson–lindenstrauss matrices: Compression with neuroscience-based constraints. *Proceedings of the National Academy of Sciences*, 111(47):16872–16876, 2014.

Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Rosa I Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

Woody Austin, Grey Ballard, and Tamara G Kolda. Parallel tensor compression for large-scale scientific data. In *Parallel and Distributed Processing Symposium, 2016 IEEE International*, pages 912–922. IEEE, 2016.

Rafael Ballester-Ripoll, Peter Lindstrom, and Renato Pajarola. Tthresh: Tensor compression for multidimensional visual data. *IEEE transactions on visualization and computer graphics*, 2019.

Muthu Baskaran, Benoît Meister, Nicolas Vasilache, and Richard Lethin. Efficient and scalable computations with sparse tensors. In *High Performance Extreme Computing (HPEC), 2012 IEEE Conference on*, pages 1–6. IEEE, 2012.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4), 2015.

Casey Battaglino, Grey Ballard, and Tamara G Kolda. A practical randomized cp tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.

Casey Battaglino, Grey Ballard, and Tamara G Kolda. Faster parallel tucker tensor decomposition using randomization. 2019.

Rajendra Bhatia, Ludwig Elsner, and Gerd Krause. Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix. *Linear algebra and its applications*, 142:195–209, 1990.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

Hilmar Böhm and Rainer Von Sachs. Structural shrinkage of nonparametric spectral estimators for multivariate time series. *Electronic Journal of Statistics*, 2:696–721, 2008.

Hilmar Böhm and Rainer von Sachs. Shrinkage estimation in the frequency domain of multivariate time series. *Journal of Multivariate Analysis*, 100(5):913–935, 2009.

Cécile Bordier, Carlo Nicolini, and Angelo Bifone. Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. *Frontiers in neuroscience*, 11:441, 2017.

Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *Geometric and Functional Analysis*, 25 (4):1009–1088, 2015.

Christos Boutsidis and Alex Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

Susan M Bowyer. Coherence a measure of the brain networks: past and present. *Neuropsychiatric Electrophysiology*, 2(1):1, 2016.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2:107–144, 2005.

David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 1981.

David R Brillinger. Time series: data analysis and theory, vol. 36. *New York, NY: Society for Industrial Mathematics. doi*, 10(1.9780898719246), 2001.

Peter J Brockwell and Richard A Davis. *Time series: theory and methods*. Springer Science & Business Media, 2013.

Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of computational biology*, 9(2):225–242, 2002.

T. Tony Cai and Harrison H. Zhou. Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica*, 22(4):1319–1349, 2012.

T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Statist.*, 10(1):1–59, 2016. doi: 10.1214/15-EJS1081.

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

Andrzej Cichocki. Tensor decompositions: a new concept in brain data analysis? *arXiv preprint arXiv:1305.0395*, 2013.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54, 2017.

Graham Cormode and Marios Hadjieleftheriou. Finding frequent items in data streams. *Proceedings of the VLDB Endowment*, 1(2):1530–1541, 2008.

Rainer Dahlhaus and Michael Eichler. Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, pages 115–137, 2003.

Rainer Dahlhaus, Michael Eichler, and Jürgen Sandkühler. Identification of synaptic connections in neural ensembles by graphical models. *Journal of neuroscience methods*, 77(1):93–107, 1997.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4): 1253–1278, 2000.

Huaian Diao, Zhao Song, Wen Sun, and David P. Woodruff. Sketching for Kronecker Product Regression and P-splines. *arXiv e-prints*, art. arXiv:1712.09473, Dec 2017.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

Michael Eichler. A frequency-domain based test for non-correlation between stationary time series. *Metrika*, 65(2):133–157, 2007.

László Erdős, Horng-Tzer Yau, and Jun Yin. Bulk universality for generalized wigner matrices. *Probability Theory and Related Fields*, pages 1–67, 2012.

Carolina Euan, Hernando Ombao, and Joaquin Ortega. The hierarchical spectral merger algorithm: a new time series clustering procedure. *arXiv preprint arXiv:1609.08569*, 2016.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Mark Fiecas and Hernando Ombao. Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111(516):1440–1453, 2016.

Mark Fiecas and Rainer von Sachs. Data-driven shrinkage of the spectral density matrix of a high-dimensional time series. *Electron. J. Statist.*, 8(2):2975–3003, 2014. doi: 10.1214/14-EJS977. URL `https://doi.org/10.1214/14-EJS977`.

Mark Fiecas, Chenlei Leng, Weidong Liu, and Yi Yu. Spectral analysis of high-dimensional time series. *arXiv preprint arXiv:1810.11223*, 2018.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522. ACM, 2003.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

James W Hurrell, Marika M Holland, Peter R Gent, Steven Ghan, Jennifer E Kay, Paul J Kushner, J-F Lamarque, William G Large, D Lawrence, Keith Lindsay, et al. The community earth system model: a framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9):1339–1360, 2013.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008.

Alexander Jung. Learning the conditional independence structure of stationary time series: A multitask learning approach. *IEEE Transactions on Signal Processing*, 63 (21):5677–5690, 2015.

Alexander Jung, Gabor Hannak, and Norbert Goertz. Graphical lasso based model selection for time series. *IEEE Signal Processing Letters*, 22(10):1781–1785, 2015.

Samuel Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural networks proceedings, 1998. ieee world congress*

on computational intelligence. the 1998 ieee international joint conference on, volume 1, pages 413–418. IEEE, 1998.

JE Kay, C Deser, A Phillips, A Mai, C Hannay, G Strand, JM Arblaster, SC Bates, G Danabasoglu, J Edwards, et al. The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96 (8):1333–1349, 2015.

Oguz Kaya and Bora Uçar. High performance parallel algorithms for the tucker decomposition of sparse tensors. In *Parallel Processing (ICPP), 2016 45th International Conference on*, pages 103–112. IEEE, 2016.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Tamara G Kolda and Jimeng Sun. Scalable tensor decompositions for multi-aspect data mining. In *2008 Eighth IEEE International Conference on Data Mining*, pages 363–372. IEEE, 2008.

Amy Kuceyeski, Keith W Jamison, Julia Owen, Ashish Raj, and Pratik Mukherjee. Functional rerouting via the structural connectome is associated with better recovery after mild tbi. *bioRxiv*, 2018. doi: 10.1101/320515. URL `https://www.biorxiv.org/content/early/2018/05/18/320515`.

Simon Lapointe, Bruno Savard, and Guillaume Blanquart. Differential diffusion effects, distributed burning, and local extinctions in high karlovitz premixed flames. *Combustion and flame*, 162(9):3341–3355, 2015.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Jiajia Li, Casey Battaglino, Ioakeim Perros, Jimeng Sun, and Richard Vuduc. An input-adaptive and in-place approach to dense tensor-times-matrix multiply. In *High Performance Computing, Networking, Storage and Analysis, 2015 SC-International Conference for*, pages 1–12. IEEE, 2015.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296. ACM, 2006.

Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1):92–106, 2006.

Osman Asif Malik and Stephen Becker. Low-rank tucker decomposition of large tensors using tensorsketch. In *Advances in Neural Information Processing Systems*, pages 10116–10126, 2018.

Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

Shanmugavelayutham Muthukrishnan et al. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236, 2005.

Hernando C. Ombao, Jonathan A. Raz, Robert L. Strawderman, and Rainer von Sachs. A simple generalised crossvalidation method of span selection for periodogram smoothing. *Biometrika*, 88(4):1186–1192, 2001.

Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 2015.

Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

Murray Rosenblatt. *Stationary sequences and random fields*. Springer, 1985.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Mark Rudelson. Row products of random matrices. *Advances in Mathematics*, 231(6): 3199–3231, 2012.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Ozgur D Sahin, Aziz Gulbeden, Fatih Emekçi, Divyakant Agrawal, and Amr El Abbadi. Prism: indexing multi-dimensional data in p2p networks using reference vectors. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 946–955. ACM, 2005.

Kathrin Schäcke. On the kronecker product. 2013.

Hai Shu and Bin Nan. Estimation of large covariance and precision matrices from temporally dependent observations. *arXiv preprint arXiv:1412.5059*, 2014.

Jimeng Sun, Dacheng Tao, Spiros Papadimitriou, Philip S Yu, and Christos Faloutsos. Incremental tensor analysis: Theory and applications. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(3):11, 2008.

Yiming Sun, Yang Guo, Joel A Tropp, and Madeleine Udell. Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*, 2018a. URL `https://r2learning.github.io/assets/papers/CameraReadySubmission%2041.pdf`.

Yiming Sun, Yige Li, Amy Kuceyeski, and Sumanta Basu. Large spectral density matrix estimation by thresholding. *arXiv preprint arXiv:1812.00532*, 2018b.

Yiming Sun, Yang Guo, Charlene Luo, Joel A Tropp, and Madeleine Udell. Low rank tucker approximation of a tensor from streaming data. *In preparation*, 2019.

J.A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. Technical report, 2019a. URL `https://arxiv.org/pdf/1902.08651.pdf`.

Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. More practical sketching algorithms for low-rank matrix approximation. Technical Report 2018-01, California Institute of Technology, Pasadena, California, 2018.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *Submitted to SISC*, 2019b.

Charalampos E Tsourakakis. Mach: Fast randomized tensor decompositions. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 689–700. SIAM, 2010.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Sara van de Geer. "lecture notes on sparsity, 2016.

Nick Vannieuwenhoven, Raf Vandebril, and Karl Meerbergen. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing*, 34(2):A1027–A1052, 2012.

M Alex O Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision*, pages 447–460. Springer, 2002.

Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for large-scale search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12): 2393–2406, 2012.

Yining Wang, Hsiao-Yu Tung, Alexander J Smola, and Anima Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *Advances in Neural Information Processing Systems*, pages 991–999, 2015.

Kam Chung Wong and Ambuj Tewari. Lasso guarantees for beta -mixing heavy tailed time series. *arXiv preprint arXiv:1708.01505*, 2017.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

Wei Biao Wu and Paolo Zaffaroni. Uniform convergence of multivariate spectral density estimates. *arXiv preprint arXiv:1505.03659*, 2015.

Alp Yurtsever, Madeleine Udell, Joel A Tropp, and Volkan Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. *arXiv preprint arXiv:1702.06838*, 2017.

Guoxu Zhou, Andrzej Cichocki, and Shengli Xie. Decomposition of big tensors with low multilinear rank. *arXiv preprint arXiv:1412.1885*, 2014.

Xi-Nian Zuo, Clare Kelly, Adriana Di Martino, Maarten Mennes, Daniel S. Margulies, Saroja Bangaru, Rebecca Grzadzinski, Alan C. Evans, Yu-Feng Zang, F. Xavier Castellanos, and Michael P. Milham. Growing together and growing apart: Regional and sex differences in the lifespan developmental trajectories of functional homotopy. *Journal of Neuroscience*, 30(45):15034–15043, 2010. doi: 10.1523/JNEUROSCI.2612-10.2010.

Antoni Zygmund. *Trigonometric series*, volume 1. Cambridge university press, 2002.