Spectral Estimation Using Multitaper Whittle Methods with a Lasso Penalty

Shuhan Tang, Peter F. Craigmile, and Yunzhang Zhu

Last updated July 25, 2019

***Abstract*—Spectral estimation provides key insights into the frequency domain characteristics of a time series. Naive nonparametric estimates of the spectral density, such as the periodogram, are inconsistent, and the more advanced lag window or multitaper estimators are often still too noisy. We propose an L_1 penalized quasi-likelihood Whittle framework based on multitaper spectral estimates which performs semiparametric spectral estimation for regularly sampled univariate stationary time series. Our new approach circumvents the problematic Gaussianity assumption required by least square approaches and achieves sparsity for a wide variety of basis functions. We present an alternating direction method of multipliers (ADMM) algorithm to efficiently solve the optimization problem, and develop universal threshold and generalized information criterion (GIC) strategies for efficient tuning parameter selection that outperform cross-validation methods. Theoretically, a fast convergence rate for the proposed spectral estimator is established. We demonstrate the utility of our methodology on simulated series and to the spectral analysis of electroencephalogram (EEG) data.**

***Index Terms*—Alternating direction method of multipliers (ADMM) algorithm; basis expansion; multitaper spectral estimates; wavelets.**

I. Introduction

ESTIMATING the spectral density function (SDF) or spectrum of a series collected over time is an important tool in time series analysis and signal processing. It is used in many fields such as astronomy, cognitive science, earth sciences, electrical engineering, and finance. Examining the SDF allows us to explore periodicities in the data (e.g., [1, ch. 10]), provides an alternative way to analyze and estimate the covariance structure of stationary time series (e.g., [1, ch. 4]), and can also be used to understand the effect of preprocessing a time series (e.g., [2]).

There are many nonparametric estimators of the SDF of a univariate stationary time series. These include the periodogram, direct spectral estimators, lag window and overlapping segment averaging spectral estimators, and multitaper (MT) spectral estimators. (See [1] for a complete review.) While many of these estimators are developed to provide an adequate tradeoff between bias and variance, often these nonparametric estimates are still too noisy when a stable estimate of the SDF is required. An alternative strategy is to use a parametric approach, however model misspecification caused by considering a limited class of models for the SDF, can compromise estimation (see [1], ch.9, and references therein).

A popular alternative approach is to consider a semiparametric model for the SDF, in which the log SDF is expressed in terms of a truncated basis expansion, where the number of basis functions are allowed to increase with the sample size. The statistical problem then becomes how to enforce sparsity by selecting the basis functions and estimating the model parameters so that we adequately estimate the SDF, but also have computational efficiency as the sample size is increased. Gao [3] [4], Moulin [5] and Walden *et al.* [6] enforce sparsity using a penalized least square (LS) approach for estimating the log SDF with wavelet soft thresholding. In terms of computational complexity, wavelet thresholding methods are typically O(N), for a time series of N regularly sampled values.

A number of approaches enforce smoothness of the SDF via an L_2 penalty: Cogburn and Davis [7], Wahba and Wold [8] and Wahba [9] use penalized LS, and Pawitan and O'Sullivan [10] uses a penalized Whittle method. To enforce sparsity, some of these L_2 methods of smoothing splines also use model selection, often in combination with cross-validation, to select the basis functions that are used to model the SDF. Alternatively, one can implement methods such as [11] to enforce sparsity on the basis expansion directly. (On a related topic, smoothness of spectral estimation can also be tuned with high-resolution approaches introduced in [12], [13] and using extended frameworks based on so-called beta and tau divergence families, such as [14]–[18]; see [19] for a general review of such divergences.)

Our method is also motivated by the need to enforce sparsity while adequately estimating the SDF. In addition, we seek computational efficiency as we increase the sample size. We develop a quasi-likelihood method for estimating SDFs using a Whittle likelihood [20] based on MT spectral estimates. A quasi-likelihood function [21] [22, ch. 9] has similar statistical properties to that of the log likelihood, and can be used for statistical inference, but does not have to match exactly to the log of the joint probability density function of the data. MT estimates [23] [1] provide a good compromise between bias and variance and can yield more efficient estimates of the SDF [6]. We demonstrate that the addition of a Whittle likelihood method [20] improves estimation over traditional LS approaches.

We use a lasso penalty [24] to enforce sparsity, deriving two strategies to optimally select the tuning parameter that is key to obtaining estimates of the SDF with low integrated root mean

S. Tang, P. F. Craigmile, Y. Zhu are with the Department of Statistics, The Ohio State University, Columbus, OH, 43210 USA (e-mail: tang.723@osu.edu).

Manuscript received December 12, 2018; revised June 27, 2019.

squared error (IRMSE): "universal threshold" and generalized information criterion (GIC)-based methods. Neither method compromises on computational or statistical efficiency by requiring the use of cross-validation to select the tuning parameter. Theoretically, we derive the rate of convergence for our proposed spectral estimator under some technical conditions on the model sparsity and the MT spectral estimator.

We introduce a computationally efficient method to estimate the parameters in our model using the alternating direction method of multipliers (ADMM) algorithm. To reach an ϵ-optimal solution with a time series of length N, our method is $O(N\epsilon^{-1})$ when using wavelet bases and $O(N^3 + \epsilon^{-1}N^2)$ for general bases. Although computationally more challenging, our method can be applied to SDF estimation using any collection of basis functions and, as mentioned above, outperforms LS-based methods such as wavelet thresholding in terms of estimation quality.

The rest of the paper is organized as follows. Section II presents models for SDFs in terms of basis methods. In Section III, we introduce the multitaper spectral estimator that we use in our penalized Whittle estimation method in Section IV. The ADMM algorithm is described in Section V and Section VI outlines two approaches for tuning parameter selection. We derive the rate of convergence for the proposed L_1 penalized MT-Whittle likelihood estimator in Section VII. Our methods are evaluated using Monte Carlo simulations in Section VIII and we perform a spectral analysis of electroencephalogram (EEG) data in Section IX. We close with some remarks in Section X. Proofs and further details of the ADMM algorithm are provided in the Appendix.

II. Basis models for SDFs

Let $\{X_t : t \in \mathbb{Z}\}$ be a univariate real-valued stationary process collected at sampling interval $\Delta > 0$. Without loss of generality assume $\Delta = 1$. Let $\gamma(h) = \mathrm{cov}(X_t, X_{t+h})$, $h \in \mathbb{Z}$, denote the (stationary) *autocovariance function* (ACVF) of $\{X_t\}$ and assume that the ACVF is absolutely summable: $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$. Then the *spectral density function (SDF)* $S(f)$ for a frequency $|f| < 1/2$ exists and is defined as the Fourier transform pair of the ACVF:

$$S(f) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i2\pi fh}, \tag{1}$$

$$\text{with} \quad \gamma(h) = \int_{-1/2}^{1/2} e^{i2\pi fh} S(f) df. \tag{2}$$

The SDF is a non-negative, even, and real-valued function and decomposes the variance of the time series $\{X_t\}$: from (2) with $h = 0$, $\mathrm{var}(X_t) = \int_{-1/2}^{1/2} S(f) df$. See [1] and [25] for further properties of the SDF.

Basis methods for the estimation of a SDF typically involve assuming that the log SDF can be expanded in terms of a set of p basis functions $\{\phi_l(f) : l = 1, \ldots, p\}$ (e.g., [9]). For each frequency f letting $\boldsymbol{\phi}(f) = (\phi_1(f), \ldots, \phi_p(f))^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, we suppose that

$$\log S(f) = \sum_{l=1}^{p} \phi_l(f)\beta_l = \boldsymbol{\phi}^T(f)\boldsymbol{\beta}. \tag{3}$$

There are a variety of options for families of basis functions $\boldsymbol{\phi}(f)$ that can be used for spectral estimation. For example, polynomial and Fourier bases can be used to capture global patterns [26], smoothing splines allow for local and smooth patterns in the SDF [7], [8], and wavelet bases model local behaviour, while capturing second order effects such as peaks, troughs, and cusps [3]–[6]. When the SDF is spatially inhomogeneous, spatially adaptive bases, such as wavelets, have theoretical optimality properties (see, e.g., [27]–[29]). We can also combine families of basis functions to form *dictionaries*.

In Section IV we propose our estimator of the coefficient vector $\boldsymbol{\beta}$, which is based on multitaper spectral estimates of time series data that we define in the next section. (In Section VII we provide assumptions on the basis representation (3) so that we can asymptotically recover the true log SDF.)

III. Multitaper spectral estimation

Suppose we observe N observations, $\boldsymbol{X} = (X_1, \ldots, X_N)^T$, from the stationary process $\{X_t\}$. Then a *multitaper (MT)* or multiple taper spectral estimate [23], is an average of a number of tapered spectral estimates. Specifically, let $\{h_{k,t} : k = 1, \ldots, K,\ t = 1, \ldots, N\}$ denote K orthonormal data tapers; i.e., $\sum_t h_{k,t}^2 = 1$ and $\sum_t h_{k,t}h_{k',t} = 0$ for $k \neq k'$. Then the standard MT spectral estimator of the SDF $\widehat{S}^{(\mathrm{mt})}(f)$, is the average of the K eigenspectra,

$$\widehat{S}^{(\mathrm{mt})}(f) \quad = \quad \frac{1}{K}\sum_{k=1}^{K} \widehat{S}_k^{(\mathrm{mt})}(f), \tag{4}$$

where the kth ($k = 1, \ldots, K$) eigenspectrum is defined by $\widehat{S}_k^{(\mathrm{mt})}(f) = |J_k(f)|^2$, with

$$J_k(f) = \sum_{t=1}^{N} h_{k,t} X_t \exp(-i2\pi ft).$$

Different tapers induce different statistical properties for the MT estimator. Discrete prolate spheroidal sequences (DPSS) and sine tapers are most commonly used [1]. DPSS tapers are designed to reduce the sidelobes in the spectral estimate. They solve the time-frequency concentration problem in which we find the time limited sequence which has most of its energy concentrated in a specified frequency band [1, ch. 8]. We use the easily calculated sine tapers [30],

$$h_{k,t} = \left(\frac{2}{N+1}\right)^{1/2} \sin\left(\frac{(k+1)\pi t}{N+1}\right), \qquad k = 1, \ldots, K,\ t = 1, \ldots, N,$$

which are designed to reduce the smoothing bias, at a compromise to sidelobe reduction. For a given K the sine tapers are concentrated in the frequency band $[-W, W]$ for half bandwidth $W = (K+1)/(2(N+1))$. As K increases we lose resolution but decrease the variance of the MT spectral estimator. Walden [31] demonstrates other classes of tapers, showing that Welch's weighted overlapped segment averaging (WOSA) estimator [32] and lag window estimators can be reformulated as MT estimators. Walden [31] also shows under known conditions that MT estimators are consistent if we let the number of tapers K increase with the sample size N.

IV. Penalized Whittle multitaper estimation

The *Whittle likelihood* [20] is widely used to approximate log-likelihoods for stationary and some nonstationary time series using the naive spectral estimator known as the periodogram. (See [33] for an introduction to the Whittle likelihood.) For stationary Gaussian processes, using results for symmetric Toeplitz matrices (e.g., [34]), the Whittle likelihood is asymptotically equal to two times the negative log-likelihood [33]. For certain non-Gaussian cases the Whittle likelihood is still a valid approximation [35]. This is because the Whittle likelihood is equal to the joint asymptotic distribution of the periodogram, assuming independence of the spectral estimates over the frequencies that the spectral estimates are evaluated at [10]. Since the periodogram often exhibit poor statistical properties such as bias due to leakage and inconsistency in estimation [1], we develop a Whittle likelihood based on MT spectral estimates. We explain why this MT-Whittle likelihood is still reasonable for parameter estimation.

Using the series $\boldsymbol{X}$, we evaluate the MT spectral estimates $\widehat{S}^{(\text{mt})}(f_j)$ on the set of $M = \lceil N/2 \rceil - 1$ non-zero, non-Nyquist (i.e., not equal to $1/2$) Fourier frequencies defined by

$$\left\{ f_j = \frac{j}{N} : j = 1, \ldots, M \right\}. \tag{5}$$

Choosing a specific basis representation, let $\boldsymbol{\Phi}$ denote the $M \times p$ design matrix of basis functions evaluated at these Fourier frequencies with row $j = 1, \ldots, M$ equal to $\boldsymbol{\phi}^T(f_j)$. The vector of log SDFs

$$\boldsymbol{\zeta} = (\log S(f_1), \ldots, \log S(f_M))^{\mathrm{T}}$$

evaluated at these frequencies is $\boldsymbol{\zeta} = \boldsymbol{\Phi}\boldsymbol{\beta}$, by (3).

Definition 1. *Using MT spectral estimators, a quasi-likelihood function with expression*

$$l_W(\boldsymbol{\Phi}\boldsymbol{\beta}) = \sum_{j=1}^{M} \left\{ \log S(f_j) + \frac{\widehat{S}^{(mt)}(f_j)}{S(f_j)} \right\} \tag{6}$$

is called the ***MT-Whittle likelihood*** *function.*

The MT-Whittle likelihood has the same functional form as the Whittle likelihood based on the periodogram, except we replace the periodogram by the MT estimator. It can be easily shown that we obtain the usual Whittle likelihood using a single $K = 1$ rectangular taper defined by $h_{1,t} = 1/\sqrt{N}$ for all $t = 1, \ldots, N$ in the MT spectral estimator; the tapered Whittle likelihood [36] is also a special case.

To see why the MT-Whittle likelihood is a reasonable quasi-likelihood, we present the following two propositions.

Proposition 1. *[31, Section 3.3] Suppose that $\{X_t : t \in \mathbb{Z}\}$ is strictly stationary with all moments existing such that*

$$\sum_{\tau_1, \ldots, \tau_{l-1} = -\infty}^{\infty} |\mathrm{cum}(X_{t+\tau_1}, \ldots, X_{t+\tau_{l-1}}, X_t)| < \infty,$$

for $l = 2, 3, \ldots$, where $\mathrm{cum}(X_{t_1}, \ldots, X_{t_l})$ denotes the joint cumulant function of order l (see, e.g., [25], sec. 2.3). Also for each N, let $\{h_{k,t} : k = 1, \ldots, K,\ t = 1, \ldots, N\}$ be a set of K orthonormal sine or DPSS data tapers. Then

$$\widehat{S}^{(mt)}(f) \to_d S(f) \frac{\chi^2_{2K}}{2K}, \ \textit{for } 0 < f < 1/2, \ \textit{as } N \to \infty,$$

where χ^2_{2K} denotes a chisquared random variable (RV) with $2K$ degrees of freedom.

Walden [31] provides a proof of this result in the multivariate case – our result has been simplified to the univariate case. This proposition tells us that at each frequency f, our MT-spectral estimators have a valid asymptotic scaled chisquared distribution that depends on the true underlying SDF.

In general MT-spectral estimators are correlated over frequencies: by tapering we reduce the sidelobes to decrease the bias, but increase the effective bandwidth of the spectral estimator, which then increases the correlation. Thomson [23] showed that with a locally slowly varying spectrum, for $0 < f < f' < 1/2$, with f close to f',

$$\mathrm{Cov}\{\widehat{S}^{(\text{mt})}(f), \widehat{S}^{(\text{mt})}(f')\} \approx \frac{S^2(f)}{K^2} \sum_{k=1}^{K} \sum_{l=1}^{K} \left| \sum_{t=1}^{N} h_{t,k} h_{t,l} e^{i2\pi(f'-f)t} \right|^2. \tag{7}$$

However, the next result shows that our MT-Whittle likelihood (6) can be reinterpreted as a gamma quasi-likelihood, which ignores these correlations between frequencies.

Proposition 2. *The MT-Whittle likelihood* (6) *corresponds to a gamma quasi-likelihood assuming the asymptotic distribution of Proposition 1 at the Fourier frequencies* (5)*, and assuming independence between the Fourier frequencies.*

The proof is given in Appendix A. In statistical science, it is common to estimate certain parameters of a model (e.g., parameters related to the mean of the distribution, such as $\boldsymbol{\beta}$) using a simplified model compared to the *true* statistical model (see, e.g., [37]). The simplified model is known as the *working model*. Proposition 2 tells us that when we write down a MT-Whittle likelihood using MT spectral estimates, we define a working model for the spectral estimates, assuming independence across the Fourier frequencies. An extensive literature on estimating $\boldsymbol{\beta}$ in this setting (e.g., [37]–[39]) indicates that we can consistently estimate the model parameters $\boldsymbol{\beta}$ under this working model, even when independence does not truly hold. Thus Proposition 2 allows us to introduce the following penalized quasi-likelihood framework for valid statistical inference for $\boldsymbol{\beta}$.

We incorporate a Lasso-type penalty with the MT-Whittle likelihood (6) to enforce sparsity as the number of basis functions, p, increases with sample size N. In Sections VIII and IX, we let $p = M + 1 = \lceil N/2 \rceil$, where M is the number of non-zero, non-Nyquist, Fourier frequencies, although depending on the form of the penalty, p could actually be larger than N. Our optimization problem is then

$$\min_{\boldsymbol{\beta}} \ l_W(\boldsymbol{\Phi}\boldsymbol{\beta}) + \sum_{l=1}^{p} \lambda_l |\beta_l|, \tag{8}$$

where $\lambda_l \geq 0$ denotes the *tuning parameter* for coefficient β_l, with $l = 1, \ldots, p$. The penalty $\sum_{l=1}^{p} \lambda_l |\beta_l|$ serves the function

of reducing the noise effect on the coefficient estimates, which in general shrinks all regression coefficients' magnitudes. We let $\widehat{\boldsymbol{\beta}}$ denote the solution to (8), and $\widehat{S}(f) = \exp(\boldsymbol{\phi}^T(f)\widehat{\boldsymbol{\beta}})$ denote the resulting estimator of the SDF.

Our L_1 penalized method performs simultaneous parameter estimation and feature selection; e.g., [40]. With some of the coefficient estimates shrunk exactly to zero, L_1 penalized method identifies a small subset of predictors, which is favored when the true model is considered to be sparse. (See Section X for a discussion of L_2 penalized methods that also carry out feature selection.)

In the next section an algorithm to solve (8) is presented. Tuning parameter selection is discussed in Section VI.

V. Computation: an ADMM algorithm

Sardy *et al.* [29] suggested using the interior point algorithm to optimize a penalized non-Gaussian likelihood incorporating wavelet bases expansions, and provided the explicit steps of the algorithm for the Poisson case. No explicit algorithm was given for the gamma distribution that makes up our MT-Whittle likelihood or for including other basis functions. We use instead the *alternating direction method of multipliers* (ADMM) algorithm, which is an approach for solving large-scale nonsmooth convex optimization problems, and was introduced by Glowinski and Marroco [41] and Gabay and Mercier [42] [see [43] for further review]. The ADMM algorithm has the advantage that the per-iteration cost is often much lower than that of the interior point algorithm, which makes it an attractive choice when solutions of medium accuracy are sufficient, such as parameter estimation problems.

The ADMM algorithm proceeds by first introducing new equality constraints to decouple the two terms in the objective function (8). Specifically, by introducing the equality constraints

$$\boldsymbol{\zeta} \;=\; \boldsymbol{\Phi}\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\eta} \;=\; \boldsymbol{\beta}\,, \tag{9}$$

the original problem (8) is equivalent to

$$\min_{\boldsymbol{\zeta},\boldsymbol{\eta},\boldsymbol{\beta}} \quad l_W(\boldsymbol{\zeta}) + \sum_{l=1}^{p} \lambda_l |\eta_l|$$
$$\text{subject to } \boldsymbol{\Phi}\boldsymbol{\beta} = \boldsymbol{\zeta} \text{ and } \boldsymbol{\beta} = \boldsymbol{\eta}\,,$$

where $\boldsymbol{\zeta} \in \mathbb{R}^M$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_p)^T \in \mathbb{R}^p$. The ADMM algorithm solves the above problem by alternately updating the *primal variables* $(\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\beta})$ and the associated *dual variables* $(\boldsymbol{u}_1, \boldsymbol{u}_2)$. The $(n+1)$-th step of the algorithm is:

$$\boldsymbol{\beta}^{(n+1)} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{I}_p)^{-1}\Big\{\boldsymbol{\Phi}^T(\boldsymbol{\zeta}^{(n)} - \boldsymbol{u}_1^{(n)}) + \boldsymbol{\eta}^{(n)} - \boldsymbol{u}_2^{(n)}\Big\};$$

$$\zeta_j^{(n+1)} = \arg\min_{\zeta_j} \Big\{\zeta_j + \widehat{S}^{(\mathrm{mt})}(f_j)\exp(-\zeta_j) + \frac{\rho}{2}\{\boldsymbol{\phi}^T(f_j)\boldsymbol{\beta}^{(n+1)} - \zeta_j + u_{1j}^{(n)}\}^2\Big\}, \quad j = 1, \ldots, M;$$

$$\eta_l^{(n+1)} = \mathrm{ST}\left(\beta_l^{(n+1)} + u_{2l}^{(n)}, \frac{\lambda_l}{\rho}\right), \; l = 1, \ldots, p;$$

$$\boldsymbol{u}_1^{(n+1)} = \boldsymbol{u}_1^{(n)} + \boldsymbol{\Phi}\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\zeta}^{(n+1)};$$

$$\boldsymbol{u}_2^{(n+1)} = \boldsymbol{u}_2^{(n)} + \boldsymbol{\beta}^{(n+1)} - \boldsymbol{\eta}^{(n+1)}.$$

In this algorithm, u_{1j} and u_{2j} denote respectively the j-th components of $\boldsymbol{u_1}$ and $\boldsymbol{u}_2$, $\rho > 0$ is a positive *penalty parameter*, and $\mathrm{ST}(x, a) = \mathrm{Sign}(x)\max(|x| - a, 0)$ denotes the *soft-thresholding function* with threshold $a \geq 0$. For any $\rho > 0$, the iterates $\boldsymbol{\beta}^{(n)}$ have been shown to converge to the global solution of the original optimization problem (8) under some mild conditions on the objective function [43]. Moreover, following Boyd *et al.* [43] the algorithm is terminated when both the primal and dual residuals are smaller than prespecified precision parameters. We use a second equality constraint $\boldsymbol{\beta} = \boldsymbol{\eta}$ to avoid the need to solve a lasso problem at each iteration of the ADMM in the cases that the basis functions are not orthogonal. A detailed derivation of the ADMM updates and the stopping criterion are presented in Appendix B.

A close inspection of the above ADMM updating scheme shows that it is ideally suited for our optimization problem (8) because all the subproblems can be solved efficiently with any basis expansion of $\log S(f)$, especially when wavelet basis functions are used. First, note that the $\boldsymbol{\beta}$-update is a linear system for general bases and can be solved in linear time when using wavelets, because $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is a diagonal matrix for a wavelet basis, and all matrix-vector multiplications involving $\boldsymbol{\Phi}$ or $\boldsymbol{\Phi}^T$ can be carried out in linear time using the cascade algorithm (e.g., [44, Section 4]). Second, the $\boldsymbol{\zeta}$-update decomposes into M independent univariate optimization problems, each of which can be solved efficiently using a bisection method (Recall that M is the number of Fourier frequencies used to formulate the MT-Whittle likelihood). For practical purposes, it is often assumed that the number of bisection iterations is a constant. Consequently, the per-iteration cost of the proposed ADMM algorithm is $O(M)$ if wavelet basis functions are used. For general basis, we perform one Cholesky factorization of $\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{I}_p$ upfront for the $\boldsymbol{\beta}$-update, which is $O(M^3)$. Given the Cholesky factorization, all subsequent updates have complexity no greater than $O(M^2)$. Therefore, the per-iteration cost when using general basis is $O(M^2)$. Further acceleration is possible using parallel computing techniques.

With regard to the rate of convergence, the total number of ADMM iterations required to reach an ϵ-optimal solution is $O(\epsilon^{-1})$ [45]. Consequently, the overall computational complexity for the ADMM algorithm to reach an ϵ-optimal solution is $O(\epsilon^{-1}M)$ for a wavelet basis, and $O(\epsilon^{-1}M^2 + M^3)$ for a general basis.

VI. Tuning parameter selection

The goodness of fit of the model is determined by the selection of the tuning parameters $\lambda_1, \lambda_2, \ldots, \lambda_p$ in (8). We assume a common tuning parameter λ for non-intercept bases and $\lambda_1 = 0$ for the intercept β_1. Traditional methods of selection is based on a Gaussian likelihood assumption. Extending to the non-Gaussian case, Sardy *et al.* [29] deduced the rules that a tuning parameter should satisfy with a concave and differentiable non-Gaussian log-likelihood, but only derived an explicit solution for the Poisson case when the interior point algorithm is used. No explicit solution is given for our case. We develop universal threshold and generalized information

criterion (GIC) [46] approaches for tuning parameter selection with a MT-Whittle likelihood.

A. Scale-calibrated Universal Threshold

Although the universal threshold was originally introduced for wavelet thresholding by Donoho and Johnstone [27] in the same paper they state that their "results apply equally well in orthogonal regression". We extend the idea of a universal threshold to our penalized MT-Whittle likelihood problem.

Starting with the unpenalized MT-Whittle likelihood with an orthonormal basis, under the assumptions of Proposition 2 we show in Appendix C that the maximum quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}^W$ has the asymptotic distribution $\widehat{\boldsymbol{\beta}}^W \sim N(\boldsymbol{\beta}, \boldsymbol{I}_p/K)$. Next we use a similar argument as the *universal threshold* derived by Donoho and Johnstone [27]. Let $\boldsymbol{\xi}^W = \widehat{\boldsymbol{\beta}}^W - \boldsymbol{\beta}$ denote the noise component of the unpenalized MT-Whittle estimate. Since each component of $\boldsymbol{\xi}^W$, ξ_l^W, are independent N(0, $1/K$) RVs,

$$P\left(\max_l |\xi_l^W| > \sqrt{1/K}\sqrt{2\log p}\right) \to 0, \quad \text{as } p \to \infty.$$

Consulting the ADMM algorithm given in Section V we note that the tuning parameter only participates in the soft thresholding step of the algorithm. Choosing a *scale-calibrated universal threshold* $\lambda^{univ} = \sqrt{1/K}\sqrt{2\log p}$ as the tuning parameter λ, the noise components of $\boldsymbol{\beta}$ will be shrunk to zero with high probability, leaving only those components that represent the true underlying SDF. This mimics the universal threshold of [27], but varies in two distinct ways: (i) we need no estimate of the noise variance: the variance (scale) is determined by the number of tapers K in the MT spectral estimator – as K increases this variance decreases; (ii) our choice of λ depends on the initial number of basis functions p, and not the sample size M as in [27].

B. Generalized information criterion

For the L_1 penalized MT-Whittle likelihood problem (8), a *generalized information criterion (GIC)* finds

$$\widehat{\lambda} = \arg\min_{\lambda} \left\{2K\, l_W(\boldsymbol{\Phi}\boldsymbol{\beta}_\lambda) + c_M|p_\lambda|\right\}, \tag{10}$$

where $\boldsymbol{\beta}_\lambda$ is the optimizer of L_1 penalized MT-Whittle likelihood with tuning parameter λ. In the penalty term for model complexity, $|p_\lambda|$ denotes the number of non-zero elements in $\boldsymbol{\beta}_\lambda$, and c_M is the penalty parameter.

The *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC) are two special cases of GIC with $c_M = 2$ and $c_M = \log M$ respectively. According to Fan and Tang [46], AIC has similar performance as cross-validation and typically overfits the statistical model and BIC only consistently selects the true model when the dimension of predictor space is fixed.

When the number of predictors, p, increases exponentially as the sample size M increases (i.e., $\log p = O(M^\kappa)$ for some $\kappa > 0$), Fan and Tang [46] suggest the choice of penalty parameter $c_M = (\log\log M)(\log p)$.

VII. Theory

We now derive the rate of convergence for the proposed L_1 penalized MT-Whittle likelihood estimator. This allows us to study consistency of spectrum estimation.

The main result is based on an extension to the arguments used to derive the so-called *fast rate* for L_1 penalized methods (see, e.g., [47], [48]). There are two main challenges for our estimator. First, typical theoretical results for the L_1 penalized problem assume independence among samples, whereas the MT-Whittle likelihood involves sums of dependent RVs. Moreover, to the best of our knowledge, all existing theory for L_1 penalized generalized linear models assumes that the canonical link is used (see, e.g., [47]). In contrast, we deal with a situation where the link function is not the canonical link, which makes the log-partition function dependent on random quantities. These would make our theoretical analysis more challenging, and the technical conditions more complicated.

For each $j = 1, \ldots, M$, let

$$R_j = \widehat{S}^{(\text{mt})}(f_j)/S(f_j), \tag{11}$$

where, by Proposition 1, $R_j - 1$ follows the distribution as $\chi^2_{2K}/2K - 1$ RVs asymptotically. We further impose the following assumptions.

Assumption (A1) Assume that

$$\log S(f_j) = \langle \phi(f_j), \beta^0 \rangle, \quad j = 1, \ldots, M, \tag{12}$$

for some sparse vector $\beta^0 \in \mathbb{R}^p$ and basis functions $\phi(f_j)$ satisfying $\|\phi(f)\|_\infty \le B$ for $f \in [-1/2, 1/2]$ and some constant B.

Assumption (A2) (Compatibility condition) Let $S = \{l : \beta_l^0 \neq 0\}$ and $s_0 = |S|$. Assume that for any $v \in \mathbb{R}^p$ with $\|v_{S^c}\|_1 \le 3\|v_S\|_1$ that

$$\frac{1}{M}\sum_{j=1}^{M} R_j \left\{\exp(v^\top \phi(f_j)) - v^\top \phi(f_j) - 1\right\} \geq \min\left\{\frac{c_0}{s_0}\|v_S\|_1^2, c_1 M^{\gamma-\frac{1}{2}}\|v_S\|_1\right\}, \tag{13}$$

with probability tending to 1 as $M \to \infty$ for some constants $c_0 > 0$, $c_1 > 0$, $0 < \gamma \le \frac{1}{2}$ and where R_j is defined by (11).

Assumption (A1) is the typical sparsity condition imposed for theoretical analysis of penalized procedures using sparsity-inducing penalty, while Assumption (A2) is similar to the so-called *compatibility condition* required for one type of theoretical analysis for lasso estimator [48]. Note that since we are dealing with non-canonical link functions, the compatibility condition does involve random quantities, which makes it more difficult to verify in practice.

Theorem 1. *Under Assumptions (A1) and (A2) and on the event that*

$$\left\|\sum_{j=1}^{M}(R_j - 1)\,\phi(f_j)\right\|_\infty < \frac{c_1}{3}M^{\gamma+\frac{1}{2}}, \tag{14}$$

we have that

$$\left\|\widehat{\beta}(\lambda) - \beta^0\right\|_1 \le \frac{3\lambda s_0}{2c_0 M} \tag{15}$$

for any λ satisfying

$$2\left\|\sum_{j=1}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty \le \lambda < \frac{2}{3}c_1 M^{\gamma+\frac{1}{2}}.$$

In particular, when $\lambda = 2\left\|\sum_{j=1}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty$, we have, on the event (14)*, that*

$$\left\|\widehat{\beta}(\lambda)-\beta^0\right\|_1 \le \frac{3s_0}{c_0 M}\left\|\sum_{j=1}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty, \qquad (16)$$

and

$$\sup_{f\in[-\frac{1}{2},\frac{1}{2}]}|\log\widehat{S}(f)-\log S(f)| \le \frac{3Bs_0}{c_0 M}\left\|\sum_{j=1}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty, \qquad (17)$$

where $\log\widehat{S}(f)$ is the L_1 penalized MT-Whittle estimator of the log SDF.

Some remarks are in order. First, unlike existing theoretical analysis for L_1 penalized methods, we do not impose any independence assumptions on the "sample": $\{\widehat{S}^{(\mathrm{mt})}(f_j); j = 1,\ldots,M\}$. In this sense, the above result is a deterministic result that holds under the event (14). To verify that the event (14) happens with probability tending to 1, we make the following heuristic argument. We assume that the 2-norm of the basis functions is equal to M. Then

$$\begin{aligned}
&\frac{1}{\sqrt{M}}\left\|\sum_{j=1}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty \\
&\le \frac{1}{\sqrt{M}}\max_{1\le j\le M}|R_j-1|\max_{1\le k\le p}\sum_{j=1}^{M}|\phi_k(f_j)| \\
&\le \max_{1\le j\le M}|R_j-1|\frac{1}{\sqrt{M}}\sqrt{\max_{1\le k\le p}\sum_{j=1}^{M}\phi_k^2(f_j)} \\
&= \max_{1\le j\le M}|R_j-1|. \qquad (18)
\end{aligned}$$

Moreover, by Proposition 1, we have that $R_j - 1$ behaves like $\chi^2_{2K}/2K - 1$ RVs asymptotically. Making this distributional assumption for all $j = 1,\ldots,M$, we have

$$\begin{aligned}
&P\left(\max_{1\le j\le M}|R_j-1| > C\log(M)\right) \\
&\le \sum_{j=1}^{M}P(|R_j-1|>C\log(M)) \\
&\le M\max_{1\le j\le M}P(|R_j-1|>C\log(M)) \\
&\le M\exp(-C\log(M)/4) = M^{-(C/4-1)} \to 0,
\end{aligned}$$

as $M\to\infty$ for any constant $C > 4$ and $M \ge 2$ (The last inequality uses Lemma 1 of [49]). Thus we conjecture that asymptotically, $\max_{1\le j\le M}|R_j - 1| = O_p(\log M)$. and hence in combination with (18), we get

$$\frac{1}{\sqrt{M}}\left\|\sum_{j=1s}^{M}(R_j-1)\,\phi(f_j)\right\|_\infty \le O_p(\log(M)). \qquad (19)$$

In view of this, condition (14) holds asymptotically for any $\gamma > 0$. This will lead to the following rate of convergence for $\widehat{\beta}$ and the log SDF:

$$\left\|\widehat{\beta}(\lambda)-\beta^0\right\|_1 = O_p\left(s_0\frac{\log(M)}{\sqrt{M}}\right),$$
$$\sup_{f\in[-\frac{1}{2},\frac{1}{2}]}|\log\widehat{S}(f)-\log S(f)| = O_p\left(s_0\frac{\log(M)}{\sqrt{M}}\right).$$

Note that if we further assume that s_0 is quite small, that is, the true log SDF has a sparse basis representation, a parametric rate $M^{-1/2}$ can be achieved (up to a log factor). This is in contrast to the slower nonparametric rate for typical one-dimensional nonparametric regression or density estimation problems (see, e.g., [50]). In summary, our theory suggests that by exploring sparsity, if it is indeed present in the signal, a significant improvement in estimation efficiency can be achieved using the proposed method.

VIII. Simulations

We use simulations to evaluate our L_1 penalized MT-Whittle method as compared to the commonly used L_1 penalized LS method. We also investigate the effect of selecting the tuning parameter and assess our theoretical rate presented after the statement of Theorem 1. We use the following processes:

1) AR(2) process: $X_t = \varphi_{1,1}X_{t-1} + \varphi_{1,2}X_{t-2} + \varepsilon_t$ with $\varphi_{1,1} = 0.97\sqrt{2}$, $\varphi_{1,2} = -0.97^2$;
2) AR(4) process: $X_t = \varphi_{2,1}X_{t-1} + \varphi_{2,2}X_{t-2} + \varphi_{2,3}X_{t-3} + \varphi_{2,4}X_{t-4} + \varepsilon_t$ with $\varphi_{2,1} = 2.7607$, $\varphi_{2,2} = -3.8106$, $\varphi_{2,3} = 2.6535$, $\varphi_{2,4} = -0.9238$;
3) High-order MA process: $X_t = \sum_{l=0}^{15000}\theta_l\varepsilon_{t-l}$ with $\theta_0 = 1$, $\theta_1 = \pi/4$, and $\theta_l = \sin(\pi(l-1)/2)/(l-1)$ for $l = 2,3,\ldots,15000$.

Plots of the true SDFs are shown in Figure 1 for each process. These processes have SDFs that exhibit a range of local structures that can be hard to estimate using simple estimators of the SDF. We demonstrate how our estimation method performs when the innovations $\{\varepsilon_t\}$ are N(0, 1) RVs, but also present the same AR(2) process case 1) with innovations generated by a shifted Exponential distribution with mean 0 and variance 1 – we want to investigate how robust our method is to departures from Gaussianity.

We use the decibel-scale *integrated root mean squared error (IRMSE)* to measure how well we estimate the true SDF. For $M = \lceil N/2\rceil - 1$ non-zero, non-Nyquist frequencies, letting $\widetilde{S}(f_j)$ denote any estimate of the SDF at Fourier frequency f_j, the IRMSE is

$$\left\{\frac{1}{M}\sum_{j=1}^{M}\left[10\log_{10}\widetilde{S}(f_j) - 10\log_{10}S(f_j)\right]^2\right\}^{1/2}.$$

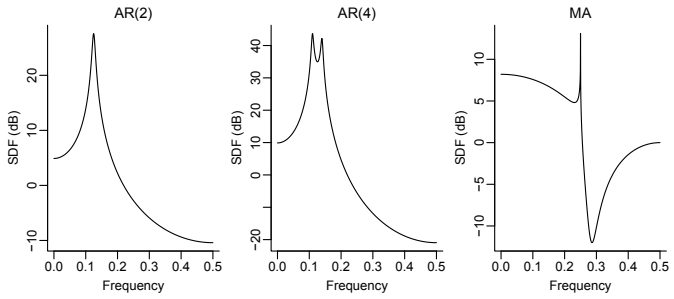


Fig. 1. Plots of the SDF for three processes on the decibel scale (dB).

We present the IRMSE averaged over 1000 realizations of each process.

In preliminary simulations we used a range of different basis function representations to model each SDF: orthogonal polynomial bases, Fourier bases, B-spline bases, wavelet bases, as well as some mixed dictionary bases. We found that a wavelet basis based on the discrete wavelet transform with the LA(8) wavelet filter had good performance across all the experimental conditions, where LA(8) represents the Daubechies least asymmetric wavelet of width 8 (see [51] for details). Following Walden *et al.* [6] we fit our models to mirrored data of frequencies on $[0, 1)$ to capture the evenness and periodic nature of the spectrum, but only summarize on the M frequencies between zero and the Nyquist frequency.

Figure 2 shows simulation results for the three processes (one for each panel) comparing our L_1 penalized MT-Whittle method (black circles) to an L_1 penalized LS approach (gray triangles). The first three panels are for Gaussian process, the last panel is for the non-Gaussian process. In each case, $N = 2048$, we use $K = 10$ sine tapers for the MT spectral estimate, and we start with the maximum number of basis functions, $p = M+1 = 1024$, including the intercept. We also show the effect of changing the method to select the tuning parameter. While we advocate choosing between the universal threshold (Univer.) and the GIC methods, we also compared to cases in which we used cross-validation (CV) to select the tuning parameter, or we did not penalize with a tuning parameter (None). We also calculated the tuning parameter that corresponds to the smallest possible IRMSE – in practice this value of the IRMSE is not known, but gives us a way to see how close our method is to the optimal value. In each figure the best IRMSEs for each method (least squares, gray; MT-Whittle, black) are denoted by the horizontal dashed lines. In terms of the uncertainty, the height of the symbols are larger than 95% bootstrap confidence intervals for each IRMSE.

Some results are coherent across the three processes. Using cross-validation to select the tuning parameters does not perform well compared to the universal and GIC methods, although it performs significantly better than using no penalization. For the AR processes, the L_1 penalized MT-Whittle outperforms L_1 penalized LS by between 4.4–6.0% in terms of the IRMSE. There was little difference between these two estimators for the MA process. By process, there were slight differences between the IRMSE values using the universal threshold and GIC to select the tuning parameter, however both methods yielded IRMSEs that were competitive with the best value possible. Also, for the non-Gaussian AR(2) example, all methods performed similarly with respect to one another. The right panel of Figure 2 shows that all penalized approaches have slightly increased IRMSE values when non-Gaussian noise is used, but that our method still outperforms the other approaches. This suggests that our L_1 MT-Whittle method is robust to non-Gaussianity.

As suggested by a referee, to further summarize how well our method performs, Figure 3 compares for each time series process, our L_1 penalized MT-Whittle to a raw MT spectral estimate. In each panel, we display the simulation of length $N = 2048$ that yields the median IRMSE. The solid black line denotes the L_1 penalized MT-Whittle estimate we obtain using the universal threshold with $K = 10$ data tapers. The gray line denotes our raw MT spectral estimate (again with $K = 10$) and the thin black line denotes the true SDF in each case. Although we slightly underestimate the spectral peaks using the L_1 method, our method better captures the general spectral features of each process relative to the raw MT estimate. The underestimation of the peaks reduces as N increases – since our estimation method is semiparametric, as N increases, p increases and we are better able to capture finer spectral features with more basis function (see also [6]).

We carried out an additional set of simulations that varied the number of sine tapers K for the MT-Whittle calculation – the results are omitted. Our simulations demonstrate that there is a quadratic relationship between the bias of our spectral estimator and K, and the IRMSE and K. We also learned that for very large values of K that the IRMSE for the L_1 penalized least squares method approached that of L_1 penalized MT-Whittle. This is not surprising since by Proposition 1, as $K \to \infty$, the distribution of log MT spectral estimate is better approximated by a normal distribution (e.g., [9]). Our selected value of $K = 10$ tapers provided a good compromise between balancing the bias and variance of the estimated SDF for these simulated processes. Also, using the L_1 penalized MT-Whittle outperformed the L_1 penalized Whittle based on the untapered periodogram.

A. Validating the empirical theoretical rate

In this section we empirically verify the conjectured rate in (19) by simulation using the four different time series processes defined above. Based on 1000 simulations we calculated the ratio comparing $\left\|\sum_{j=1}^{M}(R_j - 1)\phi(f_j)\right\|_\infty / \sqrt{M}$ with $\log(M)$ for a wide range of M values. This ratio should be bounded as we increase M. Figure 4 shows the median (solid line), 2.5th percentile (lower dashed line), and 97.5th percentile (upper dashed line) for this ratio, and demonstrates that our empirical rate indeed matches closely with the conjectured theoretical rate.

IX. Spectral analysis of EEG signals

Electroencephalogram (EEG) signals are often used to monitor brain activity and diagnose disease such as epileptic seizures. We analyze two channels of EEG data collected from the left and right front cortex of one male rat. (The

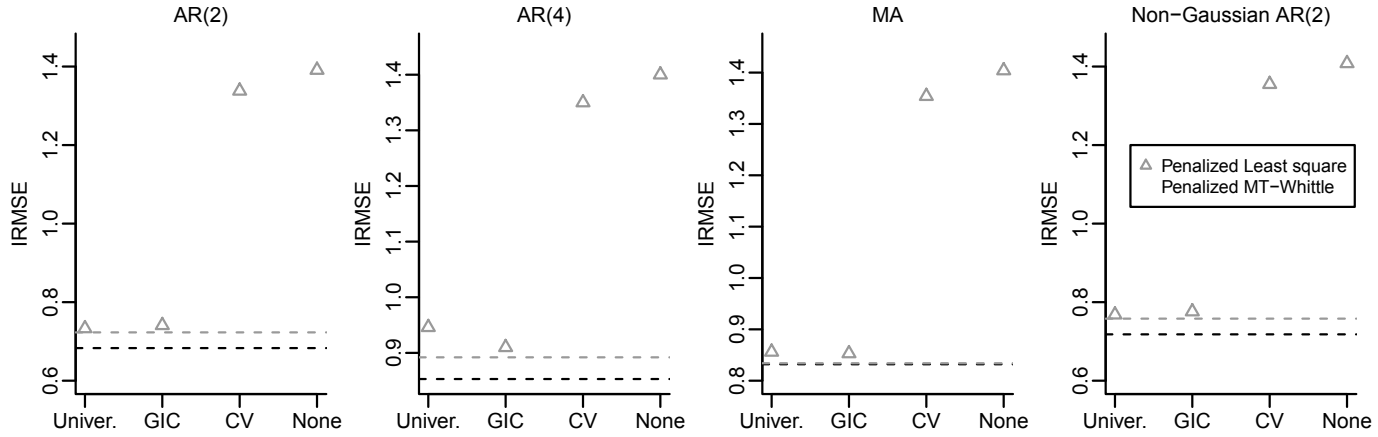


Fig. 2. A comparison of the IRMSEs for different L_1 penalizations methods for estimating the SDF. Here, the height of the symbols are larger than the widths 95% bootstrap confidence intervals for each IRMSE; the horizontal dashed lines indicate the best possible IRMSEs for each method.

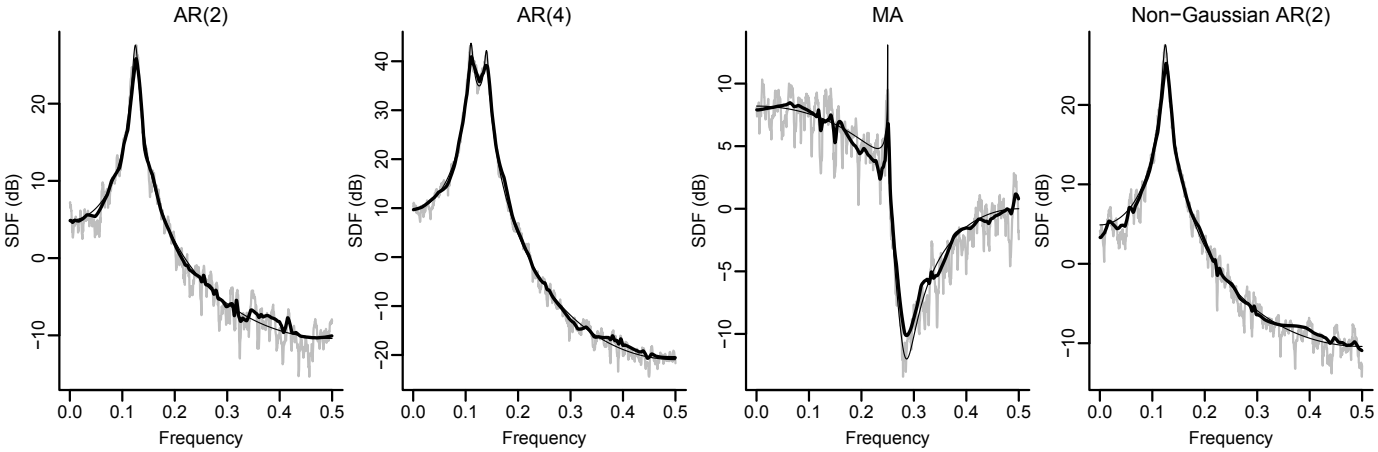


Fig. 3. A comparison of spectral estimates for the different four processes. For each process, the thick black line is the L_1 penalized MT-Whittle estimates, the gray line is the corresponding raw MT estimates with $K = 10$ tapers, and the thin black line is the true SDF.

data is presented in [52], and was downloaded from http://www.vis.caltech.edu/~rodri.) Quiroga *et al.* [53] argue that, genetically, analyzing these series is relevant to the study of human epilepsy. Each channel contains 1000 voltages recorded in units of microvolts (mV) collected at a sampling rate of 200 Hz. Time series plots of the left and right channels are shown in the panels (a) and (b) of Figure 5, and hint at strong spectral features in the two series.

In the supplement of [54] an L_2 penalized multivariate Whittle likelihood based on the periodogram is used to estimate the SDF. Using spline bases, and focusing on estimates of the SDF between 0 and 30 Hz, they discover spectral peaks at around 9 and 18 Hz which indicate a "local synchronization of neurons in both hemispheres of the frontal cortex".

Using our L_1 penalized MT-Whittle approach, we estimate the SDFs for the left and right channels separately. The MT-Whittle approach will counteract the bias due to leakage and mitigate the inconsistency of a periodogram-based approach. Compared with spline bases, we model the log SDF using a wavelet basis to better capture sharp peaks and other possible local features in the SDF. Additionally, the L_1 penalty enforces sparsity, keeping only relevant features in the estimated SDF.

Panels (c) and (d) of Figure 5 display, respectively, the SDF estimates for the left and right channels using our approach. (We mean padded the series to a length of $N = 1024$, prior to spectral estimation.) In each panel the solid line denotes the estimated SDF when we use the universal threshold method and the dashed line shows the estimate with the GIC method. In our estimates, we use MT-spectral estimates with $K = 5$ sine tapers, and we construct our wavelet basis using the LA(8) wavelet filter. Our SDF estimates for both channels contain several sharp turning points, and include features not depicted by the spline models of [54]. In terms of the tuning parameter selection, the universal-threshold- and GIC-based estimates are fairly similar to one another (the estimated SDFs are more similar for the left channel). This indicates that both methods are able to capture the interesting local features in the SDFs.

For the left channel, we estimate prominent spikes at 9 Hz and 16 Hz as well as a noticeable elbow around 5 Hz. There is more energy in the SDF estimates for the right channel, compared to the left. In the right channel, we pick up a strong broadband structure: the right channel presents clustered power between 5 Hz and 10 Hz, as well as a summit at 9 Hz, and a crest between 13 to 14 Hz.

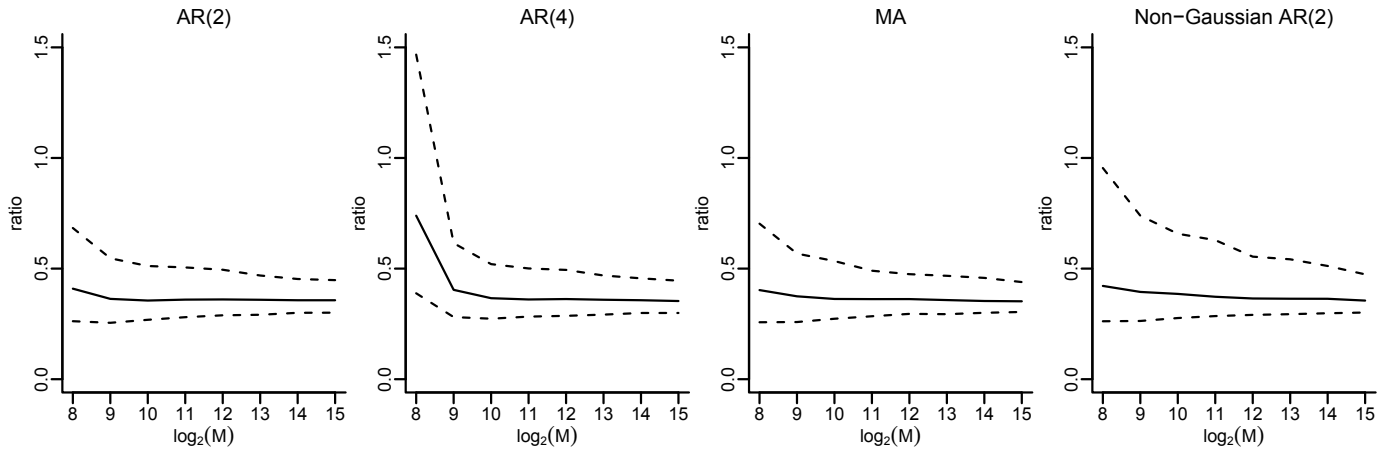


Fig. 4. Plots of the ratio $\left\|\sum_{j=1}^{M}(R_j-1)\phi(f_j)\right\|_\infty/(\sqrt{M}\log(M))$ against $\log_2(M)$, for M varying from 2^8 to 2^{15}. For each process, the solid line is the median of the ratio, and the dashed lines are the 2.5th and 97.5th percentiles of the ratio, based on 1000 simulations.

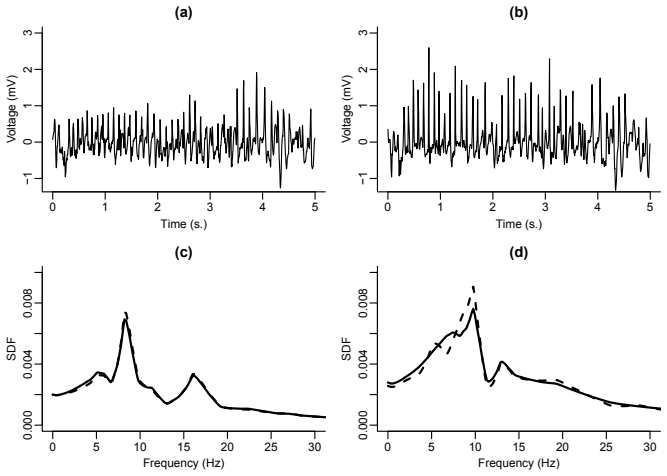


Fig. 5. Time series plots of the (a) left and (b) right EEG channels, as well as estimated SDFs for the (c) left and (d) right channels. In (c) and (d) the line solid is the estimate with the universal threshold, and the dashed line is the estimate with GIC-based threshold.

In conclusion our L_1 penalized MT-Whittle method is useful in revealing spike-wave discharges for EEG signals observed in the frontal cortex for the male rat under study. It would be interesting in the future to see how these methods apply to the spectral analysis of multiple EEG channels in different study populations.

X. Discussion

In this article we presented a L_1 penalized quasi-likelihood framework based on MT spectral estimates using a basis representation to model the SDF of a stationary time series. Our methodology allows the number of basis functions to increase with sample size, and through enforcing sparsity, our L_1 penalized MT-Whittle estimator performs better or as good as previous methods for estimating the SDF. Our method extends to the application of broader classes of basis functions and their mixtures, beyond those traditionally used with wavelet thresholding. Simulations demonstrate a clear advantage of using the GIC and calibrated universal threshold over cross-validation for tuning parameter selection, with a significant reduction in IRMSE. Computationally, the calibrated universal threshold is data-invariant (it only relies on the number of tapers K) whereas the calculation of GIC is data-dependent. However both methods are more efficient than using cross-validation. The proposed ADMM algorithm accelerates when parallel computing and orthogonal bases, such as the Daubechies class of wavelets, are employed.

There are a number of extensions that we are considering. In [6], the authors also vary the number of initial basis functions, p, that are chosen before they perform the L_1 penalization using least squares. We found in simulation studies that preselecting p can slightly reduce the IRMSE, but the problem is that the optimal choice of p depends on the underlying statistical process – it is not known how to select p for a given process. This idea is related to more general methods that can be used to simultaneously select and estimate the coefficients of the basis expansion. For example, using a truncated lasso approach proposed by Shen *et al.* [55] could lead to further reductions in IRMSE, but this requires developing a general approach to selecting the truncation parameters in the algorithm. Another extension is linked to the automatic relevance determination approach reformulated by [11], where a local minimum can be attained while achieving sparsity by solving a series of reweighted L_1 problems.

Some recent studies on inference for L_1 penalized methods in general settings (not specifically estimating SDFs) are summarized by Dezeure *et al.* [56]. Potentially, de-sparsifying the lasso and multi sample-splitting approaches can be applied to construct confidence intervals for our L_1 penalized MT-Whittle estimator, but the correlation between frequencies and the splitting strategy would need to be considered more carefully. We also note that the inference approach taken by Zhang [57] does not apply to our situation since their L_1 penalty term is based on the total variation distance.

We are also investigating methods for multivariate SDF estimation. Penalized basis expansions for multivariate SDF estimation typically involve spline basis functions, such as Whittle-based estimation for the cross-spectrum [58], as well

as penalized least squares [59] and penalized Whittle methods [54] based on a Cholesky decomposition. The extension of our L_1 penalized MT-Whittle approach to enforce sparsity in the estimation of multivariate SDFs would be beneficial.

APPENDIX A
PROOF OF PROPOSITION 2

We can rewrite the asymptotic distribution of $Z_j = \widehat{S}^{(\text{mt})}(f_j)$ $(j = 1, \ldots, M)$ stated in Proposition 1 as Gamma$(K, S(f_j)/K)$, a parametrization of the gamma typically used in generalized linear models (see, e.g., [22, ch. 8]). Here Gamma$(\nu, \mu/\nu)$ denotes a gamma RV with mean μ, shape parameter ν, and variance μ^2/ν. This asymptotic probability density function evaluated at $Z_j = z_j$ is

$$p(z_j) \;=\; \frac{z_j^{K-1}}{\Gamma(K)\left[\frac{S(f_j)}{K}\right]^K}\,\exp\left(-\frac{z_j}{S(f_j)/K}\right).$$

The proposition follows by assuming independence between the MT-spectral estimates over frequencies: the resulting quasi-likelihood, as required, is

$$\begin{aligned}\sum_{j=1}^{M}\log p(z_j) &= M\log\frac{K^K}{\Gamma(K)} + (K-1)\sum_{j=1}^{M}\log z_j \\ &\quad - K\sum_{j=1}^{M}\left[\log S(f_j) + \frac{z_j}{S(f_j)}\right] \\ &= \text{constant} - K\; l_W(\boldsymbol{\Phi}\boldsymbol{\beta}).\end{aligned}$$

APPENDIX B
DETAILS OF THE COMPUTATIONAL ALGORITHM

Following Boyd *et al.* [43], the augmented Lagrangian can be written as

$$\begin{aligned}l_A(\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\eta},\boldsymbol{u}_1,\boldsymbol{u}_2) &= l_W(\boldsymbol{\zeta}) + \sum_{l=1}^{p}\lambda_l|\eta_l| \\ &\quad + \frac{\rho}{2}\left\|\boldsymbol{\Phi}\boldsymbol{\beta} - \boldsymbol{\zeta} + \boldsymbol{u}_1\right\|_2^2 \\ &\quad + \frac{\rho}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\eta} + \boldsymbol{u}_2\right\|_2^2,\end{aligned}$$

where $\rho > 0$ is the penalty parameter. Note that the convergence is guaranteed for any positive ρ, although a carefully chosen ρ that balance the convergence of the primal and dual residuals often leads to faster convergence.

Minimizing the augmented Lagrangian over the primal and the dual variables, we obtain the ADMM updates as follows

$$\begin{aligned}&\boldsymbol{\beta}^{(n+1)} = (\boldsymbol{\Phi}^{\boldsymbol{T}}\boldsymbol{\Phi} + \boldsymbol{I})^{-1}\{\boldsymbol{\Phi}^T(\boldsymbol{\zeta}^{(n)} - \boldsymbol{u}_1^{(n)}) + (\boldsymbol{\eta}^{(n)} - \boldsymbol{u}_2^{(n)})\};\\
&\zeta_j^{(n+1)} = \underset{\zeta_j}{\arg\min}\left\{\zeta_j + \widehat{S}^{(\text{mt})}(f_j)\exp(-\zeta_j)\right.\\
&\qquad\qquad\left. + \frac{\rho}{2}\{\boldsymbol{\phi}^T(f_j)\boldsymbol{\beta}^{(n+1)} - \zeta_j + u_{1j}^{(n)}\}^2\right\}, \quad j = 1,\ldots,M;\\
&\eta_l^{(n+1)} = \text{ST}\left(\beta_l^{(n+1)} + u_{2l}^{(n)}, \frac{\lambda_l}{\rho}\right), \qquad l = 1,\ldots,p;\\
&\boldsymbol{u}_1^{(n+1)} = \boldsymbol{u}_1^{(n)} + (\boldsymbol{\Phi}\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\zeta}^{(n+1)});\\
&\boldsymbol{u}_2^{(n+1)} = \boldsymbol{u}_2^{(n)} + (\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\eta}^{(n+1)}),\end{aligned}$$

where, recall that $\text{ST}(x, a) = \text{Sign}(x)\max(|x| - a, 0)$ is the soft-thresholding function with threshold $a \geq 0$.

To obtain $\zeta_j^{(n+1)}$, we take the partial derivative of $l_A(\boldsymbol{\beta}^{(n+1)}, \boldsymbol{\zeta}, \boldsymbol{\eta}^{(n)}, \boldsymbol{u}_1^{(n)}, \boldsymbol{u}_2^{(n)})$ with respect to ζ_j and set to zero, which leads to a score equation

$$\rho\zeta_j - \widehat{S}^{(\text{mt})}(f_j)e^{-\zeta_j} + 1 - \rho(\boldsymbol{\phi}^T(f_j)\boldsymbol{\beta}^{(n+1)} + u_{1,j}^{(n)}) = 0\,,$$

which can be solved efficiently using the bisection method. The solution ensures a global minimum since the second derivative of $l_A(\boldsymbol{\beta}^{(n+1)}, \boldsymbol{\zeta}, \boldsymbol{\eta}^{(n)}, \boldsymbol{u}_1^{(n)}, \boldsymbol{u}_2^{(n)})$ with respect to ζ_j is $\rho + \widehat{S}^{(\text{mt})}(f_j)e^{-\zeta_j}$, which is positive for all $\zeta_j \in \mathbb{R}$ for any $\rho > 0$.

Following the optimality conditions and stopping criteria in [43], we define the *primal residual*

$$\boldsymbol{s}_{\text{pri}}^{(n+1)} = \begin{pmatrix}\boldsymbol{\Phi}\\ \boldsymbol{I}\end{pmatrix}\boldsymbol{\beta}^{(n+1)} - \begin{pmatrix}\boldsymbol{\zeta}^{(n+1)}\\ \boldsymbol{\eta}^{(n+1)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Phi}\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\zeta}^{(n+1)}\\ \boldsymbol{\beta}^{(n+1)} - \boldsymbol{\eta}^{(n+1)}\end{pmatrix},$$

and the *dual residual*

$$\begin{aligned}\boldsymbol{s}_{\text{dual}}^{(n+1)} &= \rho\begin{pmatrix}\boldsymbol{\Phi}\\ \boldsymbol{I}\end{pmatrix}^T(-\boldsymbol{I})\begin{pmatrix}\boldsymbol{\zeta}^{(n+1)} - \boldsymbol{\zeta}^{(n)}\\ \boldsymbol{\eta}^{(n+1)} - \boldsymbol{\eta}^{(n)}\end{pmatrix}\\ &= -\rho\left[\boldsymbol{\Phi}^T(\boldsymbol{\zeta}^{(n+1)} - \boldsymbol{\zeta}^{(n)}) + (\boldsymbol{\eta}^{(n+1)} - \boldsymbol{\eta}^{(n)})\right].\end{aligned}$$

We terminate the algorithm when both residuals are smaller than some prespecified precisions; more specifically, let

$$\begin{aligned}\epsilon^{\text{pri}} &= \sqrt{\frac{N}{2} - 1 + p}\;\epsilon^{\text{abs}}\\ &\quad + \epsilon^{\text{rel}}\;\max\Big\{\left\|\boldsymbol{\Phi}\boldsymbol{\beta}^{(n)}\right\|_2 + \left\|\boldsymbol{\beta}^{(n)}\right\|_2,\\ &\qquad\qquad \left\|\boldsymbol{\zeta}^{(n)}\right\|_2 + \left\|\boldsymbol{\eta}^{(n)}\right\|_2\Big\},\\ \epsilon^{\text{dual}} &= \sqrt{p}\;\epsilon^{\text{abs}} + \epsilon^{\text{rel}}\rho\left\|\boldsymbol{\Phi}^T\boldsymbol{u}_1^{(n)} + \boldsymbol{u}_2^{(n)}\right\|_2,\end{aligned}$$

where $\epsilon^{\text{abs}} > 0$ is an absolute tolerance and $\epsilon^{\text{rel}} > 0$ is a relative tolerance. The stopping criterion is then set to be $\|\boldsymbol{s}_{\text{pri}}^{(n)}\|_2 \leq \epsilon^{\text{pri}}$ and $\|\boldsymbol{s}_{\text{dual}}^{(n)}\|_2 \leq \epsilon^{\text{dual}}$. In our studies, setting both tolerances to 10^{-4} provided a reasonable balance between fast convergence of the optimization and satisfactory estimation accuracy.

APPENDIX C
DERIVATION OF THE ASYMPTOTIC DISTRIBUTION OF THE MAXIMUM QUASI-LIKELIHOOD ESTIMATOR

This derivation relies on writing our asymptotic distribution for the MT-spectral estimators as a generalized linear model (GLM) as given in [22]. Based on Proposition 1, the asymptotic density of $Z_j = \widehat{S}^{(\text{mt})}(f_j)$ evaluated at z_j in exponential family form is

$$p(z_j) = \exp\left\{\frac{z_j\vartheta - b(\vartheta)}{a(\varsigma)} + c(z_j, \varsigma)\right\},$$

with $\vartheta = -1/\mu_j = -1/S(f_j)$, $a(\varsigma) = 1/\nu = 1/K$, $b(\vartheta) = -\log(-\vartheta) = \log S(f_j)$, and $c(z_j, \varsigma) = K\log(Kz_j) - \log z_j - \log\Gamma(K)$. Then, as $\text{var}(Z_j) = b''(\vartheta)a(\varsigma) = V(\mu)a(\varsigma) = \mu^2/\nu$, the variance function is $V(\mu) = \mu^2$. Assuming asymptotic independence over frequencies, the Fisher Information for the GLM can be computed as

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{\Phi}^T\boldsymbol{W}\boldsymbol{\Phi},$$

where $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_M)$ with

$$w_j = \frac{1}{V(\mu_j)a(\varsigma)\big(g'(\mu_j)\big)^2} = \frac{1}{\mu_j^2(1/K)\big(1/\mu_j\big)^2} = K,$$

since our model assumes the link function $g(\mu) = \log \mu$, so that $g'(\mu) = 1/\mu$. Thus, with an orthonormal basis,

$$\boldsymbol{I}(\boldsymbol{\beta}) = K\, \boldsymbol{\Phi}^T \boldsymbol{I}_M \boldsymbol{\Phi} = K\, \boldsymbol{I}_p,$$

and under suitable regularity conditions (see, e.g., [22, appendix A]), as $M \to \infty$,

$$\widehat{\boldsymbol{\beta}}^W \to_d \boldsymbol{N}_p(\boldsymbol{\beta}, \boldsymbol{I}(\boldsymbol{\beta})^{-1}),$$

with $\mathrm{Cov}\Big(\widehat{\boldsymbol{\beta}}^W\Big) = \boldsymbol{I}(\boldsymbol{\beta})^{-1} = \boldsymbol{I}_p/K$.

APPENDIX D
PROOF OF THEOREM 1

For simplicity, we omit the dependence of $\widehat{\beta}(\lambda)$ on λ. Then

$$\widehat{\beta} \in \arg\min_{\beta} \sum_{j=1}^{M} \Big\{ (\beta - \beta^0)^\top \phi(f_j) + R_j \exp(-(\beta - \beta^0)^\top \phi(f_j)) \Big\} + \lambda \|\beta\|_1 \,,$$

since $\log S(f_j) = \phi^\top(f_j)\beta^0$ for each j. Hence,

$$\sum_{j=1}^{M} (\widehat{\beta} - \beta^0)^\top \phi(f_j) + R_j \exp(-(\widehat{\beta} - \beta^0)^\top \phi(f_j)) + \lambda\|\beta\|_1 \leq \sum_{j=1}^{M} R_j + \lambda \|\beta^0\|_1 \,. \tag{20}$$

Let $\widehat{\delta} = \widehat{\beta} - \beta^0$. Rearranging terms in (20), we obtain that

$$\sum_{j=1}^{M} R_j \Big\{ \exp(-\widehat{\delta}^\top \phi(f_j)) + \widehat{\delta}^\top \phi(f_j) - 1 \Big\} \leq \widehat{\delta}^\top \left\{ \sum_{j=1}^{M} (R_j - 1)\, \phi(f_j) \right\} + \lambda \|\beta^0\|_1 - \lambda \|\widehat{\beta}\|_1,$$

which, together with the fact that

$$\widehat{\delta}^\top \left(\sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right) \leq \widehat{\delta}\|_1 \left\| \sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right\|_\infty \leq \frac{\lambda}{2} \|\widehat{\delta}\|_1,$$

implies that

$$\begin{aligned} &\sum_{j=1}^{M} R_j \Big\{ \exp(-\widehat{\delta}^\top \phi(f_j)) + \widehat{\delta}^\top \phi(f_j) - 1 \Big\} \\ &\leq \frac{\lambda}{2} \Big(\|\widehat{\delta}\|_1 + 2\|\beta^0\|_1 - 2\|\widehat{\beta}\|_1 \Big) \\ &= \frac{\lambda}{2} \Big(\|\widehat{\beta}_S - \beta^0_S\|_1 + \|\widehat{\beta}_{S^c}\|_1 + 2\|\beta^0_S\|_1 - 2\|\widehat{\beta}_S\|_1 - 2\|\widehat{\beta}_{S^c}\|_1 \Big) \\ &\leq \frac{\lambda}{2} \Big(3\|\widehat{\beta}_S - \beta^0_S\|_1 - \|\widehat{\beta}_{S^c}\|_1 \Big) = \frac{\lambda}{2} \Big(3\|\widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1 \Big). \end{aligned}$$

Note that the LHS is nonnegative, because $e^x \geq x + 1$ for any $x \in \mathbb{R}$. It follows that $\|\widehat{\delta}_{S^c}\|_1 \leq 3\|\widehat{\delta}_S\|$. By Assumption (A2), we have that

$$\begin{aligned} &M \min \left\{ \frac{c_0}{s_0} \|\widehat{\delta}_S\|_1^2 \,, c_1 M^{\gamma - \frac{1}{2}} \|\widehat{\delta}_S\|_1 \right\} \\ &\leq \frac{\lambda}{2} \Big(3\|\widehat{\delta}_S\|_1 - \|\widehat{\delta}_{S^c}\|_1 \Big) \leq \frac{3\lambda}{2} \|\widehat{\delta}_S\|_1 \,. \end{aligned}$$

Hence, on the event that

$$\frac{2}{3} c_1 M^{\gamma + \frac{1}{2}} > \lambda \geq 2 \left\| \sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right\|_\infty ,$$

we have

$$\Big\| \widehat{\beta}(\lambda) - \beta^0 \Big\|_1 \leq \frac{3\lambda s_0}{2c_0 M} \,. \tag{21}$$

Letting $\lambda = 2 \left\| \sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right\|_\infty$, we have on event

$$\left\{ \left\| \sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right\|_\infty < 3^{-1} c_1 M^{\frac{1}{2} + \gamma} \right\},$$

that

$$\|\widehat{\delta}\|_1 \;\leq\; \frac{3\lambda s_0}{2c_0 M} \;=\; \frac{3s_0}{c_0 M} \left\| \sum_{j=1}^{M} (R_j - 1)\phi(f_j) \right\|_\infty ,$$

and

$$\begin{aligned} &\sup_{f \in [-\frac{1}{2}, \frac{1}{2}]} |\log \widehat{S}(f) - \log S(f)| \\ &\leq \sup_{f \in [-\frac{1}{2}, \frac{1}{2}]} |(\widehat{\beta} - \beta^0)^\top \phi(f)| \;\leq\; \|\widehat{\beta} - \beta^0\|_1 \|\phi(f)\|_\infty \\ &\leq \frac{3Bs_0}{c_0 M} \left\| \sum_{j=1}^{M} (R_j - 1)\, \phi(f_j) \right\|_\infty . \end{aligned}$$

This completes the proof.

ACKNOWLEDGMENT

Craigmile is supported in part by the US National Science Foundation (NSF) under grants NSF-DMS-1407604 and NSF-SES-1424481, and the National Cancer Institute of the National Institutes of Health under Award Number R21CA212308. Zhu is supported in part by the NSF under grants NSF-DMS-1721445 and NSF-DMS-1712580. We thank the Associate Editor and reviewers for suggestions that improved the article.

REFERENCES

[1] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge, England: Cambridge University Press, 1993.

[2] E. Mosley-Thompson, C. R. Readinger, P. Craigmile, L. G. Thompson, and C. A. Calder, "Regional sensitivity of Greenland precipitation to NAO variability," *Geophysical Research Letters*, vol. 32, no. L24707, 2005.

[3] H.-Y. Gao, "Wavelet estimation of spectral densities in time series analysis," Ph.D. dissertation, Department of Statistics, University of California, Berkeley, 1993.

[4] ——, "Choice of thresholds for wavelet shrinkage estimate of the spectrum," *Journal of Time Series Analysis*, vol. 18, pp. 231–251, 1997.

[5] P. Moulin, "Wavelet thresholding techniques for power spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 42, pp. 3126–3136, 1994.

[6] A. T. Walden, D. B. Percival, and E. J. McCoy, "Spectrum estimation by wavelet thresholding of multitaper estimators," *IEEE Transactions on Signal Processing*, vol. 46, pp. 3153–3165, 1998.

[7] R. Cogburn and H. T. Davis, "Periodic splines and spectral estimation," *The Annals of Statistics*, vol. 2, pp. 1108–1126, 1974.

[8] G. Wahba and S. Wold, "Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing," *Communications in Statistics - Theory and Methods*, vol. 4, pp. 125–141, 1975.

[9] G. Wahba, "Automatic smoothing of the log periodogram," *Journal of the American Statistical Association*, vol. 75, pp. 122–132, 1980.

[10] Y. Pawitan and F. O'Sullivan, "Nonparametric spectral density estimation using penalized Whittle likelihood," *Journal of the American Statistical Association*, vol. 89, pp. 600–610, 1994.

[11] D. P. Wipf and S. S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems*, 2008, pp. 1625–1632.

[12] C. Byrnes, T. T. Georgiou, and A. Lindquist, "A new approach to spectral estimation: A tunable high-resolution spectral estimator," *IEEE Transactions on Signal Processing*, vol. 48, pp. 3189–3205, 2000.

[13] T. T. Georgiou and A. Lindquist, "Kullback-leibler approximation of spectral density functions," *IEEE Transactions on Information Theory*, vol. 49, pp. 2910–2917, 2003.

[14] A. Basu, I. R. Harris, N. L. Hjort, and M. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, pp. 549–559, 1998.

[15] A. Ferrante, M. Pavon, and M. Zorzi, "A maximum entropy enhancement for a family of high-resolution spectral estimators," *IEEE Transactions on Automatic Control*, vol. 57, pp. 318–329, 2012.

[16] M. Zorzi, "A new family of high-resolution multivariate spectral estimators," *IEEE Transactions on Automatic Control*, vol. 59, pp. 892–904, 2014.

[17] ——, "Multivariate spectral estimation based on the concept of optimal prediction," *IEEE Transactions on Automatic Control*, vol. 60, pp. 1647–1652, 2015.

[18] ——, "An interpretation of the dual problem of the three-like approaches," *Automatica*, vol. 62, pp. 87–92, 2015.

[19] A. Cichocki and S. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, pp. 1532–1568, 2010.

[20] P. Whittle, "Estimation and information in stationary time series," *Arkiv för Matematik*, vol. 2, pp. 423–434, 1953.

[21] R. W. M. Wedderburn, "Quasi-likelihood functions, generalized linear models, and the gauss-newton method," *Biometrika*, vol. 61, p. 439, 1974.

[22] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. New York, NY: Chapman and Hall/CRC Press, 1999.

[23] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, pp. 1055–1096, 1982.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[25] D. Brillinger, *Time Series: Data Analysis and Theory*. New York, NY: Holt, 1981.

[26] P. Bloomfield, "An exponential model for the spectrum of a scalar time series," *Biometrika*, vol. 60, pp. 217–226, 1973.

[27] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[28] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200–1224, 1995.

[29] S. Sardy, A. Antoniadis, and P. Tseng, "Automatic smoothing with wavelets for a wide class of distributions," *Journal of Computational and Graphical Statistics*, vol. 13, pp. 399–421, 2004.

[30] K. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Transactions on Signal Processing*, vol. 43, pp. 188–195, 1995.

[31] A. T. Walden, "A unified view of multitaper multivariate spectral estimation," *Biometrika*, vol. 87, pp. 767–788, 2000.

[32] P. Welch, "The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio Electoacoustics*, vol. 15, pp. 70–73, 1967.

[33] M. Calder and R. A. Davis, "Introduction to Whittle (1953), The Analysis of Multiple Stationary Time Series," in *Breakthroughs in Statistics*. New York, NY: Springer, 1997, pp. 141–169.

[34] U. Grenander and G. Szegö, *Toeplitz forms and their applications*. Berkeley, CA: University of California Press, 1958.

[35] E. J. Hannan, "The asymptotic theory of linear time-series models," *Journal of Applied Probability*, vol. 10, pp. 130–145, 197.

[36] R. Dahlhaus, "Small sample effects in time series analysis: a new asymptotic theory and a new estimate," *The Annals of Statistics*, vol. 16, pp. 808–841, 1988.

[37] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press, 2002.

[38] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, pp. 13–22, 1986.

[39] S. L. Zeger and K.-Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, vol. 42, pp. 121–130, 1986.

[40] L. A. Wasserman, *All of Nonparametric Statistics*. New York, NY: Springer, 2006.

[41] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Revue Française d'Automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, vol. 9, pp. 41–76, 1975.

[42] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, pp. 17–40, 1976.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2011.

[44] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge, England: Cambridge University Press, 2000.

[45] B. He and X. Yuan, "On non-ergodic convergence rate of Douglas–Rachford alternating direction method of multipliers," *Numerische Mathematik*, vol. 130, pp. 567–577, 2015.

[46] Y. Fan and C. Y. Tang, "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, pp. 531–552, 2013.

[47] S. A. Van de Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, pp. 614–645, 2008.

[48] S. A. van de Geer and P. Bhlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[49] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, vol. 28, pp. 1302–1338, 2000.

[50] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY: Springer, 2009.

[51] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992, vol. 61.

[52] G. van Luijtelaar, *The WAG/Rij Rat Model of Absence Epilepsy: Ten Years of Research: a Compilation of Papers*. Nijmegen, Netherlands: Nijmegen University Press, 1997.

[53] R. Q. Quiroga, T. Kreuz, and P. Grassberger, "Event synchronization: a simple and fast method to measure synchronicity and time delay patterns," *Physical Review E*, vol. 66, no. 041904, 2002.

[54] R. T. Krafty and W. O. Collinge, "Penalized multivariate Whittle likelihood for power spectrum estimation," *Biometrika*, vol. 100, pp. 447–458, 2013.

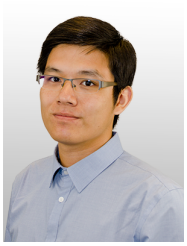
[55] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, pp. 223–232, 2012.

[56] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen, "High-dimensional inference: Confidence intervals, p-values and R-software hdi," *Statistical Science*, vol. 30, pp. 533–558, 2015.

[57] L. Zhang, "Penalized regression methods in time series and functional data analysis," Ph.D. dissertation, University of Alberta, Canada, 2017.

[58] Y. Pawitan, "Automatic estimation of the cross-spectrum of a bivariate time series," *Biometrika*, vol. 83, pp. 419–432, 1996.

[59] M. Dai and W. Guo, "Multivariate spectral analysis using Cholesky decomposition," *Biometrika*, vol. 91, pp. 629–643, 2004.

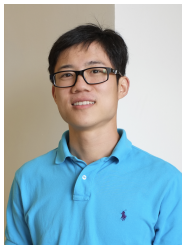


Shuhan Tang received the B.Sc. degree in mathematics from Central South University, Changsha, China, in 2011, and a M.Sc. degree in applied statistics from Bowling Green State University, Bowling Green, OH, United States, in 2014.

He is currently a Ph.D. student in statistics at The Ohio State University, Columbus. His current research interests include time series, statistical learning, and causal inference. He was the recipient of the Dr. Gary G. Koch and Mrs. Carolyn J. Koch Fellowship at The Ohio State University.

Peter F. Craigmile received a B.Sc. degree in mathematics and statistics from the University of Glasgow, Glasgow, U.K., in 1996, a Diploma in mathematical statistics from Cambridge University, Cambridge, U.K., in 1997, and a Ph.D. degree in statistics from the University of Washington, Seattle, in 2000.

He is a professor in the Department of Statistics, The Ohio State University, Columbus. His research interests include time series analysis, spatial statistics, spatio-temporal modeling, and longitudinal methods. He carries out application-oriented research in areas such as Biomedical Sciences, Climatology, Epidemiology, Environmental Sciences, and Psychology. He is a fellow of the American Statistical Association and The Royal Statistical Society.



Yunzhang Zhu received a B.S. degree in mathematics and physics from Tsinghua University, Beijing, China, in 2009 and a Ph.D. degree in statistics from University of Minnesota, Minneapolis, MN, United States, in 2014.

He is currently an assistant professor in the Department of Statistics at The Ohio State University. His research interests include statistical learning and optimization.