Missing-Data Nonparametric Coherency Estimation

C. Haley, *Member, IEEE*

***Abstract*—Chave recently proposed an estimator for multitaper spectral density where the time series contains missing values. In this article we generalize this technique to a multitaper estimator of coherence and phase and show that one can also obtain bootstrapped confidence intervals. We give two examples. The first is a toy example in which the true coherence is known. In the second example we show that the multitaper missing-data coherence estimator computed on real data with a single gap comprising 11% of the data outperforms the Daniell-smoothed coherence estimator where there are no gaps. The case where the two time series have different missing indices is also discussed.**

***Index Terms*—Coherence, Lomb-Scargle Periodogram, Missing Data Problem, Multivariate Time Series, Multitaper, Power Spectrum**

I. Introduction

Nonparametric bivariate frequency-domain analysis of time series is indispensable in numerous fields for the identification of common periodicities, time delay estimation, and linear mechanisms, and in the case of vector-valued data, determination of polarization parameters and related quantities [5], [13]. While time series are usually collected with equally spaced samples, it is common to have gaps in the data. This situation is referred to as the missing data problem. Frequently, missing data are interpolated by some method before proceeding to exploratory nonparametric spectrum or coherence analysis, and the interpolation-estimation process may be iterative [1].

A recent innovation in nonparametric spectral analysis has been to use multiple tapering functions, or generalized prolate spheroidal sequences, that incorporate data gaps in a way such that the taper functions are orthogonal on the temporal grid and yet still properly constrain spectral leakage in frequency [6]. This missing-data multitaper power spectrum inherits the desirable properties of multitaper spectrum estimation [19], namely, (i) it provides consistent estimates with bounded bias, (ii) enables optimal control of the bias-variance tradeoff [14], [4], [7], (iii) performs particularly well on spectra that have a large range or large numbers of discrete oscillations [22], and (iv) is an approximately maximum likelihood estimator of the spectrum [18]. This spectrum estimator is a special case of a more general technique due to Bronez for data observed on nonuniform temporal grids in one or more dimensions [6], [3].

This paper is concerned with coherence and phase estimation when two time series contain the same sampling and the same gaps. This allows for meaningful extension of coherence and phase estimation using multitapers described in [21]. This method does not require hand tuning, parameter selection, or any ad hoc procedure for filling gaps, and it could be automated.

Submitted mm/dd/yyyy. This material was based on work supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under contract number DE-AC02-06CH11357.

C. Haley is with the Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave, Lemont, IL 60439 USA (e-mail: haley@anl.gov).

II. Methods

A. Generalized Prolate Spheroidal Sequences

The missing-data multitaper method requires its own set of optimal tapers [19], [3], [6], the *missing-data prolate spheroidal sequences* or missing-data Slepian sequences. These sequences extend the notion of maximally bandlimited orthogonal sequences, [17], to those sampled on a grid where there is missing-data. Define the bandwidth W for the desired spectral window, and let the sequences be given on the length N index set $\{t_n\}_{n=0}^{N-1}$ with unit sampling except where there are gaps.

The missing-data Slepian sequences solve the eigenvalue problem

$$\lambda_k \, v_n^{(k)} = \sum_{m=0}^{N-1} \frac{\sin 2\pi W(t_n - t_m)}{\pi(t_n - t_m)} \, v_m^{(k)}. \tag{1}$$

The kth missing-data Slepian sequence is denoted $v_n^{(k)}$ at time t_n, and and λ_k is its associated eigenvalue, sorted in decreasing order $1 > \lambda_0 > \lambda_1 > \ldots > \lambda_{N-1} > 0$. Note that both of these explicitly depend on N and W, but for simplicity this is suppressed in the notation. When there are no missing values, the sinc-like matrix in (1) is Toeplitz and has a special form which allows for numerically effective eigenvalue routines.

By applying the nonuniform discrete Fourier transform to the generalized data taper in (1), one obtains the missing-data prolate spheroidal functions, or *missing-data Slepian functions*

$$V^k(f) = \sum_{n=0}^{N-1} v_n^{(k)} e^{-2\pi i t_n f}. \tag{2}$$

The missing-data Slepian sequences form an orthonormal set on the grid $\{t_n\}_{n=0}^{N-1}$; and the missing-data Slepian functions are orthonormal on $(-1/2, 1/2)$ and orthogonal on $(-W, W)$.

1) Missing-Data Spectrum and Cross-Spectrum Analysis: Letting $\{x_n\}_{n=0}^{N-1}$ and $\{y_n\}_{n=0}^{N-1}$ be samples from two jointly stationary, zero mean time series on the missing data grid, one constructs a multitaper estimate of the cross spectrum as follows. First, one computes the missing-data eigen-(cross)-spectra in [6] using the nonuniform discrete Fourier transform

$$\widehat{S}_{xy}^{(k)}(f) = \left(\sum_{j=0}^{N-1} x_t v_t^k e^{-i2\pi f t_j} \right) \left(\sum_{j=0}^{N-1} y_j v_t^k e^{-i2\pi f t_j} \right)^* \tag{3}$$

$$\widehat{S}_{xy}(f) = \frac{\sum_{k=0}^{K-1} d_x(f) d_y(f) \widehat{S}_{xy}^k(f)}{\sqrt{\sum_{k=0}^{K-1} d_x^2(f) \sum_{k=0}^{K-1} d_y^2(f)}}. \tag{4}$$

These are approximately independent estimators of the power spectrum. The approximate bandlimitedness of the tapers controls the bias of each $\hat{S}_{xy}^{(k)}(f)$, and each of these is exponentially distributed. One gains a significant reduction in variance by estimating the spectrum using a weighted average of the eigenspectra,

$$\widehat{S}_{xx}(f) = \frac{\sum_{k=0}^{K-1} d_x^2(f)\widehat{S}_{xx}^{(k)}(f)}{\sum_{k=0}^{K-1} d_x^2(f)}, \tag{5}$$

where we have used $0 > K \lessapprox 2NW$ tapers. The frequency-dependent weights, $d_x(f)$, are chosen so as to satisfy the equation [19, §V]

$$\sum_{k=0}^{K} \frac{\lambda_k(\widehat{S}_{xx}(f) - \widehat{S}_{xx}^{(k)}(f))}{(\lambda_k \widehat{S}_{xx}(f) + s_x^2(1-\lambda_k))^2} = 0, \tag{6}$$

where $s_x^2 = (N-1)^{-1}\sum_{j=0}^{N-1} x_j^2$ is the sample variance of the time series. The control of the bias-variance tradeoff through the adjustment of bandwidth is widely recognized as the main performance advantage of the multitaper method [4], [7].

2) Coherence Estimation: The coherency between two zero mean jointly stationary time series $x(t)$ and $y(t)$ is defined as

$$\rho(f) = \frac{S_{xy}(f)}{\sqrt{S_{xx}(f)S_{yy}(f)}}, \tag{7}$$

where $S_{xx}(f)$ denotes the spectrum of the process $x(t)$, $S_{yy}(f)$ denotes the spectrum of the process $y(t)$, and $S_{xy}(f)$ denotes the (complex-valued) cross spectrum of $x(t)$ and $y(t)$. The squared magnitude of this quantity is known as the *magnitude squared coherence* (MSC); and the angle, often reported in degrees, is the *phase*. The MSC can be interpreted as a frequency-dependent correlation coefficient indicating the proportion of variance common to both series via a linear relationship [10], [21], [16].

A simple multitaper estimator for the coherency is computed in the following way:

$$c(f) = \frac{\widehat{S}_{xy}(f)}{\sqrt{\widehat{S}_{xx}(f)\widehat{S}_{yy}(f)}}. \tag{8}$$

This estimator avoids the need for smoothing [2] through the use of the frequency-dependent weights. Using the known distribution of the coherence [9], [5], [21], one can assign significance to the estimate.

3) Jackknifing Multitaper Estimates: The jackknife variance given in [21], [20] can be used to construct confidence intervals for the spectrum, coherence, and phase [21], [11], [5], so we will not repeat the theory here. Note that (1) One obtains the independent estimates of the coherency needed for jackknifing by virtue of having used K orthogonal tapers. One jackknifes over tapers, not frequencies; and (2) the use of transformations with the jackknife is especially important when the statistics of interest are bounded [12, §5.3], so that one can use the Student-t distribution for the construction of the confidence intervals. For this reason we prefer to jackknife over logarithms of the spectrum, normalized inverse hyperbolic tangents of the coherence, and ratios or *phase factors* computed from the complex coherency.

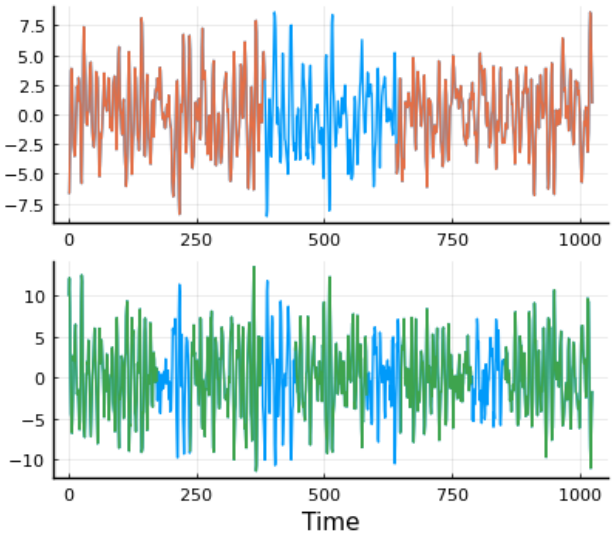

Fig. 1. Time series above show two different missing-data sampling paradigms, i.e. the top panel shows the data x_t, Eqn. (9) in blue, and the red curve shows a time series derived from x_t in which a single gap is missing from the center of the sample, the bottom plot shows the data y_t, Eqn. (10) in blue and missing data in green show a sampling paradigm having four gaps.

III. Example: Known Coherence

The simplest example on which to test a coherence estimator is the case where a noisy linear filter is applied to an input. Consider as input the AR(2) process

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} = w_t \tag{9}$$

where $\phi_1 = -1.5$ and $\phi_2 = 0.75$ and w_t is unit variance white noise. Let y_t be the result of passing x_t through a moving average filter and adding uncorrelated unit variance white noise, ζ_t:

$$z_t = \sum_{j=1}^{q} \psi_j x_{t-j}; \quad y_t = z_t + \zeta_t, \quad t = 0, \ldots, N-1, \tag{10}$$

where $q = 4$, $\psi_1 = 1.0$, $\psi_2 = -0.25$, $\psi_3 = 0.1$, $\psi_4 = -1.33$, and ζ_t is uncorrelated with the input.

The MSC between x_t and y_t is [10, p. 146]

$$\rho_{x,y}^2(f) = \frac{S_z(f)}{1 + S_z(f)}, \tag{11}$$

where $S_z(f)$ is the spectrum of the intermediate process z_t,

$$S_z(f) = \frac{\left|\sum_{j=1}^{q} \psi_j e^{-2\pi i j f}\right|^2}{\left|1 - \sum_{j=1}^{p} \phi_j e^{-i2\pi j f}\right|^2}. \tag{12}$$

Fig. 1 shows a $N = 1,024$ sample realization of x_t, y_t. In order to study the impact of different numbers of gaps on the process, three generic temporal samplings are shown, that is, where there are zero, one, and four gaps, and where data are missing, there are a total of 256 missing data points. Figure 2 shows the squared coherences between x_t and y_t in the three sampling paradigms along with the theoretical MSC in Eqn. (11). Contrasting the blue curve with the green curve in the right panel of the figure, one gets a poorer estimate as

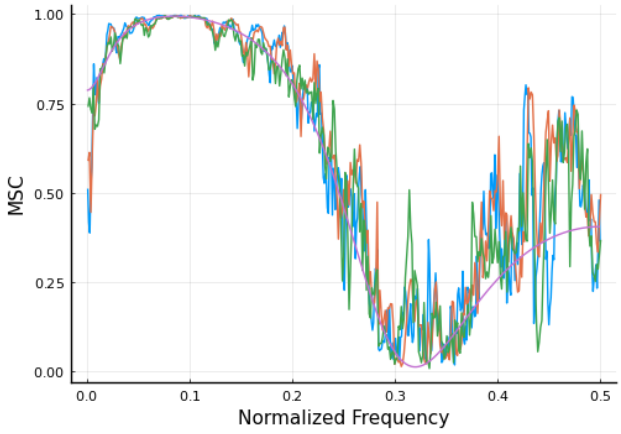

Fig. 2. The MSC estimates of the data shown in 1 between the process x_t and y_t for the two sampling schemes where the colors of the curves match with panels (a) as well as the theoretical MSC in Eqn. (11), smooth purple curve. The MSC where there are four gaps in the data appears to be the most variable, however 95% confidence intervals (not shown) overlap.

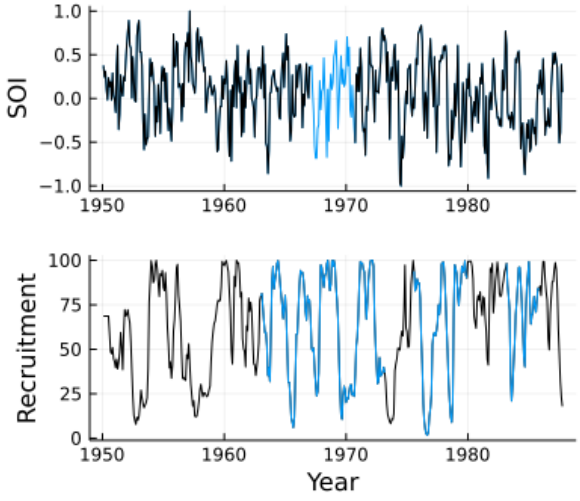

Fig. 3. Two time series measured at identical timepoints whose characteristics are described in the text (blue lines), SOI is unitless while Recruitment is in number of new fish. A second set of time series with large gaps is constructed and shown in black. Coherence estimates for the two missing-data paradigms are shown in Figs. 4, 6.

one increases the number of gaps; however the missing-data estimators track the multitaper coherence without gaps.

IV. Example: Southern Oscillation Index and Recruitment

We follow a textbook example taken from Shumway and Stoffer [16, Example 4.21] that investigates the relationship between the Southern Oscillation Index (SOI) and recruitment, the number of new tropical fish computed from a model. These data are also analyzed using Bayesian methodology in [15]. The data are sampled monthly for a total of 453 samples; time series are shown in the left panels of Fig. 4. We have created gapped time series by removing 49 samples from each dataset, about 11% of the time series length. The gapped series are shown in black, overlaid on the original data in blue.

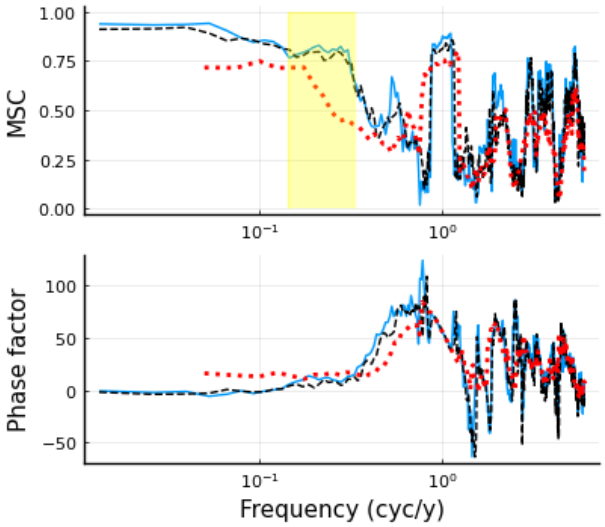

Fig. 4. Coherence and phase estimates of the data shown in Fig. 3 as a function of frequency in cycles per year. The red dotted line shows the result computed on gap-free data using a smoothed estimate of the squared coherence and phase described in [16, Example 4.21]. The blue solid line is a multitaper estimate of the same quantities, computed with a time-bandwidth product of $NW = 5$ and 9 tapers. The black dashed lines show the multitaper estimates computed on the data with gaps, with the same choices of time-bandwidth product and number of tapers. Phase estimates are similar and have been calculated by first subtracting the line $180 - 150f$ degrees as described in the text. The shaded yellow band in the top right panel shows the range in frequency over which El Niño ought to produce high coherence.

Coherence estimates and their phases are computed by using three techniques: (i) the tapered and smoothed method of [16] where there is no taper, a Daniell kernel with parameter 9 ($D(t) = 1/19$ for $t = -9, \ldots, 0, \ldots, 9$) is used as a smoother and is shown as red dotted lines in the right panels of Fig. 4; (ii) the multitaper estimate of coherence and phase for time series without gaps [21], with $NW = 5, K = 6$ shown as the blue solid line; and (iii) the jackknife mean multitaper missing-data coherence estimator described in the preceding section and shown as the black dashed line.

Coherence estimates differ enormously between the smoothed periodogram and multitaper techniques even for regularly sampled time series. In particular, effective bandwidths are larger for the periodogram estimate, resulting in wider peaks (e.g., the large peak at one cycle per year); and the smoothing across frequency results in values of squared coherence reverting toward its mean across frequency. This results in suppression of large values of the coherency and smearing of relevant features. Despite the variance-reducing properties of smoothing, a 99.9% confidence interval for this estimate spans a large width, about 0.35, around the estimate; see [16, Fig 4.15].

While the authors report that El Niño cycles should be visible in these data at periods of 3–7 years (0.14–0.33 cycles/y), shown as a shaded yellow band in the figure, a peak appears at about 9 years (0.11 cyc/y) in the squared coherence estimate (red line) in Fig. 4, dropping to average levels by the 3-year period (0.33 cyc/y). In contrast, the multitaper estimates both with and without gaps remain large in the El Niño range in frequency and also appear to detect common periodicity

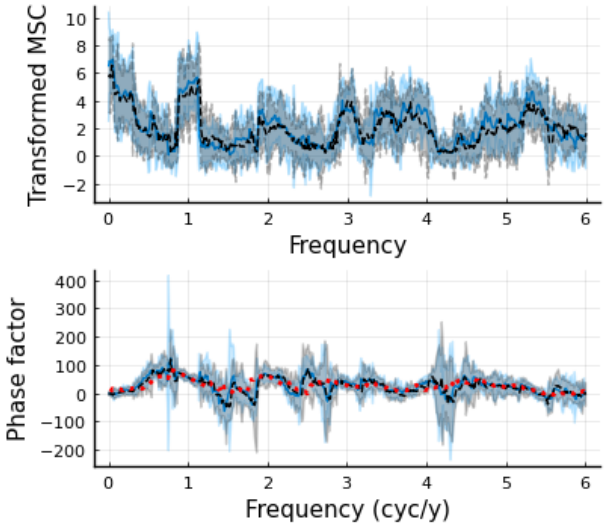

Fig. 5. Top panel shows the comparison between the two multitaper estimates on the transformed (arctanh scale) described in [21], where jackknifed confidence intervals are evaluated (95%) CI. Bottom panel shows the estimates of the phase where the line y = 150 x - 180 has been subtracted. Both figures have the same color scheme as in Fig. 4.

between SOI and recruitment at a period of 2 years (0.5 cyc/y). The multitaper estimate that allows gaps mainly outperforms the periodogram by the same metrics. Confidence intervals for the multitaper estimates are computed in Fig. 5.

The estimates of the phase in Fig. 4 have had the simple linear trend with slope approximately $150^\circ/(\text{cyc/y})$ and intercept 180° removed and phase factors calculated relative to this line in order to properly compare the phase estimates. The linear trend and 180 degree intercept indicates that fish recruitment lags the Southern Oscillation Index by about one month. If this line is removed from all of the phase estimates in the bottom right panel of Fig. 4 and phase factors (or circularly wrapped estimates) computed, one obtains the bottom panel of Fig. 5, which shows the ultimate similarity of the three estimates with their jackknifed 95% confidence intervals.

Varying the number and size of the gaps, we produce another example using the sampling pattern shown in the bottom panel of Fig 3. Spectra and squared coherence are shown in Fig. 6. It is clear that with this amount of missing data the performance of the estimator is beginning to break down, but very similar characteristics remain, even in such an extreme case.

V. Different Missing-Data

When the two time series of interest have missing data at different sets of indices, i.e. the processes $x(t)$ and $y(t)$ are observed on the temporal grid $\{t_{x,j}\}_{j=1}^{M}$ and $\{t_{y,j}\}_{j=0}^{N}$ with the same sampling rate and duration, the two sets of missing-data Slepian sequences $v_{x,n}^{(k)}$ and $v_{y,n}^{k}$, that one obtains from (1) are not orthogonal. By replacing the missing data-slepians in the cross-spectrum estimate (4) with their associated tapers, one obtains an estimator of the cross spectrum that *does not inherit most of the desirable properties of the orinary multitaper.*

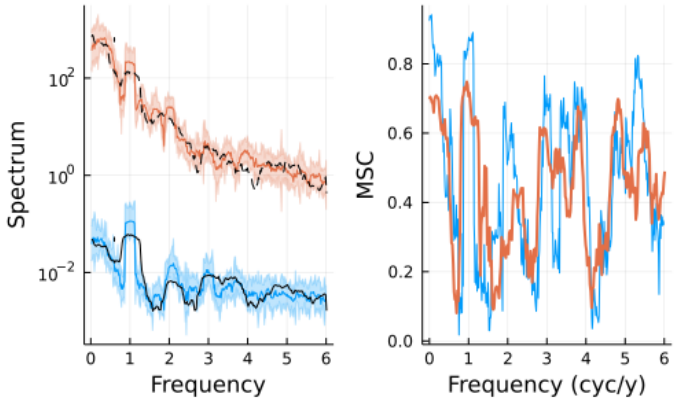

Fig. 6. Left panel shows the multitaper spectra (SOI, lower, in $(\text{cyc/y})^{-1}$ and Recruitment, upper, in units of squared numbers of fish per cyc/y) with the missing-data version lines in black and right panel shows squared coherence of the SOI and Recruitment datasets for a spectacular 45% missing data paradigm (shown on the Recruitment dataset, bottom panel of Fig. 3.) Missing data are shown in red, ordinary multitaper in blue.

While the tapers produce a cross-spectrum with good bias properties, the lack of orthogonality of the tapers implies that the reduction in variance one gets by multitapering is radically reduced. One may also find that misalignment problems are exacerbated, e.g. [10], that is, coherence may be underestimated when events with large variance are separated widely in time, or are partially covered by gaps. Jackknifing is not to be used, nor is assigning significance to transformed coherence estimates according to the Student-t distribution.

VI. Conclusion

In this paper, coherences have been computed on data containing gaps and a jackknife method is used to estimate the variance. While many have used interpolation to approach the missing data problem, this paper and the paper of Chave [6] offer an alternative nonparametric method that is not iterative or ad-hoc. The missing-data coherence can be employed in an automated fashion with the following caveat: Slepian tapers can have poor concentrations when the data have numerous small gaps, so it is useful to alert the user when this is the case.

The reader will find Jupyter notebooks in the `Multitaper.jl` package [8] to reproduce the examples in this paper.

VII. Acknowledgments

The author thanks Sally Dodson-Robinson, David J. Thomson, and Alan Chave for their input on this work. This material is based upon work supported by the US Department of Energy, Office of Science, Office of Advanced Scientific Computing Research.

REFERENCES

[1] P. Babu and P. Stoica. Spectral analysis of nonuniformly sampled data – A review. *Digital Signal Processing*, 20(2):359–378, 2010.

[2] P. J. Brockwell, R. A. Davis, and S. E. Fienberg. *Time series: Theory and methods*. Springer Science & Business Media, 1991.

[3] T. P. Bronez. Spectral estimation of irregularly sampled multidimensional processes by generalized prolate spheroidal sequences. *IEEE Trans. on Signal Processing*, 36:862–873, 1988.

[4] T. P. Bronez. On the performance advantage of multitaper spectral analysis. *IEEE Trans. on Signal Processing*, 40:2941–2946, 1992.

[5] G. C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75:236–255, 1987.

[6] A. D. Chave. A multitaper spectral estimator for time series with missing data. *Geophys. J. Inter.*, 218:2165–2178, 2019.

[7] C. L. Haley and M. Anitescu. Optimal bandwidth for multitaper spectrum estimation. *IEEE Signal Processing Letters*, 24(11):1696–1700, November 2017.

[8] C. L. Haley and C. J. Geoga. `Multitaper.jl`: a Julia library for multitaper nonparametric spectrum analysis. http://bitbucket.org/clhaley/Multitaper.jl, 2020.

[9] E. J. Hannan. *Multiple Time Series*. John Wiley and Sons, New York, 1970.

[10] L. H. Koopmans. *The Spectral Analysis of Time Series*. Academic Press, New York, second edition, 1995.

[11] J. C. Lee. On bias reduction in estimation of the magnitude–squared coherence function. *Bull. of Informatics and Cyberbetics*, 20:107–114, 1983.

[12] R. G. Miller. The jackknife – A review. *Biometrika*, 61:1–15, 1974.

[13] J. Park, F. L. Vernon, and C. R. Lindberg. Frequency dependent polarization analysis of high-frequency seismograms. *J Geophys. Res.*, 92:12,664–12,674, 1987.

[14] D. B. Percival and A. T. Walden. *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge Univ. Press, Cambridge, 1993.

[15] Ori Rosen and David S Stoffer. Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, 94(2):335–345, 2007.

[16] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, New York, 4 edition, 2001.

[17] D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: The discrete case. *Bell System Tech. J.*, 57(5):1371–1429, 1978.

[18] P. Stoica and T. Sundin. On nonparametric spectral estimation. *Circuits, Systems, and Signal Processing*, 18(2):169–181, 1999.

[19] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.

[20] D. J. Thomson. Jackknifing multitaper spectrum estimates. *IEEE Signal Processing Magazine*, 24(7):20–30, July 2007.

[21] D. J. Thomson and A. D. Chave. Jackknifed error estimates for spectra, coherences, and transfer functions. In Simon Haykin, editor, *Advances in Spectrum Analysis and Array Processing*, volume 1, chapter 2, pages 58–113. Prentice-Hall, Upper Saddle River, NJ, 1991.

[22] D. J. Thomson and C. L. Haley. Spacing and shape of random peaks in non–parametric spectrum estimates. *Proc. R. Soc. A*, 470(2167):20140101, 2014.