Submitted to the Annals of Statistics

LARGE SPECTRAL DENSITY MATRIX ESTIMATION BY THRESHOLDING

By Yiming Sun*, Yige Li†, Amy Kuceyeski‡ and Sumanta Basu*

*Department of Statistical Science, Cornell University**
Department of Epidemiology, Harvard T. H. Chan School of Public Health†
Department of Radiology, Weill Cornell Medical College ‡

Spectral density matrix estimation of multivariate time series is a classical problem in time series and signal processing. In modern neuroscience, spectral density based metrics are commonly used for analyzing functional connectivity among brain regions. In this paper, we develop a non-asymptotic theory for regularized estimation of high-dimensional spectral density matrices of Gaussian and linear processes using thresholded versions of averaged periodograms. Our theoretical analysis ensures that consistent estimation of spectral density matrix of a p-dimensional time series using n samples is 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 learning coherence networks among the components of a multivariate time series. We demonstrate the advantage of our estimators using simulation studies and a real data application on functional connectivity analysis with fMRI data.

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,

MSC 2010 subject classifications: Primary 62M15, 62M10
Keywords and phrases: High-dimensional Time Series, Multivariate Spectral Density

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 et al. [1997], Dahlhaus and Eichler [2003], 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 spetral 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, Rothman et al., 2009, Cai and Liu, 2011, Cai et al., 2016] 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, interpretable 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, we formally introduce our problem, provide a brief review of shrinkage estimators, and describe our proposed thresholding methods. In section 3, we derive nonasymptotic upper bounds on our proposed spectral density estimates for Gaussian time series. In section 4 we extend the results for Gaussian time series to general linear processes with different non-Gaussian noise distributions. In section 5, we conduct simulation studies to assess the finite sample properties of our proposed estimators. Section 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 $\sqrt{\mathrm{tr}(A^\dagger A)}$, 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 $0 \leq q < 1$, 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 $A \geq cB$. We use $A \asymp B$ to denote $A \gtrsim B$ and $B \gtrsim A$.

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 $\mathrm{Cov}(X_t, X_{t-\ell}) = \mathbb{E}X_t X_{t-\ell}^\top$ only depends on ℓ, so we can define autocovariance as function of the lag ℓ, viz., $\Gamma(\ell) = \mathrm{Cov}(X_t, X_{t-\ell})$. Spectral density aggregates information of autocovariance of different lag orders ℓ at a specific frequency $\omega \in [-\pi, \pi]$ as

$$f(\omega) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\omega}. \tag{2.1}$$

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 $q \geq 0$,

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

Following Basu and Michailidis [2015], we will also use $|||f||| := |||f|||_2 =$

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. $\max_{1\le r\le p} \text{ess sup}_\omega \|f_{rr}(\omega)\|$ 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

$$g_{rs}(\omega) = \frac{f_{rs}(\omega)}{\sqrt{f_{rr}(\omega) f_{ss}(\omega)}}, \tag{2.2}$$

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

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

$$I(\omega) = \sum_{|\ell|<n} \hat{\Gamma}(\ell) e^{-i\ell\omega}, \tag{2.3}$$

where $\hat{\Gamma}(\ell) = n^{-1} \sum_{t=\ell+1}^{n} X_t X_{t-\ell}^\top$ for $\ell \ge 0$, 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) $d(\omega) = \mathcal{X}^\top (C(\omega) - iS(\omega))$, where

$$\begin{aligned} C(\omega) &= \frac{1}{\sqrt{n}} (1, \cos\omega, \ldots, \cos(n-1)\omega)^\top, \\ S(\omega) &= \frac{1}{\sqrt{n}} (1, \sin\omega, \ldots, \sin(n-1)\omega)^\top. \end{aligned} \tag{2.4}$$

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

$$\hat{f}(\omega; m) = \frac{1}{2\pi(2m+1)} \sum_{|k|\leq m} I(\omega + \omega_k), \tag{2.5}$$

where $\omega_k = 2\pi k/n$, $k \in F_n$, the set of Fourier frequencies. To be precise, F_n denotes the set $\{-[\frac{n-1}{2}], \ldots, [\frac{n}{2}]\}$ 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 (2.5) becomes

$$\hat{f}(\omega_j; m) = \frac{1}{2\pi(2m+1)} \sum_{|k|\leq m} I(\omega_{j+k}). \tag{2.6}$$

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 $m = o(\sqrt{n})$, (2.6) is a consistent estimator. As in general nonparametric function estimation, one can replace the weights $1/(2m+1)$ in (2.6) 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

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

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 $\mu(\omega) = \frac{1}{p}\mathrm{tr}(f(\omega))$.

Then the shrinkage estimator is defined as the minimizer of the convex program

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

where

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

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

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

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.

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 (2.5). 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.,

$$T_\lambda(\hat{f}_{rs}(\omega_j)) = \begin{cases} \hat{f}_{rs}(\omega_j) & \text{if } |\hat{f}_{rs}(\omega_j)| \ge \lambda \\ 0 & \text{if } |\hat{f}_{rs}(\omega_j)| < \lambda, \end{cases} \tag{2.7}$$

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 $1 \le r, s \le p$. 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 family 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) $|S_\lambda(z)| \leq |z|$,
(2) $S_\lambda(z) = 0$ if $|z| \leq \lambda$,
(3) $|S_\lambda(z) - z| \leq \lambda$.

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

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

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, we provide a detailed discussion of the theoretical choices of these parameters that ensure consistent estimation in high-dimensional regime. In the next subsection 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 1 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 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.

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

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

In classical asymptotics (p fixed, $n \to \infty$) and the Kolmogorov asymptotics ($p \to \infty, n \to \infty, p^2/n \to 0$) [Brockwell and Davis, 2013, Böhm and von Sachs, 2009], 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 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 $\omega_j \in F_n, \omega_j \geq 0$, are asymptotically independent.

Algorithm 1: Threshold Selection by Frequency Domain Sample-splitting

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 $||J_1| - |J_2|| \leq 1$ 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: $\hat{\lambda}_j := \hat{\lambda}(\omega_j) = \operatorname{argmin}_{\lambda\in\mathcal{L}} \hat{R}(\omega_j, \lambda)$

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 $|m_1 - m_2| \leq 1$. 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 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 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 $0 \leq q < 1$. 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 $0 \leq q < 1$, $\|A\|_q^q$ satisfies the triangle inequality leading to

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

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

$$\|f(\omega)\| = \left\| \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\omega\ell} \right\| \leq \sum_{\ell=-\infty}^{\infty} \|\Gamma(\ell)\| \leq \sum_{\ell=-\infty}^{\infty} p\|\Gamma(\ell)\|_{\max}. \tag{3.1}$$

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

$$\Omega_n(f) = \max_{1\le r,s\le p} \sum_{\ell=-n}^{n} |\ell||\Gamma_{rs}(\ell)|, \qquad L_n(f) = \max_{1\le r,s\le p} \sum_{|\ell|>n} |\Gamma_{rs}(\ell)|. \tag{3.2}$$

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 3.1, both of these quantities are finite. In Proposition 3.4, we show how these quantities grow for some common classes of multivariate time series.

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 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}:

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

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 3.3 and 3.6, respectively. Finally, in Proposition 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 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 3.1. Then there exists a universal constant $c > 0$ such that for any $\eta > 0$ and any $p \times p$ real matrix A,*

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

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 $\mathrm{rk}(A)$. 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 3.3. *For any coordinate (r, s) with $1 \leq r, s \leq p$ and any Fourier frequency ω_j, $j \in F_n$, the estimation bias of averaged periodogram with a smoothing span $2m + 1$ satisfies*

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

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 [2009], Böhm and Von Sachs [2008], 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 3.4 and VAR process for condition 3 in Proposition 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

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

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

$$\rho(\ell) = \rho(\sigma(X_t, t \le 0), \sigma(X_t, t \ge \ell)). \tag{3.3}$$

The two characteristics $\|\Gamma(\ell)\|_{\max}$ and $\rho(\ell)$ are usually easy to describe for finite order VMA and VAR(1) model. For 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

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

A 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 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 $\|\tilde{A}_1\| \geq 1$ [Basu and Michailidis, 2015]. So we cannot directly verify the geometric decay conditions 1 and 2 in Proposition 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 C.2 in the Appendix.

PROPOSITION 3.4. *Consider a weakly stationary centered time series X_t.*

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

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

2. *Suppose X_t satisfies $\rho(\ell) \leq \sigma_X \rho_X^{|\ell|}$ where $\rho(\ell)$ is the ρ-mixing coefficient defined in (3.3). Then*

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

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 (3.4), and assume $\tilde{A}_1$ is diagonalizable with an eigendecomposition $\tilde{A}_1 = SDS^{-1}$. Then*

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

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

REMARK 3.5. *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 3.6. *There exist universal positive constants c_1, c_2 such that for any $\eta > 0$,*

(3.4)
$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right| \geq |||f|||\eta\right) \leq c_1 \exp\left[-c_2(2m+1)\min\{\eta, \eta^2\}\right].$$

A complete proof is provided in Appendix 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 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 3.7. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series satisfying Assumption 3.1, 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 $m \precsim n/\Omega_n(f)$ and $m \gtrsim |||f|||^2 \log p$, and any $R > 0$, there exist universal constants $c_1, c_2 > 0$ such that choosing a thresh-*

old

$$\lambda = 2R|||f|||\sqrt{\frac{\log p}{m}} + 2\left[\frac{m+1/2\pi}{n}\Omega_n(f) + \frac{1}{2\pi}L_n(f)\right], \tag{3.5}$$

the estimation error of thresholded averaged periodogram satisfies

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

Similarly, there exist universal positive constants c_1, c_2 such that for any $R > 0$, with the same choice of threshold in (3.5), we have

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

REMARK 3.8. *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.,

$$\max\left\{R|||f|||\sqrt{\frac{\log p}{m}}, \frac{m}{n}\Omega_n(f)\right\} = o(1). \tag{3.6}$$

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. 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 3.9. *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 3.7, there exist universal constants $c_i > 0$ such that*

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

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 3.3 and 3.6. After building the concentration inequality, all the proof left is exactly same as in Proposition 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 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 3.10. *Assume $X_t, t = 1, \ldots, n$, are n consecutive observations from a stable Gaussian time series satisfying Assumption 3.1, and consider a single Fourier frequency $\omega_j, j \in F_n$. Assume $n \succsim \Omega_n(f)|||f|||^2 \log p$. Then for any m satisfying $m \precsim n/\Omega_n(f)$ and $m \succsim |||f|||^2 \log p$, and any $R > 0$, if we set threshold value λ as (3.5) , then there exists universal constant c_1, c_2 s.t.*

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

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

$$\mathbb{P}\left(\exists\ (r, s) \in \mathcal{S}(\gamma) : T_\lambda(\hat{f}_{rs}(\omega_j)) = 0, f_{rs}(\omega_j) \neq 0\right)$$

is at most $\leq c_1 \exp[-(c_2(\gamma - 1)^2 R^2 - 2)\log p]$.

REMARK 3.11. *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 (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 3.10 to ensure coherence graph selection consistency.

PROPOSITION 3.12. *Assume* $X_t, t = 1, \ldots, n$, *are* n *consecutive observations from a stable Gaussian time series* X_t *satisfying Assumption 3.1, 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* $m \precsim n/\Omega_n(f)$ *and* $m \gtrsim |||f|||^2 \log p$ *and* λ *as in* (3.5), *there exist universal positive constants* c_1, c_2 *such that for any* $R > 0$,

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

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

$$\mathbb{P}\left(\exists\ (r, s) \in \mathcal{S}(\gamma) : T_{2\lambda/\tau}(\hat{g}_{rs}(\omega_j)) = 0, g_{rs}(\omega_j) \neq 0\right)$$

is at most $c_1 \exp[-(c_2(\gamma - 1)^2 R^2 - 2)\log p]$.

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 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:

$$X_t = \sum_{\ell=0}^{\infty} B_\ell \varepsilon_{t-\ell}, \tag{4.1}$$

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

$$\sum_{\ell=0}^{\infty} |B_{\ell,(r,s)}| < \infty \tag{4.2}$$

for any r, s, $1 \le r, s \le p$. 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 3.1 holds. A proof is given in Lemma C.8 for completeness.

We assume that each component ε_{tr}, $1 \le r \le p$, 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 $\eta > 0$, $\mathbb{P}[|\varepsilon_{tr}| > \eta] \le 2\exp\left(-\frac{\eta^2}{2\sigma^2}\right)$;

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

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

REMARK 4.1. *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 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 [Rudelson and Vershynin, 2013, Erdős et al., 2012]. 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 B.

LEMMA 4.2. *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, $1 \le r \le p$. Then*

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

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

$$\mathcal{T}_1(\eta, A) = 2\exp\left[-c\min\left\{\frac{\eta}{\|A\|}, \frac{\eta^2}{rk(A)\|A\|^2}\right\}\right],$$

$$\mathcal{T}_2(\eta, A) = c_1\exp\left[-c_2\left(\frac{\eta}{\sqrt{rk(A)}\|A\|}\right)^{\frac{1}{2+2\alpha}}\right],$$

$$\mathcal{T}_3(\eta, A) = \frac{c_3 rk(A)\|A\|^2}{\eta^2}.$$

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, $\ell \ge 0$.

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

PROPOSITION 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 (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*

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

where $\mathcal{T}_j(\eta, A), j = 1, 2, 3$, are tail decay functions for the three families, as defined in Lemma 4.2.

REMARK 4.4. *The main difference between the concentration inequalities in Lemma 4.2 and Proposition 4.3 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 4.2 applied to a np-dimensional random vector.

This result generalizes the Hanson-Wright type concentration inequality in Lemma 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 3.6 which we listed as following Proposition. The proof follows the same line as the proof of Proposition 3.6, by replacing Gaussian Hanson-Wright type inequality with those in Proposition 4.3. We omit this for sake of brevity.

PROPOSITION 4.5. *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 (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, $1 \le r, s \le p$, and any Fourier frequency $\omega_j \in F_n$, we have*

$$\mathbb{P}\left(\left|\hat{f}_{rs}(\omega_j) - \mathbb{E}\hat{f}_{rs}(\omega_j)\right| \ge |||f|||\eta\right) \le \mathcal{B}_k(\eta, m), \tag{4.3}$$

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 3.7 for linear processes with any of the three different types of noise distributions.

PROPOSITION 4.6. *Suppose $\{X_t\}$ is a linear process defined in (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 $n \gtrsim \Omega_n(f)|||f|||^2\mathcal{N}_k$, where $\mathcal{N}_1 = \log p$, $\mathcal{N}_2 = (\log p)^{4+4\alpha}$, 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

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

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

$$\begin{aligned} \mathcal{B}_1 &= c_1 \exp\left[-(c_2 R^2 - 2)\log p\right], \\ \mathcal{B}_2 &= c_3 \exp\left[-(c_4 R - 2)\log p\right], \\ \mathcal{B}_3 &= c_5 \exp\left[-2R \log p\right], \end{aligned} \tag{4.4}$$

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 3.7, by replacing Gaussian variance bound in Proposition 3.6 with Proposition 4.5. We omit this for sake of brevity.

REMARK 4.7. *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.*

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.

$$A^0 = B^0 = \begin{bmatrix} 0.5 & 0.9 & 0 \\ 0 & 0.5 & 0.9 \\ 0 & 0 & 0.5 \end{bmatrix}. \tag{5.1}$$

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

$$\text{precision}(\omega_j) = \frac{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0, \ |f_{rs}(\omega_j)| \neq 0\}}{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0\}}$$

$$\text{recall}(\omega_j) = \frac{\#\{(r,s) : \ |\hat{f}_{rs}(\omega_j)| \neq 0, \ |f_{rs}(\omega_j)| \neq 0\}}{\#\{(r,s) : \ |f_{rs}(\omega_j)| \neq 0\}}$$

$$\text{F1}(\omega_j) = 2 \times (\text{precision}(\omega_j) \cdot \text{recall}(\omega_j))/(\text{precision}(\omega_j) + \text{recall}(\omega_j)).$$

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, 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 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 3.10. 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. 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

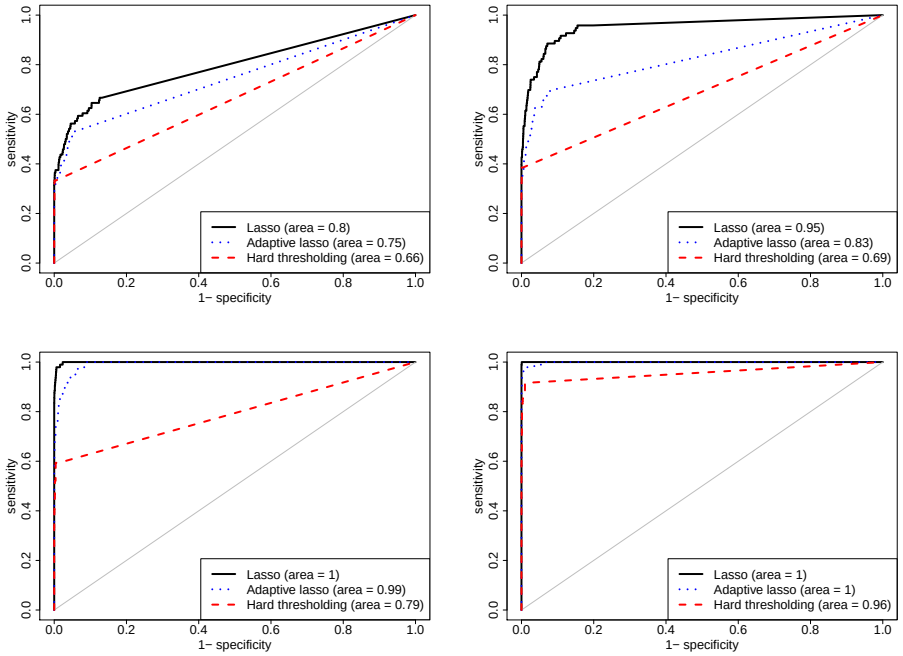

Fig 1: Receiver Operating Characeristic (ROC) curves of hard thresholding, lasso and adaptive lasso for recovering coherence network of a $p = 96$ dimensional VAR(1) model using $n = 100$ (top left), $n = 200$ (top right), $n = 400$ (bottom left) and $n = 600$ (bottom right) time series observations.

can be successfully used to estimate large spectral density matrices with same order of accuracy as shrinkage methods, and with an additional advantage of performing automatic edge selection in coherence networks.

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 traumatic 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].

	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

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.*

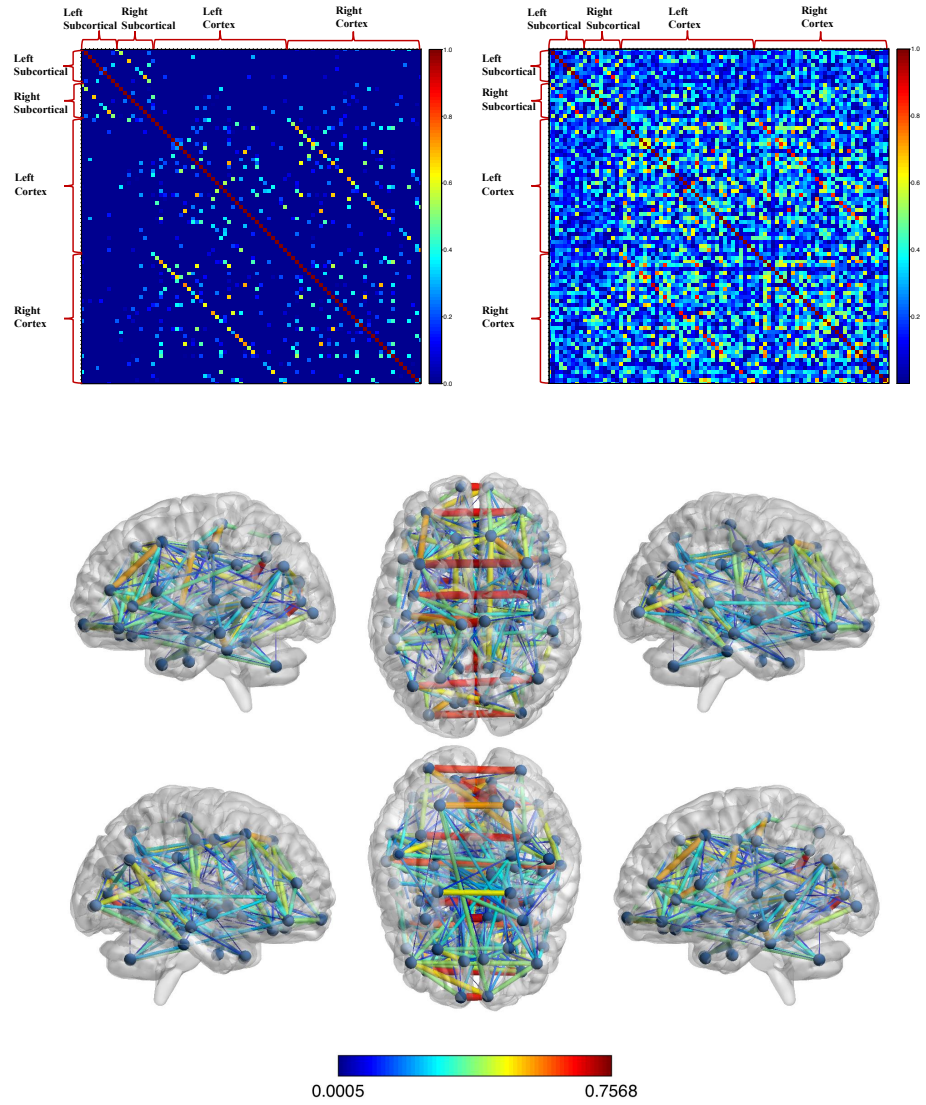

Fig 2: [top]: Heat maps of absolute coherence matrices (at frequency 0) obtained from spectral density estimated using [top left] adaptive lasso thresholding and [top right] a shrinkage method. [bottom]: Absolute coherence network among brain regions obtained using adaptive lasso and visualized using 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.

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 [Fischl and Dale, 2000]. 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 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 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 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 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 and 2 in Appendix 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 [Utevsky AV and SA, 2014]. 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 [Fransson and Marrelec, 2008]. 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.

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.

Acknowledgements. The authors wish to thank Pratik Mukherjee for providing and Keith Jamison for pre-processing the TBI patient MRI data. SB was supported by NSF award (DMS-1812128) and AK was supported by a Kellen Foundation Fellowship and the NIH (R21 NS104634-01 and R01 NS102646-01A1).

References.

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

P. J. Bickel and E. Levina. Covariance regularization by thresholding. The Annals of Statistics, pages 2577–2604, 2008.

H. Böhm and R. Von Sachs. Structural shrinkage of nonparametric spectral estimators for multivariate time series. Electronic Journal of Statistics, 2:696–721, 2008.

H. Böhm and R. von Sachs. Shrinkage estimation in the frequency domain of multivariate time series. Journal of Multivariate Analysis, 100(5):913–935, 2009.

C. Bordier, C. Nicolini, and A. Bifone. Graph analysis and modularity of brain functional connectivity networks: searching for the optimal threshold. Frontiers in neuroscience, 11:441, 2017.

S. M. Bowyer. Coherence a measure of the brain networks: past and present. Neuropsychiatric Electrophysiology, 2(1):1, 2016.

R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. Probability surveys, 2:107–144, 2005.

D. R. Brillinger. Time series: data analysis and theory, volume 36. Siam, 1981.

P. J. Brockwell and R. A. Davis. Time series: theory and methods. Springer Science & Business Media, 2013.

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

T. T. Cai and H. H. Zhou. Minimax estimation of large covariance matrices under ℓ_1-norm. Statistica Sinica, 22(4):1319–1349, 2012.

T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electron. J. Statist., 10(1):1–59, 2016. .

R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. Oxford Statistical Science Series, pages 115–137, 2003.

R. Dahlhaus, M. Eichler, and J. Sandkühler. Identification of synaptic connections in neural ensembles by graphical models. Journal of neuroscience methods, 77(1):93–107, 1997.

M. Eichler. A frequency-domain based test for non-correlation between stationary time series. Metrika, 65(2):133–157, 2007.

L. Erdős, H.-T. Yau, and J. Yin. Bulk universality for generalized wigner matrices. Probability Theory and Related Fields, pages 1–67, 2012.

C. Euan, H. Ombao, and J. Ortega. The hierarchical spectral merger algorithm: a new time series clustering procedure. arXiv preprint arXiv:1609.08569, 2016.

M. K. S. Faradonbeh, A. Tewari, and G. Michailidis. Finite time identification in unstable linear systems. Automatica, 96:342–353, 2018.

M. Fiecas and H. Ombao. Modeling the evolution of dynamic brain processes during an associative learning experiment. Journal of the American Statistical Association, 111 (516):1440–1453, 2016.

M. Fiecas and R. von Sachs. Data-driven shrinkage of the spectral density matrix of a high-dimensional time series. Electron. J. Statist., 8(2):2975–3003, 2014. . URL https://doi.org/10.1214/14-EJS977.

B. Fischl and A. M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences, 97(20): 11050–11055, 2000.

P. Fransson and G. Marrelec. The precuneus/posterior cingulate cortex plays a piv-

otal role in the default mode network: Evidence from a partial correlation network analysis. NeuroImage, 42(3):1178 – 1184, 2008. ISSN 1053-8119. . URL http://www.sciencedirect.com/science/article/pii/S1053811908007283.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, pages 424–438, 1969.

A. Jung. Learning the conditional independence structure of stationary time series: A multitask learning approach. IEEE Transactions on Signal Processing, 63(21):5677–5690, 2015.

A. Jung, G. Hannak, and N. Goertz. Graphical lasso based model selection for time series. IEEE Signal Processing Letters, 22(10):1781–1785, 2015.

A. Kuceyeski, K. W. Jamison, J. Owen, A. Raj, and P. Mukherjee. Functional rerouting via the structural connectome is associated with better recovery after mild tbi. bioRxiv, 2018. . URL https://www.biorxiv.org/content/early/2018/05/18/320515.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. Journal of multivariate analysis, 88(2):365–411, 2004.

H. C. Ombao, J. A. Raz, R. L. Strawderman, and R. von Sachs. A simple generalised crossvalidation method of span selection for periodogram smoothing. Biometrika, 88 (4):1186–1192, 2001.

M. Rosenblatt. Stationary sequences and random fields. Springer, 1985.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association, 104(485):177–186, 2009.

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. Electronic Communications in Probability, 18, 2013.

H. Shu and B. Nan. Estimation of large covariance and precision matrices from temporally dependent observations. arXiv preprint arXiv:1412.5059, 2014.

S. D. Utevsky AV and H. SA. Precuneus is a functional core of the default-mode network. Journal of Neuroscience, 34(3):932–40, 2014.

S. van de Geer. lecture notes on sparsity, 2016.

K. C. Wong and A. Tewari. Lasso guarantees for beta -mixing heavy tailed time series. arXiv preprint arXiv:1708.01505, 2017.

W. B. Wu and P. Zaffaroni. Uniform convergence of multivariate spectral density estimates. arXiv preprint arXiv:1505.03659, 2015.

X.-N. Zuo, C. Kelly, A. Di Martino, M. Mennes, D. S. Margulies, S. Bangaru, R. Grzadzinski, A. C. Evans, Y.-F. Zang, F. X. Castellanos, and M. 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. .

Yiming Sun and Sumanta Basu
Department of Statistical Science
Cornell University
Ithaca NY 14850
E-mail: ys784@cornell.edu
sumbose@cornell.edu

Yige Li
Department of Epidemiology,
Harvard T. H. Chan School of Public Health
Harvard University
Boston MA 02138
E-mail: yigeli@hsph.harvard.edu

Amy Kuceyeski
Department of Radiology
Weill Cornell Medical College
New York NY 10065
E-mail: amk2012@med.cornell.edu