Efficient Bayesian inference for ARFIMA processes

Timothy Graves* Robert B. Gramacy† Christian Franzke‡
Nicholas Watkins§

Abstract

In forecasting problems it is important to know whether or not recent events represent a regime change (low long-term predictive potential), or rather a local manifestation of longer term effects (potentially higher predictive potential). Mathematically, a key question is about whether the underlying stochastic process exhibits "memory", and if so whether the memory is "long" in a precise sense. Being able to detect or rule out such effects can have a profound impact on speculative investment (e.g., in financial markets) and inform public policy (e.g., characterising the size and timescales of the earth system's response to the anthropogenic CO_2 perturbation). Most previous work on inference of long memory effects is frequentist in nature. Here we provide a systematic treatment of Bayesian inference for long memory processes via the Autoregressive Fractional Integrated Moving Average (ARFIMA) model. In particular, we provide a new approximate likelihood for efficient parameter inference, and show how nuisance parameters (e.g., short memory effects) can be integrated over in order to focus on long memory parameters and hypothesis testing more directly than ever before. We illustrate our new methodology on both synthetic and observational data, with favorable comparison to the standard estimators.

Key words: long-range dependence, auto-regressive models, moving average models, ARFIMA, Metropolis–Hastings, reversible jump

1 Introduction

In this paper we are concerned with Bayesian analysis of specific types of stochastic processes capable of possessing 'long memory', or "long-range dependence" (LRD) (Beran, 1994b; Palma, 2007; Beran et al., 2013). Long memory is the notion of there being correlation between the present and *all* points in the past. A standard definition is that a (finite variance, stationary) process has *long memory* if its autocorrelation function (ACF) has

[*] URS Corporation, London, UK

[†] Corresponding author: The University of Chicago Booth School of Business; 5807 S. Woodlawn Avenue, Chicago, IL 60637; rbgramacy@chicagobooth.edu

[‡] Meteorologisches Institut, University of Hamburg, Germany

[§] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany; Centre for Complexity and Design, Open University, Milton Keynes, UK; Centre for the Analysis of Time Series, London School of Economics and Political Science, London, UK; and Centre for Fusion Space and Astrophysics, University of Warwick, Coventry, UK.

power-law decay: $\rho(\cdot)$ such that $\rho(k) \sim c_\rho \, k^{2d-1}$ as $k \to \infty$, for some non-zero constant c_ρ, and where $0 < d < \frac{1}{2}$. The parameter d is the memory parameter; if $d = 0$ the process does not exhibit long memory, while if $-\frac{1}{2} < d < 0$ the process is said to have *negative* memory.

The study of long memory originated in the 1950s in the field of hydrology, where studies of the levels of the river Nile (Hurst, 1951) demonstrated anomalously fast growth of the rescaled range of the time series. After protracted debates[1] about whether this was a transient (finite time) effect, the mathematical pioneer Benoît B. Mandelbrot showed that if one retained the assumption of stationarity, novel mathematics would then be essential to sufficiently explain the Hurst effect. In doing so he rigorously defined (Mandelbrot and Van Ness, 1968; Mandelbrot and Wallis, 1968) the concept of long memory.

Most research into long memory and its properties has been based on classical statistical methods, spanning parametric, semi-parametric and non-parametric modeling (see Beran et al., 2013, for a review). Very few Bayesian methods have been studied, most probably due to computational difficulties. The earliest works are parametric and include Koop et al. (1997) Pai and Ravishanker (1998), and Hsu and Breidt (2003). If computational challenges could be mitigated, the Bayesian paradigm would offer advantages over classical methods including flexibility in specification of priors (i.e., physical expertise could be used to elicit an informative prior). It would offer the ability to marginalise out aspects of a model apparatus and data, such as short memory or seasonal effects and missing observations, so that statements about long memory effects can be made unconditionally.

Towards easing the computational burden, we focus on the ARFIMA class of processes (Granger and Joyeux, 1980; Hosking, 1981) as the basis of developing a systematic and unifying Bayesian framework for modeling a variety of common time series phenomena, with particular emphasis on detecting potential long memory effects. ARFIMA has become very popular in statistics and econometrics because it is generalisable and its connection to the ARMA family (and to fractional Gaussian noise) is relatively transparent. A key property of ARFIMA is its ability to simultaneously yet separately model long and short memory. Both Liseo et al. (2001) and Holan et al. (2009) argued, echoing a sentiment in the classical literature, that full parametric long memory models (like ARFIMA) are 'too hard' to work with. Furthermore, often d is the only object of real interest, and consideration of a single class of models, such as ARFIMA, is too restrictive. They therefore developed methods which have similarities to classical periodograms.

We think ARFIMA deserves another look—that many of the above drawbacks, to ARFIMA in particular and Bayesian computation more generally, can be addressed with a careful treatment. We provide a new approximate likelihood for ARFIMA processes that can be computed quickly for repeated evaluation on large time series, and which underpins an efficient MCMC scheme for Bayesian inference. Our sampling scheme can be best described as a modernisation of a blocked MCMC scheme proposed by Pai and Ravishanker (1998)—adapting it to the approximate likelihood and extending it to handle a richer form of (known) short memory effects. We then further extend the analysis to the case where the short memory form is unknown, which requires transdimensional MCMC. This aspect is similar to the

[1] For a detailed exposition of this period of mathematical history, see Graves et al. (2014).

work of Ehlers and Brooks (2008) who considered the simpler ARIMA model class, and to Holan et al. (2009) who worked with a nonparametric long memory process. Our contribution has aspects in common with Eğrioğlu and Günay (2010) who presented a more limited method focused on model selection rather than averaging. The advantage of averaging is that the unknown form of short memory effects can be integrated out, focusing on long-memory without conditioning on nuisance parameters.

The aim of this paper is to introduce an efficient Bayesian algorithm for the inference of the parameters of the ARFIMA(p, d, q) model, with particular emphasis on the LRD parameter d. Our Bayesian inference algorithm has been designed in a flexible fashion so that, for instance, the innovations can come from a wide class of different distributions; e.g., α-stable or t-distribution. The remainder of the paper is organised as follows. Section 2 summarises of ARFIMA required for our purposes. Section 3 discusses the important numerical calculation of likelihoods, representing a hybrid between earlier classical statistical methods, and our new contributions towards a full Bayesian approach. Section 4 describes our proposed Bayesian framework and methodology method in detail, focusing on long-memory only. Then, in Section 5, we consider extensions for additional short memory. Empirical illustration and comparison of all methods is provided in Section 6. The paper concludes with a discussion in Section 7 focused on potential for further extension.

2 Time series definitions and the ARFIMA model

Following (Brockwell and Davis, 1991) a *time series* will mean a set of univariate real-valued observations $\{x_t\}$, each recorded at a specified time $t \in \mathbb{Z}$, and sampled at discrete, regular, intervals. A *process* will refer to a corresponding set of random variables $\{X_t\}$. The process $\{X_t\}$ is *strictly* stationary if the joint distributions $(X_{t_1}, \dots, X_{t_k})^\top$ and $(X_{t_1+h}, \dots, X_{t_k+h})^\top$ are the same for all positive integers k, and for all $t_1, \dots, t_k, h \in \mathbb{Z}$. It is *weakly* stationary if: (1) $\mathbb{E}X_t = \mu < \infty$ for all $t \in \mathbb{Z}$; (2) $\mathbb{E}|X_t|^2 < \infty$ for all $t \in \mathbb{Z}$; and (3) $\mathbb{C}\mathrm{ov}(X_r, X_s) = \mathbb{C}\mathrm{ov}(X_{r+t}, X_{s+t})$ for all $r, s, t \in \mathbb{Z}$. A process $\{X_t\}$ is Gaussian if the distribution of $(X_{t_1}, \dots, X_{t_k})^\top$ is multivariate normal (MVN) for all positive integers k, and for all $t_1, \dots, t_k \in \mathbb{Z}$. Throughout, stationary Gaussian processes will be assumed for convenience, where 'strong' and 'weak' are equivalent and consequently those qualifiers will be dropped.

From the above, we see that the covariance depends only on the temporal difference which motivates defining an autocovariance *ACV* $\gamma(\cdot)$ of a weakly stationary process as $\gamma(k) = \mathrm{Cov}(X_t, X_{t+k})$, where k is referred to as the (time) 'lag'. The (normalised) autocorrelation function *ACF* $\rho(\cdot)$ is defined as: $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$.

Another useful time domain tool is the 'backshift' operator $\mathcal{B}$, where $\mathcal{B}X_t = X_{t-1}$, and powers of $\mathcal{B}$ are defined iteratively: $\mathcal{B}^k X_t = \mathcal{B}^{k-1}(\mathcal{B}X_t) = \mathcal{B}^{k-1}X_{t-1} = \cdots = X_{t-k}$. A stationary process $\{X_t\}$ is said to be *causal* if there exists a sequence of coefficients $\{\psi_k\}$, with finite total mean square $\sum_{k=0}^{\infty} \psi_k^2 < \infty$ such that for all t, a given member of the process can be expanded as a power series in the backshift operator acting on the 'innovations', $\{\varepsilon_t\}$:

$$X_t = \Psi(\mathcal{B})\varepsilon_t, \quad \text{where } \Psi(z) = \sum_{k=0}^{\infty} \psi_k z^k. \tag{1}$$

The innovations are a white (i.e. stationary, zero mean, iid) noise process with variance σ^2. Causality specifies that for every t, X_t can only depend on the past and present values of the innovations $\{\varepsilon_t\}$. Furthermore Wold's theorem shows that any purely non-deterministic stationary process has a unique causal representation (referred to as the Wold expansion).

A stationary process $\{X_t\}$ is said to be *invertible* if there exists a sequence of coefficients $\{\pi_k\}$ such that $\sum_{k=0}^{\infty} \pi_k^2 < \infty$, allowing innovations to be written as a power series

$$\varepsilon_t = \Pi(\mathcal{B})X_t, \quad \text{where } \Pi(z) = \sum_{k=0}^{\infty} \pi_k z^k. \tag{2}$$

The expansion in (2) has many uses, but an additional reason for assuming invertibility is that it is closely related to identifiability—it is possible for two different processes to have the same ACF, however this cannot happen for two *invertible* ones. Therefore in what follows we restrict ourselves to models that are causal (and hence stationary) and in addition invertible.

A process $\{X_t\}$ is said to be an *auto-regressive process of order p*, AR(p), if for all t:

$$\Phi(\mathcal{B})X_t = \varepsilon_t, \quad \text{where} \quad \Phi(z) = 1 + \sum_{k=1}^{p} \phi_k z^k, \quad \text{and} \quad (\phi_1, \ldots, \phi_p) \in \mathbb{R}^p. \tag{3}$$

AR(p) processes are invertible, stationary and causal if and only if $\Phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. From (2) invertibility is equivalent to the process having an AR(∞) representation. Similarly, $\{X_t\}$ is said to be a *moving average process of order q*, MA(q), if

$$X_t = \Theta(\mathcal{B})\varepsilon_t, \quad \text{where} \quad \Theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k, \quad \text{and} \quad (\theta_1, \ldots, \theta_p) \in \mathbb{R}^q, \tag{4}$$

for all t.[2] MA(q) processes are stationary and causal, and are invertible if and only if $\Theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$.

A natural extension of the AR and MA classes arises by combining them (Box and Jenkins, 1970). The process $\{X_t\}$ is said to be an *auto-regressive moving average (ARMA) process* process of orders p and q, ARMA(p, q), if for all t:

$$\Phi(\mathcal{B})X_t = \Theta(\mathcal{B})\varepsilon_t. \tag{5}$$

Although there is no simple closed form for the ACV of an ARMA process with arbitrary p and q, so long as the process is causal and invertible, then $|\rho(k)| \leq Cr^k$, for $k > 0$, i.e., it

[2] Many authors define $\Phi(z) = 1 - \sum \phi_k z^k$. Our version emphasises connections between Φ and (3–4).

decays exponentially fast. In other words, although correlation between nearby points may be high, dependence between distant points is negligible.

Before turning to 'long memory', we require one further result. Under some extra conditions, stationary processes with ACV $\gamma(\cdot)$ possess a spectral density function (SDF) $f(\cdot)$ defined such that: $\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda)\, d\lambda$, $\forall k \in \mathbb{Z}$. This can be inverted to obtain an explicit expression for the SDF (e.g. Brockwell and Davis, 1991, §4.3): $f(\lambda) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}$, where $-\pi \leq \lambda \leq \pi$.[3] Finally, the SDF of an ARMA process is

$$f(\lambda) = \frac{\sigma^2}{2\pi}\frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}, \qquad 0 \leq \lambda \leq \pi. \tag{6}$$

The restriction $|d| < \frac{1}{2}$ is necessary to ensure stationarity; clearly if $|d| \geq \frac{1}{2}$ the ACF would not decay. The continuity between stationary and non-stationary processes around $|d| = \frac{1}{2}$ is similar to that which occurs for AR(1) process with $|\phi_1| \to 1$ (such processes are stationary for $|\phi_1| < 1$, but the case $|\phi_1| = 1$ is the non-stationary random-walk).

There are a number of alternative definitions of LRD, one of which is particularly useful, as it considers the frequency domain: A stationary process has long memory when its SDF follows $f(\lambda) \sim c_f \lambda^{-2d}$, as $\lambda \to 0^+$ for some positive constant c_f, and where $0 < d < \frac{1}{2}$. Similarly, it is said to have *negative* memory if that relationship holds for $-\frac{1}{2} < d < 0$.

The simplest way of *creating* a process which exhibits long memory is through the SDF. Consider $f(\lambda) = |1 - e^{i\lambda}|^{-2d}$, where $0 < |d| < \frac{1}{2}$. By simple algebraic manipulation, this is equivalently $f(\lambda) = \left(2\sin\frac{\lambda}{2}\right)^{-2d}$, from which we deduce that $f(\lambda) \sim \lambda^{-2d}$ as $\lambda \to 0^+$. Therefore, assuming stationarity, the process which has this SDF (or any scalar multiple of it) is a long memory process. More generally, a process having spectral density

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|1 - e^{i\lambda}\right|^{-2d}, \qquad 0 < \lambda \leq \pi. \tag{7}$$

is called *fractionally integrated* with memory parameter d, FI(d) (Barnes and Allan, 1966; Adenstedt, 1974). The full trichotomy of negative, short, and long memory is determined solely by d. When $d = 0$, the SDF is flat, yielding white noise.

In practice this model is of limited appeal to time series analysts because the entire memory structure determined by just one parameter, d. One often therefore generalises by taking any short memory SDF $f^*(\cdot)$, and defining a new SDF: $f(\lambda) = f^*(\lambda)\left|1 - e^{i\lambda}\right|^{-2d}$, $0 \leq \lambda \leq \pi$. An obvious class of short memory processes to use this way is ARMA. Taking f^* from (6) yields so-called auto-regressive fractionally integrated moving average process with parameter d, and orders p and q (ARFIMA(p, d, q)), having SDF:

$$f(\lambda) = \frac{\sigma^2}{2\pi}\frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}|1 - e^{i\lambda}|^{-2d}, \qquad 0 \leq \lambda \leq \pi. \tag{8}$$

Choosing $p = q = 0$ recovers FI(d) $\equiv$ ARFIMA($0, d, 0$).

[3] Since ACV of a stationary process is an even function of lag, the above equation implies that the associated SDF is an even function. One therefore only needs to be interested positive arguments: $0 \leq \lambda \leq \pi$.

Practical utility from the perspective of (Bayesian) inference demands finding a representation in the temporal domain. To obtain this, consider the operator $(1-\mathcal{B})^d$ for real $d > -1$, which is formally defined using the generalised form of the binomial expansion (Brockwell and Davis, 1991, Eq. 13.2.2):

$$(1-\mathcal{B})^d =: \sum_{k=0}^{\infty} \pi_k^{(d)} \mathcal{B}^k, \qquad \text{where} \qquad \pi_k^{(d)} = (-1)^k \frac{1}{\Gamma(k+1)} \frac{\Gamma(d+1)}{\Gamma(d-k+1)}. \tag{9}$$

From this observation, one can show that $X_t = (1-\mathcal{B})^{-d} Z_t$, where $\{Z_t\}$ is an ARMA process, has SDF (8). The operator $(1-\mathcal{B})^d$ is called the 'fractional differencing' operator since it allows a degree of differencing between zeroth and first order. The process $\{X_t\}$ is fractionally 'inverse-differenced', i.e. it is an 'integrated' process. The operator is used to redefine both the ARFIMA$(0, d, 0)$ and more general ARFIMA(p, d, q) processes in the time domain. A process $\{X_t\}$ is an ARFIMA$(0, d, 0)$ process if for all t: $(1-\mathcal{B})^d X_t = \varepsilon_t$. Likewise, a process $\{X_t\}$ is an ARFIMA(p, d, q) process if for all t: $\Phi(\mathcal{B})(1-\mathcal{B})^d X_t = \Theta(\mathcal{B})\varepsilon_t$, where Φ and Θ are given in (3) and (4) respectively.

Finally, to connect back to our first definition of long memory, consider the ACV of the ARFIMA$(0, d, 0)$ process. By using the definition of spectral density to directly integrate (7), and an alternative expression for $\pi_k^{(d)}$ in (9)

$$\pi_k^{(d)} = \frac{1}{\Gamma(k+1)} \frac{\Gamma(k-d)}{\Gamma(-d)}, \tag{10}$$

one can obtain the following representation of the ACV of the ARFIMA$(0, d, 0)$ process:

$$\gamma_d(k; \sigma) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}. \tag{11}$$

Because the parameter σ^2 is just a scalar multiplier, we may simplify notation by defining $\gamma_d(k) = \gamma_d(k; \sigma)/\sigma^2$, whereby $\gamma_d(\cdot) \equiv \gamma_d(\cdot; 1)$. Then the ACF is:

$$\rho_d(k) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}, \tag{12}$$

from which Stirling's approximation gives $\rho_d(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1}$, confirming a power-law relationship for the ACF. Finally, note that (10) can be used to represent ARFIMA$(0, d, 0)$ as an AR(∞) process, as $X_t + \sum_{k=1}^{\infty} \pi_k^{(d)} X_{t-k} = \varepsilon_t$. And noting that $\psi_k^{(d)} = \pi_k^{(-d)}$, leads to the following MA(∞) analog: $X_t = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+1)} \frac{\Gamma(k+d)}{\Gamma(d)} \varepsilon_{t-k}$.

3 Likelihood evaluation for Bayesian inference

For now we restrict our attention to (a Bayesian) analysis of an ARFIMA$(0, d, 0)$ process, having no short-ranged ARMA components, placing emphasis squarely on the memory parameter d. We present two alternative likelihoods, 'exact' and 'approximate'. The exact one is not original, but is presented here to highlight some important (particularly computational) issues that prevent effective use in a Bayesian context where MCMC inference requires thousands of evaluations. The approximate one represents a novel contribution.

3.1 Exact likelihood calculation

For Gaussian processes, all information is contained in the covariance structure, so inference about memory behaviour only can proceed through the covariance matrix Σ given σ and d: $\Sigma(\sigma, d)_{(i,j)} = \sigma^2\gamma_d(i-j)$, where $\gamma_d(\cdot) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}$. Therefore, the vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is MVN with mean $\mu\mathbf{1}_n$ and covariance $\Sigma(\sigma, d)$, so the likelihood is:

$$L(\mathbf{x}|\mu, \sigma, d) = (2\pi)^{-\frac{n}{2}}\{\det[\Sigma(\sigma, d)]\}^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu\mathbf{1}_n)^t\Sigma(\sigma, d)^{-1}(\mathbf{x}-\mu\mathbf{1}_n)\right].$$

To simplify the development below, write $\sigma^2\Sigma_d$ as a shorthand for $\Sigma(\sigma, d)$, whereby we have $\det[\Sigma(\sigma, d)] = \sigma^{2n}\det(\Sigma_d)$. Also, denote the quadratic term as: $Q(\mathbf{x}|\mu, d) = (\mathbf{x} - \mu\mathbf{1}_n)^t\Sigma_d^{-1}(\mathbf{x}-\mu\mathbf{1}_n)$, so the log-likelihood can be re-written as

$$\ell(\mathbf{x}|\mu, \sigma, d) = -n\log\sigma - \frac{1}{2}\log[\det(\Sigma_d)] - \frac{1}{2\sigma^2}Q(\mathbf{x}|\mu, d). \tag{13}$$

Numerical evaluation requires computing the determinant and inverse of a dense, symmetric positive-definite $n \times n$ matrix, an $\mathcal{O}(n^3)$ operation—too slow for the large n typically encountered in long memory contexts.[4] Simplifications arise upon recognising that Σ_d is symmetric *Toeplitz*, being expressible by just n scalars $c_0, \ldots, c_{n-1}$, i.e., $\Sigma(d, \sigma)_{i,j} = c_{i-j}$ for $i \geq j$. The Durbin–Levinson algorithm (Palma, 2007, §4.1.2) exploits this form, yielding an $\mathcal{O}(n^2)$ cost. However even that remains too large in practice for most applications.

3.2 Approximate likelihood calculation

Here we develop an efficient scheme for evaluating the (log) likelihood, via approximation. Throughout, suppose that we have observed the vector $\mathbf{x} = (x_1, \ldots, x_n)^\top$ as a realisation of a stationary, causal and invertible ARFIMA$(0, d, 0)$ process $\{X_t\}$ with mean $\mu \in \mathbb{R}$. The innovations will be assumed to be independent, and taken from a zero-mean *location-scale probability density* $f(\cdot; 0, \sigma, \boldsymbol{\lambda})$, which means the density can be written as $f(x; \delta, \sigma, \boldsymbol{\lambda}) \equiv \frac{1}{\sigma}f\left(\frac{x-\delta}{\sigma}; 0, 1, \boldsymbol{\lambda}\right)$. The parameters δ and σ are called the 'location' and 'scale' parameters respectively. The m–dimensional $\boldsymbol{\lambda}$ is a 'shape' parameter (if it exists, i.e. $m > 0$). An common example is the Gaussian $\mathcal{N}(\mu, \sigma^2)$, where $\delta \equiv \mu$ and there is $\boldsymbol{\lambda}$. We classify the four parameters μ, σ, $\boldsymbol{\lambda}$, and d, into three distinct classes: (1) the mean of process, μ; (2) innovation distribution parameters, $\boldsymbol{\upsilon} = (\sigma, \boldsymbol{\lambda})$; and (3) memory structure, d. Together, $\boldsymbol{\psi} = (\mu, \boldsymbol{\upsilon}, \boldsymbol{\omega})$, where $\boldsymbol{\omega}$ will later encompass the short-range parameters p and q.

Our proposed likelihood approximation uses a truncated AR(∞) approximation (cf. Haslett and Raftery (1989)). We first re-write the AR(∞) approximation of ARFIMA$(0, d, 0)$

[4] Σ_d is often also poorly conditioned, complicating decomposition (Chen et al., 2006, appendix A).

to incorporate the unknown parameter μ, and drop the (d) superscript for convenience: $X_t - \mu = \varepsilon_t - \sum_{k=1}^{\infty} \pi_k (X_{t-k} - \mu)$. Then we truncate this AR(∞) representation to obtain an AR(P) one, with P large enough to retain low frequency effects, e.g., $P = n$. We denote: $\Pi_P = \sum_{k=0}^{P} \pi_k$ and, with $\pi_0 = 1$, rearrange terms to obtain the following modified model:

$$X_t = \varepsilon_t + \Pi_P \mu - \sum_{k=1}^{P} \pi_k X_{t-k}. \tag{14}$$

It is now possible to write down a *conditional* likelihood. For convenience the notation $\mathbf{x}_k = (x_1, \ldots, x_k)^\top$ for $k = 1, \ldots, n$ will be used (and $\mathbf{x}_0$ is interpreted as appropriate where necessary). Denote the unobserved P–vector of random variables $(x_{1-P}, \ldots, x_{-1}, x_0)^\top$ by $\mathbf{x}_A$ (in the Bayesian context these will be 'auxiliary', hence 'A'). Consider the likelihood $L(\mathbf{x}|\boldsymbol{\psi})$ as a joint density which can be factorised as a product of conditionals. Writing $g_t(x_t|\mathbf{x}_{t-1}, \boldsymbol{\psi})$ for the density of X_t conditional on $\mathbf{x}_{t-1}$, we obtain $L(\mathbf{x}|\boldsymbol{\psi}) = \prod_{t=1}^{n} g_t(x_t|\mathbf{x}_{t-1}, \boldsymbol{\psi})$.

This is still of little use because the g_t may have a complicated form. However by further conditioning on $\mathbf{x}_A$, and writing $h_t(x_t|\mathbf{x}_A, \mathbf{x}_{t-1}, \boldsymbol{\psi})$ for the density of X_t conditional on $\mathbf{x}_{t-1}$ *and* $\mathbf{x}_A$, we obtain: $L(\mathbf{x}|\boldsymbol{\psi}, \mathbf{x}_A) = \prod_{t=1}^{n} h_t(x_t|\mathbf{x}_A, \mathbf{x}_{t-1}, \boldsymbol{\psi})$. Returning to (14) observe that, conditional on both the observed and *un*observed past values, X_t is simply distributed according to the innovations' density f with a suitable change in location: $X_t|\mathbf{x}_{t-1}, \mathbf{x}_A \sim f\left(\cdot; \left[\Pi_P \mu - \sum_{k=1}^{P} \pi_k x_{t-k}\right], \sigma, \boldsymbol{\lambda}\right)$. Then using location-scale representation:

$$\begin{aligned} h_t(x_t|\mathbf{x}_A, \mathbf{x}_{t-1}, \boldsymbol{\psi}) &\approx f\left(x_t; \left[\Pi_P \mu - \sum_{k=1}^{P} \pi_k x_{t-k}\right], \sigma, \boldsymbol{\lambda}\right) & (15) \\ &\equiv \frac{1}{\sigma} f\left(\frac{c_t - \Pi_P \mu}{\sigma}; 0, 1, \boldsymbol{\lambda}\right), \quad \text{where} \quad c_t = \sum_{k=0}^{P} \pi_k x_{t-k}, \qquad t = 1, \ldots, n. \end{aligned}$$

Therefore, $L(\mathbf{x}|\boldsymbol{\psi}, \mathbf{x}_A) \approx \sigma^{-n} \prod_{t=1}^{n} f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda}\right)$, or equivalently:

$$\ell(\mathbf{x}|\boldsymbol{\psi}, \mathbf{x}_A) \approx -n \log \sigma + \sum_{t=1}^{n} \log\left\{ f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda}\right)\right\}. \tag{16}$$

Evaluating this expression efficiently depends upon efficient calculation of $\boldsymbol{c} = (c_1, \ldots, c_n)^t$ and $\log f(\cdot)$. From (15), $\boldsymbol{c}$ is a convolution of the augmented data, $(\mathbf{x}, \mathbf{x}_A)$, and coefficients depending on d, which can be evaluated quickly in R via `convolve` via FFT. Consequently, evaluation of the *conditional* likelihood in the Gaussian case costs only $\mathcal{O}(n \log n)$—a clear improvement over the 'exact' method. Obtaining the *un*conditional likelihood requires marginalisation over $\mathbf{x}_A$, which is analytically infeasible. However this conditional form will suffice in the context of our Bayesian inferential scheme, presented below.

4 A Bayesian approach to long memory inference

We are now ready to consider Bayesian inference for ARFIMA$(0, d, 0)$ processes. Our method can be succinctly described as a modernisation of the blocked MCMC method of Pai and

Ravishanker (1998). Isolating parameters by blocking provides significant scope for modularisation which helps accommodate our extensions for short memory. Pairing with efficient likelihood evaluations allows much longer time series to be entertained than ever before. Our description begins with appropriate specification of priors which are more general than previous choices, yet still encourages tractable inference. We then provide the relevant updating calculations for all parameters, including those for auxiliary parameters $\mathbf{x}_A$.

We follow earlier work (Koop et al., 1997; Pai and Ravishanker, 1998) and assume *a priori* independence for components of ψ. Each component will leverage familiar prior forms with diffuse versions as limiting cases. Specifically, we use a diffuse Gaussian prior on μ: $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, with σ_0 large. The improper flat prior is obtained as the limiting distribution when $\sigma_0 \to \infty$: $p_\mu(\mu) \propto 1$. We place a gamma prior on the precision $\tau = \sigma^{-2}$ implying a *Root-Inverse Gamma* distribution $\mathcal{R}(\alpha_0, \beta_0)$ for σ, with density $f(\sigma) = \frac{2}{\Gamma(\alpha)}{\beta_0}^{\alpha_0}\sigma^{-(2\alpha_0+1)}\exp\left(-\frac{\beta_0}{y^2}\right)$, $\sigma > 0$. A diffuse/improper prior is obtained as the limiting distribution when $\alpha_0, \beta_0 \to 0$: $p_\sigma(\sigma) \propto \sigma^{-1}$. Finally, we specify $d \sim \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)$.

Updating μ: Following Pai and Ravishanker (1998), we use a symmetric random walk (RW) MH update with proposals $\xi_\mu \sim \mathcal{N}(\mu, \sigma_\mu^2)$, for some σ_μ^2. The acceptance ratio is

$$A_\mu(\mu, \xi_\mu) = \sum_{t=1}^{n} \log\left\{f\left(\frac{c_t - \Pi_P \xi_\mu}{\sigma}; \boldsymbol{\lambda}\right)\right\} - \sum_{t=1}^{n} \log\left\{f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda}\right)\right\} + \log\left[\frac{p_\mu(\xi_\mu)}{p_\mu(\mu)}\right] \quad (17)$$

under the approximate likelihood. With the exact likelihood, recall (13) to obtain:

$$A_\mu(\mu, \xi_\mu) = \frac{1}{2\sigma^2}\left[Q(\mathbf{x}|\mu, d) - Q(\mathbf{x}|\xi_\mu, d)\right] + \log\left[\frac{p_\mu(\xi_\mu)}{p_\mu(\mu)}\right].$$

Updating σ: We diverge from Pai and Ravishanker (1998) here, who suggest independent MH with moment-matched inverse gamma proposals, finding poor performance under poor moment estimates. We instead prefer a Random Walk (RW) Metropolis-Hastings (MH) approach, which we conduct in log space since the domain is $\mathbb{R}^+$. Specifically, set: $\log \xi_\sigma = \log \sigma + \upsilon$, where $\upsilon \sim \mathcal{N}(0, \sigma_\sigma^2)$ for some σ_σ^2. $\xi_\sigma|\sigma$ is log-normal and we obtain: $\frac{q(\sigma;\xi_\sigma)}{q(\xi_\sigma;\sigma)} = \frac{\xi_\sigma}{\sigma}$. Recalling (17) the MH acceptance ratio under the approximate likelihood is

$$\begin{aligned} A_\sigma(\sigma, \xi_\sigma) = &\sum_{t=1}^{n} \log\left\{f\left(\frac{c_t - \Pi_P \mu}{\xi_\sigma}; \boldsymbol{\lambda}\right)\right\} - \sum_{t=1}^{n} \log\left\{f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda}\right)\right\} \\ &+ \log\left[\frac{p_\sigma(\xi_\sigma)}{p_\sigma(\sigma)}\right] + (n-1)\log\left[\frac{\sigma}{\xi_\sigma}\right]. \end{aligned}$$

When using the exact likelihood, (13) gives

$$A_\sigma(\sigma, \xi_\sigma) = \frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\xi_\sigma^2}\right) Q(\mathbf{x}|\mu, d) + \log\left[\frac{p_\sigma(\xi_\sigma)}{p_\sigma(\sigma)}\right] + (n-1)\log\left[\frac{\sigma}{\xi_\sigma}\right].$$

The MH algorithm, applied alternately in a Metropolis-within-Gibbs fashion to the parameters μ and σ, works well. However *actual* Gibbs sampling is an efficient alternative in this two-parameter case (i.e., for known d). Since inference for d is a primary goal, we have relegated a derivation of the resulting updates to Appendix A.

Update of d: Updating the memory parameter d is far less straightforward than either μ or σ. Regardless of the innovations' distribution, the conditional posterior $\pi_{d|\psi_{-d}}(d|\boldsymbol{\psi}_{-d}, \mathbf{x})$ is not amenable to Gibbs sampling. We use RW proposals from truncated Gaussian $\xi_d \sim \mathcal{N}^{(a,b)}(\mu, \sigma^2)$, with density

$$f(x; \mu, \sigma, a, b) = \frac{1}{\sigma} \frac{\phi[(x-\mu)/\sigma]}{\Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]}, \qquad a < x < b. \tag{18}$$

In particular, we use $\xi_d|d \sim \mathcal{N}^{(-1/2,1/2)}(d, \sigma_d^2)$ via rejection sampling from $\mathcal{N}(d, \sigma_d^2)$ until $\xi_d \in (-\frac{1}{2}, \frac{1}{2})$. Although this may seem inefficient, it is perfectly acceptable: as an example, if $\sigma_d = 0.5$ the expected number of required variates is still less than 2, regardless of d. More refined methods of directly sampling from truncated normal distributions exist—see for example Robert (1995)—but we find little added benefit in our context.

A useful cancellation in $q(d;\xi_d)/q(\xi_d;d)$ obtained from (18) yields

$$A_d = \ell(\mathbf{x}|\xi_d, \boldsymbol{\psi}_{-d}) - \ell(\mathbf{x}|d, \boldsymbol{\psi}_{-d}) + \log\left[\frac{p_d(\xi_d)}{p_d(d)}\right] + \log\left\{\frac{\Phi[(\frac{1}{2}-d)/\sigma_d] - \Phi[(-\frac{1}{2}-d)/\sigma_d]}{\Phi[(\frac{1}{2}-\xi_d)/\sigma_d] - \Phi[(-\frac{1}{2}-\xi_d)/\sigma_d]}\right\}.$$

Denote $\xi_{c_t} = \sum_{k=0}^{P} \xi_{\pi_k} x_{t-k}$ for $t = 1, \ldots, n$, where $\{\xi_{\pi_k}\}$ are the proposed coefficients $\{\pi_k^{(\xi_d)}\}$; $\pi_k^{(d)} = \frac{1}{\Gamma(k+1)} \frac{\Gamma(k-d)}{\Gamma(-d)}$. Denote $\xi_{\Pi_P} = \sum_{k=0}^{P} \xi_{\pi_k}$. Then in the approximate case:

$$\begin{aligned} A_d = {} & \sum_{t=1}^{n} \log\left\{f\left(\frac{\xi_{c_t} - \xi_{\Pi_P}\mu}{\sigma}; \boldsymbol{\lambda}\right)\right\} - \sum_{t=1}^{n} \log\left\{f\left(\frac{c_t - \Pi_P\mu}{\sigma}; \boldsymbol{\lambda}\right)\right\} \\ & + \log\left[\frac{p_d(\xi_d)}{p_d(d)}\right] + \log\left\{\frac{\Phi[(\frac{1}{2}-d)/\sigma_d] - \Phi[(-\frac{1}{2}-d)/\sigma_d]}{\Phi[(\frac{1}{2}-\xi_d)/\sigma_d] - \Phi[(-\frac{1}{2}-\xi_d)/\sigma_d]}\right\}. \end{aligned} \tag{19}$$

In the exact likelihood case, from (13) we obtain:

$$\begin{aligned} A_d = {} & \frac{1}{2}\log[\det(\Sigma_d)] - \frac{1}{2}\log[\det(\Sigma_{\xi_d})] + \frac{1}{2\sigma^2}\left[Q(\mathbf{x}|\mu, d) - Q(\mathbf{x}|\mu, \xi_d)\right] \\ & + \log\left[\frac{p_d(\xi_d)}{p_d(d)}\right] + \log\left\{\frac{\Phi[(\frac{1}{2}-d)/\sigma_d] - \Phi[(-\frac{1}{2}-d)/\sigma_d]}{\Phi[(\frac{1}{2}-\xi_d)/\sigma_d] - \Phi[(-\frac{1}{2}-\xi_d)/\sigma_d]}\right\}. \end{aligned} \tag{20}$$

Optional update of $\mathbf{x}_A$: When using the approximate likelihood method, one must account for the auxiliary variables $\mathbf{x}_A$, a P–vector (where $P = n$ is sensible). We find that, in practice, it is not necessary to update all the auxiliary parameters at each iteration. In fact the method can be shown to work perfectly well, empirically, if we *never* update them, provided they are given a sensible initial value (such as the sample mean of the observed data $\bar{x}$). This is not an uncommon tactic in the long memory (big-n) context (e.g., Beran, 1994a); for further discussion refer to Graves (2013, Appendix C).

For a full MH approach, we recommend an independence sampler to 'backward project' the observed time series. Specifically, first relabel the observed data: $y_{-i} = x_{i+1}$, $i = 0, \dots n-1$. Then use the vector $(y_{-(n-1)}, \dots, y_{-1}, y_0)^t$ to generate a new vector of length n, $(Y_1, \dots, Y_n)^t$ where Y_t via (14): $Y_t = \varepsilon_t + \Pi_P \mu - \sum_{k=1}^n \pi_k Y_{t-k}$, where the coefficients $\{\pi\}$ are determined by the current value of the memory parameter(s). Then take the proposed $\mathbf{x}_A$, denoted $\boldsymbol{\xi}_{\mathbf{x}_A}$, as the reverse sequence: $\xi_{x_{-i}} = y_{i+1}$, $i = 0, \dots, n-1$. Since this is an independence sampler, calculation of the acceptance probability is straightforward. It is only necessary to evaluate the proposal density $q(\boldsymbol{\xi}_{\mathbf{x}_A} | \mathbf{x}, \boldsymbol{\psi})$. But this is easy using the results from section 3.2. For simplicity, we prefer uniform prior for $\mathbf{x}_A$.

Besides simplicity, justification for this approach lies primarily in is preservation of the auto-correlation structure—this is clear since the ACF is symmetric in time. The proposed vector has a low acceptance rate, and the potential remedies (e.g., multiple-try methods) seem unnecessarily complicated given the success of the simpler method.

5 Extensions to accommodate short memory

Simple ARFIMA$(0, d, 0)$ are mathematically convenient but have limited practical applicability because the entire memory structure is determined by just one parameter, d. Although d is often of primary interest, it may be unrealistic to assume no short memory effects. This issue is often implicitly acknowledged since semi-parametric estimation methods, such as those used as comparators in Section 6.1, are motivated by a desire to circumvent the problem of specifying precisely (and inferring) the form of short memory (i.e., the values of p and q in an ARIMA model). Full parametric Bayesian modelling of ARFIMA(p, d, q) processes represents an essentially untried alternative, primarily due to computational challenges. Related, more discrete, alternatives show potential. Pai and Ravishanker (1998) considered all four models with $p, q \leq 1$, whereas Koop et al. (1997) considered sixteen with $p, q \leq 3$.

Such approaches, especially ones allowing larger p, q, can be computationally burdensome as much effort is spent modelling unsuitable processes towards a goal (inferring p, q) which is not of primary interest (d is). To develop an efficient, fully-parametric, Bayesian method of inference that properly accounts for varying models, and to marginalise out these nuisance quantities, we use reversible-jump (RJ) MCMC (Green, 1995). We extend the parameter space to include the set of models (p and q), with chains moving *between* and within models, and focus on the marginal posterior distribution of d obtained by (Monte Carlo) integration over all models and parameters therein. RJ methods have previously been applied to both auto-regressive models (Vermaak et al., 2004), and full ARMA models (Ehlers and Brooks, 2006, 2008). In the long memory context, Holan et al. (2009) applied RJ to FEXP processes. However for ARFIMA, the only related work we are aware of is by Eğrioğlu and Günay (2010) who demonstrated a promising if limited alternative.

Below we show how the likelihood may be calculated with extra short-memory components when p and q are known, and subsequently how Bayesian inference can be applied in this case. Then, the more general case of unknown p and q via RJ is described.

5.1 Likelihood derivation and inference for known short memory

Recall that short memory components of an ARFIMA process are defined by the AR and MA polynomials, Φ and Θ respectively, (see Section 2). Here, we distinguish between the polynomial, Φ, and the vector of its coefficients, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$. When the polynomial degree is required explicitly, bracketed superscripts will be used; $\Phi^{(p)}$, $\boldsymbol{\phi}^{(p)}$, $\Theta^{(p)}$, $\boldsymbol{\theta}^{(p)}$, respectively.

We combine the short memory parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ with d to create a single 'memory' parameter, $\boldsymbol{\omega} = (\boldsymbol{\phi}, \boldsymbol{\theta}, d)$. For a given unit-variance ARFIMA(p, d, q) process, we denote its ACV by $\gamma_{\boldsymbol{\omega}}(\cdot)$, with $\gamma_d(\cdot)$ and $\gamma_{\boldsymbol{\phi},\boldsymbol{\theta}}(\cdot)$ those of the relevant unit-variance ARFIMA$(0, d, 0)$ and ARMA(p, q) processes respectively. The SDF of the unit-variance ARFIMA(p, d, q) process is written as $f_{\boldsymbol{\omega}}(\cdot)$, and its covariance matrix is $\Sigma_{\boldsymbol{\omega}}$. Therefore, in the general Gaussian ARFIMA(p, d, q) case, we can update the likelihood in (13) to obtain $\ell(\mathbf{x}|\mu, \sigma, \boldsymbol{\omega}) = -n \log \sigma - \frac{1}{2} \log[\det(\Sigma_{\boldsymbol{\omega}})] - \frac{1}{2\sigma^2} Q(\mathbf{x}|\mu, \boldsymbol{\omega})$.

An 'exact' likelihood evaluation requires an explicit calculation of the ACV $\gamma_{\boldsymbol{\omega}}(\cdot)$, however there is no simple closed form for arbitrary ARFIMA processes. Fortunately, our proposed approximate likelihood method of section 3.2 can be ported over directly. Given the coefficients $\{\pi_k^{(d)}\}$ and polynomials Φ and Θ, it is trivial to calculate the $\{\pi_k^{(\boldsymbol{\omega})}\}$ coefficients required by again applying the numerical methods of Brockwell and Davis (1991, §3.3).

To focus the exposition, consider the simple, yet useful, ARFIMA$(1, d, 0)$ model where the full memory parameter is $\boldsymbol{\omega} = (d, \phi_1)$. Because the parameter spaces of d and ϕ_1 are independent, it is simplest to update each of these parameters separately; d with the methods of section 4 and ϕ_1 similarly: $\xi_{\phi_1}|\phi_1 \sim \mathcal{N}^{(-1,1)}(\phi_1, \sigma^2_{\phi_1})$, for some $\sigma^2_{\phi_1}$. In practice however, the posteriors of d and ϕ_1 typically exhibit significant correlation so independent proposals are inefficient. One solution would be to reparametrise to some d^* and orthogonal ϕ_2^*, but the interpretation of d^* would not be clear. An alternative to explicit reparametrisation is to update the parameters jointly, but in such a way that proposals are aligned with the correlation structure. This will ensure a reasonable acceptance rate and mixing.

To propose parameters in the manner described above, a two-dimensional, suitably truncated Gaussian random walk, with covariance matrix aligned with the posterior covariance, is required. To make proposals of this sort, and indeed for arbitrary $\boldsymbol{\omega}$ in larger p and q cases, requires sampling from a *hypercuboid*-truncated MVN $\mathcal{N}_r^{(\mathbf{a},\mathbf{b})}(\boldsymbol{\omega}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$, where $(\mathbf{a}, \mathbf{b})$ describe the coordinates of the hypercube. We find that rejection sampling based unconstrained similarly parameterised MVNs samples [e.g., using `mvtnorm` (Genz et al., 2012)] works well, because in the RW setup the mode of the distribution always lies inside the hypercuboid. Returning to the specific ARFIMA$(1, d, 0)$ case, clearly $r = 2$, $\mathbf{b} = (0.5, 1)$ and $\mathbf{a} = -\mathbf{b}$, is appropriate. Calculation of the MH acceptance ratio $A_{\boldsymbol{\omega}}(\boldsymbol{\omega}, \boldsymbol{\xi}_{\boldsymbol{\omega}})$ is trivial; it simply requires numerical evaluation of $\Phi_r(\cdot; \cdot, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$, e.g., via `mvtnorm`, since the ratios of hypercuboid normalisation terms would cancel. We find that initial $\phi^{[0]}$ chosen uniformly in $\mathcal{C}_1$, i.e. the interval $(-1, 1)$, and $d^{[0]}$ are systematically from $\{-0.4, -0.2, 0, 0.2, 0.4\}$ work well. Any choice of prior for $\boldsymbol{\omega}$ can be made, although we prefer flat (proper) priors.

The only technical difficulty is the choice of proposal covariance matrix $\Sigma_{\boldsymbol{\omega}}$. Ideally, it would be aligned with the posterior covariance—however this is not *a priori* known. We find that running a 'pilot' chain with independent proposals via $\mathcal{N}_r^{(\mathbf{a},\mathbf{b})}(\boldsymbol{\omega}, \sigma^2_{\boldsymbol{\omega}}\boldsymbol{I}_r)$ can help choose a $\Sigma_{\boldsymbol{\omega}}$. A rescaled version of the sample covariance matrix from the pilot posterior chain, following Roberts and Rosenthal (2001), works well [see Section 6.2].

5.2 Unknown short memory form

We now expand the parameter space to include models $M \in \mathcal{M}$, the set of ARFIMA models with p and q short memory parameters, indexing the size of the parameter space $\Psi^{(M)}$. For our 'transdimensional moves', we only consider adjacent models, on which we will be more specific later. For now, note that the choice of bijective function mapping between models spaces (whose Jacobian term appears in the acceptance ratio), is crucial to the success of the sampler. To illustrate, consider transforming from $\Phi^{(p+1)} \in \mathcal{C}_{p+1}$ down to $\Phi^{(p)} \in \mathcal{C}_p$. This turns out to be a non-trivial problem however because, for $p > 1$, $\mathcal{C}_p$ has a very complicated shape. The most natural map would be: $(\phi_1, \ldots, \phi_p, \phi_{p+1}) \longmapsto (\phi_1, \ldots, \phi_p)$. However there is no guarantee that the image will lie in $\mathcal{C}_p$. Even if the model dimension is fixed, difficulties are still encountered; a natural proposal method would be to update each component of $\boldsymbol{\phi}$ separately but, because of the awkward shape of $\mathcal{C}_p$, the 'allowable' values for each component are a complicated function of the others. Nontrivial proposals are required.

A potential approach is to reparametrise in terms of the inverse roots (poles) of Φ, as advocated by Ehlers and Brooks (2006, 2008): By writing $\Phi(z) = \prod_{i=1}^p (1 - \alpha_i z)$, we have that $\boldsymbol{\phi}^{(p)} \in \mathcal{C}_p \iff |\alpha_i| < 1$ for all i. This looks attractive because it transforms $\mathcal{C}_p$ into $D^p = D \times \cdots \times D$ (p times) where D is the open unit disc, which is easy to sample from. But this method has serious drawbacks when we consider the RJ step. To decrease dimension, the natural map would be to remove one of the roots from the polynomial. But because it is assumed that Φ has real coefficients (otherwise the model has no realistic interpretation), any complex α_i must appear as conjugate pairs. There is then no obvious way to remove a root; a contrived method might be to remove the conjugate pair and replace it with a real root with the same modulus, however it is unclear how this new polynomial is related to the original, and to other aspects of the process, like ACV.

Reparametrisation of Φ and Θ

We therefore propose reparameterising Φ (and Θ) using the bijection between $\mathcal{C}_p$ and $(-1,1)^p$ advocated by various authors, e.g., Marriott et al. (1995) and Vermaak et al. (2004). To our knowledge, these methods have not previously been deployed towards integrating out short memory components in Bayesian analysis of ARFIMA processes.

Monahan (1984) defined a mapping $\boldsymbol{\phi}^{(p)} \longleftrightarrow \boldsymbol{\varphi}^{(p)}$ recursively as follows:

$$\phi_i^{(k-1)} = \frac{\phi_i^{(k)} - \phi_k^{(k)}\phi_{k-i}^{(k)}}{1 - \left(\phi_k^{(k)}\right)^2}, \qquad k = p, \ldots, 2, \qquad i = 1, \ldots, k-1. \tag{21}$$

Then set $\varphi_k^{(p)} = \phi_k^{(k)}$ for $k = 1, \ldots, p$. The reverse recursion is given by:

$$\phi_i^{(k)} = \begin{cases} \varphi_k^{(p)} & \text{for} \quad i = k \qquad\qquad k = 1, \ldots, p \\ \phi_i^{(k-1)} + \varphi_k^{(p)} \phi_{k-i}^{(k-1)} & \text{for} \quad i = 1, \ldots, k-1 \quad k = 2, \ldots, p \end{cases}.$$

Note that $\phi_p^{(p)} = \varphi_p^{(p)}$. Moreover, if $p = 1$, the two parametrisations are the same, i.e. $\phi_1 = \varphi_1$ (consequently the brief study of ARFIMA$(1, d, 0)$ in section 5.1 fits in this framework). The equivalent reparametrised form for $\boldsymbol{\theta}$ is $\boldsymbol{\vartheta}$. The full memory parameter $\boldsymbol{\omega}$ is reparametrised as $\bar{\Omega} = (-1/2, 1/2) \times$ (the image of $\mathcal{C}_{p,q}$). However recall that in practice, $\mathcal{C}_{p,q}$ will be assumed equivalent to $\mathcal{C}_p \times \mathcal{C}_q$, so the parameter space is effectively: $\bar{\Omega} = (-1/2, 1/2) \times (-1, 1)^{p+q}$.

Besides mathematical convenience, this bijection has a very useful property (Kay and Marple, 1981, cf.) which helps motivate its use in defining RJ maps . In Appendix B we show that the ACFs of the $\boldsymbol{\phi}^{(p)}$ and $\boldsymbol{\phi}^{(p-1)}$ are identical. In other words, if $d = q = 0$, using this parametrisation for $\boldsymbol{\varphi}$ when moving between different values of p allows one to automatically choose processes that have very closely matching ACFs at low lags. In the MCMC context this is useful because it allows the chain to propose models that have a similar correlation structure to the current one. Although this property is nice, it may be of limited value for full ARFIMA models, since the proof of the main result [Theorem 1] does not easily lend itself to the inclusion of either a MA or long memory component. Nevertheless, our empirical results similarly indicate a 'near-match' for a full ARFIMA(p, d, q) model.

Application of RJ MCMC to ARFIMA(p, d, q) processes

We now use this reparametrisation to efficiently propose new parameter values. Firstly, it is necessary to propose a new memory parameter $\boldsymbol{\varpi}$ whilst keeping the model fixed. Attempts at updating each component individually suffer from the same problems of excessive posterior correlation that were encountered in section 5.1. Therefore the simultaneous update of the entire $r = (p + q + 1)$-dimensional parameter $\boldsymbol{\varpi}$ is performed using the hypercuboid-truncated Gaussian distribution from definition $\boldsymbol{\xi}_{\boldsymbol{\varpi}} | \boldsymbol{\varpi} \sim \mathcal{N}_r^{\mathcal{H}_r}(\boldsymbol{\varpi}, \Sigma_{\boldsymbol{\varpi}})$, where $\mathcal{H}_r$ defines the r-dimensional rectangle. The covariance matrix $\Sigma_{\boldsymbol{\varpi}}$ is discussed in some detail below. The choice of prior $p_{\boldsymbol{\varpi}}(\cdot)$ is arbitrary. Pai and Ravishanker (1998) used a uniform prior for $\boldsymbol{\omega}$ which has an explicit expression in the $\boldsymbol{\varpi}$ parameterisation (Monahan, 1984). However, their expression is unnecessarily complicated since a uniform prior over Ω holds no special interpretation. We therefore prefer uniform prior over $\bar{\Omega}$: $p_{\boldsymbol{\varpi}}(\boldsymbol{\varpi}) \propto 1$, $\boldsymbol{\varpi} \in \bar{\Omega}$.

Now consider the 'between-models' transition. We must first choose a model prior $p_{\mathcal{M}}(\cdot)$. A variety of priors are possible; the simplest option would be to have a uniform prior over $\mathcal{M}$, but this would of course be improper. We may in practice want to restrict the possible values of p, q to $0 \leq p \leq P$ and $0 \leq q \leq Q$ for some P,Q (say 5), which would render the uniform prior proper. However even in this formulation, a lot of prior weight is being put onto complicated models which, in the interests of parsimony, might be undesired. We prefer a truncated joint Poisson distribution with parameter λ: $p_{\mathcal{M}}(p, q) \propto \frac{\lambda^{p+q}}{p!q!} \mathbb{I}(p \leq P, q \leq Q)$.

Now, denote the probability of jumping from model $M_{p,q}$ to model $M_{p',q'}$ by $U_{(p,q),(p',q')}$. U could allocate non-zero probability for every model pair, but for convenience we severely restrict the possible jumps (whilst retaining irreducibility) using a two-dimensional bounded birth and death process. Consider the subgraph of $\mathbb{Z}^2$: $G = \{(p, q) : 0 \leq p \leq P,\ 0 \leq q \leq Q\}$, and allocate uniform non-zero probability only to neighboring values, i.e., if and only if $|p - p'| + |q - q'| = 1$. Each point in the 'body' of G has four neighbours; each point on the 'line boundaries' has three; and each of the four 'corner points' has only two neighbours. Therefore the model transition probabilities $U_{(p,q),(p',q')}$ are either 1/4, 1/3, 1/2, or 0.

Now suppose the current $(p + q + 3)$-dimensional parameter is $\boldsymbol{\psi}^{(p,q)}$, given by $\boldsymbol{\psi}^{(p,q)} = (\mu, \sigma, d, \boldsymbol{\varphi}^{(p)}, \boldsymbol{\vartheta}^{(q)})$, using a slight abuse of notation. Because the mathematical detail of the AR and MA components are almost identical, we consider only the case of de/increasing p by 1 here; all of the following remains valid if p is replaced by q, and $\boldsymbol{\varphi}$ replaced by $\boldsymbol{\vartheta}$. We therefore seek to propose a parameter $\boldsymbol{\xi}^{(p+1,q)} = (\xi_\mu, \xi_\sigma, \xi_d, \boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p+1)}, \boldsymbol{\xi}_{\boldsymbol{\vartheta}}^{(q)})$, that is somehow based on $\boldsymbol{\psi}^{(p,q)}$. We further simplify by regarding the other three parameters (μ, σ, and d) as having the same interpretation in every model, choosing $\xi_\mu = \mu$, $\xi_\sigma = \sigma$ and $\xi_d = d$. For simplicity we also set $\boldsymbol{\xi}_{\boldsymbol{\vartheta}}^{(q)} = \boldsymbol{\vartheta}^{(q)}$. Now consider the map $\boldsymbol{\varphi}^{(p)} \to \boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p+1)}$. To specify a bijection we 'dimension-match' by adding in a random scalar u. The most obvious map is to specify u so that its support is the interval $(-1, 1)$ and then set: $\boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p+1)} = \big(\boldsymbol{\varphi}^{(p)}, u\big)$. The corresponding map for decreasing the dimension is $\boldsymbol{\varphi}^{(p+1)} \to \boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p)}$ is $\boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p)} = \Big(\varphi_1^{(p+1)}, \ldots, \varphi_p^{(p+1)}\Big)$. In other words, we either add, or remove the final parameter, whilst keeping all others fixed with the identity map, so the Jacobian is unity. The proposal $q(u|\boldsymbol{\psi}^{(p,q)})$ can be made in many ways—we prefer the simple $\mathcal{U}(-1, 1)$. With these choices the RJ acceptance ratio is

$$A = \ell_{(p',q')}(\mathbf{x}|\boldsymbol{\xi}^{(p',q')}) - \ell_{(p,q)}(\mathbf{x}|\boldsymbol{\psi}^{(p,q)}) + \log\left\{\frac{p_{\mathcal{M}}(p', q')}{p_{\mathcal{M}}(p, q)}\frac{U_{(p',q'),(p,q)}}{U_{(p,q),(p',q')}}\right\},$$

which applies to both increasing and decreasing dimensional moves.

Construction of $\Sigma_{\boldsymbol{\varpi}}$: Much of the efficiency of the above scheme, including within- and between-model moves, depends on the choice of $\Sigma_{\boldsymbol{\varpi}} \equiv \Sigma^{(p,q)}$, the within-model move RW proposal covariance matrix. We first seek an appropriate $\Sigma^{(1,1)}$, as in section 5.1, with a pilot tuning scheme. That matrix is shown on the left below, where we've 'blocked it out'

$$\Sigma^{(1,1)} = \left(\begin{array}{c|c:c} \sigma_d^2 & \sigma_{d,\varphi_1} & \sigma_{d,\vartheta_1} \\ \hline & \sigma_{\varphi_1}^2 & \sigma_{\varphi_1,\vartheta_1} \\ \hdashline & & \sigma_{\vartheta_1}^2 \end{array}\right), \quad \Sigma^{(p,q)} = \left(\begin{array}{c|c:c} \sigma_d^2 & \Sigma_{d,\boldsymbol{\varphi}^{(p)}} & \Sigma_{d,\boldsymbol{\vartheta}^{(q)}} \\ \hline & \Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\varphi}^{(p)}} & \Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\vartheta}^{(q)}} \\ \hdashline & & \Sigma_{\boldsymbol{\vartheta}^{(q)},\boldsymbol{\vartheta}^{(q)}} \end{array}\right), \qquad (22)$$

(where each block is a scalar) so that we can extend this idea to the (p, q) case in the obvious way—on the right above—where $\Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\varphi}^{(p)}}$ is a $p \times p$ matrix, $\Sigma_{\boldsymbol{\vartheta}^{(q)},\boldsymbol{\vartheta}^{(q)}}$ is a $q \times q$ matrix, etc. If either (or both) $p, q = 0$ then the relevant blocks are simply omitted. To specify the various sub-matrices, we propose $\varphi_2, \ldots, \varphi_p$ with equal variances, and *independently* of

$d, \varphi_1, \vartheta_1$, (and similarly for $\vartheta_2, \ldots, \vartheta_q$). In the context of (22), the following hold:

$$\Sigma_{d,\boldsymbol{\varphi}^{(p)}} = \left(\begin{array}{c:c} \sigma_{d,\varphi_1} & \mathbf{0} \end{array} \right), \qquad \Sigma_{d,\boldsymbol{\vartheta}^{(q)}} = \left(\begin{array}{c:c} \sigma_{d,\vartheta_1} & \mathbf{0} \end{array} \right),$$

$$\Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\varphi}^{(p)}} = \left(\begin{array}{c:c} \sigma^2_{\varphi_1} & \mathbf{0} \\ \hdashline \mathbf{0} & \sigma^2_{\boldsymbol{\varphi}} I_{p-1} \end{array} \right), \qquad \Sigma_{\boldsymbol{\vartheta}^{(q)},\boldsymbol{\vartheta}^{(q)}} = \left(\begin{array}{c:c} \sigma^2_{\vartheta_1} & \mathbf{0} \\ \hdashline \mathbf{0} & \sigma^2_{\boldsymbol{\vartheta}} I_{q-1} \end{array} \right),$$

$$\Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\vartheta}^{(q)}} = \left(\begin{array}{c:c} \sigma_{\varphi_1,\vartheta_1} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{O} \end{array} \right),$$

where the dotted lines indicate further blocking, $\mathbf{0}$ is a row-vector of zeros, and $\mathbf{O}$ is a zero matrix. This choice of $\Sigma_{\boldsymbol{\varpi}}$ is conceptually simple, computationally easy and preserves the positive-definiteness as required; this is shown by a simple relabeling of the rows/columns, and then repeated application of theorems 2 and 3 which appear in B.

6 Empirical illustration and comparison

Here we provide empirical illustrations for the methods above: for classical and Bayesian analysis of long memory models, and extensions for short memory. To ensure consistency throughout, the location and scale parameters will always be chosen as $\mu_I = 0$ and $\sigma_I = 1$. Furthermore, unless stated otherwise, the simulated series will be of length $n = 2^{10} = 1024$. This is a reasonable size for many applications; it is equivalent to 85 years' monthly observations. When using the approximate likelihood method we set $P = n$. Unless otherwise stated the priors used will be those simple defaults suggested in the previous sections.

6.1 Long memory

We begin by demonstrating that the approximate likelihood of section 3.2 is accurate. We then conduct a Monte Carlo study varying length of the input, n. Finally, we compare the Bayesian point-estimates and with common non/semi-parametric alternatives.

Standard MCMC diagnostics were used throughout to ensure, and tune for, good mixing. Because d is the parameter of primary interest, the initial values $d^{[0]}$ will be chosen to systematically cover its parameter space, usually starting five chains at the regularly-spaced points $\{-0.4, -0.2, 0, 0.2, 0.4\}$. Initial values for other parameters are not varied: μ will start at the sample mean $\bar{x}$; σ at the sample standard deviation of the observed series $\mathbf{x}$. When using the approximate likelihood method, setting each of $\mathbf{x}_A$ to $\bar{x}$ turns out to be a sufficiently good strategy. For other MCMC particulars, see Graves (2013, §4.3.3).

Efficacy of approximate likelihood method

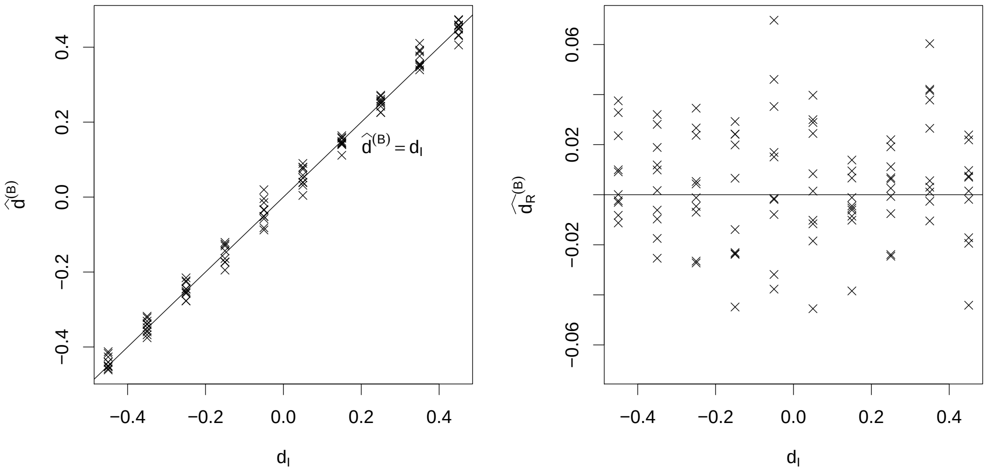
Start with the 'null case', i.e., how does the algorithm perform when the data are not from a long memory process? One hundred independent ARFIMA$(0, 0, 0)$, or Gaussian white noise, processes are simulated, from which marginal posterior means, standard deviations, and credibility interval endpoints are extracted. Table 1 shows averages over the runs.

Table 1: Posterior summary statistics for ARFIMA$(0,0,0)$ process. Average of 100 runs.

	mean	std	95% CI	
d	0.006	0.025	-0.042	0.055
μ	-0.004	0.035	-0.073	0.063
σ	1.002	0.022	0.956	1.041

The average estimate for each of the three parameters is less than a quarter of a standard deviation away from the truth. Credibility intervals are nearly symmetric about the estimate and the marginal posteriors are, to a good approximation, locally Gaussian (not shown). Upon, applying a proxy 'credible-interval-based hypothesis test' one would conclude in ninety-eight of the cases that $d = 0$ could not be ruled out. A similar analysis for μ and σ shows that hypotheses $\mu = 0$ and $\sigma = 1$ would each have been accepted ninety-six times. These results indicate that the 95% credibility intervals are approximately correctly sized.

Next, consider the more interesting case of $d_I \neq 0$. We repeat the above experiment except that ten processes are generated with d_I set to each of $\{-0.45, -0.35, \ldots, 0.45\}$, giving 100 series total. Figure 1 shows a graphical analog of results from this experiment. The plot axes involve a Bayesian residual estimate of d, $\widehat{d_R}^{(B)}$, defined as $\widehat{d_R}^{(B)} = \widehat{d}^{(B)} - d_I$, where $\widehat{d}^{(B)}$ is the Bayesian estimate of d. From the figure is clear that the estimator for d is performing well. Plot (a) shows how 'tight' the estimates of d are around the input



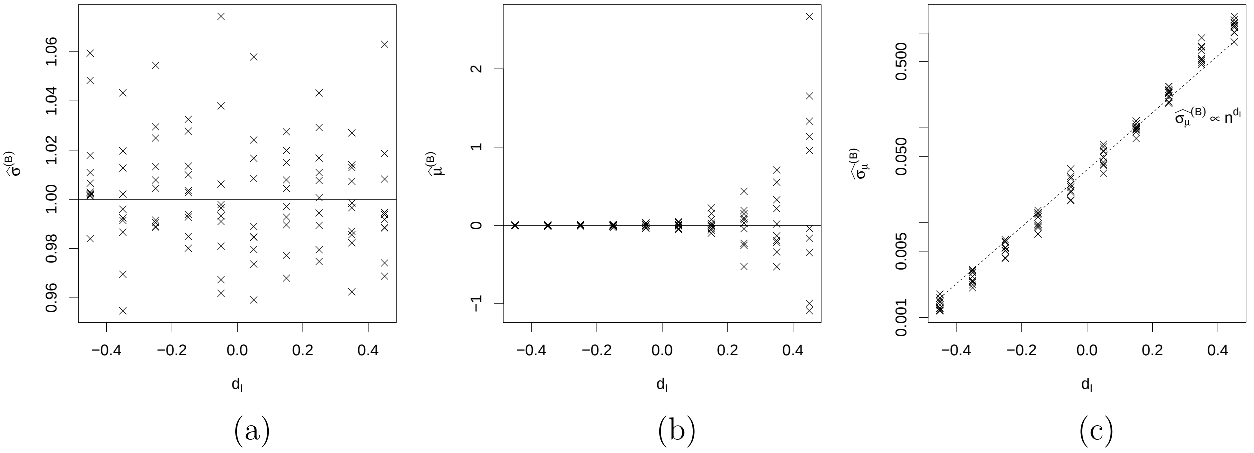
(a)

(b)

Figure 1: Posterior outputs; (a) $\widehat{d}^{(B)}$ against d_I, (b) $\widehat{d_R}^{(B)}$ against d_I.

value—recall that the parameter space for d is the whole interval $(-\frac{1}{2}, \frac{1}{2})$. Moreover, plot (b) indicates that there is no significant change of posterior bias or variance as d_I is varied.

Next, the corresponding plots for the parameters σ and μ are shown in figure 2. We see from plot (a) that the estimate of σ also appears to be unaffected by the input value d_I. The situation is different however in plot (b) for the location parameter μ. Although the bias appears to be roughly zero for all d_I, the posterior variance clearly *is* affected by d_I. To ascertain the precise functional dependence, consider plot (c) which shows, on a semi-log scale, the marginal posterior standard deviation of μ, $\widehat{\sigma_\mu}^{(B)}$, against d_I.



(a) (b) (c)

Figure 2: Posterior outputs; (a) $\widehat{\sigma}^{(B)}$ against d_I, (b) $\widehat{\mu}^{(B)}$ against d_I, and (c) $\widehat{\sigma_\mu}^{(B)}$ against d_I (semi-log scale).

It appears that the marginal posterior standard deviation $\widehat{\sigma_\mu}^{(B)}$ is a function of d_I; specifically: $\widehat{\sigma_\mu}^{(B)} \propto A^{d_I}$, for some A. The constant A could be estimated via least-squares regression. Instead however, inspired by asymptotic results in literature concerning classical estimation of long memory processes (Beran, 1994b) we set $A = n$ and plotted the best fitting such line (shown in plot (c)). Observe that, although not fitting exactly, the relation $\widehat{\sigma_\mu}^{(B)} \propto n^{d_I}$ holds reasonably well for $d_I \in (-\frac{1}{2}, \frac{1}{2})$. Indeed, Beran motivated long memory in this way, and derived asymptotic consistency results for optimum (likelihood-based) estimators and found indeed that the standard error for μ is proportional to $n^{d-1/2}$ (theorem 8.2) but the standard errors of all other parameters are proportional to $n^{-1/2}$ (theorem 5.1).

To study the impact of improper priors for μ and σ we twice repeated the analysis of the 100 ARFIMA$(0, 0, 0)$ above, changing $p_\mu(\cdot)$ in the first instance and $p_\sigma(\cdot)$ in the second. When using $p_\mu(\cdot) \sim \mathcal{N}(0, 100^2)$, the maximum difference between estimates was 0.0012 for d, 0.0017 for σ and 0.0019 for μ. When using $p_\sigma(\cdot) \sim \mathcal{R}(0.01, 0.01)$, the corresponding values were 0.0020 for d, 0.0017 for σ and 0.0027 for μ. Since these maximum differences are well below the Monte Carlo error, we conclude that there is practically no difference between using an improper prior and a vague proper prior.

Comparison of likelihood methods

To compare the methods based on approximate and exact likelihoods we consider fifty ARFIMA$(0,0,0)$ generated as described above. The output summary statistics are presented in a table within figure 3, which is accompanied by a useful visualisation. Observe that both methods produce highly correlated point estimates and credibility interval endpoints. The plot shows that the posteriors of d are similar; there is only a very slight 'shift' between the approximate and exact results, implying that the approximate method is consistently underestimating d (by less than 0.01). The credibility intervals have excellent frequentist coverage properties; 48 (96%) of the 95%-level intervals contain the input d_I. We also found (not shown) that μ and σ exhibit the same pattern of behaviour.

	mean	std	95% CI	
approx d	0.006	0.025	-0.043	0.056
exact d	0.014	0.026	-0.035	0.065
approx μ	-0.005	0.034	-0.073	0.063
exact μ	-0.001	0.036	-0.072	0.069
approx σ	1.002	0.022	0.959	1.046
exact σ	1.001	0.022	0.958	1.044

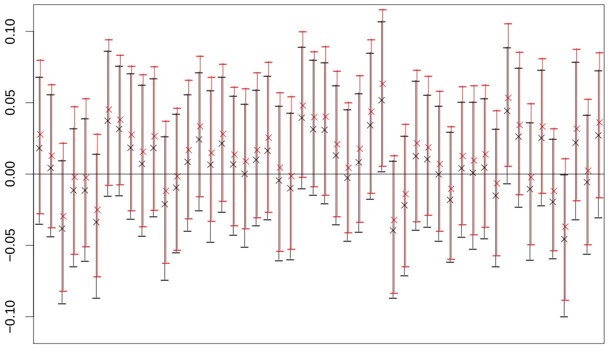


Figure 3: *Table:* Comparison of posterior summary statistics for ARFIMA$(0,0,0)$ process obtained via approximate and exact likelihood methods. Average of 50 runs. *Plot:* Comparison of 95% credibility intervals for d from ARFIMA$(0,0,0)$ processes obtained via approximate and exact likelihood methods. Black is used for approximate, red for exact. The crosses are the respective point estimates.

To check for similarity between the approximate and exact likelihood methods across the entire parameter space we repeated the ARFIMA$(0,d,0)$ simulations with varying d (analyzed in figure 1). The results for the residuals, d_R, are presented in figure 4. We observe the same pattern here as with the $d_I = 0$ case. Note that the estimates of d obtained using the two methods appear to be slightly closer for negative d_I than for positive d_I, with very low discrepancies (0.01) even for the largest $d_I = 0.45$. This suggests that the approximate method generally performs better for smaller d_I.

Effect of varying time series length

We now analyse the effect of changing the time series length. For this we conduct a similar experiment but fix $d_I = 0$ and vary n. The posterior statistics of interest are the posterior standard deviations $\widehat{\sigma_d}^{(B)}$, $\widehat{\sigma_\mu}^{(B)}$ and $\widehat{\sigma_\sigma}^{(B)}$. For each $n \in \{128 = 2^7, 2^8, \ldots, 2^{14} = 16,384\}$, 10 independent ARFIMA$(0,0,0)$ time series are generated. The resulting posterior standard deviations are plotted against n (on log-log scale) in figure 5.

	mean	std	95% CI	
approx d_R	0.005	0.025	-0.042	0.048
exact d_R	0.014	0.025	-0.034	0.057

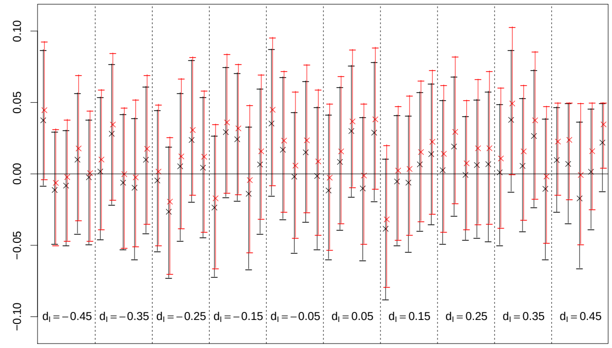
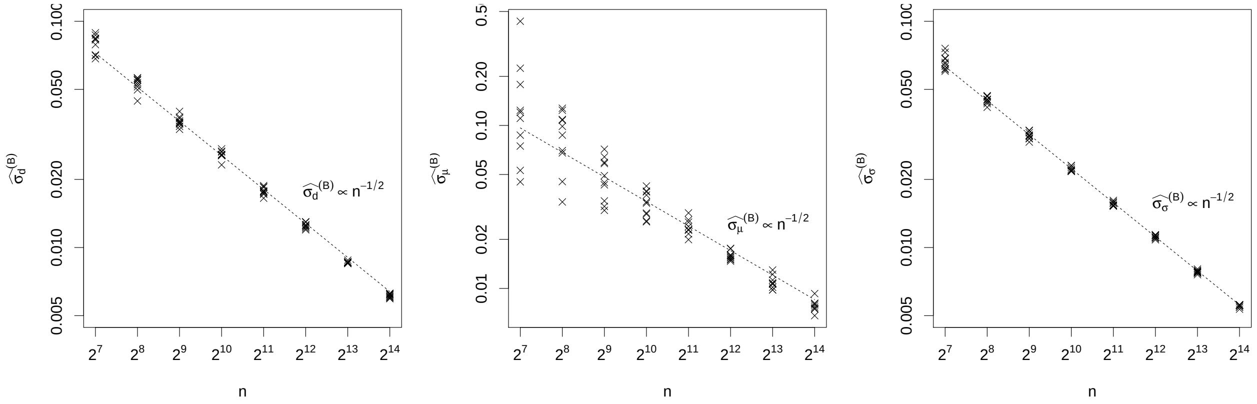


Figure 4: Analog of Figure 3 for ARFIMA$(0, d, 0)$.



(a) (b) (c)

Figure 5: Posterior outputs from ARFIMA$(0, 0, 0)$ series; (a) $\widehat{\sigma_d}^{(B)}$ against n, (b) $\widehat{\sigma_\mu}^{(B)}$ against n, (c) $\widehat{\sigma_\sigma}^{(B)}$ against n (log-log scale).

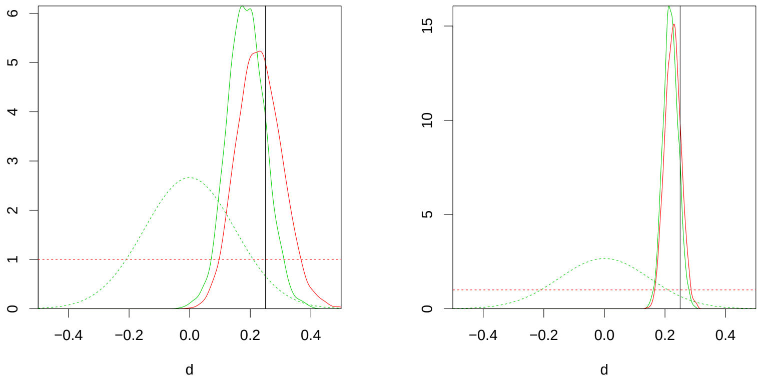
Observe that all three marginal posterior standard deviations are proportional to $\frac{1}{\sqrt{n}}$, although the posterior of μ is less 'reliable'. Combining these observations with our earlier deduction that $\sigma_\mu^{(B)} \propto n^{d_I}$, we conclude that for an ARFIMA$(0, d_I, 0)$ process of length n, the marginal posterior standard deviations follow those of Beran given previously.

Influence of prior

Throughout, uninformative priors are used by default. However, one of the principal advantages of the Bayesian approach is the ability to include genuine prior information. To explore the effect prior variations we performed a series of tests in which $p_d(d) \propto \mathcal{N}(0, 0.15^2)$ in an analysis of ARFIMA$(0, 0.25, 0)$ processes. Note that the true value $d = 0.25$ is outside of the central 90% of the prior distribution, with the density there being less than a quarter of that at the maximum (i.e. when $d = 0$). For each length of $n = 2^7, 2^8, 2^9, 2^{10}$, ten series were analysed and compared with the equivalent analyses with the flat prior.

As expected, in each case the Bayesian estimate $\hat{d}^{(B)}$ is always lower when the Gaussian prior is used, compared to the flat prior. However the average *difference* between the two estimates shows a clear inverse relationship with n; see the table in figure 6. For small n, the average difference is fairly significant, but for $n = 1024$, the difference is below the Monte Carlo error, as the likelihood has overwhelmed the prior in this case. For a graphical demonstration the convergence of the priors, refer to the plots in figure 6. Plot (a) shows that for $n = 128$, the two posteriors are quite clearly distinct. However plot (b) shows that for $n = 1024$, the posteriors have essentially coincided.

n	mean difference
128	0.057
256	0.029
512	0.015
1024	0.007



(a) (b)

Figure 6: *Table:* Mean difference of estimates $\hat{d}^{(B)}$ under alternative prior assumption. *Plots:* Comparison of posteriors (solid lines) obtained under different priors (dotted lines). Time series used: ARFIMA$(0, 0.25, 0)$; (a) $n = 2^7 = 128$, (b) $n = 2^{10} = 1024$.

Comparison with common estimators

In many practical applications, the long memory parameter is estimated using non/semi-parametric methods. These may be appropriate in many situations, where the exact form of the underlying process is unknown. However when a specific model form is known (or at least assumed) they tend to perform poorly compared with fully parametric alternatives (Franzke et al., 2012). Our aim here is to demonstrate, via a short Monte Carlo study involving ARFIMA$(0, d, 0)$ data, that our Bayesian likelihood-based method significantly outperforms other common methods in that case. We consider the following comparators: (i) rescaled adjusted range, or R/S Hurst (1951); Graves (2013)—we use the R implementation in the `FGN` (McLeod et al., 2007) package; (ii) Semi-parametric Geweke–Porter-Hudak (GPH) method (Geweke and Porter-Hudak, 1983)—implemented in R package `fracdiff` (Fraley et al., 2012); (iii) detrended fluctuation analysis (DFA), originally devised by Peng et al. (1994)—in the R package `PowerSpectrum` (Vyushin et al., 2009). (iv) wavelet-based semi-parametric estimators Abry et al. (2003) available in R package `fARMA` (Wuertz, 2012).

Each of these four methods will be applied to the same 100 time series with varying d_I as were used earlier experiments above. We extend the idea of a residual, $\widehat{d_R}^{(R)}$, $\widehat{d_R}^{(G)}$, $\widehat{d_R}^{(D)}$, and $\widehat{d_R}^{(W)}$, to accomodate the new comparators, respectively, and plot them against $\widehat{d_R}^{(B)}$ in figure 7. Observe that all four methods have a much larger variance than our Bayesian

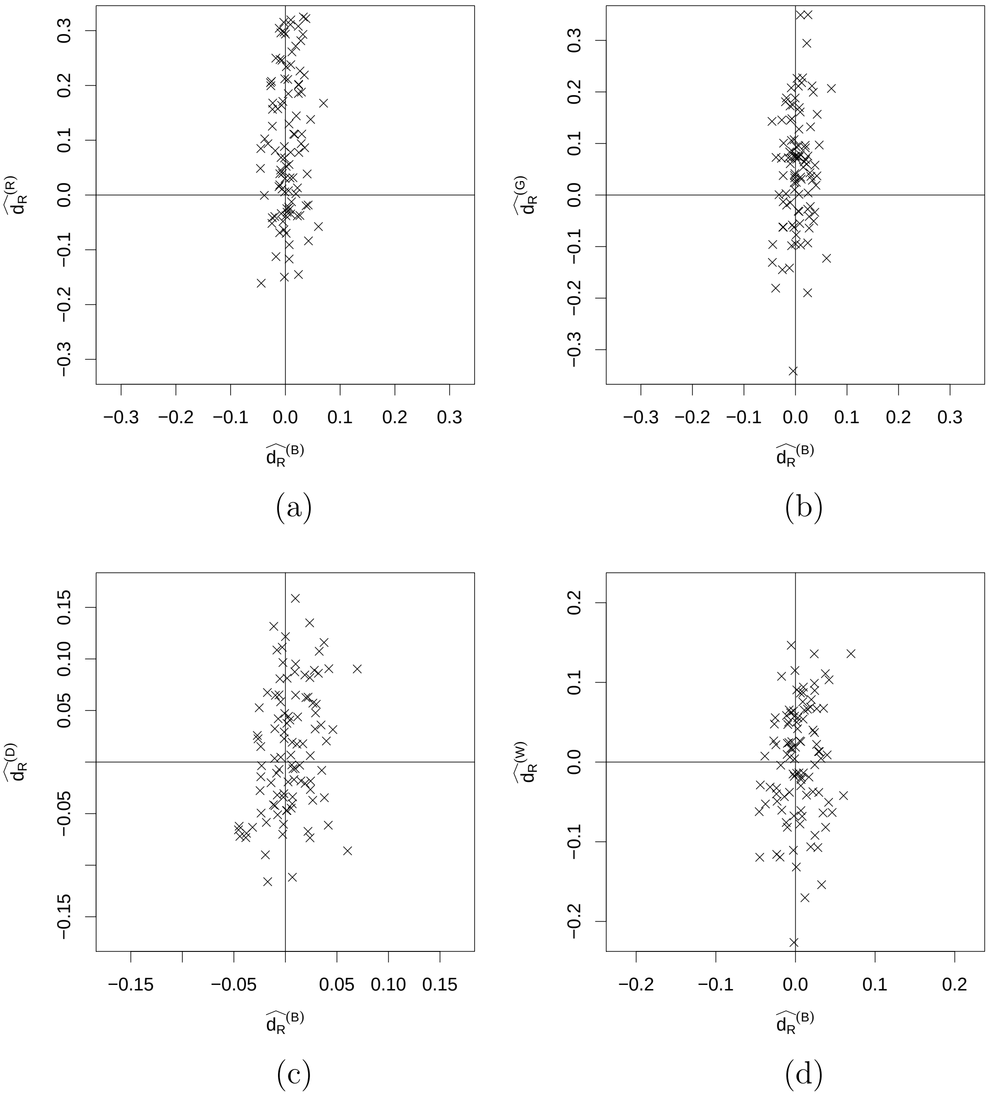


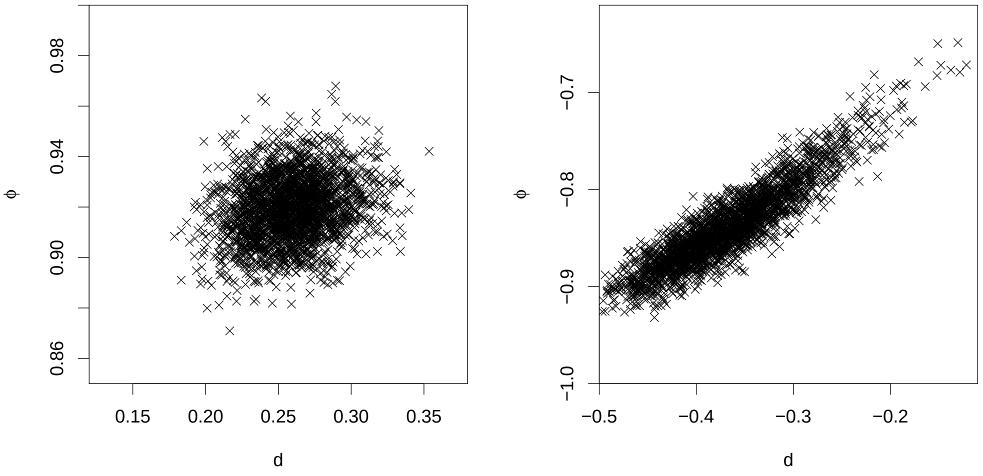
Figure 7: Comparison of Bayesian estimator with common classical estimators; (a) R/S, (b) GPH, (c) DFA, (d) Wavelet.

method, and moreover the R/S is positively biased. Actually, the bias in some cases would seem to depend on d_I: R/S is significantly (i.e. > 0.25) biased for $d_I < -0.3$ but slightly negatively biased for $d > 0.3$ (not shown); DFA is only unbiased for $d_I > 0$; both the GPH and wavelet methods are unbiased for all $d \in (-\frac{1}{2}, \frac{1}{2})$.

6.2 Extensions for short memory

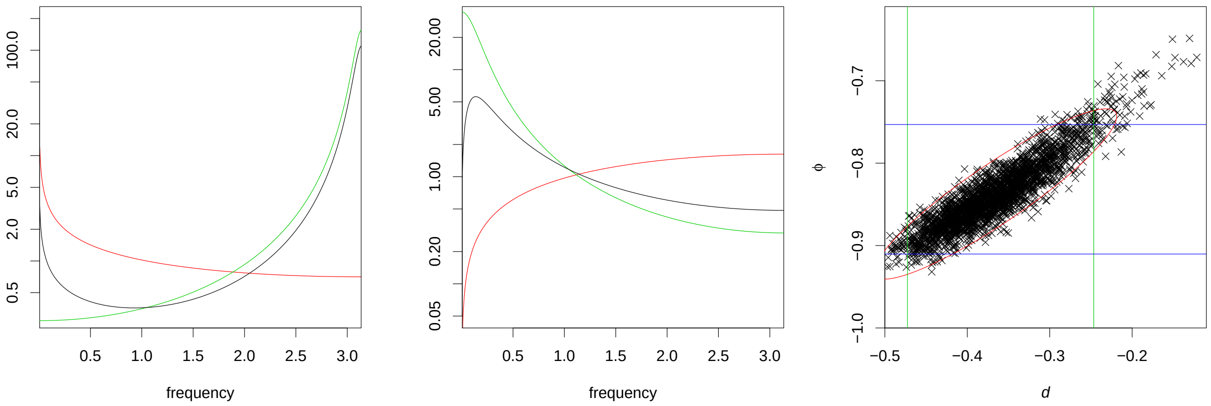
Known form: We first consider the MCMC algorithm from section 5.1 for sampling under an ARFIMA$(1, d, 0)$ model where the full memory parameter is $\boldsymbol{\omega} = (d, \phi_1)$. Recall that that method involved proposals from a hypercuboid MVN using a pilot-tuned covariance matrix. Also recall that it is a special case of the re-parametrised method from section 5.2.

In general, this method works very well; two example outputs are presented in figure 8, under two similar data generating mechanisms. Plot (a) shows relatively mild correlation ($\rho = 0.21$) compared with (b) which shows strong correlation ($\rho = 0.91$). This differential behaviour can be explained heuristically by considering the differing data-generating values. For the process in plot (a) the short memory and long memory components exhibit their effects at opposite ends of the spectrum; see figure 9(a). The resulting ARFIMA spectrum, with peaks at either end, makes it easy to distinguish between short and long memory effects, and consequently the posteriors of d and ϕ are largely uncorrelated. In contrast, the parameters of the process in plot (b) express their behaviour at the same end of the spectrum. With negative d these effects partially cancel each other out, except very near the origin where the negative memory effect dominates; see figure 9(b). Distinguishing between the effects of ϕ and d is much more difficult in this case, consequently the posteriors are much more dependent. In cases where there is significant correlation between d and ϕ, it arguably makes little sense to consider only the marginal posterior distribution of d. For example the 95% credibility interval for d from plots (b) is $(-0.473, -0.247)$, and the corresponding interval for ϕ is $(-0.910, -0.753)$, yet these clearly give a rather pessimistic view of our joint knowledge about d and ϕ—see figure 9(c). In theory an ellipsoidal credibility set could be constructed, although this is clearly less practical when $\dim \boldsymbol{\omega} > 2$.



(a) (b)

Figure 8: Posterior samples of (d, ϕ); input time series (a) $(1 + 0.92\mathcal{B})(1 - \mathcal{B})^{0.25} X_t = \varepsilon_t$, (b) $(1 - 0.83\mathcal{B})(1 - \mathcal{B})^{-0.35} X_t = \varepsilon_t$.

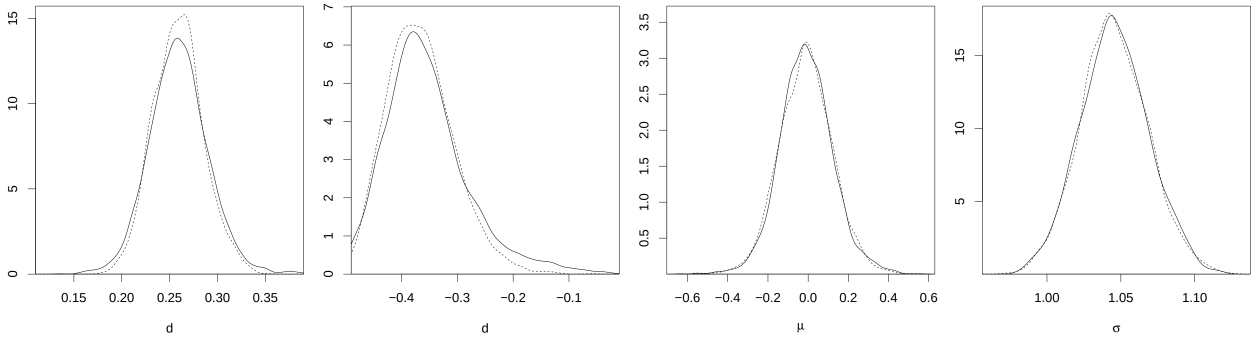


(a) (b) (c)

Figure 9: Spectra for processes in figure 8. Green line is relevant ARMA$(1, 0)$ process, red line is relevant ARFIMA$(0, d, 0)$ process, black line is ARFIMA$(1, d, 0)$ process; (a) $(1 + 0.92\mathcal{B})(1 - \mathcal{B})^{0.25}X_t = \varepsilon_t$; (b) $(1 - 0.83\mathcal{B})(1 - \mathcal{B})^{-0.35}X_t = \varepsilon_t$. Pane (c) shows posterior samples of (d, ϕ) from series considered in pane (b) with credibility sets: red is 95% credibility set for (d, ϕ), green is 95% credibility interval for d, blue is 95% credibility interval for ϕ.

Unknown form: The RJ scheme outlined in section 5.2 works well for data simulated with p and q up to 3. The marginal posteriors for d are generally roughly centred around d_I (the data generating value) and the modal posterior model probability is usually the 'correct' one. To illustrate, consider again the two example data generating contexts used above.

For both series, kernel density for the marginal posterior for d are plotted in figure 10(a)–(b), together with the equivalent density estimated assuming unknown model orders. Notice



(a) (b) (c) (d)

Figure 10: Marginal posterior density of d from series in figure 8, (a)–(b) respectively. Solid line is density obtained using reversible-jump algorithm. Dotted line is density obtained using fixed $p = 1$ and $q = 0$. Panels (c)–(d) shows the posterior densities for μ and σ, respectively, corresponding to the series in 8(a); those for 8(b) look similar.

how the densities obtained via the RJ method are very close to those obtained assuming $p = 1$ and $q = 0$. The former are slightly more heavy-tailed, reflecting a greater level of uncertainty about d. Interestingly, the corresponding plots for the posteriors of μ and σ do not appear to exhibit this effect—see figure 10(c)–(d). The posterior model probabilities are presented in table 2, showing that the 'correct' modes are being picked up consistently.

Table 2: Posterior model probabilities for time series from figures 8(a)–(b) and 10(a)–(b).

	$p\backslash q$	0	1	2	3	4	5	marginal
(a)	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	**0.805**	0.101	0.003	0.000	0.000	0.000	0.908
	2	0.038	0.043	0.001	0.000	0.000	0.000	0.082
	3	0.005	0.004	0.000	0.000	0.000	0.000	0.009
	4	0.000	0.001	0.000	0.000	0.000	0.000	0.001
	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	marginal	0.848	0.148	0.004	0.000	0.000	0.000	

	$p\backslash q$	0	1	2	3	4	5	marginal
(b)	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	**0.829**	0.125	0.002	0.000	0.000	0.000	0.956
	2	0.031	0.013	0.000	0.000	0.000	0.000	0.044
	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	marginal	0.860	0.138	0.002	0.000	0.000	0.000	

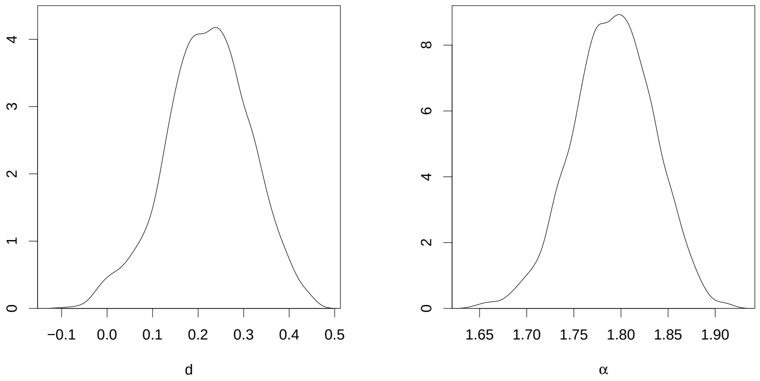
As a test of the robustness of the method, consider a complicated short memory input combined with a heavy tailed α-stable innovations distribution. Specifically, the time series that will be used is the following ARFIMA$(2, d, 1)$ process

$$\left(1 - \frac{9}{16}\mathcal{B}^2\right)(1 - \mathcal{B})^{0.25} X_t = \left(1 + \frac{1}{3}\mathcal{B}\right)\varepsilon_t, \qquad \text{where } \varepsilon_t \sim \mathcal{S}_{\alpha=1.75,0}. \tag{23}$$

For more details, see (Graves, 2013, §7.1). The marginal posterior densities of d and α are presented in figure 11. Performance looks good despite the complicated structure. The posterior estimate for d is $\widehat{d}^{(B)} = 0.22$, with 95% CI $(0.04, 0.41)$. Although this interval is admittedly rather wide, it is reasonably clear that long memory is present in the signal. The corresponding interval for α is $(1.71, 1.88)$ with estimate $\widehat{\alpha}^{(B)} = 1.79$. Finally, we see from table 3 that the algorithm is very rarely in the 'wrong' model.

The Nile Data: We conclude with an application of our methods to the famous annual Nile minima data. Because of the fundamental importance of the river to the civilisations it has supported, local rulers kept measurements of the annual maximal and minimal heights obtained by the river at certain points (called gauges). The longest uninterrupted sequence of recordings is from the Roda gauge (near Cairo), between 622 and 1284 AD ($n = 663$).[5]

[5] There is evidence (e.g. Ko and Vannucci, 2006b) that the sequence is not actually homogeneous.



(a) (b)

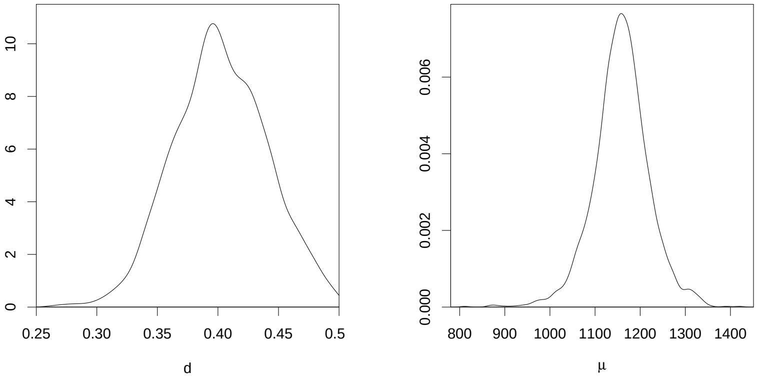
Figure 11: Marginal posterior densities (a) d, (b) α.

Table 3: Posterior model probabilities.

$p\backslash q$	0	1	2	3	4	5	marginal
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	**0.822**	0.098	0.001	0.000	0.000	0.921
3	0.014	0.056	0.004	0.000	0.000	0.000	0.075
4	0.003	0.001	0.000	0.000	0.000	0.000	0.004
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
marginal	0.017	0.880	0.102	0.002	0.000	0.000	

The posterior summary statistics and marginal densities of d and μ for the Nile data are presented in figure 12. Posterior model probabilities are presented in table 4. We see that

	mean	95% CI	
d	0.039	0.336	0.482
μ	62	1037	1284
σ	1.91	66.46	73.97



(a) (b)

Figure 12: *Table:* Summary posterior statistics for Nile minima. *Plots:* Marginal posterior densities for Nile minima; (a) d, (b) μ.

Table 4: Posterior model probabilities for Nile minima.

$p\backslash q$	0	1	2	3	4	5	marginal
0	**0.638**	0.101	0.010	0.000	0.000	0.000	0.750
1	0.097	0.124	0.011	0.000	0.000	0.000	0.232
2	0.007	0.010	0.000	0.000	0.000	0.000	0.018
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
marginal	0.742	0.236	0.022	0.000	0.000	0.000	

the model with the highest posterior probability is the ARFIMA$(0, d, 0)$ model with $d \approx 0.4$. This suggests a strong, 'pure', long memory feature. Our results compare favourably with other studies (Liseo et al., 2001; Hsu and Breidt, 2003; Ko and Vannucci, 2006a).

7 Discussion

We have provided a systematic treatment of efficient Bayesian inference for ARFIMA models, the most popular parametric model combining long and short memory effects. Through a mixture of theoretical and empirical work we have demonstrated that the methods can handle the sorts of time series data that are typically confronted with possible long memory in mind.

Many of the choices made throughout, but in particular those leading to our likelihood approximation stem from a need to accommodate further extension. For example, in future work we intend to extend them to cope with a heavy-tailed innovations distribution. For more evidence of potential in this context, see Graves (2013, §7). Along similar lines, there is scope for further generalisation to incorporate seasonal (long memory) effects. Finally, an advantage of the Bayesian approach is that it provides a natural mechanism for dealing with missing data, via data augmentation. This is particularly relevant for long historical time series which may, for a myriad of reasons, have recording gaps. For example, some of the data recorded at other gauges along the river Nile have missing observations although otherwise span a similarly long time frame. For a demonstration of how this might fit within our framework, see §5.6 of Graves dissertation.

A Gibbs sampling of μ and σ

Assuming Gaussianity, and the Gaussian and root-inverse-gamma independent priors (deliberately chosen to ensure prior-posterior conjugacy), it is possible to use Gibbs updating for the parameters μ and σ. Demonstrating this requires the following result: If $g(x) \propto \exp\left[-\frac{1}{2}(\alpha x^2 - 2\beta x)\right]$, and $\int_{\mathbb{R}} g(x)\, dx = 1$, then $g \sim \mathcal{N}(\frac{\beta}{\alpha}, \frac{1}{\alpha})$.

Now, let $\bar{c} =: \frac{1}{n}\sum_{t=1}^{n} c_t$. Then, updating μ by combining its prior with the approximate (log) likelihood (17), yields the following conditional posterior:

$$\begin{aligned}\pi_{\mu|\psi_{-\mu}}(\mu|\boldsymbol{\psi}_{-\mu},\mathbf{x}) &\propto \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}\sigma^{-n}\prod_{t=1}^{n}\left\{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(c_t-\Pi_P\mu)^2}{2\sigma^2}\right]\right\}\\ &\propto \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(c_t-\Pi_P\mu)^2\right\}\\ &\propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2-\frac{1}{2\sigma^2}\left(n\Pi_P^2\mu^2-2\mu\Pi_P n\bar{c}\right)\right\}\\ &\propto \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{\sigma_0^2}+\frac{n\Pi_P^2}{\sigma^2}\right)-2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\Pi_P n\bar{c}}{\sigma^2}\right)\right]\right\},\end{aligned}$$

Then, our result reveals

$$\mu|\boldsymbol{\psi}_{-\mu},\mathbf{x}\sim\mathcal{N}\left(\left[\frac{1}{\sigma_0^2}+\frac{n\Pi_P^2}{\sigma^2}\right]^{-1}\left[\frac{\mu_0}{\sigma_0^2}+\frac{\Pi_P n\bar{c}}{\sigma^2}\right],\left[\frac{1}{\sigma_0^2}+\frac{n\Pi_P^2}{\sigma^2}\right]^{-1}\right). \tag{24}$$

When using the exact likelihood method, we obtain

$$\mu|\boldsymbol{\psi}_{-\mu},\mathbf{x}\sim\mathcal{N}\left(\left[\frac{1}{\sigma_0^2}+\frac{\mathbf{1}_n^t\Sigma_d^{-1}\mathbf{1}_n}{\sigma^2}\right]^{-1}\left[\frac{\mu_0}{\sigma_0^2}+\frac{\mathbf{x}^t\Sigma_d^{-1}\mathbf{1}_n}{\sigma^2}\right],\left[\frac{1}{\sigma_0^2}+\frac{\mathbf{1}_n^t\Sigma_d^{-1}\mathbf{1}_n}{\sigma^2}\right]^{-1}\right). \tag{25}$$

In the limiting case of the flat prior, the conditional density (24) becomes:

$$\mu|\boldsymbol{\psi}_{-\mu},\mathbf{x}\sim\mathcal{N}\left(\frac{\bar{c}}{\Pi_P},\frac{\sigma^2}{n\Pi_P}\right),\quad\text{and (25) is}\quad\mu|\boldsymbol{\psi}_{-\mu},\mathbf{x}\sim\mathcal{N}\left(\frac{\mathbf{x}^t\Sigma_d^{-1}\mathbf{1}_n}{\mathbf{1}_n^t\Sigma_d^{-1}\mathbf{1}_n},\frac{\sigma^2}{\mathbf{1}_n^t\Sigma_d^{-1}\mathbf{1}_n}\right).$$

Likewise, when updating σ we obtain the conditional posterior:

$$\begin{aligned}\pi_{\sigma|\psi_{-\sigma}}(\sigma|\boldsymbol{\psi}_{-\sigma},\mathbf{x}) &\propto \frac{2}{\Gamma(\alpha_0)}\beta_0^{\alpha_0}\sigma^{-(2\alpha_0+1)}\exp\left(-\frac{\beta_0}{\sigma^2}\right)\sigma^{-n}\prod_{t=1}^{n}\left\{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(c_t-\Pi_P\mu)^2}{2\sigma^2}\right]\right\}\\ &\propto \sigma^{-(2\alpha_0+n+1)}\exp\left\{-\frac{1}{\sigma^2}\left[\beta_0+\frac{1}{2}\sum_{t=1}^{n}(c_t-\Pi_P\mu)^2\right]\right\},\quad\text{giving}\end{aligned}$$

$$\sigma|\boldsymbol{\psi}_{-\sigma},\mathbf{x}\sim\mathcal{R}\left(\alpha_0+\frac{n}{2},\beta_0+\frac{1}{2}\sum_{t=1}^{n}(c_t-\Pi_P\mu)^2\right). \tag{26}$$

Similarly, when using the exact likelihood (13):

$$\sigma|\boldsymbol{\psi}_{-\sigma},\mathbf{x}\sim\mathcal{R}\left(\alpha_0+\frac{n}{2},\beta_0+\frac{1}{2}Q(\mathbf{x}|\mu,d)\right), \tag{27}$$

Finally, the limiting case of the diffuse prior yields we obtain the following for (26) and (27)

$$\sigma|\boldsymbol{\psi}_{-\sigma},\mathbf{x}\sim\mathcal{R}\left(\frac{n}{2},\frac{1}{2}\sum_{t=1}^{n}(c_t-\Pi_P\mu)^2\right)\quad\text{and}\quad\sigma|\boldsymbol{\psi}_{-\sigma},\mathbf{x}\sim\mathcal{R}\left(\frac{n}{2},\frac{1}{2}Q(\mathbf{x}|\mu,d)\right).$$

B Preservation of ACF under bijection

Consider the bijection from Monahan (1984) in (21). Now, suppose we have the vector $\boldsymbol{\varphi}^{(p)} = (\varphi_1^{(p)}, \ldots, \varphi_{p-1}^{(p)}, \varphi_p^{(p)})$, and its truncated version:

$$\boldsymbol{\varphi}^{(p-1)} = (\varphi_1^{(p)}, \ldots, \varphi_{p-1}^{(p)}) =: (\varphi_1^{(p-1)}, \ldots, \varphi_{p-1}^{(p-1)}).$$

Let the 'classical' representations of these vectors be denoted $\boldsymbol{\phi}^{(p)}$ and $\boldsymbol{\phi}^{(p-1)}$ respectively, and note that $(\phi_1^{(p-1)}, \ldots, \phi_{p-1}^{(p-1)}) \neq (\phi_1^{(p)}, \ldots, \phi_{p-1}^{(p)})$. These define two polynomials $\Phi^{(p)}$ and $\Phi^{(p-1)}$, so a natural question to ask is: What is the relationship between $\Phi^{(p)}$ and $\Phi^{(p-1)}$?

Theorem 1. *Consider the pair of AR(p) and AR(p − 1) models, with polynomials $\Phi^{(p)}$ and $\Phi^{(p-1)}$ as defined above. Denote the ACV of the AR(p) model by $\gamma(\cdot)$, and that of the AR(p − 1) model by $\gamma'(\cdot)$. Then for $0 \leq k < p$: $\gamma'(k) = \gamma(k)\left[1 - \left(\phi_p^{(p)}\right)^2\right]$.*

Proof. Our strategy is to show that $\gamma'(\cdot)$ satisfies the relevant Yule–Walker equations, and then appeal to uniqueness. Firstly, from (21) we have that $\left[1 - \left(\phi_p^{(p)}\right)^2\right]\phi_i^{(p-1)} = \phi_i^{(p)} - \phi_p^{(p)}\phi_{p-i}^{(p)}$, $i = 1, \ldots, p-1$. (Note that this also is valid for special case of $i = 0$.) We are given (Brockwell and Davis, 1991, §3.3) that the Yule–Walker equations for $\boldsymbol{\phi}^{(p)}$ are $\sum_{i=0}^{p} \phi_i^{(p)} \gamma(k-i) = \sigma^2 \mathbb{I}_{\{k=0\}} i$. Now consider the following for $k = 0, \ldots, p-1$:

$$\begin{aligned}
\sum_{i=0}^{p-1} \phi_i^{(p-1)} \gamma'(k-i) &= \sum_{i=0}^{p-1} \phi_i^{(p-1)} \gamma(k-i)\left[1 - \left(\phi_p^{(p)}\right)^2\right] = \sum_{i=0}^{p-1} \gamma(k-i)\left(\phi_i^{(p)} - \phi_p^{(p)}\phi_{p-i}^{(p)}\right) \\
&= \sum_{i=0}^{p} \gamma(k-i)\left(\phi_i^{(p)} - \phi_p^{(p)}\phi_{p-i}^{(p)}\right) - \gamma(k-p)\left(\phi_p^{(p)} - \phi_p^{(p)}\phi_{p-p}^{(p)}\right) \\
&= \sum_{i=0}^{p} \phi_i^{(p)} \gamma(k-i) - \phi_p^{(p)} \sum_{i=0}^{p} \phi_i^{(p)} \gamma((p-k)-i).
\end{aligned}$$

Using Yule–Walker when $k = 0$, the last is $\sigma^2 - \phi_p^{(p)} 0 = \sigma^2$. And for $k = 1, \ldots, p-1$ it equals $0 - \phi_p^{(p)} 0 = 0$ (because $k = 1, \ldots, p-1$ corresponds to $p - k = p-1, \ldots, 1$). So we have that:

$$\sum_{i=0}^{p-1} \phi_i^{(p-1)} \gamma'(k-i) = \left\{ \begin{array}{lll} \sigma^2 & \text{for} & k = 0 \\ 0 & \text{for} & k = 1, \ldots, p-1 \end{array} \right. .$$

These equations are the Yule–Walker equations for the AR($p-1$) process so we are done. □

Corollary 1. *Under Thm. 1 assumptions, ACFs of $\boldsymbol{\phi}^{(p)}$ and $\boldsymbol{\phi}^{(p-1)}$ are equal for $0 \leq k < p$.*

Proof.

$$\rho'(k) = \frac{\gamma'(k)}{\gamma'(0)} = \frac{\gamma(k)\left[1 - \left(\phi_p^{(p)}\right)^2\right]}{\gamma(0)\left[1 - \left(\phi_p^{(p)}\right)^2\right]} = \frac{\gamma(k)}{\gamma(0)} = \rho(k).$$

□

C Positive definiteness of $\Sigma_{\boldsymbol{\varpi}}$

Lemma 1. *Suppose A is an $M \times M$ matrix which can be block divided as:*

$$A \quad = \quad \left(\begin{array}{c:c} B_{m,m} & B_{m,m'} \\ \hdashline B_{m',m} & B_{m',m'} \end{array}\right),$$

where $M = m + m'$, and $B_{m,m'} = B^t_{m',m}$. Furthermore, let $\mathbf{x} \in \mathbb{R}^M$ be block divided as $\mathbf{x}^t = \left(\begin{array}{c:c} \mathbf{y}^t & \mathbf{z}^t \end{array}\right)$, where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^{m'}$. Then:

$$\mathbf{x}^t A\mathbf{x} = \mathbf{y}^t B_{m,m}\mathbf{y} + 2\mathbf{y}^t B_{m,m'}\mathbf{z} + \mathbf{z}^t B_{m',m'}\mathbf{z}. \tag{28}$$

Theorem 2. *Let Σ_M be an $M \times M$ positive-definite matrix, and let Σ_m be the leading m-submatrix of Σ_M. Then Σ_m is also positive definite.*

Proof. Σ_M is positive definite so $\mathbf{x}^t\Sigma_M\mathbf{x} > 0$, for all $\mathbf{x} \in \mathbb{R}^M$. In particular, this is true for all block-divided $\mathbf{x}^t = \left(\begin{array}{c:c} \mathbf{y}^t & \mathbf{0}^t \end{array}\right)$, where $\mathbf{y} \in \mathbb{R}^m$. Then from (28):

$$0 < \mathbf{x}^t\Sigma_M\mathbf{x} = \mathbf{y}^t\Sigma_m\mathbf{y} + 2\mathbf{y}^t B_{m,m'}\mathbf{0} + \mathbf{0}^t B_{m',m'}\mathbf{0} = \mathbf{y}^t\Sigma_m\mathbf{y}.$$

Clearly therefore, $\mathbf{y}^t\Sigma_m\mathbf{y} > 0$ for all $\mathbf{y} \in \mathbb{R}^m$, i.e. Σ_m is positive-definite. □

Theorem 3. *Suppose the $m \times m$ matrix Σ_m and the $n \times n$ matrix Σ_n are positive definite. Then the following $(m + n) \times (m + n)$ matrix is also positive definite:*

$$\Sigma_{m,n} \quad = \quad \left(\begin{array}{c:c} \Sigma_m & \mathbf{O} \\ \hdashline \mathbf{O} & \Sigma_n \end{array}\right).$$

Proof. Let $\mathbf{x} \in \mathbb{R}^{m+n}$ be partitioned as $\mathbf{x}^t = \left(\begin{array}{c:c} \mathbf{y}^t & \mathbf{z}^t \end{array}\right)$ where $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^n$. Then from (28), $\mathbf{x}^t\Sigma_{m,n}\mathbf{x} = \mathbf{y}^t\Sigma_m\mathbf{y} + 2\mathbf{y}^t\mathbf{O}\mathbf{z} + \mathbf{z}^t\Sigma_n\mathbf{z} = \mathbf{y}^t\Sigma_m\mathbf{y} + \mathbf{z}^t\Sigma_n\mathbf{z}$. And because these two terms are positive, we have that $\mathbf{x}^t\Sigma_{m,n}\mathbf{x} > 0$, i.e. $\Sigma_{m,n}$ is positive-definite. □

References

Abry, P., Flandrin, P., Taqqu, M. S., and Veitch, D. (2003). "Self-similarity and long-range dependence through the wavelet lens." In Doukhan et al. (2003), 527–556.

Adenstedt, R. K. (1974). "On Large-sample Estimation for the Mean of a Stationary Random Sequence." *The Annals of Statistics*, 2, 1095–1107.

Barnes, J. A. and Allan, D. W. (1966). "A statistical model of flicker noise." *Proceedings of the IEEE*, 54, 2, 176–178.

Beran, J. (1994a). "On a Class of M-Estimators for Gaussian Long-Memory Models." *Biometrika*, 81, 4, 755–766.

— (1994b). *Statistics for Long Memory Processes*. Chapman & Hall, New York.

Beran, J., Feng, Y., Ghosh, S., and Kulik, R. (2013). *Long Memory Processes*. Springer, Heidelberg.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. 2nd ed. Springer; New York.

Chen, W. W., Hurvich, C. M., and Lu, Y. (2006). "On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series." *J. of the American Statistical Association*, 101, 474, 812–822.

Doukhan, P., Oppenheim, G., and Taqqu, M. S., eds. (2003). *Theory and Applications of Long-Range Dependence*. Birkhäuser; Boston, MA.

Ehlers, R. S. and Brooks, S. P. (2006). "Bayesian analysis of order uncertainty in ARIMA models." Tech. rep., Department of Statistics, Federal University of Paraná.

— (2008). "Adaptive proposal construction for reversible jump MCMC." *Scandanavian J. of Statistics*, 35, 4, 677–690.

Eğrioğlu, E. and Günay, S. (2010). "Bayesian model selection in ARFIMA models." *Expert Systems with Applications*.

Fraley, C., Leisch, F., Maechler, M., Reisen, V., and Lemonte, A. (2012). `fracdiff`*: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models*. R package version 1.4-1.

Franzke, C. L. E., Graves, T., Watkins, N. W., Gramacy, R. B., and Hughes, C. (2012). "Robustness of estimators of long-range dependence and self-similarity under non-Gaussianity." *Phil. Trans. of the Royal Society A*, 370, 1962, 1250–1267.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2012). mvtnorm*: Multivariate Normal and t Distributions*. R package version 0.9-9992.

Geweke, J. and Porter-Hudak, S. (1983). "The estimation and application of long-memory time series models." *J. of Time Series Analysis*, 4, 4, 221–238.

Granger, C. W. J. and Joyeux, R. (1980). "An Introduction to Long-memory Time Series Models and Fractional Differencing." *J. of Time Series Analysis*, 1, 1, 15–29.

Graves, T. (2013). "A systematic approach to Bayesian inference for long memory processes." Ph.D. thesis, University of Cambridge, UK.

Graves, T., Gramacy, R. B., Watkins, N. W., and Franzke, C. (2014). "A brief history of long memory." Tech. rep., The University of Chicago, http://arxiv.org/abs/1406.6018.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82, 4, 711–732.

Haslett, J. and Raftery, A. E. (1989). "Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource." *J. of the Royal Statistical Society, Series C*, 38, 1, 1–50. With discussion.

Holan, S., McElroy, T., and Chakraborty, S. (2009). "A Bayesian Approach to Estimating the Long Memory Parameter." *Bayesian Analysis*, 4, 159–190.

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

Hsu, N.-J. and Breidt, F. J. (2003). "Bayesian Analysis of Fractionally Integrated ARMA with additive noise." *J. of Forecasting*, 22, 6-7, 491–514.

Hurst, H. E. (1951). "Long-term storage capacity of reservoirs." *Transactions of the American Society of Civil Engineers*, 116, 770–808. With discussion.

Kay, S. M. and Marple, S. L. (1981). "Spectrum analysis-A modern perspective." *Proceedings of the IEEE.*, 69, 11, 1380–1419.

Ko, K. and Vannucci, M. (2006a). "Bayesian wavelet analysis of autoregressive fractionally integrated moving-average processes." *J. of Stat. Plan. and Inf.*, 136, 10, 3415–3434.

— (2006b). "Bayesian Wavelet-Based Methods for the Detection of Multiple Changes of the Long Memory Parameter." *IEEE Transactions on Signal Processing*, 54, 11, 4461–4470.

Koop, G., Ley, E., Osiewalski, J., and Steel, M. F. (1997). "Bayesian analysis of long memory and persistence using ARFIMA models." *J. of Econometrics*, 76, 1-2, 149–169.

Liseo, B., Marinucci, D., and Petrella, L. (2001). "Bayesian semiparametric inference on long-range dependence." *Biometrika*, 88, 4, 1089–1104.

Mandelbrot, B. B. and Van Ness, J. W. (1968). "Fractional Brownian Motions, Fractional Noises and Applications." *SIAM Review*, 10, 4, 422–437.

Mandelbrot, B. B. and Wallis, J. R. (1968). "Noah, Joseph and operational hydrology." *Water Resources Research*, 4, 5, 909–918.

Marriott, J. M., Ravishanker, N., Gelfand, A. E., and Pai, J. S. (1995). "Bayesian analysis for ARMA processes: complete sampling based inference under exact likelihoods." In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, K. Chaloner, and J. Geweke, 243–256. Wiley, New York.

McLeod, A. I., Yu, H., and Krougly, Z. (2007). "Algorithms for Linear Time Series Analysis: With R Package." *J. of Statistical Software*, 23, 5.

Monahan, J. F. (1984). "A note on enforcing stationarity in autoregressive-moving average models." *Biometrika*, 71, 2, 403–404.

Pai, J. S. and Ravishanker, N. (1998). "Bayesian analysis of autoregressive fractionally integrated moving-average processes." *J. of Time Series Analysis*, 19, 1, 99–112.

Palma, W. (2007). *Long Memory Time Series*. Wiley.

Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). "Mosaic organization of DNA nucleotides." *Physical Review E*, 49, 2, 1685–1689.

Robert, C. P. (1995). "Simulation of truncated normal variables." *Statistics and Computing*, 5, 121–125.

Roberts, G. O. and Rosenthal, J. S. (2001). "Optimal Scaling for Various Metropolis–Hastings Algorithms." *Statistical Science*, 16, 4, 351–367.

Vermaak, J., Andrieu, C., Doucet, A., and Godsill, S. (2004). "Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes." *J. of Time Series Analysis*, 25, 6, 785–809.

Vyushin, D., Mayer, J., and Kushner, P. (2009). `PowerSpectrum`*: Spectral Analysis of Time Series*. `http://www.atmosp.physics.utoronto.ca/people/vyushin/mysoftware.html`. R package version 0.3.

Wuertz, D. (2012). `fArma`*: ARMA Time Series Modelling*. R package version 2160.77.