*Biometrika* (2018), **xx**, x, *pp*. 1–16
*Printed in Great Britain*

# The de-biased Whittle likelihood

BY ADAM M. SYKULSKI
*Data Science Institute / Department of Mathematics & Statistics, Lancaster University, Bailrigg, Lancaster LA1 4YW, U.K.*
a.sykulski@lancaster.ac.uk

SOFIA C. OLHEDE, ARTHUR P. GUILLAUMIN
*Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.*
s.olhede@ucl.ac.uk arthur.guillaumin.14@ucl.ac.uk

JONATHAN M. LILLY AND JEFFREY J. EARLY
*NorthWest Research Associates, 4118 148th Ave NE, Redmond, Washington 98052, U.S.A.*
lilly@nwra.com jearly@nwra.com

SUMMARY

The Whittle likelihood is a widely used and computationally efficient pseudo-likelihood. However, it is known to produce biased parameter estimates with finite sample sizes for large classes of models. We propose a method for de-biasing Whittle estimates for second-order stationary stochastic processes. The de-biased Whittle likelihood can be computed in the same $\mathcal{O}(n \log n)$ operations as the standard Whittle approach. We demonstrate the superior performance of the method in simulation studies and in application to a large-scale oceanographic dataset, where in both cases the de-biased approach reduces bias by up to two orders of magnitude, achieving estimates that are close to those of exact maximum likelihood, at a fraction of the computational cost. We prove that the method yields estimates that are consistent at an optimal convergence rate of $n^{-1/2}$ for Gaussian, as well as certain classes of non-Gaussian or non-linear processes. This is established under weaker assumptions than standard theory, where the power spectral density is not required to be continuous in frequency. We describe how the method can be readily combined with standard methods of bias reduction such as tapering and differencing, to further reduce bias in parameter estimates.

*Some key words*: Parameter estimation; Pseudo-likelihood; Fast Fourier Transform; Frequency Domain; Computational efficiency

## 1. INTRODUCTION

This paper introduces an improved computationally-efficient method of estimating time series model parameters of second-order stationary processes. The standard approach is to maximize the exact time-domain likelihood, which in general has computational efficiency of order $n^2$ for regularly-spaced Gaussian observations (where $n$ is the length of the observed time series) and produces estimates that are asymptotically efficient, converging at a rate of $n^{-1/2}$. A second approach is the method of moments, which in general has a computational efficiency of smaller order but with poorer statistical performance (Brockwell & Davis, 1991, p.253), exhibiting both bias and often a higher variance. A third approach of approximating the exact likelihood, often referred to as quasi-, pseudo-, or composite-likelihoods, is receiving much recent attention across statistics, see e.g. Fan et al. (2014) and Guinness & Fuentes (2017). In time series analysis, such likelihood approximations offer the possibility of considerable improvements in computational performance (usually scaling as order $n \log n$), with only small changes in statistical behaviour, see e.g. Anitescu et al. (2016). Here we introduce a pseudo-likelihood that is based on the Whittle likelihood (Whittle, 1953) which offers dramatic decreases in bias and mean-squared error in applications, yet with no significant increase in computational cost, and no loss in consistency or rate of convergence. We refer to our pseudo-likelihood as the de-biased Whittle likelihood.

The Whittle likelihood of Whittle (1953) is a frequency-domain approximation to the exact likelihood. This method is considered a standard method in parametric spectral analysis on account of its order $n \log n$ computational efficiency (Choudhuri et al., 2004; Fuentes, 2007; Matsuda & Yajima, 2009; Krafty & Collinge, 2013; Jesus & Chandler, 2017). However, it has been observed that the Whittle likelihood, despite its desirable asymptotic properties, may exhibit poor properties when applied to real-world, finite-length time series, particularly in terms of estimation bias (Dahlhaus, 1988; Velasco & Robinson, 2000; Contreras-Cristan et al., 2006). Bias is caused by spectral blurring, sometimes referred to as spectral leakage (Percival & Walden, 1993). Furthermore, when the time series model is specified in continuous time, but observed discretely, then there is the added problem of aliasing (see also Eckley & Nason, 2018), which if unaccounted for will further increase bias in Whittle estimates. The challenge is to account for such sampling effects and de-bias Whittle estimates, while retaining the computational efficiency of the method. We here define such a procedure, which can be combined with tapering and appropriate differencing, as recommended by Dahlhaus (1988) and Velasco & Robinson (2000). This creates an automated procedure that incorporates all modifications simultaneously, without any hand-tuning or reliance on process-specific analytic derivations such as in Taniguchi (1983).

We compare pseudo-likelihood approaches using simulated and real-world time series. In our example from oceanography, the de-biased Whittle likelihood results in parameter estimates that are significantly closer to maximum likelihood than standard Whittle estimates, while reducing the computational runtime of maximum likelihood by a factor of 100, thus demonstrating the practical utility and scalability of our method. Additionally, the theoretical properties of our new estimator are studied under relatively weak assumptions, in contrast to Taniguchi (1983), Dahlhaus (1988), and Velasco & Robinson (2000). Taniguchi studies autoregressive processes that depend on a scalar unknown parameter such that analytic calculations are possible. Dahlhaus examines processes whose spectral densities are the product of a known function with peaks that increase with sample size, and a latent spectral density that is twice continuously differentiable in frequency. Velasco and Robinson investigate processes that exhibit power-law behaviour at low frequencies and require continuous differentiability of the spectrum (at all frequencies except zero). Our assumptions on the spectral density of the time series will be milder. In particular, we will not require that the spectral density is continuous in frequency. This is a useful generalisation as discontinuous spectra frequently arise for example in oceanography (e.g. Polzin & Lvov, 2011, p. 11). Despite these weaker assumptions, we are able to prove consistency of de-biased Whittle estimates, together with a convergence rate matching the optimal $n^{-1/2}$, for large classes of Gaussian as well as non-Gaussian or non-linear processes.

## 2. DEFINITIONS AND NOTATION

We shall assume that the stochastic process of interest is modelled in continuous time, however, the de-biased Whittle likelihood can be readily applied to discrete-time models, as we shall demonstrate later. We define $\{X_t\}$ as the infinite sequence obtained from sampling a continuous-time real-valued process $X(t;\theta)$ with zero mean (or a known non-zero mean such that it can be subtracted a priori), where $\theta$ is a length-$p$ vector that specifies the process. That is, we let $X_t \equiv X(t\Delta;\theta)$, where $t$ is a positive or negative integer, $t = \ldots, -2, -1, 0, 1, 2, \ldots$, and $\Delta > 0$ is the sampling interval. If the process is second-order stationary, we define the autocovariance sequence by $s(\tau;\theta) \equiv E\{X_t X_{t-\tau}\}$ for $\tau = \ldots, -2, -1, 0, 1, 2, \ldots$, where $E\{\cdot\}$ is the expectation operator. The power spectral density of $\{X_t\}$ forms a Fourier pair with the autocovariance sequence, and is almost everywhere given by

$$f(\omega;\theta) = \Delta \sum_{\tau=-\infty}^{\infty} s(\tau;\theta)\exp(-\mathrm{i}\omega\tau\Delta), \quad s(\tau;\theta) = \frac{1}{2\pi}\int_{-\pi/\Delta}^{\pi/\Delta} f(\omega;\theta)\exp(\mathrm{i}\omega\tau\Delta)d\omega. \quad (1)$$

As $\{X_t\}$ is a discrete sequence, its Fourier representation is only defined up to the Nyquist frequency $\pm\pi/\Delta$. Thus there may be departures between $f(\omega;\theta)$ and the continuous-time process spectral density, denoted as $\tilde{f}(\omega;\theta)$, which for almost all $\omega \in \mathbb{R}$ is given by

$$\tilde{f}(\omega;\theta) = \int_{-\infty}^{\infty} \tilde{s}(\lambda;\theta)\exp(-\mathrm{i}\omega\lambda)d\lambda, \quad \tilde{s}(\lambda;\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{f}(\omega;\theta)\exp(\mathrm{i}\omega\lambda)d\omega. \quad (2)$$

Here $\tilde{s}(\lambda;\theta) \equiv E\{X(t)X(t-\lambda)\}$ (for $\lambda \in \mathbb{R}$) is the continuous-time process autocovariance, which is related to $s(\tau;\theta)$ via $\tilde{s}(\tau\Delta;\theta) = s(\tau;\theta)$, when $\tau$ is an integer. It follows that

$$f(\omega;\theta) = \sum_{k=-\infty}^{\infty} \tilde{f}\left(\omega + k\frac{2\pi}{\Delta};\theta\right), \quad \omega \in [-\pi/\Delta, \pi/\Delta], \quad (3)$$

see Percival & Walden (1993). Thus contributions to $\tilde{f}(\omega;\theta)$ outside of the range of frequencies $\pm\pi/\Delta$ are said to be folded or wrapped into $f(\omega;\theta)$. We have defined both $f(\omega;\theta)$ and $\tilde{f}(\omega;\theta)$, as both quantities are important in separating aliasing from other artefacts in spectral estimation.

In addition to these theoretical quantities, we will also require certain quantities that are computed directly from a single length-$n$ sample $\{X_t\}_{t=1}^n$. A widely used but inconsistent estimate of $f(\omega;\theta)$ is the periodogram, denoted $I(\omega)$, which is the squared absolute value of the Discrete Fourier Transform defined as

$$I(\omega) \equiv |J(\omega)|^2, \quad J(\omega) \equiv \left(\frac{\Delta}{n}\right)^{1/2} \sum_{t=1}^{n} X_t \exp(-\mathrm{i}\omega t\Delta), \quad \omega \in [-\pi/\Delta, \pi/\Delta]. \quad (4)$$

Note that $I(\omega)$ and $J(\omega)$ are taken to be properties of the observed realisation and are formally not regarded as functions of $\theta$.

## 3. MAXIMUM LIKELIHOOD AND THE WHITTLE LIKELIHOOD

Consider the discrete sample $X = \{X\}_{t=1}^n$, which is organized as a length $n$ column vector. Under the assumption that $X$ is drawn from $X(t;\theta)$, the expected $n \times n$ autocovariance matrix is $C(\theta) \equiv E\left\{XX^T\right\}$, where the superscript "$T$" denotes the transpose, and the components of $C(\theta)$ are given by $C_{ij}(\theta) = s\,(i-j;\theta)$. Exact maximum likelihood inference can be performed for Gaussian data by evaluating the log-likelihood function (Brockwell & Davis, 1991, p.254) given by

$$\ell(\theta) \equiv -\log|C(\theta)| - X^T C^{-1}(\theta)\, X, \tag{5}$$

where the superscript "$-1$" denotes the matrix inverse, and $|C(\theta)|$ is the determinant of $C(\theta)$. We have removed additive and multiplicative constants not affected by $\theta$ in (5). The optimal choice of $\theta$ for our chosen model to characterize the sampled time series $X$ is then found by maximizing the log-likelihood function in (5) such that

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell(\theta),$$

where $\Theta$ defines the parameter space of $\theta$. Because the time-domain maximum likelihood is known to have optimal properties, any other estimator will be compared with the properties of this quantity.

A standard technique for avoiding expensive matrix inversions is to approximate (5) in the frequency domain, following the seminal work of Whittle (1953). This approach approximates $C(\theta)$ using a Fourier representation, and utilizes the special properties of Toeplitz matrices. Given the observed sampled time series $X$, the Whittle likelihood, denoted $\ell_W(\theta)$, is

$$\ell_W(\theta) \equiv -\sum_{\omega\in\Omega} \left\{ \log \tilde{f}(\omega;\theta) + \frac{I(\omega)}{\tilde{f}(\omega;\theta)} \right\}, \tag{6}$$

where $\Omega$ is the set of discrete Fourier frequencies given by

$$\Omega \equiv (\omega_1, \omega_2, \ldots, \omega_n) = \frac{2\pi}{n\Delta} \left(-\lceil n/2 \rceil + 1, \ldots, -1, 0, 1, \ldots, \lfloor n/2 \rfloor\right). \tag{7}$$

The subscript "$W$" in $\ell_W(\theta)$ is used to denote "Whittle." We have presented the Whittle likelihood in a discretized form here, as its usual integral representation must be approximated for finite-length time series. The Whittle likelihood approximates the time-domain likelihood, i.e. $\ell(\theta) \approx \ell_W(\theta)$, and this statement can be made precise (Dzhaparidze & Yaglom, 1983). Its computational efficiency is $\mathcal{O}(n \log n)$, as the periodogram can be computed using the Fast Fourier Transform, thus explaining its popularity in practice. Exact maximum likelihood, on the other hand, will require $\mathcal{O}(n^2)$ computations for regularly-sampled Gaussian processes (using the Trench algorithm (Trench, 1964) for example to compute (5)), and often higher complexity for non-Gaussian processes, as demonstrated in our non-Gaussian simulation example in Section S1·3 of the Supplementary Material.

The Whittle likelihood (6) is calculated using the periodogram, $I(\omega)$. This spectral estimate, however, is known to be a biased measure of the continuous-time process's spectral density for finite samples, due to blurring and aliasing effects (Percival & Walden, 1993), as discussed in the introduction. Aliasing results from the discrete sampling of the continuous-time process to generate an infinite sequence, whereas blurring is associated with the truncation of this infinite sequence over a finite-time interval. The desirable properties of the Whittle likelihood rely on the asymptotic behaviour of the periodogram for large sample sizes. The bias of the periodogram for finite samples, however, will translate into biased parameter estimates from the Whittle likelihood, as has been widely reported (see e.g. Dahlhaus, 1988). In the next section we propose a procedure for de-biasing Whittle estimates.

# 4. MODIFIED PSEUDO-LIKELIHOODS

## 4·1. *The de-biased Whittle likelihood*

We introduce the following pseudo-likelihood function given by

$$\ell_D(\theta) \equiv -\sum_{\omega\in\Omega} \left\{ \log \overline{f}_n(\omega;\theta) + \frac{I(\omega)}{\overline{f}_n(\omega;\theta)} \right\}, \tag{8}$$

$$\overline{f}_n(\omega;\theta) = \int_{-\pi/\Delta}^{\pi/\Delta} f(\nu;\theta)\mathcal{F}_{n,\Delta}(\omega-\nu)\, d\nu, \quad \mathcal{F}_{n,\Delta}(\omega) \equiv \frac{\Delta}{2\pi n}\frac{\sin^2(n\omega\Delta/2)}{\sin^2(\omega\Delta/2)}, \tag{9}$$

where the subscript "$D$" stands for "de-biased." Here $\tilde{f}(\omega;\theta)$ in (6) has been replaced by $\overline{f}_n(\omega;\theta) \equiv E\{I(\omega)\}$, which is the expected periodogram, and may be shown to be given by the convolution of the true modelled spectrum with the Fejér kernel $\mathcal{F}_{n,\Delta}(\omega)$. We call (8) the de-biased Whittle likelihood, where the set $\Omega$ is defined as in (7).

Replacing the true spectrum $\tilde{f}(\omega;\theta)$ with the expected periodogram $\overline{f}_n(\omega;\theta)$ in (8) is a straightforward concept, however, our key innovation lies in formulating its efficient computation without losing $\mathcal{O}(n\log n)$ efficiency. If we directly use (9), then this convolution would usually need to be approximated numerically, and could be computationally expensive. Instead we utilize the convolution theorem to express the frequency-domain convolution as a time-domain multiplication (Percival & Walden, 1993, p.198), such that

$$\overline{f}_n(\omega;\theta) = 2\Delta\cdot\Re\left\{\sum_{\tau=0}^{n-1}\left(1-\frac{\tau}{n}\right)s(\tau;\theta)\exp(-\mathrm{i}\omega\tau\Delta)\right\} - \Delta\cdot s(0;\theta), \tag{10}$$

where $\Re\{\cdot\}$ denotes the real part. Therefore $\overline{f}_n(\omega;\theta)$ can be exactly computed at each Fourier frequency directly from $s(\tau;\theta)$, provided its functional form is known for $\tau = 0,\ldots,n-1$, by using a Fast Fourier Transform in $\mathcal{O}(n\log n)$ operations. Care must be taken to subtract the variance term, $\Delta\cdot s(0;\theta)$, to avoid double counting contributions from $\tau=0$. Both aliasing and blurring effects are automatically and conveniently accounted for in (10) in one operation; aliasing is accounted for by sampling the theoretical autocovariance function at discrete times, while the effect of blurring is accounted for by the truncation of this sequence to finite length, and the inclusion of the triangle function $(1-\tau/n)$ in the expression. The result is that $\overline{f}_n(\omega;\theta)$ is a blurred and aliased version of the true spectrum $\tilde{f}(\omega;\theta)$, which reflects the blurring and aliasing artefacts present in the periodogram.

The de-biased Whittle likelihood can also be used with discrete-time processes, as (10) can be computed from the theoretical autocovariance sequence of the discrete process in exactly the same way. If the analytic form of $s(\tau;\theta)$ is unknown or expensive to evaluate, then it can be approximated from the spectral density using Fast Fourier Transforms, thus maintaining $\mathcal{O}(n\log n)$ computational efficiency.

Computing the standard Whittle likelihood of (6) with the aliased spectrum $f(\omega;\theta)$ defined in (1), without accounting for spectral blurring, would in general be more complicated than using the expected periodogram $\overline{f}_n(\omega;\theta)$. This is because the aliased spectrum $f(\omega;\theta)$ seldom has an analytic form for continuous processes, and must be instead approximated by either explicitly wrapping in contributions from $\tilde{f}(\omega;\theta)$ from frequencies higher than the Nyquist as in (3), or via an approximation to the Fourier transform in (1). This is in contrast to the de-biased Whittle likelihood, where the effects of aliasing and blurring have been computed exactly in one single operation using (10). Thus addressing aliasing and blurring together using the de-biased Whittle likelihood is simpler and computationally faster to implement than accounting for aliasing alone. This will become further apparent in the simulation studies of Section 5·1.

### 4·2. *Combining with differencing or tapering*

A standard technique for reducing the effects of blurring on Whittle estimates is to apply the Whittle likelihood to the differenced process (Velasco & Robinson, 2000), as often differencing will decrease the dynamic range of the spectrum and hence decrease broadband blurring. The de-biased Whittle likelihood can be readily implemented with differenced data. If we denote the differenced process by $U_t = X_{t+1} - X_t$, then the expected periodogram of (10) can be computed using the autocovariance of $U_t$, which is found from the autocovariance of $X_t$ via $s_U(\tau) \equiv E\{U_t U_{t-\tau}\} = 2s_X(\tau) - s_X(\tau+1) - s_X(\tau-1)$, such that the procedure remains $\mathcal{O}(n \log n)$.

Another standard approach of ameliorating the effects of blurring is to pre-multiply the data sequence with a weighting function known as a data taper (Thomson, 1982). The taper is chosen to have spectral properties such that broadband blurring will be minimized, and the variance of the spectral estimate at each frequency is reduced, although the trade-off is that tapering increases narrowband blurring as the correlation between neighbouring frequencies increases.

The tapered Whittle likelihood (Dahlhaus, 1988) corresponds to replacing the direct spectral estimator formed from $I(\omega)$ in (4) with one using the taper $h = \{h_t\}$

$$J(\omega; h) \equiv \Delta^{1/2} \sum_{t=1}^{n} h_t X_t \exp(-\mathrm{i}\omega t \Delta), \quad I(\omega; h) \equiv |J(\omega; h)|^2, \quad \sum_{t=1}^{n} h_t^2 = 1, \tag{11}$$

where $h_t$ is real-valued. Setting $h_t = 1/n^{1/2}$ for $t = 1, \ldots n$ recovers the periodogram estimate of (6). To estimate parameters we then maximize

$$\ell_T(\theta) \equiv -\sum_{\omega \in \Omega} \left\{ \log \tilde{f}(\omega; \theta) + \frac{I(\omega; h)}{\tilde{f}(\omega; \theta)} \right\}, \tag{12}$$

where the subscript "$T$" denotes that a taper has been used. Velasco & Robinson (2000) demonstrated that for certain discrete processes it is beneficial to use this estimator, rather than the standard Whittle likelihood, for parameter estimation, particularly when the spectrum exhibits a high dynamic range. Nevertheless, tapering will not remove all broadband blurring effects in the likelihood, because we are still comparing the tapered spectral estimate against the theoretical spectrum, and not against the expected tapered spectral estimate. Furthermore, there remain the issues of narrowband blurring, as well as aliasing effects with continuous sampled processes.

Our de-biasing procedure can be naturally combined with tapering. We define the pseudo-likelihood

$$\ell_{TD}(\theta) \equiv -\sum_{\omega \in \Omega} \left\{ \log \overline{f}_n(\omega; h, \theta) + \frac{I(\omega; h)}{\overline{f}_n(\omega; h, \theta)} \right\}, \tag{13}$$

$$\overline{f}_n(\omega; h, \theta) = \int_{-\pi/\Delta}^{\pi/\Delta} f(\nu; \theta) \mathcal{H}_\Delta(\omega - \nu)\, d\nu, \quad \mathcal{H}_\Delta(\omega) \equiv \Delta \left| \sum_{t=1}^{n} h_t \exp(-\mathrm{i}\omega t \Delta) \right|^2,$$

with $I(\omega; h)$ as defined in (11) such that $\overline{f}_n(\omega; h, \theta) \equiv E\{I(\omega; h)\}$. We call $\ell_{TD}(\theta)$ the de-biased tapered Whittle likelihood and $\overline{f}_n(\omega; h, \theta)$ the expected tapered spectral estimate. The function $\overline{f}_n(\omega; h, \theta)$ can be computed exactly using a $\mathcal{O}(n \log n)$ calculation similar to (10), i.e.

$$\overline{f}_n(\omega; h, \theta) = 2\Delta \cdot \Re \left\{ \sum_{\tau=0}^{n-1} s(\tau; \theta) \left( \sum_{t=1}^{n-\tau} h_t h_{t+\tau} \right) \exp(-\mathrm{i}\omega \tau \Delta) \right\} - \Delta \cdot s(0; \theta).$$

Accounting in $\overline{f}_n(\omega; h, \theta)$ for the particular taper used accomplishes de-biasing of the tapered Whittle likelihood, just as using the expected periodogram does for the standard Whittle likelihood. The time-domain kernel $\sum_{t=1}^{n-\tau} h_t h_{t+\tau}$, for $\tau = 0, \ldots, n-1$, can be pre-computed using FFTs or using a known analytical form. Then during optimization, an FFT of this fixed kernel multiplied by the autocovariance sequence is taken at each iteration. Thus the de-biased tapered Whittle likelihood is also an $\mathcal{O}(n \log n)$ pseudo-likelihood estimator.

Both the de-biased Whittle and de-biased tapered Whittle likelihoods have their merits, but the trade-offs are different with nonparametric spectral density estimation than they are with parametric model estimation. Specifically, although tapering decreases the variance of nonparametric estimates at each frequency, it conversely may increase the variance of estimated parameters. This is because the taper is reducing degrees of freedom in the data, which increases correlations between local frequencies. On the other hand, the periodogram creates broadband correlations between frequencies, especially for processes with a high dynamic range, which also contributes to variance in parameter estimates. We explore these trade-offs in greater detail in Section 5·1.

## 5. Simulations and Applications

### 5·1. *The Matérn Process*

In this section we investigate the performance of the de-biased Whittle likelihood in a Monte Carlo study using observations from a Matérn process (Matérn, 1960), as motivated by the simulation studies of Anitescu et al. (2012) who investigate the same process. The Matérn process is a three-parameter continuous Gaussian process defined by its continuous-time unaliased spectrum

$$\tilde{f}(\omega) = \frac{A^2}{(\omega^2 + c^2)^\alpha}, \quad \omega \in \mathbb{R}. \tag{14}$$

The parameter $A \geq 0$ sets the magnitude of the variability, $1/c > 0$ is the damping timescale, and $\alpha > 1/2$ controls the rate of spectral decay, or equivalently the smoothness or differentiability of the process. For large $\alpha$ the power spectrum exhibits a high dynamic range, and the periodogram will be a poor estimator of the spectral density due to blurring. Conversely, for small $\alpha$ there will be departures between the periodogram and the continuous-time spectral density because of aliasing. We therefore investigate the performance of estimators over a range of $\alpha$ values, and this motivates why the Matérn in a suitable process to study.

In Table 1 we display the average percentage bias, standard deviation and RMSE (relative to the true parameter values) for six different pseudo-likelihoods: standard Whittle likelihood with both the observed and differenced processes (6), the tapered Whittle likelihood (12), and the de-biased versions of each (equations (8) and (13)). Our choice of data taper is the Discrete Prolate Spheroidal Sequence (DPSS) taper (Slepian & Pollak, 1961), with bandwidth parameter equal to 4, where performance was found to be broadly similar across different choices of bandwidth (not shown). We also include results for exact maximum likelihood (5). The results are averaged over estimates of the three parameters $\{A, \alpha, c\}$ which are all assumed unknown, where the true $\alpha$ varies from $[0.6, 2.5]$ in intervals of $0.1$, and we fix $A = 1$ and $c = 0.2$. This is to explore performance over spectra that have aliasing artefacts as well as high dynamic range. For each value of $\alpha$, we simulate 10,000 time series each of length $n = 1000$. The optimization is performed in MATLAB using `fminsearch`, and uses identical settings for all likelihoods. Initialized guesses for the slope and amplitude are found using a least squares fit in the range $[\pi/4\Delta, 3\pi/4\Delta]$, and the initial guess for the damping parameter $c$ is set at a mid-range value of 100 times the Rayleigh frequency (i.e. $c = 100\pi/n = \pi/10$.)

The performance of all standard Whittle methods are significantly contaminated by bias. The de-biased variants remove this bias by an order of magnitude. The standard deviation is broadly

Table 1: Average percentage bias, standard deviation (SD), and root-mean-square-error (RMSE) (relative to the true parameter values) for a Matérn process across all estimates of $\{A, \alpha, c\}$ with $n = 1000$

| Inference Method | Eqn | Bias | SD | RMSE |
|---|---|---|---|---|
| Standard Whittle (periodogram) | (6) | 23.69% | 10.34% | 26.66% |
| De-Biased Whittle (periodogram) | (8) | 3.96% | 12.97% | 13.75% |
| Standard Whittle (tapered) | (12) | 18.11% | 12.23% | 23.12% |
| De-Biased Whittle (tapered) | (13) | 2.60% | 14.15% | 14.41% |
| Standard Whittle (differenced) | (6) | 18.99% | 9.33% | 22.09% |
| De-Biased Whittle (differenced) | (8) | 1.19% | 8.90% | 8.99% |
| Maximum Likelihood | (5) | 1.10% | 7.60% | 7.68% |

similar across all Whittle methods, where tapering results in standard deviations that are approximately twice that of maximum likelihood, which is consistent with the loss of information from using a data taper. Of all pseudo-likelihood estimators considered, the de-biased Whittle likelihood using the differenced process performs best, and yields results close to exact maximum likelihood. Overall, of the three modifications to the standard Whittle likelihood—de-biasing, tapering and differencing—the de-biasing method proposed here is the single procedure that yields the greatest overall improvement in parameter estimation.

In Figure S1 of the supplementary material, we provide a figure which separates out the bias and RMSE improvements over different values of $\alpha$, demonstrating that the de-biased Whittle likelihood can effectively address bias from aliasing when $\alpha$ is low, and bias from blurring when $\alpha$ is high. Furthermore, in Table S1 of the supplementary material, we include a comparison with a time-domain $\mathcal{O}(n \log n)$ estimator from Anitescu et al. (2012) which is found to perform similarly to the de-biased Whittle likelihood with differenced data in terms of bias and RMSE, although we note that the latter is computed in a fraction of the computational time.

In the next section we will prove that the de-biased Whittle likelihood is a consistent estimator converging at the optimal $n^{-1/2}$ rate, under assumptions which are satisfied by the Matérn process. Motivated by this, we perform an additional experiment over different lengths of time series $n = 2^k$, with $k$ taking integer values from 7 to 13, such that $n$ ranges from 128 to 8192. To isolate the convergence of the parameter estimate, we fix $A = 1$ and $c = 0.2$ as before but this time assume these are known, and now only estimate the slope parameter which we set to $\alpha = 2$.

The average bias, standard deviation and RMSE of each estimator are displayed in Fig. 1, together with average CPU times. We include results for the standard and de-biased Whittle likelihood, as well as exact maximum likelihood. Motivated by Table 1, we include the de-biased Whittle likelihood with differenced data. Finally, we include results for the standard Whittle likelihood using an approximated aliased spectrum, which we find using (3) by truncating the summation limits to $\pm 5$ to keep the computation efficient. The reason for including an approximate aliased version of standard Whittle is to show that bias corrections are made by the de-biased Whittle with regards to both blurring and aliasing. Here we see that the standard Whittle likelihood using the unaliased spectrum of (14) performs poorly with increasing $n$ due to bias—this is because for growing domain asymptotics, bias due to aliasing does not decrease as $n$ increases.

The standard deviations of all estimates converge at a rate consistent with $n^{-1/2}$, as Theorem 1 of Section 6·1 will prove for de-biased methods. Overall, the de-biased approaches provide a good balance between statistical and computational efficiency over all sample sizes, where in contrast exact maximum likelihood is only computed up to $n = 2048$ due to rapidly increasing computational costs. Standard Whittle with differenced data is not included for clarity of presen-

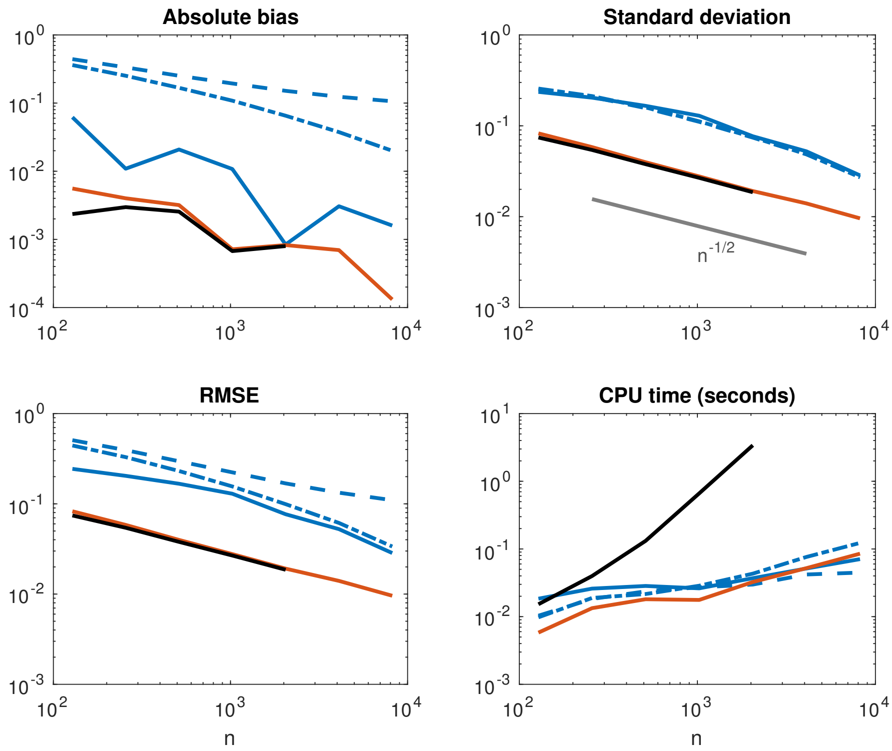


Fig. 1: Performance of five different likelihoods: Standard Whittle unaliased (- - -), Standard Whittle aliased (-·-·-) De-biased Whittle periodogram (—), De-biased Whittle differenced (—) and exact maximum likelihood (—). Repeated 1,000 times over various values of $n$ when estimating the slope parameter ($\alpha = 2$) of a Matérn process. Note the axes are on a log-log scale. CPU times are as performed on a 2.2 GHz Intel Core i7 processor.

tation, but was found to perform worse than de-biased Whittle with differenced data, consistent with Table 1.

### 5·2. *Application to Large-Scale Oceanographic Data*

In this section we examine the performance of our method when applied to a real-world large-scale dataset, by analysing data obtained from the Global Drifter Program (`http://www.aoml.noaa.gov/phod/dac/index.php`), which maintains a publicly-downloadable database of position measurements obtained from freely-drifting satellite-tracked oceanic instruments known as drifters. In total over 23,000 drifters have been deployed, with interpolated six-hourly data available since 1979 and one-hourly data since 2005 (see Elipot et al., 2016), with over 100 million data points available in total. The collection of such data is pivotal to the understanding of ocean circulation and its impact on the global climate system (Griffa et al., 2007); it is therefore essential to have computationally efficient methods for their analysis.

In Fig. 2, we display 50-day position trajectories and corresponding velocity time series for three drifters from the one-hourly data set, each from a different major ocean. These trajectories can be considered as complex-valued time series, with the real part corresponding to the east/west velocity component and the imaginary part corresponding to the north/south velocity component. We then plot the periodogram of the complex-valued series, which has different power at positive

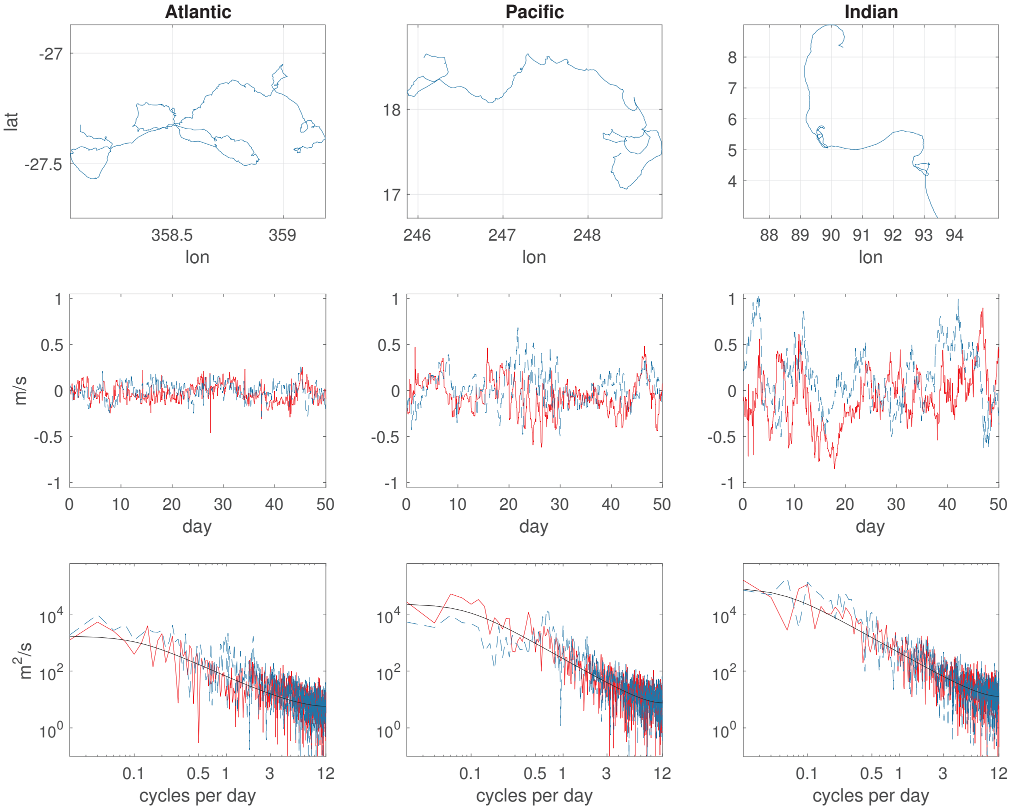


Fig. 2: The top row displays 50-day trajectories of Drifter IDs: #2339255 (Atlantic Ocean), #49566 (Pacific Ocean), and #43577 (Indian Ocean). The second row displays the east/west (red-solid) and north/south (blue-dashed) velocity time series for each trajectory. The third row displays the periodograms of the complex-valued velocity series, with the non-inertial side of the spectrum in red-solid, and the inertial side in blue-dashed. The expected periodogram, $\overline{f}_n(\omega;\hat{\theta})$, from the de-biased Whittle likelihood is overlaid in black. Note that the frequency axis is on a logarithmic scale.

Table 2: Estimated Matérn parameters (with corresponding estimated standard errors in parenthesis) using the maximum, de-biased Whittle, and standard Whittle, likelihoods for the velocity time series of Fig. 2; the parameters are given in terms of the damping timescale $(1/c)$, the slope $(2\alpha)$ and the diffusivity $(\kappa)$; CPU times are as performed on a 2.8 GHz Intel Core i7 processor

| Drifter location | Inference method | Damping (days) | Slope (dimensionless) | Diffusivity ($m^2/s\times10^3$) | CPU (s) |
|---|---|---|---|---|---|
| | ML | 10.65 (2.49) | 1.460 (0.023) | 0.49 (0.18) | 7.42 |
| Atlantic | De-biased Whittle | 9.84 (5.51) | 1.462 (0.062) | 0.44 (0.28) | 0.16 |
| | Standard Whittle | 30.19 (16.2) | 1.097 (0.043) | 0.65 (0.36) | 0.04 |
| | ML | 10.62 (1.85) | 1.829 (0.024) | 5.09 (1.71) | 7.47 |
| Pacific | De-biased Whittle | 11.82 (4.64) | 1.827 (0.048) | 6.00 (3.83) | 0.10 |
| | Standard Whittle | 19.59 (6.51) | 1.575 (0.036) | 7.18 (3.60) | 0.02 |
| | ML | 21.76 (4.83) | 1.825 (0.025) | 30.48 (12.9) | 10.06 |
| Indian | De-biased Whittle | 19.90 (9.41) | 1.802 (0.053) | 22.70 (17.2) | 0.10 |
| | Standard Whittle | 39.99 (16.9) | 1.545 (0.038) | 31.19 (19.7) | 0.02 |

and negative frequencies, distinguishing directions of rotation on the complex plane (Schreier & Scharf, 2010). The de-biased Whittle likelihood for complex-valued proper processes is exactly the same as (8)–(10) (see also Sykulski et al., 2016), where the autocovariance sequence of a complex-valued process $Z_t$ is $s(\tau;\theta) = E\{Z_t Z_{t-\tau}^*\}$. For proper processes the complementary covariance is $r(\tau;\theta) = E\{Z_t Z_{t-\tau}\} = 0$ at all lags (Schreier & Scharf, 2010), and can thus be ignored in the likelihood, as $s(\tau;\theta)$ captures all second-order structure in the zero-mean process.

We model the velocity time series as a complex-valued Matérn process, with power spectral density given in (14), as motivated by Sykulski et al. (2016) and Lilly et al. (2017). To account for a type of circular oscillations in each time series known as inertial oscillations, which create an off-zero spike on one side of the spectrum, we fit the Matérn process semi-parametrically to the opposite "non-inertial" side of the spectrum (as displayed by the red-solid line in the figure). We overlay the fit of the de-biased Whittle likelihood to the periodograms in Fig. 2. For a full parametric model of surface velocity time series, see Sykulski et al. (2016). We have selected drifters without noticeable tidal effects; for de-tiding procedures see Pawlowicz et al. (2002).

We estimate the Matérn parameters for each time series using the de-biased and regular Whittle likelihood, as well as exact maximum likelihood. The latter of these methods can be performed over only positive or negative frequencies by first decomposing the time series into analytic and anti-analytic components using the discrete Hilbert transform, see Marple (1999), and then fitting the corresponding signal to an adjusted Matérn autocovariance that accounts for the effects of the Hilbert transform. The details for this procedure are provided in the online code.

The parameter estimates from the three likelihoods are displayed in Table 2, along with the corresponding CPU times. We also provide estimated parameter standard errors using the methodology described in Section 6·2, with more details in the online code. We reparametrize the Matérn to output three important oceanographic quantities: the damping timescale, the decay rate of the spectral slope, and the diffusivity (which is the rate of particle dispersion) given by $\kappa \equiv A^2/4c^{2\alpha}$ (Lilly et al., 2017, eq.(43)). From Table 2 it can be seen that the de-biased Whittle and maximum likelihoods yield similar values for the slope and damping timescale, however, regular Whittle likelihood yields parameters that underestimate the slope by around 15%, and overestimate the damping timescale by a factor of two, which if used would incorrectly specify under-damped and rougher trajectories than expected. These biases are consistent with the significant biases reported in Section 5·1. Diffusivity estimates vary across all estimation procedures and have large standard errors, and this variability is likely due to the fact that diffusivity is a measure of the spectrum at frequency zero, hence estimation is performed over relatively few frequencies.

Maximum likelihood is two orders of magnitude slower to execute than de-biased Whittle. When this is scaled to fitting all time series in the Global Drifter Program database, then time-domain maximum likelihood becomes impractical. The de-biased Whittle likelihood, on the other hand, retains the speed of Whittle likelihood, while returning estimates that are close to maximum likelihood. This section therefore serves as a proof of concept of how the de-biased Whittle likelihood is a useful tool for efficient parameter estimation from large datasets.

## 5·3. *Autoregressive Processes*

Here we investigate the performance of the de-biased Whittle likelihood when estimating parameters of a discrete-time autoregressive (AR) process, $X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t$, where $\epsilon_t \sim^{iid} \mathcal{N}(0, \sigma^2)$. Specifically, we generate time series from the AR(4) process studied in Percival & Walden (1993), used throughout the book as a motivating example of a process that generates high spectral blurring in spectral density estimation. As the process is discrete-time, then there is no issue with aliasing, and this example therefore assesses how well the de-biased Whittle likelihood accounts for bias purely due to blurring.

Table 3: Average parameter estimates and root mean square errors of estimating all five AR(4) parameters using different estimation methods, for $n = 256$ and $n = 1024$. Results are obtained over 1,000 replicated time series for each time series length

| AR(4) Parameters: | $\phi_1 = 2.7607$ | $\phi_2 = -3.8106$ | $\phi_3 = 2.6535$ | $\phi_4 = -0.9238$ | $\sigma = 1$ |
|---|---|---|---|---|---|
| | | $n = 256$ | | | |
| | | Average parameter estimate | | | |
| Yule-Walker | 1.7669 | -1.6555 | 0.6081 | -0.1685 | 4.3621 |
| Standard Whittle | 1.8989 | -1.9485 | 0.8895 | -0.2746 | 4.0231 |
| De-Biased Whittle | 2.5309 | -3.3065 | 2.1754 | -0.7439 | 1.5341 |
| Maximum Likelihood | 2.7478 | -3.7490 | 2.5799 | -0.8798 | 1.0525 |
| | | Root mean square error | | | |
| Yule-Walker | 1.0591 | 2.2725 | 2.1473 | 0.7861 | 3.6499 |
| Standard Whittle | 0.9800 | 2.0775 | 1.9523 | 0.7076 | 3.5326 |
| De-Biased Whittle | 0.5136 | 1.0539 | 0.9777 | 0.3456 | 1.7368 |
| Maximum Likelihood | 0.0330 | 0.1618 | 0.2031 | 0.1318 | 0.2320 |
| | | $n = 1024$ | | | |
| | | Average parameter estimates | | | |
| Yule-Walker | 2.1959 | -2.5237 | 1.4004 | -0.4328 | 2.9207 |
| Standard Whittle | 2.2642 | -2.6878 | 1.5644 | -0.5008 | 2.7092 |
| De-Biased Whittle | 2.7030 | -3.6704 | 2.5161 | -0.8665 | 1.0370 |
| Maximum Likelihood | 2.7574 | -3.8006 | 2.6428 | -0.9185 | 1.0028 |
| | | Root mean square error | | | |
| Yule-Walker | 0.6409 | 1.4513 | 1.4094 | 0.5489 | 2.1214 |
| Standard Whittle | 0.6225 | 1.3923 | 1.3441 | 0.5164 | 2.0663 |
| De-Biased Whittle | 0.2001 | 0.4346 | 0.4133 | 0.1550 | 0.6632 |
| Maximum Likelihood | 0.0131 | 0.0438 | 0.0526 | 0.0326 | 0.0623 |

In Table 3 we display the average parameter estimates and root mean square errors when estimating all five parameters of the AR(4) process $\{\phi_1, \phi_2, \phi_3, \phi_4, \sigma\}$. We contrast four approaches: maximum likelihood, the standard Whittle likelihood, the de-biased Whittle likelihood, and the standard Yule-Walker estimation procedure; the latter of which is used to initialise parameter estimates for the likelihood-based methods. We do not include results using the differenced process as this was not found to yield improved parameter estimates for this particular example.

Yule-Walker and standard Whittle estimates perform similarly and quite poorly with both sample sizes considered, which is consistent with the fact that the former uses the biased sample autocovariance to solve the Yule-Walker equations, and the latter uses the periodogram, which is the Fourier pair of the biased sample autocovariance. The de-biased Whittle likelihood accounts for this bias and yields average estimates that are close to exact maximum likelihood and the true values, and eliminates around half the root mean square error when $n = 256$, and two thirds when $n = 1024$. The de-biased Whittle likelihood is therefore an effective pseudo-likelihood for discrete-time as well as continuous-time processes. In Section S1 of the supplementary material we include further simulation results, including a performance comparison for a non-Gaussian process, where again the de-biased Whittle likelihood is found to provide a good trade-off between statistical and computational efficiency.

## 6. Properties of the De-Biased Whittle Likelihood

### 6·1. *Consistency and optimal convergence rates*

In this section, we establish consistency and optimal convergence rates for de-biased Whittle estimates with Gaussian and certain classes of non-Gaussian or non-linear processes. To show that de-biased Whittle estimates converge at the optimal rate, the main challenge is that although our pseudo-likelihood accounts for the bias of the periodogram, there is still present the correlation between different frequencies caused by the leakage associated with the Fejér kernel. This is what prevents the de-biased Whittle likelihood from being exactly equal to the time-domain maximum likelihood for Gaussian data. To establish optimal convergence rates, we bound the asymptotic behaviour of this correlation. The statement is provided in Theorem 1, with the proof provided in Section S2 of the supplementary material. The proof is composed of several lemmas which, for example, place useful bounds on the expected periodogram, the variance of linear combinations of the periodogram at different frequencies, and also the first and second derivatives of the de-biased Whittle likelihood. Together these establish that the de-biased Whittle likelihood is a consistent estimator with estimates that converge in probability at an optimal rate of $n^{-1/2}$, under relatively weak assumptions.

THEOREM 1. *Assume that $\{X_t\}$ is an infinite sequence obtained from sampling a zero-mean continuous-time real-valued process $X(t;\theta)$, which satisfies the following assumptions:*

1. *The parameter set $\Theta \subset \mathbb{R}^p$ is compact with a non-null interior, and the true length-$p$ parameter vector $\theta$ lies in the interior of $\Theta$.*
2. *For all $\theta \in \Theta$ and $\omega \in [-\pi, \pi]$, the spectral density of the sequence $\{X_t\}$ is bounded below by $f(\omega;\theta) \geq f_{\min} > 0$, and bounded above by $f(\omega;\theta) \leq f_{\max}$.*
3. *$\theta_1 \neq \theta_2$ implies $f(\cdot;\theta_1) \neq f(\cdot;\theta_2)$ on a set of positive Lebesgue measure.*
4. *$f(\omega;\theta)$ is continuous in $\theta$ and Riemann integrable in $\omega$.*
5. *The expected periodogram $\overline{f}_n(\omega;\theta)$, as defined in (9) of the main body, has two continuous derivatives in $\theta$ which are bounded above in magnitude uniformly for all $n$, where the first derivative in $\theta$ also has $\Theta(n)$ frequencies in $\Omega$ that are non-zero.*
6. *$\{X_t\}$ is a fourth-order stationary process with finite fourth-order moments and absolutely summable fourth-order cumulants.*

*Then the estimator*

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_D(\theta),$$

*for a sample $\{X_t\}_{t=1}^n$, where $\ell_D(\theta)$ is the de-biased Whittle likelihood of* (8)*, satisfies*

$$\hat{\theta} = \theta + \mathcal{O}_P\left(n^{-1/2}\right).$$

The fourth-order cumulant is formally defined in (2) in Section S2 of the supplementary material. All stationary Gaussian processes automatically satisfy Assumption 6 as the fourth-order cumulant is identically zero. In Section S3 of the supplementary material, we provide a class of non-linear processes and prove that this class satisfies Assumption 6. Specifically, we study the process $Y_t = X_t^2$ where $X_t$ is a Gaussian process with bounded spectral density and absolutely summable autocovariance.

### 6·2. *Standard error estimation*

Here we provide a novel method of obtaining standard error estimates for de-biased Whittle estimates. This method was used to calculate standard errors in our application example in Table 2. From equations (24), (27), and (28) in the supplementary material we see that the $p \times p$ covariance matrix of the estimated vector $\hat{\theta}$ satisfies

$$\mathrm{var}\left\{\hat{\theta}\right\} = E\{H(\theta)\}^{-1}\,\mathrm{var}\left\{\nabla \ell_D(\theta)\right\} E\{H(\theta)\}^{-1}\left(1 + o(1)\right), \tag{15}$$

where $\nabla = [\partial/\partial\theta_1\ \partial/\partial\theta_2 \ldots \partial/\partial\theta_p]^T$, $\nabla \ell_D(\theta)$ is known as the score. The $p \times p$ matrix $H(\theta)$, known as the Hessian, is defined entrywise by $H_{ij}(\theta) = \partial^2 \ell_D(\theta)/\partial\theta_i\partial\theta_j$, and its expectation can be approximated either analytically or numerically at $\hat{\theta}$. The remaining term in (15), $\mathrm{var}\{\nabla \ell_D(\theta)\}$, is the $p \times p$ covariance matrix of the score. The diagonal elements in this matrix, which are variances of individual components of the score, can be expressed from (8) as

$$\begin{aligned}\mathrm{var}\left\{\frac{\partial}{\partial\theta_i}\ell_D(\theta)\right\} &= \mathrm{var}\left\{\sum_{\omega\in\Omega}\frac{\partial \overline{f}_n(\omega;\theta)}{\partial\theta_i}\cdot\frac{I(\omega)}{\overline{f}_n^2(\omega;\theta)}\right\} = \mathrm{var}\left\{\sum_{j=1}^{n} a_{ij}(\theta) I(\omega_j)\right\}\\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{ij}(\theta)a_{ik}(\theta)\,\mathrm{cov}\left\{I(\omega_j), I(\omega_k)\right\},\end{aligned}$$

where $w_j$ are the elements of $\Omega$ as defined in (7) and we have defined

$$a_{ij}(\theta) \equiv \frac{\partial \overline{f}_n(\omega_j;\theta)}{\partial\theta_i}\cdot\frac{1}{\overline{f}_n^2(\omega_j;\theta)}.$$

Here we have made use of the fact that the $\partial \log\{\overline{f}_n(\omega;\theta)\}/\partial\theta$ term is deterministic and therefore has no variance. As we have established asymptotic efficiency for $\hat{\theta}$ we can now use the invariance principle of maximum likelihood (Casella & Berger, 2002, p.320) to construct an estimator of the variance, that is

$$\widehat{\mathrm{var}}\left\{\frac{\partial}{\partial\theta_i}\ell_D(\theta)\right\} = \sum_{j=1}^{n}\sum_{k=1}^{n} \hat{a}_{ij}(\theta)\hat{a}_{ik}(\theta)\widehat{\mathrm{cov}}\left\{I(\omega_j), I(\omega_k)\right\},$$

and by the same reasoning we can approximate $\hat{a}_{ij}(\theta)$ by $a_{ij}(\hat{\theta})$. Then to estimate the covariance of the periodogram we compute

$$\widehat{\mathrm{cov}}\left\{I(\omega_j), I(\omega_k)\right\} = \left|\frac{1}{2\pi n}\int_{-\pi}^{\pi} f\left(\omega';\hat{\theta}\right) D_n\left(\omega_j - \omega'\right) D_n^*\left(\omega_k - \omega'\right)\, d\omega'\right|^2,$$

where the asterisk denotes the complex conjugate and $D_n(\omega)$ is the (non-centred) Dirichlet kernel defined by

$$D_n(\omega) \equiv \frac{\sin(n\omega/2)}{\sin(\omega/2)}\exp(-\mathrm{i}\omega(n+1)/2),$$

such that we arrive at estimates of the diagonal elements of $\mathrm{var}\left\{\nabla \ell_D(\theta)\right\}$. Estimates of $\mathrm{cov}\left\{\partial/\partial\theta_i(\ell_D(\theta)), \partial/\partial\theta_j(\ell_D(\theta))\right\}$, which are the off-diagonal terms of $\mathrm{var}\left\{\nabla \ell_D(\theta)\right\}$, can be found in the same way. Then substituting all estimated entries of $\mathrm{var}\{\nabla \ell_D(\theta)\}$ into (15), along with the estimate of the Hessian, provides estimates of the variance of the estimators.

### 6·3. *Discussion*

Standard theory shows that standard Whittle estimates are consistent with optimal convergence rates if the spectrum (and its first and second partial derivatives in $\theta$) are continuous in $\omega$ and bounded from above and below (see Dzhaparidze & Yaglom, 1983), as well as being twice continuously differentiable in $\theta$. In contrast, we have not required that the spectrum nor its derivatives are continuous in $\omega$; such that Theorem 1 will hold for discontinuous spectra, as long as the other assumptions are satisfied such as Riemann integrability. As detailed in Lemma 9 of the supplementary material, this is possible because the expectation of the score is now zero after de-biasing (equation (19) in the supplementary material), which would not be the case for the standard Whittle likelihood, such that we only need to consider the variance of the score and Hessian. To control these variances we make repeated use of a bound on the variance of linear combinations of the periodogram (Lemma 8)—a result previously established in Theorem 3.1 of Giraitis & Koul (2013) under a different set of assumptions.

It can be easily shown that the assumptions in Theorem 1 are weaker than standard Whittle assumptions, despite requiring statements on the behaviour of the expected periodogram $\overline{f}_n(\omega;\theta)$ in Assumption 5. This is because if the spectral density $f(\omega;\theta)$ (and its first and second partial derivatives in $\theta$) are continuous in both $\omega$ and $\theta$, then it can be shown by applying the Leibniz' integration rule to the first and second derivatives of (9) in $\theta$, that $f(\omega;\theta)$ twice continuously differentiable in $\theta$ implies that $\overline{f}_n(\omega;\theta)$ is twice continuously differentiable in $\theta$. To show this we make use of Proposition 3.1 in Stein & Shakarchi (2003), which states that the convolution of two integrable and periodic functions is itself continuous. This result can also be used to show that $\overline{f}_n(\omega;\theta)$ is always continuous in $\omega$, even if $f(\omega;\theta)$ is not, as from (9) we see that $\overline{f}_n(\omega;\theta)$ is the convolution of $f(\omega;\theta)$ and the Fejér kernel—two functions which are integrable and $2\pi$-periodic in $\omega$. Therefore, not only does $\overline{f}_n(\omega;\theta)$ remove bias from blurring and aliasing, and is computationally efficient to evaluate, but it also has desirable theoretical properties leading to consistency and optimal convergence rates of de-biased Whittle estimates under weaker assumptions.

## Acknowledgments

We would like to thank the Associate Editor and anonymous referees for their extremely helpful suggestions in improving the paper. The work of A. M. Sykulski and S. C. Olhede was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/I005250/1 and Grant EP/L025744/1. A. M. Sykulski and S. C. Olhede also acknowledge support from the 7th European Community Framework Programme via a Marie Curie International Outgoing Fellowship (Sykulski), and via a European Research Council Fellowship via Grant CoG 2015-682172NETS (Olhede). The work of J. M. Lilly was supported by Award #1459347 from the Physical Oceanography program of the United States National Science Foundation.

## Supplementary Material

The supplementary material contains: addition simulation results (Section S1), technical proofs (Section S2), and details of a class of non-Gaussian processes satisfying the assumptions of our theory (Section S3). All reported simulation and application results in Section 5 can be exactly reproduced in MATLAB, and all data can be freely downloaded, using the software available at `https://github.com/AdamSykulski/SPG`. As part of the software we provide a simple package for estimating the parameters of any time series observation modelled as a second-order stationary stochastic process specified by its autocovariance.

## References

ANITESCU, M., CHEN, J. & STEIN, M. L. (2016). An inversion-free estimating equations approach for Gaussian process models. *J. Comput. Graph. Stat.* **26**, 98–107.

ANITESCU, M., CHEN, J. & WANG, L. (2012). A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. *SIAM J. Sci. Comput.* **34**, A240–A262.

BROCKWELL, P. J. & DAVIS, R. A. (1991). *Time series: Theory and methods*. Springer.

CASELLA, G. & BERGER, R. L. (2002). *Statistical inference*. Duxbury Pacific Grove, CA.

CHOUDHURI, N., GHOSAL, S. & ROY, A. (2004). Contiguity of the Whittle measure for a Gaussian time series. *Biometrika* **91**, 211–218.

CONTRERAS-CRISTAN, A., GUTIÉRREZ-PEÑA, E. & WALKER, S. G. (2006). A note on Whittle's likelihood. *Commun. Stat.-Simul. C.* **35**, 857–875.

DAHLHAUS, R. (1988). Small sample effects in time series analysis: A new asymptotic theory and a new estimate. *Ann. Stat.* **16**, 808–841.

DZHAPARIDZE, K. O. & YAGLOM, A. M. (1983). Spectrum parameter estimation in time series analysis. In *Developments in Statistics*, P. R. Krishnaiah, ed. Academic Press, Inc., pp. 1–96.

ECKLEY, I. A. & NASON, G. P. (2018). A test for the absence of aliasing or white noise in locally stationary wavelet time series. *Biometrika (in press)* .

ELIPOT, S., LUMPKIN, R., PEREZ, R. C., LILLY, J. M., EARLY, J. J. & SYKULSKI, A. M. (2016). A global surface drifter data set at hourly resolution. *J. Geophys. Res. Oceans* **121**, 2937–2966.

FAN, J., QI, L. & XIU, D. (2014). Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *J. Bus. Econ. Stat.* **32**, 178–191.

FUENTES, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *J. Am. Stat. Soc.* **102**, 321–331.

GIRAITIS, L. & KOUL, H. L. (2013). On asymptotic distributions of weighted sums of periodograms. *Bernoulli* **19**, 2389–2413.

GRIFFA, A., KIRWAN, A. D., MARIANO, A. J., ÖZGÖKMEN, T. & ROSSBY, T. (2007). *Lagrangian analysis and prediction of coastal and ocean dynamics*. Cambridge University Press.

GUINNESS, J. & FUENTES, M. (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *J. Comput. Graph. Stat.* **26**, 88–97.

JESUS, J. & CHANDLER, R. E. (2017). Inference with the Whittle likelihood: A tractable approach using estimating functions. *J. Time Ser. Anal.* **38**, 204–224.

KRAFTY, R. T. & COLLINGE, W. O. (2013). Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika* **100**, 447–458.

LILLY, J. M., SYKULSKI, A. M., EARLY, J. J. & OLHEDE, S. C. (2017). Fractional Brownian motion, the Matérn process, and stochastic modeling of turbulent dispersion. *Nonlinear Proc. Geoph.* **24**, 481–514.

MARPLE, S. L. (1999). Computing the discrete-time analytic signal via FFT. *IEEE T. Signal Proces.* **47**, 2600–2603.

MATÉRN, B. (1960). *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. Statens Skogsforskningsinstitut.

MATSUDA, Y. & YAJIMA, Y. (2009). Fourier analysis of irregularly spaced data on $R^d$. *J. R. Statist. Soc. B* **71**, 191–217.

PAWLOWICZ, R., BEARDSLEY, B. & LENTZ, S. (2002). Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE. *Computers & Geosciences* **28**, 929–937.

PERCIVAL, D. B. & WALDEN, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and conventional univariate techniques*. Cambridge University Press.

POLZIN, K. L. & LVOV, Y. V. (2011). Toward regional characterizations of the oceanic internal wavefield. *Reviews of Geophysics* **49**.

SCHREIER, P. J. & SCHARF, L. L. (2010). *Statistical signal processing of complex-valued data: The theory of improper and noncircular signals*. Cambridge University Press.

SLEPIAN, D. & POLLAK, H. O. (1961). Prolate spheroidal wave functions, Fourier analysis and uncertainty–I. *Bell Syst. Tech. J.* **40**, 43–63.

STEIN, E. M. & SHAKARCHI, R. (2003). *Fourier analysis: An introduction*. Princeton University Press.

SYKULSKI, A. M., OLHEDE, S. C., LILLY, J. M. & DANIOUX, E. (2016). Lagrangian time series models for ocean surface drifter trajectories. *J. R. Statist. Soc. C* **65**, 29–50.

TANIGUCHI, M. (1983). On the second order asymptotic efficiency of estimators of Gaussian ARMA processes. *Ann. Stat.* **11**, 157–169.

THOMSON, D. J. (1982). Spectrum estimation and harmonic analysis. *Proc. IEEE* **70**, 1055–1096.

TRENCH, W. F. (1964). An algorithm for the inversion of finite Toeplitz matrices. *Journal of the Society for Industrial and Applied Mathematics* **12**, 515–522.

VELASCO, C. & ROBINSON, P. M. (2000). Whittle pseudo-maximum likelihood estimation for nonstationary time series. *J. Am. Stat. Soc.* **95**, 1229–1243.

WHITTLE, P. (1953). Estimation and information in stationary time series. *Ark. Mat.* **2**, 423–434.