Response to reviewers comments

Response to „Interactive comment on “Efficient Bayesian inference for ARFIMA processes” by T. Graves et al.“ by Anonymous Referee #1

The reviewers comments are in italics while our responses are in normal font.

B. Recommendation
In my opinion, the method presented in this paper is a novel way of applying Bayesian reasoning thoroughly to the chosen problem of inferring LRD in time series data. The authors have done a thorough job of grounding their approach on a rigorous mathematical foundation and presenting the theory unambiguously and carefully.
That being said, the chosen geo-scientific example in the manuscript, viz. the Nile river minima time series was given only a short, cursory treatment and no effort was made to infer the results obtained by the method in this case. Given the focus of NPG and its target audience of geoscientists, I feel that the authors could move some of the mathematically intensive sections of their method to an appendix and thereafter devote more space and attention to the application of the method to the Nile data and how the method relates to physically understandable features of the Nile river system.
I thus recommend publication in NPG only after a few major shifts in presentation and focus of the manuscript, which I hope the authors can address with a major revision. Besides these points, there are certain technical issues with the presentation that I list out below. I hope that these suggestions help the authors to better their manuscript.

Thank you very much for your positive comments. We have improved the presentation and moved some mathematical parts into an appendix. We also enhanced the geophysical examples by including an analysis of the Central England Temperature time series.

C. General Comments
1. A first major point that could be improved is the focus given to the application of the method developed by the authors to real data. In the current version, the manuscript gives much space to the theory and constructed, illustrative examples which, even though necessary and absolutely crucial, should ideally in the end lead to an equally detailed example from the real-world with a discussion on how the method performs in that case. I feel that, given the particular focus of NPG, which is at the juncture of theory and application, it is important to properly discuss the relevant real-world implications of the results obtained when applying the method to the Nile river minima for instance. This is lacking in the current version. One such difficulty for me in understanding the final application was that I was unable to relate the parameters p, d, and q to the physical system that gave rise to the Nile data set.

As stated above we have included a new section were we discuss the Central England Temperature.
A physical interpretation of the parameters d, p and q has been elusive as the model was originally introduced in econometrics as phenomenological rather than physical or structural. The simplest explanation, which is generic rather than application specific, is that d describes the behavior on very long time scales whereas p and q describe fluctuations on shorter time scales. However, very recent progress is being made in statistics and physics on bridging continuous time linear dynamical systems and the discrete time ARFIMA models. We have noted this new work by citing Ślęzak and Weron [2015] as a representative entry point.

2. Another main issue for me was that there were discrepancies between the title, abstract, and the numerical examples in the text in terms of the main focus. The title indicated that the manuscript is about "Bayesian inference for ARFIMA processes", the abstract suggested that the main focus was LRD inference, and finally the text devoted quite some space to "short memory" as well. The authors should take care to clearly state what are their objectives (which they seem to do even now) and thereafter, relate every new topic/application/idea later in the text to the stated objectives (which is lacking).

We have changed the manuscript title to „Efficient Bayesian inference for natural time series exhibiting both Short- and Long-Memory using a parametric ARFIMA process“. We also have changed long-range dependence to long-memory throughout the manuscript.

3. A last major issue that I have with the manuscript is that the Tables and Figures do not have adequate captioning. The table captions do not indicate what are the various parameters represented and for what kind of numerical experiments. The figures do not have a clear "visual" legend and neither do the captions indicate what kind of numerical analysis gave rise to the figure. I feel that this is an extremely crucial part in communicating results.

We agree with the reviewer that captions are very important. We have improved the captions of all tables and figures.

D. Specific / Technical comments
1) TITLE: Perhaps the title is a too vague. Maybe the authors can consider giving a more informative title?

We have changed the title into „Efficient Bayesian inference for natural time series exhibiting both Short- and Long-Memory using a parametric ARFIMA process“.

2) P574, L7: "Rather than Mandelbrot's fractional Gaussian noise ..." There is no clear comparison of the method from this study to Mandelbrot's fractional noise. What is the purpose of mentioning it here?

We have deleted this statement.

3) P574, L15: "We illustrate our new methodology on the Nile water level data..." What are the main results? How do they enhance our understanding of the Nile river data and LRD?

Thanks for this comment. We have augmented the abstract to explain that we treat the Nile data and as a new addition the Central England Temperature time series, noting that comparisons are favorable and that the CET involves an extension to seasonal long memory. Summarizing the results for both analysis seemed inappropriate for the Abstract section. However we have added some brief discussion into our Introduction Section.

4) P574, L26: "A standard definition..." even so, please cite a reference for this definition.

We now have added a few references here.

5) P576, L1: "It would offer the ability to marginalize out aspects of a model apparatus and data, such as short memory..." Perhaps this idea is key to the extensions of short

memory later in the text? If so, the authors should address this is more detail here and later as well.

Yes, this is a key idea. We have modified some of the text, in the following two paragraphs actually, to clarify that designing a computationally efficient means of marginalizing out nuiscance parameters (seasonal and short memory terms) is a novel and important contribution in the paper. However, rather than additing detail into an already long introduction section, we reiterate later in the paper (just before Section 4.1) the importance of this marginalization.

6) P576, L5: ARFIMA has not been defined before this in the main text (only in the abstract).

We now also define ARFIMA in the text. See line 52.

7) P576, L15: Why is it necessary to include/Cite statements about something being "too hard to work with"?

We deleted this statement.

8) P576, L18: "many of the above drawbacks..." It is not clear to me how many drawbacks were mentioned before this line.

We have changed this into „that the above drawbacks ...“

9) P577, L18: "2 Time series definitions and the ARFIMA model" I personally feel that barring the definition of ARFIMA processes, the rest can moved to an appendix.

We have moved most of this section into appendix A.

10) P577, L22: You defined covariance function ngamma (k) as "Cov(X_t, X_{t + k})" but what is "Cov". This definition is ambiguous and unclear.

„Cov“ is the standard covariance function in statistics, which can be found on Wikipedia under the heading „Covariance“.

11) P577, L25: "the “backshift” operator" - Here and later, the use of quotes for terms and definitions are a bit distracting. Maybe the authors can find some other way of emphasis?

We have deleted most quotes for terms and definitions.

12) P579, L11: I understand that this is the LRD parameter "d" being referred to here, and which was defined in the Introduction. But I find it hard to relate the $|d| < 1/2$ statement to the preceding equation, i.e., Eq. 5.

Thanks for this comment. This statement, and the surrounding discussion, has been moved into the appendix. We have adjusted the text here to clarify that we are referring back to Equation (1), which defines the ARFIMA process, keeping in mind the immediately previous discussion on ACV/ACFs.

13) P580, L15: "Choosing p = q = 0 recovers FI(d) ARFIMA(0, d , 0)". Is this the

definition of ARFIMA? If so, maybe it is possible to add a line stating this more clearly?

As noted above, ARFIMA is now defined in Equation (1), and the rest of this discussion occurs in the Appendix, i.e., much later.

14) P582, L1: "3 Likelihood evaluation for Bayesian inference". I feel it might be better if the authors added a few words here about the interpretation of the parameters p, d, and q of the ARFIMA process here (even at the cost of repetition) before starting with the model inference part. Also, why is the starting point of the inference is a ARFIMA(0, d,0) process and not some other value of p and q? What kind of a process is this?

Thanks. We now clarify that the phrase „having no short-ranged components“ is what we mean by p=q=0 and that this corresponds to a fractionally integrated process. For details we refer the reader to our new Appendix. This restriction is made for simplicitiy: if p or q were non-zero the process would be more complicated. This is, or course, generalized later.

15) P582, L9: "causal" in what sense? Maybe briefly mention here.

„Causal“ is a time series term. We povide a definition in our appendix, although it can be found in any standard time series text, e.g., Brockwell & Davis.

16) P582, L24: It might be ambiguous to infer AR(P) dependence from a time series of length n = P.

If by ambiguous you mean that there are too many degrees of freedom for accurate inference (of the variance for example), then this is of course correct under least squares or maximum likelihood inference. However, under a proper prior for the AR coefficients, or indeed one that encourages sparsity, there is of course no problem technically with Bayesian inference. It is possible, say if seasonal effects are present in the data but not explicitly accounted for in the model, then identifiability might be a concern. But it is not more of a concern than it would be in a P < n setting.

17) P588, L5: "with chains moving between and within models..." It is unclear what this means for someone unfamiliar with numerical techniques of Bayesian likelihood estimation and MCMC methods.

Thanks. We agree that it would help to be more explicit by linking „between“ moves to choices of p and q and „within“ moves to inference for phi and theta given p and q. The text has been adjusted accordingly.

18) P588, L7: What is "RJ"?

RJ denotes Reversible Jump. This is defined on line 219.

19) P588, L9: What is "FEXP"?

FEXP stands for Fractional Exponential Process. See line 224.

20) P589, L20: Does "MVN" mean Multivariate Normal"?

Yes. This is clarified where it is first used in Section 4.1.

21) P591, L11: What is a transdimensional move?

The only mention of „transdimensional“ that we could find is in Section 1 where the text reads „transdimensional MCMC, in which the model order (the p and q parameters in the ARFIMA model varies and, thus, the dimension of the problem). In other words this is the same as a between model move, in point 17. To better connect the two passages we have added text to the passage referenced above to link „transdimentional“ to „between“ model moves.

22) P593, L3: The authors should maybe add one more line on why they choose the Poisson distribution here, it is not clear from the parsimony argument they mention.

In the previous sentence we clarify that the more „complicated“ models are „larger“ ones (i.e., bigger p and/or q). Therefore, a prior that prefers parsimony is one that puts more weight on smaller models. We then adjust the text to read that „As a simple representative of potential priors that give greater weight to smaller models we prefer a truncated joint Poisson ...“ In other words, the particular form of the Poisson isn't important.

23) P595, L22: Is it not possible to show the results in a graphical way, such as a histogram instead of a table?

We think that our table provides a very concise presentation of our test results and would prefer to keep it. However, if the editor feels otherwise we are happy the exchange the table with a histogram.

24) P596, L9: Figure 1 legend is unclear. If I understand correctly, for each value of d_l, there are 1024 estimates and thus 1024 "x" markers on the vertical axis - is this correct? I suggest to use better captions to remove such ambiguities.

We have enhanced the caption of this figure to better describe the plot axis labels and the x-markers. Thanks.

25) P600, L8: For the sake of reproducibility, I suggest that the authors state the source of the data and from where it was obtained. Also, maybe they should devote a few lines on the nature of the data, and preprocessing of the data such as removal of missing values, outliers and the like.

The Nile data is part of the R package 'longmemo' and the CET time series can be downloaded from http://www.metoffice.gov.uk/hadobs/hadcet/
We now state this in the manuscript.

26) P601, L11: Reference missing to the dissertation cited here.

We have added this reference.

Response to „Interactive comment on “Efficient Bayesian inference for ARFIMA processes” by T. Graves et al.“ by M. Crucifix (Referee)

1) p. 578 : Introduce new line before "fXtg is said to be an auto-"

Done.

2) p. 579 : Before the "restriction |d|<1/2" ... : the condition sounds awkward given that the previous paragraphs concerns the ARMA process and not ARFIMA. Introduce a sentence clarifying that we return to the discussion of the more general ARFIMA process.

Done. Much of this has been moved to the appendix, and extra text has been added to clarify which part of that expression applies to the ARFIMA model, Eq (1), and which to the ACF discussion preceeding.

3) p. 581 : approximate expression after eq. 11 : there is a bit more than the Stirling’s approximation involved here, since one also needs the asymptotic limit d k.

Yes, there is more involved but we thought it might be a distraction to lay out the details considering we were summaryizing textbook results. In any case, these passages have been moved to our Appendix.

4) p. 581, l. 19 : "And noting that" : add "in this case" (to be specific).

Done.

5) p. 582, l. 11 : This "f" introduced here is not the same as the spectral density function introduced eq. (5). Consider having distinct notations for the two quantities.

We have changed the notation for the spectral density function.

6) p. 582, l. 15 : You probably meant "there is no ":

Yes, we corrected this. Thank you for spotting this.

7) p. 583, l. 21 : Is this common practice to denote the statistical software "R" using the mathbb font R? I have never seen this before.

Corrected. Thank you for spotting this.

8) p. 584, l. 1 : Make it clear that the likelihood is conditional on "xA".

Done.

9) p. 584, l. 23 : The authors may want to further justify their prior choice for by observing that the asymptotic limit is equivalent to a log-uniform prior.

Done, thanks.

10) p. 585, ll. 19-20 : The variables and may be mistaken for the process mean and standard deviation. I would propose to introduce straight away the particular case N(d; 2d) to avoid unnecessary confusion.

Done.

11) p. 585 eq 16 and following equations : the function - introduced here seems to stand for exponential of the minus squared, and the erf function. These symbols have thus not the same meaning as in equation (4). Please clarify and change notation if needed. The use of - as in eq. (4) is reestablished on page 588, further strengthening the possible confusion.

Thanks for spotting this. We have replaced Phi and phi, the usual standard normal CDF and PDF, by versions which have a N superscript in caligraphic font so they are less likeliy to be confused with the parameters to the ARMA model.

12) p. 586, l. 14 : "P = n is sensible". Please explain.

Sensible was a poor choice of words. We have replaced this with a „for example“ clause. It is important to choose P large enough to retain low-frequency effects, however having P be too big (e.g., bigger than n) complicates inference because then there are more latent variable than actual data. Choosing P = n is a common middle-ground in the literature.

13) p. 589 : words 'trivial' and 'clearly' may be felt as slightly annoying when trying to go into the details of a notation that is not always clear and trivial.

We deleted the words trivial and clearly where appropriate.

14) p. 589, l. 26 : again clarify the meaning of here.

Done, thanks.

15) p. 589, l. 28 : 'since the normalization terms would cancel' : is that so obvious ? Normalisation terms to do not cancel on eq. 17 and I must confess that it is not clear to me why they cancel here.

The normalization constant is a function of d. In Eq. 17, which is an acceptance ratio for d, they don't cancel because the normalization constants in the numerator and denominator are for different d. However the discussion here is for the short memory parameter, and so d is fixed. Therefore the constants in the numerator and denominator are the same.

16) p. 589, last sentence (wrapping on p. 590) : clumsy grammar

We have improved the wording here.

17) l. 592 l. 3 : "In other words" : withdraw

Done.

18) l. 594 : More explicit details need be given about how and - (bold and) are determined. This is with this kind of detail in mind that one can see that supporting code will be welcomed by readers wishing to reproduce the algorithms proposed here, and use it for other applications.

It may not be attractive, but by „pilot tuning scheme“ we literally mean: try a value (e.g., sigma=1) and adjust based on observed acceptance rates, autocorrelations, effective sample sizes, from the resulting Markov chains. This is obviously more of an art than a science, but is standard fare in general purpose MCMC libraries for Bayesian inference, like BUGS. We are happy to provide our code, and in fact have provided it on several occations in the past when we have gotten requests. We would like to turn it into a R package but Tim Graves who developed the code now works outside of academia, so this might take some time.

19) l. 597 : "Those of Beran" clarify or add exact reference (papers by Beran are cited a couple of times, but one needs to be specific and informative here)

We have added the specific references here.

20) l. 599 : Provide the true value of dl (sorry, if it is there I couldn’t see it).

Thanks, we have added the true values to the figure caption.

21) Figure 8 and 9: indicate true values of parameters when known (e.g.: d and on Figure 9)

Done. Of course, true values for the Nile are not known (Figure 9).

Manuscript prepared for Nonlin. Processes Geophys.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 5 September 2015

Efficient Bayesian inference for natural time series exhibiting both Short- and Long-Memory using a parametric ARFIMA ~~processes~~process

Timothy Graves[1], Robert B. Gramacy[2], Christian L. E. Franzke[3], and Nicholas W. Watkins[4,5,6,7]

[1]URS Corporation, London, UK
[2]The University of Chicago, Booth School of Business, USA
[3]Meteorological Institute and Center for Earth System Research and Sustainability (CEN), University of Hamburg, Germany
[4]Centre for the Analysis of Time Series, London School of Economics and Political Science, London, UK
[5]Centre for Fusion Space and Astrophysics, University of Warwick, Coventry, UK
[6]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[7]Centre for Complexity and Design, Open University, Milton Keynes, UK

Correspondence to: Christian Franzke (christian.franzke@uni-hamburg.de)

Abstract. Many geophysical quantities, like atmospheric temperature, water levels in rivers, and wind speeds, have shown evidence of ~~long-range dependence (LRD). LRD means~~ long memory (LM). LM implies that these quantities experience non-trivial temporal memory, which potentially enhances their predictability, but also hampers the detection of externally forced trends. Thus, it is important to reliably identify whether or not a system exhibits ~~LRD~~LM. In this paper we present a modern and systematic approach to the inference of ~~LRD. Rather than Mandelbrot's fractional Gaussian noise, we use the more~~ LM. We use the flexible Autoregressive Fractional Integrated Moving Average (ARFIMA) model which is widely used in time series analysis, and of increasing interest in climate science. Unlike most previous work on the inference of ~~LRD~~LM, which is frequentist in nature, we provide a systematic treatment of Bayesian inference. 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. We illustrate our new methodology on the Nile water level data and the Central England Temperature (CET) time series, with favorable comparison to the standard estimators. For CET we also extend our method to seasonal long memory.

1 Introduction

Many natural processes are sufficiently complex that a stochastic model is essential, or at the very least an efficient description (Watkins, 2013). Such a process will be specified by several

properties, of which a particularly important one is the degree of memory in a time series, often expressed through a characteristic autocorrelation time over which fluctuations will decay in magnitude. In this paper, however, we are concerned with specific types of stochastic processes that are capable of possessing ~~'long memory ', or "long-range dependence" (LRD~~long memory (LM) (Beran, 1994a; 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 (Beran, 1994a; Palma, 2007; Beran et al., 2013) is that a (finite variance, stationary) process has *long memory* if its autocorrelation function (ACF) has 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 be anti-persistent.

The asymptotic power law form of the ACF corresponds to an absence of a characteristic decay timescale, in striking contrast to many standard (stationary) stochastic processes where the effect of each data point decays so fast that it rapidly becomes indistinguishable from noise. An example of the latter is the exponential ACF where the e-folding time scale sets a characteristic correlation time. The study of processes that *do* possess long memory is important because they exhibit unusual properties, because many familiar mathematical results fail to hold, and because of the numerous examples of data sets where ~~LRD~~ LM is seen.

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 the concept of long memory (Mandelbrot and Van Ness, 1968; Mandelbrot and Wallis, 1968).

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 marginalize 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~~ Autoregressive Fractional Integrated Moving Average (ARFIMA) class of processes (Granger and Joyeux, 1980; Hosking,

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

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~~.~~, marginally (i.e., averaging over short memory and seasonal effects). ARFIMA has become very popular in statistics and econometrics because it is generalizable 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 statistics 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~~Here we present a Bayesian framework for the efficient and systematic estimation of the ARFIMA parameters. 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 modernization 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, in which the model order (the p and q parameters in the ARFIMA model) varies and, thus, so does the dimension of the problem. This aspect is similar to the 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~~ LM 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 (to be published in a companion paper). The remainder of the paper is organized as follows. Section ~~2.1 summarizes the ARFIMA model required for our purposes. Section~~ 2 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 3 describes our proposed Bayesian framework and methodology in detail, focusing on long-memory only. Then, in Section 4, we consider extensions for additional short memory and the computational techniques required to integrate

them out. Empirical illustration and comparison of all methods is provided in Section 5 ~~. The~~ via synthetic and real data including the Nile water level data and the Central England Temperature (CET) time series, with favorable comparison to the standard estimators. In the case of the Nile data we find strong evidence for long memory. The CET analysis requires a slight extension to handle *seasonal* long memory, and we find that the situation here is more nuanced in terms of evidence for long memory. The paper concludes with a discussion in Section 7 focused on the potential for further extension.

2 ~~Time series definitions and the ARFIMA model~~Likelihood evaluation for Bayesian inference

2.1 ARFIMA model

~~Because ARFIMA models have not yet been very widely used in the geosciences we~~ We provide here a brief review of ~~them. Readers familiar with ARFIMA models can skip this section.~~ the ARFIMA model. More details are given in appendix A.

~~We define 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 (normalized) autocorrelation function ACF$\rho(\cdot)$ is defined as: $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$. Another useful time domain tool is the 'backshift '~~ An ARFIMA model is given by:

$$\Phi(\mathcal{B})(1-\mathcal{B})^d X_t = \Theta(\mathcal{B})\varepsilon_t. \tag{1}$$

We define 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', $\{\varepsilon_t\}$:~~

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

~~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 $\{\varepsilon_t\}$~~$\mathcal{B}^k X_t = \mathcal{B}^{k-1}X_{t-1} = \cdots = X_{t-k}$. Φ is the autoregressive component and Θ is the moving average component and constitute the short-memory components of the ARFIMA model.

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

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

processes are invertible, stationary and causal if and only if $\Phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. $\{X_t\}$ is said to be a moving average process of order q, , if

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

for all t.[2] 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) processprocess of orders p and q, , if for all t:

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

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 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 $-\pi \leq \lambda \leq \pi$.[2] Finally, the SDF of an ARMA process is

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

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 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).

[2] Many authors define $\Phi(z) = 1 - \sum \phi_k z^k$. Our version emphasises connections between Φ and (A2–A3).

[2] 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: $0 \leq \lambda \leq \pi$.

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 $0 < d < \frac{1}{2}$.

The simplest way of creatinga 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

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

is called fractionally integratedwith memory parameter d, (Barnes and Allan, 1966; Adenstedt, 1974) . The full trichotomy of negative, short, and long memory is determined solely by d.

In practice this modelis of limited appeal to time series analysts because the entire memory structure is determined by just one parameter, d. One often therefore generalizes it 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 yields so-called auto-regressive fractionally integrated moving average process with parameter d, and orders p and q (), having SDF:

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

Choosing $p = q = 0$ recovers $\equiv$.

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 generalized form of the binomial expansion (Brockwell and Davis, 1991, Eq. 13.2.2) :

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

From this observation, one can show that $X_t = (1 - \mathcal{B})^{-d} Z_t$, where $\{Z_t\}$ is an ARMA process, has SDF . 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 and more general processes in the time domain. A process $\{X_t\}$ is an process if for all t: $(1 - \mathcal{B})^d X_t = \varepsilon_t$. Likewise, a process

~~$\{X_t\}$ is an 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 and respectively.~~

~~Finally, to connect back to our first definition of long memory, consider the ACV of the process. By using the definition of spectral density to directly integrate , and an alternative expression for $\pi_k^{(d)}$~~ These are defined in more detail in appendix A and in

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

~~one can obtain the following representation of the ACV of the process:~~

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

~~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:~~

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

~~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 can be used to represent as an 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 analog: $X_t = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+1)}\frac{\Gamma(k+d)}{\Gamma(d)}\varepsilon_{t-k}$.~~ Graves (2013) .

3 ~~Likelihood evaluation for Bayesian inference~~

~~For nowwe~~

2.1 Likelihood function

For now, restrict our attention to ~~(a Bayesian)~~ a Bayesian analysis of an ARFIMA$(0,d,0)$ process, having no short-ranged ARMA components ($p=q=0$), placing emphasis squarely on the memory parameter d. As we explain in our Appendix the resulting process is identical to a *fractionally integrated* processes with memory parameter d.

Here we develop an efficient and new scheme for evaluating the (log) likelihood, via approximation. ~~This scheme is very flexible in the sense that it seamlessly allows to use different noise distributions (like a t-distribution instead of a Gaussian; this will be reported elsewhere).~~ Throughout, suppose that we have observed the vector $\mathbf{x} = (x_1,\ldots,x_n)^\top$ as a realization 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'~~ location and scale parameters respectively. The m–dimensional $\boldsymbol{\lambda}$ is a ~~'shape'~~ 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 no $\boldsymbol{\lambda}$. 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$(0,d,0)$ 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:

$$X_t = \varepsilon_t + \Pi_P \mu - \sum_{k=1}^{P} \pi_k X_{t-k}. \tag{2}$$

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'~~ auxiliary, hence 'A'). Consider the likelihood $L(\mathbf{x}|\boldsymbol{\psi})$ as a joint density which can be factorized 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 (2) 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:

$$\begin{aligned} h_t(x_t|\mathbf{x}_A,\mathbf{x}_{t-1},\boldsymbol{\psi}) &\approx f\left(x_t;\left[\Pi_P\mu - \sum_{k=1}^{P}\pi_k x_{t-k}\right],\sigma,\boldsymbol{\lambda}\right) \\ &\equiv \frac{1}{\sigma} f\left(\frac{c_t - \Pi_P\mu}{\sigma};0,1,\boldsymbol{\lambda}\right), \quad \text{where} \quad c_t = \sum_{k=0}^{P}\pi_k x_{t-k}, \qquad t = 1,\ldots,n. \end{aligned} \tag{3}$$

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:

$$\ell(\mathbf{x}|\boldsymbol{\psi},\mathbf{x}_A) \approx -n\log\sigma + \sum_{t=1}^{n}\log\left\{f\left(\frac{c_t-\Pi_P\mu}{\sigma};\boldsymbol{\lambda}\right)\right\}. \tag{4}$$

Evaluating this expression efficiently depends upon efficient calculation of $\boldsymbol{c} = (c_1, \ldots, c_n)^t$ and $\log f(\cdot)$. From (3), $\boldsymbol{c}$ is a convolution of the augmented data, $(\mathbf{x}, \mathbf{x}_A)$, and coefficients depending on d, which can be evaluated quickly in ~~R~~ the R language for statistical computing via `convolve` via FFT. Consequently, evaluation of the *conditional* on $\mathbf{x}_A$ likelihood in the Gaussian case costs only $\mathcal{O}(n \log n)$—a clear improvement over the ~~'exact'~~ exact method. Obtaining the *un*conditional likelihood requires marginalization over $\mathbf{x}_A$, which is analytically infeasible. However this conditional form will suffice in the context of our Bayesian inferential scheme, presented below.

3 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 modernization of the blocked MCMC method of Pai and Ravishanker (1998). Isolating parameters by blocking provides significant scope for modularization which helps to 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 $\alpha_0, \beta_0 \to 0$: $p_\sigma(\sigma) \propto \sigma^{-1}$ which, in the asymptotic limit, is equivalent to a log uniform prior. 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

$$A_\mu(\mu, \xi_\mu) = \sum_{t=1}^{n} \log\left\{ f\left(\frac{c_t - \Pi_P \xi_\mu}{\sigma}; \boldsymbol{\lambda} \right) \right\} - \sum_{t=1}^{n} \log\left\{ f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda} \right) \right\} + \log\left[\frac{p_\mu(\xi_\mu)}{p_\mu(\mu)} \right] \tag{5}$$

under the approximate likelihood.

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, we 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 (5) the MH

acceptance ratio under the approximate likelihood is

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

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, see Graves (2013)).

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 ~~$\xi_d \sim \mathcal{N}^{(a,b)}(\mu,\sigma^2)$~~ $\xi_d \sim \mathcal{N}^{(a,b)}(d,\sigma_d^2)$, with density

$$f(x;\mu,\sigma,a,b) = \frac{1}{\sigma} \frac{\phi[(x-\mu)/\sigma]}{\Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma]} \frac{\phi^{(\mathcal{N})}[(x-\mu)/\sigma]}{\Phi^{(\mathcal{N})}[(b-\mu)/\sigma] - \Phi^{(\mathcal{N})}[(a-\mu)/\sigma]}, \qquad a < x < b, \tag{6}$$

where $\Phi^{(\mathcal{N})}$ and $\phi^{(\mathcal{N})}$ are the standard normal CDF and PDF respectively. 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 (6) yields

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

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:

$$A_d = \sum_{t=1}^{n} \log\left\{ f\left(\frac{\xi_{c_t} - \xi_{\Pi_P}\mu}{\sigma}; \boldsymbol{\lambda} \right) \right\} - \sum_{t=1}^{n} \log\left\{ f\left(\frac{c_t - \Pi_P \mu}{\sigma}; \boldsymbol{\lambda} \right) \right\} + \log\left[\frac{p_d(\xi_d)}{p_d(d)} \right] + \log\left\{ \frac{\Phi[(\frac{1}{2}-d)/\sigma_d] - \Phi[(-\frac{1}{2}-d)/\sigma_d]}{\Phi[(\frac{1}{2}-\xi_d)/\sigma_d] - \Phi[(-\frac{1}{2}-\xi_d)/\sigma_d]} \frac{\Phi^{(\mathcal{N})}[(\frac{1}{2}-d)/\sigma_d] - \Phi^{(\mathcal{N})}[(-\frac{1}{2}-d)/\sigma_d]}{\Phi^{(\mathcal{N})}[(\frac{1}{2}-\xi_d)/\sigma_d] - \Phi^{(\mathcal{N})}[(-\frac{1}{2}-\xi_d)/\sigma_d]} \right\}. \tag{7}$$

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~~ e.g., $P = n$~~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, 1994b); for further discussion refer to Graves (2013, Appendix C).

For a full MH approach, we recommend an independence sampler to ~~'backward project'~~ backward project the observed time series. Specifically, first relabel the observed data: $y_{-i} = x_{i+1}$, $i = 0, \ldots n-1$. Then use the vector $(y_{-(n-1)}, \ldots, y_{-1}, y_0)^t$ to generate a new vector of length n, $(Y_1, \ldots, Y_n)^t$ where Y_t via (2): $Y_t = \varepsilon_t + \Pi_P \mu - \sum_{k=1}^{n} \pi_k Y_{t-k}$, 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: $\xi_{x_{-i}} = y_{i+1}$, $i = 0, \ldots, n-1$. 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 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.

4 Extensions to accommodate short memory

Simple ARFIMA$(0, d, 0)$ models 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 5.1, 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 modeling 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 modeling 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 marginalize 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* (i.e., changing p and/or q) and within (sampling $\boldsymbol{\phi}$ and ~~within~~ $\boldsymbol{\theta}$ given particular fixed p and q) models, and focus on the marginal posterior distribution of d obtained by (Monte Carlo) integration over all models and parameters therein. RJ methods, which mixes so-called *transdimensional*, between-model moves with the conventional within-model ones, 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~~ Fractional Exponential processes (FEXP). 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. The result is a Monte Carlo inferential scheme that alows short memory effects to be marginalized out when summarizing inferences for the main parameter of interest: d, for long memory.

4.1 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 2.1). 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'~~ 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 $\Sigma_{\boldsymbol{\omega}}$.

An ~~'exact'~~ 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 2 can be ported over directly. Given the coefficients $\{\pi_k^{(d)}\}$ and polynomials Φ and Θ, it is ~~trivial~~ straightforward 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 3 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 parametrize 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~~ Multivariate Normal (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 parameterized 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~~ are appropriate choices. Calculation of the MH acceptance ratio $A_{\boldsymbol{\omega}}(\boldsymbol{\omega}, \boldsymbol{\xi}_{\boldsymbol{\omega}})$ is trivial; it simply requires numerical evaluation of ~~$\Phi_r(\cdot;\cdot, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$~~$\Phi_r^{(\mathcal{N})}(\cdot;\cdot, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$, e.g., via `mvtnorm`, since the ratios of hypercuboid normalization terms would cancel. We find that initial values $\phi^{[0]}$ 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 '~~ 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 5.2].

4.2 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'~~ 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~~ model 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 '~~ allowable values for each component are a complicated function of the others. Nontrivial proposals are required.

A potential approach is to parametrize 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 reparametrisation 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:

$$\phi_i^{(k-1)} = \frac{\phi_i^{(k)} - \phi_k^{(k)}\phi_{k-i}^{(k)}}{1 - \left(\phi_k^{(k)}\right)^2}, \qquad k = p, \ldots, 2, \qquad i = 1, \ldots, k-1. \tag{8}$$

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

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

Note that $\phi_p^{(p)} = \varphi_p^{(p)}$. Moreover, if $p = 1$, the two parameterizations are the same, i.e. $\phi_1 = \varphi_1$ (consequently the brief study of ARFIMA$(1, d, 0)$ in section 4.1 fits in this framework). The equivalent parametrized form for $\boldsymbol{\theta}$ is $\boldsymbol{\vartheta}$. The full memory parameter $\boldsymbol{\omega}$ is parametrized 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 other words, if~~ If $d = q = 0$, using this parametrization 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 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 4.1. 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 $\boldsymbol{\varpi}$ parametrization (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 (larger) more complicated models which, in the interests of parsimony, might be undesired. ~~We~~ As a simple representative of potential priors that give greater weight to smaller models 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 '~~ body of G has four neighbors; each point on the ~~'line boundaries '~~ line boundaries has three; and each of the four ~~'corner points '~~ corner points has only two neighbors. 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~~decreasing/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: $\boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p+1)} = \left(\boldsymbol{\varphi}^{(p)}, u\right)$. The corresponding map for decreasing the dimension is

$\boldsymbol{\varphi}^{(p+1)} \to \boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p)}$ is $\boldsymbol{\xi}_{\boldsymbol{\varphi}}^{(p)} = \left(\varphi_1^{(p+1)}, \ldots, \varphi_p^{(p+1)}\right)$. ~~In other words, we~~ 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 4.1, with a pilot tuning scheme. That matrix is shown on the left below, where we've ~~'~~blocked it out ~~'~~

$$\Sigma^{(1,1)} = \left(\begin{array}{c|c:c} \sigma_d^2 & \sigma_{d,\varphi_1} & \sigma_{d,\vartheta_1} \\ \hline & \sigma_{\varphi_1}^2 & \sigma_{\varphi_1,\vartheta_1} \\ \hdashline & & \sigma_{\vartheta_1}^2 \end{array}\right), \quad \Sigma^{(p,q)} = \left(\begin{array}{c|c:c} \sigma_d^2 & \Sigma_{d,\boldsymbol{\varphi}^{(p)}} & \Sigma_{d,\boldsymbol{\vartheta}^{(q)}} \\ \hline & \Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\varphi}^{(p)}} & \Sigma_{\boldsymbol{\varphi}^{(p)},\boldsymbol{\vartheta}^{(q)}} \\ \hdashline & & \Sigma_{\boldsymbol{\vartheta}^{(q)},\boldsymbol{\vartheta}^{(q)}} \end{array}\right), \tag{9}$$

(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 (9), the following hold:

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

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

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

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 (see Graves (2013).

5 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 $P = n$.

5.1 Long memory

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 $\mathbf{x}$.

Efficacy of approximate likelihood method

Start with the ~~'null case'~~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.

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'~~ tight the estimates of d are around the input 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.

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, 1994a) 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)~~.

Effect of varying time series length

We now analyze 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 3.

Observe that all three marginal posterior standard deviations are proportional to $\frac{1}{\sqrt{n}}$, although the posterior of μ is less ~~'reliable'~~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.~~ Beran (1994a) .

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 pack-

age `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 accommodate the new comparators, respectively, and plot them against $\widehat{d_R}^{(B)}$ in figure 5.

Observe that all four methods have a much larger variance than our Bayesian 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 $d \in (-\frac{1}{2}, \frac{1}{2})$.

5.2 Extensions for short memory

Known form: We first consider the MCMC algorithm from section 4.1 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-parametrized method from section 4.2.

In general, this method works very well; two example outputs are presented in figure 6, 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 behavior 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 7(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 behavior 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 7(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 7(c). In theory an ellipsoidal credibility set could be constructed, although this is clearly less practical when $\dim \boldsymbol{\omega} > 2$.

Unknown form: The RJ scheme outlined in section 4.2 works well for data simulated with p and q up to 3. The marginal posteriors for d are generally roughly centered around d_I (the data generating

value) and the modal posterior model probability is usually the ~~'correct'~~ 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 8(a)–(b), together with the equivalent density estimated assuming unknown model orders.

Notice 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 8(c)–(d). The posterior model probabilities are presented in table 2, showing that the ~~'correct'~~ correct modes are being picked up consistently.

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

$$\left(1 - \frac{9}{16}\mathcal{B}^2\right)(1 - \mathcal{B})^{0.25} X_t = \left(1 + \frac{1}{3}\mathcal{B}\right)\varepsilon_t, \qquad \text{where } \varepsilon_t \sim \mathcal{S}_{\alpha=1.75,0}. \tag{10}$$

For more details, see (Graves, 2013, §7.1). The marginal posterior densities of d and α are presented in figure 9.

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'~~ wrong model.

~~The Nile Data:~~

6 Observational Data Analysis

We conclude with ~~an~~ the application of our ~~methods to the famous annual Nile minima data.~~ method to two long data sets: The Nile water level minima data and the Central England Temperature (CET). The Nile data is part of the R package 'longmemo' and the CET time series can be downloaded from http://www.metoffice.gov.uk/hadobs/hadcet/

6.1 The Nile Data:

Because of the fundamental importance of the Nile river to the civilizations 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 ($n = 663$).[2] ~~The~~ These data are plotted in figure 10.

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

We immediately observe the apparent low frequency component of the data. The data appear to be on the 'verge' of being stationary, however the general consensus amongst the statistical community is that the series *is* stationary. The posterior summary statistics ~~and~~ are presented in table 5, density estimates of the marginal posteriors of d and μ are presented in figure 11, and the posterior model probabilities are presented in table 6.

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

We see that the model with the highest posterior probability is the ARFIMA$(0, d, 0)$ model with $d \approx 0.4$. These facts suggest that the memory in the signal is strong, 'pure', long memory. It is interesting to compare these findings with other literature. Liseo et al. (2001) used a semi-parametric Bayesian method on the first 512 observations of the sequence and obtained an estimate for d of 0.278. Hsu and Breidt (2003) used a similar method to Pai and Ravishanker (1998) to estimate d (within an ARFIMA$(0, d, 0)$ model) at 0.416 with approximate credibility interval of $(0.315, 0.463)$. Ko and Vannucci (2006a) similarly found using wavelets $\widehat{d^B} = 0.379$ with credibility interval $(0.327, 0.427)$. Palma (2007) obtained $\widehat{d^B} = 0.420$. Holan et al. (2009) obtained $\widehat{d^B} = 0.387$ with credibility interval $(0.316, 0.475)$ using their Bayesian FEXP method.

We note that the interpretation as persistence of the $d \approx 0.4$ ($H \approx 0.9$) value that we and others have obtained has been challenged by (Kärner, 2001) . In his view the analysis should be applied to the increments of the level heights rather than the level heights themselves, giving an antipersistent time series with a negative d value. The need for a short range dependent component that he argues for is, however, automatically included in the use of an ARFIMA model. Although ARFIMA was originally introduced in econometrics as a phenomenological model of LM, very recent progress is being made in statistics and physics on building a bridge between it and continuous time linear dynamical systems (see e.g. Slezak and Weron (2015)).

In conclusion, our findings agree with all published Bayesian long memory results (except Liseo et al. 's anomalous finding). Moreover, these findings agree with numerous classical methods of analysis (e.g. Beran, 1994a) that have found the best model fit is an ARFIMA$(0, d, 0)$ model with $d \approx 0.4$. We note that it is a result of our data analysis method that short-memory can be neglected, rather than being an a priori assumption.

6.2 Central England temperature

There is increasing evidence that surface air temperatures posseses long-memory (Gil-Alana, 2003, 2008; Bunde et al., 2014; Franzk
long time series are needed to get robust results. The central England temperature (CET) index is a famous measure of the *monthly* mean temperature in an area of southern-central England dating

back to 1659 (Manley, 1974) . Given to a precision of 0.5°C prior to 1699 and 0.1°C thereafter, the index is considered to be the longest reliable known temperature record from station data. As expected, the CET exhibits a significant seasonal signal, at least some of which must be considered as deterministic. Following the approach of Montanari et al. (2000) , the index is first deseasonalised using the additive 'STL' method (Cleveland et al., 1990) . This deseasonalised CET index is shown in figure 13.

The estimated seasonal function $S(t)$ that was removed is shown in figure 14(a). The spectrum of the deseasonalised process is shown in figure 14(b). D denotes the seasonal long memory parameter. Notice that, in addition to the obvious spectral peak at the origin, there still remains a noticeable peak at the monthly frequency $\omega = \frac{\pi}{6}$. However there are no further peaks in the spectrum which would appear to rule out a SARFIMA model. These observations therefore suggest that a simple 2-frequency Gegenbauer$(d, D; \frac{\pi}{6})_2$ process might be an appropriate model. See appendix B for more details about seasonal long memory.

Applying this model, the marginal posterior statistics are presented in table 7 and the joint posterior samples of (d, D) from this model are plotted in figure 15. These clearly indicate that both d and D are non-zero (albeit small in the case of D) suggesting the presence of long memory in both the conventional and seasonal sense.

In order to compare these results with other publications', it is important to note that to remove annual seasonality from the CET, the series of annual means is often used instead of the monthly series. This of course reduces the fidelity of the analysis. Hosking (1984) found (using rather crude estimation procedures) that the best fitting model for the annual means of the CET was the ARFIMA$(1, 0.33, 0)$ model with $\phi = 0.16$. Pai and Ravishanker (1998) used the same series as test data for their Bayesian method; they fitted each of the ARFIMA models with $p, q \leq 1$ and found that all models were suitable. Their estimates of d ranged from 0.24 for $p = q = 0$ to 0.34 for $p = 0$, $q = 1$.

Of course all these studies assume the time series is stationary, in particular that it has a constant mean. The validity of this assumption was considered by Gil-Alana (2003) who used formal hypothesis testing to consider models:

$$Y_t = \beta_0 + \beta_1 t + X_t, \tag{11}$$

where $\{X_t\}$ is an ARFIMA$(0, d, 0)$ process. For values of $d = 0,\ 0.05,\ 0.10,\ 0.15$, β_1 was found to be significantly non-zero (at about 0.23°C per century) but for $d \geq 0.20$, statistical significance was not found. Gil-Alana (2008) later extended this work by replacing the ARFIMA$(0, d, 0)$ process in (11) with a Gegenbauer$(d; \omega)$ process to obtain similar results. However, choice of ω was rather *ad hoc* likely influencing the results.

In order to consider the stationarity of the time series, we divided the series up into four blocks of length 1024 months (chosen to maximise efficiency of the fast Fourier transform) and analysed each block independently. The posterior statistics for each block are presented in table 8 with some results presented graphically in figure 16.

It is interesting to note that the degree of (conventional) long memory is roughly constant over the last three blocks but appears to be larger in the first block. Of particular concern is that there is no value of d that is included in all four 95% credibility intervals; this would suggest non-stationarity. Although this phenomenon may indeed have a physical explanation, it is more likely caused by the inhomogeneity of the time series. Recall that the first fifty years of the index are only given to an accuracy of 0.5°C compared to 0.1°C afterwards; this lack of resolution clearly has the potential to bias in favour of strong auto-correlation when compared with later periods.

Interestingly, the seasonal long memory parameter D has 95% credibility intervals that include zero for the both the second and third blocks. Finally, note that the 95% credibility intervals for μ all include the range $(9.314, 9.517)$, in other words it is entirely credible that the mean is non-varying over the time period.

7 Conclusions

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~~ our method can handle the sorts of time series data ~~that~~ with possible long memory that we 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 generalization 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~~Graves (2013) .

Acknowledgements. We thank one anonymous reviewer and M. Crucifix for their comments which helped to improve this manuscript. CF is supported by the German Research Foundation (DFG) through the cluster of excellence CliSAP (EXC177).

Appendix A: ARFIMA model

We define 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 and $\mathrm{Cov} = \mathrm{E}[(X - (E)[X])(X - (E)[X])]$ is the lag-covariance matrix. The (normalized) autocorrelation function *ACF* $\rho(\cdot)$ is defined as: $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$. 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, $\{\varepsilon_t\}$:

$$X_t = \Psi(\mathcal{B})\varepsilon_t, \quad \text{where } \Psi(z) = \sum_{k=0}^{\infty} \psi_k z^k. \tag{A1}$$

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 $\{\varepsilon_t\}$.

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

$$\Phi(\mathcal{B})X_t = \varepsilon_t, \quad \text{where} \quad \Phi(z) = 1 + \sum_{k=1}^{p} \phi_k z^k, \quad \text{and} \quad (\phi_1, \ldots, \phi_p) \in \mathbb{R}^p. \tag{A2}$$

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

$$X_t = \Theta(\mathcal{B})\varepsilon_t, \quad \text{where} \quad \Theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k, \quad \text{and} \quad (\theta_1, \ldots, \theta_p) \in \mathbb{R}^q, \tag{A3}$$

for all t.[3] 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:

$$\Phi(\mathcal{B})X_t = \Theta(\mathcal{B})\varepsilon_t. \tag{A4}$$

[3] Many authors define $\Phi(z) = 1 - \sum \phi_k z^k$. Our version emphasises connections between Φ and (A2–A3).

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 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) $f_{\text{sd}}(\cdot)$ defined such that: $\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f_{\text{sd}}(\lambda)\, d\lambda$, $\forall k \in \mathbb{Z}$. This can be inverted to obtain an explicit expression for the SDF (e.g. Brockwell and Davis, 1991, §4.3): $f_{\text{sd}}(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda}$, where $-\pi \leq \lambda \leq \pi$.[4] Finally, the SDF of an ARMA process is

$$f_{\text{sd}}(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}, \qquad 0 \leq \lambda \leq \pi. \tag{A5}$$

For an ARFIMA process (1) 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 LM, one of which is particularly useful, as it considers the frequency domain: A stationary process has long memory when its SDF follows $f_{\text{sd}}(\lambda) \sim c_f \lambda^{-2d}$, as $\lambda \to 0^+$ for some positive constant c_f, and where $0 < d < \frac{1}{2}$.

The simplest way of *creating* a process which exhibits long memory is through the SDF. Consider $f_{\text{sd}}(\lambda) = |1 - e^{i\lambda}|^{-2d}$, where $0 < |d| < \frac{1}{2}$. By simple algebraic manipulation, this is equivalently $f_{\text{sd}}(\lambda) = \left(2 \sin \frac{\lambda}{2}\right)^{-2d}$, 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

$$f_{\text{sd}}(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d}, \qquad 0 < \lambda \leq \pi. \tag{A6}$$

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.

In practice this model is of limited appeal to time series analysts because the entire memory structure is determined by just one parameter, d. One often therefore generalizes it by taking any short memory SDF $f^*_{\text{sd}}(\cdot)$, and defining a new SDF: $f_{\text{sd}}(\lambda) = f^*_{\text{sd}}(\lambda) \left|1 - e^{i\lambda}\right|^{-2d}$, $0 \leq \lambda \leq \pi$. An obvious class of short memory processes to use this way is ARMA. Taking f^* from (A5) yields so-called auto-regressive fractionally integrated moving average process with parameter d, and

[4] 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: $0 \leq \lambda \leq \pi$.

orders p and q (ARFIMA(p,d,q)), having SDF:

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2} |1 - e^{i\lambda}|^{-2d}, \qquad 0 \leq \lambda \leq \pi. \tag{A7}$$

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

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 generalized form of the binomial expansion (Brockwell and Davis, 1991, Eq. 13.2.2) :

$$(1-\mathcal{B})^d =: \sum_{k=0}^{\infty} \pi_k^{(d)} \mathcal{B}^k, \qquad \text{where} \qquad \pi_k^{(d)} = (-1)^k \frac{1}{\Gamma(k+1)} \frac{\Gamma(d+1)}{\Gamma(d-k+1)}. \tag{A8}$$

From this observation, one can show that $X_t = (1-\mathcal{B})^{-d} Z_t$, where $\{Z_t\}$ is an ARMA process, has SDF (A7). 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 (A2) and (A3) 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 (A6), and an alternative expression for $\pi_k^{(d)}$ in (A8)

$$\pi_k^{(d)} = \frac{1}{\Gamma(k+1)} \frac{\Gamma(k-d)}{\Gamma(-d)}, \tag{A9}$$

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

$$\gamma_d(k;\sigma) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}. \tag{A10}$$

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:

$$\rho_d(k) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}, \tag{A11}$$

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 (A9) 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 in this case $\psi_k^{(d)} = \pi_k^{(-d)}$, leads to the following MA(∞) analog: $X_t = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+1)} \frac{\Gamma(k+d)}{\Gamma(d)} \varepsilon_{t-k}$.

Appendix B: Seasonal long memory models

We define a seasonal differencing operator $(1-\mathcal{B}^s)$, as a natural extension to a Seasonal ARFIMA (SARFIMA) processes by combining seasonal and non-seasonal fractional differencing operators (Porter-Hudak, 1990) :

$$(1-\mathcal{B})^d (1-\mathcal{B}^s)^D X_t = \varepsilon_t.$$

The generalisation to include both seasonal and non-seasonal short memory components is obvious (Porter-Hudak, 1990) :

$$\Phi^{(p)}(\mathcal{B}) \Phi_s^{(P)}(\mathcal{B}^s) (1-\mathcal{B})^d (1-\mathcal{B}^s)^D X_t = \Theta^{(q)}(\mathcal{B}) \Theta_s^{(Q)}(\mathcal{B}^s) \varepsilon_t,$$

Focusing on the first of these issues, Hosking (1981) considered generalising the ARFIMA$(0,d,0)$ process in a different manner by retaining only one pole but at any given frequency in $[0,\pi]$. The model he suggested was later studied and popularised by Anděl (1986) and Gray et al. (1989, 1994) , and became known as the 'Gegenbauer process'.

A process $\{X_t\}$ is a Gegenbauer$(d;\omega)$ process if for all t:

$$(1-2u\mathcal{B}+\mathcal{B}^2)^d X_t = \varepsilon_t, \tag{B1}$$

where $\omega = \cos^{-1} u$ is called the Gegenbauer frequency. The obvious extension to include short memory components $\Phi^{(p)}$ and $\Theta^{(p)}$ is denoted GARMA$(p,d,q;\omega)$.

The term 'Gegenbauer' derives from the close relationship to the Gegenbauer polynomials, a set of orthogonal polynomials useful in applied mathematics. The Gegenbauer polynomials are most usefully defined in terms of their generating function. The Gegenbauer polynomial of order k with parameter d, $G_k^{(d)}(\cdot)$ satisfies:

$$(1-2uz+z^2)^{-d} \equiv \sum_{k=0}^{\infty} G_k^{(d)}(u) z^k. \tag{B2}$$

The spectral density function of the Gegenbauer$(d;\omega)$ process is (Gray et al., 1989) :

$$f(\lambda) = \frac{\sigma^2}{2\pi} |2(\cos\lambda - \cos\omega)|^{-2d}, \qquad 0 \leq \lambda \leq \pi.$$

Note that Gegenbauer$(d;\omega)$ processes possess a pole at the Gegenbauer frequency ω. Gegenbauer processes may be considered to be somewhat ambiguous in terms of long memory. Non-trivial (i.e. $\omega \neq 0$) Gegenbauer processes have bounded spectral density functions at the origin, and therefore do not have long memory according to our strict definition. Consequently a more general Gegenbauer process was developed:

Let $\mathbf{d} = (d_1,\ldots,d_k)$ and $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_k)$, and for all j, $u_j = \cos\omega_j$ (assumed distinct). Then a process $\{X_t\}$ is a k-factor Gegenbauer$(\mathbf{d};\boldsymbol{\omega})$ process if for all t Woodward et al., 1998 :

$$\prod_{j=1}^{k} (1 - 2u_j\mathcal{B} + \mathcal{B}^2)^{d_j} X_t = \varepsilon_t. \tag{B3}$$

The spectral density function of the k-factor Gegenbauer$(\mathbf{d};\boldsymbol{\omega})$ process is Woodward et al., 1998 :

$$f(\lambda) = \frac{\sigma^2}{2\pi} \prod_{j=1}^{k} |2(\cos\lambda - \cos\omega_j)|^{-2d_j}, \qquad 0 \leq \lambda \leq \pi.$$

K-factor Gegenbauer models are very flexible, and include nearly all other seasonal variants of ARFIMA processes such as the flexible-seasonal ARFIMA (Hassler, 1994) and fractional ARUMA (Robinson, 1994; Giraitis and Leipus, 1995) processes. Importantly, they also includes SARFIMA processes Reisen et al., 2006 :

A SARFIMA$(0,d,0)\times(0,D,0)_s$ process is equivalent to a $\lfloor \frac{s+2}{2} \rfloor$-factor Gegenbauer$(\mathbf{d};\boldsymbol{\omega})$ process where:

$$\omega_j = \frac{2\pi(j-1)}{s}, \qquad j = 1,\ldots,k,$$

and $d_1 = \frac{d+D}{2}$, $d_j = D$ for $j = 2,\ldots,k$, unless s is even in which case $d_k = \frac{D}{2}$.

Although k-factor Gegenbauer models are very general, one particular sub-model is potentially very appealing. This is the 2-factor model, with one pole at the origin and one at a non-zero frequency. In order to conform with notation for ARFIMA$(0,d,0)$ processes, we will slightly re-define this model: A process $\{X_t\}$ is a *simple 2-frequency* Gegenbauer process with parameters d, D, and ω, denoted Gegenbauer$(d,D;\omega)_2$ if for all t:

$$(1 - (2\cos\omega)\mathcal{B} + \mathcal{B}^2)^D (1-\mathcal{B})^d X_t = \varepsilon_t.$$

The Bayesian MCMC methodology developed here is easily extended to incorporate these seasonal fractional models. It is assumed that the frequency ω, or seasonal period s, is *a priori* known.

References

Abry, P., Flandrin, P., Taqqu, M. S., and Veitch, D.: Self-similarity and long-range dependence through the wavelet lens., in: Theory and Applications of Long-Range Dependence, pp. 527–556, 2003.

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

Anděl, J.: Long Memory Time Series Models, Kybernetika, 22, 105–123, 1986.

Barnes, J. A. and Allan, D. W.: A statistical model of flicker noise, Proceedings of the IEEE, 54, 176–178, 1966.

Beran, J.: Statistics for Long Memory Processes, Chapman & Hall, New York, 1994a.

Beran, J.: On a Class of M-Estimators for Gaussian Long-Memory Models, Biometrika, 81, 755–766, 1994b.

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

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

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

Bunde, A., Ludescher, J., Franzke, C., and Schellnhuber, H.-J.: How significant is West Antarctic warming?, Nature Geoscience, 7, 246–247, 2014.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A Seasonal-Trend Decomposition Procedure Based on Loess, J. of Official Statistics, 6, 3–73, with discussion, 1990.

Ehlers, R. S. and Brooks, S. P.: Bayesian analysis of order uncertainty in ARIMA models, Tech. rep., Department of Statistics, Federal University of Paraná, http://www.est.ufpr.br/rt/ehlb06.pdf, 2006.

Ehlers, R. S. and Brooks, S. P.: Adaptive proposal construction for reversible jump MCMC, Scandanavian J. of Statistics, 35, 677–690, 2008.

Eğrioğlu, E. and Günay, S.: Bayesian model selection in ARFIMA models, Expert Systems with Applications, 2010.

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

Franzke, C.: Long-range Dependence and Climate Noise Characteristics of Antarctic Temperature Data, J. Climate, 23, 6074–6081, 2010.

Franzke, C.: Nonlinear trends, long-range dependence and climate noise properties of surface air temperature, J. Climate, 25, 4172–4183, 2012.

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

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T.: `mvtnorm`: Multivariate Normal and t Distributions, http://CRAN.R-project.org/package=mvtnorm, R package version 0.9-9992, 2012.

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

Gil-Alana, L. A.: An Application of Fractional Integration to a Long Temperature Series, International J. of Climatology, 23, 1699–1710, 2003.

Gil-Alana, L. A.: Cyclical long-range dependence and the warming effect in a long temperature time series, International J. of Climatology, 28, 1435–1443, 2008.

Giraitis, L. and Leipus, R.: A generalized fractionally differencing approach in long-memory modeling, Lithuanian Mathematical Journal, 35, 53–65, 1995.

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

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

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

Gray, H. L., Zhang, N.-F., and Woodward, W. A.: On generalized fractional processes, J. of Time Series Analysis, 10, 233–257, 1989.

Gray, H. L., Zhang, N.-F., and Woodward, W. A.: Correction to: Gray et al. (1989), J. of Time Series Analysis, 15, 561–562, 1994.

Green, P. J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination., Biometrika, 82, 711–732, 1995.

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

Hassler, U.: (Mis)specification of long memory in seasonal time, J. of Time Series Analysis, 15, 19–30, 1994.

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

Hosking, J. R. M.: Fractional differencing, Biometrika, 68, 165–176, 1981.

Hosking, J. R. M.: Modeling persistence in hydrological time series using fractional differencing, Water Resources Research, 20, 1898–1908, 1984.

Hsu, N.-J. and Breidt, F. J.: Bayesian Analysis of Fractionally Integrated ARMA with additive noise., J. of Forecasting, 22, 491–514, 2003.

Hurst, H. E.: Long-term storage capacity of reservoirs, Transactions of the American Society of Civil Engineers, 116, 770–808, with discussion, 1951.

Kärner, O.: Comment on Hurst Exponent, Geophysical Research Letters, 28, 3825–3826, doi:10.1029/2001GL013104, http://dx.doi.org/10.1029/2001GL013104, 2001.

Kay, S. M. and Marple, S. L.: Spectrum analysis-A modern perspective., Proceedings of the IEEE., 69, 1380–1419, 1981.

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

Ko, K. and Vannucci, M.: Bayesian Wavelet-Based Methods for the Detection of Multiple Changes of the Long Memory Parameter, IEEE Transactions on Signal Processing, 54, 4461–4470, 2006b.

Koop, G., Ley, E., Osiewalski, J., and Steel, M. F.: Bayesian analysis of long memory and persistence using ARFIMA models., J. of Econometrics, 76, 149–169, 1997.

Liseo, B., Marinucci, D., and Petrella, L.: Bayesian semiparametric inference on long-range dependence., Biometrika, 88, 1089–1104, 2001.

Mandelbrot, B. B. and Van Ness, J. W.: Fractional Brownian Motions, Fractional Noises and Applications, SIAM Review, 10, 422–437, 1968.

Mandelbrot, B. B. and Wallis, J. R.: Noah, Joseph and operational hydrology, Water Resources Research, 4, 909–918, 1968.

Manley, G.: Central England temperatures: monthly means 1659 to 1973, Quarterly J. of the Royal Meteorological Society, 100, 389–405, 1974.

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

McLeod, A. I., Yu, H., and Krougly, Z.: Algorithms for Linear Time Series Analysis: With R Package, J. of Statistical Software, 23, http://www.jstatsoft.org/v23/i05/, 2007.

Monahan, J. F.: A note on enforcing stationarity in autoregressive-moving average models., Biometrika, 71, 403–404, 1984.

Montanari, A., Rosso, R., and Taqqu, M. S.: A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan, Water Resources Research, 36, 1249–1259, 2000.

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

Palma, W.: Long Memory Time Series, Wiley, 2007.

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

Porter-Hudak, S.: An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates, J. of the American Statistical Association, 85, 338–344, 1990.

Reisen, V. A., Rodrigues, A. L., and Palma, W.: Estimation of seasonal fractionally integrated processes, Computational Statistics and Data Analysis, 50, 568–582, 2006.

Robert, C. P.: Simulation of truncated normal variables, Statistics and Computing, 5, 121–125, 1995.

Roberts, G. O. and Rosenthal, J. S.: Optimal Scaling for Various Metropolis–Hastings Algorithms, Statistical Science, 16, 351–367, 2001.

Robinson, P. M.: Efficient Tests of Nonstationary Hypotheses, J. of the American Statistical Association, 89, 1420–1437, 1994.

Slezak, J. and Weron, A.: From physical linear systems to discrete-time series. A guide for analysis of the sampled experimental data, Phys. Rev. E, 91, 053 302, 2015.

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

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

Watkins, N. W.: Bunched black (and grouped grey) swans: Dissipative and non-dissipative models of correlated extreme fluctuations in complex geosystems., Geophys. Res. Lett., 40, 1–9, doi:10.1002/GRL.50103, 2013.

Woodward, W. A., Cheng, Q. C., and Gray, H. L.: A k-factor GARMA long-memory model, J. of Time Series Analysis, 19, 485–504, 1998.

Wuertz, D.: `fArma`: ARMA Time Series Modelling, http://CRAN.R-project.org/package=fArma, R package version 2160.77, 2012.

Table 1. Posterior summary statistics for ARFIMA$(0,0,0)$ process. ~~Average of~~ Results are based on averaging over 100 ~~runs~~independent ARFIMA$(0,0,0)$ simulations for the long-memory parameter d, mean μ and noise variance σ.

	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

Table 2. Posterior model probabilities for time series from figures 6(a)–(b) and 8(a)–(b) for the autoregressive parameter p and moving average parameter q.

(a)

$p\backslash q$	0	1	2	3	4	5	marginal
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	**0.805**	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	

(b)

$p\backslash q$	0	1	2	3	4	5	marginal
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	**0.829**	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	

Table 3. Posterior model probabilities based on simulations of model Eq. (10) for the autoregressive parameter p and moving average parameter q.

$p\backslash q$	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	

Table 4. Posterior model probabilities for Nile minima time series for the autoregressive parameter p and moving average parameter q..

$p\backslash q$	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	

Table 5. Summary posterior statistics for Nile minima time series for the long-memory parameter d, mean μ and noise variance σ.

	mean	std	95% CI endpoints	
d	0.402	0.039	0.336	0.482
μ	1158	62	1037	1284
σ	70.15	1.91	66.46	73.97

Table 6. Posterior model probabilities for Nile minima time series for the autoregressive parameter p and moving average parameter q.

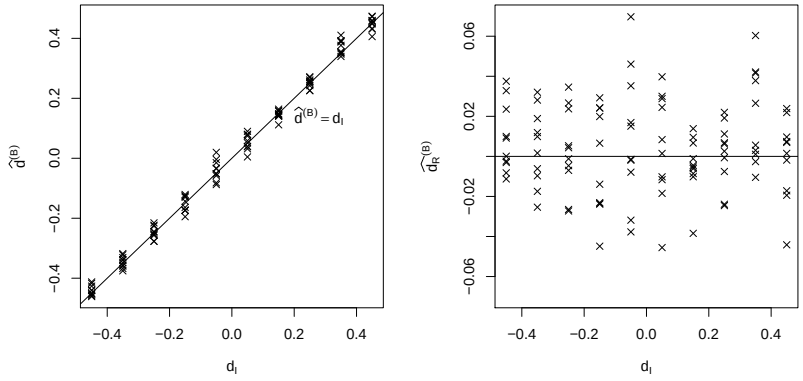
$p \backslash q$	0	1	2	3	4	5	marginal
0	0.638	0.101	0.010	0.000	0.000	0.000	0.750
1	0.097	0.124	0.011	0.000	0.000	0.000	0.232
2	0.007	0.010	0.000	0.000	0.000	0.000	0.018
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000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

Table 7. Posterior summary statistics for CET index for the long-memory parameter d, seasonal long-memory parameter D, mean μ and noise variance σ.

	mean	std	95% CI endpoints	
d	0.209	0.013	0.186	0.235
D	0.040	0.011	0.018	0.062
μ	9.266	0.144	9.010	9.576
σ	1.322	0.015	1.294	1.353

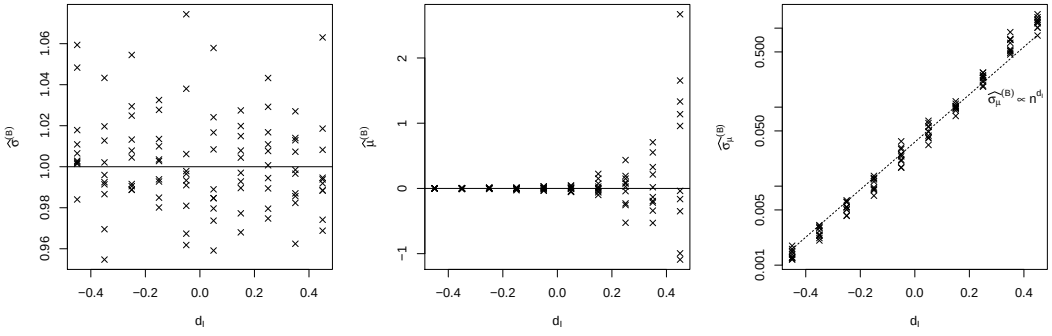
Table 8. Posterior summary statistics for four blocks of CET index for the long-memory parameter d, seasonal long-memory parameter D, mean μ and noise variance σ.

			mean	std	95% CI endpoints	
1659–1744		d	0.277	0.026	0.231	0.332
		D	0.054	0.022	0.013	0.097
		μ	9.036	0.347	8.332	9.702
		σ	1.217	0.027	1.167	1.271
1744–1829		d	0.204	0.028	0.151	0.259
		D	0.017	0.023	−0.028	0.063
		μ	9.107	0.216	8.671	9.533
00		σ	1.348	0.031	1.290	1.409
1829–1914		d	0.172	0.027	0.118	0.223
		D	0.036	0.022	−0.010	0.076
		μ	9.172	0.168	8.859	9.517
		σ	1.364	0.030	1.312	1.429
1914–2000		d	0.163	0.027	0.108	0.213
		D	0.063	0.022	0.023	0.109
		μ	9.591	0.152	9.314	9.906
		σ	1.348	0.030	1.291	1.406



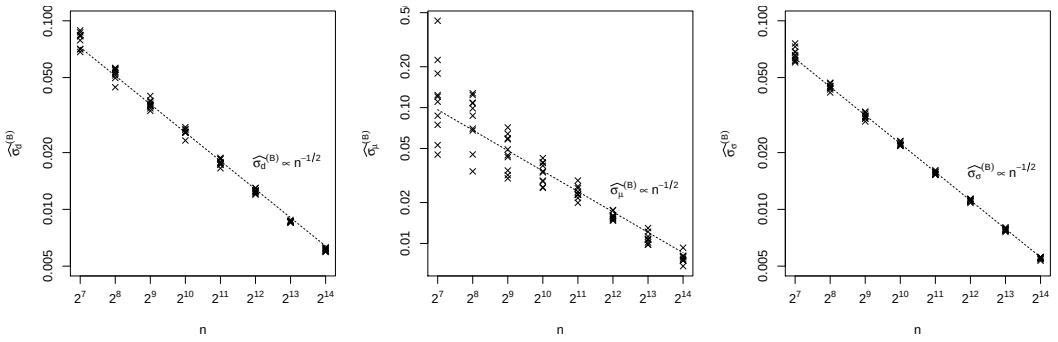
(a) (b)

Figure 1. Posterior outputs; (a) Bayesian estimate $\widehat{d}^{(B)}$ values on the y-axis against the true d_I on the x-axis, (b) residuals $\widehat{d_R}^{(B)}$ from the Bayesian estimate from the truth against that truth, d_I. Each "x" plotted represts one estimate or residual.



(a) (b) (c)

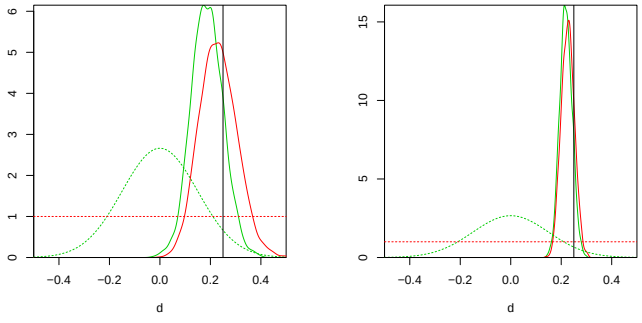
Figure 2. Posterior outputs; (a) Bayesian estimated standard deviation $\widehat{\sigma}^{(B)}$ against true d_I values, (b) Bayesian estimated mean $\widehat{\mu}^{(B)}$ against d_I, and (c) Uncertainty in the posterior for μ, the standard deviation $\widehat{\sigma_\mu}^{(B)}$ against d_I (semi-log scale). Each "x" plotted corresponds to an estimate.



(a) (b) (c)

Figure 3. Posterior outputs from ARFIMA$(0,0,0)$ series; (a) the posterior standard deviation in d, $\widehat{\sigma_d}^{(B)}$ against the sample size n, (b) posterior standard deviation in μ, $\widehat{\sigma_\mu}^{(B)}$ against n, (c) $\widehat{\sigma_\sigma}^{(B)}$ against n (log-log scale).

n	mean difference
128	0.057
256	0.029
512	0.015
1024	0.007



(a) (b)

Figure 4. *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) $n = 2^{10} = 1024$.

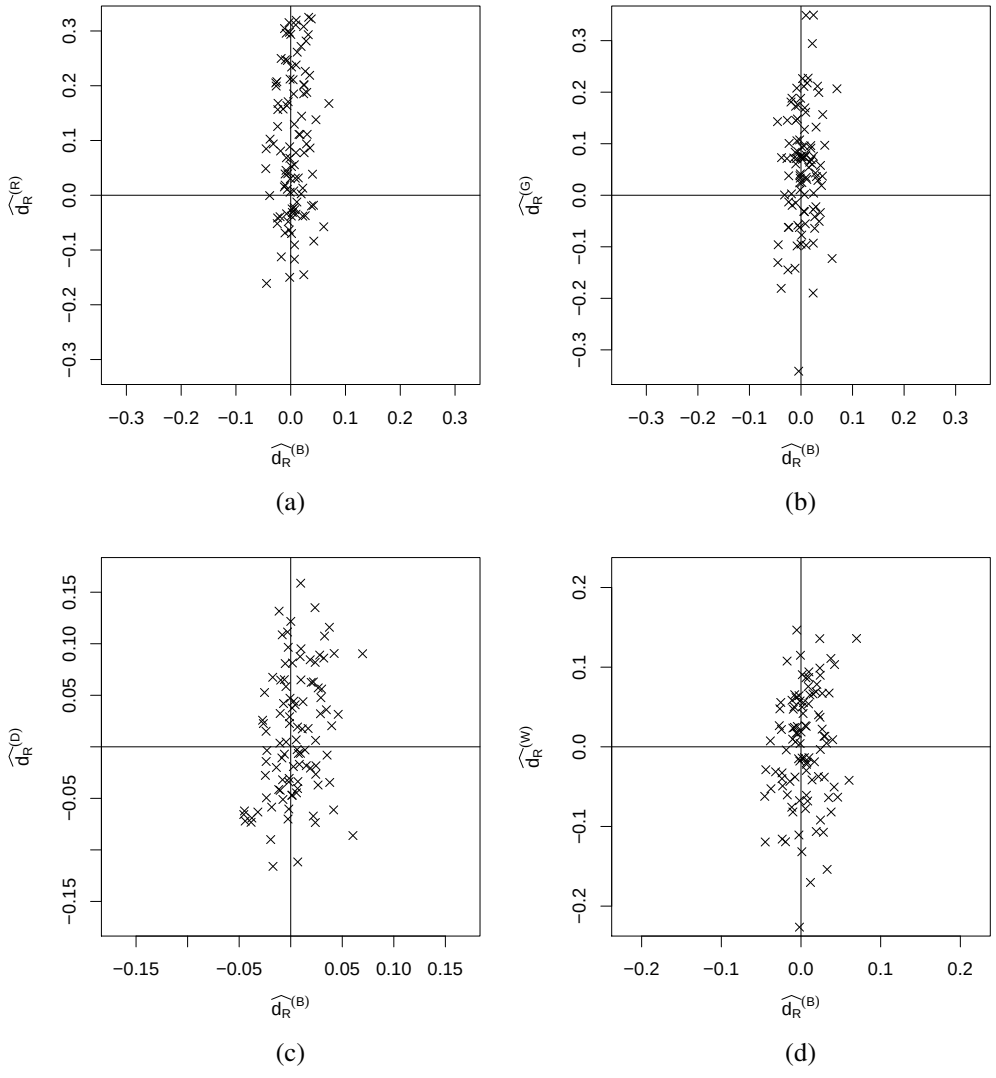


Figure 5. Comparison of Bayesian estimator with common classical estimators; (a) R/S, (b) GPH, (c) DFA, (d) Wavelet.

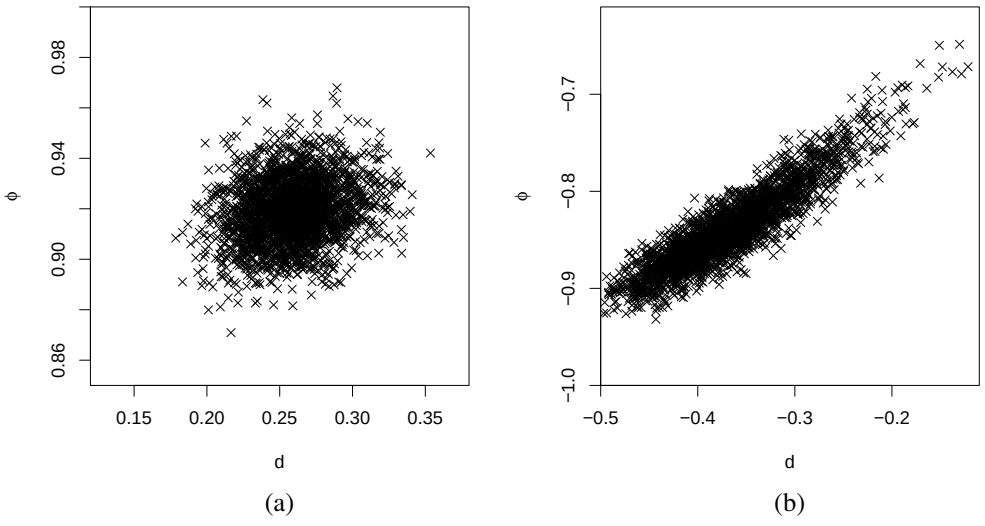


Figure 6. Posterior samples of (d, ϕ); input time series (a) $(1+0.92\mathcal{B})(1-\mathcal{B})^{0.25}X_t = \varepsilon_t$, (b) $(1-0.83\mathcal{B})(1-\mathcal{B})^{-0.35}X_t = \varepsilon_t$.

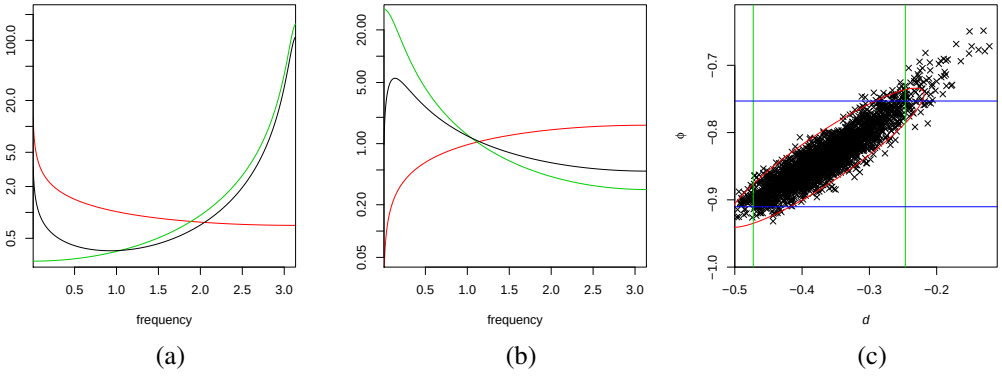
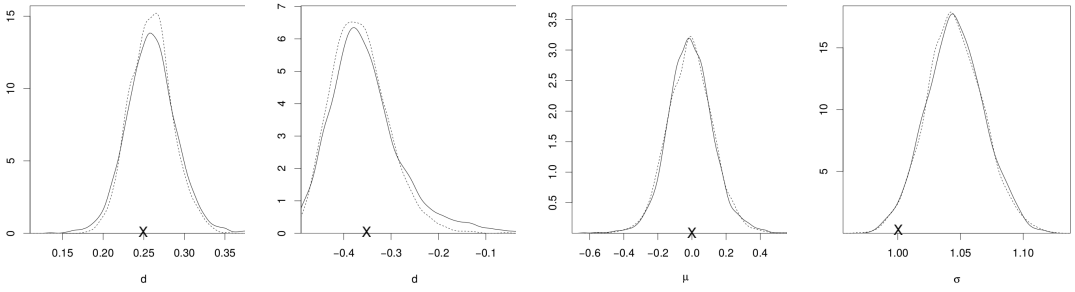
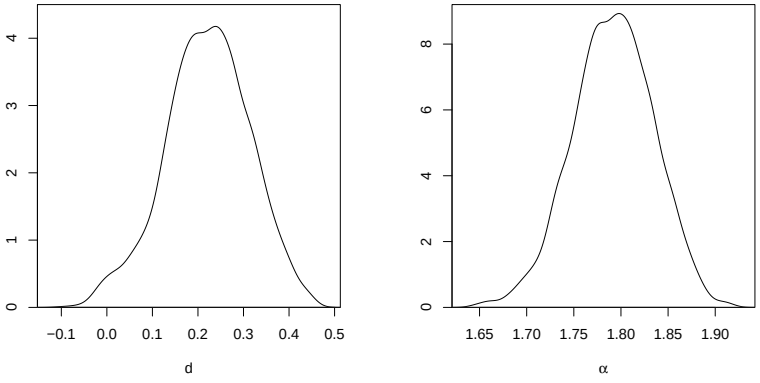


Figure 7. Spectra for processes in figure 6. Green line is relevant ARMA$(1,0)$ process, red line is relevant ARFIMA$(0,d,0)$ process, black line is ARFIMA$(1,d,0)$ process; (a) $(1+0.92\mathcal{B})(1-\mathcal{B})^{0.25}X_t = \varepsilon_t$; (b) $(1-0.83\mathcal{B})(1-\mathcal{B})^{-0.35}X_t = \varepsilon_t$. 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.



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

Figure 8. Marginal posterior density of d from series in figure 6, (a)–(b) respectively. Solid line is density obtained using reversible-jump algorithm. Dotted line is density obtained using fixed $p = 1$ and $q = 0$. The true values are $d_l = 0.25$ and -0.35, respectively. Panels (c)–(d) shows the posterior densities for μ and σ, respectively, corresponding to the series in 6(a); those for 6(b) look similar. The true values are $\mu = 0$ and $\sigma = 1$. True values are marked by an X.



(a) (b)

Figure 9. Marginal posterior densities (a) d, (b) α from the model Eq. (10).

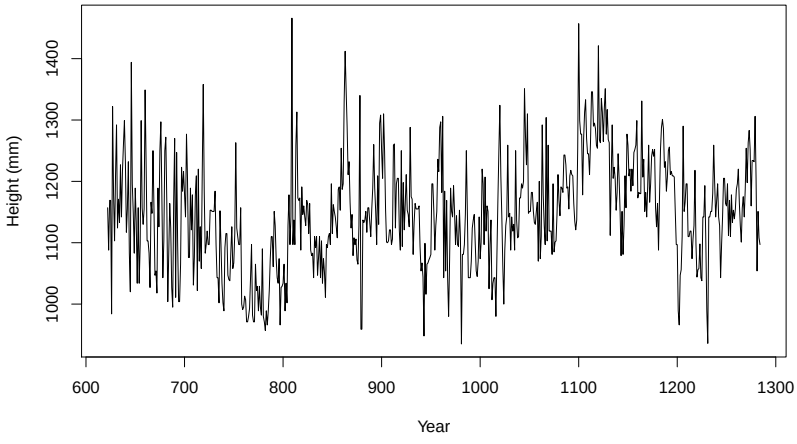
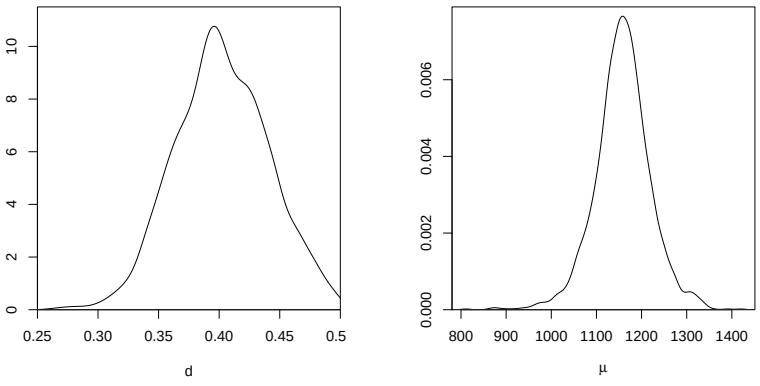


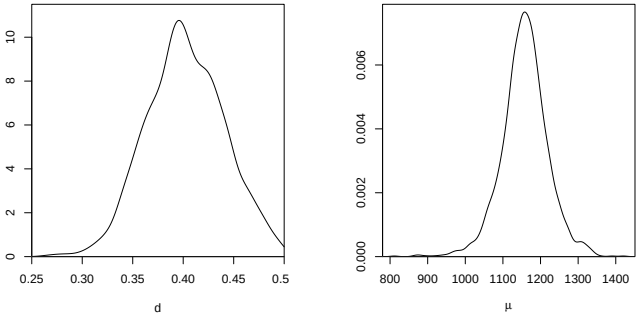
Figure 10. Annual Nile minima time series.



(a) (b)

Figure 11. Marginal posterior densities for Nile minima; (a) d, (b) μ.

	mean	95% CI	
d	~~0.039~~ 0.402	0.336	0.482
μ	~~62~~ 1158	1037	1284
σ	~~1.91~~ 70.15	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.

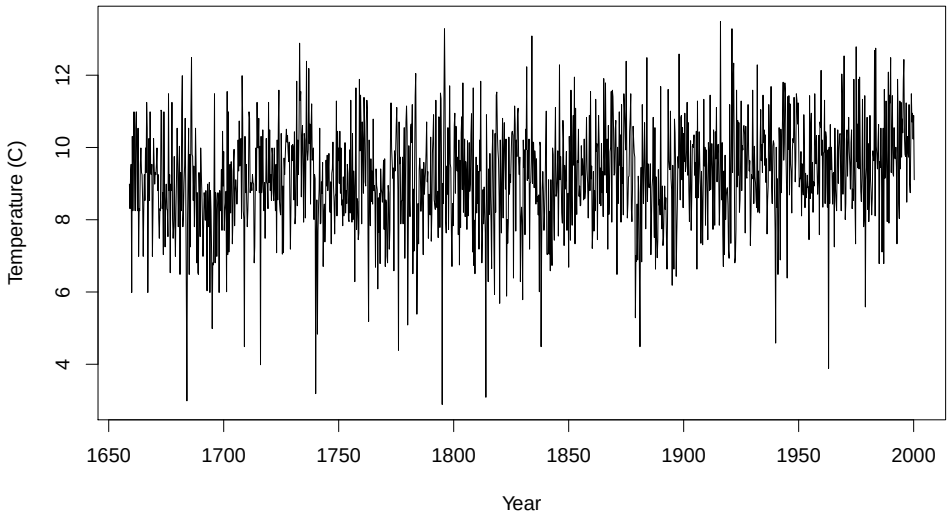
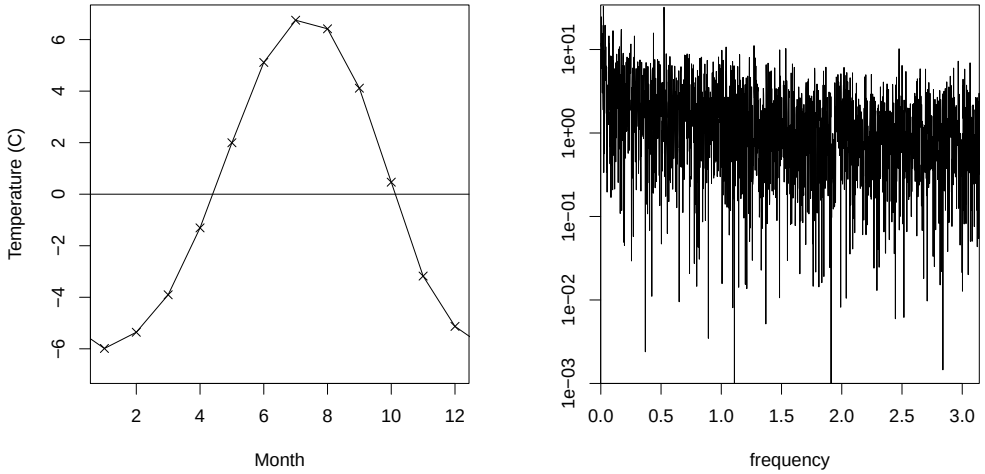


Figure 13. CET time series (deseasonalised).



(a) (b)

Figure 14. CET time series; (a) assumed deterministic seasonal component $S(t)$, (b) spectrum of deseasonalised index.

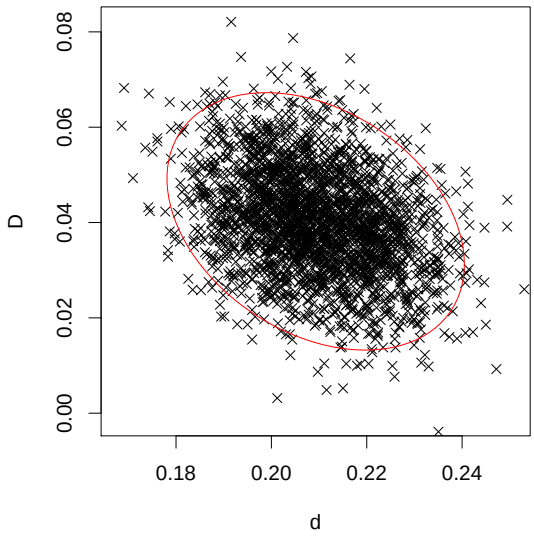
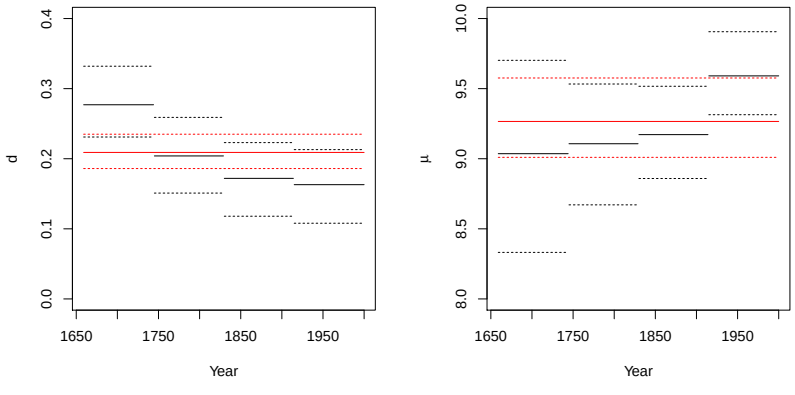


Figure 15. Joint posterior samples of (d, D) with 95% credibility set in red for CET time series.



(a) (b)

Figure 16. CET time series: posterior estimate (solid line) and 95% credibility interval (dotted line) for four blocks (black) and whole index (red) for (a) d, (b) μ.