Bayesian Time Series Modelling and Prediction with Long-Range Dependence

Giovanni Petris and Mike West
Carnegie Mellon University and Duke University

Summary We present a class of models for trend plus stationary component time series, in which the spectral densities of stationary components are represented via non-parametric smoothness priors combined with long-range dependence components. We discuss model fitting and computational issues underlying Bayesian inference under such models, and provide illustration in studies of a climatological time series. These models are of interest to address the questions of existence and extent of apparent long-range effects in time series arising in specific scientific applications.

Some key words: Bayesian time series analysis; Non-parametric models; Long memory; Spectral analysis.

1 INTRODUCTION

We develop a class of Bayesian time series models for data that may exhibit both structured trends and long range dependencies. Interest in long range dependence, or long memory effects in time series, has been high in recent years in connection with problems of assessing observed patterns in climatological series and the issues of global warming. We contribute to this discussion, presenting the first nonparametric Bayesian approach to studying time series models incorporating both deterministic trend and long memory components, and exploring analyses of a climatological series coming from a report of the Intergovernmental Panel on Climate Change. Our models for stationary components combine flexible, essentially nonparametric smoothness prior models for short memory stationary processes with a globally applicable long memory term. In part this extends previous models for long range effects, though our handling of the short memory components are different, and are closely related to the parallel Bayesian modelling developments in Carter and Kohn (1996). We present the model class and prior structures in Section 2, followed by brief details of posterior computation and analysis in Section 3 and in the appendix. This follows on the preliminary development briefly reported in Petris (1996). Section 4 proceeds to an application in climatology.

The development will use the following basic assumptions, notation and terminology. Consider a zero-mean, stationary process x_t observed over equally-spaced time intervals $t = 1, \dots, n$. The process is assumed to have a continuous spectral density function $f(\omega)$ over frequencies $0 < \omega < \pi$. A standard, short memory process has a spectral density that is bounded above as $\omega \to 0$. A long memory process of index d, $(0 < d < 1)$, behaves like $f(\omega) \sim \omega^{-d}$ as $\omega \to 0$, by comparison. For the given period of observations $t = 1, \dots, n$, and every $\omega \in (0, \pi)$, define the periodogram

$$I(\omega) = \frac{1}{2\pi n}\Big|\sum_{t=1}^{n} X_t e^{-it\omega}\Big|^2.$$

Under quite general assumptions (Künsch 1986, Petris 1996), for fixed $0 < \omega_1 < \dots < \omega_k$, the quantities $2I(\omega_j)/f(\omega_j)$ are asymptotically independent χ^2 variates (as n goes to infinity). If the spectral density is continuous on $[0, \pi]$, hence bounded, the convergence is uniform in ω. Therefore one can take the ω_j's to vary with n. Most of the standard inferential procedures in spectral analysis, resting on this observation, are based on the periodogram at the Fourier frequencies $2\pi j/n$, $j = 0, \dots, [n/2]$. Recent work of Robinson (1995) and others has been concerned with the breakdown of the standard asymptotic theory in the case of long range dependence. The basic issue is that the asymptotic distributional form and the asymptotic independence of the periodogram at the Fourier frequencies are theoretically questionable, especially for the very low Fourier frequencies. A traditional approach to handling this is to analyse the periodogram transforms of the data ignoring some of the low Fourier frequencies (Robinson, 1995). However, that Petris (1996, section 6) explored this issue via simulation studies and concluded that, in several models simulated, departures from the standard asymptotics are, for all practical purposes, quite negligible. This is clearly an area for further investigation.

We now introduce the model class and prior structures for data exhibiting trend and essentially arbitrary stationary time series behaviour, and then proceed to analysis and application.

2 TIME SERIES MODEL CLASS AND PRIOR STRUCTURE

Consider an equally-spaced time series z_t to be observed over $t = 1, \ldots, n$. We focus on models of the form $z_t = \mu_t + x_t$ where μ_t represents a parametrised trend term and x_t is a zero-mean, stationary process with spectral density function $f(\omega)$, continuous over frequencies $0 < \omega < \pi$. We assume that $\mu_t = h_t'\beta$ for some regression parameter vector β and known regression vectors h_t for each t. The linear trend case is two-dimensional with $h_t = (1, t)'$ for each t; this is the case of interest in our application in Section 4. The spectral component assumes an underlying smooth spectral density $g(\cdot)$ that may be modified near the origin by a long-memory component; specifically, we adopt $f(\omega) = \omega^{-d} \exp(g(\omega))$ where $d \in [0, 1)$ and $g(\omega)$ is continuous and bounded on $[0, \pi)$. Thus, if $d > 0$ the spectral density function has an explicit long-memory component, otherwise the continuous spectral density is that of an essentially arbitrary short-memory process. We now discretise the model, following previous authors (Geweke and Porter-Hudak 1983, Smith 19??, Petris 1996). However, in order to use the standard asymptotic results, instead of the Fourier frequencies, we consider the periodogram at a number s of equally spaced frequencies, with $s/n \to 0$.

Over the given period of observations $t = 1, \ldots, n$, write $m = n^\nu$, with $0 < \nu < 1$, $s = \lfloor m/2 \rfloor - 1$, $\omega_j = 2\pi j/m$ and define $y_j = \log(I(\omega_j))$. Then $y_j = \log(f(\omega_j)) + \epsilon_j$ where the ϵ_j are independent and distributed as $\log(\chi^2/2)$ under the asymptotic theory. This results in the regression model

$$y_j = -d \log(\omega_j) + g_j + \epsilon_j \tag{1}$$

where $g_j = g(\omega_j)$ for $j = 1, \ldots, s$. Given this representation, we work in the discretised framework. Based on data $z = \{z_1, \ldots, z_n\}$ with $z_t = h_t'\beta + x_t$, and assuming the representation (1), we are interested in inferences on β, d and the discrete version of $g(\omega)$ represented by $g = \{g_1, \ldots, g_s\}$. This extends standard spectral analysis to include the regression parameter β and the long memory parameter d. The framework is very similar to that in Smith (19??), but significantly different in the sense that our approach from here on is fully

Bayesian. To proceed, therefore, requires model completion in terms of prior distributions for $\{\beta, d, g\}$.

Our analyses are applicable under ranges of priors for β, and priors that are context dependent are encouraged. For the illustrative application in Section 4, we assume a locally uniform prior as a reference, noting that the resulting posteriors are essentially unchanged under diffuse multivariate normal priors. Conditionally normal priors are convenient assumptions from the viewpoint of posterior computations, as will be made clear in Section 3. We now detail the classes of priors for the spectral elements d and g.

2.1 Priors for d

Following Petris (1996), we assume a class of priors in which β, d and g are independent. We assume a normal prior $N(\beta|b_0, B_0)$ for β, with specified moments. For d, we explicitly recognise the interest in testing for long range dependence through priors with point masses at $d = 0$. Our class of priors for d is given by

$$\alpha\delta_0 + (1-\alpha)Be(d|a_d, b_d) \tag{2}$$

where δ_0 is the point mass at zero, α is the prior probability that $d = 0$, and $Be(\cdot|a_d, b_d)$ is the beta distribution on $(0, 1)$. Our analyses reported later are all based on the choice $\alpha = 0.5$, though other choices could be made.

2.2 Priors for g and associated hyperparameters

We use a specific class of smoothness priors for the discretised log spectral density $g(\cdot)$. Our development follows Petris (1996), and is similar to the parallel development in Carter and Kohn (1997), though based on different choices of smoothness priors. Specifically, we model g as a Gaussian autoregressive process of order p. For some hyperparameters (m, σ^2, γ), we assume the elements of g to be a finite set of observations from the the autoregression $(1 - \gamma B)^p g_t = w_t$ started at m. More explicitely,

$$g_t = \sum_{j=1}^{p} \binom{p}{j} (-\gamma)^j g_{t-j} + w_t$$

where the innovations w_t are conditionally independent $N(\cdot|0, \sigma^2)$,, $0 < \gamma < 1$ and $g_t = m$ for $t \leq 0$ We treat (m, γ, σ^2) as hyperparameters to be estimated. This prior implies, for values of γ close to one, that the g_j's are nearly equal to m for the first few j's, which forces the asymptotic behaviour of $\omega^{-d} g(\omega)$ at zero, in the long-memory case, to be picked by d. This is an example of a more general class of autoregressive smoothness priors, and our development could be followed with easy and direct modifications using other members of this class, as it could with alternative smoothness priors such as those of Carter and

Kohn (1997). The practical relevance of more complex models is very questionable in many applications from the viewpoint of parsimony and model complexity, however, and our experience with this relatively simple class support this view from an empirical viewpoint.

This model implies a multivariate normal prior for the full set of spectral elements $g = (g_1, \dots, g_s)$, given the hyperparameters (m, γ, σ^2). Namely, the joint conditional density of g is

$$p(g|m, \gamma, \sigma^2) = (2\pi\sigma^2)^{s/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{s} \left(g_j - \sum_{k=1}^{r} \binom{r}{k} (-\gamma)^k g_{j-k} \right)^2 \right), \qquad (3)$$

where $g_0 = g_{-1} = \dots = g_{-r+1} = m$. The variance of g_j, for j large, is approximately a multiple of σ^2: $\tau^2 = U_\gamma \sigma^2$, with

$$U_\gamma = \sum_{k=0}^{\infty} \left(\gamma^k \frac{r(r+1)\dots(r+k-1)}{k!} \right)^2.$$

More precisely, τ^2 is the variance of each g_j before conditioning on the starting values being equal to m.

The class of priors we use for the hyperparameters assumes m, γ and τ^2 to be independent, with a normal margin for m, an inverse gamma margin for τ^2, and a beta margin for γ. The choices made for specific applications are context dependent, and we describe our choices in the application later. Note that we could alteratively adopt standard reference priors for (m, τ).

3 MODEL FITTING AND POSTERIOR SAMPLING

The previous section outlines the joint prior for all model parameters and hyperparameters, denoted by $\phi = \{\beta, d, g, m, \gamma, \tau\}$. The prior independence assumptions imply the structure

$$p(\phi) = p(\beta)p(d)p(g|m, \gamma, \tau)p(m)p(\gamma)p(\tau).$$

Bayesian inference requires summary analysis of the joint posterior $p(\phi|z)$ based on observations $z = \{z_1, \dots, z_n\}$. We explore this by iterative simulation, sequencing through simulations of sets of conditional posteriors in a standard Markov chain Monte Carlo context. For any element or subset of elements θ of ϕ, denote by θ_- the remaining elements. Then analysis iteratively simulates the sequence $p(\beta|z, \beta_-)$, $p(d|z, d_-)$, $p(g|z, g_-)$, $p(m|z, m_-)$, $p(\gamma|z, \gamma_-)$ and $p(\tau|z, \tau_-)$. At each simulation, the required conditioning parameters are set are their most recently sampled values. Some details of these conditional distributions and methods of sampling are now provided.

3.1 Conditional posteriors for spectral elements

For a fixed value of β the time series model implies that $x_t = z_t - h_t'\beta$ is observed. Hence, conditional on β, we are in a context of observing the zero-mean stationary process so defined, the regression term being removed. This is essentially the context of Petris (1996), and details of the resulting conditional posteriors for elements d, g, m, γ and τ are given there. We briefly comment here.

We now have a regression model of the form (1) where the y_j are now the logged values of the ordinates of the periodiogram of the imputed data $x_t = z_t - h_t'\beta$, $(t = 1, \ldots, n)$. This is a linear regression in the elements of g which, together with the autoregressive prior model for g, may be couched in a state-space, or dynamic linear model representation as in Carter and Kohn (1996). (Note that in Petris (1996) a slightly different autoregressive prior for g is used, resulting in a different state-space representation.) The non-normality of the error terms ϵ_j in the regression is handled by direct approximation as a mixture of several fixed normal distributions (Carter and Kohn 1997, Petris 1996). For each $j = 1, \ldots, s$, multinomial indicator variables are introduced to select mixture components, and the model is augmented with these indicator variables to enable posterior simulation using the very efficient forward-filtering, backward sampling algorithm for mixtures of normal state-space models (Carter and Kohn, 1994). This provides conditional posterior samples for the full g vector at each iteration, rather than sampling individual elements one at a time, and so is conducive to rapid convergence of the Markov chain sampler.

The conditionals for m and τ are standard normal and inverse gamma distributions, respectively. The conditional for d, based on the prior in Section 2.1, is a mixture of a point mass at $d = 0$ with a continuous component on $0 < d < 1$. An ingenious device introduced in Petris (1996) and further developed in Petris and Tardella (1998) allows for direct sampling of this posterior that avoids an implict numerical integration over the continuous part of the conditional posterior. Note that repeat sampling of d values builds up a profile of the posterior distributions for d conditional on $d > 0$ as well as a Monte Carlo estimate of the posterior probability that $d = 0$. This feeds into direct inferences on the existence of long memory effects as well as its extent if indeed it exists.

Finally, the conditional posterior for the spectral smoothing parameter γ is simulated via a Metropolis-Hastings strategy. This is quite involved, requiring evaluation of the conditional likelihood function (3) as a function of γ, and the introduction of appropriate proposal distributions; see Petris (1996) for futher details of the construction used.

3.2 Conditional posteriors for regression parameters

Now consider the conditional posterior for β. Write the n model equations $z_t = h_t'\beta + x_t$ in the usual vector-matrix form $z = H\beta + x$ where H is now the design matrix whose rows are

h'_t, $(t = 1, \ldots, n)$, and $x = (x_1, \ldots, x_n)'$. Under the assumed Gaussianity of the x_t process, we have $x \sim N(x|0, V)$ where the $n \times n$ variance matrix V is to be determined. Hence we have a normal linear regression model for β, with variance matrix known when we fix the elements of V. A normal prior for β updates to a conditional normal posterior that is easily sampled. Alternatively, a locally uniform reference prior for β implies a conditional normal posterior $N(\beta|Bb, B)$ where $B = (H'V^{-1}H)^{-1}$ and $b = H'V^{-1}z$.

The issues of implementation here are those of efficiently computing the posterior moments b and B. Conditioning on current values of d and g implies a resulting discrete approximation to the spectral density function of x_t at the Fourier frequencies, namely $f_j = \omega_j^{-d} g_j$ for each $j = 1, \ldots, s$. The corresponding values of variances and covariances are then implied by Fourier transforms. Writing v_{ij} for the $i - j$ element of V we have

$$v_{i,i+k} = \int_{-\pi}^{\pi} e^{ik\omega} f(\omega)\, d\omega$$

for $k = 0, \pm 1, \ldots$. The above integrals can be approximated numerically by computing the fast Fourier transform of the discretized version of the spectral density f. Unfortunately, this is likely to produce unaccurate results, especially for large values of k and non-zero values of d, due to the pole of f and the rapidly oscillating behaviour of $e^{ik\omega}$. Furthermore, following this direct approach one would have then to invert the variance matrix V. From the point of view of computational efficiency, unless the observed series is unrealistically short, the inversion of an $n \times n$ matrix at each cycle of the MCMC simulation, would adversely effect the speed of computation, and is practically infeasible. Carter and Kohn (1997) proposed an alternative approach, based on Whittle approximation of the likelihood of a stationary process. We believe, however, that in their approach the posterior distribution is not invariant for the transition kernel of the Markov Chain. To sample from the regression parameter β we use a Metropolis-Hastings step.

4 AN APPLICATION

In this section we present an application to climatology of the methodology described above. The data, coming from a report of the Intergovernmental Panel on Climate Change, consists of centered and deseasonalised monthly values for the mean overall land and sea temperature (in centigrade degrees) over the Southern hemisphere during the period 1854-1989. A similar data set for the Northern hemisphere is analysed in Smith (1993). The series, of total length 1627, is plotted in Figure 1, together with a five-year moving average (solid line).

4.1 Short memory model.

First, to fix a baseline for comparisons with the our full model, we perform an analysis based on the submodel obtained by setting $\alpha = 1$ in (2). We assume that the mean of the process is a linear function of time: $\mu_t = \beta(2t/N - 1)$. For the g_j's an autoregressive prior of order $p = 4$ is assumed. We choose $\mu_m = 0$ and $s_m^2 = 4$. The other parameters of the prior distribution are $\gamma = 0.9$, $a_G = b_G = 10$. All the posterior estimates are based on 1500 iterations of the Markov chain. The chain seems to reach the stationary state very quickly, but we use nonetheless a burn in period of 500 iterations. We observed that the starting point has no practical influence on the inferences drawn from the Markov chain. In Figure 2 the value of the log posterior density is plotted against the iteration number. The plot is consistent with a stationary behaviour of the chain. Also the plots of the sampled values of each parameter are not in contrast with the assumed stationarity of the chain.

Figure 3 shows the histogram of the posterior distribution of the trend parameter. A summary of the distribution is given in table 1.

	Min.	1st Qu.	Median	Median	3rd Qu.	Max.
β	0.15	0.27	0.30	0.30	0.33	0.48

Table 1: Summary of the posterior of β, short memory model.

Histograms of the posterior distributions of the other parameters of the model are reported in Figures 4 and 5.

4.2 Long Memory model.

In this second part of our analysis of the temperature data, we modify our prior to allow for long memory in the process underlying the observations. To represent prior indifference about short versus long memory and about any particular value of the memory parameter, we specify the distribution of d as in (2), with $\alpha = 0.5$, $a_d = b_d = 1$. The specification of the other components of the prior is unchanged.

	Min.	1st Qu.	Median	Median	3rd Qu.	Max.
d	0.40	0.75	0.85	0.82	0.93	1

Table 2: Summary of the posterior of d.

	Min.	1st Qu.	Median	Median	3rd Qu.	Max.
β	0.16	0.26	0.30	0.30	0.34	0.47

Table 3: Summary of the posterior of β, long memory model.

Figure 6 provides the histogram of the posterior distribution of d. A summary of the posterior distribution of the memory parameter is reported in table 2. Note that, after the burn in period, the minimum value of d visited by the chain is 0.4. This implies that, within the accuracy provided by the Monte Carlo estimates, the posterior probability that the underlying process has short memory is zero.

With this more general model we can obtain an estimate of the spectral density, on a log scale, as the sum of the two components $-\hat{d}\log(\omega)+\hat{g}(\omega)$, where the hat simply denotes the average of the Monte Carlo samples. Figure 7 shows, on a logarithmic scale, the posterior estimate of the spectral density f (solid line). In a separate graph (Figure 8) the reader can find an estimate of the continuous component on the spectral density, g, on a log scale. Some samples from the chain are also provided to give a sense for the variability of the realizations of g.

As far as the trend is concerned, the posterior distribution is summarised by the histogram (Figure 9) and table 3.

Histograms of the posterior distributions of the other parameters of the model for this more general case are reported in Figures 10 and 11.

ACKNOWLEDGEMENT

The authors thank Richard L. Smith for discussions that initiated this research.

References

[1] Carter, C,. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

[2] Carter, C,. and Kohn, R. (1997). Semiparametric Bayesian inference for time series with mixed spectra. *Journal of the Royal Statistical Society, B* **59**, 255–268

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

[4] Künsch, H. (1986). Discrimination between monotonic trend and long-range dependence. *Journal of Applied Probability* **23**, 1025–1030.

[5] Petris, G. (1996). Bayesian spectral analysis of long memory time series, In *Proceedings of the 1996 Joint Statistical Meetings*, ASA Publications, Alexandria VA.

[6] Petris, G and Tardella, L. (1998). Simulating from mixture distributions. *Technical Report 681*, Department of Statistics, Carnegie Mellon University.

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

[8] Smith, R.L. (1993). Long-range dependence and global warming. *In* Barnett, V. and Turkman, F. (eds), *Statistics for the environment*, 141–161.

[9] Smith, R.L. (199?).

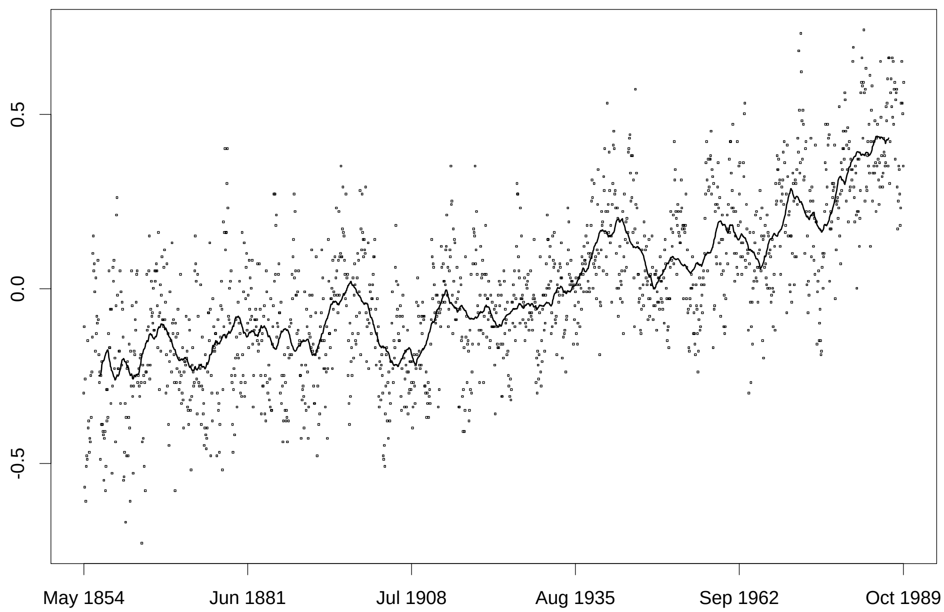

Figure 1: Temperature data.

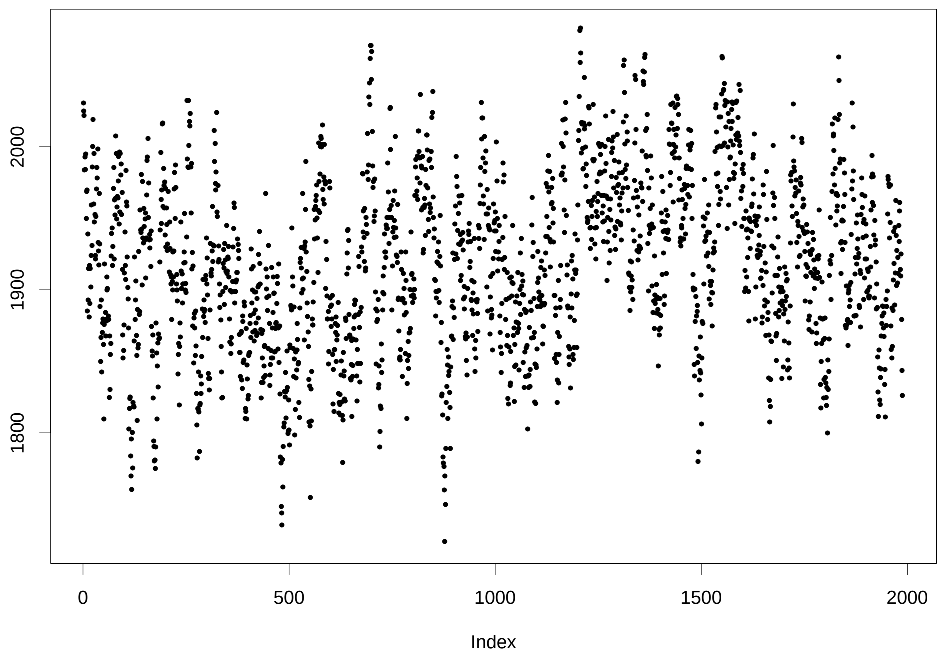

Figure 2: Posterior log density, short memory model.

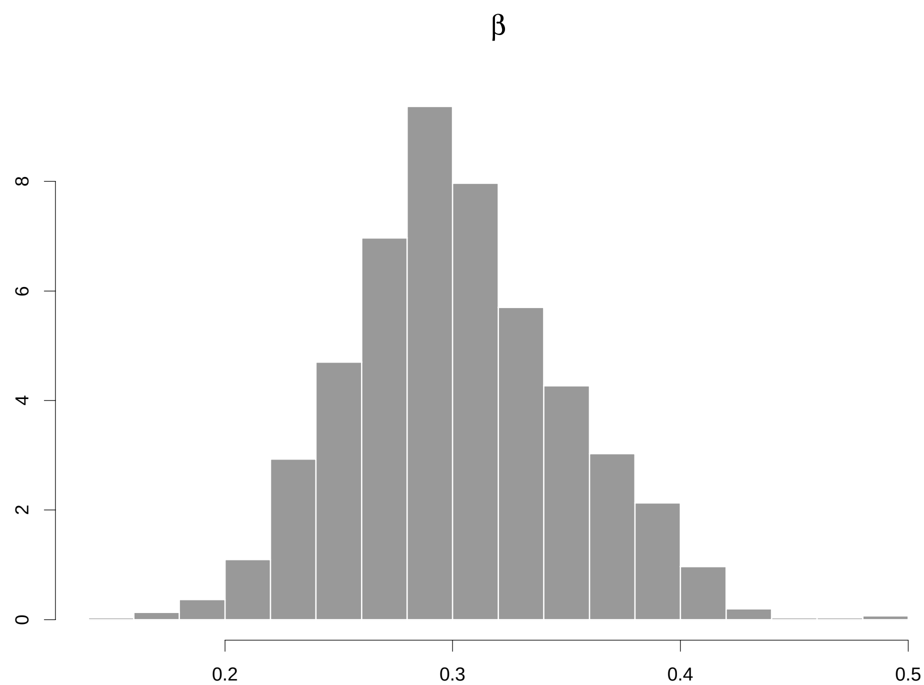

Figure 3: Posterior of β, short memory model.

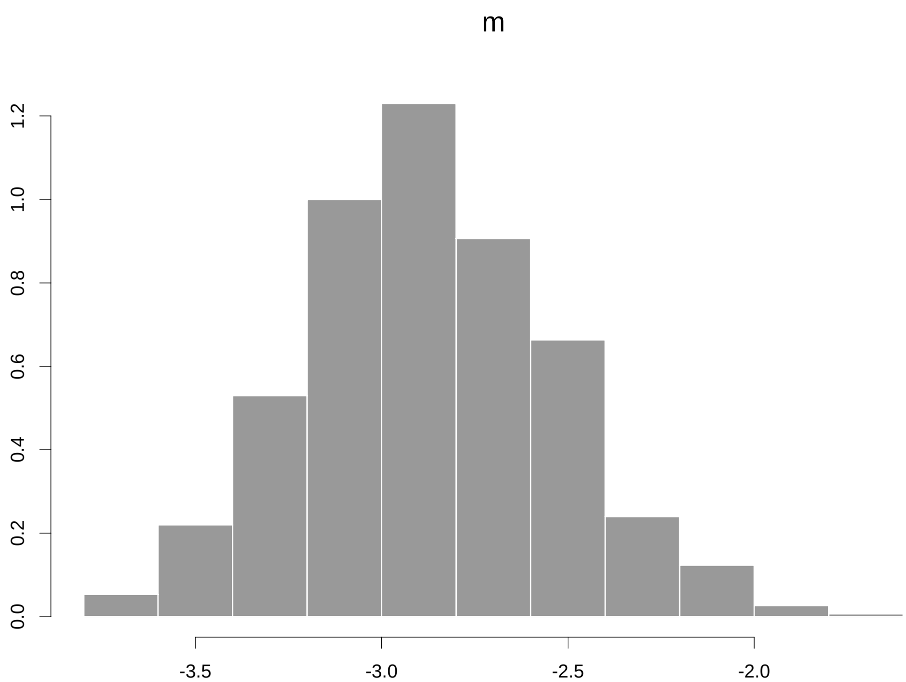

Figure 4: Posterior of m, short memory model.

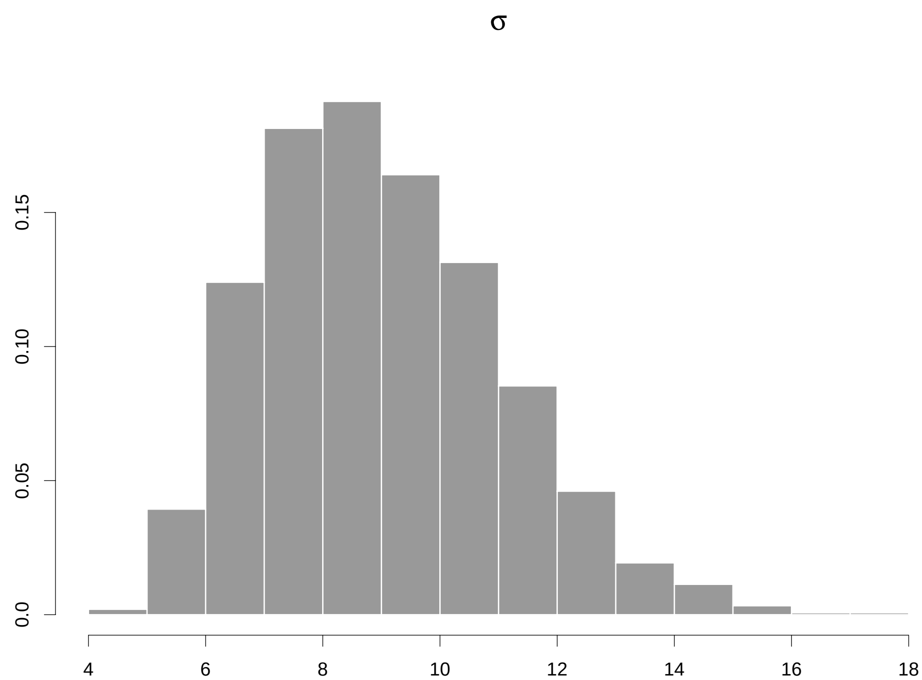

Figure 5: Posterior of σ, short memory model.

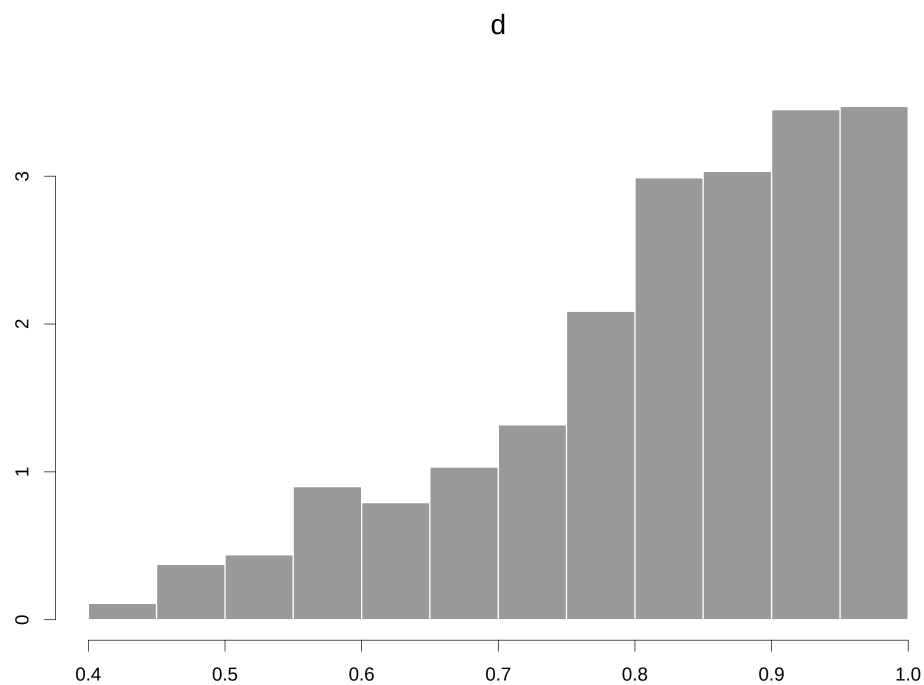

Figure 6: Posterior of d, long memory model.

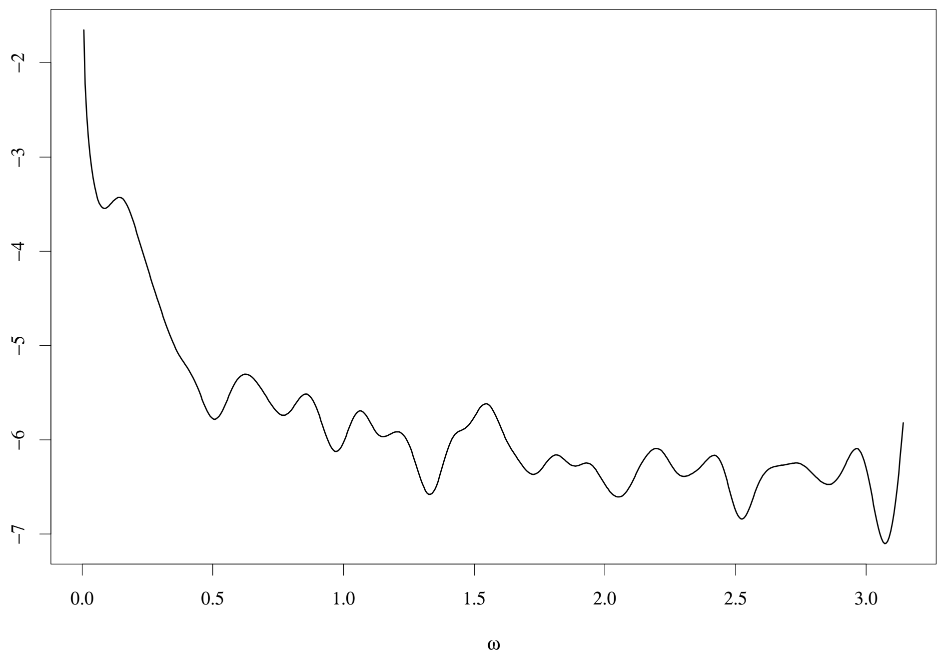

Figure 7: Estimate of $\log f$, long memory model.

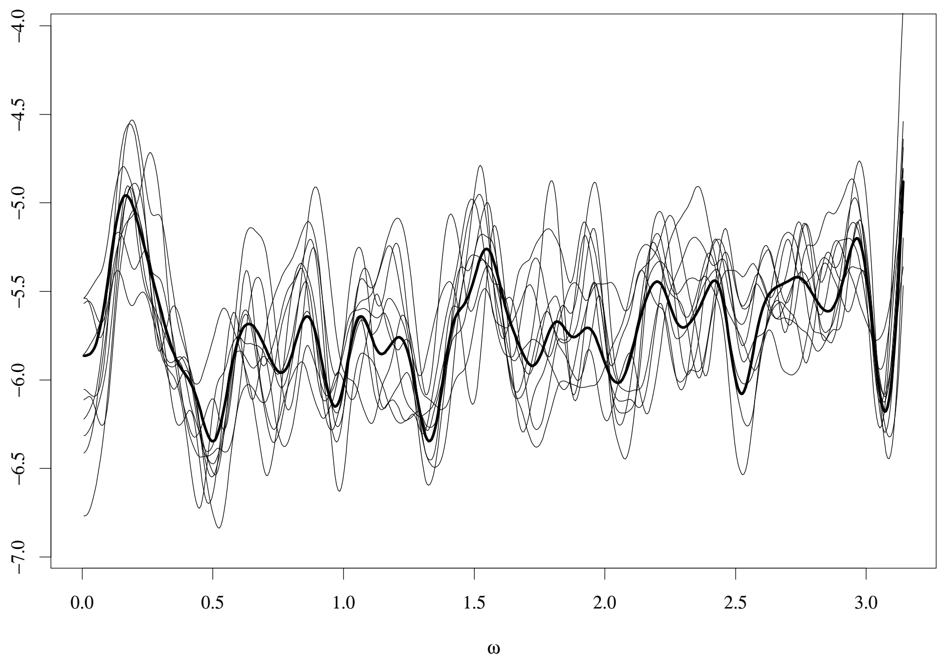

Figure 8: Estimate of g, long memory model.

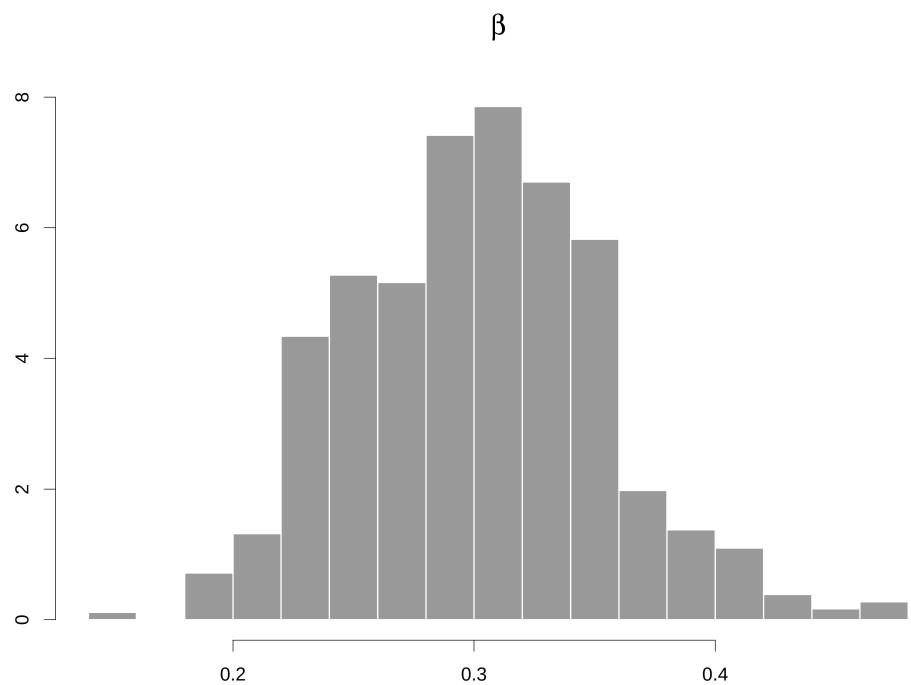

Figure 9: Posterior of β, long memory model.

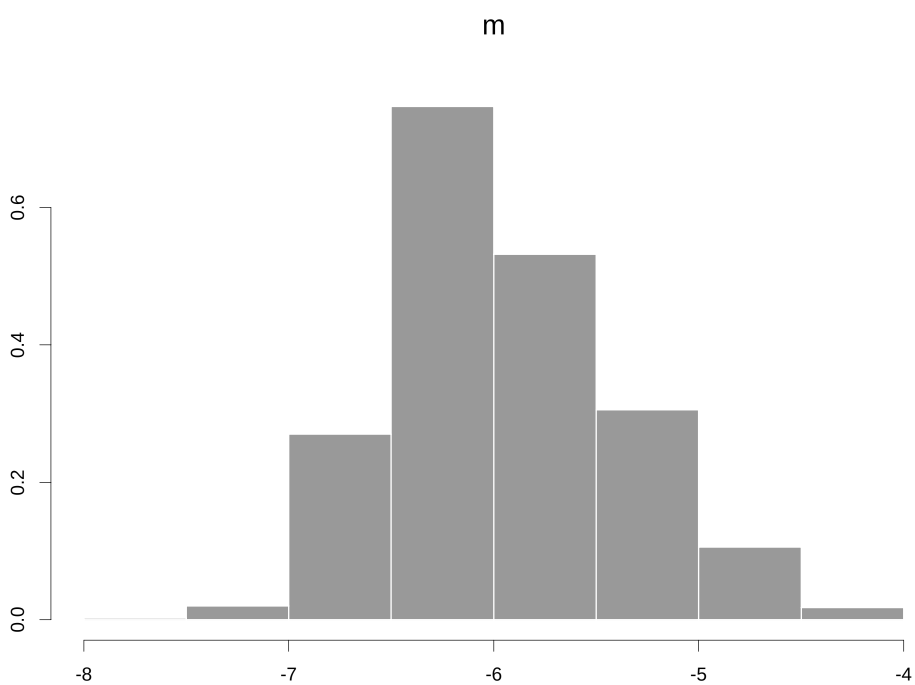

Figure 10: Posterior of m, long memory model.

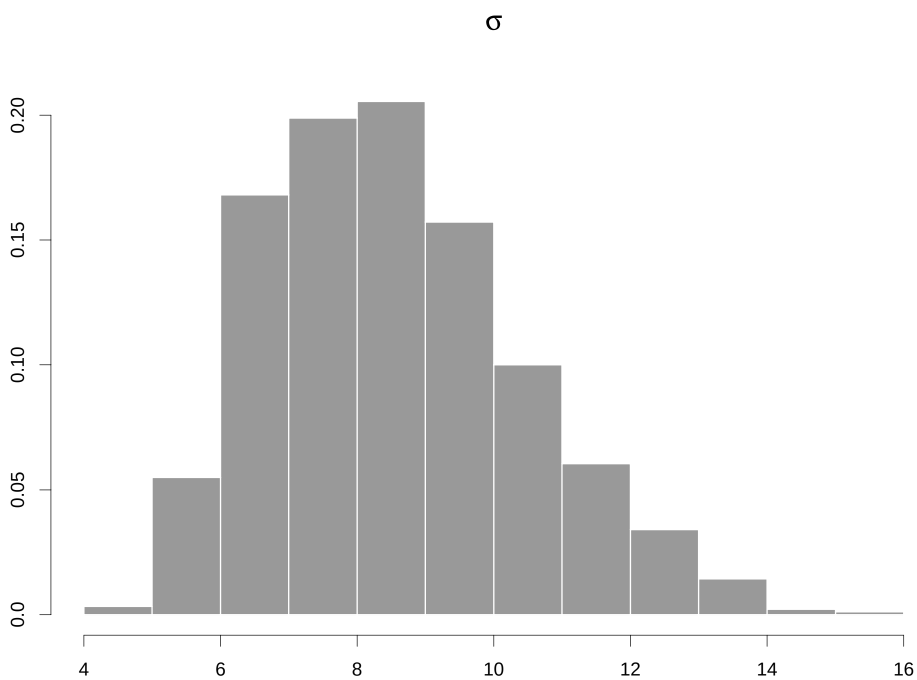

Figure 11: Posterior of σ, long memory model.