UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Contributions to Bayesian inference via spectral methods

by Thomas Goodwin

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Matias Quiroz, Prof. James Brown.

University of Technology Sydney
Faculty of Science

June 2024

Certificate of original authorship

I, Thomas Goodwin, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the Faculty of Science at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 03/06/2024

Acknowledgements

I would like to thank God for helping me complete this thesis.

Next, I would like to thank my supervisor, Dr. Matias Quiroz. Matias took me on as a main supervisor at the end of the first year of my PhD. I didn't know it at the time, but this was probably the most consequential and, in the end, the best outcome that could've happened.

I express my gratitude to my academic grandfather, Prof. Mattias Villani, at the University of Stockholm, Sweden. It was truly a pleasure working with you. I spent five weeks at the University of Stockholm, Sweden, on a research trip with Matias and Mattias. I loved my time in Sweden, and I hope to get back there someday and work with Mattias again.

I would also like to thank Scientia Professor Robert Kohn. Robert hired me as a postdoc in November of 2023, roughly six months before submitting this thesis. At the time he hired me, I was almost broke since my scholarship had run out, and I needed a position. Robert took a leap of faith and hired me, and for that, I thank him. I also have to thank Matias, as he made the introduction and gave me a reference, which, I presume, made the decision to hire me easier for Robert.

I thank Professor James Brown, distinguished Professor Matt Wand, and Associate Professor Stephen Woodcock, who were my co-supervisors at different times. Although not involved in the research, they often gave me great advice and made it a pleasure to work at UTS. Also, I'd like to thank Matt for providing me with additional funding through a scholarship via the ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS).

I thank my family, particularly my parents, who always supported and encouraged me.

Finally, I'd like to express my deepest gratitude to my partner, Mina. She has put up with me and supported me throughout this journey. This thesis is dedicated to her.

Abstract

This thesis investigates Bayesian inference methods for time series and spatial models in the frequency domain. One of the main drawbacks of Bayesian inference in this setting is the computational burden, especially for large data. Using ideas from Fourier analysis, the original signal (data) domain can be transformed into the frequency domain, which portrays how the signal is decomposed across different frequencies, which is known as the spectrum. A key property of the spectrum is the asymptotic independence of the spectrum ordinates, which can be used to form an approximate likelihood known as the Whittle likelihood, which is computationally faster than the corresponding time domain likelihood. We explore this computationally faster likelihood for three Bayesian models. First, we explore linear dynamic regression with semi-long memory disturbance processes. Second, spectral subsampling of continuous-time models for large data. Third, the estimation of stationary random fields for latticed spatial data.

List of papers

I: Goodwin, T., Quiroz, M. and Kohn, R. (2024), Improving forecasting in dynamic linear regression models by a semi-long memory model for the error process, *Manuscript*.

II: Goodwin, T., Quiroz, M. and Kohn, R. (2024), Bayesian inference via spectral subsampling MCMC for continuous-time ARMA processes, *Manuscript*.

III: Goodwin, T., Guillaumin, A., Villani, M., Quiroz, M. and Kohn, R. (2024), Bayesian inference for random fields on a lattice via the debiased spatial Whittle likelihood, *Manuscript*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims of the thesis

The overall aim of the thesis is to perform reliable Bayesian inference for time series and spatial models for large data sets in the frequency domain. The main motivation arises due to the heavy computational burden of conventional estimation techniques for temporal and spatial models with thousands, hundreds of thousands or sometimes millions of observations. To this end, we use tools from Fourier analysis, such as the Fourier transform, to investigate temporal and spatial data in the frequency domain.

The thesis provides three main contributions to the literature stated below.

1. Propose dynamic regression models with semi-long memory disturbance processes. We demonstrate that using models that capture semi-long memory provides a better fit to data that has long-memory characteristics. We show that traditional time-domain estimation methods are slow for large data and propose a new frequency domain approach to perform Bayesian inference. Also, the forecasting capabilities of this proposed model are on par, if not better than the existing models.

2. Spectral subsampling for continuous-time series auto-regressive moving average models for large data. By transforming the data into the frequency domain, we can construct a so-called spectral subsampling Markov chain Monte Carlo scheme, drastically reducing the computational time of performing Bayesian inference, while preserving much of the statical efficiency.

3. Frequency domain estimation of stationary random fields for spatial data on a lattice. Fitting spatial models to data is notorious burdensome due to the requirement of solving large systems of equations. Furthermore, frequency domain estimation for two or more dimensions can lead to substantial bias. We developed a frequency domain methodology

for Bayesian inference which significantly reduces the bias and does so in a computationally efficient manner.

1.2 Preliminaries and notation

Let X be a real-valued continuous random variable with *cumulative distribution function* (cdf),

$$F_X(x) = P(X \leq x),$$

which gives the probability that X takes a value less than or equal to some outcome x. The cdf can be defined by the integral,

$$F_X(x) = \int_{-\infty}^{x} f(t)dt,$$

where $f(x)$ is the *probability density function* (pdf) of the random variable X. Furthermore, let Y be a real-valued random variable defined over the real line $\mathbb{R}$. The first moment or *expected value* of Y is given by

$$\mu_Y = \mathrm{E}[Y] = \int_{\infty}^{\infty} yf(y)dy,$$

where $f(y)$ is a probability density function. The second central moment, or variance of Y is given by

$$\mathrm{Var}[Y] = \mathrm{E}[(Y - \mu)^2]. \tag{1.1}$$

The variance measures how large, on average, the squared difference between the values of Y and its average. The covariance is a measure of the degree of the linear relationship between X and Y, defined by

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}[XY] - \mu_X\mu_Y. \tag{1.2}$$

1.3 Stochastic processes

Denote a time series, or *stochastic process* as a sequence of random variables $\{X_t, t \in T\}$ where t is an index of time in a discrete or continuous interval T. The observations of a time series are taken at discrete time points, which, for this thesis, we assume are equally spaced. Furthermore, time series data can be modelled in continuous time, $T \in \mathbb{R}$, or discrete time, when T is a discrete set of points.

1.3.1 Stationary time series

We introduce the notion of *stationarity* for statistical analysis of stochastic processes. A *strictly* stationary process is a stochastic process whose statistical distribution does not change when

shifted in time. Mathematically, let $F(x_{t_1+\tau}, \ldots, x_{t_n+\tau})$ denote the unconditional cumulative distribution function of X_t with any lag such that $t_i + \tau \in T$. Then X_t is strictly stationary if

$$F(x_{t_1}, \ldots, x_{t_n}) = F(x_{t_1+\tau}, \ldots, x_{t_n+\tau}).$$

It is important to note for any τ, $F_X(\cdot)$ remains unchanged; hence F_X is not a function of time. In practical applications, strict stationarity is too restrictive and may not be appropriate, and therefore, our analysis in this thesis is primarily concerned with the first two moments of time series. Therefore, we use a weaker form of stationarity, referred to as weak stationarity.

A process is *weakly stationary* if the mean and covariance do not vary with time, and the second moment is finite for all times. Define the mean function as $m(t) = \mathrm{E}[X_t]$ and the autocovariance function for any two time points s, t as $\gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathrm{E}[X_s X_t] - m(t)^2$. Formally, a stochastic process X_t is weakly stationary if

$$\begin{aligned}
&1.\ m(t+\tau) = m(t), && \forall \tau, t \in T,\\
&2.\ \gamma(s,t) = \gamma(\tau, 0) = \gamma(s-t, 0), \quad \text{where } \tau = |s-t|, && \forall s, t \in T,\\
&3.\ \mathrm{E}[|X_t|^2] < \infty && \forall t \in T.
\end{aligned}$$

The first point describes how the mean function $m(t)$ is constant for all time. The second point states the covariance function depends only on the time difference $\tau = |t-s|$, i.e., $r(\tau) = r(s,t)$, not on the specific choice of s, t themselves. Finally, the third point ensures the second moment of the process is finite.

Strict stationarity encompasses the full cumulative distribution function $F_X(\cdot)$, whereas weak stationarity only describes the first and second moments being time-invariant. This is clearly a less restrictive version than strict stationarity; however, stationary Gaussian processes that are weakly stationary are also strictly stationary since their distribution is completely defined by the first two moments. Through the rest of this thesis, stationarity will be assumed to be weakly stationary unless stated otherwise.

Autoregressive integrated moving average models

A common time series model used in economics and climatology is the autoregressive moving average (ARMA) model first proposed in Whittle (1951). The ARMA model comprises two simpler models: autoregressive and moving average models. An autoregressive process of order p, denoted as AR(p), is a stochastic linear difference equation,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2),$$

where $\phi_1 \ldots, \phi_p$ are the parameters multiplied up to the pth 'lags' of X_t, and ε_t is white noise with variance σ^2. This definition for X_t is a zero-mean process (Shumway et al., 2000). To write compactly, the backshift operator L is defined as $L^k X_t = X_{t-k}$ for $t > k$. The backshift operator is convenient as an AR(p) process can be written as

$$\boldsymbol{\phi}(L)X_t = \varepsilon_t,$$

where $\boldsymbol{\phi}(z) = 1 - \sum_{i=1}^{p} \phi_i z^i$ is the autoregressive lag polynomial. To ensure the stationarity for a general AR(p) process, the roots of $\boldsymbol{\phi}(z)$ must lie outside the unit circle, i.e., $|z_i| > 1$, where $z_1, \ldots, z_p$ are the roots of $\boldsymbol{\phi}(z)$ (Cryer and Chan, 2008).

The second component of an ARMA model is the moving average model. The purpose of the moving-average model is to model the random shocks. The random component of the process, known as the *innovation error* or error term, can themselves have an auto-regressive component. Formally, an MA(q) process is written as

$$X_t = \boldsymbol{\psi}(L)\varepsilon_t = \sum_{i=1}^{q} \psi_i \varepsilon_{t-i} + \varepsilon_t,$$

where $\boldsymbol{\psi}(z) = 1 + \psi_1 z, \ldots, \psi_q z^q$ is the moving-average lag polynomial. One can show that an AR(p) model can be formulated as an MA(∞) model. If the reverse is true, that an MA(q) model can be written as an AR(∞) model, i.e.

$$\varepsilon_t = \boldsymbol{\psi}(L)^{-1} X_t = X_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots,$$

then the MA process is *invertible*. Representing an AR process as an infinite MA process and visa versa is convenient as it allows one to use useful statistical properties of both representations. Note that any MA process with Gaussian error terms is stationary; however, invertible MA processes are desirable as they further ensure the model is *identifiable*. A statistical model is identifiable if there is a one-to-one mapping between the parameters of the statistical model and the probability distribution generated by the data (Lehmann and Casella, 2006). In the context of MA models, if the model is invertible, then there is only one set of parameter values $(\psi_1, \ldots, \psi_q, \sigma)$ that can generate a given autocorrelation function (Cryer and Chan, 2008). To summarize, an AR(p) process is always invertible but not necessarily stationary, while a MA(q) process is always stationary but not necessarily invertible. Similar to AR models, if the roots of the $\boldsymbol{\psi}(z)$ polynomial are outside the unit circle, then the MA model is invertible (Lindgren et al., 2013).

An ARMA(p, q) model combines the two aforementioned models into a single compact form,

$$\boldsymbol{\phi}(L)X_t = \boldsymbol{\psi}(L)\varepsilon_t, \tag{1.3}$$

to directly describe the X_t regressed against lagged values of itself and the contemporaneous error term ε_t as a linear combination of its past values. Let us consider a simple ARMA$(1,1)$ model

$$X_t - \phi_1 X_{t-1} = \varepsilon_t + \psi_1 \varepsilon_{t-1}, \tag{1.4}$$

where the stationarity condition is $|\phi_1| < 1$ and invertibility condition is $|\psi_1| < 1$. The autocovariance function is given by

$$\begin{aligned}
\gamma(0) &= \sigma^2 \left(1 + \frac{(\phi_1 + \psi_1)^2}{1 - \phi_1^2}\right), \\
\gamma(1) &= \sigma^2 \left((\phi_1 + \psi_1) + \frac{(\phi_1 + \psi_1)^2 \phi}{1 - \phi_1^2}\right), \\
\gamma(\tau) &= \phi_1^{\tau - 1} \gamma(1), \quad \text{for } \tau \geq 2,
\end{aligned}$$

which does not depend on time t. Much more can be said about ARMA models. For a comprehensive treatment of ARMA models, refer to Box et al. (2015). Chapter 2 will discuss extensions of the ARMA model, mainly to incorporate exogenous covariates and so-called 'semi-long memory' where the autocorrelation function decays exponentially to zero only when the distance between observations is large.

1.3.2 Spatial data

Chapter 4 is concerned with spatial data, i.e. observations from an underlying stochastic process usually indexed by a geographic location/coordinate. The underlying process is known as a *random field*, a type of random function, i.e. stochastic process, with a multidimensional domain. This can be seen as a multivariate generalization of a time series where the x-axis of time is extended to a d-dimensional space. Spatial data are common in natural science areas such as earth sciences (Christakos, 2012), geological sciences (Simons and Olhede, 2013), and forestry Matérn (2013), among others. For example, one might be interested in modelling the spatial dependence of the surface temperature of the sea over some region of the ocean (Gelfand et al., 2010). In neuroscience, random fields have been used to model areas of brain activation measured via fMRI (Worsley et al., 1992). Certain simplified random field models are also popular in areas of machine learning, including computer vision (He et al., 2004) and classification (Cohen et al., 1991), and natural language processing (Lafferty et al., 2001). Spatial data on a grid are usually captured in the form of an image, where the values of the image (greyscale or RGB) represent the phenomena being observed corresponding to a location intrinsic to where the pixels of the image lay.

Let $\{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{D} \subseteq \mathbb{R}^d\}$ be a zero-mean random field, where $\mathcal{D} \subseteq \mathbb{R}$ is some continuous domain of interest. As mentioned above, the random field $Y(\boldsymbol{s})$ is indexed by the set $\boldsymbol{s} \in \mathbb{R}^d$, which

usually represents spatial locations in $d = 2$ or $d = 3$. This is not to be confused with multivariate time series, where the data are a k-dimensional vector, $\boldsymbol{y}_t \in \mathbb{R}^k$ for all (one-dimensional) t. The spatial version of this case is called multivariate spatial data, where the output for a given location is a vector $\boldsymbol{Y}(\boldsymbol{s}) \in \mathbb{R}^k$.

The process $Y(\boldsymbol{s})$ can be thought of as a collection of random variables at all locations in $\mathcal{D}$ that has a well-defined joint distribution (Gelfand et al., 2010). The covariance function describes the spatial dependence between points in the domain, defined as

$$c_{\boldsymbol{\theta}}(\boldsymbol{u}) = \mathrm{E}\left[Y(\boldsymbol{s})Y(\boldsymbol{s}+\boldsymbol{u})\right], \tag{1.5}$$

where $\boldsymbol{u} \in \mathbb{R}^d$ are known as the spatial lags. The aforementioned equation depends on a vector of unknown parameters $\boldsymbol{\theta}$. A common assumption is to assume the covariance function is *isotropic*, i.e. given two arbitrary points $\boldsymbol{v}$ and $\boldsymbol{v}'$, the covariance function is only a function of $|\boldsymbol{v} - \boldsymbol{v}'|$. Isotropic covariance functions imply stationarity since the covariance function does not depend on the specific point in space, only on the 'distance' between two points (Rasmussen and Williams, 2006). Chapter 4 is concerned with estimating parameters $\boldsymbol{\theta}$ in the covariance kernel given a realization of $Y(\boldsymbol{s})$. A widely known covariance kernel is the Mátern kernel (Matérn, 2013), given by,

$$c_{\boldsymbol{\theta}}(\boldsymbol{u}|\rho,\sigma,\nu) = \sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{||\boldsymbol{u}||}{\rho}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{||\boldsymbol{u}||}{\rho}\right), \tag{1.6}$$

where ρ, σ, and ν are positive parameters and K_ν is a modified Bessel function (Abramowitz and Stegun, 1968). The advantage of the aforementioned covariance kernel is its flexibility, i.e. for different values of ν, it exhibits drastically different behaviour in the realizations of $Y(\boldsymbol{s})$. For specific values of ν, special cases become apparent; for example, two common covariance kernels are

$$c(\boldsymbol{u}|\rho,\sigma,\nu = 1/2) = \sigma^2\exp\left(-\frac{||\boldsymbol{u}||}{\rho}\right),$$

referred to as the exponential kernel and

$$c(\boldsymbol{u}|\rho,\sigma,\nu \to \infty) = \sigma^2\exp\left(-\frac{||\boldsymbol{u}||^2}{2\rho^2}\right),$$

known as the squared-exponential kernel (Rasmussen and Williams, 2006). For estimation, the covariance must be computed between all distinct points in the observed domain, which becomes computationally expensive for large spatial data sets. The main contribution of this thesis is to address this computational burden by analysing the data in the frequency domain, described in the section below.

1.4 Fourier analysis

In the early 19th century, while studying heat flow, the French mathematician Joseph Fourier discovered that any function, possibly discontinuous, can be expressed in terms of an infinite series of sine functions (Joseph and Freeman, 2003). Once refined and expanded, this idea became the base of Fourier analysis, one of the most important areas of science and mathematics. At the heart of Fourier analysis lies the *Fourier transform* (FT). This can be seen as decomposing a function or signal in time into its frequencies, i.e. spectrum, which make up that function or signal. The Fourier transform relies on basis functions in the form of complex exponentials,

$$e^{\mathrm{i}x} = \cos(\mathrm{x}) + \mathrm{i}\sin(\mathrm{x}), \tag{1.7}$$

which is simply a unit circle in the complex plane, from Euler's formula.

Given a continuous function $g(t)$, the Fourier transform is defined as,

$$G(\omega) = \int_{\mathbb{R}} g(t)\exp(-\mathrm{i}\omega t)dt, \tag{1.8}$$

where the *angular frequency* is $\omega = 2\pi\varsigma$, with the frequency ς in Hertz. Intuitively, the signal $g(t)$ in (1.8) is being 'wrapped' around the unit circle at different frequencies in the complex plane. The crucial insight in the transformation is the integration with respect to (wrt) time t, such that the resulting function $G(\omega)$ is now only a function of ω. For the rest of this thesis, we will refer to angular frequencies as frequencies for simplicity. Its corresponding *inverse* Fourier transform is given by,

$$g(t) = \frac{1}{2\pi}\int_{\mathbb{R}} G(\omega)\exp(-\mathrm{i}\omega t)d\omega, \tag{1.9}$$

where integration wrt the frequencies recovers the original $g(t)$.

To better connect the content within this thesis, we introduce the Fourier transform through the lens of statistical signal processing. Suppose a stationary process $\{X_t, t \in \mathbb{R}\}$ has a continuous covariance function $\gamma(\tau)$, which is absolutely integrable, i.e. $\int_{-\infty}^{\infty}|\gamma(\tau)|d\tau < \infty$. Bochner's theorem provides conditions under which the covariance function of the process corresponds to the Fourier transform of a spectral measure on the real line (Brockwell and Davis, 2009). Formally,

$$\gamma(\tau) = \mathrm{E}[X_{t+\tau}X_t] = \int_{-\pi}^{\pi} e^{\mathrm{i}\omega\tau}dF(\omega), \tag{1.10}$$

where $F(\omega)$ is a non-negative, non-decreasing bounded function known as the spectral measure or *spectral distribution function*, and this representation of the covariance function is unique. If

the spectral distribution function is absolutely continuous, it can be written as

$$F(\omega) = \int_{-\pi}^{\omega} f(\lambda) d\lambda,$$

where $f(\cdot)$ is the *spectral density* and (1.10) becomes,

$$\gamma(\tau) = \int_{-\infty}^{\infty} f(\omega) e^{\mathrm{i}\omega\tau} d\omega. \tag{1.11}$$

Hence, the Fourier transform of $\gamma(\tau)$ exists and is a *Fourier pair* or *dual* with the spectral density,

$$f(\omega) = \int_{-\infty}^{\infty} \gamma(\tau) e^{-\mathrm{i}\omega\tau} d\tau. \tag{1.12}$$

A key insight to the interpretation of the spectral density is that the variance of the process is the integral over the whole spectral density,

$$\mathrm{Var}[X_t] = \gamma(0) = \int_{-\infty}^{\infty} f(\omega) d\omega. \tag{1.13}$$

Hence, the power spectrum can be considered as the distribution of total variance, where the variance (power) is decomposed over different frequencies. This is critical as the spectral density gives us insight into which frequencies are the main contributors to the variance of the process of interest. Significant spectral 'mass' at the lower frequencies corresponds to persistent signals, slowly changing over time. In contrast, higher frequency dominant signals are fast changing, usually with lower or negative correlations between neighbouring time points. Additionally, constant spectra across the frequency range correspond to white noise processes.

For discrete models such as ARMA models described in the previous section, the process is discrete in time for $t = 0, 1, 2, \ldots$, and hence the covariance function is only defined for $\tau \in \mathbb{N}$. Satisfying the condition $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| < \infty$, the spectral density is positive, continuous and integrable. The discrete Fourier transform (DFT) is defined as

$$f(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-\mathrm{i}\omega\tau}, \tag{1.14}$$

for frequencies in the interval $\omega \in (-\pi, \pi]$. It can be shown that the spectral density of an ARMA(p, q) model is a rational function of the form

$$f(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{\boldsymbol{\psi}(e^{-\mathrm{i}\omega})}{\boldsymbol{\phi}(e^{-\mathrm{i}\omega})} \right|^2.$$

The Fourier transforms we have discussed have been in the one-dimensional time series setting.

For $d > 1$, the Fourier transform can be extended to a multi-dimensional setting, where the spectrum is defined on a multivariate frequency domain, $f(\boldsymbol{\omega})$, $\boldsymbol{\omega} \in \mathbb{R}^d$. The multivariate Fourier transform is

$$\gamma(\boldsymbol{\tau}) = \mathrm{E}[X(\boldsymbol{s})X(\boldsymbol{s}+\boldsymbol{u})] = \int_{\mathbb{R}^d} f(\boldsymbol{\omega})\exp(\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{u})d\boldsymbol{\omega} \tag{1.15}$$

where $\cdot$ is the dot product. The covariance function in (1.15) forms a Fourier dual with $f(\boldsymbol{\omega}) = \int_{-\infty}^{\infty}\gamma(\boldsymbol{\tau})\exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{u})d\boldsymbol{\tau}$. The spectral decomposition of the autocovariance of a stationary process into its power spectral density is known as *Wiener-Khinchin* theorem (Chatfield and Xing, 2019). By applying the Fourier transform to the Mátern kernel in (1.6), its spectral density is given by

$$f(\boldsymbol{\omega}) = \sigma^2\frac{2^d\pi^{d/2}\Gamma(\nu+\frac{d}{2})(2\nu)^\nu}{\Gamma(\nu)\rho^{2\nu}}\left(\frac{2\nu}{\rho^2}+4\pi^2\boldsymbol{\omega}^2\right)^{-\left(\nu+\frac{d}{2}\right)}, \tag{1.16}$$

where d is the number of dimensions (Rasmussen and Williams, 2006).

1.4.1 Fourier transform of data

So far, we have defined theoretical quantities such as the covariance function and spectral density, which dictate the second-order statistical properties of a stochastic process that generates data. With some foundational underpinnings, we now turn to the main focus of statistical signal processing, which is the analysis of discretely sampled signals/data to estimate their spectra, covariances/correlations and underlying parameters in the stochastic model. Assume X_t is a zero mean, stationary time series for $t = 1, \ldots, T$, the *discrete Fourier transform*, DFT, is defined as

$$J(\omega_k) = \sum_{t=0}^{T-1} X_t\ \exp(-\mathrm{i}\omega_k t), \tag{1.17}$$

with ω_k is in the set of natural Fourier frequencies

$$\Omega \equiv \{2\pi k/T,\ \text{for k} = -\lceil T/2\rceil + 1, \ldots, \lfloor T/2\rfloor\}.$$

For a given $\omega_k \in \Omega$, the DFT is a sum of the data X_t multiplied by a deterministic constant (the complex exponential). This is analogous to the central limit theorem, which loosely states that the normalized sum of independent random variables is normally distributed, even when the distribution of the random variables is not normal. Of course, the DFT in (1.17) is a complex sum of a dependent sequence of random variables, but a similar central limit theorem exists for the DFT. This central limit theorem for the DFT for stationary processes is foundational in the statistical analysis of signals, described below.

If X_t is normal, its DFT in (1.17) is complex Gaussian. The real and imaginary parts of the DFT are asymptotically independently identically distributed normal distributions. Formally,

Let $CN(0,\sigma^2)$ denote a complex Gaussian, if $Z \sim CN(0,\sigma^2)$, then,

$$\begin{pmatrix} \mathrm{Re}\{Z\} \\ \mathrm{Im}\{Z\} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right],$$

hence the real and imaginary parts are independent, and Z has density

$$p_Z(z) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|z|^2}{\sigma^2} \right).$$

(Goodman, 1963). Hence the distribution of the DFT of X_t is

$$\frac{1}{\sqrt{T}} J(\omega_k) \sim CN(0, 2\pi f(\omega_k)), \tag{1.18}$$

as $T \to \infty$, except for frequencies $\omega_k = 0$ and $\omega_k = \pi$, then $(1/\sqrt{T})J(\omega_k) \sim N(0, 2\pi f(\omega_k))$ (Brillinger, 2001). If X_t is not Gaussian, under certain mild conditions, the distribution of the normalized DFT follows (1.18) asymptotically, as $T \to \infty$ (Peligrad and Wu, 2010). Furthermore, $J(\omega_k)$ are asymptotically independent for each frequency ω_k (Shao and Wei, 2007). A similar central limit theorem for the multidimensional DFT can be extended for random fields (Peligrad and Zhang, 2019). In practice, the DFT is computed efficiently via the *Fast Fourier Transform*, FFT, (Cooley and Tukey, 1965). The FFT exploits the structure of the DFT and reduces the number of floating point operations from $\mathcal{O}(T^2)$ to $\mathcal{O}(T\log T)$.

The *periodogram*, or the observed power spectrum of X_t, is given by,

$$\mathcal{I}(\omega_k) = T^{-1} \Big| J(\omega_k) \Big|^2, \tag{1.19}$$

and is an estimate of its spectral density. The scaled periodogram is $\mathcal{I}(\omega_k)/f(\omega_k) \sim \chi^2_r$ asymptotically, where χ^2_r is a chi-squared random variable with r degrees of freedom where $r = 2$ (i.e. standard exponential) for all $k \neq 0, T$; and $r = 1$ for $k = 0, T$. The periodogram in (1.19) is an asymptotically unbiased estimate of $f(\omega_k)$; however, it is not a consistent estimator of the spectral density. More specifically,

$$\mathcal{I}(\omega_k) \sim \mathrm{Exp}(f(\omega_k)), \qquad \omega_k \in \Omega, \tag{1.20}$$

as $T \to \infty$ with the exponential distribution parameterized by its mean (Shao and Wei, 2007). The periodogram ordinates in (1.20) are asymptotically independent for each frequency ω_k. The spectral density and periodogram are symmetric about the origin and are repeating for frequencies outside the interval $\omega_k \in (-\pi/2, \pi/2)$, hence we only need to consider the interval $\omega_k \in (0, \pi/2)$. In this thesis, we assume that the time series process has a data-generating process with a well-

defined spectral density function. However, the spectral density of a process can be modelled from its periodogram ordinates without explicitly specifying its data-generating process in the time domain; see Mallick et al. (2002); Gangopadhyay et al. (1999). For more details about the multivariate DFT and its asymptotic properties, see Chapter 4.

1.4.2 Sampling and aliasing

Time series data can be sampled at different rates depending on the application. For example, financial time series data can be sampled at extremely high-frequency rates (milliseconds), and temperature data in climatology can be sampled every hour, day, or year. As stated, we assume the time series is observed at equidistant time points. Likewise, for two-dimensional spatial data in this thesis, we assume site locations are equidistant apart or that the process can be modelled on a lattice or grid structure. Statistical inference is possible for irregularly spaced data, see Benedetto (1992); however, that is the topic for future work and beyond the scope of this thesis.

Consider a continuous stationary process $\{Y_t,\ t \in \mathbb{R}\}$ with covariance $\gamma(\tau)$ and spectral density $f(\omega)$. Define the sampling interval δ such that the observed process is $\{Y_t,\ t = 0, \delta, 2\delta, \ldots, T\delta\}$. The sampling rate is $1/\delta$, and the *Nyquist* frequency is π/δ. When a continuous process is discretely sampled at regular intervals, this introduces the *aliasing* effect, which describes how the spectral density becomes an infinite sequence. The aliased spectral density is

$$f_\delta(\omega) = \sum_{k=-\infty}^{\infty} f\big(\omega + \frac{2\pi k}{\delta}\big), \qquad \omega \in [-\pi/\delta, \pi/\delta]. \tag{1.21}$$

The proof of this can be found in Shannon (1949). Here, contributions from the infinite sum is the spectrum being 'folded' or 'wrapped' for frequencies outside the interval $[-\pi/\delta, \pi/\delta]$. Intuitively, suppose one obtains samples of a sine wave at regular intervals. In that case, reconstruction of the original signal can be at the original frequency or all integer multiples of the original frequency. Hence, there are infinite ways to reconstruct the signal from the observed data.

The periodogram of a regularly sampled continuous-time process is only defined up to the Nyquist frequency π/δ. If the sampling interval is too large (the sampling rate is too low), the periodogram will not contain reliable information about significant portions of its spectrum. Ideally, we wish to sample the process finely enough to capture all non-negligible mass in the spectrum. Furthermore, if aliasing is not appropriately addressed, this can result in bias of the parameter estimates (Sykulski et al., 2019; Guillaumin et al., 2022). For a more detailed discussion, see Chapter 3, and for the extension to multidimensional data, see Chapter 4.

1.4.3 The Whittle likelihood

The main parameter estimation method this thesis is concerned with is the Whittle likelihood (Whittle, 1951). Throughout this thesis, we refer to the Whittle likelihood as the log-likelihood version unless stated otherwise. The Whittle likelihood is a frequency domain estimation method for stationary processes that is valid for large samples. The asymptotic independence of the DFT at each frequency makes it possible to derive the iid density for the periodogram using its distribution in (1.20). For discrete-time models, the Whittle likelihood is given as

$$\ell(\boldsymbol{\theta}) = -\sum_{k=1}^{\lfloor (T-1)/2 \rfloor} \left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} \right), \tag{1.22}$$

where the dependence on the unknown parameters $\boldsymbol{\theta}$ is explicit. For continuous models, the spectral density in (1.22) is replaced by the aliased spectral density $f_{\boldsymbol{\delta}}(\omega)$. For stationary Gaussian data, the Whittle likelihood is asymptotically equivalent to the exact Gaussian likelihood (Guyon, 1982). Moreover, the Whittle likelihood is robust as it can handle stationary non-Gaussian data (under mild conditions), leveraging the previously mentioned central limit theorem for the DFT.

To compute the Whittle likelihood in (1.22), one must first compute the periodogram, which has a one-time cost of $\mathcal{O}(T \log T)$. Then, after storing the result, the evaluation of (1.22) has a subsequent cost of $\mathcal{O}(T)$, which involves computing the spectral density and the summation over the frequencies. This is a huge advantage of the Whittle likelihood compared to the exact Gaussian likelihood, which has, at best, a cost of $\mathcal{O}(T \log^2 T)$, due to expensive matrix inversion (Ammar and Gragg, 1988). The Whittle likelihood will be discussed in depth throughout the remainder of this thesis; with the presence of long memory and exogenous inputs in Chapter 2, for continuous-time time series in Chapter 3, and for spatial models in Chapter 4.

1.5 The Bayesian paradigm

The Bayesian approach to statistical inference is to treat an unknown quantity $\boldsymbol{\theta}$ as a random variable. For example, $\boldsymbol{\theta}$ could be a hypothesis, a statistical model, a parameter in a statistical model or a future prediction for a given statistical model. In this thesis, we consider the case when $\boldsymbol{\theta}$ is a vector of unknown parameters that govern a statistical model. Since $\boldsymbol{\theta}$ is unknown, it is treated as random but typically incorporates prior knowledge of $\boldsymbol{\theta}$ before conducting an experiment, known as the *prior* distribution $p(\boldsymbol{\theta})$. The prior distribution is a *subjective* belief, which can reflect previous experiments or experiences based on the problem at hand. The scientist conducts the experiment and collects/observes the data $\boldsymbol{y}$, which is then used to obtain the *likelihood* function $p(y|\boldsymbol{\theta})$ which contains information of $\boldsymbol{\theta}$ given the observed day. The subjective belief of $\boldsymbol{\theta}$ is *objectively* updated by combining likelihood and prior via *Bayes' theorem* to obtain

the main object of inference, the *posterior* distribution $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y})$,

$$\pi(\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \qquad \text{where } p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.23}$$

This fundamentally differs from frequentist inference in two ways: 1) the posterior distribution is a valid probability distribution of the unknown quantity $\boldsymbol{\theta}$, and 2) the posterior is conditional on the actual observed data. The latter is a primary difference compared to frequentist analysis, as any uncertainty of point estimates is over all possible data, observed or not. Hence, one of the main advantages of the Bayesian paradigm is the intrinsic uncertainty about the quantity $\boldsymbol{\theta}$, ingrained in the notion of probability, conditioned on the data $\boldsymbol{y}$ we observed. This aforementioned uncertainty makes Bayesian inference attractive to real-world problems.

One of the main criticisms of Bayesian analysis is the subjectivity inherited with the prior distribution $p(\boldsymbol{\theta})$. Ideally, the prior distribution is constructed from past information such as previous experiments or expert knowledge; when no relevant information is present, uninformative priors may be employed to portray our assumption of lack of information (Chaloner, 1996). Furthermore, when more information (data) becomes available, the influence of the likelihood dominates that of the prior distribution. In reality, the choice of the statistical model is a subjective decision. Hence, one must ensure the model is consistent with the data to carry out appropriate statistical analysis. The same logic should be applied to choosing the prior distribution. For more criticism and discussion thereof, and the inherent link between Bayesian inference and decision theory, see Berger (2013).

Computation

An important problem in Bayesian inference is the computation of expectations with respect to the posterior

$$\mathrm{E}_{\pi}[\psi(\boldsymbol{\theta})] = \int_{\Theta} \psi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{1.24}$$

over the parameter space $\boldsymbol{\theta} \in \Theta$. In practice, closed-form solutions to the above equation seldom exist. Only in certain cases does a closed-form posterior distribution exist—for a few choices of the likelihood $p(y|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$, known as *conjugate priors*. Here, the posterior takes the same closed-form distribution as the prior, giving a particular form for the likelihood. Simulation techniques such as Monte Carlo integration can be used to estimate (1.24). Here, samples from the posterior $\pi(\boldsymbol{\theta})$ must first be obtained for Monte Carlo integration. However, sampling the posterior in (1.23) is generally a complex problem and discussed further in Section 1.5.3.

1.5.1 Prediction

A quantity of interest in Bayesian inference, particularly for time series analysis, is the *posterior predictive* distribution, defined as

$$p(\widetilde{\boldsymbol{y}}|\boldsymbol{y}) = \int_{\Theta} p(\widetilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}. \tag{1.25}$$

The main goal of (1.25) is to make inferences for the future by simulating new data while taking into account the uncertainty around the parameters $\boldsymbol{\theta}$. The integrand of the above equation is the joint distribution of new data $\widetilde{\boldsymbol{y}}$ and the parameters $\boldsymbol{\theta}$ conditional on the observed $\boldsymbol{y}$. The analytic form of the posterior predictive distribution is seldom known in practice but can be sampled via simulation. At first glance, the integrand may look unwieldy; however, the joint density can be decomposed as

$$p(\widetilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y}) = p(\widetilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}).$$

This decomposition makes it clear that one can sample from $p(\widetilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y})$ by generating new data from the model, conditional on samples from the posterior and the observed data. Once again, samples from the posterior $\pi(\boldsymbol{\theta})$ must be obtained to sample from $p(\widetilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y})$. Thus, the posterior predictive can be rewritten as

$$p(\widetilde{\boldsymbol{y}}|\boldsymbol{y}) = \int_{\Theta} p(\widetilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.26}$$

The posterior predictive is not to be confused with a posterior expectation in (1.24), which usually estimates an intractable integral. Instead, we want to obtain samples from the distribution $p(\widetilde{\boldsymbol{y}}|\boldsymbol{y})$, which is all possible future values (conditioned on the observed data). This can be achieved by sampling from the $p(\widetilde{\boldsymbol{y}}, \boldsymbol{\theta}|\boldsymbol{y})$ as mentioned above and simply discarding samples from $\pi(\boldsymbol{\theta})$.

An important objective of time series analysis is the forecasting ability of the model in question. This gives us an understanding of how well the model fits the data. A common strategy is to divide the observed data into a training set $y_{1:T} = (y_1, \ldots, y_T)$ and testing set $y_{T+1:T+k} = (y_{T+1}, \ldots, y_{T+k})$. We fit the model and compute the Bayesian predictive posterior for each model we've fit. Then, we can compare the forecasted values to the testing set using a forecasting metric.

Suppose we want the h set-ahead forecast, $\widehat{y}_{T+h}$, conditional on all previous values up to time T. Then, to obtain point forecasts, define the conditional expectation as

$$\widehat{y}_{T+h} = \mathrm{E}[\widetilde{y}_{T+h}|y_{1:T}] = \int \widetilde{y}_{T+h} p(\widetilde{y}_{T+h}|y_{1:T}) d\widetilde{y}_{T+h},$$

where $p(\widetilde{y}_{T+h}|y_{1:T})$ is the posterior predictive distribution for time series data. Here, the posterior

predictive is computed over the length of the testing set $\widetilde{y}_{T+i}$, $i = 1, \ldots, k$. To evaluate the performance of the point forecasts, the root mean square error (RMSE) is computed as

$$\text{RMSE}_h = \sqrt{\frac{1}{k-h+1}\sum_{i=0}^{k-h}\left(y_{T+h+i} - \widehat{y}_{T+h+i}\right)^2}.$$

for each forecast horizon h. The RMSE is the root of the average squared distance between the testing set and forecasted values, and hence, the lower the value, the better. Distributional forecasts, i.e. the full posterior predictive distribution, can also be assessed, which is discussed in more detail in Chapter 2. For a wide selection of forecasting metrics to assess forecasting performance, see Hyndman and Athanasopoulos (2018).

Another common measure of forecasting performance is the h-step-ahead log posterior predictive distribution given as

$$\begin{aligned}\log p(y_{T+1:T+h}|y_{1:T}) &= \log \int_\Theta p(y_{T+1:T+h}|\boldsymbol{\theta}, y_{1:T})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \\ &\approx \log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_{T+1:T+h}|\boldsymbol{\theta}^{(m)}, y_{1:T})\right),\end{aligned}$$

where the last line is a Monte Carlo approximation thereof, with $\boldsymbol{\theta}^{(m)} \sim \pi(\boldsymbol{\theta})$. Once the LPDS has been computed, the training and testing sets are updated (expanded) by $t+1$. The model is re-estimated from the new training set, and the LPDS is computed for the updated testing set. This process is repeated until observations of the testing set $y_{T+1:T+k}$ have been used in the computation of the LPDS. This method takes the time series structure into account and is known as time series cross-validation, described in detail in Bürkner et al. (2020). Note, that this technique is also used to compute the RMSE_h for the whole testing set as described above.

1.5.2 Model selection

An important consideration for any practitioner is selecting an appropriate model for the observed data. Consider observations from an ARMA(p, q) model. Choosing the number of AR terms p and moving-average terms q is not obvious. There are heuristics to determine the p, q from looking at the empirical autocorrelations and partial autocorrelations of the data. Still, in practice, determining the appropriate model is not always straightforward due to complicated observed autocorrelation structures. A well-known phenomenon is that the maximum likelihood value increases as more parameters are added to the model, which can result in over-fitting. Hence, we seek to find a model that fits the data well but penalises the model with a large number of parameters, e.g. for large p or q.

For a given statistic model $\mathcal{M}$, the *marginal likelihood* or *model evidence* can be written

explicitly as

$$p(\boldsymbol{y}|\mathcal{M}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \tag{1.27}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})$ is the likelihood and $p(\boldsymbol{\theta}|\mathcal{M})$ is the prior for $\boldsymbol{\theta}$ under model $\mathcal{M}$. In practice, this integral is intractable, and one must resort to approximations of the above integral. There are many techniques to approximate or estimate the marginal likelihood, such as importance sampling, thermodynamic integration (Lartillot and Philippe, 2006), and bridge sampling (Gronau et al., 2017). For an overview, see Chapter 7 of Gamerman and Lopes (2006), and for a comparative study, see DiCiccio et al. (1997).

One such way is Laplace's method evaluates the integral above to obtain an approximate solution. This is achieved by performing a second-order Taylor expansion on the log-likelihood about the maximum likelihood estimate (MLE) denoted as $\widehat{\boldsymbol{\theta}}$, and integrating out $\boldsymbol{\theta}$ while ignoring $\mathcal{O}(1)$ terms. The result is known as the *Bayesian information criterion* (BIC), given as

$$\text{BIC} \equiv -2 \log p(\boldsymbol{y}|\mathcal{M}) \approx -2 \log \widehat{L} - k \log n, \tag{1.28}$$

where k is the number of parameters in the model, $\widehat{L}$ is the likelihood function evaluated at the MLE, and n is the number of observations. Bayesian information criterion was first introduced in 1978 Schwarz (1978) as a Bayesian adaptation of the frequentist AIC (Akaike information criterion) (Akaike, 1998). Furthermore, BIC combats overfitting by incurring a penalty for the number of parameters in the model. Another main advantage of BIC for model selection is the simplicity of use, which requires only the value of the maximized log-likelihood, the number of parameters, and the total number of observations. Since the BIC in (1.28) is an approximation of the $-2 \log p(\boldsymbol{y}|\mathcal{M})$ and hence models with lower BIC values are usually preferred.

Another well-known method for model selection is the deviance information criterion (DIC) (Spiegelhalter et al., 2002). Deviance is defined as

$$D(\boldsymbol{\theta}) = -2 \log p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}) + 2 \log h(\boldsymbol{y}),$$

where $h(\boldsymbol{y})$ is a fully specified standardizing term which is a function of only the data, which we set to be $h(\boldsymbol{y}) = 1$ for all models. The DIC is comprised on two components,

$$DIC = \overline{D}(\boldsymbol{\theta}) + p_D,$$

where the first term is the posterior expectation of the deviance,

$$\overline{D(\boldsymbol{\theta})} = \mathrm{E}_{\theta|y}\left[D(\boldsymbol{\theta})\right] = \mathrm{E}_{\theta|y}\left[-2 \log p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})\right].$$

This first term assesses how well the model fits the data, with smaller values being a 'better' fit.

The second component, p_D, is the effective number of parameters of the model, which penalises the complexity of the model,

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\boldsymbol{\theta}^*) = \mathrm{E}_{\theta|y}\left[-2\log p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})\right] + 2\log p(\boldsymbol{y}|\boldsymbol{\theta}^*, \mathcal{M}),$$

where $\boldsymbol{\theta}^*$ is the posterior model from the observed data. Thus, the DIC is a trade-off between the complexity and adequacy of a model Chan and Grant (2016). The DIC can be re-written for model $\mathcal{M}$ as

$$DIC \equiv -4\mathrm{E}_{\theta|y}\left[\log p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})\right] + 2\log p(\boldsymbol{y}|\boldsymbol{\theta}^*, \mathcal{M}),$$

where $\boldsymbol{\theta}^*$ is the posterior model. The DIC generalizes AIC, since AIC $= D(\widehat{\boldsymbol{\theta}}) + 2p$ where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (Berg et al., 2004). Furthermore, the DIC is straightforward to estimate, since we only require draws from the posterior as well as log-likelihood evaluated at the posterior mean. For more on predictive information criterion, see Gelman et al. (2014).

1.5.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are an important class of algorithms in Bayesian computation. The MCMC algorithm samples from an unknown probability distribution, e.g. the posterior, via random simulation of a Markov chain with its equilibrium distribution as the posterior (Gelman et al., 1995). These methods are attractive due to their asymptotic properties. Let M be the number of iterations, then as $M \to \infty$, the MCMC algorithm samples the posterior distribution without approximation. MCMC algorithms output a set of posterior samples $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{M}$. As mentioned previously, posterior samples are crucial in computing quantities such as the posterior predictive distribution, the BIC for model selection, and high-dimensional integrals in the form of (1.24). Given iid posterior samples, the law of large numbers states that

$$\frac{1}{M}\sum_{j=1}^{M}\psi(\boldsymbol{\theta}^{(j)}) \xrightarrow{\text{a.s.}} \mathrm{E}_{\pi}[\psi(\boldsymbol{\theta})], \tag{1.29}$$

where a.s. denotes almost sure convergence. In reality, drawing iid samples from a posterior in moderate to high dimensions is not possible, and hence, one must resort to MCMC methods. The algorithm generates a correlated sequence of samples $\{\boldsymbol{\theta}^{(j)}\}_{j\geq J}^{M}$, which is distributed according to the target $\pi(\boldsymbol{\theta})$ for large J. Despite the samples no longer being iid, MCMC algorithms can be employed to sample high-dimensional posteriors. Furthermore, the presence of auto-correlation of the samples, (1.29) still holds; however, the statistical efficiency is reduced because of this

dependence (Geyer, 2011). It can be shown, as $M \to \infty$,

$$\text{Var}[\sqrt{M}\ \overline{\boldsymbol{\theta}}] \to \sigma^2(1 + 2\sum_{k=1}^{\infty} \rho_k),$$

where $\sigma^2 = \text{Var}[\boldsymbol{\theta}^{(j)}]$ and ρ_k is the auto-correlation at lag k of the sequence. Estimating the statistical efficiency of the M correlated samples from an MCMC output is often useful. This motivates the effective sample size (ESS), defined as

$$\text{ESS} \equiv \frac{M}{1 + 2\sum_{k=1}^{\infty} \rho_k}. \tag{1.30}$$

For independent samples, the Monte Carlo estimate has $\text{Var}[\overline{\boldsymbol{\theta}}] = \sigma^2/M$, which implies that the ESS is M. Thus, ESS is the equivalent number of iid samples from the posterior that our M correlated samples have produced. The quantity in the denominator of (1.30) is known as the *inefficiency factor*. The higher the inefficiency factor, the less effective the sampling algorithm and the more samples one has to draw to have the same estimation power for M iid samples from the posterior. In practice, ESS is non-trivial to estimate due to the possibly infinite number of auto-correlation terms in (1.30). See Gong and Flegal (2016); Flegal and Jones (2010) for a more thorough treatment. For a modern review of MCMC, see Brooks et al. (2011) and Gelman et al. (1995).

Metropolis-Hastings algorithm

The Metropolis algorithm, first proposed by Metropolis et al. (1953), then later developed into the Metropolis-Hastings (MH) algorithm in Hastings (1970), is an important MCMC algorithm. The MH algorithm is a randomized algorithm that constructs a Markov Chain to sample from the invariant distribution $\pi(\boldsymbol{\theta})$.

We consider a common variant known as *Random Walk* Metropolis-Hasting (RW-MH), which uses a *proposal* distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_c)$ which generates a new state $\boldsymbol{\theta}_p$ which depends only on the state at the previous iteration $\boldsymbol{\theta}_c$. The RW-MH constructs a Markov chain in the following way:

1. Initialize the chain $\boldsymbol{\theta}_c = \boldsymbol{\theta}^{(0)}$
2. Generate new state $\boldsymbol{\theta}_p$ from the proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_c)$.
3. Compute the *Metropolis-Hastings ratio*,

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}_p)q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}{\pi(\boldsymbol{\theta}_c)q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}\right). \tag{1.31}$$

4. Accept

$$\boldsymbol{\theta}^{(j)} = \begin{cases} \boldsymbol{\theta}_p \text{ with probability } \alpha, \\ \boldsymbol{\theta}_c \text{ with probability } 1-\alpha. \end{cases} \tag{1.32}$$

5. Set $\boldsymbol{\theta}_c = \boldsymbol{\theta}^{(j)}$.

6. Repeat steps 2-5, M times.

The algorithm obtains a sequence of posterior draws $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{N}$. One benefit of the MH algorithm is the avoidance of computing the possibly intractable normalizing constant in (1.23) due to the cancellation from the MH ratio in Step 3. The choice of the proposal distribution q is crucial to the algorithm's efficiency. Intuitively, the proposal distribution should generate new states in areas of high posterior. One common proposal distribution is the multivariate normal $q(\boldsymbol{\theta}) = N(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1})$ where $\boldsymbol{\theta}^*$ is the maximum a posteriori probability (MAP) estimate and $\mathcal{I}(\boldsymbol{\theta}^*)$ is the inverse covariance matrix given by

$$\mathcal{I}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where $\nabla_{\boldsymbol{\theta}}$ is the gradient wrt the parameters. This is useful because the proposal $q(\boldsymbol{\theta})$ mimics the curvature of the log posterior. Furthermore, much work has been done for assessing the optimal scalings of the proposal distributions, see Gelman et al. (1996), Roberts and Rosenthal (2001), Sherlock and Roberts (2009), and for MCMC convergence diagnostics see (Cowles and Carlin, 1996).

1.5.4 Pseudo-marginal MCMC

The Metropolis-Hastings algorithm requires the posterior distribution, up to a normalising constant, to be evaluated in the acceptance ratio (1.31) at each iteration. However, for many models, the posterior is expensive or intractable to evaluate. The pseudo-marginal MCMC algorithm proposed by Andrieu and Roberts (2009) enables one to estimate $\pi(\boldsymbol{\theta})$ to alleviate this burden. It does so by augmenting the parameter space with auxiliary, latent variables $\boldsymbol{u}$, which generate an unbiased estimate of the likelihood $\widehat{p}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta})$. Similar to the marginal MH algorithm, the pseudo-marginal MCMC approach constructs a Markov chain, but on the augmented space $(\boldsymbol{\theta}, \boldsymbol{u})$. This gives the desirable property that the true posterior $\pi(\boldsymbol{\theta})$ is the stationary distribution of the PMMH algorithm once marginalized over $\boldsymbol{u}$. Given an unbiased likelihood estimator $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})$ with the property

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{u}}\Big[\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})\Big] = \int \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u}) p(\boldsymbol{u}) d\boldsymbol{u}, \tag{1.33}$$

where $\boldsymbol{u} \sim p(\boldsymbol{u})$ are the auxiliary, latent variables $\boldsymbol{u}$ which are responsible for generating the estimated likelihood. From Bayes' theorem, the joint posterior of $(\boldsymbol{\theta}, \boldsymbol{u})$ is given by

$$\widehat{\pi}(\boldsymbol{\theta}, \boldsymbol{u}) = \frac{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{u})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \qquad \text{where} \qquad p(\boldsymbol{y}) = \int_{\Theta}\int_{\boldsymbol{u}} \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}d\boldsymbol{\theta}.$$

The unbiasedness condition in (1.33) is important since integrating w.r.t. $\boldsymbol{u}$ gives

$$\begin{aligned}\int \widehat{\pi}(\boldsymbol{\theta}, \boldsymbol{u})d\boldsymbol{u} &= \frac{(\int \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u})p(\boldsymbol{\theta})}{\int_{\Theta}\int_{\boldsymbol{u}} \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}p(\boldsymbol{\theta})d\boldsymbol{\theta}}, \\ &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}, \\ &= \pi(\boldsymbol{\theta}),\end{aligned}$$

which is the exact posterior. Hence, running MCMC on the artificially augmented parameter space produces samples $\{\boldsymbol{\theta}^{(j)}, \boldsymbol{u}^{(j)}\}_{j=1}^{M}$ and disregarding the samples pertaining to $\boldsymbol{u}$ corresponds to integrating out $\boldsymbol{u}$. Thus, as $M \to \infty$, samples are generated from the exact posterior $\pi(\boldsymbol{\theta})$ and exact Bayesian inference for $\boldsymbol{\theta}$ is performed. The PMMH algorithm is attractive since the implementation is straightforward, provided the user has an unbiased estimator of the likelihood. The steps to draw from $\pi(\boldsymbol{\theta}, \boldsymbol{u})$ are described below:

1. Propose $\boldsymbol{u}_p \sim p(\boldsymbol{u})$ and $\boldsymbol{\theta}_p \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}_c)$

2. Compute the Pseudo-marginal Metropolis-Hastings ratio,

$$\alpha = \min\left(1, \frac{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}_p, \boldsymbol{u}_p)p(\boldsymbol{\theta}_p)q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}{\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}_c, \boldsymbol{u}_c)p(\boldsymbol{\theta}_c)q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_p)}\right). \tag{1.34}$$

3. Accept

$$(\boldsymbol{\theta}^{(j)}, \boldsymbol{u}^{(j)}) = \begin{cases} (\boldsymbol{\theta}_p, \boldsymbol{u}_p) \text{ with probability } \alpha, \\ (\boldsymbol{\theta}_c, \boldsymbol{u}_c) \text{ with probability } 1-\alpha. \end{cases} \tag{1.35}$$

4. Set $(\boldsymbol{\theta}_c, \boldsymbol{u}_c) = (\boldsymbol{\theta}^{(j)}, \boldsymbol{u}^{(j)})$.

In a Bayesian setting, unbiased estimators of the likelihood are constructed via importance sampling (Beaumont, 2003), particle filters for state space models (Andrieu et al., 2010), and for subsampling (Quiroz et al., 2021). Despite the benefits of using an estimated likelihood to avoid computation of an intractable likelihood, PMMH is very sensitive to the $\text{Var}\left[\log \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})\right]$. Notice that for extreme overestimates of $p(\boldsymbol{y}|\boldsymbol{\theta})$, the acceptance probability in (1.34) can become 1 for the current iteration, then close to 0 for the subsequent iterations. This results in the chain getting stuck for any number of iterations. Increasing the number of particles, i.e. decreasing

the variance of the likelihood estimator, will increase the efficiency of the Markov chain, but at the cost of computation time and likewise for visa-versa. This computation time-variance trade-off has been studied in Pitt et al. (2012) and Doucet et al. (2015) with the latter finding $\text{Var}\left[\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})\right] \approx 1$ the optimal trade-off. The pseudo-marginal approach is a foundational part of large data subsampling in Quiroz et al. (2019), which we employ in Chapter 3 for large time series data in the frequency domain.

References

Abramowitz, M. and Stegun, I. A. (1968). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government Printing Office.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Ammar, G. S. and Gragg, W. B. (1988). Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9(1):61–76.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Benedetto, J. J. (1992). Irregular sampling and frames. *Wavelets: A Tutorial in Theory and Applications*, 2:445–507.

Berg, A., Meyer, R., and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics*, 22(1):107–120.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. SIAM.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods*. Springer Science & Business Media.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, 90(14):2499–2523.

Chaloner, K. (1996). Elicitation of prior distributions. In *Bayesian Biostatistics*, pages 141–156. CRC Press.

Chan, J. C. and Grant, A. L. (2016). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, 100:847–859.

Chatfield, C. and Xing, H. (2019). *The Analysis of Time Series: An Introduction with R*, volume 7. CRC press.

Christakos, G. (2012). *Random Field Models in Earth Sciences*. Dover Publications.

Cohen, F. S., Fan, Z., and Patel, M. A. (1991). Classification of rotated and scaled textured images using Gaussian Markov random field models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(02):192–202.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91:883–904.

Cryer, J. and Chan, K. (2008). *Time Series Analysis: With Applications in R*. Springer Texts in Statistics. Springer New York.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38:1034–1070.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC.

Gangopadhyay, A., Mallick, B., and Denison, D. (1999). Estimation of spectral density of a stationary time series via an asymptotic representation of the periodogram. *Journal of Statistical Planning and Inference*, 75(2):281–290.

Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*, 20116022(45):22.

Gong, L. and Flegal, J. M. (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25(3):684–700.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1):152–177.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97.

Guillaumin, A. P., Sykulski, A. M., Olhede, S. C., and Simons, F. J. (2022). The debiased spatial Whittle likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1526–1557.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105.

Hastings, W. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109.

He, X., Zemel, R. S., and Carreira-Perpinán, M. A. (2004). Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Joseph, J. B. and Freeman, A. (2003). *The Analytical Theory of Heat*. Courier Corporation.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, page 282–289.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207.

Lehmann, E. L. and Casella, G. (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Lindgren, G., Rootzen, H., and Sandsten, M. (2013). *Stationary Stochastic Processes for Scientists and Engineers*. CRC Press.

Mallick, B., Denison, D., and Gangopadhyay, A. (2002). A Bayesian curve fitting approach to power spectrum estimation. *Journal of Nonparametric Statistics*, 14(1-2):141–153.

Matérn, B. (2013). *Spatial Variation*, volume 36. Springer Science & Business Media.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Peligrad, M. and Wu, W. B. (2010). Central limit theorem for Fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022.

Peligrad, M. and Zhang, N. (2019). Central limit theorem for Fourier transform and periodogram of random fields. *Bernoulli*, 25(1):499–520.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021). The block-Poisson estimator for optimally tuned exact subsampling MCMC. *Journal of Computational and Graphical Statistics*, 30(4):877–888.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Shao, X. and Wei, B. W. (2007). Asymptotic spectral theory for nonlinear time series. *Annals of Statistics*, 35(4):1773–1801.

Sherlock, C. and Roberts, G. (2009). Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15:774–798.

Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.

Simons, F. J. and Olhede, S. C. (2013). Maximum-likelihood estimation of lithospheric flexural rigidity, initial-loading fraction and load correlation, under isotropy. *Geophysical Journal International*, 193(3):1300–1342.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Sykulski, A. M., Olhede, S. C., Guillaumin, A. P., Lilly, J. M., and Early, J. J. (2019). The debiased Whittle likelihood. *Biometrika*, 106(2):251–266.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Statistics / Uppsala universitet. Almqvist & Wiksells boktr.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918.

Chapter 2

Improving forecasting in dynamic linear regression models by a semi-long memory model for the error process

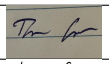
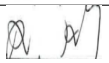
Thomas Goodwin, Matias Quiroz and Robert Kohn [1]

Abstract

Dynamic linear regression models forecast the values of one time series based on a linear combination of a set of exogenous time series while incorporating a time series process for the error term. This error process is often assumed to follow an autoregressive integrated moving average (ARIMA) model, or seasonal variants thereof, which are unable to capture a long-range dependency structure of the error process. We propose a novel dynamic linear regression model that incorporates the long-range dependency feature of the errors and shows that it improves the model's forecasting ability. We develop a Markov chain Monte Carlo method to fit general dynamic linear regression models based on a frequency domain approach that enables fast, asymptotically exact Bayesian inference for large datasets. We demonstrate that our approximate algorithm is faster than the traditional time-domain approaches, such as the Kalman filter and the multivariate Gaussian likelihood, while retaining a high accuracy when approximating the posterior distribution. We illustrate the method in simulated examples and two energy forecasting applications.

Keywords: Dynamic linear regression, forecasting, Bayesian inference.

[1]Goodwin: *School of Mathematical and Physical Sciences, University of Technology Sydney.* Quiroz: *Department of Statistics, Stockholm University.* Kohn: *School of Economics, University of New South Wales.*

Author	Project conception	Work	Manuscript writing	Manuscript editing	Signature
Tom Goodwin		X	X	X	
Matias Quiroz	X		X	X	
Robert Kohn				X	

Status of paper

This chapter is presented as a draft manuscript under preparation to submit to the International Journal of Forecasting. In the coming weeks, this manuscript will be uploaded to arXiv. I certify that the work in Chapter 2 has not been submitted as part of any other documents required for a degree.

2.1 Introduction

Forecasting time series data plays an important role in various fields, such as engineering, economics, and climate sciences. Forecasting a single output time series may be more accurate by using a linear combination of one or more exogenous time series that explains some of its historical variation (Hyndman and Athanasopoulos, 2018). However, the linear combination of the exogenous time series often does not capture all the serial correlation present in the output time series, resulting in errors that are autocorrelated, i.e. serially correlated. Dynamic linear regression models (Pankratz, 2012) provide a framework that relates the output time series to a linear combination of the exogenous time series and, moreover, models the resulting error term as a time series process to account for serially correlated errors. Dynamic linear regression models are a particular case of the more general class transfer function models (Box et al., 2015).

The standard approach assumes that the error process, i.e. the time series process for the error terms, follows an autoregressive integrated moving average process (ARIMA). This approach allows the error process to have an autoregressive component and a moving average component for the white noise error. While this error process accommodates a wide range of stationary processes, it may require a large number of autoregressive components in order to accommodate error processes that show significant autocorrelation. To model the autocorrelation of the error process parsimoniously, we use the autoregressive tempered fractionally integrated moving average (ARTFIMA) (Meerschaert et al., 2014; Sabzikar et al., 2019). The ARTFIMA class nests the well-known autoregressive fractionally integrated moving average (ARFIMA) (Granger and Joyeux, 1980) model, which is useful for modelling time series with so-called long memory. The ARTFIMA model has the following three advantages compared to the ARFIMA model, where the latter two are of particular importance for the frequency domain estimation approach for

dynamic linear regression models proposed in this paper. First, the autocovariance function of the ARFIMA model decays at such a slow rate that it is not absolutely summable, making it hard to analyse (Sabzikar et al., 2019). The autocorrelation function of the ARFIMA process exhibits a power-law decay as the lags increase, which decays much slower than exponential decay. Such a process is referred to as having *long memory*. In contrast, the ARTFIMA model has a summable autocovariance function, which exhibits long-range dependence for a number of lags but eventually decays exponentially fast. Such a process is referred to as having *semi-long memory*. Second, the spectral density of the ARFIMA process diverges as the frequency approaches zero. In practice, however, the empirical power spectrum (the estimate of the spectral density) is bounded for small frequencies. This is empirically illustrated in Meerschaert et al. (2014); Sabzikar et al. (2019), who show that the ARTFIMA process fits the power spectrum better than the ARFIMA process for low frequencies; see also Figure 2.7 in our paper. This stylised fact is important for our purpose as we develop an estimation method based on a parametric Whittle likelihood (Whittle, 1953), which is a frequency domain approximation of the time domain likelihood. Third, the Whittle approximation fails to hold for long memory processes, especially for small frequencies (Robinson, 1995; Rousseau et al., 2012), which we demonstrate this with a simulation study in Section 2.4.2. Thus, we show that resorting to the time domain likelihood for estimating a dynamic linear regression with ARFIMA errors is computationally much more costly than our approach of a frequency domain likelihood with ARTFIMA errors. Moreover, in terms of prediction accuracy, our model is on par with, if not better for longer prediction horizons for the empirical data sets analysed. Our model's distributional forecasts are also on par with the existing models.

The model parameters in dynamic linear regression models consist of the regression coefficients that form the linear combination of the exogenous time series and the parameters of the error process. For the standard dynamic linear regression models with ARIMA errors that are normally distributed, efficient likelihood-based inference can be carried out by finding the finite-dimensional state space representation of the model and using the Kalman filter to integrate out the unobserved time-varying error terms. This may still be computationally costly with many time observations, especially in Bayesian inference, which typically requires many posterior samples for reliable inference. For each such sample, the Kalman filter needs to cycle through all observations, which can be prohibitively expensive, especially for large time series. The Whittle log-likelihood can be derived using large sample properties of the so-called periodogram data, which are formed via the discrete Fourier transform of the time domain data. The Whittle log-likelihood is directly a function of the regression coefficients and the error process parameters without needing to integrate out unobserved error terms. However, in dynamic linear regression models, the periodogram data becomes a function of the regression coefficients, which requires recomputing the discrete Fourier transform in every iteration. We show how this can be cir-

cumvented, and thus, our algorithm requires computing the discrete Fourier transform only once before inference, thereby obtaining significant computational gains compared to the time domain log-likelihood based on the Kalman filter. Our frequency domain approach is also applicable to a dynamic linear regression model with ARTFIMA errors, where a finite-dimensional state space representation is not readily available.

To summarise, our article has two contributions. First, we propose a frequency domain estimation approach for dynamic linear models that significantly outperforms estimation approaches based on the time domain likelihood in terms of computing time, especially when a finite-dimensional state space representation of the model is not available. Second, we utilise a semi-long memory process for the error process and show that it provides more accurate forecasts compared to both the standard dynamic linear model and that of using a long memory process for the error term.

The rest of the article is organised as follows. Section 2.2 reviews existing dynamic linear models and presents our extension. Section 2.3 introduces the necessary frequency domain tools, outlines our estimation method and validates its performance relative to the time domain likelihood. Section 2.5 presents applications for two real-world electricity demand data sets and improved forecasts via dynamic linear regression models with ARTFIMA errors.

2.2 Dynamic linear regression models

2.2.1 Standard dynamic linear regression models

Let $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{mt}) \in \mathbb{R}^m$ be a set of m exogenous stationary time series observed at time t. A dynamic linear regression model models the single output time series $Y_t \in \mathbb{R}$ as a linear combination of the exogenous $\boldsymbol{X}_t$, i.e.

$$Y_t = \boldsymbol{X}_t^\top \boldsymbol{\beta} + \eta_t, \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ is a vector of regression coefficients and $\eta_t \in \mathbb{R}$ is a zero-mean error process. The error process contains all unobserved factors that affect Y_t, and we assume that $\mathrm{E}(\eta_t|\boldsymbol{X}_t) = 0$, i.e. it is uncorrelated with each of the elements in $\boldsymbol{X}_t$ and hence $\boldsymbol{X}_t$ is exogenous.

When the error process η_t does not show any temporal dependence, (2.1) is a standard linear regression (given it has a constant variance) and can easily be estimated using standard linear regression approaches (or modified versions thereof if heteroscedasticity is present). However, in many applications, the unobserved factors are time-varying, resulting in serially correlated errors η_t. The standard dynamic linear regression model assumes that η_t is an autoregressive integrated

moving average process, denoted ARIMA(p, d, q) and defined as

$$\phi_p(B)\Delta^d \eta_t = \psi_q(B)\varepsilon_t, \tag{2.2}$$

where $\phi_p(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ and $\psi_q(B) = 1 + \sum_{i=1}^{q} \psi_i B^i$ are the autoregressive and moving average lag polynomials respectively with the lag operator B such that $B^i \eta_t = \eta_{t-i}$. The differencing operator Δ^d, for $d = 0, 1, 2, \ldots$, is defined as $\Delta^d \eta_t = (1 - B)^d \eta_t$. When the error process exhibits seasonality with seasonal period s, this can be modelled by a seasonal ARIMA process, denoted ARIMA$(p, d, q)(P, D, Q)_s$ and defined as

$$\phi_p(B)\phi_P^\star(B^s)\Delta^d \Delta_s^D \eta_t = \psi_q(B)\psi_Q^\star(B^s)\varepsilon_t, \tag{2.3}$$

where $\phi_P^\star(B^s) = 1 - \sum_{i=1}^{P} \phi_i^\star B^{is}$ and $\psi_Q^\star(B^s) = 1 + \sum_{i=1}^{Q} \psi_i^\star B^{is}$ are the seasonal autoregressive and seasonal moving average lag polynomials, and $\Delta_s^D = (1 - B^s)^D$, for $D = 0, 1, 2, \ldots$, is the seasonal differencing operator (Box et al., 2015).

To carry out inference in (2.1), one typically assumes that η_t follows a normal distribution. The resulting log-likelihood is then a multivariate Gaussian distribution with a Toeplitz covariance matrix given that η_t is stationary (Doornik and Ooms, 2003). Inversion of such a matrix is usually performed via the Levinson-Durbin algorithm (Levinson, 1946; Durbin, 1960) in $\mathcal{O}(T^2)$ operations. However, recent approaches, known as *superfast* Toeplitz algorithms, can solve the matrix system in $\mathcal{O}(T \log^2 T)$ operations, which becomes more efficient than the Levinson-Durbin algorithm when $T > 256$ (Ammar and Gragg, 1988). A more computationally efficient approach is to find the finite-dimensional state space representation of the model in (2.1) with the errors following either (2.2) or (2.3), in which the resulting likelihood can be evaluated in $\mathcal{O}(T)$ using the Kalman filter.

2.2.2 Long memory dynamic linear regression models

In many applications, a pair of observations separated by a long time interval exhibit a non-negligible correlation. The standard dynamic linear regression model with the errors in (2.2) has an autocovariance function whose absolute value decays exponentially fast and can thus not capture this feature. A common approach to model time series with long memory is to consider so-called fractional differencing.

In a dynamic linear regression setting, Doornik and Ooms (2004) propose to model the error process η_t in (2.1) with an autoregressive fractionally integrated moving average process (Granger and Joyeux, 1980; Hosking, 1981), denoted ARFIMA(p, d, q). This model is, for $d \notin \mathbb{Z}$,

$$\phi_p(B)\Delta^d \eta_t = \psi_q(B)\varepsilon_t, \tag{2.4}$$

where $\phi_p(B)$ and $\psi_q(B)$ are defined as in (2.2). For $d \notin \mathbb{Z}$ the fractional differencing operator is defined via the fractional binomial theorem

$$\begin{aligned}\Delta^d \eta_t &= (1-B)^d \eta_t \\ &= \sum_{j=0}^{\infty} \binom{d}{j} (-B)^j \eta_t \\ &= \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} \eta_{t-j},\end{aligned}$$

where $\Gamma(s) = \int \exp(-t) t^{s-1} dt$ and we have used that $s! = \Gamma(1+s)$. Provided that the roots of the polynomial $\phi_p(z)$ are outside the unit circle in the complex plane, the ARFIMA process is stationary if $-0.5 < d < 0.5$ and has long memory when $0 < d < 0.5$ (Granger and Joyeux, 1980). Doornik and Ooms (2004) also propose a seasonal extension by modelling η_t as a seasonal ARFIMA$(p, d, q)(P, D, Q)_s$ with $d \notin \mathbb{Z}$ and integer D. A possible extension of this model is to allow for non-integer D as in Bisognin and Lopes (2009).

Doornik and Ooms (2003) show that maximum likelihood estimation for ARFIMA models can be carried out in $\mathcal{O}(T^2)$ time, and this also applies for the model in (2.1) with the errors following (2.4). As mentioned previously, more recent algorithms can estimate these models in $\mathcal{O}(T \log^2 T)$ time (Ammar and Gragg, 1988). Chan and Palma (1998) show that there is no finite-dimensional state space representation of an ARFIMA process. They propose an infinite-dimensional state space representation for which the Kalman filter can be computed in a finite number of steps equal to T; however, the resulting computation is $\mathcal{O}(T^3)$.

2.2.3 Semi long memory dynamic linear regression models

The autoregressive tempered fractional integrated moving-average (ARTFIMA) (Meerschaert et al., 2014; Sabzikar et al., 2019) is an extension of the ARFIMA model that incorporates a tempering parameter λ. The model for the error process is then

$$\phi_p(B) \Delta^{d,\lambda} \eta_t = \psi_q(B) \varepsilon_t, \tag{2.5}$$

where $\phi_p(B)$ and $\psi_q(B)$ are defined as in (2.2) and for $d \notin \mathbb{Z}$ and $\lambda > 0$ the tempered fractional differencing operator is

$$\begin{aligned}\Delta^{d,\lambda}\eta_t &= (1 - \exp(-\lambda)B)^d \eta_t \\ &= \sum_{j=0}^{\infty} \binom{d}{j} (-B)^j \eta_t \\ &= \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} \exp(-\lambda j)\eta_{t-j}. \end{aligned} \tag{2.6}$$

Note that when $\lambda = 0$, the ARTFIMA model becomes a stationary ARFIMA model if $-0.5 < d < 0.5$, provided the roots of the polynomial $\phi_p(z)$ are outside the unit circle in the complex plane. When $\lambda = 0$ and d is integer-valued, the ARTFIMA model becomes an ARIMA model. The ARTFIMA model is stationary for all $d \notin \mathbb{Z}$ and $\lambda > 0$ provided that the root condition of the polynomial $\phi_p(z)$ is satisfied (Sabzikar et al., 2019). The ARTFIMA process has several advantages over the ARFIMA process, which are listed in Section 2.1.

Similar to the ARFIMA model, a finite-dimensional state space representation that allows efficient Kalman filter computations is not readily available. Instead, the log-likelihood can be computed using a multivariate Gaussian distribution with a Toeplitz covariance matrix formed via the autocovariance function in Sabzikar et al. (2019, Theorem 2.5 (b)). However, a more computationally efficient approach is forming a frequency domain log-likelihood for the so-called periodogram data described in Section 2.3. Let ω denote the angular frequency. The approach uses the process's spectral density $f(\omega)$, which is the Fourier transform of the covariance function. The spectral density shows how the variation of the time series is distributed in the frequency spectrum. The spectral density when η_t follows the ARTFIMA process in (2.5) is (Sabzikar et al., 2019, Theorem 2.5 (a))

$$f(\omega; \boldsymbol{\vartheta}) = \frac{\sigma_\varepsilon^2}{2\pi} \left|1 - e^{-(\lambda + i\omega)}\right|^{-2d} \left|\frac{\psi_q(e^{-\mathrm{i}\omega})}{\phi_p(e^{-i\omega})}\right|^2, \tag{2.7}$$

with parameter vector $\boldsymbol{\vartheta} = (\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q, d, \lambda, \sigma_\varepsilon^2)$, where $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_t)$ and i is the imaginary number. The spectral density for the seasonal ARTFIMA$(p,d,q)(P,D,Q)_s$ with $d \notin \mathbb{Z}$ and integer D, i.e.

$$\phi_p(B)\phi_P^\star(B^s)\Delta^{d,\lambda}\Delta_s^D\eta_t = \psi_q(B)\psi_Q^\star(B^s)\varepsilon_t, \tag{2.8}$$

where all quantities have been previously defined, is

$$f(\omega;\boldsymbol{\vartheta}) = \frac{\sigma_\varepsilon^2}{2\pi}\left|1-e^{-(\lambda+i\omega)}\right|^{-2d}\left|\frac{\psi_q(e^{-i\omega})\psi_Q^\star(e^{-is\omega})}{\phi_p(e^{-i\omega})\phi_P^\star(e^{-is\omega})}\right|^2. \tag{2.9}$$

Figure 2.1 shows the spectral density, also known as the power, for both the ARFIMA and ARTFIMA process for different values of the fractional differencing parameter d. An important difference is the limiting behaviour as $\omega \to 0$, where the spectral density of the ARFIMA process diverges when $d > 0$, whereas that of the ARTFIMA process does not. When estimating the power for real data, it is often observed to be bounded. Hence the ARTFIMA model will provide a better fit; see Figure 2.7.

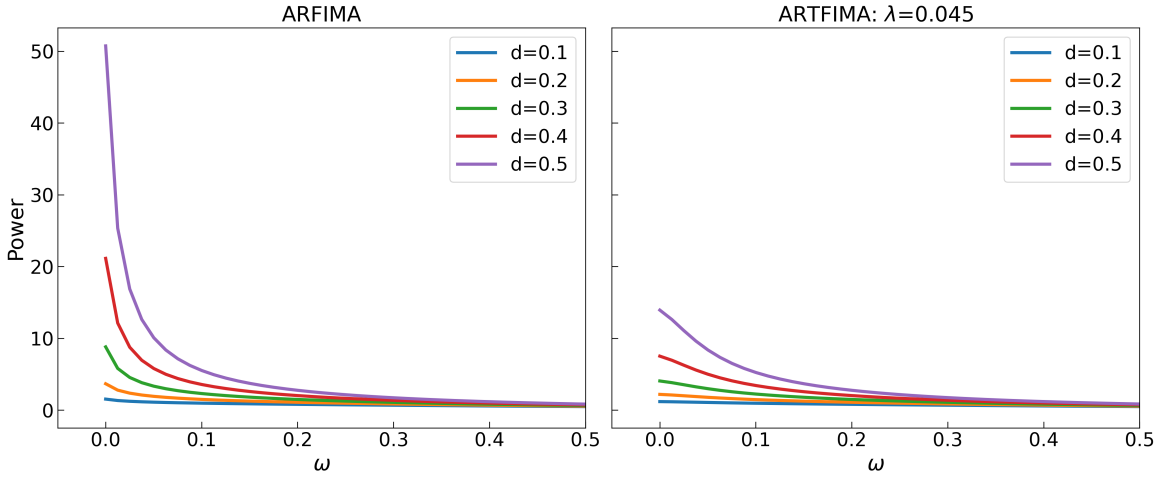


Figure 2.1: Spectral densities of different ARFIMA and ARTFIMA models with $p = 1$ and $q = 0$. For both processes, the parameter $\phi = 0.5$ and ARTFIMA $\lambda = 0.045$.

2.3 Methodology

2.3.1 Frequency domain likelihood

The frequency domain approach to inference relies on the asymptotic properties of the frequency representation of the time series process in (2.1). Engle (1974) shows how to rewrite (2.1) in terms of the periodograms of Y_t and $\boldsymbol{X}_t$ and derive the ordinary least squares estimator of the transformed regression, which is also shown to be the best linear unbiased estimator. Our Bayesian approach requires a log-likelihood function, which can be derived using asymptotic properties of the periodogram data, which we now outline in detail.

Let Y_t and $\boldsymbol{X}_t$ be zero-mean processes. Suppose that $\boldsymbol{\beta}$ is known and define the pseudo data

$$Z_t = Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta},\ t = 1, \ldots, T.$$

The general case assumes that η_t in (2.1) is a seasonal ARTFIMA process and hence the spectral density of Z_t is given by (2.9). The spectral density of the ARTFIMA process in (2.7) is obtained using $Q = P = 0$ in (2.9). The spectral density of the seasonal ARMA process is obtained by setting $\lambda = 0$ and $d = 0$ in (2.9), and for ARMA, in addition, $Q = P = 0$.

Let $\omega \in [-\pi, \pi]$ be the angular frequency and denote the natural Fourier frequencies as $\omega_k = \frac{2\pi k}{T}$ for $k \in \mathcal{K}$, where

$$\mathcal{K} = \begin{cases} -\frac{T}{2}, -\frac{T}{2}+1, \ldots, \frac{T}{2}-1, & \text{if } T \text{ is even,} \\ -\frac{(T-1)}{2}, -\frac{(T-1)}{2}+1, \ldots, \frac{(T-1)}{2}, & \text{if } T \text{ is odd.} \end{cases}$$

The frequency representation of the time series process Z_t is obtained via its discrete Fourier transform (DFT), which is the complex-valued transform

$$J_Z(\omega_k) = \sum_{t=1}^{T} Z_t \exp(-i\omega_k t), \tag{2.10}$$

which can be computed for all T frequencies ω_k, $k \in \mathcal{K}$, using $\mathcal{O}(T\log(T))$ operations via the fast Fourier transform (Cooley and Tukey, 1965). Since the pseudo data depends on $\boldsymbol{\beta}$, the DFT in (2.10) needs to be recomputed for each new sample in the Markov chain Monte Carlo algorithm. However, following Matsuda and Yajima (2009),

$$\begin{aligned} J_Z(\omega_k) &= \sum_{t=1}^{T} (Y_t - \boldsymbol{X}_t^\top \boldsymbol{\beta}) \exp(-i\omega_k t) \\ &= \sum_{t=1}^{T} Y_t \exp(-i\omega_k t) - \left(\sum_{t=1}^{T} \boldsymbol{X}_t \exp(-i\omega_k t) \right)^\top \boldsymbol{\beta} \\ &= J_Y(\omega_k) - \boldsymbol{J}_{\boldsymbol{X}}(\omega_k)^\top \boldsymbol{\beta}, \end{aligned}$$

where J_Y is the DFT of Y_t, and $\boldsymbol{J}_{\boldsymbol{X}}(\omega_k)$ is the m-dimensional row-vector with the jth element being the DFT of the (univariate) time series X_{jt}. Thus, since $J_Y(\omega_k)$ and $J_{\boldsymbol{X}}(\omega_k)$ only depend on the data Y_t and $\boldsymbol{X}_t$, and can be pre-computed before the Markov chain Monte Carlo algorithm, $J_Z(\omega_k)$ can be evaluated in $\mathcal{O}(T)$ for all frequencies when $\boldsymbol{\beta}$ changes.

The DFT of $J_Z(\omega_k)$ is a weighted complex valued sum of the pseudo data. Peligrad and Wu

(2010, Theorem 2.1) show that under quite regular conditions,

$$\frac{1}{\sqrt{T}}(\Re(J_Z(\omega_k)), \Im(J_Z(\omega_k))), \quad T \to \infty, \tag{2.11}$$

where, respectively, $\Re(z)$ and $\Im(z)$ denote the real and imaginary parts of z, converge in distribution to a bivariate normal distribution with (asymptotically) independent components having expected value 0 and variance $\pi f(\omega_k)$, with f being the spectral density of Z_t. Moreover, they show that $\frac{1}{\sqrt{T}} J_Z(\omega_k)$ are asymptotically independent for all $k \in \mathcal{K}$.

Define the periodogram

$$I_Z(\omega_k) = \frac{1}{2\pi} \left| \frac{1}{\sqrt{T}} J_Z(\omega_k) \right|^2. \tag{2.12}$$

Then, for $k \in \mathcal{K} \setminus 0$, by (2.11),

$$\frac{I_Z(\omega_k)}{f(\omega_k)} = \frac{1}{2} \left| \frac{1}{\sqrt{\pi f(\omega_k)}} \frac{1}{\sqrt{T}} J_Z(\omega_k) \right|^2 \sim \frac{\chi^2(2)}{2}, \tag{2.13}$$

as $T \to \infty$, where $\chi^2(\nu)$ denotes the chi-squared distribution with ν degrees of freedom. Since $\frac{\chi^2(2)}{2}$ is a standard exponential random variable, it follows that

$$I_Z(\omega_k) \sim \text{Exp}(f(\omega_k)), \ k \in \mathcal{K} \setminus 0, \tag{2.14}$$

where Exp denotes an exponential random variable parameterised by its mean. Emphasising the dependence on the parameters, the log-density of (2.14) is

$$\log p(I_Z(\omega_k; \boldsymbol{\beta}) | \boldsymbol{\theta}) = -\log(f(\omega_k; \boldsymbol{\vartheta})) - \frac{I_Z(\omega_k; \boldsymbol{\beta})}{f(\omega_k; \boldsymbol{\vartheta})}, \ k \in \mathcal{K} \setminus 0, \tag{2.15}$$

where $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\beta})$ contains all unknown parameters. That $\nu = 2$ follows from $|\cdot|^2$ in (2.13) being a sum of two squared (asymptotically) independent standard normal random variables when $k \in \mathcal{K} \setminus 0$. When $k = 0$, $J_Z(0) = 0$ since $J_Y(0) = \sum_{t=1}^T Y_t = 0$ and $\boldsymbol{J}_{\boldsymbol{X}}(0) = \sum_{t=1}^T \boldsymbol{X}_t = \boldsymbol{0}$ for demeaned data.

The asymptotic distributions of the periodogram data underlie the idea of the so-called Whittle log-likelihood (Whittle, 1953): Form the log-likelihood via the distributions of the frequency domain data, i.e. the periodogram ordinates $I_Z(\omega_k; \boldsymbol{\beta})$. For a real-valued process Z_t, both the periodogram and spectral density are symmetric around the origin; hence, only non-negative frequencies are considered. The Whittle log-likelihood is obtained by adding (due to the asymptotic independence) the log-densities in (2.15) for all the positive frequencies (zeroth frequency

excluded with demeaned data). The Whittle log-likelihood is, for odd T,

$$\ell_W(\boldsymbol{\theta}) = -\sum_{k=1}^{(T-1)/2} \left(\log(f(\omega_k; \boldsymbol{\vartheta})) + \frac{I_Z(\omega_k; \boldsymbol{\beta})}{f(\omega_k; \boldsymbol{\vartheta})} \right), \tag{2.16}$$

and the summation runs to $T/2 - 1$ instead if T is even.

The Whittle log-likelihood is an approximation of the Gaussian time domain likelihood. Guyon (1982) and Kent and Mardia (1996) investigate the rates of asymptotic equivalence for the aforementioned log-likelihoods,

$$|\mathcal{L}_{\text{true}}(\boldsymbol{\theta}) - \mathcal{L}_W(\boldsymbol{\theta})| = \mathcal{O}_p(1), \tag{2.17}$$

as $T \to \infty$. The result in (2.17) also holds for the first and second-order derivatives of the likelihoods. Further, it can be shown that the bias incurred from the Whittle likelihood is smaller than the standard error.

Finally, we note that a frequency domain approach for dynamic linear regression models with the Whittle log-likelihood is not appropriate when using the ARFIMA model because the Whittle approximation fails to hold for long memory processes, in particular for the small frequencies (Robinson, 1995; Rousseau et al., 2012). This is further investigated in a simulation study in Section 2.4.2. Thus, to carry out inference in dynamic linear regression models with ARFIMA errors, one has to resort to the time domain likelihood, which is considerably slower to compute than a frequency domain approach that uses the ARTFIMA process.

2.3.2 Bayesian inference via Markov chain Monte Carlo

An important objective in time series is to learn $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \boldsymbol{\beta})$ given realisations of the time series processes Y_t and $\boldsymbol{X}_t$. Let $p(\boldsymbol{Z}|\boldsymbol{\theta})$ denote the likelihood function of $\boldsymbol{\theta}$ given the pseudo data $\boldsymbol{Z} = (Z_1, \ldots, Z_T)$, which depends on the subset $\boldsymbol{\beta}$ of the parameter vector $\boldsymbol{\theta}$, but we suppress this dependence for simplicity. When the likelihood is obtained via the Whittle log-likelihood in (2.16), $p(\boldsymbol{Z}|\boldsymbol{\theta}) = \exp(\ell_W(\boldsymbol{\theta}))$. The likelihood function in this case is given the periodogram ordinates $I(\omega_0), \ldots, I(\omega_{(T-1)/2})$, but since they are functions of $\boldsymbol{Z}$ we keep using the notation $p(\boldsymbol{Z}|\boldsymbol{\theta})$.

The cornerstone of Bayesian inference is the posterior distribution obtained via Bayes' theorem,

$$p(\boldsymbol{\theta}|\boldsymbol{Z}) = \frac{p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{Z})}, \; p(\boldsymbol{Z}) = \int p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{2.18}$$

where $p(\boldsymbol{\theta})$ is the prior distribution. Markov chain Monte Carlo is a class of iterative procedures

to sample from (2.18), with the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) arguably being the most used. The Metropolis-Hastings algorithm constructs a Markov chain $\left\{\boldsymbol{\theta}^{(j)}\right\}$ by starting at some initial value $\boldsymbol{\theta}^{(0)}$ and then, recursively, proposes a candidate draw $\boldsymbol{\theta}'$ from a proposal density $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j-1)})$ and sets $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}'$ with acceptance probability

$$\alpha_{\text{MH}} = \min\left(1, \frac{p(\boldsymbol{Z}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')/q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(j-1)})}{p(\boldsymbol{Z}|\boldsymbol{\theta}^{(j-1)})p(\boldsymbol{\theta}^{(j-1)})/q(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}')}\right). \tag{2.19}$$

If a proposed draw is rejected, $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$. The acceptance probability in (2.19) is set so that, informally speaking, the Markov chain consists of samples from the posterior in (2.18) after a warm-up period referred to as the burn-in. Let $\left\{\boldsymbol{\theta}^{(j)}\right\}_{j=1,\ldots,N}$ be the samples after discarding the burn-in.

If the algorithm rejects too often, the Markov chain becomes "sticky", which causes inefficient estimates of posterior expectations. To explain how this inefficiency is measured, suppose θ is scalar-valued and consider estimating the expectation of a function h, i.e.

$$\text{E}(h(\theta)) = \int h(\theta)p(\theta|\boldsymbol{Z})d\theta. \tag{2.20}$$

By the law of large numbers,

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(j)}) \xrightarrow{p} \text{E}(h(\theta)). \tag{2.21}$$

If the samples were independent, then the asymptotic variance of $\sqrt{N}\widehat{I}_N$ is σ_h^2, where $\sigma_h^2 = \text{V}(h(\theta))$. However, Markov chain Monte Carlo results in correlated samples and then the asymptotic variance of $\sqrt{N}\widehat{I}_N$ is $\sigma_h^2(1+\sum_{i=1}^{\infty}\rho_i)$, with ρ_i is the autocorrelation at the ith lag the Markov chain. The term $(1+\sum_{i=1}^{\infty}\rho_i)$ is called the inefficiency factor. It measures how much the asymptotic variance of the estimate of a posterior expectation is inflated compared to an ideal sampler that produces independent draws. The effective sample size (ESS) of the resulting Markov chain is defined as $\text{ESS} = N/(1+\sum_{i=1}^{\infty}\rho_i)$. Thus, the effective sample size is an estimate of the number of independent samples that has the same precision as our N autocorrelated sample.

The proposal density is often taken to be a random walk, e.g.

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j-1)}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j-1)}, c\boldsymbol{\Sigma}_{\text{prop}}),$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density of $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma}$. Moreover, $\boldsymbol{\Sigma}_{\text{prop}}$ is an approximation of the posterior covariance matrix and $c = 2.38/\sqrt{\dim(\boldsymbol{\theta})}$ for optimality (Gelman et al., 1997).

Despite selecting a proposal distribution, as previously mentioned, sampling high-dimensional or difficult geometries/ correlation structures in the posterior can still yield low acceptance probabilities. One possibility to overcome this issue is to adaptively tune the proposal covariance matrix (Haario et al., 2001). This technique uses the full information accumulated thus far using the previous draws from the algorithm to continually update the covariance proposal matrix. The adaptive approach can lead to more accurate simulated approximations of the target posterior and less auto-correlation in the resulting chain.

To target a specified overall acceptance rate of the sampler, we follow the approach from Garthwaite et al. (2016). This approach utilizes the Robbins-Munro search process to quickly identify the scaling constant c for the proposal covariance matrix. For Gaussian target posteriors, Gelman et al. (1997) show that the optimal acceptance rate is 0.234. Garthwaite et al. (2016) propose a simple estimator for c which targets the optimal acceptance rate of 0.234 and shows that the optimal step length for the search process is robust for a wide range of distributions. We combine this method with the adaptive covariance technique described above for a sampler for DLR models with ARMA or ARTFIMA disturbance processes. We demonstrate this MCMC algorithm on a simulated data set in Section 2.4.1 and for the real data examples in Section 2.5.

2.4 Simulation study

For all examples and applications, stationarity is enforced through the prior. The auto-regressive parameters ϕ_i for $i = 1, \ldots, p$ are reparameterized according to Barndorff-Nielsen and Schou (1973) in terms of *partial auto-correlations* $\widehat{\boldsymbol{\phi}}_p = (\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$ and enforcing $|\widehat{\phi}_i| < 1$ for $i = 1, \ldots, p$ ensures stationary. The same reparameterization is used for the moving average parameters to ensure invertibility.

The regression parameters $\boldsymbol{\beta}$ have uninformative priors, $\boldsymbol{\beta} \sim \mathrm{N}(0, 100 \cdot I)$ where I is the identity matrix. Each of the lag polynomial parameters have independent uniform priors, e.g. $\widehat{\boldsymbol{\phi}}_p \sim \mathrm{Unif}(-1, 1)^p$. Furthermore, a log-transformation was used for σ_ε^2 with prior $\log(\sigma^2) \sim \mathrm{N}(0, 100)$.

2.4.1 Simulated data

As an illustrative example, we compare three different posteriors based on different likelihoods: the Whittle likelihood, the exact Gaussian likelihood and the Kalman filter likelihood, all of which are under the same non-informative prior for dynamic regression models on simulated data. By the exact Gaussian likelihood, we mean evaluating a multivariate normal density where the covariance matrix is computed from the autocovariance function. The Kalman filter likelihood, on the other hand, performs the Kalman filter recursions to compute the likelihood. Unlike the

Gaussian likelihood approach, the Kalman filter avoids inverting a large covariance matrix. Note that both these methods are exact.

Let $y_t = \beta_1 x_t + \eta_t$ where η_t is a ARMA(3,1) model with $n = 5001$. The true values of the parameter vector are set to $\beta = 3.0$, with the reparameterized vector $\boldsymbol{\theta} = (0.5,\ -0.248,\ 0.1,\ 0.2,\ 2)$ where the first two elements are the AR parameters $\boldsymbol{\phi}$, the next two elements are the MA parameters $\boldsymbol{\theta}_{\text{MA}}$ and the last element is the variance of the white noise of the process σ_ϵ^2. The exogenous predictor x_t is known and follows an ARMA(1,1) process. The MCMC algorithm described in Section 2.3.2 was used to sample the three posteriors for 10000 iterations with a burn-in of 3000 samples and the chain was thinned, keeping every other iteration. Table 2.1 reports the MSE of the posterior mean for 1000 replicates of the data for each parameter. As can be seen, the MSE of both the Whittle likelihood and the Gaussian likelihood are small. The differences between the MSE for each method are likely due to Monte Carlo error.

MSE	β_1	ϕ_1	ϕ_2	ϕ_3	θ_1	σ^2
Gaussian	0.013	0.076	0.007	0.003	0.072	0.002
Whittle	0.014	0.065	0.006	0.002	0.063	0.002

Table 2.1: The mean square error (MSE) for posterior means with the Gaussian likelihood and Whittle likelihood under the same prior for 1000 data simulation replicates.

Figure 2.2 plots kernel density estimates of the marginal posteriors for each parameter from one simulation. As seen in the aforementioned figure, the three posteriors are almost identical. Hence, the $n = 2500$ periodogram ordinates were enough to get a very close approximation of the Whittle likelihood to the true time domain likelihood.

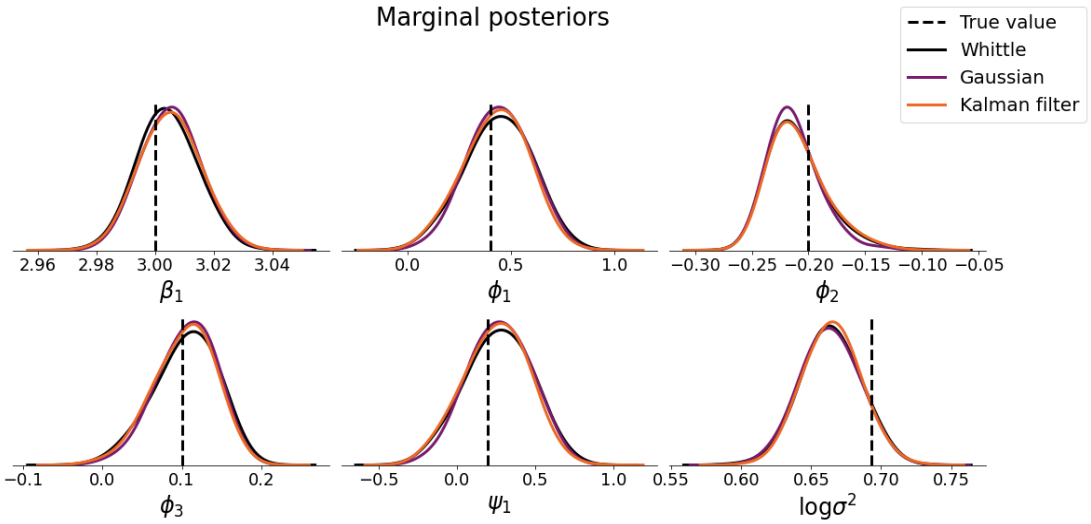


Figure 2.2: Kernel density estimates of the marginal Whittle, Kalman filter and Gaussian posteriors for a DLR model with $\eta_t \sim$ ARMA(3,1).

To measure the sampling efficiency of the resulting Markov chain from MCMC, we report the effective sample sizes for each parameter in Figure 2.3. It is shown that the Gaussian likelihood has a 25% increase in the ESS for the β_1 parameter compared to the Whittle and Kalman filter. The Whittle posterior has a lower ESS for $\log \sigma^2$ compared to its time-domain counterparts. The AR and MA parameters are comparable for all three posteriors, with the Whittle posterior having slightly higher ESS for ϕ_1, ϕ_3 and ψ_1.

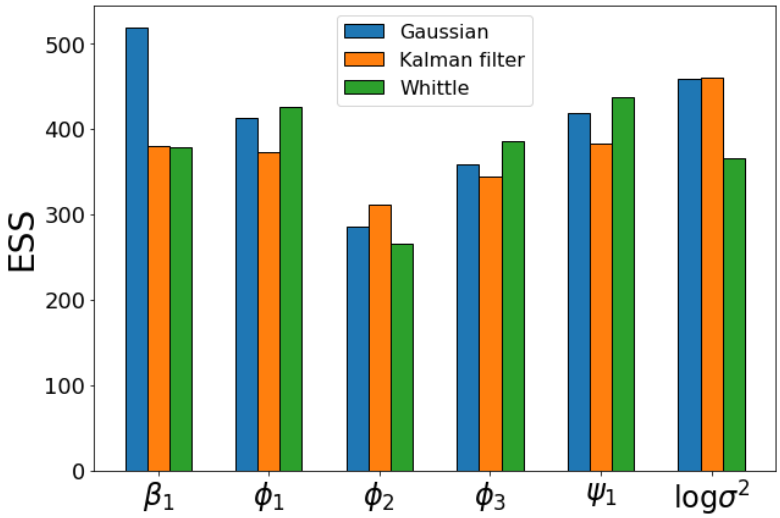


Figure 2.3: Effective sample sizes for 10,000 MCMC iterations for the three posteriors for each parameter.

The run times (in seconds) of each method are presented in Table 2.2. Our proposed Whittle approach is roughly 8 times faster than the Gaussian likelihood and 94 times faster than the Kalman filter. Here, a Python implementation is used for the Kalman filter. In contrast, the Gaussian likelihood uses the `SuperGauss` package in R (Ling and Lysy, 2017), which scales as $\mathcal{O}(T \log^2 T)$ from the superfast Toeplitz algorithm proposed in Ammar and Gragg (1988). Despite Kalman filtering being performed in $\mathcal{O}(T)$ operations, one must sweep through all observations at each iteration, whereas the Whittle likelihood in (2.16) is a sum, only requiring the computation of $I_Z(\omega)$ in (2.12) and the relevant spectral density, which takes much less operations than filtering.

Run time	Whittle	Gaussian	Kalman filter
total (s)	10.0	80.54	942.58

Table 2.2: Computation time of each likelihood method for 10000 MCMC iterations.

2.4.2 Periodogram simulations

In this section, we empirically verify that the ARTFIMA likelihood assumptions hold for small frequencies, which are violated for ARFIMA. The latter has been shown theoretically in Robinson (1995); Rousseau et al. (2012) and empirically in Meerschaert et al. (2014); Sabzikar et al. (2019). Both models are tested without the dynamic regression component. Restating Equation (2.14), the ratio of the periodogram and its spectrum has distribution,

$$\frac{I_Z(\omega_k)}{f(\omega_k)} \sim \frac{\chi^2(2)}{2}, \tag{2.22}$$

for $\omega_k \in \Omega$ as $T \to \infty$. We verify this asymptotic result via a simulation study. We consider two models

(M1) ARTFIMA$(2, d, \lambda, 0)$ with $\boldsymbol{\theta} = (\phi_1, \phi_2, d, \lambda, \sigma^2) = (0.742, 0.227, 2.139, 0.616, 1)$.

(M2) ARFIMA$(2, d, 0)$ with $\boldsymbol{\theta} = (\phi_1, \phi_2, d, \sigma^2) = (1.466, -0.525, 0.493, 1)$.

These parameter values for both models were chosen using the real data presented in Section 2.5.2.

We simulate 10000 realizations of both models for $T = 1001, 10001, 20001$. The three smallest positive Fourier frequencies ω_k (excluding zero) are chosen. Figure 2.4 displays the QQ plots of the simulated quantiles vs theoretical quantiles (2.22) for all three cases of T.

The first and second column of Figure 2.4 displays the QQ plots for the ARTFIMA and ARFIMA models, respectively. The first row corresponds to $T = 1001$, the second row is $T = 10001$, and the third is $T = 20001$. The first column shows that the ARTFIMA models' periodogram ratio is almost identical to the theoretical density for all T. The second column shows that regardless of T, the periodogram ratio of the ARFIMA model has greater dispersion than the theoretical density for all three frequencies. Thus, this is consistent with the previously mentioned findings in Rousseau et al. (2012); Robinson (1995).

The violation of (2.22) is partly attributed to the fact that the spectrum of the ARFIMA model behaves as a power law at the frequencies and diverges at the zeroth frequency. Sabzikar et al. (2019) and Meerschaert et al. (2014) show for applications such as hydrology, finance and climatology, the ARFIMA spectral density provides a poor fit of the periodogram at the low frequencies. This suggests the use of a low-frequency cutoff in the estimation of ARFIMA models in the frequency domain. However, by removing the low-frequencies of the periodogram, there would be a substantial loss of information pertaining to the long-memory parameter d. Moreover, we argue that this procedure is not automatic and would heavily depend on the parameter values of the model and the number of data points. Furthermore, the Kalman filter is not efficiently applicable for long-memory processes, as mentioned in Section 2.1. For these reasons, we use

the exact Gaussian likelihood for estimation and Bayesian inference for ARFIMA models, which incur an additional computational cost for large data we consider in Section 2.5.

2.5 Applications

This section analyses electricity demand with relevant covariates for two real-world datasets. Below we discuss how to choose and evaluate each model.

We consider three types of error processes: ARTFIMA, ARFIMA and ARMA. To control for the effect of the exogenous variables on y_t, all exogenous variables $\boldsymbol{x}_t$ are assumed to be known when forecasting. To choose the order of p and q, a search over the model space was considered up to a maximum of $p = 2, q = 1$. For various orders of p and q, we select the model with the smallest deviance information criterion (DIC) value (Spiegelhalter et al., 2002). Once the best model is chosen, we compare all three error processes with the same number of autoregressive and moving-average lags. We do this only for ARMA and ARTFIMA error processes; this is not done for ARFIMA since performing MCMC for each model under the Gaussian likelihood is intractable.

To evaluate forecasting performance, we perform time series leave-future-out-cross-validation. We use a sliding window approach with a fixed training size of T with $k = 100$ out-of-sample observations for testing. To compare models, use three different metrics: log-predictive density score (LPDS), root mean square error (RMSE) and continuous rank probability score (CRPS). We compute each metric for different forecast horizons, $h = 1, \ldots, 15$. Each time the window is rolled forward, we re-estimate the model and forecast h-steps-ahead. However, for ARFIMA, the model is only estimated once on the initial training set; this is because re-estimating the model is very costly due to the evaluation of the Gaussian likelihood. Approximate leave-future-out cross-validation techniques have been employed in Bürkner et al. (2020); however, this is only for the computation of the LPDS and cannot be used to sample the posterior, which will needed for the other metrics described below. Furthermore, due to the large amount of data, the sliding window approach only impacts the posterior negligably.

The h-step-ahead log-predictive density is given as

$$\begin{aligned}\log p(y_{T+h}|y_{1:T}) &= \log \int_{\Theta} p(y_{T+h}|\boldsymbol{\theta}, y_{1:T}) p(\boldsymbol{\theta}|y_{1:T}) d\boldsymbol{\theta}, \\ &\approx \log \left(\frac{1}{M} \sum_{m=1}^{M} p(y_{T+h}|\boldsymbol{\theta}^{(m)}, y_{1:T}) \right),\end{aligned}$$

where the last line is a Monte Carlo approximation thereof, with $\boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}|y_{1:T})$. The LPDS gives a measure of would well the model fits the out-of-sample data (Gelman et al., 2014). For the sliding window approach, the LPDS is re-computed from posterior samples based on the training

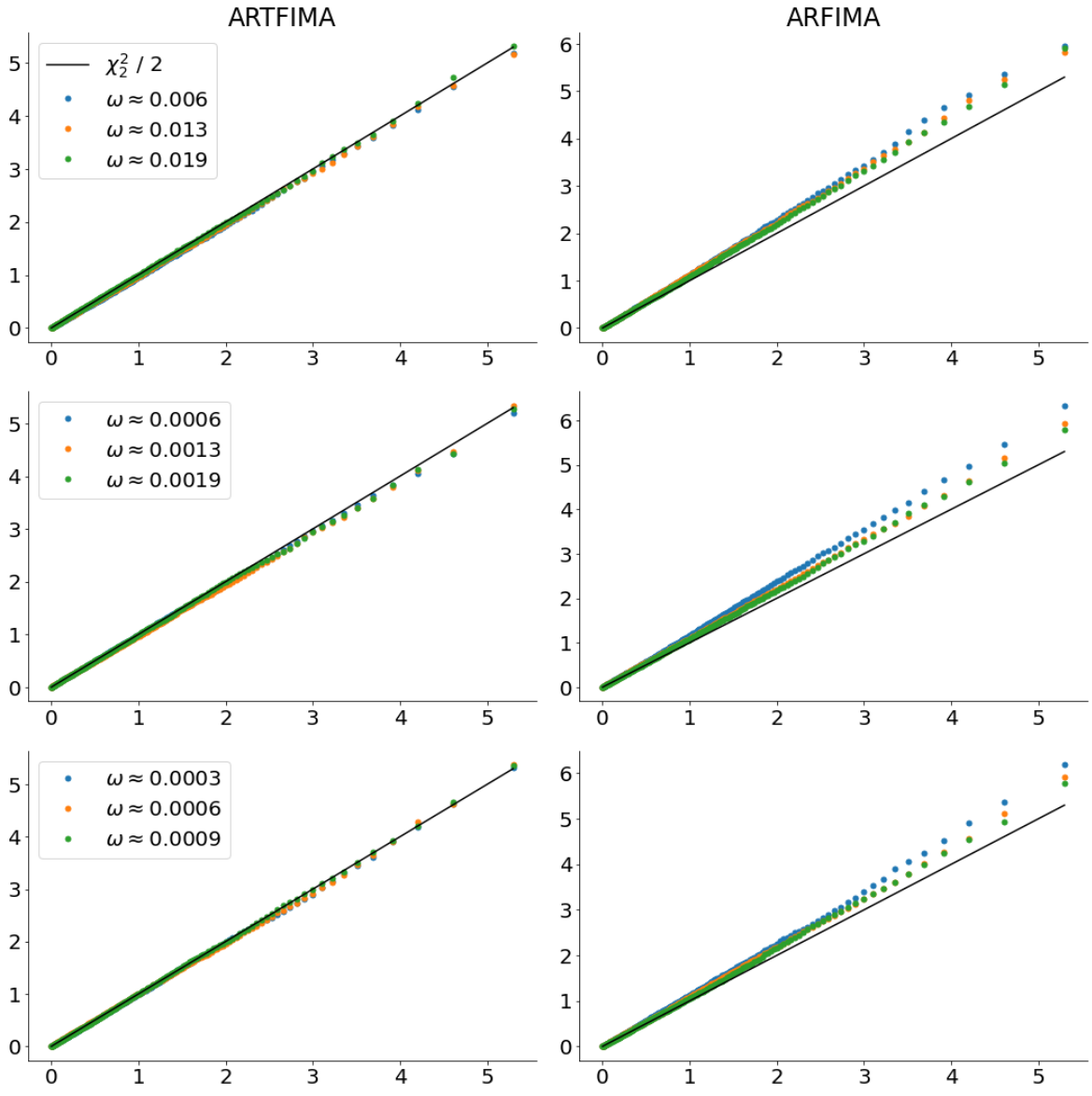


Figure 2.4: QQ plots of the ratio $I_Z(\omega_k)/f(\omega_k)$ for simulated ARTFIMA vs ARFIMA models for three lowest positive frequencies. The top row is $T = 1001$, the second row is $T = 10001$, and the third is $T = 20001$.

data that is rolled forward for all $T+1, \ldots, T+k$.

To assess point forecasts from a given model, define the conditional expectation as

$$\widehat{y}_{T+h} = \mathrm{E}[\widetilde{y}_{T+h}|y_{1:T}] = \int \widetilde{y}_{T+h} p(\widetilde{y}_{T+h}|y_{1:T}) d\widetilde{y}_{T+h},$$

where $p(\widetilde{y}_{T+h}|y_{1:T})$ is the posterior predictive distribution. Here, the posterior predictive is computed over the length of the testing set $\widetilde{y}_{T+i}$, $i = 1, \ldots, k$. To evaluate the performance of the point forecasts, the root mean square error (RMSE) is computed as

$$\mathrm{RMSE}^{(h)} = \sqrt{\frac{1}{k-h+1} \sum_{i=0}^{k-h} (y_{T+h+i} - \widehat{y}_{T+h+i})^2}.$$

for each forecast horizon h. The RMSE is the root of the average squared distance between the testing set and forecasted values, and hence, the lower the value, the better.

To assess distributional forecasts, we obtain the h-step-ahead predictive posterior density The distributional forecasts are assessed via the continuous rank probability score (CRPS), which is defined as

$$\mathrm{CRPS}_i^{(h)}(F, y_{T+h+i}) = \int_{\mathbb{R}} \Big(F(\widetilde{y}_{T+h+i}) - \mathbf{1}\{y_{T+h+i} \leq \widetilde{y}_{T+h+i}\} \Big)^2 d\widetilde{y}_{T+h+i},$$

where $\mathbf{1}$ is the indicator function, taking value one if $y_{T+h+i} \leq \widetilde{y}_{T+h+i}$ or zero elsewhere (Matheson and Winkler, 1976). When the distributional forecasts F and the data distribution are equal, the CRPS achieves its minimum. We use samples from the posterior predictive distribution to estimate the empirical cumulative distribution function $\widehat{F}(\widehat{y}_{T+h+i})$ for all, $i = 1, \ldots, k-h$. The mean of each $\mathrm{CRPS}_i^{(h)}$ is taken over all observations in the testing set, i.e. for $i = 1, \ldots, k-h$, to obtain $\mathrm{CRPS}^{(h)}$,

$$\mathrm{CRPS}^{(h)} = \frac{1}{k-h+1} \sum_{i=0}^{k-h} \mathrm{CRPS}_i^{(h)}(F, y_{T+h+i}).$$

To compute the RMSE and CRPS, we used 900 samples from the posterior predictive and 900 samples from the posterior to obtain the LPDS.

It is known that the log-likelihood function for the AR and MA parameters of the ARMA process can exhibit multimodality. To alleviate this, each model was fit with global optimization via Basin-hopping (Li and Scheraga, 1987). The same optimization procedure is used for the ARTFIMA and ARFIMA models.

We also look at the fitted spectrum to validate our model choice, as done in Sabzikar et al. (2019).

2.5.1 New England electricity demand

New England's electricity demand data consists of hourly electricity demand and temperature data for the states of Maine and Vermont from 2003 to 2017, with $T = 124200$ observations. The response, Maine electricity demand, y_t, is modelled as a linear combination of lagged covariates, Maine temperature $x_{1,t-1}$ and Vermont electricity demand $x_{2,t-1}$. Strong multi-seasonal and trend components exist within each variable. The `mstl` function in the R package `Forecast` (Hyndman and Khandakar, 2008) was used to remove these effects. Figure 2.5 depicts each variable after removing the multi-seasonal and trend components. As seen from the figure, each variable exhibits long memory, which makes it a suitable candidate for ARFIMA or ARTFIMA-type error terms.

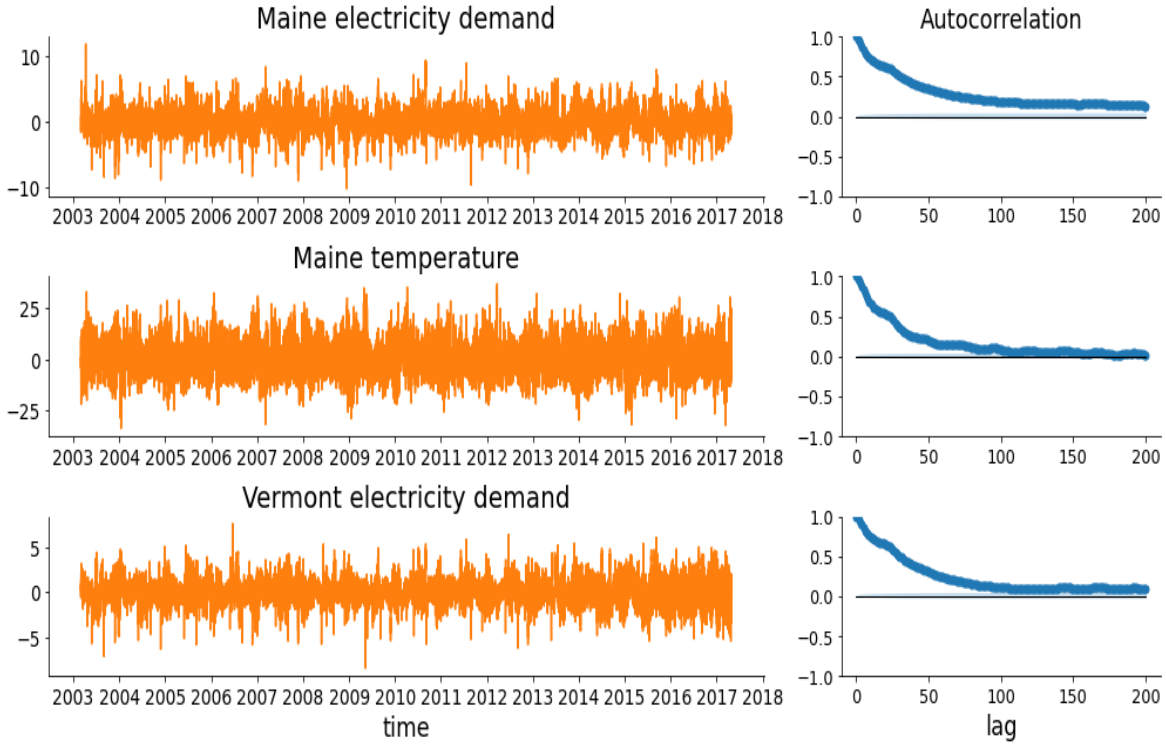


Figure 2.5: Maine and Vermont electricity demand (in megawatt) and temperature (degrees Celsius) time series data after removing multi-seasonal and trend components alongside their corresponding autocorrelation plots.

The lowest DIC for ARMA and ARTFIMA error processes was $p = 2$ and $q = 1$.Table 2.3 reports the DIC values for the three models as well as the average time for one evaluation of the log-likelihood function. As seen, ARTFIMA has the lowest DIC value, followed by ARFIMA, with ARMA being the worst. However, ARFIMA is the most computationally demanding model, which is roughly 40 times slower than the other two models since the Whittle likelihood cannot

be used. ARTFIMA and ARMA have comparable computational time per iteration of the log-likelihood of 11.5ms and 10.5ms, respectively. Note that the DIC values are positive in this example since the value of T is large.

	ARTFIMA	ARFIMA	ARMA
$p = 2, q = 1$	71987.56	72155.15	73030.91
Time (ms)	11.3	403.0	10.5

Table 2.3: DIC values and the average time for one log-likelihood evaluation for each dynamic regression model for New England electricity demand.

Figure 2.6 plots the negative LPDS, RMSE and CRPS for each model over the forecast horizon h. The negative LPDS favours models with lower scores, and thus, ARFIMA performs the best here after $h = 4$. Although hard to see by the plot, the negative LPDS for ARMA and ARTFIMA are very similar. The RMSE for all models are similar for small h; however, $h > 6$ shows a lower RMSE score for the ARMA and ARTFIMA models compared to ARFIMA. The CRPS is similar for all models, with no model out-performing the others for all forecast horizons.

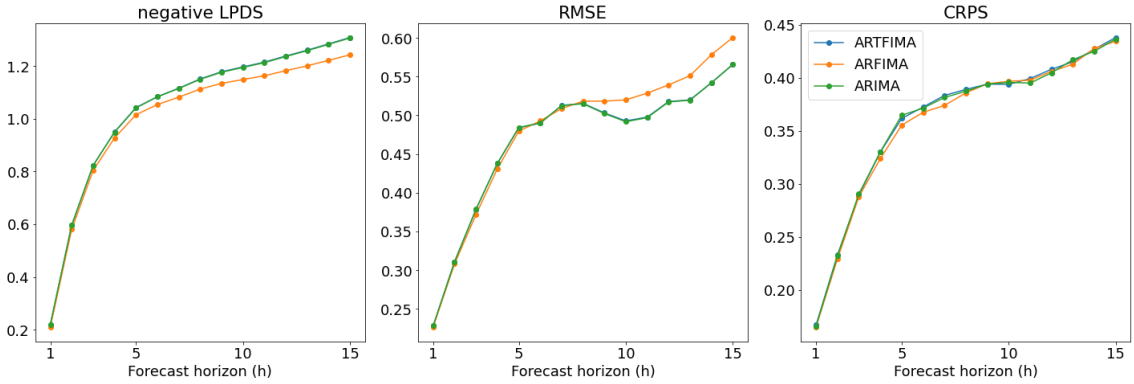


Figure 2.6: New England electricity data: negative log-predictive density score (LPDS), root mean square error (RMSE) and the continuous rank probability score (CRPS) for all models based on h-step ahead forecasts for $p = 2, q = 1$.

The fitted spectral densities of the dynamic regression models with ARTFIMA and ARFIMA errors are plotted alongside their corresponding periodograms in Figure 2.7, using a log-log scale. Due to the dependence on $\boldsymbol{\beta}$, the two periodograms are different for every frequency ω, but are almost identical at the lower frequencies. Consistent with the results of Sabzikar et al. (2019), the ARTFIMA spectrum follows a power law at moderate frequencies. Still, it flattens off at the lower frequencies, providing a better overall fit to its periodogram. The ARFIMA spectrum fits poorly at the lower frequencies due to its divergence as $\omega \to 0$. The 95% credible interval for each spectral density estimate is plotted. Since the periodogram depends on the values of $\boldsymbol{\beta}$, it is fixed at the posterior mean. The credible intervals are barely visible due to a large amount of

data, and thus, parameter uncertainty is small.

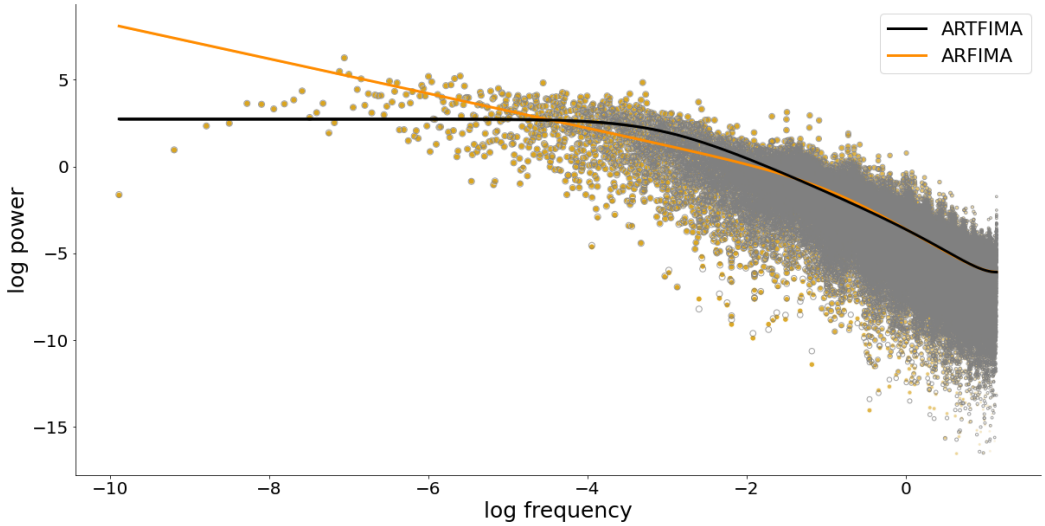


Figure 2.7: Spectral densities at the MAP and their respective periodograms. The spectrum of DLR with ARTFIMA$(2, d, \lambda, 1)$ errors (black line) with its periodogram (grey circles) and the DLR with ARFIMA$(2, d, 1)$ errors (orange line) with its corresponding periodogram (orange dots).

2.5.2 Victorian electricity demand

The second application we consider is the half-hourly electricity demand for Victoria, Australia. The data consists of $T = 52608$ observations of operation electricity demand (in megawatts) and Melbourne's temperature (degrees Celsius) between 1st January 2012 - 31st December 2014. The response y_t, Victorian electricity demand, is modelled as a linear combination of lagged covariates, Melbourne temperature x_{t-1}. Strong multi-seasonal and trend components exist within each variable and were removed via the `mstl` function. Figure 2.8 depicts each variable after removing the multi-seasonal and trend components.

Despite removing seasonality, as seen from Figure 2.8, all seasonality cannot be removed and still exists at multiples of the 48th lag. This is also present in the periodogram, see Figure 2.9, which peaks at the corresponding frequencies. We select the error processes in the same way as the previous example. We also include a seasonal ARMA (SARMA) and seasonal ARTFIMA (SARTFIMA) to account for the unexplained seasonality in y_t. One seasonal MA term of lag 48 for both seasonal models was included after an appropriate p, q was found.

The lowest DIC for the error processes was as follows: ARTFIMA$(2, d, \lambda, 0)$ and ARMA$(2, 1)$. Table 2.4 displays the DIC values and the average time for one log-likelihood evaluation for all models. Looking at the $p = 2, q = 0$ case, the lowest DIC (in bold) is the SARTFIMA model,

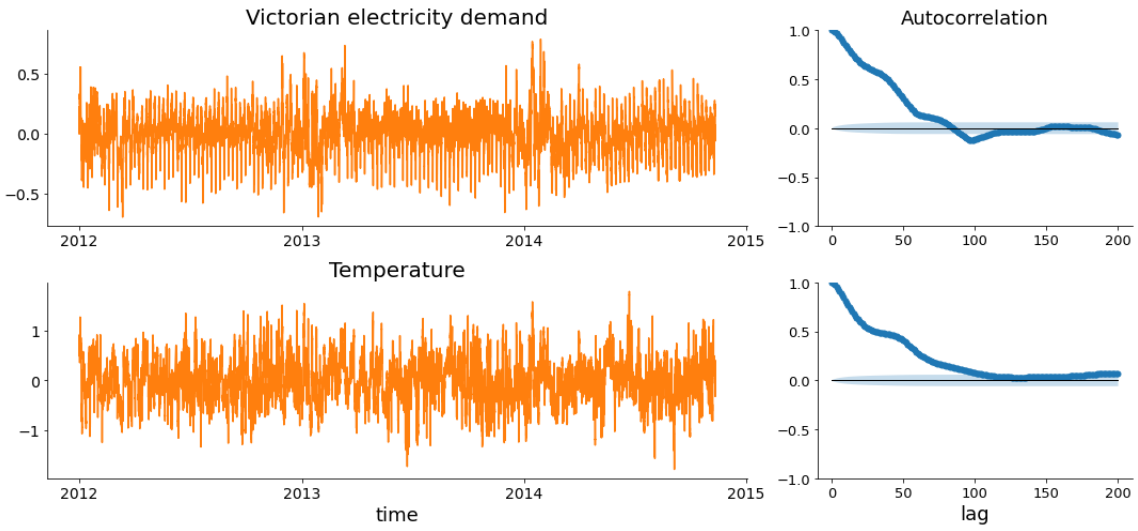


Figure 2.8: Victoria, Australia electricity demand (in megawatt) and temperature (degrees Celsius) time series data after removing multi-seasonal and trend components alongside their corresponding autocorrelation plots.

	SARTFIMA	ARTFIMA	ARFIMA	SARMA	ARMA
$p = 2, q = 0$	**−341016.47**	−328624.02	−328612.70	−338582.68	−327041.24
$p = 2, q = 1$	−276966.52	−328514.60	−328587.06	**−338646.40**	−328085.01
Time (ms)	35.1ms	3.96ms	203.0ms	35.1ms	3.79ms

Table 2.4: DIC values and the average time for one log-likelihood evaluation for each dynamic regression model for Victorian electricity demand.

and the worst (highest) was ARMA. Here ARTFIMA and ARFIMA have similar values for DIC. For the $p = 2, q = 1$ case, SARMA obtained the lowest DIC (in bold), with perhaps surprisingly, SARTFIMA being the highest DIC value. Again, ARTFIMA and ARFIMA produced similar DIC values.

Looking at the computation time of Table 2.4, significant computational speedups are gained for ARTFIMA and ARMA models, which had the shortest run time (ms) out of all models due to estimation via the Whittle likelihood. This is followed by the seasonal models, which are roughly 10 times slower than their non-seasonal counterparts. However, the run time for the ARFIMA model via the exact Gaussian log-likelihood was approximately 50 times slower than the non-seasonal ARMA and ARTFIMA models.

The ARTFIMA error process with a seasonal term gives the overall lowest DIC, followed by SARMA. This is echoed in Figure 2.9, which displays the log-log plot of the periodogram and spectral density at the MAP for the SARTFIMA model. It is apparent that the seasonal ARTFIMA model fits the lower frequencies, $\log(\omega) < -4$, and also models the seasonal compo-

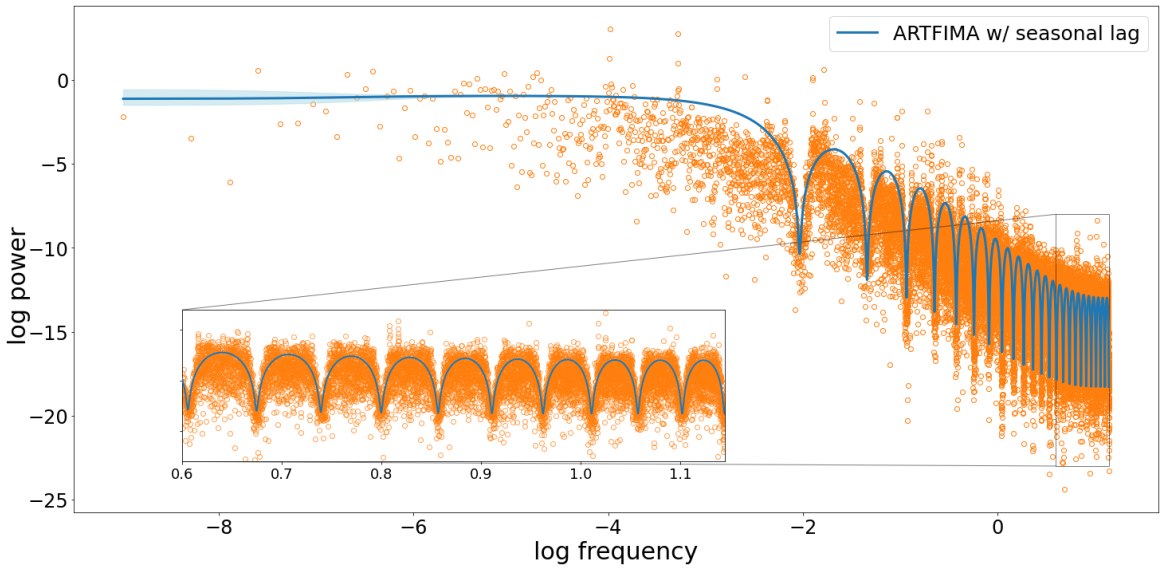


Figure 2.9: Outer plot: Log frequency vs log power (log-log) plot of the periodogram and the spectral density function at the MAP for DLR with SARTFIMA$(2, d, \lambda, 1)(0, 0, 1)_{48}$ errors. Inner plot: Same image but displaying moderate frequencies on a linear scale. The 95% credible interval of the spectral density estimated is the shaded blue region.

nents, shown in the zoomed-in section of the plot, as it captures the peaks and troughs of the periodogram appropriately.

Figure 2.10 reports the forecast negative LPDS, RMSE and CRPS of the five models for the $p = 2, q = 0$ case for all horizons $h = 1, \ldots, 15$. The seasonal models are better across all three metrics compared to the non-seasonal models. The seasonal ARTFIMA model has the best forecasting ability for all three metrics, followed closely by SARMA. In contrast, the ARMA model has the highest for all three metrics. The ARTFIMA and ARFIMA models were virtually identical, having lower negative LPDS, RMSE and CRPS scores compared to ARMA.

For the $p = 2, q = 1$ case, Figure 2.10 reports the forecast negative LPDS, RMSE and CRPS of the five models for all horizons $h = 1, \ldots, 15$. Surprisingly, the SARMA model has the highest (worst) negative LPDS out of all models for all h. The SARTFIMA is by far the best compared to all models for each metric. Similar to the $p = 2, q = 0$ case, the ARTFIMA and ARFIMA models are similar for each metric.

2.6 Conclusion and future research

This paper proposes dynamic regression models with ARTFIMA errors. Here, a response y_t is a linear combination of known exogenous predictors, which are stationary processes, plus a serially correlated error term with semi-long memory. We also propose a frequency domain estimation

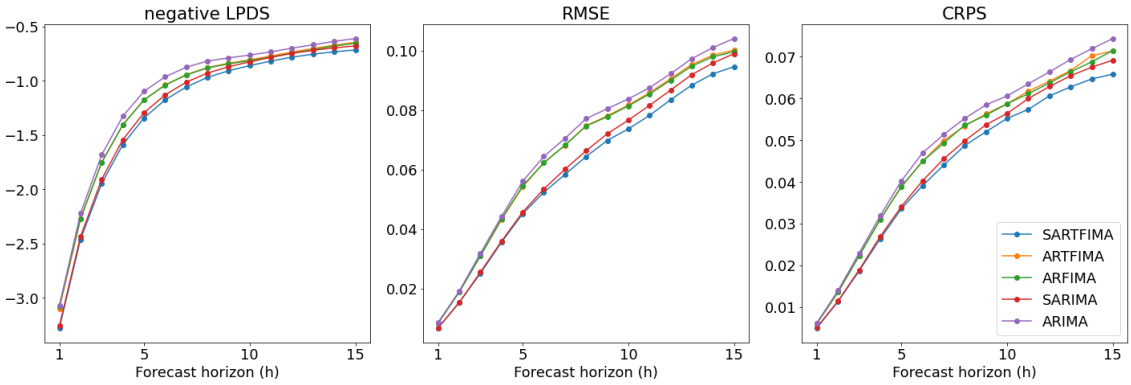


Figure 2.10: Victorian electricity data: negative log-predictive density score (LPDS), root mean square error (RMSE) and the continuous rank probability score (CRPS) for all models based on h-step ahead forecasts for $p = 2, q = 0$.

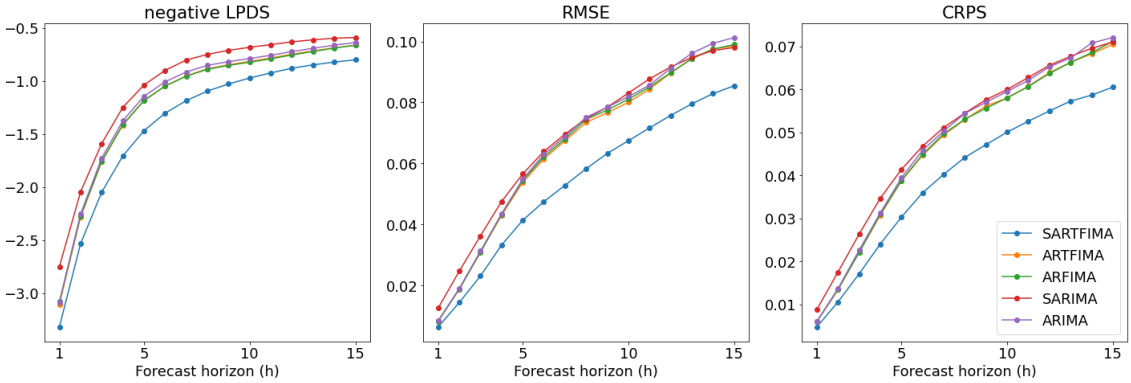


Figure 2.11: Victorian electricity data: negative log-predictive density score (LPDS), root mean square error (RMSE) and the continuous rank probability score (CRPS) for all models based on h-step ahead forecasts for $p = 2, q = 1$.

method that is asymptotically exact and utilizes the computationally efficient FFT. Furthermore, DLR with ARTFIMA errors can be shown to have improved forecasts compared to DLR with ARFIMA errors.

For computationally efficient estimation, as explained in Section 2.3, the linearity of the DFT is exploited to precompute the FFT of the response y_t and exogenous variables $\boldsymbol{x}_t$. This results in a procedure with $\mathcal{O}(T \log T)$ operations (for the FFT) once, before MCMC, then a subsequent cost of $\mathcal{O}(T)$ after that by computing the Whittle sum to evaluate the log-likelihood. The simulation study in Section 2.4.2 demonstrates that the Whittle approximation for ARFIMA models for low frequencies is not valid; thus, the exact Gaussian likelihood is used but at an additional cost of $\mathcal{O}(T \log^2 T)$ at each iteration.

Section 2.4 demonstrates our method on simulated data. Here, the error process is an

ARMA$(3,1)$ process with one exogenous predictor, given as an ARMA$(1,1)$ process. Figure 2.2 shows the Whittle, the Gaussian and the Kalman filter likelihoods are almost identical for this model with $T = 10001$. Also, the effective sample sizes are similar between the three likelihoods for all six parameters. Additionally, we demonstrate that the proposed Whittle approach requires less computation time than the Gaussian and Kalman filter likelihoods.

Finally, we illustrate our method on real-world data sets. The first is New England electricity demand data, and the second is Victorian electricity demand. In both cases, modelling the serially correlated error term as an ARTFIMA process gives on-par, if not better, root mean square error forecasts and continuous rank probability score probabilistic forecasts than existing ARMA and ARFIMA error models. Furthermore, we find substantial speed-ups for Whittle estimation of ARTFIMA models compared to ARFIMA models via the Gaussian likelihood.

Future research will extend frequency domain estimation methods of dynamic regression models to higher dimensions. Multi-dimensionality can be considered for any (or all) of the following: the response variable, the explanatory predictors, and the error process. For small to moderate data sets, exploring recent advances such as the debiased Whittle likelihood (Sykulski et al., 2019) for dynamic linear regression models can potentially provide less bias than the standard Whittle approach while still exploiting the computational efficiency of the FFT. Moreover, our estimation algorithm is amenable to data subsampling for ultra-long time series, in contrast to the time domain approach via the Kalman filter, which can provide further computational gains. Another interesting extension is the case when the regression coefficients are time-varying. In this case, an MCMC scheme would estimate the time-varying regression coefficients β_t, for $t = 1, \ldots, T$ as well as the parameters of the disturbance term η_t. Lastly, an exciting extension of this work would be frequency domain estimation techniques for more general transfer function models (Box et al., 2015).

References

Ammar, G. S. and Gragg, W. B. (1988). Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9(1):61–76.

Barndorff-Nielsen, O. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3(4):408–419.

Bisognin, C. and Lopes, S. (2009). Properties of seasonal long memory processes. *Mathematical and Computer Modelling*, 49(9):1837–1851.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, 90(14):2499–2523.

Chan, N. H. and Palma, W. (1998). State space modeling of long-memory processes. *The Annals of Statistics*, 26(2):719–740.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Doornik, J. A. and Ooms, M. (2003). Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics & Data Analysis*, 42(3):333–348.

Doornik, J. A. and Ooms, M. (2004). Inference and forecasting for ARFIMA models with an application to US and UK inflation. *Studies in Nonlinear Dynamics & Econometrics*, 8(2).

Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique*, 28:233–244.

Engle, R. F. (1974). Band spectrum regression. *International Economic Review*, 15(1):1–11.

Garthwaite, P. H., Fan, Y., and Sisson, S. A. (2016). Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *Communications in Statistics-Theory and Methods*, 45(17):5098–5111.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Granger, C. W. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105.

Haario, H., Saksman, E., and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223–242.

Hastings, W. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109.

Hosking, J. R. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27:1–22.

Kent, J. T. and Mardia, K. V. (1996). Spectral and circulant approximations to the likelihood for stationary Gaussian random fields. *Journal of Statistical Planning and Inference*, 50(3):379–394.

Levinson, N. (1946). The Wiener (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1-4):261–278.

Li, Z. and Scheraga, H. A. (1987). Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615.

Ling, Y. and Lysy, M. (2017). SuperGauss: Superfast Likelihood Inference for Stationary Gaussian Time Series. *R Package version*, 1.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

Matsuda, Y. and Yajima, Y. (2009). Fourier analysis of irregularly spaced data on R^d. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):191–217.

Meerschaert, M. M., Sabzikar, F., Phanikumar, M. S., and Zeleke, A. (2014). Tempered fractional time series model for turbulence in geophysical flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(9):P09023.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Pankratz, A. (2012). *Forecasting with Dynamic Regression Models*. John Wiley & Sons.

Peligrad, M. and Wu, W. B. (2010). Central limit theorem for Fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022.

Robinson, P. M. (1995). Log-periodogram regression of time series with long range dependence. *The Annals of Statistics*, 23(3):1048–1072.

Rousseau, J., Chopin, N., and Liseo, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory Gaussian process. *The Annals of Statistics*, 40(2):964 – 995.

Sabzikar, F., McLeod, A. I., and Meerschaert, M. M. (2019). Parameter estimation for ARTFIMA time series. *Journal of Statistical Planning and Inference*, 200:129–145.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Sykulski, A. M., Olhede, S. C., Guillaumin, A. P., Lilly, J. M., and Early, J. J. (2019). The debiased Whittle likelihood. *Biometrika*, 106(2):251–266.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434.

Chapter 3

Bayesian inference via spectral subsampling MCMC for continuous-time ARMA processes

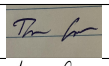
Thomas Goodwin, Matias Quiroz and Robert Kohn [1]

Abstract

Despite its rapid development in recent years, Bayesian inference via Markov chain Monte Carlo (MCMC) methods has traditionally been slow for ultra-large time series data. This is because the log-likelihood for time series data requires iterative methods such as Kalman filtering or particle filters for more complicated models. In this paper, we consider ultra-long, regularly sampled Lévy-driven continuous-time auto-regressive moving-average (CARMA) models. For stationary time series, the Whittle likelihood is an approximate frequency domain method based on the Fast Fourier Transform (FFT). It is appealing due to its computational efficiency of $\mathcal{O}(N \log N)$ where N is the number of data points. We first show that the bias incurred by the frequency domain approximation compared to the true log-likelihood is negligible for ultra-long, non-Gaussian CARMA models due to the central limit theorem for the DFT for stationary processes (Peligrad and Wu, 2010). Then, we consider Bayesian inference methodology for spectral subsampling for these aforementioned CARMA models. This method takes advantage of the approximate independence of the transformed process, which is amenable to subsampling MCMC based on Quiroz et al. (2019). We demonstrate that spectral subsampling produces speed-ups of up to two orders of magnitude compared to MCMC on the full dataset while also inducing negligible bias. We demonstrate our method for Bitcoin returns.

[1]Goodwin: *School of Mathematical and Physical Sciences, University of Technology Sydney.* Quiroz: *Department of Statistics, Stockholm University.* Kohn: *School of Economics, University of New South Wales.*

Keywords: Continuous-time models, subsampling, Bayesian inference.

Author	Project conception	Work	Manuscript writing	Manuscript editing	Signature
Tom Goodwin	X	X	X	X	
Matias Quiroz	X			X	
Robert Kohn				X	

Status of paper

This chapter is presented as a draft manuscript. Further research is necessary to add to the scope of the paper to deem it suitable for submission. I certify that the work in Chapter 3 has not been submitted as part of any other documents required for a degree.

3.1 Introduction

This chapter extends Bayesian inference methodology for spectral subsampling for regularly sampled Lévy-driven continuous-time auto-regressive moving-average (CARMA) models.

Time series analysis is an integral part of statistics and broader fields of science. Due to recent technological advancements in remote sensors, it is not uncommon to have time series data that contains hundreds of thousands or millions of observations. This introduces computational challenges for estimation techniques for time series, e.g. Kalman filtering, particle filtering, and the Gaussian likelihood. Filtering techniques require a full sweep of the observations sequentially for each log-likelihood evaluation. On the other hand, the Gaussian likelihood computes large systems for each evaluation, which is expensive for large data. This computation challenge is further exaggerated in Bayesian inference, as MCMC algorithms require a large number of iterations and, thus, many likelihood evaluations.

Observed data is a discretization of the actual data due to approximations from measuring instruments. Despite data being observed at discrete time intervals as discrete numeric values, many real-world phenomena have underlying continuous-time-generating processes. CARMA processes have been employed in many fields such as econometrics (Bergstrom, 1988; Brockwell et al., 2011), control theory (Gillberg and Ljung, 2009) and engineering (Mossberg and Larsson, 2004). An underlying continuous-time data-generating process modelled by a discrete-time model is an approximation, as discrete-time models tend to be more tractable. Despite this, continuous-time models are attractive when observing high-frequency sampled data such as a stock price or a biological system since they are a more faithful representation of the underlying process.

Despite modelling physical phenomena as a continuous-time process, the data is *sampled* at discrete times. This has two benefits. The first is that irregularly spaced data can be tedious for discrete-time models as these models assume regularly spaced intervals, and missing data can be, in some cases, non-trivial. The second benefit is that high-frequency sampled data are incorporated naturally into continuous-time models and variability above the Nyquist frequency can be handled appropriately via the aliased spectral density, whereas discrete-time models ignore the contribution of higher frequency components due to aliasing (Tómasson, 2015).

Another important aspect of continuous-time stochastic processes is the inclusion of non-Gaussian noise processes such as variance gamma and jump processes. Lévy-driven CARMA processes are generalizations of Gaussian CARMA processes with Lévy process noise terms introduced in Brockwell (2001). The inclusion of heavy-tailed or jump process extends the usefulness of CARMA processes to more complicated real-world phenomena.

Frequency domain estimation of CARMA processes has been studied in several recent papers. Authors Fasen and Fuchs (2013b) propose a consistent estimator of the spectral density for high-frequency, regularly sampled Lévy-driven CARMA processes based on the smoothed, normalized periodogram. Assuming an equidistant sampled CARMA processes with a symmetric α-stable Lévy driving process, Fasen and Fuchs (2013a) show the asymptotic distribution of the normalized periodogram at different frequencies converges to a function of a multivariate stable random vector. However, this is unsuitable for Whittle estimation, which relies on asymptotic normality of the Discrete Fourier Transform (DFT) (Peligrad and Wu, 2010). Fasen-Hartmann and Mayer (2022) consider Whittle estimation for multivariate Lévy-driven CARMA process with finite second moments sampled at low frequencies with the variance of the driving process fixed (not estimated). Gillberg and Ljung (2009) perform estimation of CARMA models via the Whittle likelihood for regularly sampled data. For irregularly spaced data, Fechner and Stelzer (2018) study the limit behaviour of a modified truncated Fourier transform for Lévy-driven CARMA processes.

Several popular time-domain quasi-likelihood estimators have been proposed in Schlemm and Stelzer (2012) and Brockwell et al. (2011); however, it is unclear how to perform Bayesian inference in the former. To our knowledge, only two papers have described in detail a Bayesian approach for CARMA processes. First, Kelly et al. (2014) explore MLE and MCMC methodology for regularly spaced Gaussian CARMA processes via Kalman recursions. Second, Sharifi et al. (2024) consider Bayes estimators for Whittle estimation of irregularly spaced Lévy-driven CARMA processes using theoretical results from Fechner and Stelzer (2018).

The contributions of this chapter are two-fold. First, we extend the methodology developed in Gillberg and Ljung (2009) to perform Bayesian inference for regularly spaced Lévy-driven CARMA models in the frequency domain via the Whittle likelihood for large data, and second, we demonstrate how to speed up MCMC via efficient spectral subsampling MCMC.

3.2 Lévy processes

Before formally describing the CARMA process, we present some basic facts about the Lévy process, which are the driving noise process behind Lévy-driven CARMA processes.

Definition 1: A stochastic process $\{L_t, t \geq 0\}$ with $L_0 = 0$ is said to be a Lévy process if it satisfies the following three properties:

- For any collection of time-points $0 \leq t_1 < t_2 < \cdots < t_n$, the random variables $(L_{t_2} - L_{t_1}), (L_{t_3} - L_{t_2}), \ldots, (L_{t_n} - L_{t_{n-1}})$ are all independent.
- The distribution of $L_t - L_s$ has the same distribution as L_{t-s} for any $0 \leq s < t < \infty$.
- For any $t \geq 0$ and $\epsilon > 0$, $\lim_{u \to 0} P(|L_{t+u} - L_t| > \epsilon) = 0$.

The first two points are the required independent and stationary increments, which intuitively describe how the change of the process L_t only depends on the length of the time integral considered. The third property is the continuity of probability, which bounds the probability of the absolute difference between observations L_t as the intervals shrink to 0, a probabilistic analogue to regular continuity. The celebrated Lévy-Khintchine theorem (Applebaum, 2009) states that the characteristic function of a Lévy process L_t is given by

$$\mathrm{E}\left[e^{-\mathrm{i}uL_t}\right] = e^{t\chi(u)}, \tag{3.1}$$

where

$$\chi(u) = \left(\mathrm{i}uk - \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}ux} - 1 - \frac{\mathrm{i}xu}{1+x^2}\right) \nu(dx)\right) \tag{3.2}$$

for $k \in \mathbb{R}$ and $\sigma \geq 0$ and ν is a Borel measure on $\mathbb{R}$, and we refer to Equation (3.2) as the Lévy exponential. The measure ν is known as the *Lévy measure* of the process L_t if it satisfies the following two conditions,

$$\nu(\{0\}) = 0, \quad \text{and} \quad \int_{\mathrm{R}} \min(1, x^2)\nu(dx) < \infty. \tag{3.3}$$

The Lévy-Khintchine theorem states that the process L_t can be decomposed into three independent components: a deterministic drift term, a Brownian motion and a Lévy jump process. For the case $\nu(\mathbb{R}) = 0$, the characteristic function simplifies to $\mathrm{i}uk - \frac{1}{2}\sigma^2 u^2$ which is recognizable as a Brownian motion with $\mathrm{E}[L_t] = kt$ and $\mathrm{Var}[L_t] = u^2 t$. Further analysis of the last term of the Lévy exponential in (3.2), pertaining to the Lévy jump component of L_t, can be split into two

parts,

$$\int_{\mathbb{R}\setminus\{0\}} \left(e^{\mathrm{i}ux} - 1 - \frac{\mathrm{i}xu}{1+x^2} \right) \nu(dx) = \int_{|x|\geq 1} (e^{iux} - 1)\nu(dx) + \int_{|x|<1} (e^{iux} - 1 - iux)\nu(dx).$$

This implies that the Lévy process L_t can be further decomposed into four parts, $L_t = L_t^{(1)} + L_t^{(2)} + L_t^{(3)} + L_t^{(4)}$. The first two terms are the deterministic and Brownian motion components, respectively. The third term, $L_t^{(3)}$, is a compound Poisson process. The fourth term, $L_t^{(4)}$, is a so-called *compensated* compound Poisson process; if N_t is a Poisson process with intensity λ, then compensated Poisson process is $\widetilde{N}_t = N_t - \lambda t$ with $\mathrm{E}[\widetilde{N}_t] = 0$. This decomposition is known as the Lévy-Itô decomposition, and for more information about Lévy processes, the reader is referred to Papapantoleon (2008) and Applebaum (2009). Following Brockwell and Marquardt (2005), we consider Lévy processes with the following properties for $t \geq 0$,

$$\mathrm{E}[L_1] < \infty, \quad \mathrm{E}[L_t] = \mu t \quad \text{and} \quad \mathrm{Var}[L_t] = \sigma^2 t.$$

Lévy processes can take the form of many known distribution functions by choosing an appropriate ν; examples include the Gamma process and the generalized inverse Gaussian (Barndorff-Nielsen and Shephard, 2001).

3.3 Model description

A Lévy-driven CARMA(p, q) model with $p > q$, is defined by a p-th order linear differential equation

$$\boldsymbol{\alpha}(D)Y_t = \boldsymbol{\beta}(D)DL_t, \quad t \geq 0, \tag{3.4}$$

where D is the differential operator with respect to t and $\{L_t\}$ is a Lévy process as described above with polynomials,

$$\alpha(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p,$$
$$\beta(z) = 1 + \beta_1 z + \cdots + \beta_q z^q,$$

where $\alpha(z)$ and $\beta(z)$ are the auto-regressive and moving average lag polynomials respectively. The definition from (3.4) cannot be analyzed directly since the derivatives DL_t exist nowhere due to the non-deterministic paths of the Lévy process. However, the CARMA process can be

defined via its state-space representation,

$$Y_t = \boldsymbol{\beta}^\top \boldsymbol{X}_t, \tag{3.5}$$

$$d\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_t dt + \boldsymbol{R} dL_t, \tag{3.6}$$

with quantities,

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_p & -\alpha_{p-1} & -\alpha_{p-2} & \cdots & -\alpha_1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 = 1 \\ \beta_1 \\ \vdots \\ \beta_q \\ \mathbf{0} \end{bmatrix}, \tag{3.7}$$

where $\mathbf{0}$ is a vector of $(p-q-1)$ zeroes. Equations (3.5) and (3.6) are the observations and state equations, respectively. The state equation is a multivariate linear stochastic differential equation with the solution,

$$\boldsymbol{X}_t = e^{\boldsymbol{A}(t-s)}\boldsymbol{X}_{\boldsymbol{s}} + \int_s^t e^{\boldsymbol{A}(t-u)}\boldsymbol{R} dL_u, \qquad \text{for } t > s \geq 0. \tag{3.8}$$

The solution in (3.8) is not generally stationary, Brockwell (2001) gives necessary and sufficient conditions for stationarity. Given L_t has independent increments, and the initial condition $\boldsymbol{X}_0$ is independent of $\{L_t, t \geq 0\}$, then $\boldsymbol{X}_t$ has a weakly stationary solution if and only if the eigenvalues $\lambda_1, \ldots, \lambda_p$ of matrix A all have negative real parts, i.e.,

$$\mathrm{Re}(\lambda_i) < 0, \quad i = 1, \ldots, p. \tag{3.9}$$

The stationary solution in (3.8) can be written as

$$\boldsymbol{X}_t = \int_{-\infty}^t e^{A(t-u)}\boldsymbol{R} dL_u \stackrel{d}{=} \int_0^{+\infty} e^{A(t-u)}\boldsymbol{R} dL_u, \tag{3.10}$$

where $\stackrel{d}{=}$ is equality in distribution. The mean and covariance of (3.10) are given as

$$\mathrm{E}[\boldsymbol{X}_t] = \frac{\mu}{\alpha_a}\boldsymbol{R}$$

$$\mathrm{Cov}[\boldsymbol{X}_{t+\delta}, \boldsymbol{X}_t] = \sigma^2 e^{\boldsymbol{A}\delta} \int_s^t e^{Au}\boldsymbol{R}\boldsymbol{R}^\top e^{A^\top u} du, \quad \delta \geq 0,$$

where $\mu = \mathrm{E}[L_1] = 0$ and $\sigma^2 = \mathrm{Var}[L_1]$ are the first and second moments of the driving process at time 1. The aforementioned covariance function of the transition equation is an important

quantity when computing the Kalman recursions when L_t is Brownian motion, as discussed later. Furthermore, the representation of Brockwell (2004), the CARMA process from Equation (3.5) can also be written as

$$Y_t = \int_{-\infty}^{t} \boldsymbol{\beta}^\top e^{A(t-u)} \boldsymbol{R} dL_u = \int_{\infty}^{\infty} g(t-u) dL_u, \quad -\infty < t < \infty,$$

where $g(t-u)$ is known as the kernel of the CARMA process defined as

$$g(u) = \frac{1}{2\pi} \int_{\infty}^{\infty} e^{iu\lambda} \frac{\beta(i\lambda)}{\alpha(i\lambda)} d\lambda. \tag{3.11}$$

The auto-covariance function of Y_t

$$\gamma(\tau) = \mathrm{Cov}(Y_{t+\tau}, Y_t) = \sigma^2 \int_{-\infty}^{\infty} \overline{g}(\tau - u) g(u) du, \tag{3.12}$$

where $\overline{g}(x) = g(-x)$. If the roots $\lambda_1, \ldots, \lambda_p$ satisfy Equation (3.9) and are distinct, the covariance can be expressed as

$$\gamma(\tau) = \sigma^2 \sum_{j=1}^{p} \frac{\beta(\lambda_j)\beta(-\lambda_j)}{\alpha(\lambda_j)\alpha(-\lambda_j)} e^{\lambda_j |\tau|}. \tag{3.13}$$

The spectral density of Y_t is a byproduct of Equation (3.12) via the convolution theorem, i.e.

$$f(\omega) = \int_{-\infty}^{\infty} \gamma(\tau) e^{\mathrm{i}\tau\omega} d\tau = \frac{\sigma^2}{2\pi} \left| \frac{\beta(\mathrm{i}\omega)}{\alpha(\mathrm{i}\omega)} \right|^2, \tag{3.14}$$

which is a rational function. The spectral density of the continuous-time processes is the Fourier transform of the continuous autocovariance function. However, the spectral density from (3.14) will be altered accordingly due to observing the continuous-time process at regularly sampled discrete times. This phenomenon, known as aliasing, will be discussed in detail in Section 3.5.1.

3.4 Aspects of CARMA models

Consider a Gaussian CAR(1) process,

$$dY_t + \alpha Y_t dt = dW_t, \tag{3.15}$$

where $\alpha \in \mathbb{R}$ and W_t is standard Brownian motion. Equation (3.15) is a stochastic differential equation (SDE), well-known as the Ornstein-Uhlenbeck (OU) process. The OU process has a

unique solution in terms of a stochastic integral (Oksendal, 2013),

$$Y_t = e^{\alpha t} Y_0 + e^{\alpha t} \int_0^t e^{-au} dW_u. \tag{3.16}$$

As mentioned briefly in the previous section, the paths of dW_t are nowhere differentiable, but the stochastic integral in (3.16) is well-defined as the limit of Riemann-Stieltjes sums (Karatzas and Shreve, 1988). Notice that differentiation of Y_t via the auto-regressive polynomial actually integrates the continuous process on the right-hand side of Equation (3.16) and, therefore, offsets the differential operator on the moving average side. Hence, to ensure the existence of a CARMA(p, q) process, the necessary condition $p > q$ must be satisfied (Stelzer, 2011).

Throughout this chapter, we only consider observed CARMA(p, q) process y_t at regularly sampled time points,

$$t, t + \delta, t + 2\delta, \ldots, t + (N - 1)\delta, \tag{3.17}$$

where $\delta > 0$, N is the total number of observations, and $T = \delta N$ is the number of periods/cycles. Unlike its discrete-time counterpart, sampling and, ultimately, estimation of continuous-time processes have two aspects: how many total cycles to observe the process T and how closely spaced the observations are δ, since both are functions of N. Kutoyants (2004) shows that if the whole path was observed (continuously sampled) of a CAR(1) process, the asymptotic MLE distribution of α is

$$\frac{1}{\sqrt{T}}(\widehat{\alpha} - \alpha) \sim N(0, 2a),$$

which makes it clear that the variance is a function of T. Hence, when $\delta \to 0$, the variability of $\widehat{\alpha}$ is explained by the number of cycles T. Furthermore, a simulation study in Tómasson (2015) shows that increasing the number of cycles increases the precision of the MLE estimates. In contrast, $\delta \to 0$ has only a marginal impact on the parameter estimates. This can be interpreted in a Bayesian setting as the posterior precision of the parameters depends heavily on the time length T, as opposed to the effect of $\delta \to 0$, which is minor. This is discussed further in Section 3.5.1.

3.5 Estimation

Numerous methods have been proposed in the literature to estimate the parameters of CARMA models; see Brockwell (2014) for a review. For Gaussian CARMA models, the Gaussian likelihood is the exact likelihood given in the log-scale as

$$\mathcal{L}_{\text{true}}(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\det\{\Sigma(\boldsymbol{\theta})\} - \frac{1}{2}Y_t^\top \Sigma^{-1}(\boldsymbol{\theta}) Y_t, \tag{3.18}$$

where $\Sigma(\boldsymbol{\theta})$ is the $N \times N$, where $N = T/\delta$, covariance matrix corresponding to the covariance function $\gamma(\tau)$ and $\det\{\Sigma(\boldsymbol{\theta})\}$ is the determinant of the covariance matrix (Kelly et al., 2014). Computing the last term in (3.18) requires solving the linear system requires $\mathcal{O}(N^3)$ operations for irregularly spaced data and $\mathcal{O}(N \log^2 N)$ for regularly spaced data (Ammar and Gragg, 1988). The exact likelihood can also be computed via the Kalman filter, which scales as $\mathcal{O}(N)$ and is more tractable for the data considered in this paper. However, the dependency between neighbouring observations implies the Kalman recursions must be computed in a sequential manner to evaluate the log-likelihood, and hence, in general, not suitable for the subsampling MCMC approach in Quiroz et al. (2019). Furthermore, for Lévy-driven CARMA processes, more elaborate techniques such as particle filters can be used but are more costly than the Kalman filter because the distributions of interest need to be computed using simulation, e.g. sequential Monte Carlo.

3.5.1 Frequency domain estimation

Suppose we have N regularly sampled demeaned observations from a Lévy-driven CARMA(p, q) process y_t^δ with spacing $\delta > 0$. Frequency domain estimation analyzes the DFT of the data via the Fast Fourier transform (FFT), and its resulting periodogram is defined as

$$J_\delta(\omega_k) = \sqrt{\frac{\delta}{N}} \sum_{n=0}^{N-1} y_t^\delta \, \exp(-\mathrm{i}\omega_k t), \tag{3.19}$$

$$\mathcal{I}_\delta(\omega_k) = \left| J_\delta(\omega_k) \right|^2, \tag{3.20}$$

respectively, with ω_k in the set of natural Fourier frequencies

$$\Omega \equiv \{2\pi k/(\delta N), \text{ for } k = 1, \ldots, \lfloor N/2 \rfloor\}. \tag{3.21}$$

For discrete-time models, such as ARMA models, the spacing between observations is assumed to be $\delta = 1$. In constant, for high frequency-sampled data, the Fourier frequencies in (3.21) usually extend further than the length of 2π since $\delta < 1$. The periodogram, or modification thereof, is sometimes called the empirical spectrum. Recall from (3.14) that the theoretical spectral density of the process is

$$f(\omega; \boldsymbol{\theta}) = \frac{\sigma^2}{2\pi} \left| \frac{\beta(\mathrm{i}\omega)}{\alpha(\mathrm{i}\omega)} \right|^2, \tag{3.22}$$

where $\boldsymbol{\theta}$ is the unknown parameters of interest, $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \sigma)$.

Regularly spaced sampling of a continuous-time process is defined up to the *Nyquist frequency* π/δ. Hence, there is no information about its underlying spectrum above the Nyquist frequency in the empirical spectrum. Furthermore, regular sampling gives rise to the aliasing phenomenon. Aliasing is when the spectral density is 'folded' into the density $f(\omega; \boldsymbol{\theta})$, for frequencies outside

$\pm\pi/\delta$. Each frequency has infinitely many aliases of itself at $2\pi k/\delta,\ k \in \mathbb{Z}$. The aliased spectral density is defined as,

$$f_\delta(\omega;\boldsymbol{\theta}) = \sum_{k=-\infty}^{\infty} f\big(\omega + \frac{2\pi k}{\delta};\boldsymbol{\theta}\big), \qquad \omega \in [-\pi/\delta, \pi/\delta]. \tag{3.23}$$

In general, the operation of 'folding' in (3.23) is not automatic and depends on the spectrum and its governing parameters. In practice, truncation of the aforementioned summation is performed to approximate the aliased spectrum, which generally does not have a closed-form solution (Sykulski et al., 2019). Ideally, the spectral density has negligible mass above the Nyquist frequency π/δ so that the information in the empirical spectrum captures a substantial amount of the total power of the process. This is why decreasing $\delta \to 0$ past a particular point yields a negligible impact on the precision of estimates (or variance of the posterior) since the very high-frequency components of the spectral density have insignificant mass past a particular frequency. Hence, for a fixed, small enough δ, we want the asymptotic condition $T \to \infty$.

The periodogram $\mathcal{I}_\delta(\omega)$ is an asymptotically unbiased estimate of the aliased spectral density $f_\delta(\omega)$ (Gelfand et al., 2010), namely,

$$\mathcal{I}_\delta(\omega_k) \sim \text{Exp}(f_\delta(\omega_k)), \quad \omega_k \in \Omega, \tag{3.24}$$

where $\text{Exp}(\cdot)$ is the exponential distribution parameterized by its mean. A key property of the periodogram is the asymptotic independence between periodogram ordinates at different frequencies. This gives rise to the Whittle likelihood (Whittle, 1953) for parameter estimation and Bayesian inference,

$$\mathcal{L}_{\text{W}}(\boldsymbol{\theta}) = -\sum_{k=1}^{(N-1)/2} \left(\log f_\delta(\omega_k;\boldsymbol{\theta}) + \frac{\mathcal{I}_\delta(\omega_k)}{f_\delta(\omega_k;\boldsymbol{\theta})}\right). \tag{3.25}$$

The Whittle likelihood is an asymptotic approximation to the exact log-likelihood in (3.18). Whittle's approximation via its maximum likelihood estimates has been studied in the frequentist setting in Contreras-Cristán et al. (2006). However, for Bayesian inference, the quality of approximation of the Whittle log-likelihood compared to the exact log-likelihood is of primary importance. Authors Guyon (1982) and Kent and Mardia (1996) give rates of asymptotic equivalence of the aforementioned log-likelihoods,

$$|\mathcal{L}_{\text{true}}(\boldsymbol{\theta}) - \mathcal{L}_W(\boldsymbol{\theta})| = \mathcal{O}_p(1),$$

as $T \to \infty$ for a fixed δ. The Whittle likelihood is robust in large samples due to the fact the data-generating process does not have to be Gaussian; instead, only the real and imaginary

parts J_δ in (3.19) are asymptotically normal. This results from the central limit theorem for the discrete Fourier transform of stationary processes studied in Brillinger (2001) and Peligrad and Wu (2010). The assumption of asymptotic normality of the DFT is tested in a simulation study for Lévy-driven CARMA(p, q) processes in Section 3.8

3.6 Spectral subsampling MCMC

Subsampling MCMC is based on the pseudo-marginal MCMC framework of Andrieu and Roberts (2009). Here, an estimator of the likelihood is used in place of the exact likelihood within a Metropolis-Hasting algorithm. To perform subsampling MCMC, we use the methodology developed from Quiroz et al. (2019).

Let $\pi(\boldsymbol{\theta}) \propto L_n(\boldsymbol{\theta})p(\boldsymbol{\theta})$ denote the posterior distribution of the model parameters $\boldsymbol{\theta}$ with a likelihood function $L_n(\boldsymbol{\theta})$ based on n samples, and prior distribution $p(\boldsymbol{\theta})$. The Metropolis-Hastings algorithm samples iteratively from $\pi(\boldsymbol{\theta})$ by proposing a parameter vector $\boldsymbol{\theta}^{(j)}$ at the jth iteration from some proposal distribution $g(\cdot|\cdot)$ and the probability of acceptance is

$$\min\left\{1, \frac{L_n(\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)})}{L_n(\boldsymbol{\theta}^{(j-1)})p(\boldsymbol{\theta}^{(j-1)})} \cdot \frac{g(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^{(j)})}{g(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j-1)})}\right\}. \tag{3.26}$$

Evaluation of the likelihood $L_n(\boldsymbol{\theta})$ can be costly for large n at each iteration. Quiroz et al. (2019) propose replacing $L_n(\boldsymbol{\theta})$ with an estimator $\widehat{L}(\boldsymbol{\theta}, \mathbf{u})$ based on a small random subsample of $m \ll n$ observations, where $\mathbf{u} = (u_1, \ldots, u_m)$ are the indices of the selected data.

The pseudo-marginal approach samples from $\boldsymbol{\theta}$ and $\mathbf{u}$ jointly from an augmented target distribution $\tilde{\pi}(\boldsymbol{\theta}, \boldsymbol{u})$. Andrieu and Roberts (2009) prove for an unbiased estimator, i.e. $\mathbb{E}_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}, \boldsymbol{u}) = L_n(\boldsymbol{\theta})$, the pseudo-marginal MCMC algorithm samples from the true posterior $\pi(\boldsymbol{\theta})$. An unbiased estimator of the log-likelihood $\widehat{\ell}(\boldsymbol{\theta}, \boldsymbol{u})$ and subsequently debiased $\exp(\widehat{\ell}(\boldsymbol{\theta}, \boldsymbol{u}))$ is used to estimate the full data likelihood. The debiasing approach does not eliminate all bias, and the pseudo-marginal sampler targets a slightly perturbed posterior, which is shown to be within $O(n^{-1}m^{-2})$ distance in total variation norm to the true posterior. In applications, Quiroz et al. (2019), Dang et al. (2019), Salomone et al. (2020) and Villani et al. (2022) show negligible bias. The so-called *difference estimator*

$$\widehat{\ell}_{\text{diff}}(\boldsymbol{\theta}) = \sum_{k=1}^{n} q_k(\boldsymbol{\theta}) + \frac{n}{m}\sum_{i=1}^{m}\left(\ell_{u_i}(\boldsymbol{\theta}) - q_{u_i}(\boldsymbol{\theta})\right), \tag{3.27}$$

with control variates $q_k(\boldsymbol{\theta})$ and indices $u_1, \ldots, u_m \sim \text{Uni}(\{1, \ldots, m\})$ is an unbiased estimate the log-likelihood based on a random subsample of m observations. Pseudo-marginal MCMC, and by extension, subsampling MCMC, requires a small variance of the likelihood estimator; otherwise, the resulting Markov chain can get stuck. The variance of (3.27) is reduced when

$q_k(\boldsymbol{\theta})$ approximates $\ell_k(\boldsymbol{\theta})$ well. An approach of Bardenet et al. (2017) constructs a second order Taylor expansion of $\ell_k(\boldsymbol{\theta})$ around some central value $\boldsymbol{\theta}^*$, defining the control variates as

$$q_k(\boldsymbol{\theta}) = \ell_k(\boldsymbol{\theta}^*) + \nabla_{\boldsymbol{\theta}}\ell_k(\boldsymbol{\theta}^*)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top\nabla^2_{\boldsymbol{\theta}}\ell_k(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \tag{3.28}$$

For complex models and large parameter spaces, the quadratic assumption of the control variates in (3.28) may give a poor approximation of the individual log-density $\ell_k(\boldsymbol{\theta})$ terms. To alleviate this, Salomone et al. (2020) propose a grouped quadratic control variate where the observations are divided into groups and the log-likelihood contribution for each group is approximated by a quadratic Taylor expansion. Tamaki (2008) proves the Bernstein-von Mises theorem (asymptotic normality of the posterior) via the Whittle likelihood, which suggests the Whittle log-likelihood is approximately quadratic for groups with large enough observation numbers. Spectral subsampling MCMC via the Whittle likelihood has been employed in Salomone et al. (2020) for discrete-time auto-regressive tempered fractionally integrated moving average (ARTFIMA) models and in Villani et al. (2022) for vector ARTFIMA models.

3.6.1 Kalman filter

We review the topic of time domain parameter estimation of Gaussian CARMA(p, q) models via the Kalman filter from Iacus and Mercuri (2015) and Tómasson (2015). The regularly spaced observations y_t^δ have the state space representation:

$$y_t^\delta = \boldsymbol{\beta}^\top\mathbf{x}_t^\delta \quad \text{and} \quad X_t^\delta = e^{A\Delta}\mathbf{x}_{t-1}^\delta + M_t^\delta, \tag{3.29}$$

where M^δ is a sequence of iid random vectors from

$$M_t^\delta = \int_{(t-1)\delta}^{t\delta} e^{A(t\delta-u)}\boldsymbol{R}dW_u, \quad t \in \mathbb{N}^+,$$

which has zero mean and covariance matrix

$$Q = \int_0^\delta e^{Au}\boldsymbol{R}\boldsymbol{R}^\top e^{A^\top u}du, \tag{3.30}$$

which can be computed using techniques from linear algebra. First, we define the unconditional covariance matrix Q_∞ of (3.10) which satisfies the continuous-time *Lypunov* or *Riccati* equations

$$AQ_\infty + Q_\infty A = -\sigma^2\boldsymbol{R}\boldsymbol{R}^\top, \tag{3.31}$$

then Q is obtained via

$$Q = Q_\infty - e^{\boldsymbol{A}u} Q_\infty e^{\boldsymbol{A}^\top u}. \tag{3.32}$$

Tsai and Chan (2000) give an efficient method to compute Equation (3.31); however, Q_∞ is a $p \times p$ matrix and needs to be computed for each distinct sampling interval, i.e. once for regularly spaced data. Furthermore, $e^{\boldsymbol{A}}$ is understood to be the matrix exponential

$$e^{\boldsymbol{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{A}^k, \tag{3.33}$$

which is a power series with $\boldsymbol{A}^0 = I$. See Moler and Van Loan (1978) for a comprehensive guide for the computation of matrix exponential.

Kalman filtering for parameter estimation consists of three steps: prediction, updating, and evaluation (Iacus and Mercuri, 2015). Filtering gives estimates of innovation residuals and their corresponding variances, which are used to evaluate the Gaussian log-likelihood. For ease of notation, we define preliminaries,

$$\mathbf{x}^\delta_{t|t-1} = \mathrm{E}[\mathbf{x}^\delta_t | \mathscr{F}_{t-1}] \tag{3.34}$$

$$\boldsymbol{\Sigma}^\delta_{t|t-1} = \mathrm{Var}[M^\delta_t | \mathscr{F}_{t-1}], \tag{3.35}$$

where $\mathscr{F}_{t-1}$ is the σ-algebra generated from the observations y_t and the estimates of the state space variables up to time N. Intuitively, $\mathbf{x}^\delta_{t|t-1}$ and $\boldsymbol{\Sigma}^\delta_{t|t-1}$ are the estimates of $\boldsymbol{X}$ and the variance of the state innovation process M^δ_t, respectively, given observation and estimates up to time N.

First, at $t = 1$ initialize $\mathbf{x}^\delta_{t-1|t-1} = 0$ and $\boldsymbol{\Sigma}^\delta_{t-1|t-1} = Q_\infty$. The prediction step is performed carrying forward from the state solution equation without the noise term in (3.8) to predict the unobservable process $\mathbf{x}^\delta_{t|t-1}$ and its covariance matrix:

$$\mathbf{x}^\delta_{t|t-1} = e^{\boldsymbol{A}\delta} \mathbf{x}^\delta_{t-1|t-1} \tag{3.36}$$

$$\boldsymbol{\Sigma}^\delta_{t|t-1} = e^{\boldsymbol{A}\delta} \boldsymbol{\Sigma}^\delta_{t-1|t-1} e^{\boldsymbol{A}^\top \delta} + Q. \tag{3.37}$$

By applying the observation Equation (3.5) to the aforementioned predicted state forecasts, the observable process is simply:

$$\widehat{y}_{t|t-1} = \boldsymbol{\beta}^\top \mathbf{x}^\delta_{t|t-1},$$

which gives the error/residuals term and its corresponding distribution,

$$\varepsilon_t = y^\delta_t - \widehat{y}_{t|t-1}, \qquad \varepsilon \sim N(0, \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^\delta_{t|t-1} \boldsymbol{\beta}). \tag{3.38}$$

Finally, to update the forecasted observable process with the current observation, the posterior update step is,

$$\mathbf{x}_{t|t}^{\delta} = \mathbf{x}_{t|t-1}^{\delta} + \boldsymbol{K}_t(y_t^{\delta} - \widehat{y}_{t|t-1}) \tag{3.39}$$

$$\boldsymbol{\Sigma}_{t|t}^{\delta} = \boldsymbol{\Sigma}_{t|t-1}^{\delta} - \boldsymbol{K}_t\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{t|t-1}^{\delta}. \tag{3.40}$$

where the Kalman gain matrix is defined by,

$$\boldsymbol{K}_t = \boldsymbol{\Sigma}_{t|t-1}^{\delta}\boldsymbol{\beta}(\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{t|t-1}^{\delta}\boldsymbol{\beta})^{-1}. \tag{3.41}$$

Intuitively, this is the weighting between the observation and the estimate $\boldsymbol{K}_t$, i.e., it tells us how much to change the estimate by the observation y_t. This algorithm outputs the residuals ε_t for $t = 1, \ldots, N$ which is used to construct the log-likelihood from (3.38) as,

$$\mathcal{L}_{\mathrm{kf}}(\boldsymbol{\theta}) = -\frac{1}{2}\Big(N\ln(2\pi) + \sum_{t=1}^{N}\ln\big(\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{t|t-1}^{\delta}\boldsymbol{\beta}\big) + \sum_{t=1}^{N}\frac{\varepsilon_t^2}{\boldsymbol{\beta}^{\top}\boldsymbol{\Sigma}_{t|t-1}^{\delta}\boldsymbol{\beta}}\Big). \tag{3.42}$$

Note that Kalman filtering can also be performed with irregularly spaced data by the same procedure; however, the covariance matrix in (3.30) must be computed for different spacing between observations.

3.7 Enforcing stationarity

Stationarity is enforced through the prior by performing two transformations on the CAR(p) parameters. We follow the approach of Tómasson (2015) to enforce the stationarity of CARMA(p, q) processes. For estimation and Bayesian inference, the stationarity pertaining to (3.9) must be enforced at proposed values of $\boldsymbol{\alpha}_p = (\alpha_1, \ldots \alpha_p)$. A CAR($p$) process is stationary if the roots of

$$\boldsymbol{\alpha}(z) = z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p \tag{3.43}$$

have negative real parts. This is related to its discrete-time counterpart, the AR(p) process, which is stationary if the roots

$$\boldsymbol{\phi}(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \tag{3.44}$$

lie outside the unit circle. Belcher et al. (1994) constructs a method based on the Cayley-Hamilton transform, which converts stationary AR(p) parameters to stationary CAR(p) parameters. Widely used in control theory to convert between continuous and discrete time models, the Cayley-Hamilton transform converts points inside the unit circle to the left side of the complex

plane. Define the complex number z such that $|z| < 1$, then

$$s = -\lambda \frac{1-z}{1+z},$$

lies on the left side of the complex half-plane. Then consider the polynomial

$$k(s) = k_0 s^p + k_1 s^{p-1} + \cdots + k_{p-1} s + k^p = \sum_{i=0}^{p} \phi_i (1 - s/\lambda)^i (1 + s/\lambda)^{p-i}, \tag{3.45}$$

with $\phi_0 = 1$. Then the CAR(p) polynomial is $\alpha_i = k_i/k_0$. If $w_1, \ldots, w_p$ are the roots of the polynomial $z^p \boldsymbol{\phi}(1/z)$, then it is guaranteed that the roots of $\boldsymbol{\alpha}(z)$ lie in the left-half complex plane, thus making the CAR(p) parameters stationary. Here, we set the time-scaling parameter $\lambda = 1$.

Next, the auto-regressive parameters $\boldsymbol{\phi}_p = (\phi_1, \ldots, \phi_p)$ are reparameterized in the space of partial auto-correlations $\boldsymbol{\varphi}_p = (\varphi_1, \ldots, \varphi_p) \in (-1, 1)$ from Barndorff-Nielsen and Schou (1973). Putting these two transformations together, denote $\widetilde{\boldsymbol{\alpha}}_p = (\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_p) \in (-1, 1)^p$ as the reparameterized CAR parameters. The composite transformation

$$\pi : (-1, 1)^p \to \text{stationary region of } \boldsymbol{\alpha}_p, \tag{3.46}$$

first maps parameters $\widetilde{\boldsymbol{\alpha}}_p$ in the continuous space $(-1, 1)^p$ to the space of stationary AR parameters $\boldsymbol{\phi}_p$ then maps to the space of stationary CAR parameters $\boldsymbol{\alpha}_p$.

We use the prior $\widetilde{\boldsymbol{\alpha}}_p \sim \text{Unif}(-1, 1)^p$ to enforce stationarity by satisfying the condition $|\widetilde{\alpha}_i| < 1$ (Salomone et al., 2020). The same transformation π described above is applied to the moving average component $\widetilde{\boldsymbol{b}}_q$ to ensure invertibility of the CARMA(p, q) process with prior $\widetilde{\boldsymbol{\beta}}_q \sim \text{Unif}(-1, 1)^p$. The variance parameter is transformed with prior $\log(\sigma^2) \sim \mathcal{N}(0, 1)$.

3.8 Simulation study

The purpose of this section is to assess the normality and independence between the real and imaginary parts of Equation (3.19), i.e.

$$\begin{pmatrix} \text{Re}\{J_\delta(\omega)\} \\ \text{Im}\{J_\delta(\omega)\} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \pi f_\delta(\omega; \boldsymbol{\theta}) & 0 \\ 0 & \pi f_\delta(\omega; \boldsymbol{\theta}) \end{pmatrix} \right],$$

for large T. We also assess the theoretical distribution of the periodogram in (3.24). The independence between the DFT's real and imaginary parts and the periodogram's distribution are asymptotic results as $T \to \infty$. These assumptions underpin the Whittle log-likelihood in (3.25); hence, it is crucial to validate this assumption for simulated data. This study follows roughly the

simulation study conducted in Fechner and Stelzer (2018). We present three examples. The first is a Gaussian CARMA(p, q) process. The second, the background Lévy process is a standardized Gamma process from Graf (2009), and the third is a so-called 'two-sided' Poisson from Fechner and Stelzer (2018). To avoid confusion with the variance of the CARMA process and the variance of the underlying driving Lévy process, we assume a unit variance of the Lévy process such that $\text{Var}[L_t] = t$.

For all examples, we simulate the CARMA process on an interval $[0, T]$ for a specified δ via the Euler-Maryuama scheme for stochastic differential equations. For simulation, the process is generated on a 10x finer grid to mitigate the effects of simulation error and sample well above the Nyquist frequency. The initial values are $y_0 = 0,\ \boldsymbol{X}_0 = \mathbf{0}$. For each example, 2000 independent paths of the CARMA process are simulated and $J_\delta(\omega)$ is computed for different specified frequencies. Furthermore, the infinite sum of the aliased spectral density in (3.23) is truncated to $k = -5, \ldots, 5$ for all simulations. This ensures we have captured enough terms in the infinite sum that any outside this range has a negligible contribution. We inspect four frequencies $\omega_k \in \Omega$ for $k \in \{1, 10, 100, 1000\}$ in all examples. Figure 3.1 displays the QQ plots for each example. The first two columns of Figure 3.1 display the normal QQ plots, and the last column displays the standard exponential QQ plots.

Example 3.1. Consider a Gaussian CARMA$(2, 1)$ model with the form

$$y_t = \boldsymbol{\beta}^\top \boldsymbol{X}_t, \qquad d\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_t dt + \sigma^2 \boldsymbol{R} dW_t, \tag{3.47}$$

where dW_t is standard Brownian motion and

$$A = \begin{bmatrix} 0 & 1 \\ -\alpha_2 & -\alpha_1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 = 1 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{X_t} = \begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix}. \tag{3.48}$$

The true parameters are $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \sigma^2) = (1, 2, 1, 1)$ and we set $T = 100$ and $\delta = 0.01$. The top row of Figure 3.1 shows the associated QQ plots with the exact frequencies. For the real part, on the left panel, the QQ plots shows almost no difference from normality for all four frequencies. The imaginary part on the right panel looks normal except for minor deviation at the tails. The periodogram appears exponentially distributed, however there is slight deviation at the theoretical quantiles of the upper tail for $\omega = 0.69$ and $\omega = 6.35$.

Example 3.2. Consider a CARMA$(2, 0)$ process with a standardised Gamma driving process G_t with representation

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \boldsymbol{X}_t, \qquad d\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_t dt + \sigma^2 \boldsymbol{R} dG_t, \tag{3.49}$$

where $\boldsymbol{A}$, $\boldsymbol{X_t}$ and $\boldsymbol{R}$ defined the same as in (3.47). We use the same standardised Gamma process

defined in Graf (2009). The standardised Gamma process G_t has density according to

$$f_{G_t}(x) = \frac{\mu^{1/2\mu t}}{\Gamma(\mu t)} x^{\mu t - 1} e^{-x\mu^{1/2}}, \quad x \in (0, \infty), \tag{3.50}$$

which is the form of a Gamma density with first and second moments

$$\mathrm{E}[G_t] = \mu^{1/2} t \quad \mathrm{Var}[G_t] = t, \tag{3.51}$$

respectively. The increments are also gamma-distributed, with density

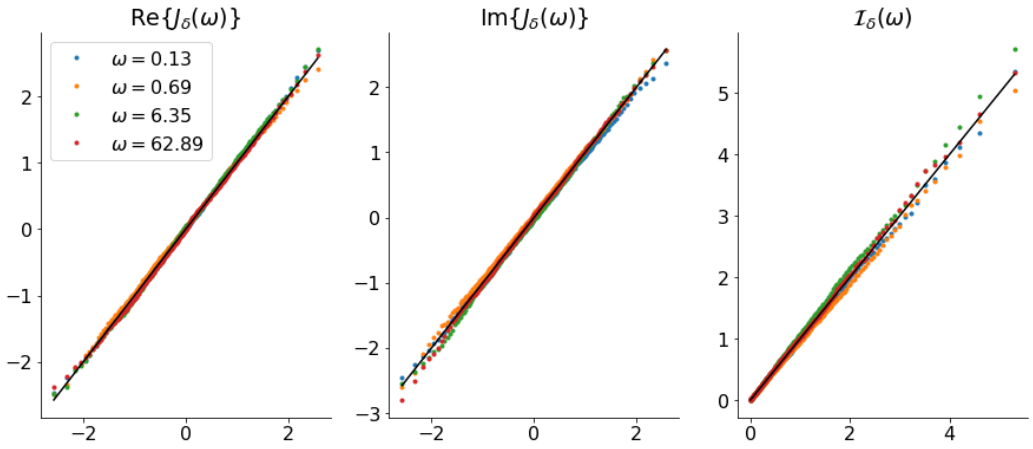
$$G_t - G_{t-1} \sim \text{Gamma}\left(a = \mu(t_n - t_{n-1}),\ b = \mu^{1/2}\right) \tag{3.52}$$

where a and b are the shape and inverse scale parameters, respectively. To satisfy the unit variance, each $\mu = 1$, and the true parameters are $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \sigma^2) = (1.5, 4, 1)$. This example is more challenging due to the skewness of the Gamma distribution; hence, we set $T = 1000$ and $\delta = 0.1$. The middle row of Figure 3.1 displays the corresponding QQ plots with the exact frequencies. The story is similar to Example 3.1; the real part (left) looks normal and the imaginary part (middle) looks indistinguishable from the standard normal distribution. For the periodogram, there is some departure from the theoretical quantiles in the tails (> 4). One possible explanation is the skewness in the gamma distribution implies a greater number of samples for the asymptotic normality to be fully satisfied.

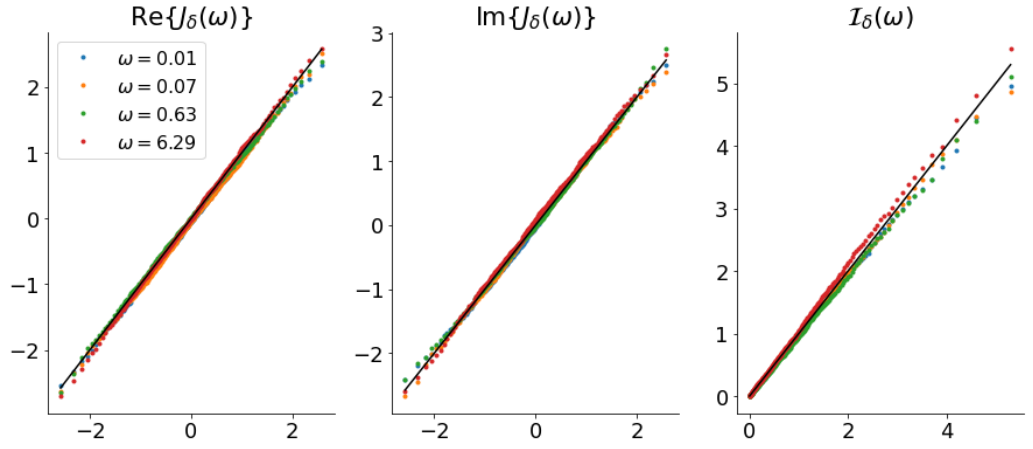
Example 3.3. In the final simulation, we consider a CARMA$(2, 1)$ with an underlying two-sided Poisson process (Fechner and Stelzer, 2018). The state space representation is defined the same as in (3.47). Let N_t be a a Poisson process for $t \geq 0$ with rate parameter $\lambda \in (0, \infty)$ satisfying

$$Pr(N_t = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x \in \mathbb{N}.$$

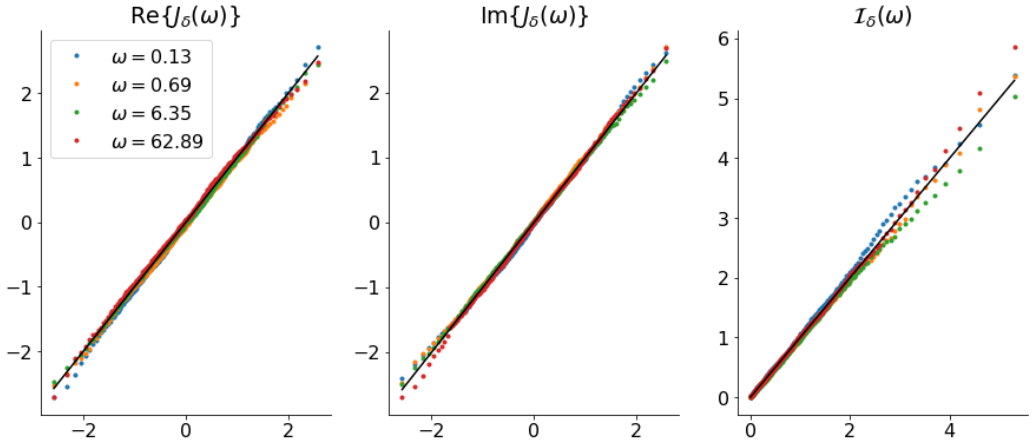
The two-sided Poisson process is defined as the difference of two independent Poisson processes, $V_t = N_{1,t} - N_{2,t}$, each with rate parameter $\lambda = \delta/2$. The two-sided Poisson process is equivalent to a compound Poisson process with rate δ and jumps ± 1 with equal probability. The true parameters are $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \sigma^2) = (1, 2, 1, 1)$. We set $T = 100$ and $\delta = 0.01$. The bottom row of Figure 3.1 shows the QQ plots. The real part (left) looks normal with minor deviation at the tails, and the imaginary part looks indistinguishable from the standard normal distribution. The periodogram plot on the right shows some departure from the theoretical quantiles in the tails (> 2). particularly for the lowest frequency and $\omega = 6.35$. This suggests that a greater number of samples is needed to obtain asymptotic independence between the real and imaginary parts of the DFT.



(a) $T = 100, \delta = 0.01$



(b) $T = 1000, \delta = 0.1$



(c) $T = 100, \delta = 0.01$

Figure 3.1: QQ plots of CARMA$(2, 1)$ process with Brownian motion driving process (top), standardised Gamma-driven CARMA$(2, 0)$ process (middle) and two-sided Poisson driving CARMA$(2, 1)$ process (bottom).

3.9 Applications

This section demonstrates Bayesian inference via MCMC and spectral subsampling MCMC on simulated and real data for large T. First, we present simulated data from the same models as the three examples in the previous section. Then, for the real data application, we consider Bitcoin prices.

For the simulated data, we compare the Whittle likelihood to the exact Kalman filter likelihood for the Gaussian CARMA(p, q) process. When the background Lévy process is the standardized Gamma process and the 'two-sided' Poisson, we compare the Whittle likelihood to the subsampled Whittle likelihood since the exact likelihood is difficult to compute. For MCMC, we compute 10000 iterations of a Random Walk Metropolis-Hasting for the Whittle and Kalman filter and 10000 iterations for PMMH for subsampling while discarding the first 3000 samples as burn-in for all MCMC runs. For subsampling, we used $G = 1000$ groups with an equal number of observations (periodogram ordinates). The tuning parameters T and δ are selected specifically for each example, but true parameters, settings and generation of the paths are the same as in the previous section.

To evaluate the computational efficiency gain in subsampling, we define a metric that incorporates the (saved) cost of estimating the likelihood and the inefficiency of the resulting MCMC samples. The computational time (CT) is

$$\text{CT} \equiv \text{IACT} \times \text{number density evaluations},$$

where $\text{IACT} \equiv 1 + 2\sum_{k=1}^{\infty} \rho_k$ is integrated autocorrelation time of the MCMC chain, where ρ_k is the autocorrelation at the kth lag. The IACT or inefficiency factor can be thought of as the number of correlated MCMC draws needed to obtain a single iid draw of the posterior. We estimate the IACT via the `TensorFlow` package from Abadi et al. (2015). The relative CT (RCT) metric is defined as the ratio of the CT of full-data Whittle MCMC and subsampling Whittle MCMC for each parameter. Values greater than one indicate that subsampling Whittle MCMC is more efficient when considering the computational cost of likelihood estimates and the inefficiency factor for MCMC. Figure 3.2 displays simulated realizations of the three models considered.

3.9.1 Simulated data

Example 3.4. Consider a Gaussian CARMA$(2, 1)$ model as in Example 3.1. We set $\delta = 0.1$ with $T = 8000$ and use and compare the Kalman filter posterior with the Whittle posterior and subsampled Whittle posterior. To reduce the bias incurred by the truncation of the aliased spectral density, it is truncated, we set $k = -50, \ldots, 50$. The same parameter values are used as in Example 3.1. Figure 3.3 displays the kernel density estimates of the marginal posteriors. The

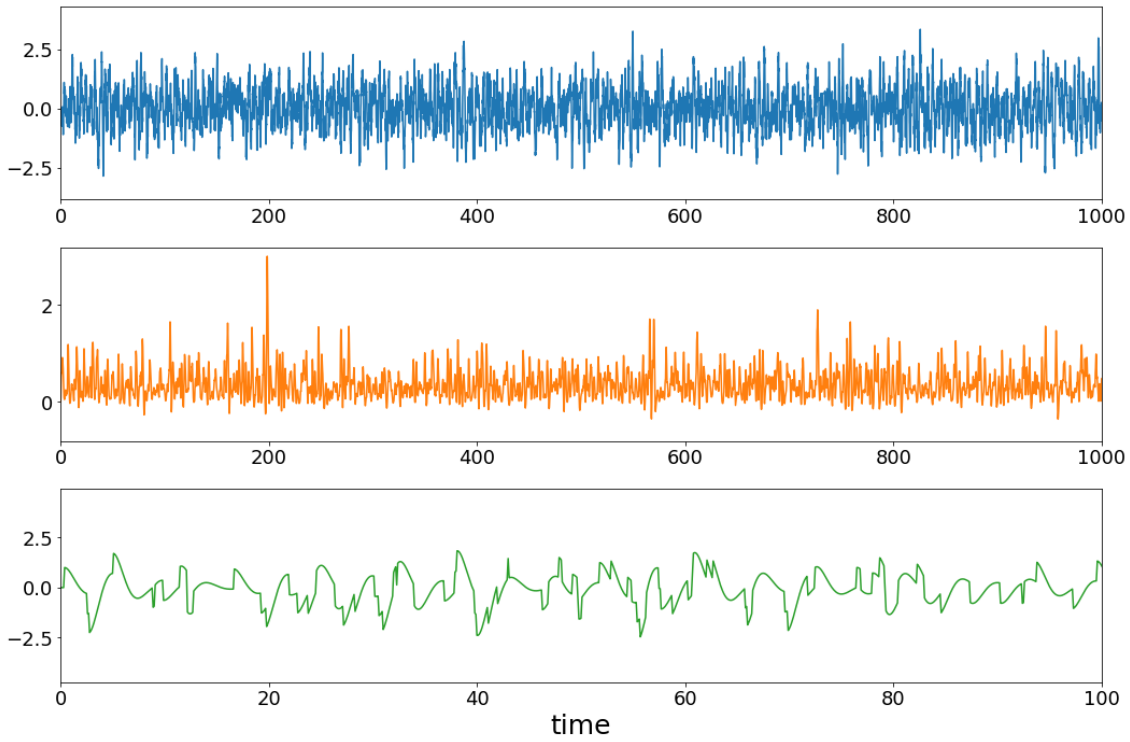


Figure 3.2: Simulated data from Examples 3.4, 3.5, and 3.6. From top to bottom, a Gaussian CARMA$(2,1)$, standardised Gamma CARMA$(2,0)$ and a two-sided Poisson CARMA$(2,1)$ process.

subsampled Whittle posterior (blue) and standard Whittle posterior (orange) are indistinguishable, suggesting that the bias incurred by subsampling is negligible. Furthermore, the Kalman filter (green) marginal posteriors are almost identical to their Whittle counterparts. The RCTs for each parameter is $(109, 94, 126, 97)$.

Example 3.5. Suppose we have observations from a CARMA$(2,0)$ with an underlying standardised Gamma process, as in Example 3.2. We set $T = 10000$, $\delta = 0.1$ and the same parameters values as Example 3.2, $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \sigma^2) = (1.5, 4, 1)$. Note, the Kalman filter is mis-specified in this model as this state space model is not Gaussian, and hence we do not use it. Figure 3.4, the variance parameter σ^2 is slightly overestimated; however, the standard Whittle posterior (black) and the subsampled Whittle posterior (red) are almost identical. The RCT for each parameter is (86, 99, 106).

Example 3.6. We consider the same two-sided Poisson-driven CARMA$(2,1)$ process in Example 3.3 with $T = 10000$, $\delta = 0.1$ and the same parameter values. Figure 3.5 compares the standard Whittle posterior (black) with the subsampled Whittle posterior (red). As seen in the plot, the

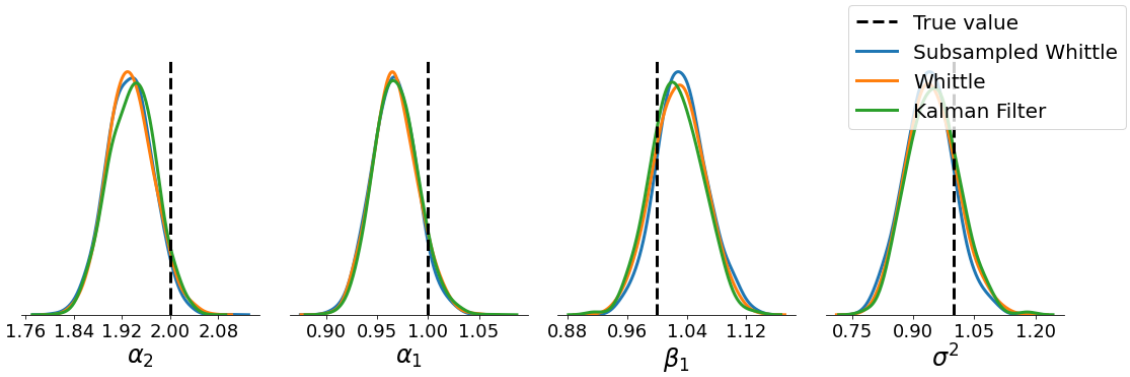


Figure 3.3: Kernel density estimates of the marginal posteriors. Gaussian CARMA$(2,1)$ process from Example 3.4, with $\delta = 0.1$ and $T = 8000$.

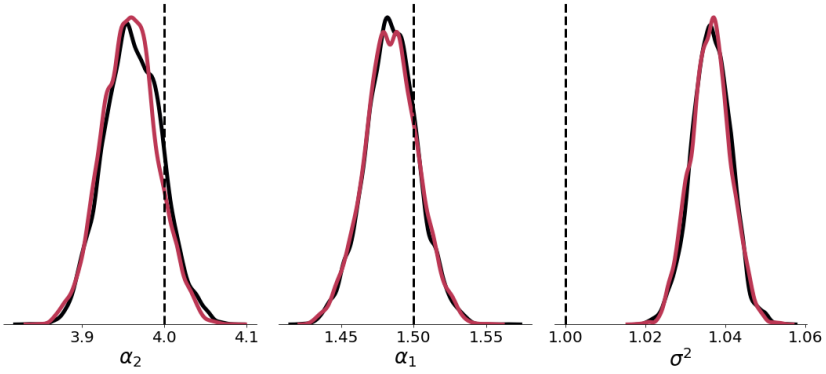


Figure 3.4: Marginal kernel density estimates of the posterior for standardised Gamma CARMA$(2,0)$ process from Example 3.5. The standard Whittle posterior is in black, and the subsampled Whittle is in red. The dashed vertical lines are the true parameter values.

marginal posteriors are almost identical with the subsampled Whittle posterior, which seems to have slightly thinner tails for α_1 and α_2. The RCT for each parameter $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \sigma^2) = (118, 173, 186, 174)$.

3.9.2 Bitcoin volatilities

Here, we consider the log-squared returns of minutely Bitcoin prices from `http://www.coinbase.com`. The total length of the series is $N = 1000001$, and we set $\delta = 1$. Here, the choice of $\delta = 1$ is due to the fact that the data were sampled at one-minute intervals as opposed to seconds/milliseconds and implies that the information contained in the empirical spectrum lies between $\pm\pi$.

For model selection, we fit all the valid models up to $p = 3$ and choose the one with the

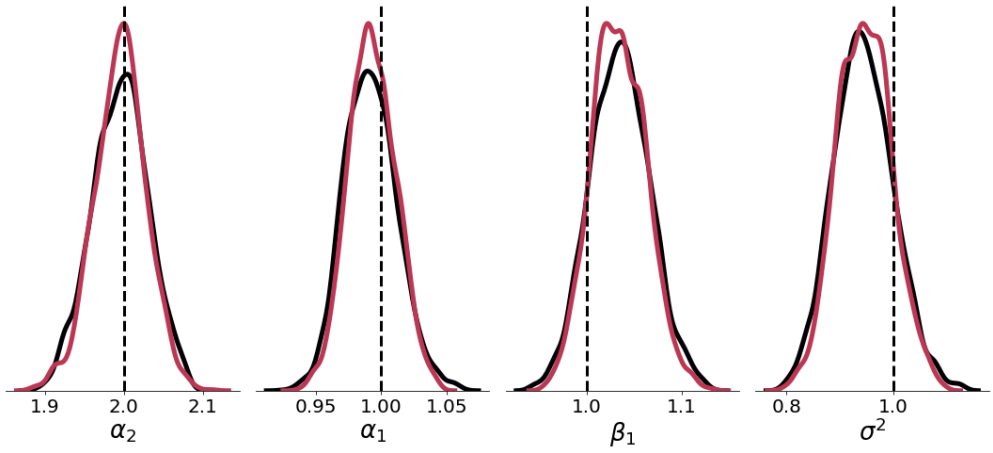


Figure 3.5: Example 3: Marginal kernel density estimates of the posterior for two-sided Poisson CARMA$(2, 1)$. The standard Whittle posterior is in black, and the subsampled Whittle is in red. The dashed vertical lines are the true parameter values.

smallest BIC. Table 3.1 shows the BIC values for all models with $p = 3$, $q = 2$ being the smallest. This is a challenging data problem due to the long memory present in the observations, which manifests itself linearly in the empirical spectrum at the lower frequencies as seen in Figure 3.6. As a result, the autoregressive parameters were close to one, hence close to the bound, making sampling difficult. Instead, we choose the next lowest BIC value, which is $p = 2, q = 1$.

CARMA	$q = 0$	$q = 1$	$q = 2$
$p = 1$	4392006.60	-	-
$p = 2$	124406137.04	4274959.12	-
$p = 3$	402002138.0	4308032.68	4271028.10

Table 3.1: BIC values for each CARMA model for Bitcoin data.

Figure 3.7 shows the posteriors of the standard Whittle posterior (black) and the subsampled Whittle posterior (blue), and there is virtually no difference between them. Figure 3.8 displays the RCT for each parameter of subsampling vs Whittle MCMC with the full dataset. The aforementioned plot indicates that Whittle subsampling is, on average, roughly 100 times more efficient than full-data MCMC. The total time in seconds for 10000 iterations of MCMC for the standard Whittle was 4925.673 seconds, whereas the subsampled Whittle was 61.045 seconds, more than an 80x increase in raw computation time. Furthermore, Figure 3.6 also displays the periodogram along with the fitted spectral density at the posterior mean. As can be seen, the credible interval in red is small due to the large amount of data.

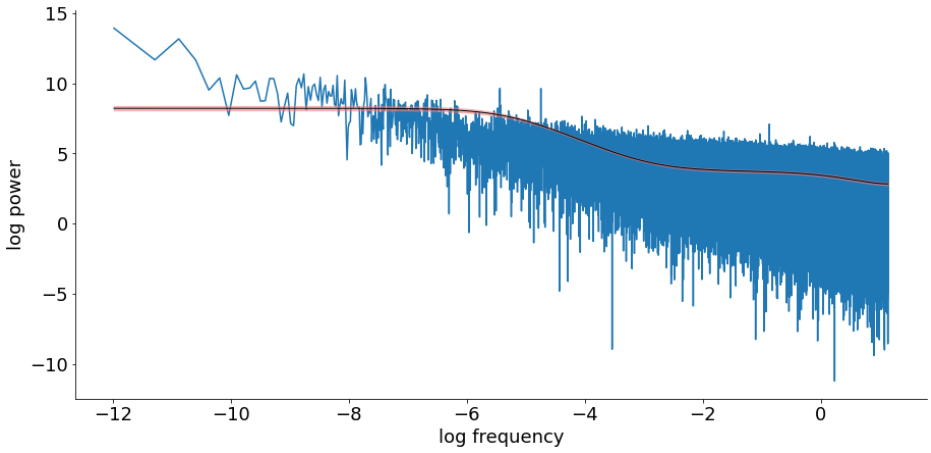


Figure 3.6: The periodogram (blue) of minutely Bitcoin returns with the fitted aliased spectral density at the posterior mean (black), and the thin shaded region (red) is the 95% credible interval from subsampling MCMC.

3.10 Conclusion and future research

In this paper, we perform Bayesian inference for Lévy-driven CARMA processes from Gillberg and Ljung (2009). Furthermore, we propose spectral subsampling of the Whittle likelihood for large regularly-spaced data from Lévy-driven CARMA processes. We demonstrate an average of 100x speed up in RCT compared to the full-data Whittle MCMC for both simulated data and the real data application of Bitcoin returns.

We first consider a simulation study in Section 3.8 to verify the asymptotic normality assumption of the parameter likelihood, which includes non-Gaussian driving processes. Then, to our knowledge, we provide a novel methodology to perform Bayesian inference for Lévy-driven CARMA models. The main contribution of this paper is to extend efficient spectral subsampling MCMC for Lévy-driven CARMA, as seen in Section 3.9.

Future research includes a fractionally integrated (FI) parameter which incorporates long-memory, fractional Lévy processes, known as CARFIMA models Brockwell and Marquardt (2005). Additionally, Bayesian estimation and subsampling can be explored for irregularly spaced Lévy-driven CARMA processes from Fechner and Stelzer (2018). Finally, recent advancements in Bayesian methodologies, such as variational inference Blei et al. (2017), provide exciting possibilities for spectral subsampling MCMC.

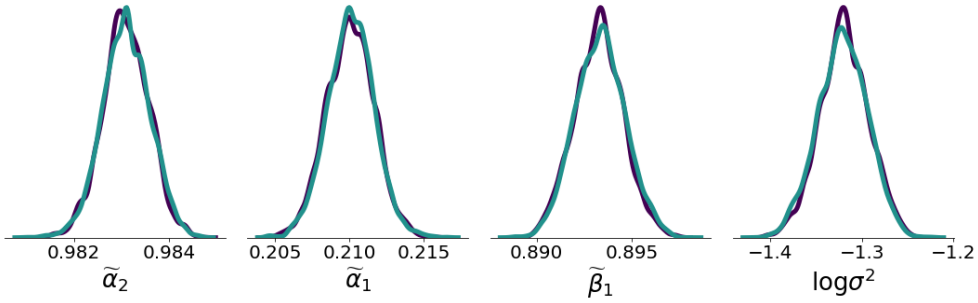


Figure 3.7: Marginal kernel density estimates of the posterior for minutely Bitcoin data for a CARMA$(2, 1)$ model. The standard Whittle posterior is in black, and the subsampled Whittle is in black.

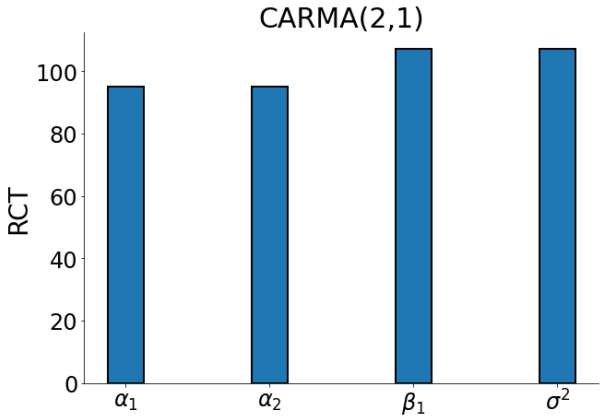


Figure 3.8: Relative computation time of the subsampled Whittle posterior vs full-data Whittle MCMC posterior.

References

Abadi, M., Agarwal, A., Barham, P., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Ammar, G. S. and Gragg, W. B. (1988). Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9(1):61–76.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37(2):697–725.

Applebaum, D. (2009). *Lévy Processes and Stochastic Calculus*. Cambridge University Press.

Bardenet, R., Doucet, A., and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43.

Barndorff-Nielsen, O. and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3(4):408–419.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241.

Belcher, J., Hampton, J., and Wilson, G. T. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):141–155.

Bergstrom, A. R. (1988). The history of continuous-time econometric models. *Econometric Theory*, 4(3):365–383.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. SIAM.

Brockwell, P. (2014). Recent results in the theory and applications of CARMA processes. *Annals of the Institute of Statistical Mathematics*, 66:647–685.

Brockwell, P. J. (2001). Lévy-driven CARMA processes. *Annals of the Institute of Statistical Mathematics*, 53:113–124.

Brockwell, P. J. (2004). Representations of continuous-time ARMA processes. *Journal of Applied Probability*, 41(A):375–382.

Brockwell, P. J., Davis, R. A., and Yang, Y. (2011). Estimation for non-negative Lévy-driven CARMA processes. *Journal of Business & Economic Statistics*, 29(2):250–259.

Brockwell, P. J. and Marquardt, T. (2005). Lévy-driven and fractionally integrated ARMA processes with continuous time parameter. *Statistica Sinica*, 15:477–494.

Contreras-Cristán, A., Gutiérrez-Peña, E., and Walker, S. G. (2006). A note on Whittle's likelihood. *Communications in Statistics-Simulation and Computation*, 35(4):857–875.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. (2019). Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of Machine Learning Research*, 20(100):1–31.

Fasen, V. and Fuchs, F. (2013a). On the limit behavior of the periodogram of high-frequency sampled stable CARMA processes. *Stochastic Processes and their Applications*, 123(1):229–273.

Fasen, V. and Fuchs, F. (2013b). Spectral estimates for high-frequency sampled continuous-time autoregressive moving average processes. *Journal of Time Series Analysis*, 34(5):532–551.

Fasen-Hartmann, V. and Mayer, C. (2022). Whittle estimation for continuous-time stationary state space models with finite second moments. *Annals of the Institute of Statistical Mathematics*, 74:223–270.

Fechner, Ż. and Stelzer, R. (2018). Limit behaviour of the truncated pathwise Fourier-transformation of Lévy-driven CARMA processes for non-equidistant discrete time observations. *Statistica Sinica*, 28(3):1633–1650.

Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press.

Gillberg, J. and Ljung, L. (2009). Frequency-domain identification of continuous-time ARMA models from sampled data. *Automatica*, 45(6):1371–1378.

Graf, M. (2009). Parametric and nonparametric estimation of positive Ornstein-Uhlenbeck type processes. *Diploma Thesis, Centre for Mathematical Sciences, TU München*.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105.

Iacus, S. M. and Mercuri, L. (2015). Implementation of Lévy CARMA model in Yuima package. *Computational Statistics*, 30:1111–1141.

Karatzas, I. and Shreve, S. (1988). *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. World Publishing Company.

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., and Uttley, P. (2014). Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets. *The Astrophysical Journal*, 788(1):33.

Kent, J. T. and Mardia, K. V. (1996). Spectral and circulant approximations to the likelihood for stationary Gaussian random fields. *Journal of Statistical Planning and Inference*, 50(3):379–394.

Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer Science & Business Media.

Moler, C. and Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836.

Mossberg, M. and Larsson, E. K. (2004). Fast and approximative estimation of continuous-time stochastic signals from discrete-time data. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–529. IEEE.

Oksendal, B. (2013). *Stochastic Differential Equations: an Introduction with Applications*. Springer Science & Business Media.

Papapantoleon, A. (2008). An introduction to Lévy processes with applications in finance. *arXiv:0804.0482*.

Peligrad, M. and Wu, W. B. (2010). Central limit theorem for Fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843.

Salomone, R., Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2020). Spectral subsampling MCMC for stationary time series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8449–8458.

Schlemm, E. and Stelzer, R. (2012). Quasi maximum likelihood estimation for strongly mixing state space models and multivariate Lévy-driven CARMA processes. *Electronic Journal of Statistics*, 6:2185–2234.

Sharifi, A., Taheriyoun, A. R., and Hamedani, H. D. (2024). A Bayesian paradigm in a large class of Lévy-driven CARMA models for high frequency data. *Communications in Statistics-Simulation and Computation*, 53(4):1824–1836.

Stelzer, R. (2011). CARMA processes driven by non-Gaussian noise. *arXiv preprint arXiv:1201.0155*.

Sykulski, A. M., Olhede, S. C., Guillaumin, A. P., Lilly, J. M., and Early, J. J. (2019). The debiased Whittle likelihood. *Biometrika*, 106(2):251–266.

Tamaki, K. (2008). The Bernstein-von Mises theorem for stationary processes. *Journal of the Japan Statistical Society*, 38(2):311–323.

Tómasson, H. (2015). Some computational aspects of Gaussian CARMA modelling. *Statistics and Computing*, 25:375–387.

Tsai, H. and Chan, K. (2000). A note on the covariance structure of a continuous-time ARMA process. *Statistica Sinica*, 10(3):989–998.

Villani, M., Quiroz, M., Kohn, R., and Salomone, R. (2022). Spectral Subsampling MCMC for Stationary Multivariate Time Series with Applications to Vector ARTFIMA Processes. *Econometrics and Statistics*.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434.

Chapter 4

Bayesian inference for random fields on a lattice via the debiased spatial Whittle likelihood

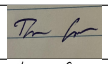
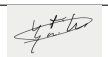
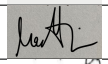
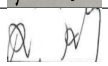
Thomas Goodwin, Arthur Guillaumin, Mattias Villani, Matias Quiroz and Robert Kohn [1]

Abstract

Estimation of regularly spaced, latticed stationary random fields is computationally demanding. Likelihood evaluations for Gaussian random fields have a cost of $\mathcal{O}(|\mathbf{n}|^2)$, where $|\mathbf{n}|$ is the number of data points via efficient algorithms that solve large systems of equations, which quickly becomes intractable for large data. Approximate frequency domain methods have been proposed for parameter estimation based on the Fast Fourier Transform (FFT) with a computational complexity of $\mathcal{O}(|\mathbf{n}| \log |\mathbf{n}|)$. However, it is well known that these Fourier methods suffer from bias. Here, we propose a methodology for Bayesian inference for the debiased spatial Whittle likelihood, which is a frequency domain estimation procedure that reduces bias and accounts for aliasing via the expected periodogram in one $\mathcal{O}(|\mathbf{n}| \log |\mathbf{n}|)$ procedure. This method is based on previous composite-likelihood work, which adjusts the curvature of the likelihood based on the sampling distribution of its maximum likelihood estimator (MLE). It can be shown that this adjustment provides asymptotically 'valid' inference by satisfying the coverage of posterior sets without sacrificing the quasi-linear computation time. We demonstrate our method using two real-data examples: sea surface temperature and Venus topography data.

Keywords: Random fields, Whittle likelihood, Spatial data, Bayesian inference.

[1]Goodwin, Quiroz: *School of Mathematical and Physical Sciences, University of Technology Sydney.* Guillaumin: *School of Mathematical Sciences, Queen Mary University of London.* Villani: *Department of Statistics, Stockholm University.* Kohn: *School of Economics, University of New South Wales.*

Author	Project conception	Work	Manuscript writing	Manuscript editing	Signature
Tom Goodwin		X	X	X	
Matias Quiroz	X			X	
Arthur Guillaumin		X	X		
Mattias Villani	X			X	
Robert Kohn	X				

Status of paper

This chapter is presented as a draft manuscript under preparation to submit to Bayesian Analysis. In the coming months, this manuscript will be uploaded to arXiv. I certify that the work in Chapter 4 has not been submitted as part of any other documents required for a degree.

4.1 Introduction

The collection and analysis of spatial data is crucial in many fields such as geology (Cressie, 1989; Matheron, 1963), climatology (Berliner et al., 2000; Hrafnkelsson and Cressie, 2003), and epidemiology (Tolbert et al., 2000; Best et al., 2001). Technological advancements make it possible to collect and store large amounts of spatial data easily and, in turn, the importance of fitting spatial models is paramount. Fast estimation and Bayesian inference of random fields are key drivers in the spatial statistics literature. Maximum likelihood estimation of Gaussian processes involves computing large systems of equations, which is computationally demanding. To avoid this bottleneck, several approximations have been studied. Time domain methods such as Anitescu et al. (2017); Stein et al. (2013) focus on estimating equation approaches to side-step expensive matrix computations via optimization and stochastic approximations, respectively. Circulant embedding techniques via data imputation reduce computation time by exploiting properties of circulant matrices have been studied in Stroud et al. (2017); Guinness and Fuentes (2017).

Fourier-based methods are attractive as they enable fast estimation and handle large amounts of data via the FFT. The Whittle likelihood is a well-known Fourier-based method first introduced in Whittle (1954). Whittle estimation is explored further in Gelfand et al. (2010), Guinness (2019), and also for the case of irregularly spaced data in Matsuda and Yajima (2009). However, these methods are approximate and typically only asymptotically equivalent, and results in substantial bias for the dimension of the lattice $d > 1$ (Dahlhaus and Künsch, 1987). Guillaumin et al. (2022) alleviates this issue by debiasing the Whittle likelihood but only considers point estimators.

For random fields, uncertainty quantification of parameters for approximate methods proves challenging when performing Bayesian inference. Only a few of the aforementioned papers explicitly study the Bayesian approach for random fields (Stroud et al., 2017; Guinness and Fuentes, 2017). Performing Bayesian inference adds another layer of complexity as the quality of the approximation of the likelihood function and the corresponding posterior distribution over the parameter space (or subset of) is crucial. Kent and Mardia (1996); Guyon (1982) analyzes the quality of the spectral and circulant approximation of the log-likelihoods for stationary Gaussian random fields. They concluded that the bias incurred by this approximation is of the same order as the standard error, and the bias is dominant for $d > 3$. This is discussed further in Section 4.4. Furthermore, coverage of posterior sets, i.e. the probability α that the unknown parameter lies within a certain region based on the posterior distribution, accounts for parameter uncertainty and is essential for Bayesian inference, particularly for prediction. In some cases, it is unclear to perform Bayesian inference with approximate methods (e.g. Stein et al. (2013)).

The contributions of this paper are as follows: we build on the spatial Debiased Whittle likelihood in Guillaumin et al. (2022) to perform 'valid' Bayesian inference, based on the notion of proper likelihoods in Monahan and Boos (1992), for covariance kernel parameters. To include proper parameter posterior uncertainty, we use ideas from Ribatet et al. (2012) to make necessary adjustments to the posterior. To verify the validity of the adjusted posteriors, we use computational schemes from Monahan and Boos (1992) and Cook et al. (2006) in a simulation study in Section 4.5.4. We demonstrate our approach on two real data applications and compare it with the standard Whittle approach.

4.2 Notation and assumptions

Throughout this paper, we assume the spatial data is observed on a regularly spaced lattice, with the possibility of missing data and irregular sampling domains/boundaries. This is described more formally below.

Let $X_{\boldsymbol{s}} \in \mathbb{R}$ be a finite variance, zero-mean random field indexed by the spatial location for $\boldsymbol{s} \in \mathbb{R}^d$ where $d \geq 1$ is a positive integer. Assume $X_{\boldsymbol{s}}$ is homogeneous — but not necessarily isometric — and denote its parametric covariance function by $c_{\boldsymbol{\theta}}(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^d$, which is governed by some unknown parameters of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, with $p \geq 1$, the number of parameters. According to Bochner's theorem (Brockwell and Davis, 2009), there exists a spectral distribution function $F_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ such that,

$$c_{\boldsymbol{\theta}}(\boldsymbol{u}) = \mathrm{E}\left[X_{\boldsymbol{s}} X_{\boldsymbol{s}+\boldsymbol{u}}\right] = \int_{\mathbb{R}^d} \exp(\mathrm{i}\boldsymbol{\omega} \cdot \boldsymbol{u}) dF_{\boldsymbol{\theta}}(\boldsymbol{\omega}), \quad \forall \boldsymbol{u} \in \mathbb{R}^d, \tag{4.1}$$

where $\cdot$ is the dot product and we shall assume that $F_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ is absolutely continuous, such that it

admits a density called the spectral density function, $f_{\boldsymbol{\theta}}(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} c_{\boldsymbol{\theta}}(\boldsymbol{u})\exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{u})d\boldsymbol{u}$.

We observe the random field $X_{\boldsymbol{s}}$ on an orthogonal rectangular grid $\mathcal{G}_{\mathbf{n}}$ with size $\mathbf{n} = (n_1, \ldots, n_d) \in (\mathbb{N}^+)^d$. Without loss of generality, we assume that the grid has a unit step size in all d dimensions. The total number of grid points is denoted by $|\mathbf{n}| = \prod_{i=1}^d n_i$. The *aliased* spectral density $f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})$ of the sampled random field is,

$$f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega}) = \sum_{\boldsymbol{u}\in\mathbb{Z}^d} f_{\boldsymbol{\theta}}(\boldsymbol{\omega} + 2\pi\boldsymbol{u}), \qquad \boldsymbol{\omega} \in \mathbb{R}^d, \tag{4.2}$$

which is a Fourier dual with $c_{\boldsymbol{\theta}}(\boldsymbol{u}) = \int_{\mathcal{T}^d} f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})\exp(\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{u})d\boldsymbol{\omega},\ \forall \boldsymbol{u} \in \mathbb{Z}^d$ and $\mathcal{T} = [0, 2\pi)$.

4.3 The debiased Whittle likelihood

This section covers the definition of the debiased spatial Whittle likelihood, introduced in Guillaumin et al. (2022).

4.3.1 Frequentist estimation

The periodogram of the sampled random field is given by

$$I_{\mathbf{n}}(\boldsymbol{\omega}) = \frac{(2\pi)^{-d}}{n}\left|\sum_{\boldsymbol{s}\in\mathcal{G}_{\mathbf{n}}} X_{\boldsymbol{s}}\exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{s})\right|^2, \qquad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{4.3}$$

This quantity can be efficiently evaluated at a computation cost of $\mathcal{O}(|\mathbf{n}|\log|\mathbf{n}|)$ operations via the Fast Fourier Transform (FFT) on the multidimensional grid of Fourier frequencies associated with the spatial grid $\mathcal{G}_{\mathbf{n}}$,

$$\mathcal{G}_{\mathbf{n}} = \prod_{j=1}^d \left\{2\pi k n_j^{-1} : k = 0, \ldots, n_j - 1\right\}.$$

The Whittle likelihood (Whittle, 1954) is a computationally efficient approximation to the Gaussian likelihood that relies on the approximate independence of periodogram ordinates on the Fourier grid. More recently, corrections to the standard Whittle likelihood for time series were proposed by Sykulski et al. (2019) that alleviate some of its limitations, such as finite sampling effects and severe bias in small sample sizes. This was extended to domains of higher dimensions in Guillaumin et al. (2022), e.g. for the analysis of spatial or spatiotemporal random processes. The approach taken is to estimate the parametric expectation of the periodogram rather than

the spectral density, according to the debiased spatial Whittle likelihood,

$$\ell_{dW}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{\boldsymbol{\omega}\in\Omega_{\mathbf{n}}}\left\{\log \overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta}) + \frac{I_{\mathbf{n}}(\boldsymbol{\omega})}{\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta})}\right\}, \tag{4.4}$$

for all $\boldsymbol{\theta} \in \Theta$, where,

$$\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left[I_{\mathbf{n}}(\boldsymbol{\omega})\right], \qquad \forall \boldsymbol{\omega} \in \mathcal{T}^d,$$

is the expected periodogram. The expected periodogram is a convolution

$$\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta}) = \left\{f_{\boldsymbol{\theta}} * \mathcal{F}_{\mathbf{n}}\right\}(\boldsymbol{\omega}) = \int_{\mathcal{T}^d} f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega}-\boldsymbol{\omega}')\mathcal{F}_{\mathbf{n}}(\boldsymbol{\omega}')d\boldsymbol{\omega}', \tag{4.5}$$

between the spectral density of the process and the multi-dimensional kernel

$$\mathcal{F}_{\mathbf{n}}(\boldsymbol{\omega}) = \frac{(2\pi)^{-d}}{\sum g_{\mathbf{s}}^2}\left|\sum_{\boldsymbol{s}\in\mathcal{G}_{\mathbf{n}}} g_{\boldsymbol{s}}\exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{s})\right|^2, \qquad \boldsymbol{\omega}\in\mathbb{R}^d, \tag{4.6}$$

where $g_{\boldsymbol{s}}$, $\forall \boldsymbol{s} \in \mathcal{G}_{\mathbf{n}}$ is a masking grid, which takes value 0 if an observation is missing, and 1 otherwise. The kernel in (4.6) is known as the *modified* Féjer kernel. In the case of a fully observed domain, this kernel becomes the multidimensional rectangular Féjer kernel, i.e. a separable product of one-dimensional Féjer kernels.

Maximization of (4.4) over Θ defines the maximum debiased Whittle likelihood estimate (MdWLE) of the data, given as,

$$\widehat{\boldsymbol{\theta}}_{\mathrm{dW}} = \arg\max_{\boldsymbol{\theta}\in\Theta}\left\{\ell_{\mathrm{dW}}(\boldsymbol{\theta})\right\}. \tag{4.7}$$

The replacement of $f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})$ with $\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta})$ yields the parametric model

$$I_{\mathbf{n}}(\boldsymbol{\omega}) \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}\left\{\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta})\right\}, \qquad \boldsymbol{\omega}\in\Omega_{\mathbf{n}}, \tag{4.8}$$

where $\mathrm{Exp}(\lambda)$ is an exponential distribution parameterized by its mean. Despite the inclusion as the expected periodogram, the iid assumption in (4.8) is only asymptotic, as $|\mathbf{n}| \to \infty$ (Bandyopadhyay and Lahiri, 2009). Thus, for finite samples, the likelihood in (4.4) can be considered misspecified and follows the framework of composite likelihoods (Varin et al., 2011) and hence (4.7) is a maximum composite likelihood estimator. Observe that $\mathrm{E}_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta}) = \mathbf{0}$, where $\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})$ is the score function, which fits the methodology of estimating equations in Heyde (1997). Thus, Bayesian inference, more precisely, correct uncertainty quantification of the parameters, is not straightforward, see Ribatet et al. (2012). This will be explored further in Section 4.5.

The computation of the expected periodogram in (4.5) takes advantage of the FFT and hence is a $\mathcal{O}(|\mathbf{n}| \log |\mathbf{n}|)$ routine. This is for any d, regardless of missing data or irregular domain patterns. The expected periodogram can directly account for finite sampling effects, mainly aliasing and spectral leakage. The quantity $\overline{I}_{\mathbf{n}}(\boldsymbol{\omega}; \boldsymbol{\theta})$ is computed via the FFT of the discretely sampled theoretical covariance function combined with, in the case of fully observed grids, a multidimensional triangular kernel (Percival and Walden, 1993). Thus, the expected periodogram accounts for aliasing via discrete sampling of covariance function and spectral leakage by including the triangular kernel.

In the case of fully observed grids, the expected periodogram is related to the spectral density in the sense that

$$\mathrm{E}_{\boldsymbol{\theta}} \left[I_{\mathbf{n}}(\boldsymbol{\omega}) \right] \xrightarrow[\mathbf{n} \to \infty]{} f_{\boldsymbol{\theta}, \boldsymbol{\delta}}(\boldsymbol{\omega}), \tag{4.9}$$

where $\mathbf{n} \to \infty$ denotes $n_i \to \infty$ for $i = 1, \ldots, d$. Thus, as the number of observations grows infinitely on all sides, the expected periodogram converges to the spectral density, and hence there is an equivalence asymptotic equivalence of the Debiased Whittle and standard Whittle likelihoods. For more details on the expected periodogram and the computation thereof, refer to Guillaumin et al. (2022).

Approaches such as Guinness and Fuentes (2017); Stroud et al. (2017) use procedures that impute missing observations via circulant embedding, which may not be appropriate when the data violates the Gaussian assumption. Instead, the debiased Whittle handles missing observations and irregular sampling domains via the modulation values $g_{\boldsymbol{s}}$.

It is important to note, as discussed in Sykulski et al. (2019); Guillaumin et al. (2022), that the computation of $f_{\boldsymbol{\theta}, \boldsymbol{\delta}}(\boldsymbol{\omega})$ in the standard Whittle estimation is not automatic and more complicated in general. This is because the aliased spectral density $f_{\boldsymbol{\theta}, \boldsymbol{\delta}}(\boldsymbol{\omega})$ seldom has an analytical form and is usually approximated in practice, see Chapter 5, Gelfand et al. (2010), by truncation of the infinite sum in (4.2) via 'wrapping' contributions of $f_{\boldsymbol{\theta}}(\boldsymbol{\omega})$ from frequencies higher than the Nyquist. Although this aforementioned process may be computationally less expensive on large grids for specific cases, computation of $\overline{I}_{\mathbf{n}}(\boldsymbol{\omega}; \boldsymbol{\theta})$ yields a procedure for automatic and exact calculation of aliasing, spectral leakage, missing data and irregular sampling domains, in a convenient way, rather than accounting for each of these effects individually.

4.4 Likelihood comparison and issues

A primary objective of Bayesian computation is to infer the posterior $\pi(\boldsymbol{\theta}|X_{\boldsymbol{s}}) \propto \mathcal{L}(\boldsymbol{\theta}) p(\boldsymbol{\theta})$ which requires knowledge of the likelihood function $\mathcal{L}(\boldsymbol{\theta})$ over a non-negligible probability area in Θ. It is not enough to only consider the maximum likelihood estimator of the debiased spatial Whittle likelihood. We need to assess the approximation of $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$ to the exact likelihood over the parameter space Θ, or at least in a locally compact region of non-negligible probability around

$\widehat{\boldsymbol{\theta}}_{\text{dW}}$. The exact likelihood for Gaussian random fields is

$$\ell_{\text{true}}(\boldsymbol{\theta}) = -\frac{|\mathbf{n}|}{2}\log(2\pi) - \frac{1}{2}\log\det\{\Sigma(\boldsymbol{\theta})\} - \frac{1}{2}X_{\boldsymbol{s}}^T\Sigma^{-1}(\boldsymbol{\theta})X_{\boldsymbol{s}}, \tag{4.10}$$

where $\Sigma(\boldsymbol{\theta})$ is the $|\mathbf{n}| \times |\mathbf{n}|$ covariance matrix corresponding to $c_{\boldsymbol{\theta}}(\boldsymbol{u})$ and $\det\{\Sigma(\boldsymbol{\theta})\}$ is the determinant of the covariance matrix. However, this is computationally challenging since $\det\{\Sigma(\boldsymbol{\theta})\}$ is $\mathcal{O}(|\mathbf{n}|^3)$ in general, or $\mathcal{O}(|\mathbf{n}|^{5/2})$ in structured cases (Sowell, 1989; Akaike, 1973). Furthermore, this likelihood is restrictive as it assumes the data-generating process is Gaussian, which may not be appropriate when modelling real data (Guilleminot, 2020).

For a rectangular grid that increases in all directions, the periodogram $I_{\mathbf{n}}(\boldsymbol{\omega})$ is an asymptotically unbiased estimate of $f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})$ with distribution $I_{\mathbf{n}}(\boldsymbol{\omega}) \overset{\text{i.i.d.}}{\sim} \text{Exp}\left\{f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})\right\}$, for $\boldsymbol{\omega} \in \Omega_{\mathbf{n}}$. This motivates the original spatial Whittle likelihood (Whittle, 1954),

$$\ell_W(\boldsymbol{\theta}) = \frac{1}{2}\sum_{\boldsymbol{\omega}\in\Omega_{\mathbf{n}}}\left\{\log f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega}) + \frac{I_{\mathbf{n}}(\boldsymbol{\omega})}{f_{\boldsymbol{\theta},\boldsymbol{\delta}}(\boldsymbol{\omega})}\right\}. \tag{4.11}$$

Denote $\widehat{\boldsymbol{\theta}}_{\text{W}}$ as the maximum likelihood estimators of (4.11).

The exact likelihood in (4.10) and the aforementioned Whittle likelihood (4.11) have the same purpose, to estimate the parameters that govern the second order structure of the underlying process but do so in different ways. The exact likelihood computes the density of the original data given a covariance model through the covariance matrix $\Sigma(\boldsymbol{\theta})$. In contrast, the Whittle approach, or frequency domain methods in general, studies the periodogram to estimate the parameters of the spectral density or the spectral density itself. Hence, an important benefit of frequency domain estimation methods is no explicit assumption of the data-generating process is required, only an appropriate model for the periodogram.

Guyon (1982) and Kent and Mardia (1996) studied the quality of the Whittle likelihood approximation to the exact likelihood. Based on Proposition 1 from Guyon (1982), assume $n_1 \leq \cdots \leq n_d$, as $n_1 \to \infty$, then,

$$|\ell_{\text{true}}(\boldsymbol{\theta}) - \ell_W(\boldsymbol{\theta})| = \mathcal{O}_p(|\mathbf{n}|/n_1). \tag{4.12}$$

From Proposition 1, for $d = 2$, the bias of $\widehat{\boldsymbol{\theta}}_{\text{W}}$ is of the same order as the standard error; however, for $d > 2$ the bias is of larger order than the standard error. Hence, for $d \geq 2$, $\widehat{\boldsymbol{\theta}}_{\text{W}}$ is not an efficient estimator. For tapered data, Dahlhaus and Künsch (1987) show that $\widehat{\boldsymbol{\theta}}_{\text{W}}$ is efficient for $d = 2, 3$. Despite the efficiency of $\widehat{\boldsymbol{\theta}}_{\text{W}}$ for tapered data, the rate of convergence is of the same order as the smallest side n_1. This is a drawback of the regular Whittle approximation as the bias of the MLE and its likelihood approximation is limited by the smallest side. Simulation studies suggest that the bias of $\widehat{\boldsymbol{\theta}}_{\text{W}}$ decreases slowly with respect to the grid size, see Figures 1 and 2 of

Guillaumin et al. (2022).

To assess the approximation of the likelihood function over Θ or around $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$, we are interested in the relative order of error, i.e.

$$\left|\frac{\ell_{\text{true}}(\boldsymbol{\theta}) - \ell_W(\boldsymbol{\theta})}{\ell_{\text{true}}(\boldsymbol{\theta})}\right| = \mathcal{O}_{\mathrm{p}}(1/n_1). \tag{4.13}$$

The above equation illustrates the convergence of the Whittle likelihood to the exact likelihood irrespective of a normalization constant. Thus for fully observed square grids that increase asymptotically in all directions, the relative error of the standard Whittle likelihood goes to zero. It follows from (4.9) that the same statement can be made for the debiased Whittle likelihood.

Empirically, the rate of $1/n_1$ where n_1 is the smallest side of the hypercube, leads to poor approximations of the true likelihood function. For demonstration, we consider a simulated example with an isotropic Mátern kernel,

$$c_{\boldsymbol{\theta}}(\boldsymbol{u}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{||\boldsymbol{u}||}{\rho}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{||\boldsymbol{u}||}{\rho}\right), \tag{4.14}$$

where $K_\nu(x)$ is a Bessel function of the second kind. The three governing parameters are the range ρ, amplitude σ and ν smoothness. Here, we set true values of $\rho = 5$ and $\sigma = 1$, and we fix $\nu = \infty$, which is not estimated. This corresponds to the squared-exponential kernel. This kernel is well known to be impractical for most real-world applications as it exhibits overly smooth processes (Stein, 2012). Furthermore, the spectral density of this process has negligible power for higher frequencies, resulting in difficulty in the estimation of the range parameter, particularly for higher values of ρ relative to the domain size. The data is simulated on a square grid of $n = (64, 64)$, and we consider three likelihoods: the debiased spatial Whittle likelihood, the standard Whittle likelihood and the exact Gaussian likelihood, all with the same non-informative prior. Figure 4.1 shows marginal kernel density estimates of each of the three posteriors.

As seen in Figure 4.1, the standard Whittle likelihood underestimates both ρ and σ, whereas the debiased spatial Whittle likelihood underestimates only σ. Both Whittle-type likelihoods are not 'close' to the exact Gaussian likelihood. Furthermore, the posteriors of both Whittle-type likelihoods, particularly in the σ parameter, are more concentrated, a common phenomenon for composite likelihoods, see Ribatet et al. (2012). Equation (4.13) suggests that n_1 to be in the thousands for $d = 2$ to have a faithful approximation of the true log-likelihood function (up to a normalizing constant), which is rarely achieved in practice.

Instead of approximating the exact likelihood function via the Whittle likelihood, we leverage properties of the Debiased Whittle likelihood estimator to include appropriate uncertainty by performing valid Bayesian inference discussed in the next section.

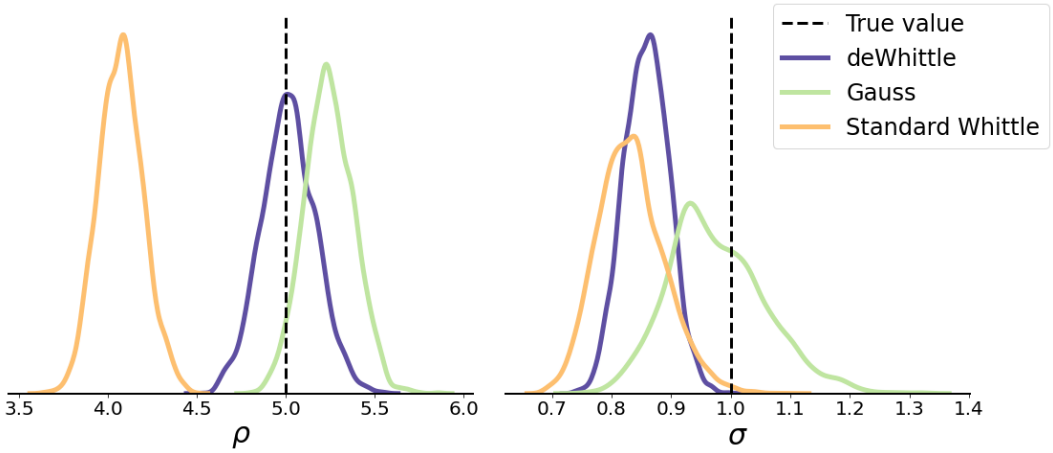


Figure 4.1: Kernel density estimates of the marginal posterior comparison of simulated data example with a squared-exponential kernel with grid size $n = (64, 64)$.

4.5 Bayesian coverage

In real-world applications, it is sometimes necessary to make approximations. In the context of Bayesian inference, approximating the likelihood or prior, and hence the posterior, is often done for computational convenience due to some intractability. As a natural consequence of employing approximate posteriors, the coverage of the approximate credible posterior sets is changed. Bayesian calibration procedures correct for or quantify the error induced by the approximation (Lee et al., 2019; Frazier et al., 2023). Furthermore, approaches such as Yao et al. (2018) and Prangle et al. (2014) provide diagnostic tools based on coverages for different posterior approximation methods. Other methods use theoretical coverage properties to provide tests/checks for correct software implementation, see Cook et al. (2006); Geweke (2004). At its core, these aforementioned methods borrow ideas from Monahan and Boos (1992), which is briefly described below.

We follow the approach of Monahan and Boos (1992), which defines valid posteriors based on coverage properties of posterior sets. Assume the data are generated from the model $X \sim p(X_{\boldsymbol{s}}|\boldsymbol{\theta})$ and suppose $\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})$ is the likelihood of interest (highlighting the dependence on the dataset), and we wish to perform inference via Bayes' theorem,

$$p(\boldsymbol{\theta}|X_{\boldsymbol{s}}) = \frac{\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})p(\boldsymbol{\theta})}{p(X_{\boldsymbol{s}})}, \quad p(X_{\boldsymbol{s}}) = \int_{\Theta} \mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.15}$$

Define $K_\alpha(X_{\boldsymbol{s}})$ as a posterior coverage set function of level α if, for every $X_{\boldsymbol{s}}$,

$$\Pr\{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})|X_{\boldsymbol{s}}\} = \alpha, \quad (X_{\boldsymbol{s}} \text{ conditioned upon}), \tag{4.16}$$

where the probability is with respect to the posterior (conditional on $X_{\boldsymbol{s}}$) with likelihood $\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})$ and prior $p(\boldsymbol{\theta})$. A simple example of a posterior coverage set function when the parameter is one-dimensional is a one-sided interval that contains α of the posterior mass, also known as a credible interval. The posterior is said to be valid by coverage if, for every α,

$$\Pr\{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})\} = \alpha, \quad (X_{\boldsymbol{s}} \text{ is random}), \tag{4.17}$$

where the probability is now with respect to the joint distribution $p(X_{\boldsymbol{s}}, \boldsymbol{\theta}) = p(X_{\boldsymbol{s}}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Note the similarity to the classical statistical coverage, however, the sampling distribution of $X_{\boldsymbol{s}}$ is averaged over the prior (instead of keeping the parameter at its true value). When the likelihood comes from the data generating process ($\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}}) = p(X_{\boldsymbol{s}}|\boldsymbol{\theta})$), the posterior is valid by coverage since (4.17) holds,

$$\begin{aligned}
\Pr\{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})\} &= \int_{X_{\boldsymbol{s}}} \int_{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})} p(X_{\boldsymbol{s}}, \boldsymbol{\theta}) d\boldsymbol{\theta} dX_{\boldsymbol{s}} \\
&= \int_{X_{\boldsymbol{s}}} \int_{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})} \mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} dX_{\boldsymbol{s}} \\
&= \int_{X_{\boldsymbol{s}}} p(X_{\boldsymbol{s}}) \left(\int_{\boldsymbol{\theta} \in K_\alpha(X_{\boldsymbol{s}})} p(\boldsymbol{\theta}|X_{\boldsymbol{s}}) d\boldsymbol{\theta} \right) dX_{\boldsymbol{s}} \\
&= \int_{X_{\boldsymbol{s}}} p(X_{\boldsymbol{s}}) \alpha dX_{\boldsymbol{s}} \\
&= \alpha,
\end{aligned}$$

using (4.15) and the definition of a posterior coverage set function in (4.16). For the simplest case $\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}}) = p(X_{\boldsymbol{s}}|\boldsymbol{\theta})$, the posterior is validated by Bayes's theorem since the likelihood is the exact conditional density of the data $X_{\boldsymbol{s}}$ given $\boldsymbol{\theta}$. As in Monahan and Boos (1992), we are interested in the case when $\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})$ differs from $p(X_{\boldsymbol{s}}|\boldsymbol{\theta})$; due to the intractability of the true likelihood $p(X_{\boldsymbol{s}}|\boldsymbol{\theta})$ for Gaussian random fields in (4.10), the likelihood function we consider is the debiased spatial Whittle likelihood $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$.

4.5.1 Posterior adjustments

Simply substituting $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$ for $\mathcal{L}(\boldsymbol{\theta}; X_{\boldsymbol{s}})$ in (4.15) will not yield proper coverage posteriors, as demonstrated in the simulation study in Section 4.5.4. Ribatet et al. (2012) propose an asymptotic curvature adjustment for composite likelihoods. This curvature adjustment is based on the Bernstein Von-Mises theorem (Van der Vaart, 2000), which loosely states that the posterior converges to the sampling distribution of the MLE. Thus, the curvature adjustment corrects the variance of a composite posterior to equal (asymptotically) the variance of its corresponding maximum composite likelihood estimator (MCLE). In our case, Guillaumin et al. (2022) show for finite

grid sizes, the debiased spatial Whittle likelihood is a composite likelihood (Varin et al., 2011; Bevilacqua and Gaetan, 2015) and fits within the framework of estimation equations (Heyde, 1997). We briefly review how to perform these adjustments for composite likelihoods.

Denote the composite likelihood as $\ell_c(\boldsymbol{\theta})$ with its maximum composite likelihood estimator as $\widehat{\boldsymbol{\theta}}_c$ and the true parameter as $\boldsymbol{\theta}_0$. Moreover, denote

$$\boldsymbol{H}(\boldsymbol{\theta}_0) = -\mathrm{E}\left[\nabla^2_{\boldsymbol{\theta}}\ell_c(\boldsymbol{\theta})\right], \quad \boldsymbol{J}(\boldsymbol{\theta}_0) = \mathrm{Var}\left[\nabla_{\boldsymbol{\theta}}\ell_c(\boldsymbol{\theta})\right],$$

as the Fisher information matrix and the covariance of the score function, respectively. The distribution of the maximum composite likelihood estimator (MCLE) $\widehat{\boldsymbol{\theta}}_c$ is

$$\sqrt{|\mathbf{n}|}\left\{\boldsymbol{G}(\boldsymbol{\theta}_0)\right\}^{1/2}(\widehat{\boldsymbol{\theta}}_c - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathrm{I}). \tag{4.18}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is often referred to as 'sandwich' variance matrices (Varin et al., 2011), defined as

$$\boldsymbol{G}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{J}^{-1}(\boldsymbol{\theta})\boldsymbol{H}(\boldsymbol{\theta}). \tag{4.19}$$

Maximum likelihood estimation of composite likelihoods can also be viewed as a maximum likelihood estimation for mis-specified models (White, 1982). However, for large enough $|\mathbf{n}|$, it can be shown (Appendix A of Ribatet et al. (2012)) that the composite posterior is

$$\pi_c(\boldsymbol{\theta}|y) \sim \mathrm{N}\left\{\boldsymbol{\theta}_0, |\mathbf{n}|^{-1}\boldsymbol{H}^{-1}(\boldsymbol{\theta}_0)\right\}. \tag{4.20}$$

For proper likelihoods, $\boldsymbol{H}(\boldsymbol{\theta}) = -\boldsymbol{J}(\boldsymbol{\theta})$, and for large enough $|\mathbf{n}|$, the posterior is roughly equal to the sampling distribution of the MLE's (Bernstein von-Mises theorem). However, from (4.18), the variance of MCLE is the sandwich estimator, and thus, the posterior under the composite likelihood does not converge to the sampling distribution of the MCLE. Intuitively, the composite marginal posterior distributions of the parameters are too concentrated (small variance) compared to that of the full posterior (see Figure 1 of Ribatet et al. (2012)), as is the case for σ in Figure 4.1. To alleviate this, the proposed asymptotic curvature adjustment to the composite log-likelihood $\ell_c(\boldsymbol{\theta})$ given by

$$\ell_{\mathrm{curv}}(\boldsymbol{\theta}|y) = \ell_c(\boldsymbol{\theta}^*|y), \qquad \boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}_c + \boldsymbol{C}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_c), \tag{4.21}$$

where C is a positive semi-definite 'adjustment' matrix

$$\boldsymbol{C}^\top\boldsymbol{H}(\boldsymbol{\theta}_0)\boldsymbol{C} = \boldsymbol{H}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}(\boldsymbol{\theta}_0). \tag{4.22}$$

One possible choice is $\boldsymbol{C} = \boldsymbol{M}^{-1}\boldsymbol{M}_A$ where $\boldsymbol{M}_A^\top\boldsymbol{M}_A = \boldsymbol{G}(\boldsymbol{\theta}_0)$ and $\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{H}(\boldsymbol{\theta}_0)$. The purpose of this adjustment is for $\ell_{\mathrm{curv}}(\boldsymbol{\theta}|y)$ at $\widehat{\boldsymbol{\theta}}_c$ to match the curvature of the large-sample

density of $\widehat{\boldsymbol{\theta}}_c$. As a consequence, the curvature adjustment changes the location of any local maxima except the global maximum at $\widehat{\boldsymbol{\theta}}_c$, which may not be appropriate if the full posterior is multi-modal. The asymptotic distribution of the curvature adjusted posterior (see Appendix A of Ribatet et al. (2012)) is

$$\pi_{\text{curv}}(\boldsymbol{\theta}|y) \sim \text{N}\left\{\boldsymbol{\theta}_0, |\mathbf{n}|^{-1}\boldsymbol{G}^{-1}(\boldsymbol{\theta}_0)\right\}.$$

As shown in Guillaumin et al. (2022), the variance of the sampling distribution of the maximum spatial debiased Whittle likelihood estimator (MdWLE) $\widehat{\boldsymbol{\theta}}_{\text{dW}}$ is

$$\text{Var}\left\{\widehat{\boldsymbol{\theta}}_{\text{dW}}\right\} \approx \boldsymbol{G}^{-1}(\boldsymbol{\theta}), \tag{4.23}$$

which has the same sandwich structure as the covariance of the MCLE in (4.18). Note, Guillaumin et al. (2022) gives an analytical form of the asymptotic distribution of the MdWLE for Gaussian random fields when the observation domain $\mathbf{n}$ grows to infinity in all directions; however, this form is seldom reached in practice. Simons and Olhede (2013) gives a practical large-sample case where the asymptotic form has not been reached. In addition, empirical findings via simulations and, in the case of missing data, prevent the use of the exact form of the asymptotic variance for posterior adjustments.

4.5.2 Computation of curvature adjustments

This section explains the computation of the sandwich matrix in (4.23) specific to the debiased spatial Whittle likelihood. The computation of the analytic form of (4.23) is intractable for larger grids for the reasons explained below. Instead, we provide two methods to estimate the adjustment matrix $\boldsymbol{C}$ via Monte Carlo simulation. It is important to note that the computation of the adjustment matrices is performed once before MCMC.

Guillaumin et al. (2022) give an analytic approximation $\boldsymbol{H}$ to the Fisher matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{\boldsymbol{\omega}\in\Omega_{\mathbf{n}_k}} \overline{I}_{\mathbf{n}_k}(\boldsymbol{\omega};\boldsymbol{\theta})^{-2}\nabla_{\boldsymbol{\theta}}\overline{I}_{\mathbf{n}_k}(\boldsymbol{\omega};\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\overline{I}_{\mathbf{n}_k}(\boldsymbol{\omega};\boldsymbol{\theta})^{\top},$$

where the gradient of the expected periodogram is

$$\nabla_{\boldsymbol{\theta}}\overline{I}_{\mathbf{n}}(\boldsymbol{\omega};\boldsymbol{\theta}) = \sum_{\boldsymbol{u}\in\mathrm{Z}^d} \nabla_{\boldsymbol{\theta}}\overline{c}_{\mathbf{n}}(\boldsymbol{u};\boldsymbol{\theta})\exp(-\mathrm{i}\boldsymbol{\omega}\cdot\boldsymbol{u}).$$

The aforementioned equation can be computed efficiently via the FFT and using the same procedure as the expected periodogram, replacing the covariance function with its gradient wrt the parameters.

The i,jth element in the variance of the score $\boldsymbol{J}(\boldsymbol{\theta})$ matrix is given as

$$\operatorname{cov}\left\{\frac{\partial \ell_{\mathrm{dW}}(\boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial \ell_{\mathrm{dW}}(\boldsymbol{\theta})}{\partial \theta_j}\right\} = |\mathbf{n}|^{-2} \sum_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2 \in \Omega_{\mathbf{n}}} \frac{\operatorname{cov}\{I_{\mathbf{n}}(\boldsymbol{\omega}_1), I_{\mathbf{n}}(\boldsymbol{\omega}_2)\}}{\overline{I}^2_{\mathbf{n}}(\boldsymbol{\omega}_1;\boldsymbol{\theta}), \overline{I}^2_{\mathbf{n}}(\boldsymbol{\omega}_2;\boldsymbol{\theta})} \frac{\partial \overline{I}_{\mathbf{n}}(\boldsymbol{\omega}_1;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \overline{I}_{\mathbf{n}}(\boldsymbol{\omega}_2;\boldsymbol{\theta})}{\partial \theta_j}. \quad (4.24)$$

The covariance between the periodogram at two different Fourier frequencies on the RHS of (4.24) involves complicated convolutions of the spectral density and Dirichlet kernel (see Section 5.5 of Guillaumin et al. (2022)). Furthermore, computation of (4.24) scales as $\mathcal{O}(|\mathbf{n}|^2)$, which becomes intractable for even moderate grid sizes. Guillaumin et al. (2022) propose a speedup to compute this quantity, which still involves the computation of integrals in the form of convolutions. Furthermore, missing data or irregular domains add another layer of complexity as (4.24) is not known exactly.

A simple but effective solution is to use a Monte Carlo estimate of $\operatorname{Var}\{\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{dW}}(\boldsymbol{\theta})\}$, namely,

$$\boldsymbol{J}(\boldsymbol{\theta}) \approx \widehat{\boldsymbol{J}}(\boldsymbol{\theta}) = \frac{1}{k-1} \sum_{i=1}^{k} \left(\boldsymbol{g}^{(i)} - \overline{\boldsymbol{g}}\right) \left(\boldsymbol{g}^{(i)} - \overline{\boldsymbol{g}}\right)^{\top}, \quad (4.25)$$

where $\boldsymbol{g}^{(i)} = \nabla_{\boldsymbol{\theta}} \ell_{\mathrm{dW}}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}} | X^{(i)}_{\boldsymbol{s}})$. Once the MdWLE given the observed data is found, one must simulate k random fields from the specified model conditional on $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$. Then, $\widehat{\boldsymbol{J}}(\boldsymbol{\theta})$ is obtained as the empirical covariance estimator based on k Monte Carlo samples of the gradient $\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{dW}}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}} | X^{(i)}_{\boldsymbol{s}})$. Note that simulation of Gaussian random fields can be performed efficiently in $\mathcal{O}(|\mathbf{n}|\log|\mathbf{n}|)$ time via circulant embedding (Dietrich and Newsam, 1997). The modified adjustment is

$$\boldsymbol{C}_1 = \boldsymbol{M}^{-1} \boldsymbol{M}_A, \quad (4.26)$$

$$\boldsymbol{M}_A \boldsymbol{M}_A^{\top} = \boldsymbol{H}(\boldsymbol{\theta}_0) \widehat{\boldsymbol{J}}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{H}(\boldsymbol{\theta}_0), \quad \boldsymbol{M}\boldsymbol{M}^{\top} = \boldsymbol{H}(\boldsymbol{\theta}_0), \quad (4.27)$$

where $\boldsymbol{M}$ and $\boldsymbol{M}_A$ are lower triangular Cholesky decompositions.

The second adjustment is obtained by replacing the estimated sandwich matrix in (4.27) with a Monte Carlo estimate of $\operatorname{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$. An estimate thereof is

$$\operatorname{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\} \approx \widehat{\boldsymbol{G}^{-1}(\boldsymbol{\theta})} = \frac{1}{k-1} \sum_{i=1}^{k} \left(\widetilde{\boldsymbol{\theta}}^{(i)}_{\mathrm{dW}} - \overline{\widetilde{\boldsymbol{\theta}}_{\mathrm{dW}}}\right) \left(\widetilde{\boldsymbol{\theta}}^{(i)}_{\mathrm{dW}} - \overline{\widetilde{\boldsymbol{\theta}}_{\mathrm{dW}}}\right)^{\top}, \quad (4.28)$$

where $\widetilde{\boldsymbol{\theta}}^{(i)}_{\mathrm{dW}}$, for $i = 1, \ldots, k$ are the MdWLE from k datasets simulated at $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$. Simulating data $X^{(i)}_{\boldsymbol{s}}$ from the likelihood at $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$ and finding the corresponding MdWLE for$k$ iterations gives the simulation approximation of the MdWLE distribution. Furthermore, the observed Fisher $\mathcal{H}(\boldsymbol{\theta})$ of $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$ is used in place the $\boldsymbol{H}(\boldsymbol{\theta})$ for the construction of $\boldsymbol{M}$. This is due to the fact that in

finite samples, the observed Fisher will match the curvature of the un-adjusted log-likelihood and hence provide a tailored curvature adjustment for the specific dataset. The adjustment is

$$\boldsymbol{C}_2 = \boldsymbol{M}\boldsymbol{M}_A^{-1}, \tag{4.29}$$

$$\boldsymbol{M}_A\boldsymbol{M}_A^\top = \widehat{\boldsymbol{G}^{-1}(\boldsymbol{\theta})}, \quad \boldsymbol{M}\boldsymbol{M}^\top = \mathcal{H}^{-1}(\boldsymbol{\theta}). \tag{4.30}$$

We denote the curvature adjusted debiased Whittle likelihoods as

$$\ell_{\mathrm{dW}}^{(i)}(\boldsymbol{\theta}) = \ell_{\mathrm{dW}}(\boldsymbol{\theta}^*), \qquad \boldsymbol{\theta}^* = \widehat{\boldsymbol{\theta}}_{\mathrm{dW}} + C_i(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{dW}}), \quad \text{for } i = 1, 2. \tag{4.31}$$

Combining $\ell_{\mathrm{dW}}^{(i)}(\boldsymbol{\theta})$ with a specified prior in a Metropolis-Hastings algorithm will target the desired posterior density $\pi_{\mathrm{dW}}^{(i)}(\boldsymbol{\theta}|X_{\boldsymbol{s}})$ for $i = 1, 2$. Algorithms 1 and 2 describe the computation necessary to obtain adjustments $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, respectively.

4.5.3 Considerations

We briefly discuss some considerations and recommendations for both curvature adjustments.

The $\boldsymbol{C}_1$ matrix requires the evaluation of $\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})$, which requires the gradient of the covariance function wrt the parameters. The gradient of the Mátern kernel in (4.14) wrt ν exists analytically but is difficult and computationally expensive (Geoga et al., 2023). For this reason, we restricted the use of $\boldsymbol{C}_1$ for fixed ν.

Both adjustments employ Monte Carlo estimation. The variance of the Monte Carlo estimates is an important consideration when choosing k to compute the adjustments. Thus, for cases when the variance is higher, a larger k is required. Generally, small and moderate grid sizes require larger k.

The $\boldsymbol{C}_1$ adjustment performs a Monte Carlo estimate of $\mathrm{Var}\{\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})\}$, whereas the $\boldsymbol{C}_2$ adjustment estimates $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$. The variance of $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$ is large when the domain size is small relative to the value of ρ. In this case, we recommend using $\boldsymbol{C}_1$ since $\mathrm{Var}\{\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})\}$, and the variance thereof is less sensitive to ρ compared to the domain size.

However, for more difficult settings such as missing data and irregular domains, the estimate of $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$ will be a better approximation of $\boldsymbol{G}^{-1}(\boldsymbol{\theta})$ compared to the $\boldsymbol{C}_1$ adjustment. Additionally, the computation of $\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})$ for $\boldsymbol{C}_1$ may not be known in closed form for missing data and/or irregular domains. Note that while $\boldsymbol{C}_2$ may be more accurate, it is also computationally more burdensome, particularly for larger grids, due to the optimization of the likelihood for each $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}^{(i)}$ in (4.28) which can require multiple evaluations of $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$.

Another important phenomenon is when the adjusted likelihood becomes flat. If elements of $\boldsymbol{C}$, particularly on the diagonal, are small, the vector $\boldsymbol{C}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$ in (4.31), becomes small. Thus for a proposed $\boldsymbol{\theta}$ that is far from $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$, the adjusted parameter $\boldsymbol{\theta}^*$ will still be close to $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$. As

a consequence, the adjusted likelihood becomes flat around $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$ and the corresponding posterior will be dominated by the prior. This phenomenon can occur when $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$ is large for the $\boldsymbol{C}_2$ adjustment, for cases when the value of ρ is large relative to the domain size. Hence, in this case, we recommend the $\boldsymbol{C}_1$ adjustment.

Algorithm 1 Adjustment $\boldsymbol{C}_1$.

Require: MdWLE $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$

1: **for** $i = 1, \ldots, k$ **do**
2: Simulate $X_{\boldsymbol{s}}^{(i)} \sim p(X_{\boldsymbol{s}} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$.
3: Compute $\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{dW}}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}} | X_{\boldsymbol{s}}^{(i)})$.
4: **end for**
5: Compute $\widehat{J}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$ from (4.25).
6: Factor $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$,
7: $\boldsymbol{M}_A\boldsymbol{M}_A^\top = \boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}})\widehat{\boldsymbol{J}}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}})^{-1}\boldsymbol{H}(\widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$.
8: **Return** $\boldsymbol{C}_1 = \boldsymbol{M}^{-1}\boldsymbol{M}_A$.

Algorithm 2 Adjustment $\boldsymbol{C}_2$

Require: MdWLE $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}$.

1: **for** $i = 1, \ldots, k$ **do**
2: Simulate $X_{\boldsymbol{s}}^{(i)} \sim p(X_{\boldsymbol{s}} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{dW}})$.
3: $\widetilde{\boldsymbol{\theta}}^{(i)} = \arg\min_{\boldsymbol{\theta}\in\Theta}\{\ell_{\mathrm{dW}}(\boldsymbol{\theta}|X_{\boldsymbol{s}}^{(i)})\}$.
4: **end for**
5: Compute $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\} \approx \widehat{\boldsymbol{G}^{-1}(\boldsymbol{\theta})}$ from (4.28).
6: Factor $\boldsymbol{M}_A\boldsymbol{M}_A^\top = \widehat{\boldsymbol{G}^{-1}(\boldsymbol{\theta})}$
7: Factor $\boldsymbol{M}\boldsymbol{M}^\top = -\left[\nabla^2_{\boldsymbol{\theta}}\ell_{\mathrm{dW}}(\boldsymbol{\theta})\right]^{-1}$.
8: **Return** $\boldsymbol{C}_2 = \boldsymbol{M}\boldsymbol{M}_A^{-1}$.

Table 1: Algorithms for computing the adjustments $\boldsymbol{C}_1$ (Algorithm 1) and $\boldsymbol{C}_2$ (Algorithm 2).

4.5.4 Simulation study

We consider the proposed computational approach of Monahan and Boos (1992) to validate the coverages of the aforementioned curvature adjustments. This approach was later formalized into a software validation algorithm in Cook et al. (2006). First, simulate j independent samples from the prior, $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta})$, and for each $\boldsymbol{\theta}^{(i)}$, generate the random field $X_{\boldsymbol{s}}^{(i)} \sim p(X_{\boldsymbol{s}}|\boldsymbol{\theta}^{(i)})$ and compute the integral

$$U^{(i)} = \int_{-\infty}^{\boldsymbol{\theta}^{(i)}} p(\boldsymbol{\theta}|X_{\boldsymbol{s}}^{(i)})d\boldsymbol{\theta}, \quad \text{for } i = 1, \ldots, j. \tag{4.32}$$

A Monte Carlo estimate of the above integral $\widehat{U}^{(i)}$ is performed with samples from the posterior via the Random Walk Metropolis-Hasting algorithm. Cook et al. (2006) prove that as the number of posterior samples used to approximate (4.32) approach infinity, then $\widehat{U}^{(i)} \sim \mathrm{Unif}(0, 1)$ for each $i = 1, \ldots, j$.

Since the two proposed curvature adjustments are valid asymptotically, we construct a simulation study to verify this for increasing grid sizes, $|\mathbf{n}| \to \infty$. We consider various priors, covariance kernels and sampling schemes. The three different likelihoods considered are:

1. $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$, the un-adjusted debiased spatial Whittle likelihood.
2. $\ell^{(1)}_{\mathrm{dW}}(\boldsymbol{\theta})$, the adjusted $\boldsymbol{C}_1$ debiased spatial Whittle likelihood.
3. $\ell^{(2)}_{\mathrm{dW}}(\boldsymbol{\theta})$, the adjusted $\boldsymbol{C}_2$ debiased spatial Whittle likelihood.

We simulate $k = 250$ datasets from a two-dimensional Gaussian random field with Mátern covariance kernel for both simulation studies.

Simulation 4.1. Consider a two-dimensional Gaussian random field from a Mátern covariance kernel with $\nu \to \infty$, known as the squared-exponential covariance kernel, defined as

$$c(\boldsymbol{u}|\rho, \sigma) = \sigma^2 \exp\left(-\frac{\boldsymbol{u}^2}{2\rho^2}\right).$$

The spectral density is $2\pi\rho\exp(-2\pi^2\rho^2\boldsymbol{\omega}^2)$ (Rasmussen and Williams, 2006). This produces overly-smooth simulations due to high spectral mass at the lower frequencies, whereas negligible mass at the higher frequencies. Estimation with this kernel is difficult since the periodogram contains correlations between Fourier frequencies. We use independent Gamma priors for both parameters,

$$\rho \sim \text{Gamma}(\alpha = 60, \beta = 10),\ \sigma \sim \text{Gamma}(\alpha = 60, \beta = 50),$$

with means $\text{E}[\rho] = 6$ and $\text{E}[\sigma] = 1.2$, respectively. We also use square grids with no missing values and grid sizes $|\mathbf{n}| = (256^2, 512^2, 1024^2)$.

Figure 4.2 displays the QQ plots for each grid size and model. The un-adjusted debiased Whittle posterior yields (top row) quantiles far from uniform for all three grid sizes. This is not surprising as pseudo-likelihoods generally do not yield proper coverage. The S shape of the QQ plots suggests that the posteriors rarely cover the true parameter $\boldsymbol{\theta}_0$ due to the over-concentration of the posteriors. The $\boldsymbol{C}_1$ adjustment (middle row) and the $\boldsymbol{C}_2$ adjustment (bottom row) are similar and will be described as one. Firstly, for $|\mathbf{n}| = 256^2$, the tails (towards 0 and 1) are heavier than those of a standard uniform, suggesting that parameters are often overestimated or underestimated. Furthermore, σ performs worse than ρ. The coverages for $|\mathbf{n}| = 512^2$ are much closer to standard uniform for both parameters. This is also the case for $|\mathbf{n}| = 1024^2$, as both parameters are indistinguishable from a standard uniform.

To conclude, the un-adjusted model violates the coverage of posterior sets for the aforementioned prior for all three grid sizes and only gets marginally better as grid sizes increase. Hence, it is unsuitable for Bayesian inference. In contrast, adjustments $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ produce simulations that are close to uniform for increasing grid sizes, 512^2 and 1024^2. Thus, asymptotically (as $\mathbf{n} \to \infty$), under this prior, the adjusted spatial debiased Whittle likelihoods, $\ell^{(1)}_{\text{dW}}(\boldsymbol{\theta})$ and $\ell^{(2)}_{\text{dW}}(\boldsymbol{\theta})$, yield proper coverage posteriors.

Simulation 4.2. We consider an irregular domain shape for this simulation. The domain shape is a grid of France. Here, a grid of size $\mathbf{n} = (500, 500)$ contains observations inside the border of France, with missing values outside the border. This is challenging since roughly 62% of the observed domain are missing values. We use a squared-exponential covariance kernel with independent gamma priors, $\rho \sim \text{Gamma}(\alpha = 120, \beta = 20), \sigma \sim \text{Gamma}(\alpha = 50, \beta = 50)$ with

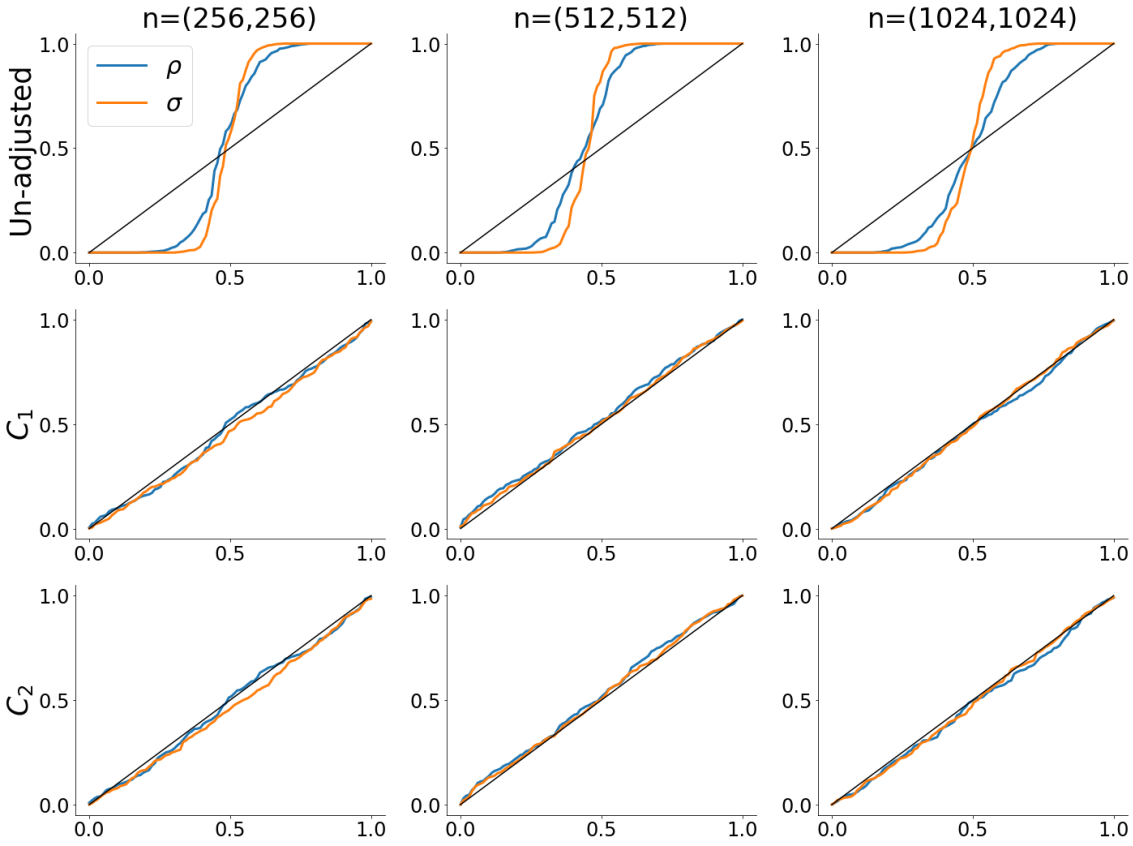


Figure 4.2: Standard uniform QQ plots for the coverage of posteriors for Gaussian random fields with independent Gamma priors.

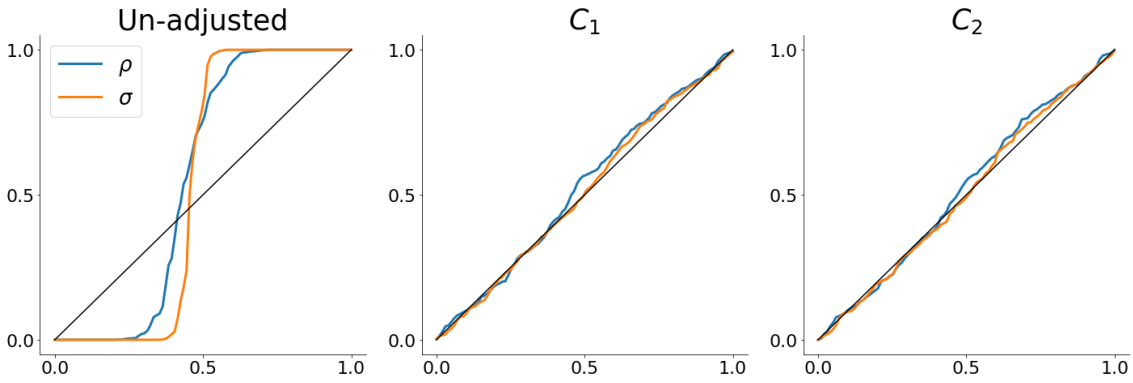


Figure 4.3: Standard uniform QQ plots for the coverage of posteriors for Gaussian random fields with independent Gamma priors, for an irregular domain shape of France.

means $\mathrm{E}[\rho] = 6$ and $\mathrm{E}[\sigma] = 1$.

The posterior quantiles are plotted in Figure 4.3. As expected, the un-adjusted debiased Whittle model fails to provide proper coverages for both parameters. The $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ adjustments are similar for both parameters, with $\boldsymbol{C}_2$ performing marginally better than $\boldsymbol{C}_1$. Both adjustments are indistinguishable from a standard uniform between $(0, 0.5)$; however, the top half interval seems to have more concentration of mass compared to the bottom interval, with ρ performing slightly worse than σ.

4.6 Applications

We illustrate our method on two data sets that are relevant in the literature and compare our approach to the standard Whittle likelihood.

4.6.1 Sea surface temperature

The first application is Tropical Rainfall Measuring Mission (TRMM) microwave imager (TMI) satellite data from the Pacific Ocean presented in Chapter 5 of Gelfand et al. (2010). Sea surface temperature (SST) data are used for climate modelling and meteorology and are essential for evaluating climate change. They are also helpful for comparison with oceanic climate models as a diagnostic tool. Identifying spatial patterns of SST is a critical factor in the formation of hurricanes in the Pacific Ocean, which strike Central America. Furthermore, the transfer of water between the northern and southern equatorial currents is an important application of the analysis of the spatial structure of SST. Quantifying spatial variability and making informed predictions about SST is crucial for research on the world's ocean and the broader climate. The data is available at www.remss.com/tmi/tmi_browse.html.

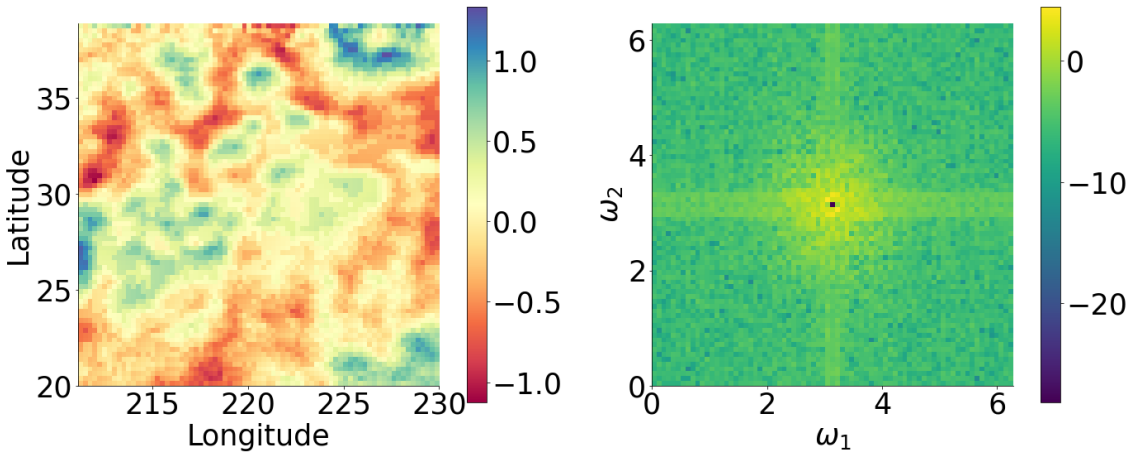


Figure 4.4: Sea surface temperatures over the Pacific Ocean. The left plots show the processed data after removing the trend. The right plot is the un-tapered log-periodogram of the data.

The SST data, in degrees Celsius, is from March 1998 and has roughly a 25km $\times$ 25km spatial resolution defined by latitude and longitude. The data are on a rectangular grid of size 75 $\times$ 75. Due to the large number of observations, maximum likelihood estimation, let alone Bayesian inference via the Gaussian likelihood, is computationally intractable. Instead, we use frequency domain methods for computationally efficient estimation and Bayesian inference for periodogram data.

To satisfy the stationarity assumption, the authors suggest a second-order polynomial mean trend be removed,

$$\beta_0 + \beta_1 \odot u(\boldsymbol{s}) + \beta_2 \odot v(\boldsymbol{s}) + \beta_3 \odot u(\boldsymbol{s})^2 + \beta_4 \odot v(\boldsymbol{s})^2 + \beta_5 \odot u(\boldsymbol{s}) \odot v(\boldsymbol{s})$$

where $u(\boldsymbol{s})$ and $v(\boldsymbol{s})$ are the longitude and latitude at each observation respectively and $\odot$ is element-wise multiplication. Figure 4.4 displays the stationary random field and its corresponding log-periodogram.

The log-periodogram in Figure 4.4 suggests that Mátern covariance kernel in (4.14) is an appropriate model. Initial optimizations were performed to obtain a sensible fixed value for the nugget parameter $\sigma_\varepsilon^2 = 10^{-10}$. We perform Bayesian inference over the joint parameter space $\boldsymbol{\theta} = (\rho, \sigma, \nu)$. This is a challenging problem as it is well known that the smoothness parameter ν is difficult to estimate due to its lack of information (De Oliveira and Han, 2022). Nonetheless, we compare three posteriors: the un-adjusted Debiased Whittle, the adjusted Debiased Whittle with $\boldsymbol{C}_2$ and the standard Whittle. Note that the $\boldsymbol{C}_1$ adjustment is not valid here as this requires the derivatives of the Mátern covariance w.r.t. the parameters, which do not always exist. We simulate $k = 500$ data sets to compute the adjustment $\boldsymbol{C}_1$ matrix. A marginal Gamma prior was

used for all three parameters, with hyper-parameters,

$$\rho \sim \text{Gamma}(\alpha = 5, \beta = 1.0),\ \sigma \sim \text{Gamma}(\alpha = 0.7, \beta = 1/0.7),\ \nu \sim \text{Gamma}(\alpha = 1.0, \beta = 2).$$

The marginal posteriors are plotted in Figure 4.5. The figure shows that the standard Whittle underestimates the ρ and σ compared to the debiased Whittle. The $\boldsymbol{C}_2$ adjustment in orange inflates the variance compared to the un-adjusted debiased Whittle in blue.

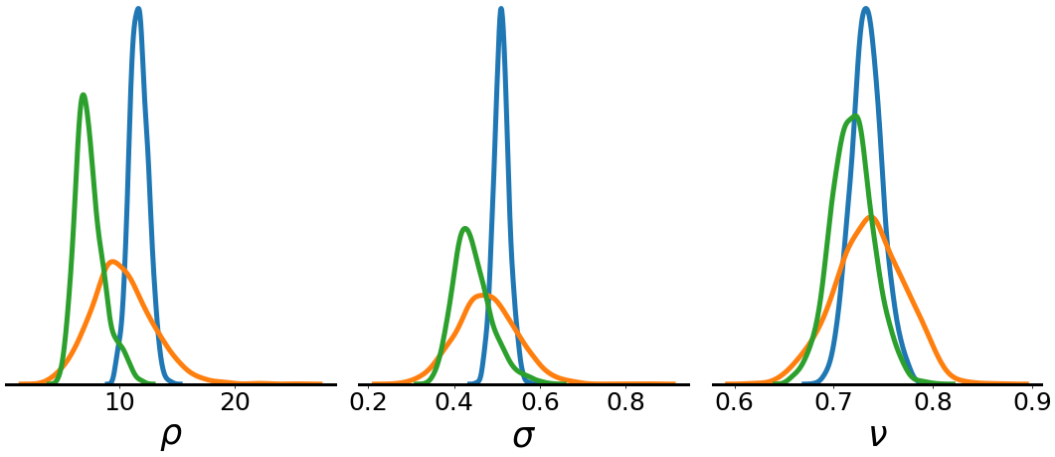


Figure 4.5: Kernel density estimates of the marginal posterior comparison of sea surface temperature data with grid size $n = (75, 75)$. The blue line is the un-adjusted debiased Whittle, the orange is the adjusted debiased Whittle with $\boldsymbol{C}_2$, and the green is the standard Whittle.

As a diagnostic check, we use the parametric model of (4.8) to define the frequency domain residual spectrum as

$$I_{\mathbf{n}}(\boldsymbol{\omega}) \ / \ \overline{I}_{\mathbf{n}}(\boldsymbol{\omega}; \boldsymbol{\theta}) \overset{\text{i.i.d.}}{\sim} \text{Exp}\{1\}, \qquad \boldsymbol{\omega} \in \Omega_{\mathbf{n}}, \tag{4.33}$$

for the $\boldsymbol{C}_2$ adjusted debiased Whittle likelihood where the division is performed element-wise. Similarly, the residuals of the standard Whittle are obtained by replacing the expected periodogram with the spectral density. Ideally, given the correct model, the residuals should be a standard exponential distribution throughout the entire observable domain. Figure 4.6 plots the debiased Whittle residuals on the left and the standard Whittle on the right using their associated posterior means. The side lobes on both panels are visible but more pronounced on the right panel. Furthermore, the values of the standard Whittle residual spectrum are more extreme, suggesting portions of the residual spectrum are not exponentially distributed. Hence, the standard Whittle is misspecified.

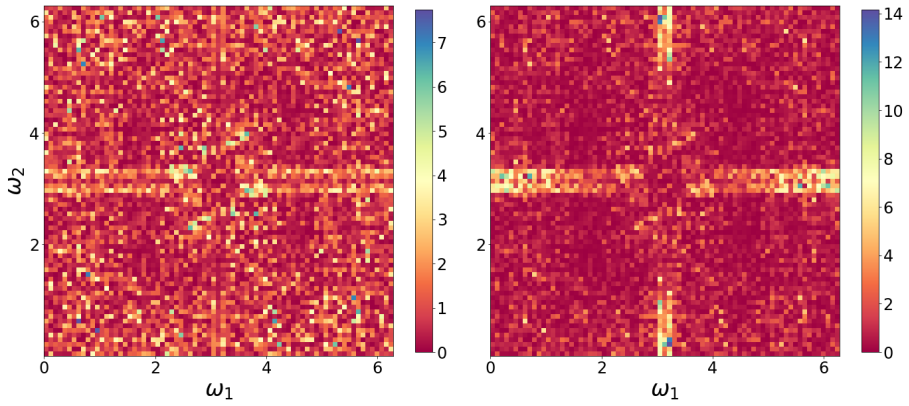


Figure 4.6: Residual spectrum: the periodogram divided by the estimated spectral density (Equation 4.33). The left plot is $\boldsymbol{C}_2$ adjusted debiased Whittle, and the right plot is the standard Whittle. The estimated spectra are based on the posterior mean.

4.6.2 Venus topography data

In this example, we consider an application of Venus topography data from Rappaport et al. (1999). Measurements of the topography and gravity field of a planet play an important role in understanding a planet's interior density structure. These measures can provide a greater understanding of the planet's thermal evaluation when combined with additional information such as surface geology (Rappaport et al., 1999). Geospatial analysis via parametric modelling of the covariance between distinct locations of extraterrestrial planets such as Venus is important in two ways. First, the parameters associated with the covariance function carry interpretable meanings of the physical phenomena. For example, in the Mátern kernel, the slope parameter ν relates to the smoothness of the terrain, the parameter σ describes the amplitude or modulation of the terrain and the range parameter ρ is associated with the distance which two spatial locations are uncorrelated, i.e. the oscillation of the process. The second is that a parametric covariance model provides a natural framework for prediction, i.e. interpolation or extrapolation, for non-observed regions.

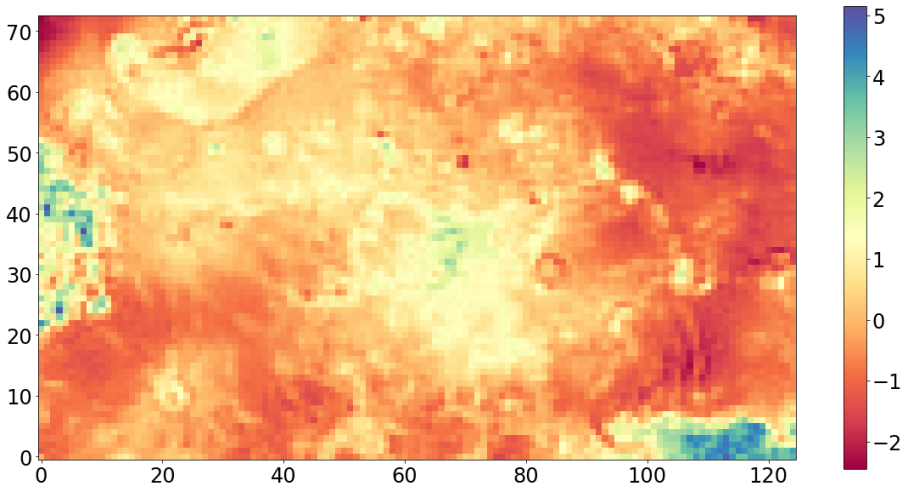


Figure 4.7: Venus topography data after standardization.

As data on Venus is not easily accessible, prediction may be the best/necessary option for a better understanding of the topography on Venus. The data, in meters, is observed on a rectangular $(73, 125)$ grid with no missing values and is displayed after standardization in Figure 4.7. Here, we analyze the same data, patch 3, as in Guillaumin et al. (2022) except in a Bayesian context. Four models are compared:

1. $\ell_{\mathrm{W}}(\boldsymbol{\theta})$, the standard Whittle likelihood.
2. $\ell_{\mathrm{dW}}(\boldsymbol{\theta})$, the un-adjusted debiased spatial Whittle likelihood.
3. $\ell^{(2)}_{\mathrm{dW}}(\boldsymbol{\theta})$, the adjusted $\boldsymbol{C}_1$ debiased spatial Whittle likelihood.
4. $\ell^{(3)}_{\mathrm{dW}}(\boldsymbol{\theta})$, the adjusted $\boldsymbol{C}_2$ debiased spatial Whittle likelihood.

We use the same prior throughout, $\rho \sim \mathrm{Gamma}(\alpha = 15, \beta = 1)$ and $\sigma \sim \mathrm{Gamma}(\alpha = 10, \beta = 10)$.

This data set is challenging due to the long-range dependence compared to the domain size. We consider an exponential covariance kernel to relieve identifiability issues related to the smoothness of the process. To find an appropriate value for the smoothness parameter, we maximize the debiased Whittle likelihood and found $\widehat{\nu} = 0.55$ and set $\sigma_\epsilon = 10^{-10}$. We set $k = 500$ for both adjustments $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. The maximum debiased Whittle likelihood estimate is $\widehat{\boldsymbol{\theta}}_{\mathrm{dW}} = (24.6012, 1.557)$ for ρ and σ respectively.

Obtaining a reasonable $\boldsymbol{C}_2$ adjustment proved difficult since the variance of the Monte Carlo estimate of $\mathrm{Var}\{\widehat{\boldsymbol{\theta}}_{\mathrm{dW}}\}$ is large. This is due to outliers in the simulated maximum likelihood estimates. This stems from larger values of ρ compared to the domain size; to alleviate this, 10% of the largest maximum likelihood estimates of ρ were removed.

The kernel density estimates of the marginal posteriors are plotted in Figure 4.8. Along with adjusting the curvature of the posterior, the $\boldsymbol{C}_2$ adjustment also adjusted the mean of the posterior. The $\text{Var}\{\widehat{\boldsymbol{\theta}}_{\text{dW}}\}$ was still large enough to make the adjusted likelihood flat. As a result, the prior has a large contribution to the adjusted posterior, as mentioned in Section 4.5.3. Likewise, for the adjusted $\boldsymbol{C}_1$ posterior, its mean also shifted to a lesser extent than the abovementioned adjustment due to the magnitude of covariance of the score function in (4.25). Despite this, the two curvature-adjusted posteriors exhibit a higher variance than the un-adjusted debiased Whittle posterior, which was shown to have improper coverages in Section 4.5.4.

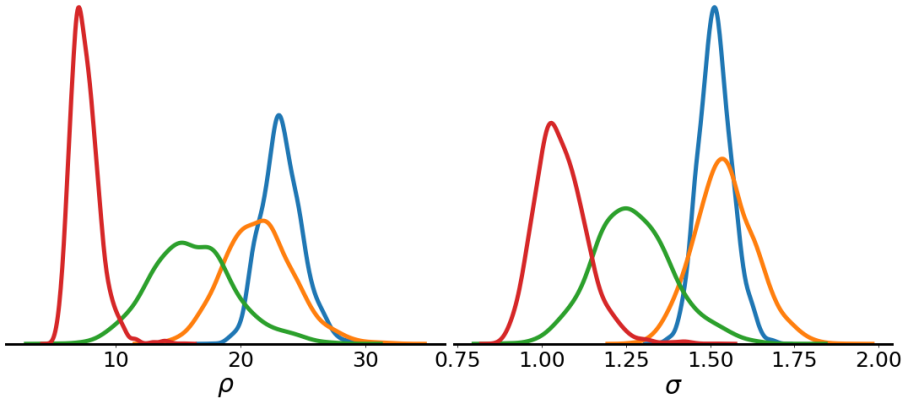


Figure 4.8: Kernel density estimates of the marginal posterior for Venus topography data. The un-adjusted debiased Whittle in blue, the adjusted $\boldsymbol{C}_1$ debiased Whittle in orange, the adjusted $\boldsymbol{C}_2$ in green and the standard Whittle in red.

4.7 Conclusion and discussion

This paper investigates Bayesian inference for covariance-stationary random fields for the debiased spatial Whittle likelihood. The debiased Whittle likelihood for frequentist parameter estimation has many benefits discussed in Guillaumin et al. (2022) as opposed to other estimation domain methods (e.g. Whittle (1954), Gelfand et al. (2010)). Not the least of which is the computationally efficiency of the debiased Whittle likelihood via the Fast Fourier Transform.

We extend the debiased Whittle likelihood to construct an asymptotically proper coverage likelihood suitable for Bayesian inference. Initially introduced in Monahan and Boos (1992), proper likelihoods for inference satisfy the Bayesian credible interval in (4.17) for every level of coverage α. They propose a simulation-based algorithm, as mirrored in Cook et al. (2006) to validate the posterior coverages by assessing the uniformity of the quantiles of the true generating parameter, which we assess in a simulation study in Section 4.5.4.

Leveraging the fact that the debiased spatial Whittle likelihood falls within the framework

of a pseudo-likelihood in finite samples, we use ideas from Ribatet et al. (2012) to construct posterior curvature adjustments that asymptotically satisfy the Bernstein Von-Mises theorem for the subsequent adjusted debiased Whittle likelihood. Unfortunately, the quantities needed to compute the posterior adjustments are not computationally tractable for the debiased spatial Whittle likelihood; hence, we rely on estimating the quantities mentioned above to propose two unique adjustments $\boldsymbol{C}_1, \boldsymbol{C}_2$. The former adjustment works by computing the variance of the score function, which is more robust to processes with large length scales; however, this adjustment relies on the derivative of the covariance function, which may not be analytically available for general Mátern class covariance functions. The latter adjustment estimates the adjustment via a simulation approximation of the sampling distribution of the MDWLE and the observed Fisher information, which provides a tailored adjustment useful in smaller grid sizes.

Our method can be used for non-Gaussian data with a specified model (by simulating the non-Gaussian field) using the same adjustments. This is because frequency domain techniques are generally more robust to non-Gaussian data-generating processes; only the real and imaginary parts of the periodogram are assumed to be asymptotically normal (e.g. Peligrad and Zhang (2019)).

This work can also be applied as a particular case for time series, $d = 1$, studied in Sykulski et al. (2019). For series that don't satisfy the assumptions of the parametric Whittle model, the Bayesian debiased Whittle likelihood can be a useful computationally efficient alternative.

Estimation techniques for large spatial data are at the forefront of statistical research. For ultra-large spatial data, spectral subsampling MCMC (Quiroz et al., 2019; Salomone et al., 2020) is an attractive alternative if the log-likelihood function can be computed independently for each data point. As it stands, this is not possible/inefficient for the debiased Whittle likelihood since the expected periodogram is computed on the whole grid of Fourier frequencies via the FFT or in the frequency domain by a convolution that requires the whole spectrum. Future research will also focus on extending non-latticed cases when the data are irregularly spaced.

References

Akaike, H. (1973). Block Toeplitz matrix inversion. *SIAM Journal on Applied Mathematics*, 24(2):234–241.

Anitescu, M., Chen, J., and Stein, M. L. (2017). An inversion-free estimating equations approach for Gaussian process models. *Journal of Computational and Graphical Statistics*, 26(1):98–107.

Bandyopadhyay, S. and Lahiri, S. N. (2009). Asymptotic properties of discrete Fourier transforms for spatial data. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 71:221–259.

Berliner, L. M., Wikle, C. K., and Cressie, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate*, 13(22):3953–3968.

Best, N., Ickstadt, K., Wolpert, R., and Briggs, D. (2001). Combining models of health and exposure data: the SAVIAH study. In *Spatial Epidemiology: Methods and Applications*, pages 393–414. Oxford University Press.

Bevilacqua, M. and Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing*, 25:877–892.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods*. Springer Science & Business Media.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692.

Cressie, N. (1989). Geostatistics. *The American Statistician*, 43(4):197–202.

Dahlhaus, R. and Künsch, H. (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, 74(4):877–882.

De Oliveira, V. and Han, Z. (2022). On information about covariance parameters in Gaussian Mátern random fields. *Journal of Agricultural, Biological and Environmental Statistics*, 27(4):690–712.

Dietrich, C. R. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107.

Frazier, D. T., Drovandi, C., and Kohn, R. (2023). Calibrated Generalized Bayesian Inference. *arXiv e-prints*, pages arXiv–2311.

Gelfand, A., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press.

Geoga, C. J., Marin, O., Schanen, M., and Stein, M. L. (2023). Fitting matérn smoothness parameters using automatic differentiation. *Statistics and Computing*, 33(2):48.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467):799–804.

Guillaumin, A. P., Sykulski, A. M., Olhede, S. C., and Simons, F. J. (2022). The debiased spatial Whittle likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1526–1557.

Guilleminot, J. (2020). Modeling non-Gaussian random fields of material properties in multiscale mechanics of materials. In *Uncertainty Quantification in Multiscale Materials Modeling*, pages 385–420. Elsevier.

Guinness, J. (2019). Spectral density estimation for random fields via periodic embeddings. *Biometrika*, 106(2):267–286.

Guinness, J. and Fuentes, M. (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *Journal of Computational and Graphical Statistics*, 26(1):88–97.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1):95–105.

Heyde, C. C. (1997). *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer.

Hrafnkelsson, B. and Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics*, 10:179–200.

Kent, J. T. and Mardia, K. V. (1996). Spectral and circulant approximations to the likelihood for stationary Gaussian random fields. *Journal of Statistical Planning and Inference*, 50(3):379–394.

Lee, J. E., Nicholls, G. K., and Ryder, R. J. (2019). Calibration Procedures for Approximate Bayesian Credible Sets. *Bayesian Analysis*, 14(4):1245–1269.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

Matsuda, Y. and Yajima, Y. (2009). Fourier analysis of irregularly spaced data on R^d. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):191–217.

Monahan, J. F. and Boos, D. D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Peligrad, M. and Zhang, N. (2019). Central limit theorem for Fourier transform and periodogram of random fields. *Bernoulli*, 25(1):499–520.

Percival, D. B. and Walden, A. T. (1993). *Spectral Analysis for Physical Applications*. Cambridge University Press.

Prangle, D., Blum, M. G., Popovic, G., and Sisson, S. (2014). Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, 56(4):309–329.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843.

Rappaport, N. J., Konopliv, A. S., Kucinskas, A. B., and Ford, P. G. (1999). An improved 360 degree and order model of Venus topography. *Icarus*, 139(1):19–31.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Salomone, R., Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2020). Spectral subsampling MCMC for stationary time series. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8449–8458.

Simons, F. J. and Olhede, S. C. (2013). Maximum-likelihood estimation of lithospheric flexural rigidity, initial-loading fraction and load correlation, under isotropy. *Geophysical Journal International*, 193(3):1300–1342.

Sowell, F. (1989). A decomposition of block Toeplitz matrices with applications to vector time series. *Unpublished manuscript*.

Stein, M. (2012). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York.

Stein, M. L., Chen, J., and Anitescu, M. (2013). Stochastic approximation of score function for Gaussian procceses. *The Annals of Applied Statistics*, 7:1162–1191.

Stroud, J. R., Stein, M. L., and Lysen, S. (2017). Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice. *Journal of Computational and Graphical Statistics*, 26(1):108–120.

Sykulski, A. M., Olhede, S. C., Guillaumin, A. P., Lilly, J. M., and Early, J. J. (2019). The debiased Whittle likelihood. *Biometrika*, 106(2):251–266.

Tolbert, P. E., Mulholland, J. A., Macintosh, D. L., et al. (2000). Air quality and pediatric emergency room visits for asthma and Atlanta, Georgia. *American Journal of Epidemiology*, 151(8):798–810.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR.

Chapter 5

Conclusion and future research

This thesis investigates Bayesian inference via frequency domain methodologies for stationary time series models and spatial data. We employ the Whittle likelihood (Whittle, 1953) or variants thereof (Guillaumin et al., 2022), to approximate the likelihood function of a stationary process with a covariance function governed by unknown parameters. Frequency domain estimation methods are generally asymptotically equivalent to their time domain counterparts but trade computational efficiency for bias in parameter estimation in finite samples. At its heart, frequency domain estimation techniques rely on the central limit theorem for the Discrete Fourier transform (DFT), studied in Peligrad and Wu (2010); Peligrad and Zhang (2019); Shao and Wei (2007). Loosely speaking, the central limit theorem for the DFT guarantees asymptotic normality for the real and imaginary parts of the DFT for stationary processes. Furthermore, the DFT of the process at each frequency are asymptotically independent. For the applications considered in this thesis, we show that the bias incurred by the frequency domain approximation is negligible or consider methods that explicitly reduce bias (Guillaumin et al., 2022).

Chapter 2 studies dynamic regression models with semi-long memory error terms. Here, we model the response variable as a linear combination of exogenous stationary predictors with an ARTFIMA process for the error term to capture the leftover, unexplained semi-long memory in the residuals. Here, the ARTFIMA model generalizes the well-known ARFIMA and ARIMA models by introducing tempering, a way of incorporating semi-long memory into an ARMA process while regularising the slow decay of the ARFIMA covariance function. We propose a fast, asymptotically exact Whittle likelihood method to fit DLR models with ARTFIMA errors. We show that this model can improve forecasts compared to the DLR with ARFIMA errors with applications in electricity demand.

Chapter 3 considers general Lévy-driven continuous-time ARMA models. Here, Bayesian inference is performed via spectral subsampling on the Whittle likelihood with the aliased spectral density. This approach relies heavily on large data to satisfy the asymptotic normal of the DFT.

We first demonstrate asymptotic normality in a simulated study for Gaussian and two non-Gaussian-driven CARMA models. We show that the Whittle likelihood approximates the Kalman filter likelihood well for Gaussian CARMA models. Subsampling was performed on Gaussian and non-Gaussian-driven CARMA processes for simulated data, and for both cases, subsampling led to a significant decrease in relative computation time. We considered an application with minute Bitcoin prices and showed, on average, a 100x increase in computational efficiency compared to the standard Whittle approach.

Chapter 4 investigates the Bayesian inference of random fields via the debiased spatial Whittle likelihood (Guillaumin et al., 2022). It is well known that frequency-domain estimation methods can cause substantial bias, particularly for $d = 2, d = 3$, (Dahlhaus and Künsch, 1987). The debiased Whittle likelihood relies on the expected periodogram, as opposed to the usual spectral density. This approach results in favourable asymptotic properties of point estimates and is still computationally efficient as it employs the FFT. We propose posterior curvature adjustments based on previous work Ribatet et al. (2012) to perform Bayesian inference to obtain proper posteriors based on coverages of posterior sets. We show in a simulation study that these adjustments satisfy posterior coverages with increasing grid sizes and for different domain shapes. We apply our method to two real-world datasets, sea surface temperature and Venus topography data.

Future research will extend the frequency domain methodology to multivariate dynamic regression models, yielding multi-dimensionality in the response or explanatory predictors or error process. Additionally, modelling the response as a non-linear transformation of the exogenous predictors is a possible extension. For continuous-time models, a natural extension to CARMA models considered in Chapter 3 is the fractionally integrated CARMA (CARFIMA) model proposed in Brockwell and Marquardt (2005). Furthermore, Whittle estimation and spectral subsampling of irregularly sampled CARMA models can be considered due to modified versions of the Fourier Transform in Fechner and Stelzer (2018). Finally, due to its computational efficiency, a debiased Whittle approach would be appealing for large, irregularly spaced spatial data.

References

Brockwell, P. J. and Marquardt, T. (2005). Lévy-driven and fractionally integrated ARMA processes with continuous time parameter. *Statistica Sinica*, 15:477–494.

Dahlhaus, R. and Künsch, H. (1987). Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, 74(4):877–882.

Fechner, Ż. and Stelzer, R. (2018). Limit behaviour of the truncated pathwise Fourier-

transformation of Lévy-driven CARMA processes for non-equidistant discrete time observations. *Statistica Sinica*, 28(3):1633–1650.

Guillaumin, A. P., Sykulski, A. M., Olhede, S. C., and Simons, F. J. (2022). The debiased spatial Whittle likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1526–1557.

Peligrad, M. and Wu, W. B. (2010). Central limit theorem for Fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022.

Peligrad, M. and Zhang, N. (2019). Central limit theorem for Fourier transform and periodogram of random fields. *Bernoulli*, 25(1):499–520.

Ribatet, M., Cooley, D., and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2):813–845.

Shao, X. and Wei, B. W. (2007). Asymptotic spectral theory for nonlinear time series. *Annals of Statistics*, 35(4):1773–1801.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434.