Bayesian Analysis (2009) **4**, Number 1, pp. 159–190

A Bayesian Approach to Estimating the Long Memory Parameter

Scott Holan*, Tucker McElroy† and Sounak Chakraborty‡

Abstract. We develop a Bayesian procedure for analyzing stationary long-range dependent processes. Specifically, we consider the fractional exponential model (FEXP) to estimate the memory parameter of a stationary long-memory Gaussian time series. In particular, we propose a hierarchical Bayesian model and make it fully adaptive by imposing a prior distribution on the model order. Further, we describe a reversible jump Markov chain Monte Carlo algorithm for variable dimension estimation and show that, in our context, the algorithm provides a reasonable method of model selection (within each repetition of the chain). Therefore, through an application of Bayesian model averaging, we incorporate *all* possible models from the FEXP class (up to a given finite order). As a result we reduce the underestimation of uncertainty at the model-selection stage as well as achieve better estimates of the long memory parameter. Additionally, we establish Bayesian consistency of the memory parameter under mild conditions on the data process. Finally, through simulation and the analysis of two data sets, we demonstrate the effectiveness of our approach.

Keywords: Bayesian model averaging; FEXP; hierarchical Bayes; long-range dependence; reversible jump Markov chain Monte Carlo; Spectral density

1 Introduction

Motivated by applications in econometrics, hydrology and other scientific disciplines, fractionally differenced models have been the subject of extensive research over the past two decades. The differencing parameter characterizes the long-memory behavior of a time series and thus has far-reaching utility in areas such as telecommunications, economics, finance, and geophysics among others. Therefore, it is of considerable interest to provide estimators of the memory parameter that have good asymptotic properties, while remaining practical for use in finite samples.

The long memory process that we consider is Gaussian and has spectral density

$$f(\lambda) = |1 - e^{-i\lambda}|^{-2d} f^*(\lambda), \;\; \lambda \in (-\pi, \pi). \tag{1}$$

Further, we say that the stationary time series $\{x_t\}$ is *long range dependent* or *long*

*Department of Statistics, University of Missouri, Columbia, MO, mailto:holans@missouri.edu

†Statistical Research Division, U.S. Census Bureau, Washington, D.C., mailto:tucker.s.mcelroy@census.gov

‡Department of Statistics, University of Missouri, Columbia, MO, mailto:chakrabortys@missouri.edu

DOI:10.1214/09-BA406

memory if there exists a $d \in (0, .5)$ such that

$$\lim_{\lambda \to 0^+} \frac{f(\lambda)}{\lambda^{-2d}} = k$$

for some constant $k > 0$. Likewise for $d = 0$ we say that the process is short range dependent or short memory. We assume that the function $f^*(\lambda)$ is even, continuous, and bounded away from zero on $[-\pi, \pi]$. Thus d and f^* respectively govern the long-term and short-term behavior of the process.

Although several methods have been proposed for estimating d, two main methods have emerged. The first method is the so-called parametric approach. One example of the parametric approach was introduced by Fox and Taqqu (1986) where the authors provide a fully parametric maximum likelihood procedure for the case where $\{x_t\}$ is strictly Gaussian. Subsequently this method was developed in greater detail by Giraitis and Taqqu (1999). Geweke and Porter-Hudak (1983) suggested a regression approach to the estimation of the long memory parameter using the spectral domain, while Beran (1993) introduced a procedure based on generalized linear regression. The second method commonly used employs semiparametric estimators and has been the subject of extensive research. Some examples are Robinson (1995a,b) and Giraitis and Surgailis (1990), where both authors justified the use of semiparametric inferential procedures. For a comprehensive discussion on long memory time series and fractionally differenced models see Beran (1994), Robinson (2003), Palma (2007) and the references therein.

Although some research on long memory has taken a Bayesian viewpoint, the literature is still very heavily frequentist (see Robinson (2003) for a discussion). The majority of the Bayesian literature consists of parametric characterizations of long-range dependence based on a class of models. Perhaps the most popular approach is to consider the class of all finite order stationary and invertible ARFIMA (Auto-Regressive Fractionally Integrated Moving Average) processes (Hosking 1981). Some examples are the models proposed by Ravishanker and Ray (1997), Pai and Ravishanker (1998), and recently Ko and Vannucci (2006a).

The first nonparametric Bayesian analysis of long memory data was considered by Petris (1997). This work is posed under the assumption that, for $\lambda \in [0, \pi]$,

$$f(\lambda) \propto \lambda^{-2d} \exp\{g(\lambda)\}, \tag{2}$$

where $g(\lambda)$ is meant to capture the short-range dependence in $\{x_t\}$. Further, at the Fourier frequencies, λ_j, the prior probability measure on $g(\cdot)$ is defined by a centered autoregressive process of order p with parameters having a Beta prior distribution. As noted in Liseo et al. (2001), in order for the two components λ^{-2d} and $g(\lambda)$ in (2) to be identifiable, appropriate conditions on $g(\lambda)$ need to be imposed.

In this paper we consider an explicitly defined semiparametric estimator of d based on fitting potentially misspecified parametric models. Specifically, the models we fit are from the FEXP class described by Beran (1993, 1994). The FEXP estimator was originally proposed by Janacek (1982) and later discussed by Robinson (1994). The spectral density of the FEXP model has the form of (1) with $f^*(\lambda)$ given by the exponential model of Bloomfield (1973), i.e.,

$$\log f^*(\lambda) = \sum_{k=0}^{m} b_k \cos(k\lambda), \tag{3}$$

where $b_0, \ldots, b_m$ are real constants and m is a positive integer.

In particular, we propose a hierarchical Bayesian method to estimate the long memory parameter d, by putting near diffuse priors on the model parameters $b_0, \ldots, b_m$ and a restricted flat prior on d. Moreover, the model order parameter m is selected adaptively and thus provides increased flexibility and more precise estimation of d, by helping to effectively discriminate between short range and long range components. The approach we suggest is similar to the Bayesian semiparametric approach of Liseo et al. (2001) to the extent that they consider a spectral method with the short-range dependent portion of the model related (but not equal) to the class of fractional exponential processes introduced by Beran (1993) and to the work of Bloomfield (1973) in the short memory case. However, the method we propose differs from Liseo et al. (2001) in several ways. One distinction between the two methods is that our method is *broad-band* in that the spectral model we impose for the short-range dependent process is modeled on the frequency band $[0, \pi]$; in contrast, the work of Liseo et al. (2001) considers a *narrow-band* estimator of the short memory process by restricting this process to the frequency band $[\lambda^*, \pi]$. This representation induces a model selection problem on the choice of λ^*. The authors suggest treating λ^* as a hyperparameter whose value is ultimately determined by the data. The selection procedure is facilitated via comparison of prior predictive distributions of the observed data set. Further, the method they propose uses a periodogram based Whittle likelihood approximation to the true likelihood (Palma 2007, chap. 4), whereas we consider the exact Gaussian likelihood. Using the exact Gaussian likelihood provides a significant distinction between our models and other models in the FEXP class, as the Toeplitz autocovariance matrix is difficult to deal with computationally. Toward this direction, we provide explicit details of algorithms that facilitate its use. Further, our approach explicitly accounts for the mean by including it as a parameter to be estimated in the model. This increases the utility of our method by providing a time domain representation of the FEXP model that avoids having to account for the true mean by way of the sample mean (Hurvich 2002). This is particularly important when using the FEXP model for the purposes of forecasting (Hurvich 2002), as the sample mean is slow to converge to the true mean when $d \in (0, .5)$. Specifically, if $\overline{x}$ denotes the sample mean then, for $d \in (0, .5)$, $\mathrm{var}(\overline{x}) \sim Cn^{2d-1}$ $(C > 0)$ as $n \to \infty$ (Cheung and Diebold 1994; Chen et al. 2006) and thus potentially degrades model estimates. Additionally, by avoiding the use of a periodogram based approximation to the likelihood (i.e., the Whittle likelihood) and by using the exact Gaussian likelihood in a Bayesian framework instead, our method can accommodate missing data with relatively little extra computational burden.

Another striking advantage of our method comes in the way of model selection (i.e., the order m in (3)); the approach we propose avoids explicit order selection by making the order random through a prior distribution and then estimating it through the model. This variable dimension approach (Sisson 2005) provides added flexibility and reduces uncertainty at the model-selection stage by integrating over *all* possible models from the FEXP class (up to a given finite order). This procedure can be seen as a type of Bayesian model averaging (Hoeting et al. 1999). In this case our estimate of d results in a weighted average of the estimates of d from our adaptive model, with the weights proportional to the probability of occurrence of the different selected model orders. Thus, we consider m a nuisance parameter. Finally, using the exact Gaussian likelihood and allowing the dimension of the parameter space to change, as well as explicitly accounting for the mean, requires a non-trivial extension to the proof of the asymptotic properties.

The use of variable dimension (reversible jump Markov chain Monte Carlo - RJMCMC) methods in long memory models can be found in Breidt and Hsu (2002). In particular, Breidt and Hsu (2002) consider an autoregressive regime-switching model for the dynamic mean structure of a univariate time series. The model they propose allows for the possibility of approximate long memory behavior and uses a RJMCMC method for Bayesian inference on unknown model parameters. Specifically, in the setting they consider, the use of RJMCMC is to allow the location and number of level shifts to be of variable dimension.

Another example of RJMCMC in long memory is provided by Ko and Vannucci (2006b) where the authors develop a wavelet-based procedure for the detection of multiple changes of the long memory parameter. The method they propose uses RJMCMC to detect the location and number of change points in an ARFIMA (p, d_t, q) model. Specifically, in the approach they propose, p and q are of fixed whereas the number of change points is of variable dimension.

In contrast the model we develop[1] is a Bayesian frequency domain long memory model where the model parameters governing the short memory dynamics (i.e., the coefficients of the EXP(m) portion of the model) are of variable dimension and estimated through RJMCMC. As such our FEXP modeling approach provides a type of automatic model order selection along with accurate Bayesian model averaged estimates of the long memory parameter. In addition we argue that our model order selection can be roughly approximated by the BIC (Kass and Raferty 1995) and provide asymptotic properties for the long memory parameter d.

A related approach to Holan et al. (2007) was proposed by Rousseau and Liseo (2007). This work presents theoretical properties for Bayesian nonparametric estimation of the spectral density of a long memory Gaussian time series. In particular the authors prove, under specific assumptions on the prior distribution, posterior consistency on $f(\cdot)$ and d under the exact Gaussian likelihood. Additionally, the authors describe the rate of convergence of the posterior sequence, which depends on the structure of the prior in a significant way, and consider what they call the fractionally exponential (FEXP) family of priors.

This paper is organized as follows. In Section 2 we describe the modeling framework,

[1] An earlier version of this research appears as Holan et al. (2007).

including choice of priors and discussion of implementation of the likelihood calculations. Section 3 provides the details of the Markov Chain Monte Carlo (MCMC) and Reversible Jump Markov Chain Monte Carlo procedures used for estimation. Specifically we present algorithms for both the fixed dimension and variable dimension cases. Section 4 develops the asymptotic properties of the *a posteriori* estimates of the long memory parameter. In Section 5, our methodology is investigated through simulation as well as applied to two real data sets, the Nile River Minima data (Beran 1994) and the Central England Temperature data (Manley 1974; Pai and Ravishanker 1998). Finally, Section 6 contains discussion. For convenience of exposition all proofs are left to the appendix.

2 Hierarchical Bayesian FEXP Model

2.1 Bayesian FEXP

Suppose that, after accounting for the mean, the time series of interest follows the model

$$f(\lambda) = |1 - e^{-i\lambda}|^{-2d} \exp\left\{\sum_{k=0}^{m} b_k \cos(k\lambda)\right\} \tag{4}$$

for $\lambda \in [-\pi, \pi]$. Here f is the spectral density of the stationary process, d is the long memory parameter, and the coefficients b_l parameterize the short memory aspects of the process. Stationarity is maintained by considering $d < 1/2$; since we focus on long memory, we also assume that $d \geq 0$. Note that when $d = 0$, f reduces to the spectral density of an EXP(m) process, which has short memory (Bloomfield, 1973). We start by defining the parameter space $\Theta_m = [0, .5) \times \mathbb{R}^{m+1}$, which explicitly depends on the order m. We use the notation $\theta_m = (d, b_0, b_1, \cdots, b_m)$ for a typical element of Θ_m. Henceforth we use the notation f_{θ_m} to denote the spectral density given in (4) whenever we wish to make the dependence on model parameters explicit. Conditional on m being known, it is simple to write down the Gaussian likelihood for a sample $x^{(n)} = \{x_1, x_2, \cdots, x_n\}$ of size n:

$$p(x^{(n)}|\theta_m, \mu) = (2\pi)^{-n/2} \, |\Sigma(f_{\theta_m})|^{-1/2} \exp\left\{-\frac{1}{2}\Big(x^{(n)} - \mu\underline{1}\Big)' \Sigma^{-1}(f_{\theta_m})\Big(x^{(n)} - \mu\underline{1}\Big)\right\}, \tag{5}$$

where $\underline{1}$ denotes a vector of ones. We use the following notation for covariance matrices: the Toeplitz covariance matrix of dimension n of a stationary process with spectral density f is written $\Sigma(f)$. Its jkth entry is the autocovariance function at lag $j - k$ given by

$$\Sigma_{jk}(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) \, e^{i\lambda(j-k)} \, \mathrm{d}\lambda.$$

Perhaps one of the most salient features of our approach arises in the choice of the model order m. In a frequentist setting it is most common to choose the order of a FEXP model through order selection criteria (see Hurvich 2001). Similarly, in the

Bayesian context one could maintain the model selection approach and choose m using Bayes factor (Gelman et al. 2004). However, our approach is more flexible as it takes full advantage of the Bayesian paradigm by putting a prior on m and selecting the number of terms to be included in the model adaptively. Since the main goal of our analysis is estimation of d, we incorporate the choice of m into the modeling framework. This added layer of flexibility in terms of m enhances the accuracy of the estimate of d by reducing the underestimation of uncertainty at the model-selection stage.

2.2 Likelihood Calculations

Implementation of the Bayesian approach under the exact Gaussian likelihood poses several challenges. First one must form the Toeplitz autocovariance matrix $\Sigma(f_{\theta_m})$ in (5). This requires calculation of the autocovariances of lag L, $\{c_L\}_{L=0}^{n-1}$, given only knowledge of f_{θ_m}. Further, once the autocovariance matrix is constructed there is still a need to calculate both $|\Sigma(f_{\theta_m})|$ and $(x^{(n)} - \mu\underline{1})'\Sigma^{-1}(f_{\theta_m})(x^{(n)} - \mu\underline{1})$ (conditional on μ). These quantities present computational difficulties because here the autocovariances decline slowly (at a power law rate) and $\Sigma(f_{\theta_m})$ is quite ill-conditioned (Chen et al., 2006).

Calculation of the autocovariances proceeds via a spectral factorization method on the short memory portion of (1) (Pourahmadi 1983; Hurvich 2002). Specifically, given $b_0, \ldots, b_m$ one can write down an equivalent MA(∞) representation (under mild conditions) using the formula

$$\beta_j = \frac{1}{2j}\sum_{k=1}^{j} k b_k \beta_{j-k},$$

with $\beta_0 \equiv 1$ and β_j ($j = 1, 2, 3, \ldots$) equal to the MA(∞) coefficients. Furthermore, the innovation variance, σ_ϵ^2 can be expressed as $\sigma_\epsilon^2 = 2\pi \exp(b_0)$. Next, let $\widetilde{c}_h$ denote the theoretical autocovariance associated with a specific short memory EXP(m) model. Here,

$$\widetilde{c}_h = \sigma_\epsilon^2 \sum_{j=h}^{\infty} \beta_j \beta_{j-h} \ \ (h \geq 0).$$

Since β_j decay exponentially fast, a good approximation to $\widetilde{c}_h$ can be obtained using the truncated version $\widetilde{c}_{h,N}$ for N sufficiently large. Note that $\widetilde{c}_{h,N}$ may be efficiently calculated by taking advantage of the FFT (Fast Fourier Transform); see Hurvich (2002) for a comprehensive discussion.

Similarly, let $\{c_s^*\}$ denote the autocovariances of an ARFIMA$(0, d, 0)$ process with unit innovation variance. Again, since $\widetilde{c}_h$ decay exponentially fast, c_L can be approximated accurately ($c_{L,Approx.}$) using $\{c_s^*\}$ and $\widetilde{c}_{h,N}$ in a "splitting method" (Hurvich 2002; Bertelli and Caporin 2002). Additionally, as suggested by Hurvich (2002) the sequence $\{c_{L,Approx.}\}_{L=0}^{n-1}$ can be used to simulate a zero-mean Gaussian realization $x_1, \ldots, x_n$ from the FEXP process $\{x_t\}$ using the Davies-Harte algorithm (Davies and

Harte 1987). The algorithm is exact, in the sense that the autocovariances of the simulated time series exactly match those used as inputs to the algorithm. This algorithm will be used in Section 5 where our methodology is investigated through simulated data. Finally, the Toeplitz matrix $\Sigma(f_{\theta_m})$ can be constructed using $\{c_{L,Approx.}\}_{L=0}^{n-1}$.

Ultimately the likelihood will be embedded in an MCMC procedure, Gibbs sampling with Metropolis-Hasting steps (Robert and Casella, 2004). For this purpose it is more convenient to work with the log likelihood rather than the likelihood itself. Thus the quantities needed to calculate $\log p(x^{(n)}|\theta_m, \mu)$ are the quadratic form $y^{(n)'}\Sigma^{-1}(f_{\theta_m})y^{(n)}$ and $\log|\Sigma(f_{\theta_m})|$, where $y^{(n)} = (x^{(n)} - \mu\underline{1})$. Conditional on μ, the first of these quantities, the quadratic form, can be accurately calculated to any desired degree of precision using conjugate gradient (CG) or preconditioned conjugate gradient algorithms (PCG) (Chen et al. 2006). Similarly $\log|\Sigma(f_{\theta_m})|$ can be accurately approximated (numerically) in $O(1)$ operations (Bötcher and Silberman 1999; Chen et al. 2006). Therefore we can calculate the exact log likelihood with a high degree of precision. For a detailed discussion see Chen et al. (2006).

2.3 Hierarchical Priors and Posteriors

We now fix ideas by setting out some specific hierarchical priors. Let $b = (b_0, b_1, b_2, \cdots)$ and $\sigma^2 = (\sigma_0^2, \sigma_1^2, \sigma_2^2, \cdots)$ be infinite sequences. Then we assign hierarchical priors on the unknown parameters as follows:

$$\begin{aligned}
b_k|m, \sigma_k^2 &\sim \begin{cases} \mathcal{N}(0, \sigma_k^2); & \text{if } k \leq m, \\ \Delta_0; & \text{if } k > m, \end{cases} \\
d &\sim \text{Uniform}(0, 1/2), \\
\sigma_k^2|m &\sim \begin{cases} IG(\alpha, \beta); & \text{if } k \leq m, \\ \Delta_0; & \text{if } k > m, \end{cases} \\
m &\sim \text{Discrete Uniform}\{1, M\}, \\
\mu &\sim \mathcal{N}(\mu_0, \sigma_\mu^2).
\end{aligned}$$

Alternatively, if one has prior information on the order of the model, this information can be incorporated easily into this framework by choosing a Truncated Poisson$\{\eta, M\}$ prior for m, where η is the rate parameter and M represents the maximum possible order of the model. In this context typically the rate parameter is user defined based on prior knowledge, see Denison et al. (1998) for a related discussion. Additionally, the support for the prior on d is chosen in order to maintain stationarity. Here, we denote the point mass at zero via Δ_0, the issue being that b_k and σ_k^2 for k exceeding a given m must be defined to be identically zero. As usual IG denotes the Inverse Gamma distribution, so that σ_k^2 has pdf

$$p(\sigma_k^2) \propto \exp\left(-\frac{\beta}{\sigma_k^2}\right)(\sigma_k^2)^{-(\alpha+1)}.$$

Note that the choices of α, β, M, μ_0, σ^2_μ and η (if appropriate) are determined by the practitioner. The selection of priors is usually problem-specific. Ideally, one would like to elicit these priors from past history. However, for most practical applications, such information is unavailable. In such cases it seems reasonable, at least from robustness considerations, to use near-diffuse priors for the hyperparameters of the hierarchical model. Although this would provide a reasonable approach under a fairly broad range of settings we find that FEXP model coefficients tend to be "relatively" small. For this reason one appropriate strategy is to define the notion of near-diffuse based on the context of the problem. Further this choice of priors can be justified by the heuristic argument that typically in long memory settings one requires long time series in order to be able to adequately estimate quantities of interest. Therefore under most prior specifications there is enough data to overcome the effect due to the prior. As we will find in the next section, we have taken "non-informative" proper priors for b_k, d, σ^2_k, μ and m, in the sense that on the scale of typical FEXP coefficients the prior provides relatively little information. Moreover, having chosen proper priors, we maintain propriety of the posterior and ensure the objectivity of our approach. Note that in Section 4 we present asymptotic results under relaxed specification of the hierarchical priors.

In general, hierarchical Bayesian models generate a richer class of models that automatically produce shrinkage estimators of parameters. These shrinkage estimators, by borrowing strength, usually perform better than the regular estimators. Additionally, taking m random we have increased the precision of our estimates as well as added flexibility to our model by allowing the model to adaptively choose its own order given the data. This will be revealed in Section 5, when we illustrate our methodology using both simulated data and actual data analyses. Below, we consider the joint posterior distribution of the parameters. Noting that in the prior distributions b, σ^2 and μ do not depend on d, and the prior distribution of d does not depend on m, we have

$$\begin{aligned} p(\mu, b, \sigma^2, d, m|x^{(n)}) &\propto p(x^{(n)}|b, \sigma^2, d, m, \mu)p(b|\sigma^2, m)p(\sigma^2|m)p(d)p(m)p(\mu) \\ &\propto (2\pi)^{-n/2}\left|\Sigma(f_{\theta_m})\right|^{-1/2} \exp\left\{-\frac{1}{2}y^{(n)\prime}\Sigma^{-1}(f_{\theta_m})y^{(n)}\right\} \\ &\times \prod_{k=0}^{m}(\sigma^2_k)^{-1/2}\exp\left\{-\frac{b^2_k}{2\sigma^2_k}\right\} \times \prod_{k=0}^{m}\exp\left\{-\frac{\beta}{\sigma^2_k}\right\}(\sigma^2_k)^{-(\alpha+1)} \\ &\times 1_{[0,1/2)}(d) \times \frac{1}{M} \times (2\pi\sigma^2_\mu)^{-1/2}\exp\left\{-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma^2_\mu}\right\}, \end{aligned} \tag{6}$$

where $y^{(n)} = (x^{(n)} - \mu\underline{1})$. Our primary objective is to find the marginal posterior distribution of d given $x^{(n)}$, and use the posterior mean to obtain an estimate for d. The exact posterior distribution $p(d|x^{(n)})$ is given in (15) of Section 4, and the posterior mean is

$$\widehat{d} = \int d\,\mathrm{d}P(d|x^{(n)}), \tag{7}$$

where $\mathrm{d}P(z|\cdots) = p(z|\cdots)\mathrm{d}z$ denotes a shorthand notation and z is some vector of parameters. Asymptotic properties of this estimate are considered in Section 4. In practice, the marginal posterior $p(d|x^{(n)})$ is intractable, as its evaluation requires high dimensional numerical integration. Thus we adopt the popular alternative approach of MCMC, which requires generating samples from the full conditionals of the different parameters given the remaining parameters and the data. Some further details in this direction are provided in Section 3 below.

3 Estimation

3.1 Full Conditionals and Fixed Dimensional MCMC

Since some of the full conditionals are not of standard form we have used a Metropolis within Gibbs algorithm (Robert and Casella 2004). Below we list these full conditionals. Although our general proposed model makes adaptive selection of m (the model order), in this section we take m to be fixed and describe our implementation procedure. Subsequently in Section 3.2 we describe a RJMCMC approach (Green 1995) for the case in which m is random.

First, since the prior distribution of σ^2 does not depend on d or μ, and the prior distributions of d and μ do not depend on m, it is easy to see that the full conditional of σ_k^2 is conjugate. As a result, it is straight forward to show that the full conditional of σ_k^2 is IG and is given by

$$p(\sigma_k^2|d, m, b_0, b_1, \ldots, b_k, \ldots, b_m) \quad \sim \quad IG\left\{(\alpha + 1/2),\, (b_k^2/2 + \beta)\right\}. \qquad (8)$$

Although the full conditional of $b = (b_0, \ldots, b_m)$ is not conjugate under this model it is equally straight forward to derive and is given by

$$\begin{aligned} p(b|m, \sigma^2, d, \mu, x^{(n)}) \quad &\propto \quad (2\pi)^{-n/2}\, |\Sigma(f_{\theta_m})|^{-1/2} \exp\left\{-\frac{1}{2} y^{(n)\prime} \Sigma^{-1}(f_{\theta_m}) y^{(n)}\right\} \\ &\times \quad \prod_{k=0}^{m} (\sigma_k^2)^{-1/2} \exp\left\{-\frac{b_k^2}{2\sigma_k^2}\right\}. \end{aligned} \qquad (9)$$

Furthermore, the full conditional of d can be expressed as

$$\begin{aligned} p(d|\sigma^2, m, b, \mu, x^{(n)}) \quad &\propto \quad (2\pi)^{-n/2}\, |\Sigma(f_{\theta_m})|^{-1/2} \exp\left\{-\frac{1}{2} y^{(n)\prime} \Sigma^{-1}(f_{\theta_m}) y^{(n)}\right\} \\ &\times \quad 1_{[0,1/2)}(d). \end{aligned} \qquad (10)$$

Finally the full conditional of μ is given by

$$p(\mu|b, \sigma^2, d, m, x^{(n)}) \sim \mathcal{N}(\mu^*, \sigma_\mu^{*2}), \qquad (11)$$

where $\mu^* = c_2/c_1$, and $\sigma_\mu^{*2} = 1/c_1$ with

$$c_1 = \left\{\underline{1}'\Sigma^{-1}(f_{\theta_m})\underline{1} + \sigma_\mu^{-2}\right\} \qquad \text{and} \qquad c_2 = \left\{x^{(n)\prime}\Sigma^{-1}(f_{\theta_m})\underline{1} + \mu_0\sigma_\mu^{-2}\right\}.$$

In the case of fixed model dimension implementation, MCMC is relatively straightforward given the calculated likelihood. The general algorithm we use is a Gibbs sampler (Gelfand and Smith, 1990), although several Metropolis-Hastings (M-H) steps will be required. The pdfs given in (8) and (11) are standard and so can be sampled from directly (given c_1 and c_2). Unfortunately, generating samples from (9) and (10) is more complicated and requires M-H updates. The algorithm proceeds as follows:

Step 1: Set initial values for all parameter values.

Step 2: Generate samples from (8).

Step 3: Using a Random-Walk, M-H sample the b_k's one at a time from (9) holding all b_j fixed for $j \neq k$.

Step 4: Using an Independent M-H step sample from (10).

Step 5: Generate samples from (11).

Step 6: Repeat.

For the implementation of the M-H algorithm, one needs a candidate generating density. Chib and Greenberg (1995) have several proposals in this regard. For specificity, let b_k denote the current state for the parameter b_k; we then draw a candidate value b_k^* using a $\mathcal{N}(b_k, \delta_k^2)$ proposal distribution, where the δ_k are user defined tuning parameters. The algorithm accepts b_k^* as a new value of b_k with acceptance probability

$$\alpha_{b_k} = \min\left\{1, \frac{p(b_k^*|m, \sigma^2, d, \mu, x^{(n)})}{p(b_k|m, \sigma^2, d, \mu, x^{(n)})}\right\},$$

where $p(b_k|m, \sigma^2, d, \mu, x^{(n)})$ is given in (9). Similarly for d^* draw a candidate value from a $U(0, 1/2)$ proposal distribution and accept d^* as a new value of d with acceptance probability

$$\alpha_d = \min\left\{1, \frac{p(d^*|\sigma^2, m, b, \mu, x^{(n)})}{p(d|\sigma^2, m, b, \mu, x^{(n)})}\right\},$$

where $p(d|\sigma^2, m, b, \mu, x^{(n)})$ is given in (10).

The algorithm proposed here is not unique. In fact, one alternative would be to sample the b_k's jointly using a multivariate proposal distribution. Often, in cases where the parameters are correlated blocking will improve convergence. However, on occasion blocking parameters in a single group becomes more detrimental than beneficial. Specifically, as the number of component b_k's increases it becomes more likely to have components fall in the tails of the full conditional distribution. This in turn will produce a small value for the test ratio and hence very few proposed values will be accepted, thus slowing convergence (Gamerman and Lopes 2006, chap. 6). Balancing these different factors, we sample the b_k's one at a time. Further, specific choices of proposal distributions for b_k and d can be modified. It should be noted that the choices presented here are a matter of preference as their implementation is straight forward while yielding good acceptance probabilities.

In the case where the model dimension is fixed *a priori* it is natural to estimate models using several different orders (values of m). Then, with these models in hand, one can choose a final model through the use of Bayes factor or DIC; see Gelman et al. (2004) for further discussion. Since we propose variable dimension modeling of d, where the order m is taken as a model parameter, we do not pursue discussion of order selection further in this section.

3.2 Adaptive Model Selection Procedure and RJMCMC

The variable dimension case is substantially more difficult to handle as we are potentially changing dimensions on each MCMC iteration. In this case, m is a random parameter to be adaptively estimated from the data. To facilitate model estimation we require the full conditional distribution of m. Thus, for $m = 1, \ldots, M$, the full conditional of m can be written as

$$\begin{aligned} p(m|b,\, \sigma^2, d, \mu, x^{(n)}) \quad &\propto \quad (2\pi)^{-n/2}\, |\Sigma(f_{\theta_m})|^{-1/2} \exp\left\{ -\frac{1}{2} y^{(n)\prime} \Sigma^{-1}(f_{\theta_m}) y^{(n)} \right\} \\ &\times \quad \prod_{k=0}^{m} (\sigma_k^2)^{-1/2} \exp\left\{ -\frac{b_k^2}{2\sigma_k^2} \right\} \times \prod_{k=0}^{m} \exp\left\{ -\frac{\beta}{\sigma_k^2} \right\} (\sigma_k^2)^{-(\alpha+1)} \\ &\times \quad \frac{1}{M}, \end{aligned} \tag{12}$$

where $y^{(n)} = (x^{(n)} - \mu \underline{1})$. Now, for notational convenience, let Ψ denote the collection of parameters in our model (i.e., $\Psi = \Psi_m = (d, b_0, \ldots, b_m, \sigma_0^2, \ldots, \sigma_m^2, \mu)$). To sample from $p(\Psi|x^{(n)})$ we need to use the reversible jump algorithm (Green 1995). Here we outline the steps we employ to traverse the posterior probability surface; see Denison et al. (1998), Denison et al. (2002) and references therein for a general discussion. Under the reversible jump setup, we allow three types of moves: BIRTH (B), DEATH (D), and MOVE (S) where these moves are defined by

B: propose with probability p_m^b to increase model dimension by 1,

D: propose with probability p_m^d to decrease model dimension by 1,

S: propose with probability p_m^s to move in the same model dimension.

Note that in the models we consider we require at least one term always in the model (in addition to b_0, i.e., b_0 and b_1). Therefore, the probabilities are given as follows:

$$\begin{aligned} &p_m^b = p_m^d = p_m^s = 1/3; \;\; \text{for} \;\; m = 2, \ldots, M-1, \\ &p_1^b = p_1^s = p_M^d = p_M^s = 1/2, \\ &p_M^b = p_1^d = 0. \end{aligned}$$

Note that $p_m^b + p_m^d + p_m^s = 1$, for all m. Finally, let Ψ' denote a candidate parameter vector and $\Psi^{(t)}$ the value of the parameter vector at the tth iteration. The algorithm proceeds as follows:

Step 1: Set starting values.

Step 2: Generate $u_1 \sim U(0,1)$

- if $u_1 < p_m^b$ then goto B move and obtain Ψ'
- otherwise if $p_m^b \leq u_1 \leq p_m^b + p_m^d$ goto D move and obtain Ψ'
- otherwise goto S move and obtain Ψ'.

Step 3: Suppose we goto B move

(i.) generate $b_{m+1} \sim \mathcal{N}(0,\tau)$
(ii.) generate entire ($m+1$ dimensional) σ^2 vector
(iii.) complete a MCMC iteration as in fixed dimensional case and obtain Ψ'.

Step 3$'$: In D move delete b_m, σ_m^2 and obtain Ψ' via a MCMC iteration. In the S move no addition or deletion is needed, we simply perform a standard MCMC to obtain new Ψ'.

Step 4: Generate $u_2 \sim U(0,1)$

- if $u_2 < \min\{1, BF(\Psi', \Psi^{(t)}) \times R\}$ then $\Psi^{(t+1)} = \Psi'$
- else $\Psi^{(t+1)} = \Psi^{(t)}$.

Step 5: Repeat.

The value τ in the above algorithm is a user defined tuning parameter chosen, for each data set, in order to obtain an approximate acceptance rate of 20-30% in our B move. In the simulations and applications pursued here (Section 5) we found the value $\tau = 1$ provided satisfactory acceptance rates (20-30%). Here,

$$R = \begin{cases} p_{m+1}^d / p_m^b; & \text{if B move,} \\ p_{m-1}^b / p_m^d; & \text{if D move,} \\ 1; & \text{if S move,} \end{cases} \tag{13}$$

and

$$BF(\Psi', \Psi^{(t)}) = \frac{\int_{\mathbb{R}^{2m'+4}} \mathrm{d}P(\Psi' | x^{(n)})}{\int_{\mathbb{R}^{2m+4}} \mathrm{d}P(\Psi^{(t)} | x^{(n)})}, \tag{14}$$

where R and BF denote the proposal ratio and Bayes factor respectively. An analytic form for BF and a related discussion regarding its appropriateness as an order selector in our setting can be found in Appendix B. Note that in the case where a Truncated Poisson$\{\eta, M\}$ prior is imposed on m, R requires slight modification.

Remark 1. In Appendix B we argue that in our context, for large samples, the BIC provides an approximation to the logarithm of the Bayes factor. Thus, heuristically, the heavy penalty term associated with the BIC enables us to choose a parsimonious model. However, it is important to note that although we show, in our context, BIC can serve as a large-sample approximation to the logarithm of Bayes factor, our model selection is performed at each iteration of the RJMCMC using the Bayes factor (calculated using the marginal likelihood from the Metropolis-Hastings output, see Chib and Jeliazkov (2001)) not the BIC approximation. Using the BIC approximation would discard prior information, including the prior on the model dimension and thus would be difficult to justify in a fully Bayesian procedure.

Remark 2. Although our algorithm does not appear to be explicitly sampling from the full conditional of m, it is straight-forward to show that $p(\Psi'|x^{(n)})$ is equivalent to (12). Thus the comparison made in Step 4 is implicitly acting as an M-H step for the parameter m (after integrating out the effects of the nuisance parameter σ^2) with our proposal distribution equal to the birth/death process.

3.3 Missing Data

One advantage of our approach is that missing data can be handled with relative ease. In previous studies involving the FEXP model there has been a tendency to use Whittle approximations to the exact Gaussian likelihood. Most of these approximations do not facilitate estimation in settings where there is missing data. The reason for this is that usually estimation is based on periodogram estimates obtained from the full data. Using the exact Gaussian likelihood the estimation we consider is facilitated via the autocovariance structure and not a periodogram based estimate. Therefore, we can tackle missing data with little added effort.

The method we propose only requires initial (starting) values for the missing data points (in addition to starting values for the algorithm). With these values in hand we can obtain estimates for x_{mis} based on the predictive distribution. Using obvious notation this can be expressed as

$$f(x_{mis}|x_{obs}) = \int_{\Psi} f(x_{mis}|x_{obs}, \Psi)\mathrm{d}P(\Psi|x_{obs}).$$

Although we do not pursue such estimation here, this simply amounts to adding an extra Metropolis within Gibbs step to the RJMCMC algorithm proposed in Section 3.2. That is, suppose after step t we have a sample $\widetilde{\Psi}^{(t)}$ of the parameter vector. Then we may obtain a sample $x_{mis}^{(t)}$ by drawing from the full conditional distribution $f(x_{mis}|x_{obs}, \widetilde{\Psi}^{(t)})$; see Palma (2007, chap. 11) and Gelman et al. (2004, chap. 21) for further details.

4 Asymptotic Properties

Here we assume the general setup of Section 2.1, so that the priors on the parameters are allowed to be more general than the specific choices of Section 2.3. The prior on m is some discrete distribution that is compactly supported, denoted by $p(m)$. We also have priors $p(b_k|\sigma_k^2, m)$ and $p(\sigma_k^2|m)$ for all $k \geq 0$, but with these distributions equalling the point mass at zero if $k > m$. From these we define the joint priors $p(b|\sigma^2, m)$ and $p(\sigma^2|m)$. We also have a prior $p(d)$ on d, but this does not depend on m (this can be relaxed in the proof). Thus, similar to (6), in general the posterior distribution of all the parameters is

$$p(\mu, b, \sigma^2, d, m|x^{(n)}) \propto p(x^{(n)}|\mu, b, \sigma^2, d, m)p(b|\sigma^2, m)p(\sigma^2|m)p(d)p(m)p(\mu).$$

The last five quantities on the right hand side are known priors, and the first quantity is just the likelihood function. For notational convenience – since $p(b|\sigma^2, m)\times$ $p(\sigma^2|m) = p(b, \sigma^2|m)$ anyways – let $\phi = (b, \sigma^2)$ denote the joint variable, with distribution $p(\phi|m)$. Conditional on m, ϕ takes values in $\mathbb{R}^{m+1} \times \mathbb{R}^{m+1}$ (though the variance parameters σ^2 are always non-negative). Denote this space by Φ_m. Using the shorthand $\mathrm{d}P(z|\cdots)$ for $p(z|\cdots)\mathrm{d}z$ (where z is some vector of parameters), the posterior distribution for d is given by

$$p(d|x^{(n)}) = \frac{\sum_m p(d) \int_{\Phi_m \times \mathbb{R}} p(x^{(n)}|\mu, \phi, d, m)\mathrm{d}P(\phi|m)\mathrm{d}P(\mu)p(m)}{\sum_m \int_{\Phi_m \times \mathbb{R}^2} p(x^{(n)}|\mu, \phi, d, m)\mathrm{d}P(\phi|m)\mathrm{d}P(\mu)\mathrm{d}P(d)p(m)}. \tag{15}$$

Recall that the posterior mean $\widehat{d}$ is given by (7). We next present a theoretical result for the asymptotics of $\widehat{d}$. The data itself is generated from a Gaussian process with spectral density given by (4), with fixed parameters $\widetilde{m}$, $\widetilde{\mu}$, and $\widetilde{\theta}_{\widetilde{m}}$ that are unknown to the practitioner (here $\widetilde{\theta}_m = (\widetilde{d}, \widetilde{b}_0, \widetilde{b}_1, \cdots, \widetilde{b}_m)$ for any m). We next establish asymptotic consistency of $\widehat{d}$ for the true parameter $\widetilde{d}$; our method follows that of Liseo et al. (2001), appropriately generalized to our hierarchical model.

Theorem 1. *Let $p(m)$ and $p(d)$ denote the prior distributions for m and d respectively where $p(m)$ is a compactly supported discrete distribution and $p(d)$ is a compactly supported continuous distribution on $(0, .5)$. Additionally, define priors $p(b_k|\sigma_k^2, m)$ and $p(\sigma_k^2|m)$ for all $k \geq 0$ such that for $k > m$ these priors equal the point mass at zero. From these priors define the joint priors $p(b|\sigma^2, m)$ and $p(\sigma^2|m)$ and let ϕ denote the joint variable (b, σ^2) with distribution $p(\phi|m)$. Conditional on m, $p(\phi|m)$ takes values in $\mathbb{R}^{m+1} \times \mathbb{R}^{m+1} = \Phi_m$ (though the variance parameters are always non-negative), and*

$$p(d|x^{(n)}) \propto \sum_m p(d) \int_{\Phi_m \times \mathbb{R}} p(x^{(n)}|\mu, \phi, d, m)\, dP(\phi|m)\, dP(\mu)p(m).$$

Assuming that the data are generated from a Gaussian process with spectral density given by (4) with fixed unknown parameters $\widetilde{m}$ and $\widetilde{\theta}_{\widetilde{m}} = \widetilde{d}, \widetilde{b}_0, \widetilde{b}_1, \cdots, \widetilde{b}_{\widetilde{m}}$, $\widehat{d}$ given by (7) converges to $\widetilde{d}$ in probability.

Remark 3. The proof of Theorem 1 still holds when the variance of b_i are not considered i.i.d. as specified in Section 2.3. In fact, the proof is based off of $p(\sigma^2|m)$, and nowhere uses the idea that the variances are i.i.d. Thus, σ_k^2 could be taken dependent and not identically distributed. Further, each distribution could also depend on m explicitly, e.g., if they are specified as identically distributed Inverse Gamma, the α and β parameters could depend on m. This is permissible since the distribution of m has finite support. Finally, making these changes to the prior specification would increase the generality of the model without substantially increasing the computational complexity.

5 Case Studies

In this section we illustrate our proposed Bayesian FEXP model for estimation of the long memory parameter d. We first begin by demonstrating the effectiveness of our approach by way of three simulation studies using both fixed and variable dimension estimation. Subsequently we analyze the benchmark Nile River Minima data (Beran 1994), and the Central England Temperature data (Manley 1974; Pai and Ravishanker 1998). The prior parameters α and β from the IG prior on σ_k^2 (for all k) were fixed at 2.333 and 1.333 respectively, thus providing an IG prior having mean equal to 1 and variance equal to 3. For the prior on μ the value of μ_0 and σ_μ^2 were fixed at 0 and 10^5 respectively and for the prior on m we fix $M = 6$ to allow models with up to 7 coefficients (the simulation in Section 5.2 had $M = 10$). Next, for all model fitting, both simulated and actual data, we run a single MCMC chain for 10,000 iterations discarding the first half for burn in. Convergence of the MCMC is verified through trace plots of the posterior. The starting values for the b_k parameters, d and μ were chosen randomly while the starting values for the σ_k^2's were set equal to 1. Specifically, the starting values for the b_k coefficients and μ were generated from a $\mathcal{N}(0,1)$ distribution and d was generated from a uniform distribution with support $(0,1/2)$. Furthermore the tuning parameters, δ_k, in the random walk M-H step were fixed at .5 for b_0 and .2 for b_k $(k \neq 0)$. Lastly, the posterior expectation (7) is approximated by $\widehat{d} = {n^*}^{-1} \sum_{i=1}^{n^*} d_{(i)}$, where $d_{(i)}$ are generated from the ith MCMC sample, and n^* denotes the number of such samples used to estimate the parameters after the initial burn in.

5.1 Simulation Study

To illustrate the performance of our method we consider two simulation studies. The data used for this empirical investigation are generated using the Davies-Harte algorithm (Davies and Harte 1987), see Section 2.2. Although the Davies-Harte algorithm simulates a zero-mean Gaussian process, we estimated d via (5) since in practice, for any given realization, the mean of the simulated process may be non-zero. Furthermore, in cases where the simulated data has mean zero our model should be able to provide a suitable estimate (similar to the non-zero mean case). The first simulation we consider consists of a parameter space where the dimension is held fixed. We denote the first simulation as the "Nile-F" simulation as the parameter values chosen for this model were obtained from log periodogram regression estimates of a FEXP(3) model fit to the

Nile River data (to be analyzed in Section 5.3). Note that log periodogram regression is mean invariant (Chen et al. 2006) and thus the mean does not need to be taken into account. The exact parameter values used to generate the data for the Nile-F simulation are $b = (b_0, b_1, b_2, b_3) = (6.640, -.121, -.232, -.044)$ and $d = .479$. The second simulation, denoted the "Synthetic" simulation, consists of two models, one having a fixed dimension parameter space and the other with a variable dimension parameter space. The parameter values chosen for this simulation are $b = (b_0, b_1, b_2) = (2, -1, 1)$ and $d = .35$.

Further for both simulations we generated 10 different realizations of the data, with sample size $n = 500$, and estimated d and b_0 using our fixed dimension approach, our variable dimension approach, log periodogram FEXP (lpFEXP) (Hurvich 2002), and the Geweke Porter-Hudak estimator (GPH) (Geweke and Porter-Hudak 1983). For the Synthetic simulation we also provide maximum likelihood estimates (mle). The mles were calculated by numerically optimizing the likelihood surface using the *optim* command in R (R Development Core Team 2007). For the Nile-F simulation we do not provide mles due to convergence issues in the numerical optimization routine owing to the fact that d is near the boundary of the parameter space for stationary processes. Although more sophisticated likelihood estimation methods, such as the EM algorithm, may be able to improve convergence we do not pursue them here. Note that both lpFEXP and GPH are mean invariant and as such the mean does not need to be taken into account in order to estimate d. For convenience of exposition we denote the fixed dimension estimation in the Synthetic simulation as "Synthetic-F" and the variable dimension estimates as "Synthetic-V". Note that in these simulations we assumed the model order was correctly specified for the lpFEXP and mle as well as our fixed dimension method. In addition, it should be emphasized that all models were fit to the same 10 realizations of the simulated data, including our variable dimension approach. The results for the Synthetic and Nile simulations are reported in Table 1. In particular, Table 1 compares the frequentist properties of our estimator to the log periodogram regression and GPH estimates. That is we compare the mean and standard deviations of the posterior mean, $\widehat{d}^{(1)}_{BFEXP}$, the lpFEXP estimate, $\widehat{d}_{lpFEXP}$, GPH estimator, $\widehat{d}_{GPH}$ and the mle estimator, $\widehat{d}_{mle}$. For completeness we also provide the median of the posterior medians, $\widehat{d}^{(2)}_{BFEXP}$ and a comparison of $\widehat{b}^{(1)}_{0,BFEXP}$, $\widehat{b}^{(2)}_{0,BFEXP}$, $\widehat{b}_{0,lpFEXP}$ and $\widehat{b}_{0,mle}$ since the innovation variance, $2\pi \exp(b_0)$, may be another parameter of interest.

As a result of this simulation, we find that our method out-performs both log periodogram regression FEXP and GPH, Table 1. Specifically, our estimators exhibit substantially lower standard error in all cases and are in close agreement with the truth. Moreover, the improvement experienced using our estimators can be easily quantified by considering their respective mean squared errors (MSEs). First in the Nile-F simulation the MSEs for $\widehat{d}_{lpFEXP}$ and $\widehat{d}_{GPH}$ are equal to .0301 and .0256 respectively while the MSE for $\widehat{d}^{(1)}_{BFEXP}$ equals .0035. This constitutes a reduction in MSE of 86%. Additionally, the Nile-F simulation demonstrates that using lpFEXP for estimation of d can lead to estimates in the nonstationary region. Conversely, our estimates are guaranteed to have $d \in [0, 1/2)$. Similarly in the Synthetic simulation the MSEs for both $\widehat{d}_{lpFEXP}$ and

Posterior Results: Simulated Data

	Synthetic-F	Synthetic-V	Nile-F
		(mode $m = 2$)	
d_{true}	.350	.350	.479
$\widehat{d}^{(1)}_{BFEXP}$	.364 (.064)	.351 (.086)	.429 (.030)
$\widehat{d}^{(2)}_{BFEXP}$	.381	.368	.431
$\widehat{d}_{lpFEXP}$	.395 (.151)	-	.501 (.172)
$\widehat{d}_{GPH}$	.423 (.140)	-	.499 (.159)
$\widehat{d}_{mle}$	.335 (.081)	-	-
$b_{0,true}$	2.000	2.000	6.640
$\widehat{b}^{(1)}_{0,BFEXP}$	1.975 (.069)	1.973 (.073)	6.623 (.074)
$\widehat{b}^{(2)}_{0,BFEXP}$	1.977	1.977	6.624
$\widehat{b}_{0,lpFEXP}$	1.986 (.107)	-	6.616 (.101)
$\widehat{b}_{0,mle}$	1.985 (.073)	-	-

Table 1: Results of simulation study for the Synthetic-F, Synthetic-V and Nile simulations. Note that 10 simulations were used in the construction of these estimates and that for the Bayesian estimates one MCMC chain was used with 10,000 iterations and the first half of the iterations were discarded for burn in. Additionally, the superscripts 1 and 2 denote the mean of the posterior means and the median of the posterior medians respectively. The number in parenthesis denotes the standard deviation, where applicable.

$\widehat{d}_{GPH}$ are equal to .025 whereas $\widehat{d}_{mle}$ equals .0069. In contrast, the MSE for our fixed dimension estimate (Synthetic-F) $\widehat{d}^{(1)}_{BFEXP}$ equals .004. This constitutes reduction in MSE of 84% over $\widehat{d}_{lpFEXP}$ and $\widehat{d}_{GPH}$ and a reduction in MSE of 42% over $\widehat{d}_{mle}$.

On the other hand, using our variable dimension approach (Synthetic-V), the MSE for $\widehat{d}^{(1)}_{BFEXP}$ equals .0073. It should be emphasized that this estimate is extraordinary in that the reduction in MSE is 70% over both $\widehat{d}_{GPH}$ and a correctly specified model order estimate $\widehat{d}_{lpFEXP}$. As expected, in the case of $\widehat{d}_{mle}$ we see that there is a slight price to be paid for using our variable dimension approach over a correctly specified mle. This price amounts to a reduction in MSE of 5.8% over $\widehat{d}^{(1)}_{BFEXP}$. This slightly higher MSE, for $\widehat{d}^{(1)}_{BFEXP}$ versus $\widehat{d}_{mle}$, can be attributed to the additional source of uncertainty experienced from estimating m (the model order). Of course, this comparison assumes that we would know the correct model order for the mle *a priori* which we would not in practice. Additionally, we notice from this simulation study that for the variable dimension case our model is able to pick up the right order over 70% of the time. This in a way provides protection against model misspecification. That is, while benefiting from the protection of model order misspecification we still obtain tremendous gains in MSE over several popular (correctly specified) alternative estimators. This “automatic model order selection” is indeed a novelty of our approach.

Alternatively, looking at each individual simulation (results available on request) for the variable dimension case we identify that estimation is less precise for the 30% of the cases where the model has not picked up the correct order. This, in some sense, sends a warning that when working with real data it is important to choose the order of the model carefully. Particularly, when we have no prior notion about the correct model order it is dangerous to choose the order in an *ad hoc* fashion using log periodogram regression. Although there has been some research undertaken for model selection in this setting (Moulines and Soulier 2000), the order selection criteria usually rely on asymptotic theory and thus may be unsuitable for use in many practical situations.

5.2 Model Misspecification

For the types of models we consider it is fairly typical to assume that the coefficients b_k decay to zero at an exponential rate (Hurvich and Brodsky 2001). As such, in practice, the maximum model order M can be chosen either to contain the true model order or be large enough that the remaining model coefficients are not significantly different from zero. In fact, in practice, one would choose M large enough so that the estimated model orders are always less than M. If the estimated model order achieves the value M frequently during the MCMC iterations one would choose M larger and re-estimate the model. Nonetheless, it is possible that the true model order is greater than M. For example, this occurs in the case where the true model is a FAR(1) (fractional autoregressive model of order 1). The FAR(1) model can be equivalently expressed as a FEXP(∞) model. In this case the coefficients $b_k = \xi^k/k!$ for $k = 1, 2, \ldots$, $b_0 = \log(\sigma^2/2\pi)$ for σ^2 equal to the innovation variance and ξ equal to the autoregressive parameter. Therefore, estimating a FAR(1) model using our approach leads to a model misspecification.

To illustrate the effect that this type of model misspecification has on our estimation procedure we conducted a simulation. In particular we generated 10 realizations from a FAR(1) model with $d = .35$, $\xi = .5$, sample size $n = 500$, and innovation variance equal to one. The FAR(1) models were generated using the R contributed package "Fracdiff". In particular, we chose $M = 10$ to allow models with up to 11 coefficients, since $b_{10} = 2.7 \times 10^{-10}$; however, note that the true model order is infinite. We ran our MCMC for 10,000 iterations discarding the first 5000 for burn in. Convergence was assessed through trace plots of the posterior. For comparison, we also estimated d with a correctly specified FAR(1) model using the *fracdiff* command (see the R contributed package "Fracdiff"). This command estimates d using maximum likelihood with the likelihood approximated using the fast and accurate method of Haslett and Raftery (1989). The mean of our estimator for d was .427 with standard deviation of .043 and MSE=.008 (across the ten simulated data sets), whereas the FAR(1) estimator for d had mean=.211 with standard deviation=.146 and MSE=.041. Thus our estimator exhibits a 80.49% reduction in MSE over the FAR(1) estimator. A histogram of the model dimensions over all 50,000 MCMC iterations is shown in Figure 1. One thing to note is that only one model had dimension greater than or equal to 8 during the MCMC iterations and no model ever had dimension greater than 9. Specifically, one model had dimension 8 a total of 78 iterations and dimension 9 during 5 iterations. In this case, even with a misspecified model our method performs well.

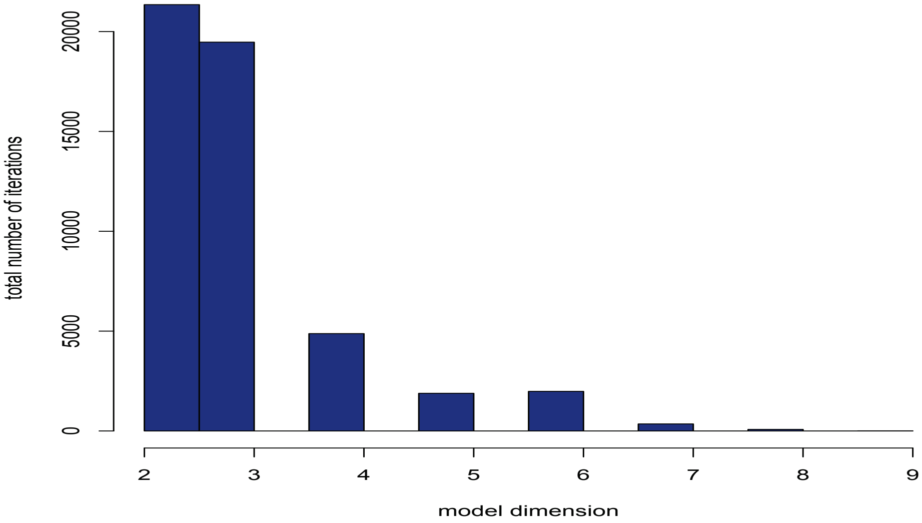

Figure 1: Histogram of Number of Model Coefficients selected from all 50,000 MCMC iterations for the misspecification simulation.

The results of this simulation should be compared with the results presented in Section 4. In particular, it may seem conflicting that consistency results hold despite model misspecification. However, it should be emphasized that the consistency result is for the estimate of the long memory parameter and in our case the model order parameter is not directly connected with d. Therefore, higher values of m simply help discriminate between short range and long range components. Thus, m and $(b_0, b_1, \cdots, b_m)$ are essentially nuisance parameters that have a more limited effect on the distribution of d.

5.3 Nile River Data

We now illustrate our proposed methodology with the benchmark Nile River minima data (Beran 1994). This data set contains yearly minimal water levels of the Nile River for the years 622 to 1281 ($n = 663$). A detailed description of the data can be obtained from Beran (1994). The first model we estimate has m fixed *a priori* and equal to 3. This analysis produces interesting results. In particular, when we simulate a process using coefficients equal to the log periodogram coefficients (i.e., Nile-F simulation, Section 5.1) we find close agreement between $\widehat{d}^{(1)}_{BFEXP}$ and d, see Table 1. However, the parameter estimate of d obtained in our analysis differs from that obtained using the lpFEXP.

Posterior Results: Nile River Minima Analysis		
	Nile - Fixed ($m = 3$)	Nile - Variable (mode $m = 1$)
$\widehat{d}^{(1)}_{BFEXP}$	.419 (.053)	.387 (.042)
95% HPD - $\widehat{d}^{(1)}_{BFEXP}$	(.322,.499)	(.316,.475)
$\widehat{d}^{(2)}_{BFEXP}$	.428	.383
$\widehat{d}_{lpFEXP}$	.479	-
$\widehat{d}_{GPH}$	.504	.504
$\widehat{b}^{(1)}_{0,BFEXP}$	6.664 (.052)	6.659 (.039)
95% HPD - $\widehat{b}^{(1)}_{0,BFEXP}$	(6.551,6.753)	(6.593,6.733)
$\widehat{b}^{(2)}_{0,BFEXP}$	6.665	-
$\widehat{b}_{0,lpFEXP}$	6.640	-

Table 2: Results of Nile River Minima analysis. Note that for the Bayesian estimates one MCMC chain was used with 10,000 iterations and the first half of the iterations were discarded for burn in. Additionally, the superscripts 1 and 2 denote the mean of the posterior means and the median of the posterior medians respectively. The number in parenthesis denotes the standard deviation.

In fact, the estimates produced by lpFEXP are highly variable and often lie outside the stationary region. This leads us to believe that, for this data set (with $m = 3$), the estimate of d based on a FEXP model estimated via log periodogram regression is perhaps overstated. In short, this phenomenon may be due in part to having fixed the model order *a priori*. One way around this peculiarity is to let m be a model parameter to be estimated from the data and to estimate d via our variable dimensional approach (Section 3.2). Again, it is important to note that similar to the Nile-F simulation we do not provide mles for this analysis due to convergence issues in the numerical optimization routine.

In order to demonstrate the utility of our variable dimension approach we re-analyzed the Nile River data using the reversible jump algorithm of Section 3.2. In terms of estimating d and b_0, the results obtained from this analysis are consistent with the results from the fixed dimension case, see Table 2. Furthermore, even though we chose a non-informative prior on m, we obtained parsimonious models from our MCMC samples with the mode of m equal to 1 (i.e., a model with only 2 coefficients b_0 and b_1), see Figure 2. This aspect of our approach is notable as it exemplifies the fact that the data is driving the model rather than the prior distribution. Additionally, our variable dimension approach preferred the model with 2 coefficients in over 75% of the RJMCMC iterations after the initial burn in, see Figure 2. Lastly, the estimates we obtain (Table 2) agree with the estimates of Ko and Vannucci (2006a) ($\widehat{d} = .3793$, with 95% credible interval (.327,.427)) and with Beran (1994, Section 10.3) ($\widehat{d} = .38$).

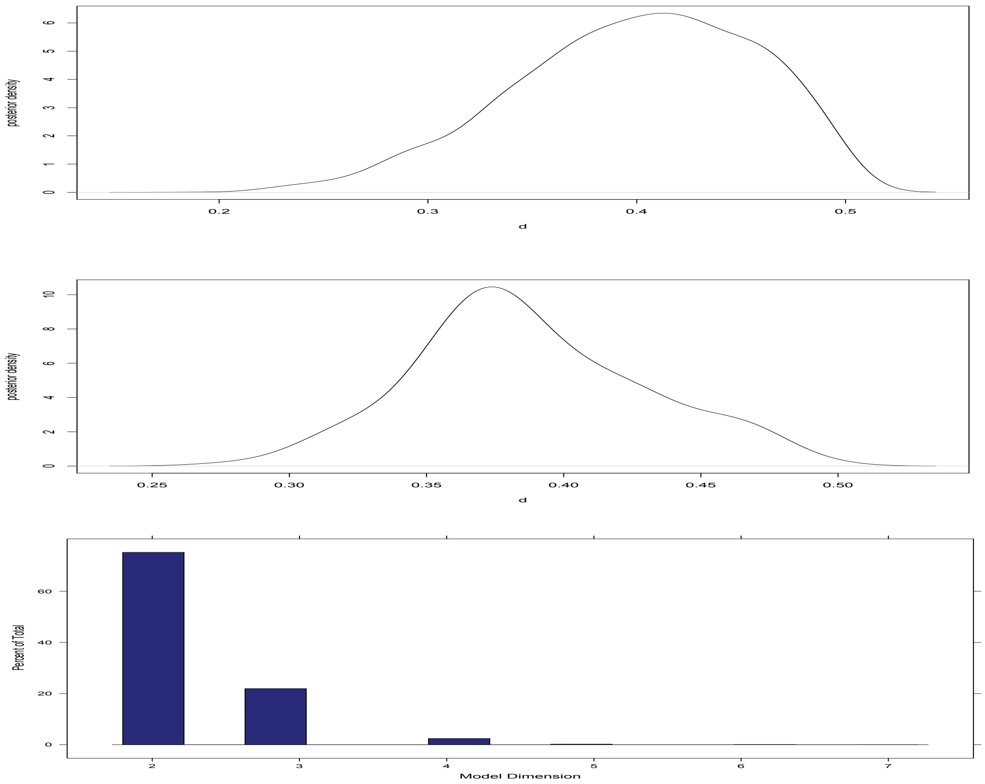

Figure 2: Nile River analysis - 95% HPD - d and histogram of Number of Model Coefficients selected in variable dimension analysis. Top panel is 95% HPD - d for fixed dimensional analysis, middle plot is 95% HPD - d for variable dimensional analysis, and lower plot is a histogram of Number of Model Coefficients selected in variable dimension analysis.

5.4 Hadley Centre Central England Temperature

As a final illustration of our methodology we analyzed the annual mean temperatures for Central England (Manley 1974; Pai and Ravishanker 1998). We use the first 300 observations so that estimates of d can be fairly compared to the ARFIMA estimates of Pai and Ravishanker (1998). In total we estimated 3 models, 2 models had fixed dimension while 1 model is of variable dimension. The fixed dimension models were chosen to have $m = 4$ and $m = 1$. The model having $m = 4$ was chosen via exploratory analysis using log periodogram regression FEXP (Hurvich 2002) while $m = 1$ was chosen to equal the posterior mode of m from the variable dimension analysis. As was alluded to previously, choosing the correct model order in the fixed dimension case can be crucial to proper analysis. For instance, when the model order is chosen as $m = 4$

Posterior Results: Hadley Temperature Analysis			
	Hadley - Fixed		Hadley - Variable
	$(m = 4)$	$(m = 1)$	(mode $m = 1$)
$\widehat{d}^{(1)}_{BFEXP}$	.256 (.113)	.345 (.070)	.330 (.057)
95% HPD - $\widehat{d}^{(1)}_{BFEXP}$	(.058,.472)	(.225,.483)	(.216,.435)
$\widehat{d}^{(2)}_{BFEXP}$	.254	.344	.333
$\widehat{d}_{lpFEXP}$	.445	.401	-
$\widehat{d}_{GPH}$	.280	.280	-
$\widehat{d}_{mle}$	.163 (.127)	.315 (.074)	-
$\widehat{b}^{(1)}_{0,BFEXP}$	-2.899 (.086)	-2.895 (.085)	-2.897 (.057)
95% HPD - $\widehat{b}^{(1)}_{0,BFEXP}$	(-3.061,-2.727)	(-3.034,-2.705)	(-2.998,-2.781)
$\widehat{b}^{(2)}_{0,BFEXP}$	-2.902	-2.898	-2.900
$\widehat{b}_{0,lpFEXP}$	-2.902	-2.903	-
$\widehat{b}_{0,mle}$	-2.928 (.082)	-2.916 (.082)	-

Table 3: Results of Hadley Temperature analysis. Note that for the Bayesian estimates one MCMC chain was used with 10,000 iterations and the first half of the iterations were discarded for burn in. Additionally, the superscripts 1 and 2 denote the mean of the posterior means and the median of the posterior medians respectively. The numbers in parenthesis denote the posterior standard deviation for the Bayesian estimates and asymptotic standard errors for the mle.

we obtain radically different results using our method, lpFEXP, GPH and mle, see Table 3. The question then becomes which estimate do we trust? This illustrates the complexity and need for effective model order selection. In contrast, under our variable dimension approach we get an estimate $\widehat{d}^{(1)}_{BFEXP} = .330$ with standard deviation of .057 and 95% HPD equal to (.216, .435). The mean and standard deviations of the marginal posterior distribution of d obtained by Pai and Ravishanker (1998) are .24(.04) for ARFIMA(0,d,0), .34(.09) for ARFIMA(0,d,1), .28(.05) for ARFIMA(1,d,1), and .32(.07) for ARFIMA(1,d,0). Moreover, the authors obtain similar estimates using maximum likelihood estimation. Our estimate corroborates the estimates obtained in Pai and Ravishanker (1998) while also eliminating the need for model selection, instilling confidence in us that our estimate is valid. Further, under our fixed dimension analysis, with the model order taken equal to the posterior mode of m from the variable dimension analysis, both analyses (fixed and variable dimension) produce similar results and were consistent with the mle ($m = 1$). The posterior densities and histogram of the model order can be seen in Figure 3. Finally, the estimate we obtain for μ (using the posterior mean) equals 9.14 with posterior standard deviation equal to .005 (variable dimension analysis). This agrees with the estimates obtained in Pai and Ravishanker (1998) as well as with maximum likelihood.

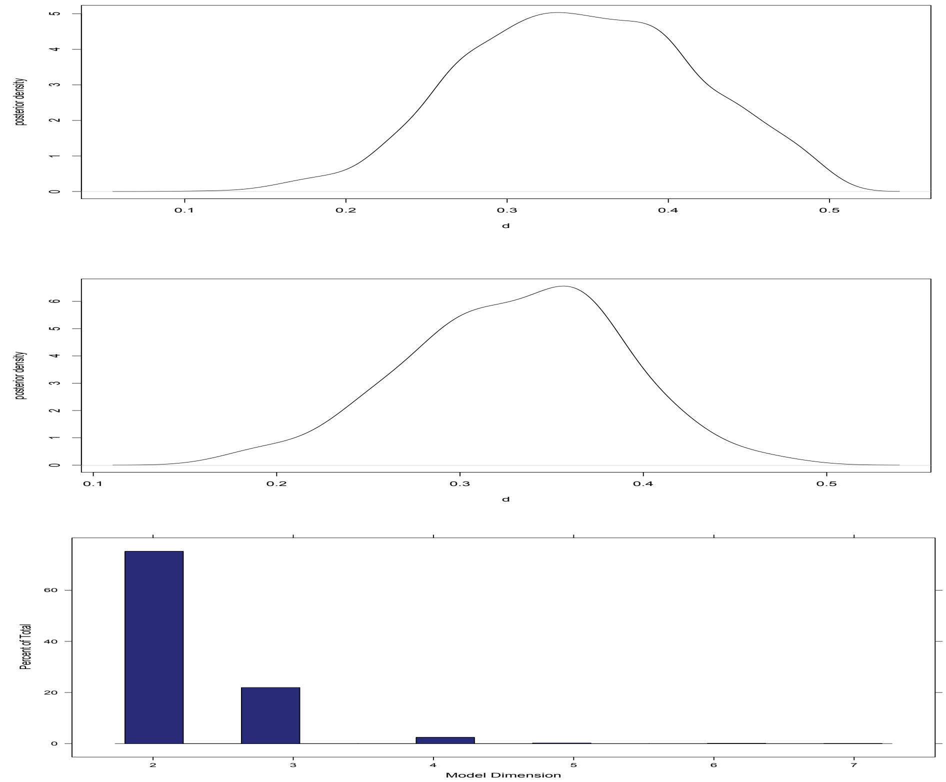

Figure 3: Hadley Temperature analysis - 95% HPD - d and histogram of Number of Model Coefficients selected in variable dimension analysis. Top panel is 95% HPD - d for fixed dimensional analysis, middle plot is 95% HPD - d for variable dimensional analysis, and lower plot is a histogram of Number of Model Coefficients selected in variable dimension analysis.

6 Discussion

The general modeling approach we propose extends the current long memory literature in several ways. First we provide a Bayesian estimate of the FEXP model based on the exact Gaussian likelihood. Second we provide the details for algorithms that facilitate quick computation. Further, we allow the dimension of our parameter space to take the form of a random parameter to be estimated in the model. Therefore, the method we propose is hierarchical and integrates over *all* possible models, thus reducing underestimation of uncertainty at the model-selection stage.

In our implementation we impose a non-informative prior on the model order m; however, in the setting where it is advantageous to incorporate prior beliefs about m into the model we can choose the Truncated Poisson prior for m. Thus, our method is quite flexible and can be implemented on a broad range of long range dependent processes. Additionally, our constructed RJMCMC performs the model selection via the estimated Bayes factor at each iteration. We argue (see Appendix B) that in our context, for large samples, the BIC provides an approximation to the logarithm of the Bayes factor. Thus, heuristically, the heavy penalty term associated with the BIC enables us to chose a parsimonious model. This is reflected in both Hadley and Nile River data where our variable dimension model selects a model with 2 coefficients (b_0, b_1) more than 70% of the time in the MCMC calculations respectively.

The hierarchical structure we propose can be implemented with two different goals in mind. First, if the goal of the analysis is strictly estimation of d, our method can be viewed as Bayesian model averaging (Hoeting et al. 1999). In this case our estimate of d results in a weighted average of the estimates of d from our adaptive model, with the weights proportional to the probability of occurrence of the different selected model orders. This approach improves the precision of our estimate in contexts where the "true" model order is unknown *a priori*. Second, if the goal of the analysis is to produce a spectral model then one can appeal to the FEXP$(\widehat{m})$ model, where $\widehat{m}$ denotes the posterior mode of m. With this model, and the estimated mean, one can also obtain an equivalent AR (autoregressive) representation using the formulas in Hurvich (2002). Since the primary goal of our procedure is estimation of d, the long memory parameter, we do not pursue detailed discussion of this model usage here. However, both of these uses constitute novel contributions to the current state of FEXP modeling.

Another striking distinction between our FEXP models and most of the methods currently being implemented for estimation of the long memory parameter is that by using a Bayesian approach we gain more information about d, since we can obtain the whole posterior distribution rather than just a point estimate and standard error. This is something log periodogram regression and GPH (Geweke and Porter-Hudak 1983) do not achieve. This distinction is fundamental and is often one of the advantages of appealing to the Bayesian paradigm. One illustration of the benefits of having a posterior density of d arises in Section 5.3 where we analyze the Nile River minima data. In this setting, examining the posterior density of d (Figure 2), we notice that although we started from a non-informative prior on d we are able to extract the embedded information about d from the likelihood. That is, by taking advantage of the Bayesian machinery we end up with a highly informative posterior. Therefore having this posterior density definitely provides an added boost to the performance of our estimate of d. Similar results followed from the posterior density plot for the Hadley data (Figure 3). Finally, our approach accommodates missing data with relative ease; this is something current FEXP models do not achieve.

Additionally, using the exact Gaussian likelihood, we establish Bayesian consistency of the memory parameter under mild conditions on the data process. This poses a non-trivial extension to the proof presented in Liseo et al. (2001). Specifically, in the proof we provide, we allow the dimension of the parameter space to change as well as explicitly account for the mean.

In summary, we have developed a flexible modeling approach for Gaussian long range dependent processes. The algorithms we provide are computationally straight forward and relieve the practitioner from the burden of order selection when the main focus is estimation of the long memory parameter. Lastly, we develop asymptotic properties of our estimator under mild conditions on the data process.

References

Beran, J. (1993). "Fitting long-memory models by generalized linear regression." *Biometrika*, 80(4): 817–822. 160, 161

— (1994). *Statistics for Long Memory Processes*. New York: Chapman & Hall. 160, 163, 173, 177

Bertelli, S. and Caporin, M. (2002). "A Note on Calculating Autocovariances of Long-memory Processes." *Journal of Time Series Analysis*, 23(5): 503–508. 164

Bloomfield, P. (1973). "An exponential model for the spectrum of a scalar time series." *Biometrika*, 60: 217–226. 161

Bötcher, A. and Silberman, B. (1999). *Introduction to Large Truncated Toeplitz Matrices*. New York: Springer-Verlag. 165

Breidt, F. J. and Hsu, N.-J. (2002). "A Class of Nearly Long-memory Time Series Models." *International Journal of Forecasting*, 18(2): 265–281. 162

Chen, W., Hurvich, C. M., and Lu, Y. (2006). "On the Correlation Matrix of the Discrete Fourier Transform and the Fast Solution of Large Toeplitz Systems for Long-memory Time Series." *Journal of the American Statistical Association*, 101(474): 812–822. 161, 165, 174

Cheung, Y.-W. and Diebold, F. X. (1994). "On Maximum Likelihood Estimation of the Differencing Parameter of Fractionally-integrated Noise with Unknown Mean." *Journal of Econometrics*, 62: 301–316. 161

Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis-Hastings algorithm." *The American Statistician*, 49: 327–335. 168

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *Journal of the American Statistical Association*, 96: 270–281. 171

Dahlhaus, R. (1989). "Efficient Parameter Estimation for Self-similar Processes." *The Annals of Statistics*, 17: 1749–1766. 187, 188

Davies, R. B. and Harte, D. S. (1987). "Tests for Hurst Effect." *Biometrika*, 74: 95–101. 164, 173

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian methods for nonlinear classification and regression*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons Ltd. 169

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). “A Bayesian CART Algorithm.” *Biometrika*, 85: 363–377. 165, 169

Fox, R. and Taqqu, M. S. (1986). “Large-sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series.” *The Annals of Statistics*, 14: 517–532. 160

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference Second Ed.*. New York: Chapman & Hall/CRC. 168

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition. 164, 169, 171

Geweke, J. and Porter-Hudak, S. (1983). “The Estimation and Application of Long Memory Time Series Models.” *Journal of Time Series Analysis*, 4: 221–238. 160, 174, 182

Giraitis, L. and Surgailis, D. (1990). “A Central Limit Theorem for Quadratic Forms in Strongly Dependent Linear Variables and Its Application to Asymptotic Normality of Whittle’s Estimate.” *Probability Theory and Related Fields*, 86: 87–104. 160

Giraitis, L. and Taqqu, M. (1999). “Whittle Estimator for Finite-variance Non-Gaussian Time Series with Long Memory.” *The Annals of Statistics*, 27(1): 178–203. 160

Green, P. J. (1995). “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination.” *Biometrika*, 82: 711–732. 167, 169

Haslett, J. and Raftery, A. E. (1989). “Space-time Modelling with Long-memory Dependence: Assessing Ireland’s Wind Power Resource.” *Applied Statistics*, 38: 1–21. 176

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). “Bayesian Model Averaging: a Tutorial.” *Statistical Science*, 14(4): 382–417. 162, 182

Holan, S., McElroy, T., and Chakraborty, S. (2007). “A Bayesian approach to estimating the long memory parameter.” *SRD Research Report No. 2007/13, US Census Bureau*, http://www.census.gov/srd/papers/pdf/rrs2007–13.pdf. 162

Hosking, J. R. M. (1981). “Fractional Differencing.” *Biometrika*, 68: 165–176. 160

Hurvich, C. M. (2001). “Model Selection for Broadband Semiparametric Estimation of Long Memory in Time Series.” *Journal of Time Series Analysis*, 22(6): 679–709. 163

— (2002). “Multistep Forecasting of Long Memory Series Using Fractional Exponential Models.” *International Journal of Forecasting*, 18(2): 167–179. 161, 164, 174, 179, 182

Hurvich, C. M. and Brodsky, J. (2001). “Broadband Semiparametric Estimation of the Memory Parameter of a Long-memory Time Series Using Fractional Exponential Models.” *Journal of Time Series Analysis*, 22(2): 221–249. 176

Janacek, G. J. (1982). "Determining the Degree of Differencing for Time Series Via the Log Spectrum." *Journal of Time Series Analysis*, 3: 177–183. 160

Kass, R. and Raferty, A. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. 162

Ko, K. and Vannucci, M. (2006a). "Bayesian wavelet analysis of autoregressive fractionally integrated moving-average processes." *Journal of Statistical Planning and Inference*, 136: 3415–3434. 160, 178

— (2006b). "Bayesian wavelet-based methods for the detection of multiple changes of the long memory parameter." *IEEE Transactions on Signal Processing*, 54: 4461–4470. 162

Liseo, B., Marinucci, D., and Petrella, L. (2001). "Bayesian semiparametric inference on long-range dependence." *Biometrika*, 88(4): 1089–1104. 160, 161, 172, 182, 187

Manley, G. (1974). "Central England temperatures: monthly means 1659 to 1973." *Q. J. R. Meteorology Soc.*, 100: 389–405. 163, 173, 179

Moulines, E. and Soulier, P. (2000). "Data Driven Order Selection for Projection Estimator of the Spectral Density of Time Series with Long Range Dependence." *Journal of Time Series Analysis*, 21(2): 193–218. 176

Pai, J. S. and Ravishanker, N. (1998). "Bayesian Analysis of Autoregressive Fractionally Integrated Moving-average Processes." *Journal of Time Series Analysis*, 19: 99–112. 160, 163, 173, 179, 180

Palma, W. (2007). *Long-Memory Time Series*. New York: John Wiley and Sons. 160, 161, 171

Petris, G. (1997). "Bayesian Analysis of Long Memory Time Series." Ph.D. thesis, Duke University. 160

Pourahmadi, M. (1983). "Exact Factorization of the Spectral Density and Its Application to Forecasting and Time Series Analysis." *Communications in Statistics: Theory and Methods*, 12: 2085–2094. 164

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL http://www.R-project.org 174

Ravishanker, N. and Ray, B. K. (1997). "Bayesian Analysis of Vector ARFIMA Processes." *The Australian & New Zealand Journal of Statistics*, 39: 295–311. 160

Robert, C. (2001). *Bayesian Choice*. New York: Springer-Verlag. 189, 190

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. New York: Springer-Verlag, second edition. 167

Robinson, P. M. (1994). "Efficient Tests of Nonstationary Hypotheses." *Journal of the American Statistical Association*, 89: 1420–1437. 160

— (1995a). "Log-periodogram Regression of Time Series with Long Range Dependence." *The Annals of Statistics*, 23: 1048–1072. 160

— (1995b). "Gaussian Semiparametric Estimation of Long Range Dependence." *The Annals of Statistics*, 23: 1630–1661. 160

— (2003). *Time Series With Long Memory*. Oxford: Oxford University Press. 160

Rousseau, J. and Liseo, B. (2007). "Bayesian nonparametric estimation of the spectral density of a long memory Gaussian time series." http://www.citebase.org/abstract?id=oai:arXiv.org:0711.0876. 162, 186

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6: 461–464. 190

Sisson, S. A. (2005). "Transdimensional Markov Chains: A Decade of Progress and Future Perspectives." *Journal of the American Statistical Association*, 100(471): 1077–1089. 162

Tierney, L. and Kadane, J. B. (1986). "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association*, 81: 82–86. 189

Acknowledgments

The authors would like to thank an associate editor and two anonymous referees whose suggestions helped improve this paper.

Disclaimer This paper is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

Supplementary material

Appendix A - Asymptotics

Proof of Theorem 1. A more general result is given in Rousseau and Liseo (2007), but we provide our own direct proof here because the latter result does not include the case of unknown mean. Moreover, some things are simpler in our case, since the order of the EXP model is always bounded. The Rousseau and Liseo (2007) results are presented in a very general and abstract manner, making it difficult (also due to some typos and a terse style) to determine whether a particular model satisfies the conditions of their main theorem. For these reasons we adopt a direct approach, which uses the general strategy of Liseo et al. (2001), but replacing the Whittle likelihood with the exact Gaussian likelihood by using some concepts from Dahlhaus (1989).

Firstly, the class of spectral densities described by (4) and our priors is seen to satisfy all the conditions of Dahlhaus (1989) except the compactness of the parameter space, which actually will not be needed. We therefore are able to apply Lemmas 5.3 and 5.4, as well as results from the proof of Theorem 3.1 of that paper, among other results. Next, note that we have

$$\widehat{d}-\widetilde{d}=\frac{\sum_m \int_{\Phi_m\times\mathbb{R}^2}(d-\widetilde{d})p(x^{(n)}|\mu,\phi,d,m)\mathrm{d}P(\phi|m)\mathrm{d}P(d)\mathrm{d}P(\mu)p(m)}{\sum_m \int_{\Phi_m\times\mathbb{R}^2}p(x^{(n)}|\mu,\phi,d,m)\mathrm{d}P(\phi|m)\mathrm{d}P(d)\mathrm{d}P(\mu)p(m)}.$$

Now the likelihood only depends on ϕ through d and the b_k coefficients, given m, so we can substitute θ_m for (d,ϕ,m); in particular, $\mathrm{d}P(\phi|m)\mathrm{d}P(d)$ can be replaced by $\mathrm{d}P(\theta_m)$. Using (5) we can write

$$\begin{aligned} p(x^{(n)}|\theta_m,\mu) &= \exp\left[-\frac{n}{2}\left\{\log 2\pi+\mathcal{L}_n(\theta_m,\mu)\right\}\right] \\ \mathcal{L}_n(\theta_m,\mu) &= \frac{1}{n}\log|\Sigma(f_{\theta_m})|+\left(x^{(n)}-\mu\underline{1}\right)'\Sigma^{-1}(f_{\theta_m})\left(x^{(n)}-\mu\underline{1}\right). \end{aligned}$$

Now, the quadratic form in $\mathcal{L}_n$ can be written as

$$\begin{aligned} (x^{(n)}-\mu\underline{1})'\Sigma^{-1}(f_{\theta_m})(x^{(n)}-\mu\underline{1}) &= (x^{(n)}-\widetilde{\mu}\underline{1})'\Sigma^{-1}(f_{\theta_m})(x^{(n)}-\widetilde{\mu}\underline{1}) \\ &+ 2(\widetilde{\mu}-\mu)\underline{1}'\Sigma^{-1}(f_{\theta_m})(x^{(n)}-\widetilde{\mu}\underline{1}) \\ &+ (\widetilde{\mu}-\mu)^2\underline{1}'\Sigma^{-1}(f_{\theta_m})\underline{1}. \end{aligned}$$

The second term is $O_P(n^{(\delta+1-2d)/2})$ by Lemmas 5.3 and 5.4 of Dahlhaus (1989), for any $\delta>0$. This will yield a negligible term when divided by n; moreover, this is uniform over all parameters (including μ), since $d<1/2$ for all d (and any m). Likewise, the third term is $O(n^{\delta+1-2d})$. Hence we can exchange $\mathcal{L}_n(\theta_m,\widetilde{\mu})$ for $\mathcal{L}_n(\theta_m,\mu)$ at a cost of $O_P(n^{(\delta-1-2d)/2})$; denote these remainder terms by $R_n(\theta_m,\mu)$. Hence

$$\widehat{d}-\widetilde{d}=\frac{\sum_m \int_{\Theta_m\times\mathbb{R}}(d-\widetilde{d})\exp\left[-\frac{n}{2}\left\{\mathcal{L}_n(\theta_m,\widetilde{\mu})+R_n(\theta_m,\mu)\right\}\right]\mathrm{d}P(\theta_m)\mathrm{d}P(\mu)p(m)}{\sum_m \int_{\Theta_m\times\mathbb{R}}\exp\left[-\frac{n}{2}\left\{\mathcal{L}_n(\theta_m,\widetilde{\mu})+R_n(\theta_m,\mu)\right\}\right]\mathrm{d}P(\theta_m)\mathrm{d}P(\mu)p(m)}.$$

Now let $\epsilon>0$ and for each $m\geq 1$ let $\epsilon_m=\epsilon 2^{-m}$, and consider the set $\{(\theta_m,\mu)\in\Theta_m\times\mathbb{R}:\|(\theta_m,\mu)-(\widetilde{\theta}_m,\widetilde{\mu})\|<\epsilon_m\}$. This is an ϵ_m-neighborhood – with respect to the Euclidean norm $\|\cdot\|$ – of $(\widetilde{\theta}_m,\widetilde{\mu})$, which is the vector consisting of the first $m+3$ correct parameters (but if $m>\widetilde{m}$, pad out with zeroes). Denote this set by $N_{\epsilon_m}(\widetilde{\theta}_m,\widetilde{\mu})$; we split the numerator integral in the expression for $\widehat{d}-\widetilde{d}$ into an integration over $N_{\epsilon_m}(\widetilde{\theta}_m,\widetilde{\mu})$ and its complement. In the former term, we obtain the strict upper bound of $\sum_{m\geq 1}\epsilon_m=\epsilon$, using $|d-\widetilde{d}|\leq\|(\theta_m,\mu)-(\widetilde{\theta}_m,\widetilde{\mu})\|\leq\epsilon_m$. The latter term is bounded by

$$\frac{\sum_m \int_{N^c_{\epsilon_m}(\widetilde{\theta}_m,\widetilde{\mu})}(d-\widetilde{d})\exp\left[-\frac{n}{2}\left\{\mathcal{L}_n(\theta_m,\widetilde{\mu})+R_n(\theta_m,\mu)\right\}\right]\mathrm{d}P(\theta_m)\mathrm{d}P(\mu)p(m)}{\sum_m \int_{N_{\epsilon_m/2}(\widetilde{\theta}_m,\widetilde{\mu})}\exp\left[-\frac{n}{2}\left(\mathcal{L}_n(\theta_m,\widetilde{\mu})+R_n(\theta_m,\mu)\right\}\right]\mathrm{d}P(\theta_m)\mathrm{d}P(\mu)p(m)}. \qquad \text{(A.1)}$$

Here we have made the denominator smaller by restricting the region of integration. Now on a subset Ω_m of the full probability space Ω with $\mathbb{P}[\Omega_m] \to 1$, it follows from the proof of Theorem 3.1 of Dahlhaus (1989) that

$$\mathcal{L}_n(\theta_m, \widetilde{\mu}) \geq \mathcal{L}_n(\widetilde{\theta}_{\widetilde{m}}, \widetilde{\mu}) + \frac{1}{2} \inf_{\theta_m \in N^c_{\epsilon_m}(\widetilde{\theta}_m)} c_m(\theta_m)$$

with $c_m(\theta_m) \propto \min\{\int_{-\pi}^{\pi} (f_{\widetilde{\theta}_{\widetilde{m}}}/f_{\theta_m} - 1)^2 \mathrm{d}\lambda, \int_{-\pi}^{\pi} (f_{\theta_m}/f_{\widetilde{\theta}_{\widetilde{m}}} - 1)^2 \mathrm{d}\lambda\}$. Note that this bounding function does not depend on μ, which has already been handled through $R_n(\theta_m, \mu)$. Using the exponential structure of the model, it is easy to see that this function is convex in θ_m even when $m \neq \widetilde{m}$ (but if $m \neq \widetilde{m}$, the function cannot attain its minimal value of zero). It also follows from Theorem 3.1 of Dahlhaus (1989) that on other sets Ω'_m tending to one in probability

$$\mathcal{L}_n(\theta_m, \widetilde{\mu}) \leq \mathcal{L}_n(\widetilde{\theta}_{\widetilde{m}}, \widetilde{\mu}) + \frac{1}{4} \sup_{\theta_m \in N_{\epsilon_{m/2}}(\widetilde{\theta}_m)} c_m(\theta_m).$$

Now by the convexity of $c_m(\theta_m)$ and the choice of the neighborhood radii ϵ_m and $\epsilon_{m.2}$ we have

$$\frac{1}{4} \sup_{\theta_m \in N_{\epsilon_{m/2}}(\widetilde{\theta}_m)} c_m(\theta_m) < \frac{1}{2} \inf_{\theta_m \in N^c_{\epsilon_m}(\widetilde{\theta}_m)} c_m(\theta_m).$$

There are only finitely many nonzero m's, so we can find an ϵ^* such that the above inequality holds with ϵ^* in place of ϵ_m, with the inequality holding uniformly in m. Noting that the intersection of finitely many sets with probability tending to one also has this property, choose some $\delta < (\frac{1}{2} \inf_{\theta_m \in N^c_{\epsilon^*}(\widetilde{\theta}_m)} c_m(\theta_m) - \frac{1}{4} \sup_{\theta_m \in N_{\epsilon^*}(\widetilde{\theta}_m)} c_m(\theta_m))/2$ and we obtain a bound of

$$\mathbb{E}|d - \widetilde{d}| \frac{\sum_m \exp\{-\frac{n}{2} \left(\frac{1}{2} \inf_{\theta_m \in N^c_{\epsilon^*}(\widetilde{\theta}_m)} c_m(\theta_m) - \frac{1}{4} \sup_{\theta_m \in N_{\epsilon^*}(\widetilde{\theta}_m)} c_m(\theta_m) - 2\delta\right)\} p(m)}{\sum_m \int_{N_{\epsilon_{m/2}}(\widetilde{\theta}_m, \widetilde{\mu})} \mathrm{d}P(\theta_m) \mathrm{d}P(\mu) p(m)}$$

on a set tending to one in probability. As $n \to \infty$ this bound can be made as small as desired, so the convergence in probability follows. □

Appendix B - Bayes Factor and BIC

Here we provide justification, in our context, for the reversible jump algorithm as a method of model selection within each repetition of the chain. First let Q denote $BF(\Psi', \Psi^{(t)}) \times R$ where R and BF are defined similar to (13) and (14) respectively.

First we consider the numerator of (14):

$$\begin{aligned}\int_{\mathbb{R}^{2m'+4}} \mathrm{d}P(\Psi'|x^{(n)}) &= \int_{\mathbb{R}^{m'+3}} L(\Psi'|x^{(n)}) \times \frac{\prod_{i=0}^{m'} \Gamma(\alpha+1/2)}{\prod_{i=0}^{m'} \left(\frac{b_i'^2}{2}+\beta\right)^{\alpha+1/2}} \\ &\times \; (2\pi\sigma_\mu^2)^{-1/2} e^{-\frac{1}{2}\frac{(\mu'-\mu_0)^2}{\sigma_\mu^2}} \mathrm{d}\theta_{m'} \mathrm{d}\mu', \\ &= \int_{\mathbb{R}^{m'+3}} e^{nh(b',d',\mu')} \mathrm{d}\theta_{m'} \mathrm{d}\mu', \qquad \text{(B.1)}\end{aligned}$$

where

$$\begin{aligned}h(b',d',\mu') &= \frac{1}{n}\Big[\log\{L(\Psi'|x^{(n)})\} + (m'+1)\left[\log\{\Gamma(\alpha+1/2)\}\right] \\ &- \; (\alpha+1/2)\left\{\sum_{i=0}^{m'} \log\left(\frac{b_i'^2}{2}+\beta\right)\right\} - \frac{1}{2}\log(2\pi\sigma_\mu^2) - \frac{1}{2}\frac{(\mu'-\mu_0)^2}{\sigma_\mu^2}\Big],\end{aligned}$$

and $L(\cdot|x^{(n)})$ denotes the likelihood function. Applying the Laplace approximation Tierney and Kadane (1986); Robert (2001) in (B.1) we get

$$\begin{aligned}\int_{\mathbb{R}^{m'+3}} e^{nh(b',d',\mu')} \mathrm{d}\theta_{m'} \mathrm{d}\mu' &= e^{nh(\widehat{b}',\widehat{d}',\widehat{\mu}')}(2\pi)^{(m'+3)/2} n^{-(m'+3)/2} |H^{-1}(\widehat{b}',\widehat{d}',\widehat{\mu}')|^{1/2} + O(n^{-1}) \\ &= \frac{L(\widehat{b}',\widehat{d}',\widehat{\mu}'|x^{(n)})}{\prod_{i=0}^{m'}\left(\frac{\widehat{b}_i'^2}{2}+\beta\right)^{\alpha+1/2}} (2\pi)^{(m'+3)/2} n^{-(m'+3)/2} \\ &\times \; |H^{-1}(\widehat{b}',\widehat{d}'\widehat{\mu}')|^{1/2} + O(n^{-1}), \qquad \text{(B.2)}\end{aligned}$$

where $\widehat{b}'$, $\widehat{d}'$, $\widehat{\mu}'$ are the maximum likelihood estimates of b, d, μ and H is the Hessian matrix for the function $h(\cdot)$ under the model with $(m'+1)$ coefficients. Similarly from the denominator of (14), using similar logic, we get

$$\begin{aligned}\int_{\mathbb{R}^{m+3}} e^{nh(b^{(t)},d^{(t)}\mu^{(t)})} \mathrm{d}\theta_m^{(t)} \mathrm{d}\mu^{(t)} &= \frac{L(\widehat{b}^{(t)},\widehat{d}^{(t)}\widehat{\mu}^{(t)}|x^{(n)})}{\prod_{i=0}^{m}\left(\frac{\widehat{b}_i^{(t)2}}{2}+\beta\right)^{\alpha+1/2}} (2\pi)^{(m+3)/2} n^{-(m+3)/2} \\ &\times \; |H^{-1}(\widehat{b}^{(t)},\widehat{d}^{(t)},\widehat{\mu}^{(t)})|^{1/2} + O(n^{-1}), \qquad \text{(B.3)}\end{aligned}$$

where $\widehat{b}^{(t)}$, $\widehat{d}^{(t)}$, $\widehat{\mu}^{(t)}$ are the maximum likelihood estimates of b, d, μ and H is the Hessian matrix for the function $h(\cdot)$ under the model with $(m+1)$ coefficients.

Now replacing the numerator and denominator of (14) by (B.2) and (B.3) and then taking the logarithm of (14) yields the following expression

$$\begin{aligned} \log\{BF(\Psi', \Psi^{(t)})\} &= \left[\log\left\{L(\widehat{b}', \widehat{d}', \widehat{\mu}'|x^{(n)})\right\} - \log\left\{L(\widehat{b}^{(t)}, \widehat{d}^{(t)}, \widehat{\mu}^{(t)}|x^{(n)})\right\}\right] \\ &+ \frac{(m-m')}{2}\log(n) \\ &- (\alpha+1/2)\left\{\sum_{i+1}^{m'}\log\left(\frac{\widehat{b}_i'^2}{2}+\beta\right) - \sum_{i+1}^{m}\log\left(\frac{\widehat{b}_i^{(t)2}}{2}+\beta\right)\right\} \\ &+ R_n\{(\widehat{b}', \widehat{d}', \widehat{\mu}'), (\widehat{b}^{(t)}, \widehat{d}^{(t)}, \widehat{\mu}^{(t)})\}. \qquad \text{(B.4)} \end{aligned}$$

where $R_n\{(\widehat{b}', \widehat{d}', \widehat{\mu}'), (\widehat{b}^{(t)}, \widehat{d}^{(t)}, \widehat{\mu}^{(t)})\}$ is obtained by gathering all remainder terms. Now assuming that the remainder term is negligible (since the models are nested - Robert (2001, pg. 353)) and, since the coefficients of the FEXP model decay exponentially, the third term is also negligible (for m sufficiently large). Hence we get

$$\begin{aligned} \log\{BF(\Psi', \Psi^{(t)})\} &\approx \left[\log\left\{L(\widehat{b}', \widehat{d}', \widehat{\mu}'|x^{(n)})\right\} - \log\left\{L(\widehat{b}^{(t)}, \widehat{d}^{(t)}, \widehat{\mu}^{(t)}|x^{(n)})\right\}\right] \\ &+ \frac{(m-m')}{2}\log(n). \qquad \text{(B.5)} \end{aligned}$$

Now consider the Bayesian information criterion (BIC) (Schwarz 1978) of Ψ' and $\Psi^{(t)}$ respectively (note that according to our RJMCMC construction of the d and σ^2 parameters are fixed at the iteration so we are only dealing with the b),

$$\text{BIC}(\Psi') = -2\log\left\{L(\widehat{b}', \widehat{d}', \widehat{\mu}'|x^{(n)})\right\} + (m'+3)\log(n), \qquad \text{(B.6)}$$

and

$$\text{BIC}(\Psi^{(t)}) = -2\log\left\{L(\widehat{b}^{(t)}, \widehat{d}^{(t)}, \widehat{\mu}^{(t)}|x^{(n)})\right\} + (m+3)\log(n), \qquad \text{(B.7)}$$

Combining B.5, B.6, and B.7 we get

$$\text{BIC}(\Psi') - \text{BIC}(\Psi^{(t)}) \approx -2\log\left\{BF(\Psi', \Psi^{(t)})\right\}.$$

Therefore, in large samples, the BIC (or Schwarz criterion) provides a reasonably rough approximation to the logarithm of the Bayes factor which is embedded in the RJMCMC algorithm. Now, given a family of models, the BIC will select the true model with probability approaching to one as $n \to \infty$. Thus, our model selection approach based on Bayes factor in Section 3.2 provides a reasonable method of model selection.