BAYESIAN ANALYSIS OF LONG MEMORY TIME SERIES

by

Giovanni Petris

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Supervisor

_____
Dr. Michael Lavine

_____
Dr. Peter Müller

_____
Dr. David Rios Insua

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

1997

ABSTRACT

(Statistics)

BAYESIAN ANALYSIS OF LONG MEMORY TIME SERIES

by

Giovanni Petris

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Supervisor

_____
Dr. Michael Lavine

_____
Dr. Peter Müller

_____
Dr. David Rios Insua

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

1997

Abstract

In recent years there has been a growing interest for the statistical analysis of long memory processes, i.e., stationary processes with a spectral density presenting a pole at the origin. In this dissertation, I will present a Bayesian nonparametric approach to the problem. I propose a class of prior distributions on a family of spectral densities which properly includes a set of densities having a pole at the zero frequency. This allows to test for the presence of a long memory behaviour and to compare the prior and posterior probability of the data coming from a long memory process, in order to see how strongly the data support this hypothesis. The basic model is then extended to include a linear regression term for the mean of the process.

Using Markov Chain Monte Carlo techniques, samples of the spectral density and the regression parameters from the posterior distribution can be obtained. With some modifications, the simulation scheme can be used to obtain also a sample from the predictive distribution of the future observations.

The methodology described is applied to the analysis of the average temperature of the Southern emisphere over the last century.

Acknowledgements

I would like to thank my advisor, Professor Mike West, who introduced me to the field of time series analysis. I owe to him some very valuable suggestions which are now part of this thesis. I would like to thank Professor Richard L. Smith for providing me with the data analysed in Chapter 5. I also wish to express my appreciation to the faculty and fellow graduate students in the Institute of Statistics and Decision Sciences for providing a stimulating and cooperative environment of research and study. Last, but not least, I thank Valeria for her constant support and understanding, without which this thesis could not have been written.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and preliminary results

1.1 Introduction.

Long-memory time series models are used to formalise the notion of strongly dependent series of observations. It has been known for a while that in several areas of application the correlations between observations taken at different times or locations may decay very slowly as a function of the time or distance which separate the observations. A consequence is that, for instance, the standard deviation of the sample mean decays with the number of data points n at a rate slower than $n^{-\frac{1}{2}}$. Smith (1938) reports this kind of behaviour in agricultural data. This is only one consequence of the intuitive fact that the amount of information carried by a data set of strongly correlated observations is much less than that contained in a set of independent or weakly dependent observations. Another early reference on long-memory is the work of the hydrologist Hurst (Hurst, 1951), where he empirically observed that, if x_i is the inflow to a reservoir at time i and $y_j = \sum_{i=1}^{j} x_i$, then the

rescaled adjusted range, or R/S-statistic, defined by

$$R/S = \frac{\max\limits_{0\le i\le k}\left(y_i - y_0 - \dfrac{i}{k}(y_k - y_0)\right) - \min\limits_{0\le i\le k}\left(y_i - y_0 - \dfrac{i}{k}(y_k - y_0)\right)}{\sqrt{\dfrac{1}{k}\displaystyle\sum_{i=1}^{k}(x_i - \overline{x})^2}}, \tag{1.1}$$

behaves, for large values of k approximately as a constant times k^H for some $H > \frac{1}{2}$. This behaviour, found also in many geophysical and climatological data sets, and now known as *Hurst effect*, is in contradiction to results holding for short memory processes. Mandelbrot and his co-workers (Mandelbrot and van Ness, 1968; Mandelbrot and Wallis, 1968; Mandelbrot and Wallis, 1969) introduced the so-called fractional Gaussian noise (see Section 1.3) and were able to show that the Hurst effect can be explained using that model. The use of standard short memory models, such as the ARMA family, to analyse strongly correlated data generally results in underestimating the uncertainty inherent to the results of the analysis. The fact that the variance of the sample mean is of order n^{d-1} for some $d \in (0, 1)$ is a good example of this behaviour. A more subtle way in which long-memory processes can be deceiving is in suggesting different kinds of nonstationarity. Let us illustrate this point with an example. Figure 1.1 shows a plot of a simulated long-memory time series of length 1000. The underlying process is stationary, with zero mean and unit variance. There is a clear decreasing trend in the first part of the data, at least for the observations 100–500. It is very likely that with only that segment of the data at hand, the ill-advised practitioner would estimate very precisely a nonzero linear trend. One may also see in the data a low frequency periodicity. All these type of nonstationarity are purely illusory. Even more basically, the overall level of the process does not seem to be zero (the sample mean is 2.14). Of course, knowing that the underlying process has long memory would not allow one to make an estimate of the mean closer to the true value, unless some prior information is used. The

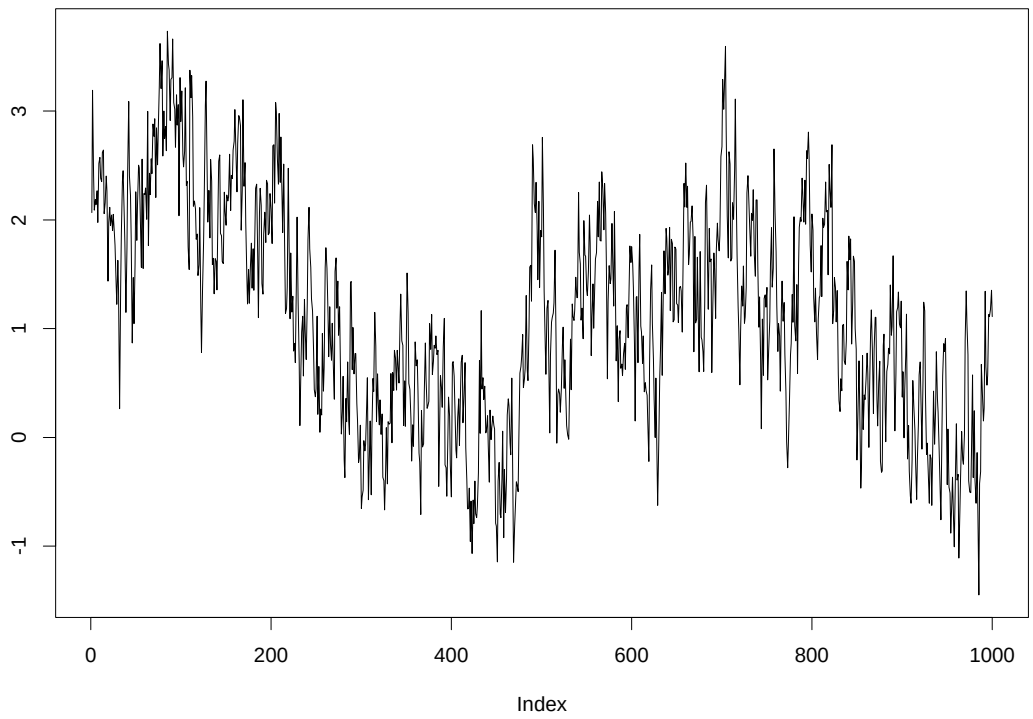

Figure 1.1: Simulated long-memory time series

difference would be more in the uncertainty about the estimate, which, taking into account the long-memory character of the process, would be larger.

Beran (1994a) is a good introduction to the Statistics of long-memory processes. From a more probabilistic point of view, a bibliography, up to 1985, is contained in Taqqu (1986). Recently the Journal of Econometrics has published a special issue entirely devoted to fractional differencing and long-memory processes (Baillie and King (1996)).

1.2 Mathematical preliminaries and notations.

In this Section we recall some well known facts about a second-order stationary discrete time stochastic processes and we set some of the notations we will be using throughout the paper.

Let $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ be a sequence of real random variables defined on a common probability space which, without real loss of generality, can be taken to be the canonical one. Assume that they all possess a finite second moment. $\mathbf{X}$ is said to be *second-order stationary* (or *stationary in the wide sense*) if the functions

$$\mathbb{E}(X_s) = M \tag{1.2}$$

$$\mathrm{Cov}(X_{s+t}, X_s) = R(t) \tag{1.3}$$

do not depend on s. The function R is called the *covariance function* of the process. For every covariance function R there exists a unique positive bounded measure μ on the Borel sets of the torus $\mathbb{T} = [0, 2\pi)$ such that

$$R(t) = \int_{\mathbb{T}} \cos(t\lambda)\, \mu(d\lambda) \tag{1.4}$$

for every integer t (see for example Doob (1953)). The measure μ is called the *spectral measure* of the process. In case it is absolutely continuous with respect to Lebesgue measure, its density f is called the *spectral density* of the process, and clearly we have

$$R(t) = \int_{\mathbb{T}} \cos(t\lambda) f(\lambda)\, d\lambda\,. \tag{1.5}$$

Let $(\alpha_k)_{k\in\mathbb{Z}}$ be a sequence of real numbers such that $\sum |\alpha_k|^2 < \infty$, and $(\epsilon_k)_{k\in\mathbb{Z}}$ a stationary sequence of uncorrelated random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. Than the series

$$Y_n = \sum_{k\in\mathbb{Z}} \alpha_{n-k}\epsilon_k, \qquad n \in \mathbb{Z} \tag{1.6}$$

converges in $\mathrm{L}^2(\Omega, \mathcal{F}, P)$ for every integer n, defining a stationary process possessing a spectral density. It turns out that the converse is also true: any stationary process possessing a spectral density is of the form (1.6) for some $(\alpha_k)_{k\in\mathbb{Z}}$ and $(\epsilon_k)_{k\in\mathbb{Z}}$. In order to obtain such a representation it may be necessary to enlarge the underlying probability space, see Doob (1953).

Given a sample of size N from a stationary process $\mathbf{X}$, the *periodogram* based on $(X_1, \ldots, X_N)$ is defined by

$$I_N(\lambda) = \frac{1}{N\pi} \left| \sum_{j=1}^{N} X_j e^{-ij\lambda} \right|^2, \qquad \lambda \in \mathbb{T}. \tag{1.7}$$

Assume that $\mathbf{X}$ has a spectral density f which is continuous on a closed subinterval $\mathcal{I}$ of $\mathbb{T}$ and a representation of X_n of the form (1.6) exists which satisfies the following conditions:

1. the ϵ_k's are independent,

2. (ϵ_k) obey the Central Limit Theorem, i.e.

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} \epsilon_n \xrightarrow{d} \mathcal{N}(0,1)\,, \tag{1.8}$$

3. (ϵ_k) have a spectral density continuous on $\mathcal{I}$.

(Given that f is continuous on $\mathcal{I}$, a representation satisfying 1.-3. always exists if $\mathbf{X}$ is a Gaussian process.) Then for any finite subset $\{\lambda_1, \ldots, \lambda_k\}$ of $\mathcal{I}$, the distribution of $(I_N(\lambda_1), \ldots, I_N(\lambda_k))$ tends, as N goes to infinity, to that of k independent random variables, distributed as $\frac{1}{2} f(\lambda_j)\chi_2^2$ if $\lambda_j \neq 0, \pi$, and as $f(\lambda_j)\chi_1^2$ otherwise $(j = 1, ..., k)$. For a proof, see Olshen (1967). To this result is related the so called Whittle approximation to the likelihood of a stationary process, which states that, for large N, the likelihood is proportional to

$$\prod_{j=1}^{N} f(\omega_j)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{N} \frac{I_N(\omega_j)}{f(\omega_j)} \right\}, \tag{1.9}$$

where $\omega_j = 2\pi j/N$, $j = 1, \ldots, N$ are the so-called Fourier frequencies. Also, under the same asymptotic theory, the joint density of $I_N(\omega_1), \ldots, I_N(\omega_{N/2})$ is approximately of the form (1.9). Note that from the definition of the periodogram it follows that $I_N(\omega) = I_N(2\pi - \omega)$, therefore (1.9) can be rewritten in terms of the periodogram ordinates at the first $N/2$ Fourier frequencies only. For details on Whittle approximation the reader is referred to Whittle (1953) and, in the context of long-memory processes, Beran (1994b).

1.3 Long-memory processes: definition and examples.

Let $\mathbf{X}$ be a stationary process with spectral density f. Assume that, for some number d in the interval (0,1), $f(\omega)$ is asymptotically equivalent to ω^{-d} times a constant, as ω approaches zero, that is to say:

$$\lim_{\omega \to 0} \frac{f(\omega)}{c\omega^{-d}} = 1, \qquad \text{for some } c \in \mathbb{R}. \tag{1.10}$$

Then $\mathbf{X}$ is said to be a *long-memory process*, with *long-memory parameter* d. Condition (1.10) can be restated in terms of the covariance function R of $\mathbf{X}$. In fact, (1.10) is equivalent of requiring that

$$\lim_{n \to \infty} \frac{R(n)}{cn^{d-1}} = 1, \qquad \text{for some } c \in \mathbb{R}. \tag{1.11}$$

The proof of the equivalence can be deduced from standard results in the theory of Fourier series, see for example Zygmund (1959). Another parameter which is often used in the place of d is the so-called Hurst parameter H, which, for a long-memory process, ranges from $\frac{1}{2}$ to 1. The relationship between H and d is given by the formula $d = 2H - 1$. Let us point out that a slightly more general definition of long-memory process can be given by substituting the constant c in (1.10) with a slowly varying

function of ω. A slowly varying function of ω, as ω goes to zero, is a function L such that

$$\lim_{\omega \to 0} \frac{L(t\omega)}{L(\omega)} = 1, \qquad \text{for every } t > 0. \tag{1.12}$$

Thus, the condition expressed by (1.10) becomes:

$$\lim_{\omega \to 0} \frac{f(\omega)}{L(\omega)\omega^{-d}} = 1, \qquad \text{for some slowly varying function } L, \tag{1.13}$$

see Beran (1992). We will not make use of this generalisation.

Two classic examples of long-memory processes are the fractionally differenced white noise and the fractional Gaussian noise. Let us start with the latter. The idea is to consider the increments of a continuous-time self-similar stochastic process. Let $\mathbf{Y} = (Y_t)$ be a continuous-time stochastic process with stationary increments, H-self-similar ($0 < H < 1$) in the sense that

$$\left(Y_{ct_1}, \ldots, Y_{ct_n}\right) \stackrel{\ell}{=} c^H \left(Y_{t_1}, \ldots, Y_{t_n}\right), \tag{1.14}$$

for every $t_1, \ldots, t_n$, n and $c > 0$, where $\stackrel{\ell}{=}$ denotes equality in law. Then, for any s, t:

$$\mathrm{Var}(Y_t - Y_s) = \mathrm{Var}(Y_{t-s} - Y_0) = \mathrm{Var}(Y_1)\,|t-s|^{2H}. \tag{1.15}$$

On the other hand

$$\mathrm{Var}(Y_t - Y_s) = \mathrm{Var}(Y_t) + \mathrm{Var}(Y_s) - 2\mathrm{Cov}(Y_t, Y_s), \tag{1.16}$$

and therefore the covariance function of $\mathbf{Y}$ is given by:

$$\mathrm{Cov}(Y_t, Y_s) = \frac{1}{2}\mathrm{Var}(Y_1)\,\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right). \tag{1.17}$$

It can be proved that a process $\mathbf{Y}$ with the properties of being H-self-similar with stationary increments exists. In particular, since its covariance function is determined up to a multiplicative factor by (1.17), there is, up to a scaling factor, only one Gaussian H-self-similar process with stationary increments, which is commonly called

fractional Brownian motion. Moreover, this process is centered and it assumes the value zero at time zero. For the proofs and a more detailed description of self-similar processes as well as their relations with long-memory processes, the reader is referred to Samorodnitsky and Taqqu (1994, Ch. 7). Fractional Brownian motion was introduced by Kolmogorov (1940), but the first statistical use of this model is due to Mandelbrot and his co-workers in the analysis of the Hurst effect. Let $\mathbf{Y}$ be a fractional Brownian motion, with self-similarity index H, and consider the sequence of the increments:

$$X_t = Y_t - Y_{t-1}, \qquad t \in \mathbb{Z}. \tag{1.18}$$

The discrete time process $\mathbf{X} = (X_t)$ is called *fractional Gaussian noise.* It is Gaussian, centered, stationary, with covariance function

$$R(t) = \frac{\mathrm{Var}(Y_1)}{2}\left(|t-1|^{2H} - 2|t|^{2H} + |t+1|^{2H}\right), \qquad t \in \mathbb{Z}, \tag{1.19}$$

and spectral density

$$f(\omega) = \mathrm{Var}(Y_1)\frac{2H\Gamma(2H)\sin(H\pi)}{\pi}\left|e^{i\omega}-1\right|^2 \sum_{k\in\mathbb{Z}} |\omega + 2k\pi|^{-2H-1}, \qquad \omega \in \mathbb{T}. \tag{1.20}$$

A fractional Brownian motion with self-similarity parameter $H = 1/2$ is just an ordinary Brownian motion, and therefore its increments are independent, i.e. the corresponding fractional Gaussian noise is just Gaussian white noise. For $1/2 < H < 1$, on the other hand, fractional Gaussian noise is a long-memory process: in fact,

$$R(t) \sim \mathrm{Var}(Y_1)H(2H-1)t^{2H-2}, \qquad \text{as } t \to \infty \tag{1.21}$$

and

$$f(\omega) \sim \mathrm{Var}(Y_1)\frac{2H\Gamma(2H)\sin(H\pi)}{\pi}|\omega|^{1-2H}, \qquad \text{as } \omega \to 0. \tag{1.22}$$

Comparing these expressions with (1.10) and (1.11) it is easy to see that, for $1/2 < H < 1$, fractional Gaussian noise satisfies the definition of a long-memory

process with long-memory parameter $d = 2H - 1$. A modification of fractional Gaussian noise has been used in econometric applications by Carlin, Dempster and Jonas (1985) and Carlin and Dempster (1989).

Let us describe now the other example of long-memory process that we mentioned, i.e. fractionally differenced white noise. The standard references here are Granger and Joyeux (1980) and Hosking (1981). Consider a stationary process $\mathbf{X} = (X_t)$ defined on $(\Omega, \mathcal{F}, P)$, and let $\mathbb{H}_X$ be the closure in $L^2(\Omega, \mathcal{F}, P)$ of the linear envelope of $\mathbf{X}$. One can then introduce the backward shift operator of $\mathbf{X}$, denoted by B, defined on $\mathbb{H}_X$ by $BX_t = X_{t-1}$, $t \in \mathbb{Z}$. Let ∇ denote the difference operator, i.e. $\nabla = 1 - B$. The *difference operator of order* α can be formally defined by:

$$\nabla^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} (-1)^k B^k. \tag{1.23}$$

Let

$$h(\omega) = \sum_{k=0}^{\infty} \binom{\alpha}{k} (-1)^k e^{-ik\omega} \tag{1.24}$$

$$= \left(1 - e^{-i\omega}\right)^\alpha. \tag{1.25}$$

From spectral considerations it follows that ∇^α is well-defined if and only if $h \in L^2([0, \pi], \mu_X)$, where μ_X is the spectral measure of $\mathbf{X}$, and in this case the spectral measure of the process $(\nabla^\alpha X_t)$ is given by $|h|^2 d\mu_X$. The process $\mathbf{X}$ is called a *fractionally differenced white noise* with parameter $H = 1/2 - \alpha$ if, for some white noise $\mathbf{Y}$, it satisfy the equations:

$$X_t = \nabla^{-\alpha} Y_t \qquad t \in \mathbb{Z}. \tag{1.26}$$

In other words, $\mathbf{X}$ is the $(H - 1/2)$-th fractional difference of the white noise $\mathbf{Y}$. For $1/2 < H < 1$, fractionally differenced white noise is a long-memory process, since for its covariance function and spectral density the following asymptotic relations hold:

$$R(t) = \frac{(-1)^t \Gamma(1 - 2\alpha)}{\Gamma(1 + t - \alpha)\Gamma(1 - t - \alpha)} \sim t^{2\alpha - 1}, \qquad \text{as } t \to \infty, \tag{1.27}$$

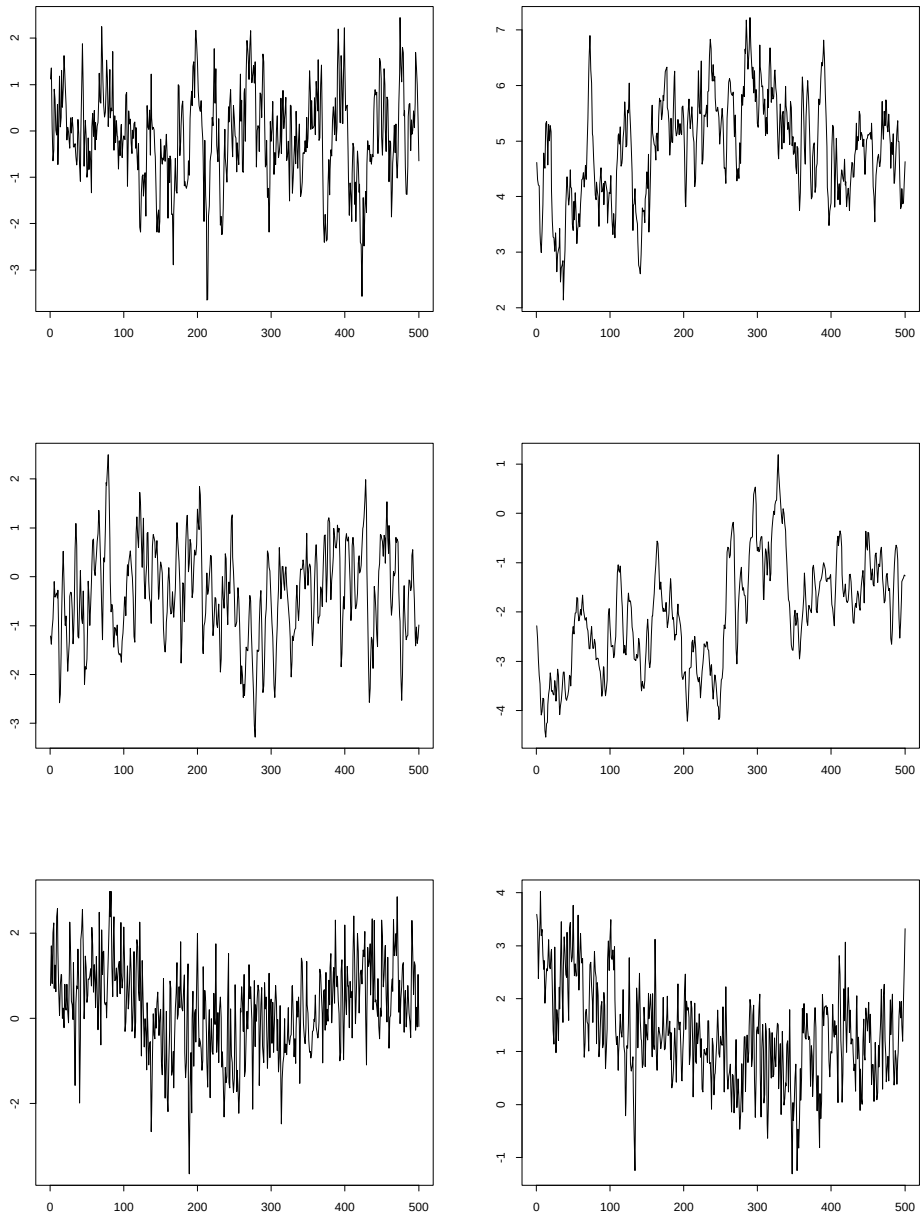

Figure 1.2: Simulated ARFIMA processes.

$$f(\omega) = \left(2 \sin \frac{\omega}{2}\right)^{-2\alpha} \sim \omega^{-2\alpha}, \qquad \text{as } \omega \to 0. \tag{1.28}$$

From a different point of view, (1.26) can be equivalently written as

$$\nabla^\alpha X_t = Y_t \qquad t \in \mathbb{Z}, \tag{1.29}$$

which state that the α-th fractional difference of $\mathbf{X}$ is a white noise, or $\mathbf{X}$ is a fractionally integrated white noise. In the terminology of Box and Jenkins (1970), $\mathbf{X}$ would be termed an ARIMA$(0, \alpha, 0)$ process. From this point of view, a natural generalisation is given by the class of ARIMA(p, α, q) processes. If ϕ and ψ are polynomials of order p and q, respectively, and $\mathbf{X}$ satisfies the relation:

$$\phi(B)\nabla^\alpha X_t = \psi(B) Y_t \qquad t \in \mathbb{Z}, \tag{1.30}$$

then $\mathbf{X}$ is called a *fractional* ARIMA(p, α, q) processes (see Beran (1994b)). Fractional ARIMA models, also known as ARFIMA, have become quite popular among practitioners, to the point that the statistical software package S-plus include several routines to simulate and analyse this kind of models. The function `arima.fracdiff.sim` has been used to generate the artificial realisations from different fractional ARIMA(p, α, q) processes plotted in Figure 1.2. The model equations, from top to bottom and left to right, are:

$$\begin{aligned}
(1 - 0.5B)\nabla^{0.2} X_t &= Y_t, \\
(1 - 0.5B)\nabla^{0.45} X_t &= Y_t, \\
(1 - 0.6B)\nabla^{0.2} X_t &= (1 - 0.9B + 0.08B) Y_t, \\
(1 - 0.6B)\nabla^{0.45} X_t &= (1 - 0.9B + 0.08B) Y_t, \\
\nabla^{0.3} X_t &= (1 - B + 0.6B) Y_t, \\
\nabla^{0.45} X_t &= (1 - B + 0.6B) Y_t.
\end{aligned}$$

All the processes are standardised to have mean zero and unit variance. It can be seen from the examples in Figure 1.2 that the ARFIMA family provide a quite flexible class

of models, with the long-term behaviour accounted for by the fractional differencing parameter α, while $\phi(B)$ and $\psi(B)$ provide an explanation of short-term properties.

1.4 Some recent asymptotic results.

Starting with the pioneering work of Geweke and Porter-Hudak (1983), people have started to estimate the long-memory parameter d of a long-memory process using a regression equation involving the periodogram ordinates at the Fourier frequencies as well as d itself. The basic idea is to use (1.10) and to consider $I_N(\omega)$ as an estimate of $f(\omega)$. From (1.10), in a neighborhood of the origin, one can write $f(\omega) \approx c\omega^{-d}$. Therefore, considering only a limited number $k \ll N/2$ of Fourier frequencies, substituting $I_N(\omega_j)$ for $f(\omega_j)$ and taking logarithms, one can try to estimate d from the set of regression equations:

$$\begin{cases} \log I_N(\omega_1) = -d \log \omega_1 + \log c + \eta_1 \\ \log I_N(\omega_2) = -d \log \omega_2 + \log c + \eta_2 \\ \dots \\ \log I_N(\omega_k) = -d \log \omega_k + \log c + \eta_k \end{cases} \tag{1.31}$$

where $\eta_1, \dots, \eta_k$ represent the error terms. The exact distribution of $\eta_1, \dots, \eta_k$ is in general very difficult to find, and people usually resort to asymptotic approximations: Geweke and Porter-Hudak, using the classical asymptotic theory illustrated in Section 1.2, considered the η_j's to be independent and identically distributed, with a log χ^2_1 distribution, and proposed to estimate d using the method of least squares. More recently, a number of studies have appeared dealing with alternative asymptotic results for the periodogram ordinates of a long-memory process. We will review in the present section the results of Hurvich and Beltrao (1993) (see also Hurvich and Beltrao (1994)) and Terrin and Hurvich (1994). Related results, in the context of a semiparametric estimation of d, can be found in Robinson (1995b).

The focus is on finding, for a long-memory process, the asymptotic distribution, and/or its first moments, of the normalised periodogram ordinates at a set of frequencies $\omega_{j_1}, \ldots, \omega_{j_k}$, where $\omega_{j_r} = 2\pi j_r/N$ $(r = 1, \ldots, k)$. Here $j_1, \ldots, j_k$ are fixed and the asymptotic behaviour is studied for N going to infinity.

Hurvich and Beltrao (1993) assume a spectral density of the form

$$f(\omega) = \left|1 - e^{-i\omega}\right|^{-d} f^*(\omega), \tag{1.32}$$

where $d \in (-1, 1)$ and f^* is a positive, continuous function on $[0, \pi]$, bounded away from 0. In their terminology the definition of a long-memory process include the case when d is in $(-1, 0)$. Other authors refer to a process with negative d as having intermediate memory. This is justified by the fact that for such a process, the covariance decays, as the lag goes to infinity, at a rate slower than exponential, but fast enough for the series of the covariances to absolutely converge. The normalised periodogram ordinates are defined as

$$\frac{I_N(\omega_{j_r})}{f(\omega_{j_r})}, \qquad r = 1, \ldots, k. \tag{1.33}$$

The authors find that the asymptotic mean of the normalised periodogram ordinates can be given the expression

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{I_N(\omega_{j_r})}{f(\omega_{j_r})}\right) = \frac{2}{\pi} \int_{\mathbb{R}} \frac{\sin^2(\lambda/2)}{(2\pi j_r - \lambda)^2} \left|\frac{\lambda}{2\pi j_r}\right|^{-d} d\lambda. \tag{1.34}$$

The right-hand side of (1.34) is a function of both d and j_r. A plot of this function for $d \in (-1, 1)$ and $j_r \in (1, 10)$ (Figure 1.3) reveals that the asymptotic relative bias in $I_N(\omega_{j_r})$:

1. decreases rather quickly, for any fixed value of d, as j_r increases,
2. increases, for any fixed j_r, as $|d|$ goes to one,
3. is typically greater at $-|d|$ than at $|d|$.

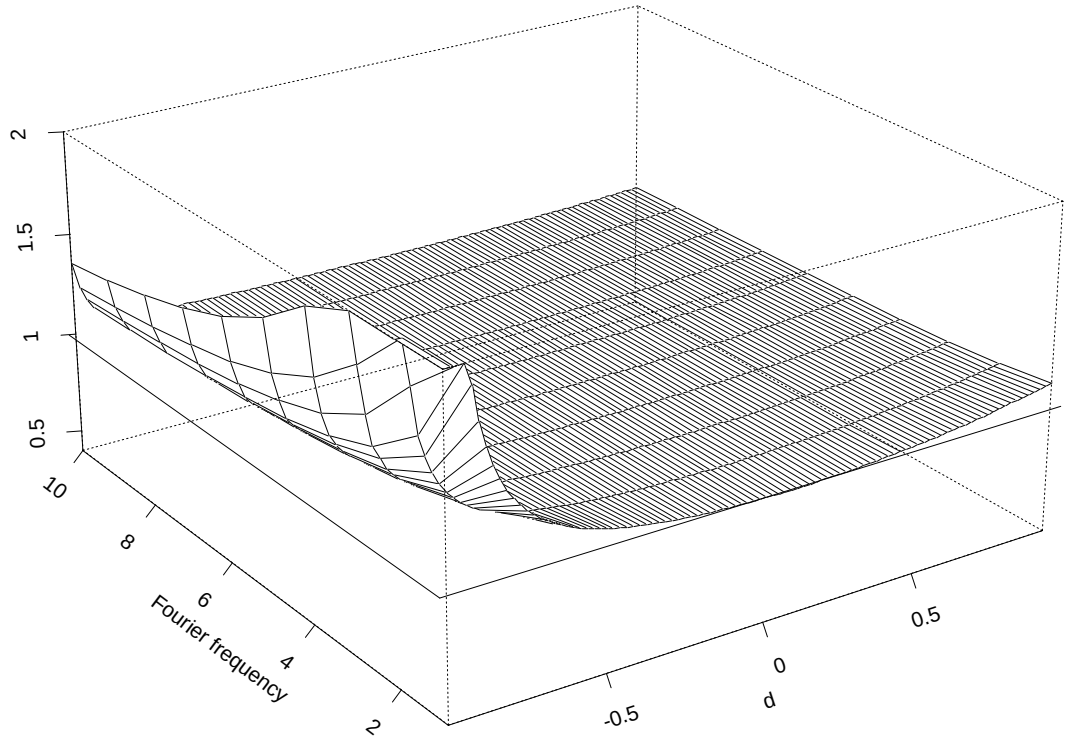

Figure 1.3: Asymptotic relative bias of the periodogram

To give an idea about the actual values, for $j_r = 1$ the function attains the value 1.14 at $d = 0.98$ and the value 1.65 at $d = -0.98$, which are essentially the maximum asymptotic relative bias for positive and negative d respectively. On the other hand, the minimum obtained with numerical methods is 0.9993. The authors also provide expressions for the asymptotic second moments of the normalised Fourier coefficients. The Fourier coefficient are defined by

$$A_N(\omega_{j_r}) = \frac{1}{\sqrt{N\pi}} \sum_{t=1}^{N} X_t \cos(\omega_{j_r} t), \tag{1.35}$$

$$B_N(\omega_{j_r}) = \frac{1}{\sqrt{N\pi}} \sum_{t=1}^{N} X_t \sin(\omega_{j_r} t), \tag{1.36}$$

and the normalised Fourier coefficient are given by

$$A_N(\omega_{j_r}) \Big/ \sqrt{f(\omega_{j_r})}, \tag{1.37}$$

$$B_N(\omega_{j_r}) \Big/ \sqrt{f(\omega_{j_r})} \tag{1.38}$$

respectively. Note that $I_N = A_N^2 + B_N^2$. The main result is that, for $j_r \neq j_s$:

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{A_N(\omega_{j_r}) B_N(\omega_{j_s})}{\sqrt{f(\omega_{j_r}) f(\omega_{j_s})}} \right) = 0; \tag{1.39}$$

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{A_N(\omega_{j_r}) A_N(\omega_{j_s})}{\sqrt{f(\omega_{j_r}) f(\omega_{j_s})}} \right) = (-1)^{j_r+j_s+1} |(2\pi j_r)(2\pi j_s)|^{d/2} \frac{1}{\pi} \cdot$$

$$\int_{\mathbb{R}} \left(\frac{\sin^2(\lambda/2)}{(2\pi j_r - \lambda)(2\pi j_s + \lambda)} + \frac{\sin^2(\lambda/2)}{(2\pi j_r - \lambda)(-2\pi j_s + \lambda)} \right) |\lambda|^{-d} \, d\lambda; \tag{1.40}$$

$$\lim_{N\to\infty} \mathbb{E}\left(\frac{B_N(\omega_{j_r}) B_N(\omega_{j_s})}{\sqrt{f(\omega_{j_r}) f(\omega_{j_s})}} \right) = (-1)^{j_r+j_s+1} |(2\pi j_r)(2\pi j_s)|^{d/2} \frac{1}{\pi} \cdot$$

$$\int_{\mathbb{R}} \left(\frac{\sin^2(\lambda/2)}{(2\pi j_r - \lambda)(-2\pi j_s + \lambda)} - \frac{\sin^2(\lambda/2)}{(2\pi j_r - \lambda)(2\pi j_s + \lambda)} \right) |\lambda|^{-d} \, d\lambda; \tag{1.41}$$

An interesting consequence is that the normalised periodogram ordinates $I_N(\omega_{j_r})/f(\omega_{j_r})$ and $I_N(\omega_{j_s})/f(\omega_{j_s})$ are asymptotically correlated if $d \neq 0$.

Another result reported in the same paper is related to the asymptotic distribution of the normalised periodogram at one frequency ω_{j_r}, in the case of a Gaussian process. This distributional result has been extended in Terrin and Hurvich (1994). The extension is twofold: on one side the Normality assumption is weakened, while on the other they consider the joint asymptotic distribution of a set of normalised periodogram ordinates. We report below the main result of the paper. Let $(\xi_t)_{t\in\mathbb{Z}}$ be

a sequence of independent identically distributed random variables with finite fourth moment and variance σ^2, and let $\mathbf{X} = (X_t)_{t\in\mathbb{Z}}$ be the linear process defined by:

$$X_t = \sum_{k=-\infty}^{\infty} b_k \xi_{t-k}, \qquad t \in \mathbb{Z}, \tag{1.42}$$

where

$$b_k = \frac{1}{2\pi}\int_0^{2\pi} e^{ik\lambda} B(\lambda)\, d\lambda, \qquad k \in \mathbb{Z} \tag{1.43}$$

for some complex valued function $B \in L^2([0,2\pi], d\lambda)$ satisfying $B(\lambda) = \overline{B(2\pi-\lambda)}$, and $\int B(\lambda)\, d\lambda = 2\pi$. Assume that the spectral density of $\mathbf{X}$,

$$f(\lambda) = \frac{\sigma^2}{2\pi}|B(\lambda)|^2 \tag{1.44}$$

can be written in the form

$$f(\lambda) = \frac{\sigma^2}{2\pi}\lambda^{-d} L(\lambda), \tag{1.45}$$

for some $d \in (-1,1)$ and some function L, slowly varying as $\lambda \to 0$. Note that the form (1.45) generalises (1.32). In order to state the main theorem we need some additional notation. Let W_λ a zero mean complex valued Gaussian process, indexed in $\mathbb{R}$, such that

$$W_{\lambda_2} - W_{\lambda_1} = \overline{W_{2\pi-\lambda_1} - W_{2\pi-\lambda_2}}, \tag{1.46}$$

$$\mathrm{Cov}(W_{\lambda_2} - W_{\lambda_1}, W_{\lambda_4} - W_{\lambda_3}) = \big((\lambda_2 \wedge \lambda_1) - (\lambda_1 \vee \lambda_3)\big)^+, \quad \lambda_1 \le \lambda_2, \lambda_3 \le \lambda_4. \tag{1.47}$$

Define

$$Y(j) = \frac{(2\pi j)^d}{2\pi}\left|\int_{\mathbb{R}} \frac{e^{i(\lambda+2\pi j)}-1}{i(\lambda+2\pi j)}|\lambda|^{-d/2}\, dW_\lambda\right|^2. \tag{1.48}$$

The main result is that, for any fixed integers $j_1, \ldots, j_k$, as the sample size N goes to infinity,

$$\left(\frac{I_N(\omega_{j_1})}{f(\omega_{j_1})}, \ldots, \frac{I_N(\omega_{j_k})}{f(\omega_{j_k})} \right) \xrightarrow{\ell} (Y_{j_1}, \ldots, Y_{j_k}), \tag{1.49}$$

where "$\xrightarrow{\ell}$" denotes convergence in law.

1.5 A simulation study.

The aim of the study of asymptotic distributions is to find a probability law which approximates the actual distribution af a statistic of interest. In the present case, we are interested in an approximation to the distribution, τ, say, of the normalised periodogram ordinates (1.33), for a given sample size N. Let us explicitly take into account the dependence of the frequencies $\omega_{j_1}, \ldots, \omega_{j_k}$ on N in the notation by writing $\omega_{j_1}^{(N)}, \ldots, \omega_{j_k}^{(N)}$. Note that, in general, there exist many sequences of distributions, converging to different limits, and containing τ as an element. Therefore there is a choice to make when using an asymptotic approximation. To be specific, in our case one can use, for example, either one of the two asymptotic distributions

$$\lim_{n\to\infty} \mathcal{L} \left(\frac{I_n(\omega_{j_1}^{(n)})}{f(\omega_{j_1}^{(n)})}, \ldots, \frac{I_n(\omega_{j_k}^{(n)})}{f(\omega_{j_k})^{(n)}} \right), \tag{1.50}$$

$$\lim_{n\to\infty} \mathcal{L} \left(\frac{I_n(\omega_{j_1}^{(N)})}{f(\omega_{j_1}^{(N)})}, \ldots, \frac{I_n(\omega_{j_k}^{(N)})}{f(\omega_{j_k}^{(N)})} \right), \tag{1.51}$$

where, for any random vector V, we denote by $\mathcal{L}(V)$ its the probability law. The difference in the two expressions is that in the first one the frequencies at which the normalised periodogram is computed are not kept fixed to the ones we are interested in. The second one correspond to the classical approximation, while the first one is the one analysed in the papers reported above.

The problem of choosing one asymptotic approximation instead of the other is in general a very difficult one, at least if one wants to solve it on the basis of theoretical considerations about which approximation is better. In fact, an exact solution would require an evaluation of the distance between each of the approximations and the distribution of interest. As far as we know nobody has given an explicit formula for that, for any fixed sampling distribution of the periodogram ordinates. Moreover, we are dealing with a model, i.e. a class of sampling distributions. Therefore it may well be that for some values of the parameters one approximation is better, while for other values the converse is true. For reasons of mathematical convenience, we will use for our model the classic asymptotic approximation. In the present section we argue that our choice is a reasonable one or, in other words, that the sampling distribution of the periodogram ordinates at the Fourier frequencies is well approximated by the classical distributional approximation. Since, as we already mentioned, the theory does not help on this point, we will resort to a simulation study.

We present in the following of the section an empirical study, aimed at supporting the choice we made in favour of the classical approximation versus the nonstandard one. In particular we will show that, even at low Fourier frequencies, the periodogram ordinates of a long-memory process can be considered, from a practical point of view, independent. We considered a fractionally differenced white noise with long-memory parameter 0.998, and we generated an independent sample of size 10,000 from the joint distribution of the periodogram ordinates at the first 20 Fourier frequencies corresponding to 2,000 observations from such a process. We then tried to detect if there is any dependence among the simulated periodogram ordinates. First of all we looked for pairwise dependence. We computed the empirical covariances of the simulated periodogram ordinates. Of the 190 distinct covariances, 90% are less then 0.02 in absolute value, and the maximum is 0.04. Figure 1.4 represents an image plot

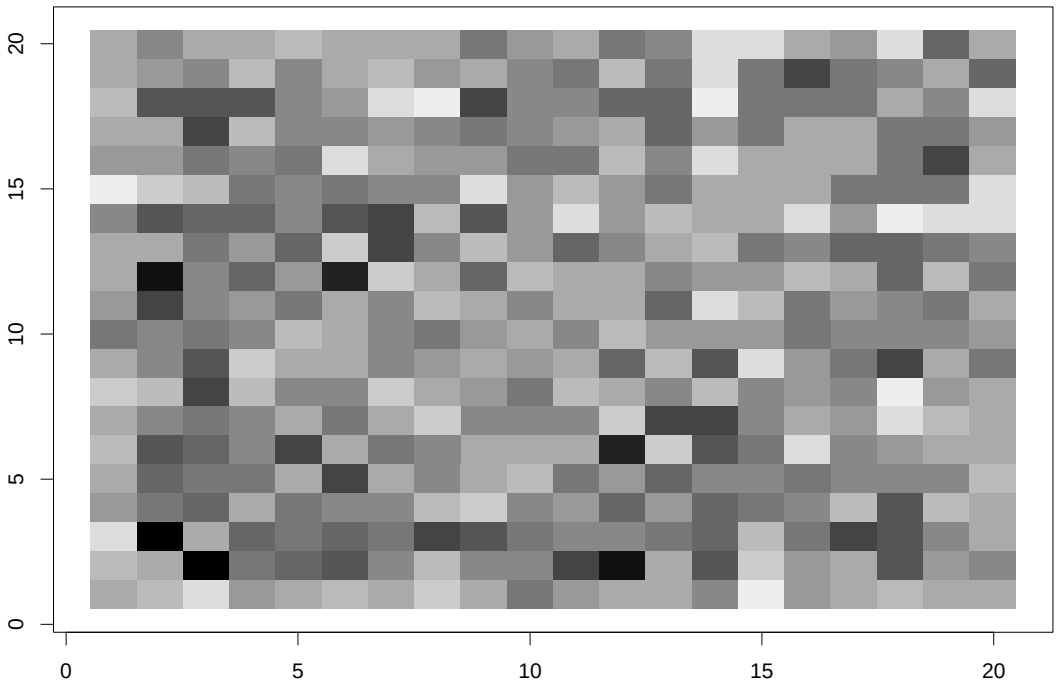

Figure 1.4: Image plot of the empirical correlations.

of the empirical correlation matrix, with the diagonal set to zero. Darker regions correspond to smaller values. It does not seem to be any particular structure in the correlations. The correlations between $I(\omega_1), \ldots, I(\omega_6)$ and the other simulated periodogram ordinates are shown in Figure 1.5. It is well known that the correlation is an indicator of a linear relationship between two random variables. To detect a pairwise nonlinear dependence we looked at the the conditional histograms of one periodogram ordinate, given that the other is in a specified interval. Consider, for example, $I(\omega_1)$ and $I(\omega_2)$ (see Figure 1.6). Table 1.1 gives the set of intervals considered for $I(\omega_1)$ together with the appropriate counts. Figure 1.7 shows the

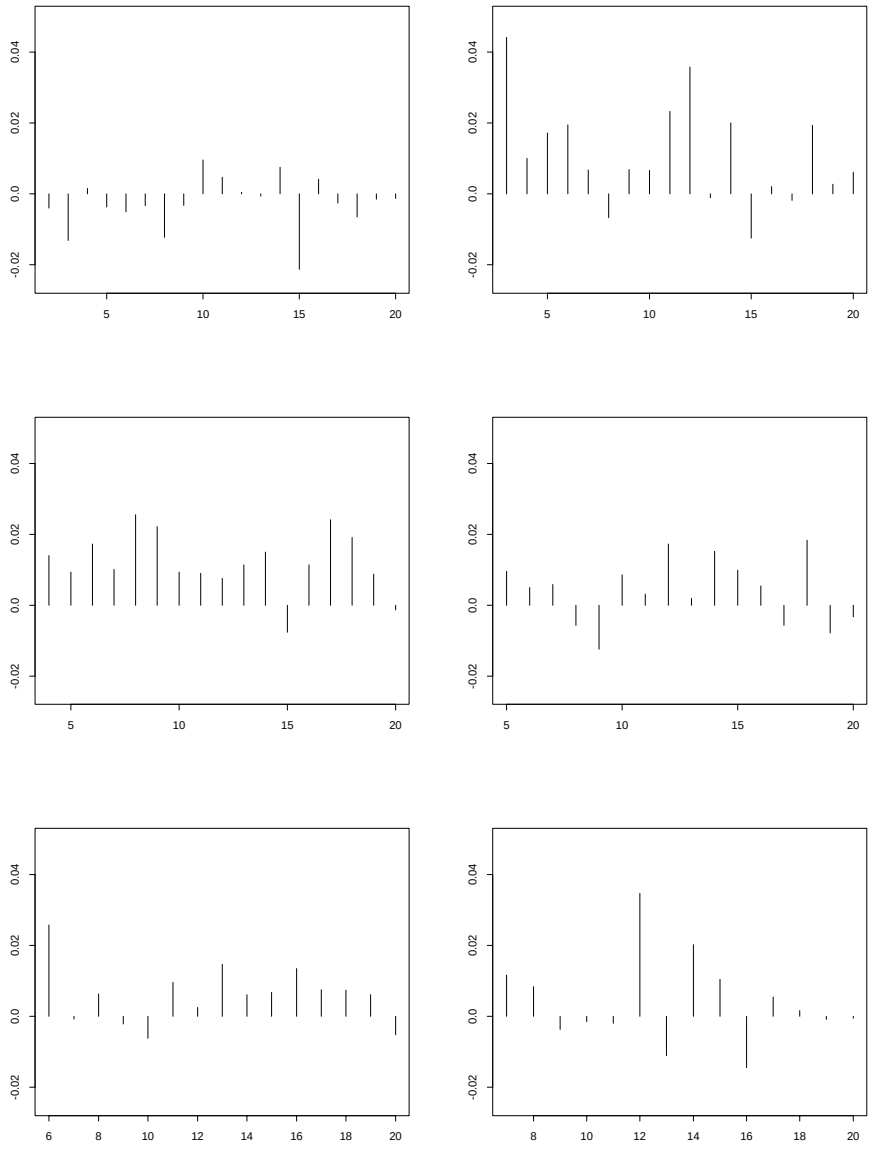

Figure 1.5: Correlation between periodogram ordinates.

conditional histogram of $I(\omega_2)$ given that $I(\omega_1)$ is in the i-th class, $i = 1, \ldots, 5$. It also contains the unconditional histogram of $I(\omega_2)$ for comparison. The influence of the value of $I(\omega_1)$ on the distribution of $I(\omega_2)$ seems to be negligible, since the plot clearly suggest that the distribution of $I(\omega_2)$ conditional on $I(\omega_1)$ is close to the

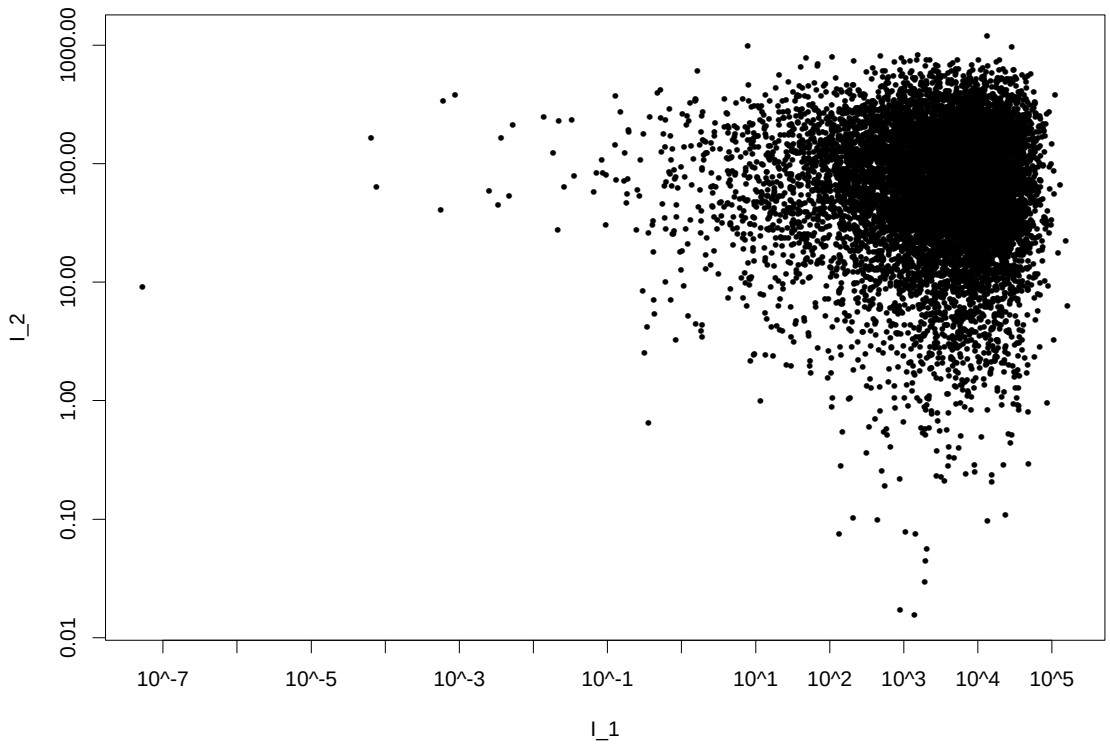

Figure 1.6: $I(\omega_2)$ versus $I(\omega_1)$, log scale.

unconditional distribution of $I(\omega_2)$.

After this graphical qualitative analysis we performed the classical χ-square test of independence. The test confirmed the feelings one can get from the graphical analysis. Grouping into ten classes the realisations of $I(\omega_1)$ and $I(\omega_2)$, see Table 1.2, we performed Pearson independence test. We obtained for the chi-squared statistic, on 81 degrees of freedom, the value of 63.82, which corresponds to a p-value of 0.92. Similar results were obtained for other pairs of periodogram ordinates.

Looking for a possible dependence among all the first twenty periodogram ordinates, we fitted a linear model using $I(\omega_1)$ as response variable and

Class	1	2	3	4	5
Interval	[0,566)	[566,2412)	[2412,6282)	[6282,14400)	$[14400, \infty)$

Table 1.1: Breakpoints.

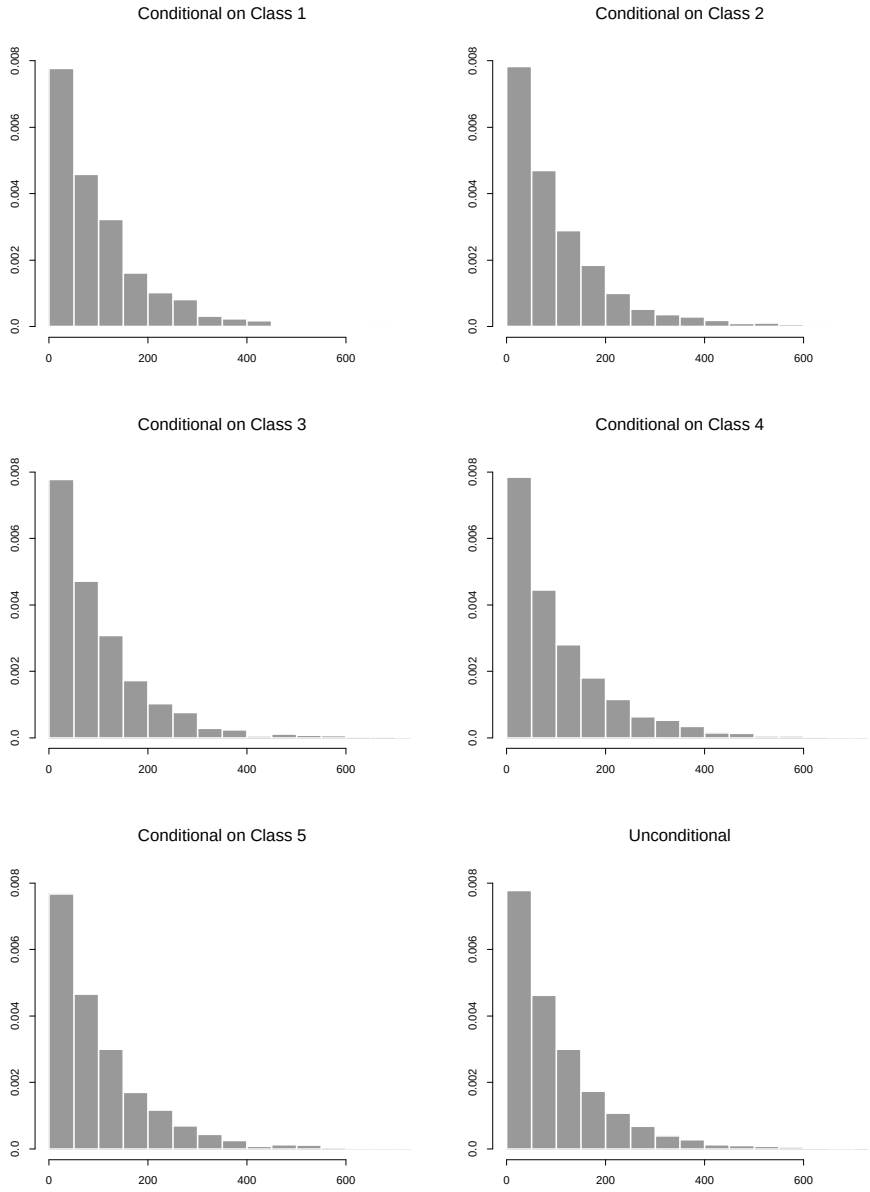

Figure 1.7: Conditional histograms

	[0,10)	[10,23)	[23,37)	[37,52)	[52,72)	[72,95)	[95,125)	[125,166)	[166,238)	[238,∞)
[0, 938)	236	243	274	280	245	248	268	265	235	255
[938, 1988)	103	116	102	114	117	104	125	92	114	113
[1988, 3178)	91	88	99	83	87	72	95	89	86	83
[3178, 4552)	62	72	79	95	89	69	97	79	64	71
[4552, 6177)	85	62	68	77	61	51	79	60	62	58
[6177, 8165)	65	57	72	70	61	55	78	62	69	74
[8165, 10729)	60	84	69	72	64	60	68	71	62	72
[10729, 14343)	57	81	65	83	65	62	71	58	74	69
[14343,20520)	60	60	65	76	66	68	76	67	65	73
[20520, ∞)	126	139	140	132	139	129	121	146	128	132

Table 1.2: Counts for the χ-square test of independence.

$I(\omega_2), \ldots, I(\omega_{20})$ as covariates. The proportion of the variation of $I(\omega_1)$ explained by the variation of the covariates (R-square) was only 0.1%, with a p-value for the standard F-test of 0.93.

As far as the distribution of every single periodogram ordinate is concerned, one can plot the quantiles of the empirical distribution of the sampled periodogram ordinate versus the quantiles of the exponential distribution (qqplot). In Figure 1.8, for example, the qqplot for the periodogram ordinate corresponding to the first Fourier frequency is shown on a log-log scale. Similar plots can be obtained for the empirical distribution of the other periodogram ordinates.

Though the study is limited, this empirical analysis suggests that the distributional assumptions implicit in the approximate likelihood that we will use in next chapter can be considered satisfied for practical purposes.

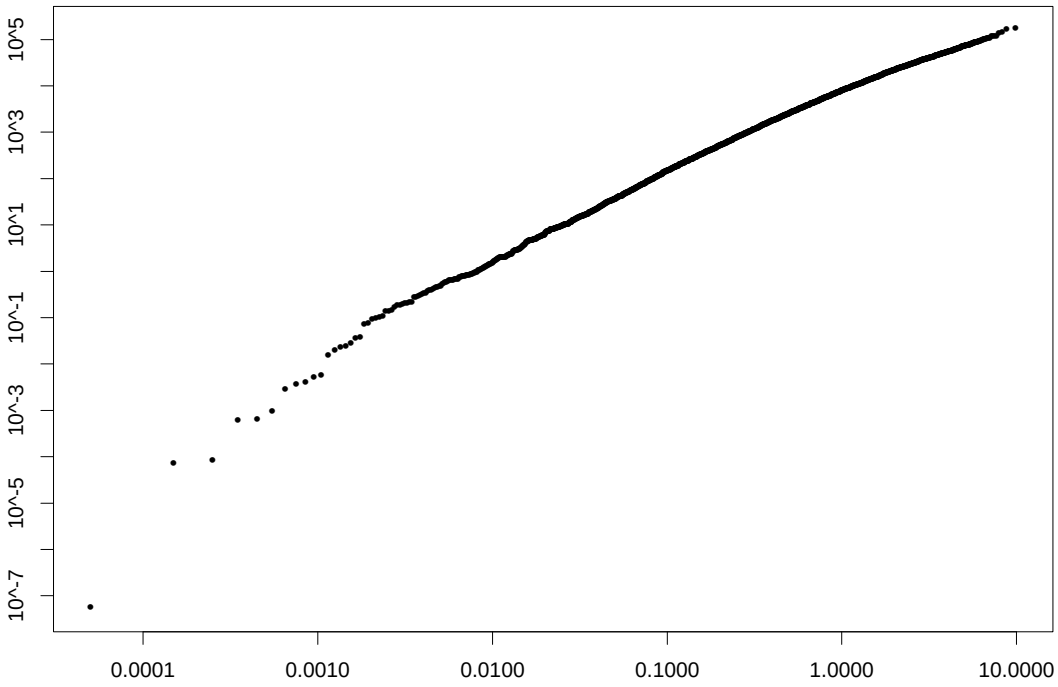

Figure 1.8: QQplot.

Chapter 2

A Class of Models for Spectral Densities

2.1 Description of the Model

We are interested in the study of a discrete time second order stationary stochastic process, possessing a spectral density f which may present a pole at the origin and is otherwise continuous on $(0, 2\pi)$. Assume that f can be written as

$$f(\lambda) = \lambda^{-d} \cdot \exp(g(\lambda)) \,, \tag{2.1}$$

with $d \in [0, 1)$ and g a function in $C([0, 2\pi])$. A decomposition of the spectral density of the form (2.1) is useful when trying to make inference about the long-memory parameter, since d is explicitly present as a parameter of the model. Although presented in a different form, (2.1) is equivalent to the model used in Geweke and Porter-Hudak (1983). Models involving a finite-dimensional parameter as well as an infinite-dimensional one are sometimes referred to as semiparametric models. Let us set $s = \lfloor N/2 \rfloor - 1$, $\lambda_j = 2j\pi/N$, $y_j = \log(I_N(\lambda_j))$ and $g_j = g(\lambda_j)$, for $j = 1, ..., s$. Then, using the asymptotic result mentioned in the previous chapter, one obtain that, approximately,

$$y_j = -d \log \lambda_j + g_j + \log(\tfrac{1}{2}\chi^2_2) \,, \tag{2.2}$$

for $j = 1, ..., s$. This set of equations can be interpreted as a regression problem with independent errors and error distribution $\log(\frac{1}{2}\chi_2^2)$. A similar approach is taken in Carter and Kohn (1997), although they deal with short memory models only. For the purpose of simulation it is convenient, following Carter and Kohn (1997), to approximate the distribution of $\log(\frac{1}{2}\chi_2^2)$ with a mixture of normal distributions:

$$\log(\tfrac{1}{2}\chi_2^2) \simeq \sum_{j=1}^{J} p_j N(\mu_j, V_j)\,. \tag{2.3}$$

The approximation can be made to any degree of accuracy, since the set of finite mixtures of normal distributions is dense in the space of all the probability distributions on the real line, endowed with Kolmogorov's metric . We found that a mixture of 5 components, with the constants $p_j, \mu_j, V_j,\ j = 1, ..., 5$ given in Table 2.1, is accurate enough for most practical purposes. If one introduces a new vector

p_j	μ_j	V_j
0.19	–2.200	1.93
0.11	–0.800	1.01
0.27	–0.550	0.69
0.25	–0.035	0.60
0.18	0.480	0.29

Table 2.1: Mixture of normals approximation to $\log(\frac{1}{2}\chi_2^2)$.

of parameters $\mathbf{K} = (K_1, ..., K_s)$, whose components are independent and identically distributed with $P(K_j = k) = p_k,\ k = 1, ..., 5$, then conditionally on $\mathbf{K}$ the y_j's are independent with distribution

$$y_j \sim N(-d \log \lambda_j + g_j + \mu_{K_j}, V_{K_j})\,, \tag{2.4}$$

$j = 1, ..., s$.

2.2 Prior distribution.

The present section contains the description of a class of prior distributions for the parameters of the model. We will assume that g and d are independent, and present the prior for each of them in two distinct subsections.

2.2.1 Smoothness prior for g.

This subsection contains the description of a class of smoothness priors for the g_j's. We assume that $g_j = \tilde{g}_j + m$, where m is a constant and $\tilde{g}$ is a centered stationary Gaussian process which satisfy the autoregressive relation

$$(1 - \gamma B)^p \tilde{g} = \epsilon, \tag{2.5}$$

where B is the backshift operator, γ is a number in $(0, 1)$, p is a positive integer and ϵ is a Gaussian white noise with variance σ_ϵ^2. A prior opinion about the spectral density may be incorporated in the model by considering a sequence (m_j) instead of a single constant m as a baseline for the process g. Our choice corresponds to a flat function as *a priori* estimate for the spectral density. We consider m, γ and σ_ϵ^2 as hyperparameters with the following prior distribution. Let m be normally distributed with mean μ_m and variance s_m^2. For γ we assume a Beta distribution with parameters a_γ, b_γ. Instead of putting a prior directly on σ_ϵ^2, we prefer to work with the variance of $\tilde{g}$, which will be denoted by σ_G^2. In general σ_G^2 is a multiple of σ_ϵ^2 and the proportionality constant can be found by solving a system of linear equations (Brockwell and Davis, 1991, p. 91). For example, for p equal 4:

$$\sigma_G^2 = \frac{(1+\gamma^2)(\gamma^4 + 8\gamma^2 + 1)}{(1-\gamma)^7(1+\gamma)^7}\,\sigma_\epsilon^2. \tag{2.6}$$

We will denote by κ the ratio of σ_G^2 to σ_ϵ^2. We assume that σ_G^2 has an Inverse Gamma distribution with parameters a_G, b_G and that σ_G^2, γ and m are *a priori* independent.

This completely specify the prior distribution. Let $\mathbf{G}$ denotes the vector whose j-th component is g_j and $\mathbf{1}$ be the vector whose components are all ones. Then, given γ, m and σ_G^2, the prior for $\mathbf{G}$ has the Gaussian density

$$(2\pi\sigma_G^2)^{-\frac{n}{2}}|\mathbf{D}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma_G^2}(\mathbf{G}-\mathbf{1}m)^t\mathbf{D}^{-1}(\mathbf{G}-\mathbf{1}m)\right), \tag{2.7}$$

where n is the dimension of the vector $\mathbf{G}$ and $\mathbf{D}$ is a positive definite correlation matrix which depends on γ.

Let us spend a couple of words on the reasons which dictate the choice of the class of priors described above. In general, the elicitation of a prior estimate for the spectral density of a stationary time series is a pretty hard task. By far more hard is to elicit directly in a meaningful way an entire probability distribution for it. At a more fundamental level, following the point of view of de Finetti, one may argue if there is any meaning at all in making probability statements about unobservable parameters like, in the present case, the spectral density (see for example de Finetti (1937)). We will not take that point of view here, but see Regazzini and Petris (1992) for a discussion about the hypothetical and predictive approaches to statistics within a Bayesian framework. In spite of the difficulty of eliciting the prior, we try to express a qualitative opinion about the smoothness of g. The realisations of an autoregressive process are likely to be smooth, and the choice of the particular form (2.5), instead of a more general one, is due to a search for a parsimonious parametrisation. In fact, keeping fixed the order p, the parameter γ may be thought as a 'smoothness parameter': the higher γ is and the smoother, on average, the realisations of the g_j's will be. On the other hand, increasing the order p will also give a smoother g. Figure 2.1 illustrates the typical behaviour of the smoothness prior showing some realisations of length 1000 from an AR(p) series having the form (2.5) for different values of p and γ. We fixed $\sigma_G^2 = 1$. In each row p is constant, equal to 2, 4, 6, from top to bottom. Similarly, γ is constant along columns, being 0.9 and 0.99, left to

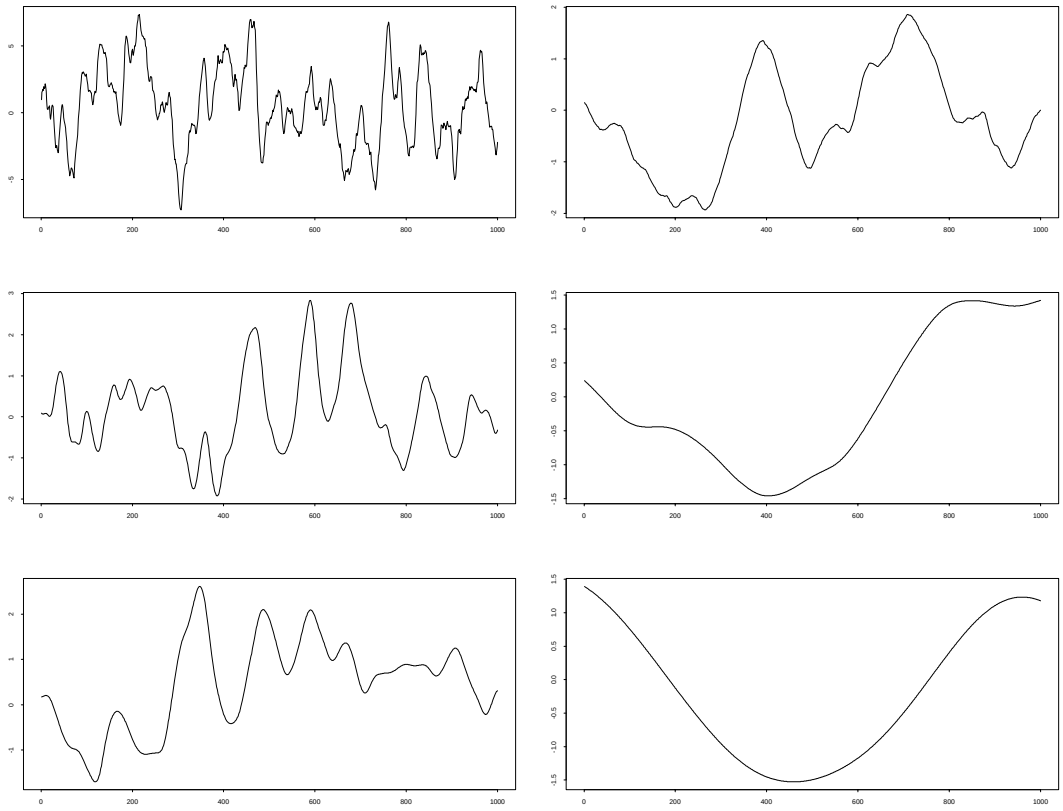

Figure 2.1: Samples from smoothness priors.

right. Smoothness priors, translating vague qualitative opinions about the behaviour of a curve, have been used in nonlinear regression by Wahba (1978) and Ansley and Kohn (1985). Carter and Kohn (1997) use an integrated Wiener process to model the log spectral density of a short memory time series.

2.2.2 Prior for d.

As far as d is concerned, we assume that it is independent on $\mathbf{G}$ and its prior distribution is a mixture with weights α, $1-\alpha$ of a point mass at 0 and a continuous distribution on $(0, 1)$. In this way we are giving a positive prior probability to the subset of the parameter space corresponding to long-memory processes. This

allow us to compute posterior probabilities of events such as $\{d = 0\}$, which are needed, for example, to perform a Bayesian test of hypothesis of $H_0 : d = 0$ versus $H_a : d \neq 0$. The continuous part of the prior will be taken to be a Beta distribution with parameters a_d and b_d. This distribution can accommodate a variety of different prior opinions about the long-memory parameter. For reasons that will become clear in the following sections, related to the need of simulating from the posterior distribution, it is convenient to have the prior distribution given in terms of a density with respect to a measure. If we denote by q the measure given by the sum of Lebesgue measure on $(0, 1)$ and a probability measure on $\{0\}$, we can write the density of the prior distribution of d with respect to q as

$$\pi(d) \;:=\; \alpha \mathbf{1}_{\{0\}}(d) + (1-\alpha)\frac{1}{B(a_d, b_d)}\mathbf{1}_{(0,1]}(d) d^{a_d - 1}(1-d)^{b_d - 1}\,, \tag{2.8}$$

where $\mathbf{1}_A$ denotes the indicator function of the set A and B is the beta function defined by $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1}\, du$.

2.3 Structure of the posterior distribution.

For the model described in the previous section the posterior density of all the parameters, i.e., $\mathbf{G}$, γ, m, σ_G^2, d, $\mathbf{K}$, considering the approximate likelihood implied by (2.4), is proportional to

$$\begin{aligned} &\pi(d)(\sigma_G^2)^{-(a_G + \frac{n}{2} + 1)} |\mathbf{D}|^{-\frac{1}{2}} \gamma^{a_\gamma - 1}(1-\gamma)^{b_\gamma - 1} \prod_j p_{K_j} \\ &\cdot \frac{1}{(\prod_j V_{K_j})^{1/2}} \exp\left[-\frac{1}{2}\sum_j \frac{(y_j + d\log\lambda_j - g_j - \mu_{K_j})^2}{V_{K_j}} \right. \\ &\left. -\frac{1}{2\sigma_G^2}(\mathbf{G} - \mathbf{1}m)^t \mathbf{D}^{-1}(\mathbf{G} - \mathbf{1}m) - \frac{1}{2s_m^2}(m - \mu_m)^2 - \frac{1}{\sigma_G^2 b_G}\right]. \end{aligned} \tag{2.9}$$

The form of the posterior is quite involved, and computing summary statistics, such as first and second moments, using an analytic approach has not been proven feasible,

as far as we know. One may think, as an alternative, to resort to the machinery of numerical analysis. Unfortunately, also this approach is not going to succeed, due to the very high dimensionality of the problem. Typically, in real applications, the number of the g_j's is of the order of several hundreds or thousands: a number that any standard numerical methods of integration or optimisation cannot possibly handle. We will approach the problem of analysing the posterior distribution via simulation. Specifically, we will resort to a *Markov chain Monte Carlo* method, to be described in this section and in the next, in order to generate a sample from approximately the posterior distribution. If the chain satisfies an appropriate ergodicity condition, one can then approximate means and variances of the posterior using the corresponding empirical moments of the simulated sample. Markov chain Monte Carlo methods have gained great popularity among Bayesian statisticians in recent years, since they often offer the only feasible way to analyse complicate models with a large number of parameters. A good review of Markov chain Monte Carlo methods is provided in Tierney (1994). The method we propose here belongs to the class of the so-called *successive substitution sampling* schemes (see Schervish (1995)), with some Metropolis steps embedded. In order to implement it, we need the full conditional distributions, i.e., the one-dimensional distributions of each parameter given all the other (and the data).

Full conditional distribution of m and σ_G^2. From (2.9), one can easily deduce that the full conditional distributions of m and σ_G^2 are respectively Normal with mean μ_m^* and variance ${\sigma_m^*}^2$ and Inverse Gamma with parameters a_G^* and b_G^*, where

$$\mu_m^* = {\sigma_m^*}^2 \left(\frac{\mathbf{1}^t \mathbf{D}^{-1} \mathbf{G}}{\sigma_G^2} + \frac{\mu_m}{s_m^2} \right), \tag{2.10}$$

$$\sigma_m^{*\,2} = \left(\frac{\mathbf{1}^t \mathbf{D}^{-1} \mathbf{1}}{\sigma_G^2} + \frac{1}{s_m^2} \right)^{-1} \tag{2.11}$$

are the parameters of the full conditional distribution of m, and

$$a_G^* = a_G + \frac{n}{2}, \tag{2.12}$$

$$b_G^* = \left(\frac{1}{b_G} + \frac{1}{2} (\mathbf{G} - \mathbf{1}m)^t \mathbf{D}^{-1} (\mathbf{G} - \mathbf{1}m) \right)^{-1}, \tag{2.13}$$

are the parameters of the full conditional distribution of σ_G^2. In this respect, the prior distributions of m and σ_G^2 behave like conjugate priors. Draws from these distributions can be easily obtained directly using readily available methods, see for example Devroye (1986).

Full conditional distribution of γ. The full conditional distribution of γ, on the other hand, is quite involved, because of the presence of γ in the matrix $\mathbf{D}$. The density has the following form, up to a proportionality constant:

$$|\mathbf{D}|^{-\frac{1}{2}} \gamma^{a_\gamma - 1} (1-\gamma)^{b_\gamma - 1} \exp\left[-\frac{1}{2\sigma_G^2} (\mathbf{G} - \mathbf{1}m)^t \mathbf{D}^{-1} (\mathbf{G} - \mathbf{1}m) \right]. \tag{2.14}$$

The details on the Metropolis step that we propose to sample from it will be explained in the next section.

Full conditional distribution of d. The full conditional distribution of d has density π^* with respect to the measure q, where

$$\pi^*(d) \;:=\; c_0 \pi(d) \prod_j \phi\left(\frac{y_j + d \log \lambda_j - g_j - \mu_k}{\sqrt{V_r}} \right) \tag{2.15}$$

and c_0 is a normalising constant. The method we use to generate from this distribution, suggested by Tardella (1996), is of some interest because it does not require the integration of π^*, which can be expensive in terms of computational time, in order to determine the relative weights of the discrete and absolutely continuous components. The details are postponed to the next section.

Full conditional distribution of K. The components of the vector $\mathbf{K}$ remain independent in the full conditional, as it can be easily seen from the form of the full conditional, which is, up to a normalising constant:

$$\prod_j p_{K_j} \cdot \frac{1}{(\prod_j V_{K_j})^{1/2}} \exp\left[-\frac{1}{2}\sum_j \frac{(y_j + d\log\lambda_j - g_j - \mu_{K_j})^2}{V_{K_j}}\right]. \tag{2.16}$$

Each K_j can therefore be sampled independently from the distribution defined by

$$\text{Prob}(K_j = k) \propto p_k \phi\left(\frac{y_j + d\log\lambda_j - g_j - \mu_k}{\sqrt{V_r}}\right), \tag{2.17}$$

where ϕ denotes the density of the standard normal distribution. Drawing a sample from a discrete distribution with five possible values is straightforward.

Full conditional distribution of G. The details on how we generate a sample from the full conditional distribution of $\mathbf{G}$ will be explained in the next section. Here we limit ourselves to report the form of the density of that distribution, which is given, up to a normalising factor, by

$$\exp\left[-\frac{1}{2}\sum_j \frac{(y_j + d\log\lambda_j - g_j - \mu_{K_j})^2}{V_{K_j}} - \frac{1}{2\sigma_G^2}(\mathbf{G} - \mathbf{1}m)^t \mathbf{D}^{-1}(\mathbf{G} - \mathbf{1}m)\right]. \tag{2.18}$$

From (2.18) it follows immediately that $\mathbf{G}$ has a full conditional distribution which is Gaussian.

2.4 Further details about the simulation scheme.

We give in this section the details about the techniques we use to sample from the full conditional distributions described in the previous section. We will limit ourselves to the ones which require some special attention.

Drawing from the full conditional distribution of γ. We use the Metropolis algorithm (Tierney, 1994, Sec. 2.3) to draw from the full conditional distribution of γ. This distribution is absolutely continuous with respect to Lebesgue measure on $(0,1)$, with density given, up to a constant factor, by (2.14). The key issue is to choose a 'good' proposal distribution, that is one which is close enough to the target distribution to keep the proportion of rejections reasonably low and, on the other hand, one which is easy to sample from. The method we use to choose the proposal distribution, based on a piecewise approximation of the density of the target distribution, can be applied every time the goal distribution has finite support and has a continuous positive density with respect to Lebesgue measure on an interval. For this reason we will describe it in general terms. Let ℓ be a continuous function in $[0,1]$, and suppose we want to use the Metropolis algorithm to sample from the distribution with cumulative distribution function given by:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \displaystyle\int_0^x e^{\ell(t)}\,dt & \text{if } 0 \le x \le 1 \\ 1 & \text{if } 1 < x. \end{cases} \tag{2.19}$$

In other words, ℓ is the logarithm of the density f of goal distribution F. Choose points $0 = x_0 < \cdots < x_n = 1$ and let $y_i = \ell(x_i)$, $i = 0, \ldots, n$. Let $y = a_i x + b_i$ the line through the points (x_{i-1}, y_{i-1}) and (x_i, y_i), $i = 1, \ldots, n$. In this way we obtain an exponential approximation to f in $[x_{i-1}, x_i]$:

$$f(x) \approx e^{a_i x + b_i} \qquad x \in [x_{i-1}, x_i]. \tag{2.20}$$

Define

$$\phi_i = \int_{x_{i-1}}^{x_i} e^{a_i t + b_i}\,dt \qquad i = 1, \ldots, n \tag{2.21}$$

and set $\Phi_i = \sum_{j=1}^{i} \phi_j$, $i = 1, \ldots, n$. The proposal distribution we suggest to use for the Metropolis algorithm is the one with cumulative distribution function G given by:

$$G(x) = \frac{1}{\Phi_n} \sum_{i=1}^{n} \mathbf{1}_{[x_{i-1}, x_i)}(x) \Big(\Phi_{i-1} + \frac{e^{b_i}}{a_i} \big(e^{a_i x} - e^{a_i x_{i-1}} \big) \Big) + \mathbf{1}_{[1, +\infty)}(x). \tag{2.22}$$

Sampling from this distribution is relatively easy using the *inverse cumulative distribution function* method. The method, in the present case, consists in drawing U from a uniform distribution on $(0, 1)$, and returning V such that $G(V) = U$. Then the distribution of V is precisely G. It is straightforward to show that, if $U \in \big(\frac{\Phi_{i-1}}{\Phi_n}, \frac{\Phi_i}{\Phi_n} \big)$, then

$$V = \frac{1}{a_i} \log \Big(\frac{a_i}{e^{b_i}} (\Phi_n U - \Phi_{i-1}) + e^{a_i x_{i-1}} \Big). \tag{2.23}$$

The density function of the proposal distribution G can be easily computed deriving (2.22) and turns out to be given by the following expression:

$$\frac{dG}{dx}(x) = \frac{1}{\Phi_n} \sum_{i=1}^{n} \mathbf{1}_{[x_{i-1}, x_i)}(x) e^{a_i x + b_i}. \tag{2.24}$$

Drawing from the full conditional distribution of d. Let us denote with $f(d)$ the density, up to a constant, of the absolutely continuous part of the prior distribution of d, i.e. a Beta density, and with $l(d)$ the product over j appearing in the right-hand side of (2.15). Let a be a point in $[0, 1]$, Δ a positive quantity and set $b = a + \Delta$. Consider an auxiliary random variable V, whose distribution is absolutely continuous with density

$$\begin{aligned} c_1 \big\{ & \mathbf{1}_{(0,a)}(v) f(v) l(v) + \mathbf{1}_{(a,b)}(v) f(a) l(a) \\ + \quad & \mathbf{1}_{(1+\Delta)}(v) f(v - \Delta) l(v - \Delta) \big\}, \qquad v \in \mathbb{R}. \end{aligned} \tag{2.25}$$

Fix any a such that $f(a)l(a) > 0$ and choose Δ in such a way that the random variable

$$V\mathbf{1}_{(0,a)}(V) + (V - \Delta)\mathbf{1}_{(b,1+\Delta)}(V) \tag{2.26}$$

has distribution $\pi^* \, dq$. The only condition we need to impose is that

$$\mathrm{P}rob(a < V < b) \; = \; \int\limits_{\{0\}} \pi^*(u) \, q(du) \, . \tag{2.27}$$

It can be easily verified that (2.27) is satisfied if

$$\Delta \; = \; \frac{\alpha l(0)}{(1-\alpha) f(a) l(a)} \, . \tag{2.28}$$

The problem in this way is reduced to that of generating a random variable with absolutely continuous distribution on a compact support. Observe that when the density of the absolutely continuous part of $\pi^* \, dq$ is log-concave, as in our case, with maximum at a, this feature is retained in the density of V. We did not exploit this feature; instead, to sample from the full conditional distribution of d in our Markov chain Monte Carlo simulation we used a Metropolis step, with a proposal distribution made up, *mutatis mutandis*, as the one described for the full conditional distribution of γ.

Drawing from the full conditional distribution of G. When all the parameters but the g_j's are given, one can reinterpret the model described in the previous section in terms of a Dynamic Linear Model (DLM). We need the following definitions, given for p equal 4:

$$x_j = \begin{bmatrix} g_{j-3} & g_{j-2} & g_{j-1} & g_j \end{bmatrix}^t , \tag{2.29}$$

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\gamma^4 & 4\gamma^3 & -6\gamma^2 & 4\gamma \end{bmatrix} , \tag{2.30}$$

$$\mathbf{F} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.31}$$

$$v_j \sim N(-d \log \lambda_j + \mu_{K_j}, V_{K_j}), \tag{2.32}$$

$$w_j \sim N\left(\begin{bmatrix} 0 & 0 & 0 & (1-\gamma)^4 m \end{bmatrix}^t, \mathrm{diag}(0,0,0,\sigma_\epsilon^2) \right) . \tag{2.33}$$

We assume that the v_j's and w_j's are all independent. Then our model is equivalent to the DLM described by the system

$$\begin{cases} y_i &= \mathbf{F} x_i + v_i \\ x_{i+1} &= \mathbf{H} x_i + w_{i+1} \,, \end{cases} \tag{2.34}$$

with $i = 1, ..., s$. The parameters g_j's are now the components on the state vector of a DLM and can be efficiently sampled using the Kalman filter as described in Carter and Kohn (1994). (See also Frühwirth-Schnatter (1994).) Let us explain in some detail how the procedure works. Note that the vector $(x_1, \ldots, x_s)$ has a distribution which is singular with respect to Lebesgue measure on $\mathbb{R}^{sp}$. For any two disjoint subsets U, V of indices in $\{1, \ldots, s\}$, let $\Psi_{U|V}$ be the conditional distribution of $(x_j)_{j \in U}$, given $(x_j)_{j \in V}$ and $(y_1, \ldots, y_s)$. For example, in this notation, for any Borel set B of $\mathbb{R}^p$ and any $\xi_1, \xi_2, \xi_3, \xi_4$ in $\mathbb{R}^p$, $\Psi_{(5)|(1,2,3,4)}(B|\xi_1, \xi_2, \xi_3, \xi_4)$ is the probability that $x_5 \in B$ given that $x_j = \xi_j$, $j = 1, 2, 3, 4$, and given $(y_1, \ldots, y_s)$. Since $y_1, \ldots, y_s$ are kept fixed, we will not make explicit the dependence on them. To simplify the notation, we will also write $\Psi_{5|1,2,3,4}$ instead of $\Psi_{(5)|(1,2,3,4)}$, and Ψ_s instead of $\Psi_{s|\varnothing}$. Furthermore, Ψ, without subscripts, will denote the distribution of $(x_1, \ldots, x_s)$ given $(y_1, \ldots, y_s)$. The following relation holds in general:

$$\Psi(d\xi_1, \ldots, d\xi_s) = \Psi_s(d\xi_s) \prod_{j=1}^{s-1} \Psi_{j|j+1,\ldots,s}(d\xi_j | \xi_{j+1}, \ldots, \xi_s). \tag{2.35}$$

This suggests that sampling from Ψ can be done by first drawing x_s from Ψ_s and then drawing each x_j from $\Psi_{j|j+1,\ldots,s}(|x_{j+1}, \ldots, x_s)$, $j = s-1, \ldots, 1$. From the Markovian structure implied by (2.34), it follows that, for any j,

$$\Psi_{j|j+1,\ldots,s}(d\xi_j | \xi_{j+1}, \ldots, \xi_s) = \Psi_{j|j+1}(d\xi_j | \xi_{j+1}). \tag{2.36}$$

Therefore we can rewrite (2.35) as

$$\Psi(d\xi_1, \ldots, d\xi_s) = \Psi_s(d\xi_s) \prod_{j=1}^{s-1} \Psi_{j|j+1}(d\xi_j|\xi_{j+1}). \tag{2.37}$$

The problem is then transformed in one of sampling from Ψ_s and from $\Psi_{j|j+1}$, $j = s-1, \ldots, 1$. Since the joint distribution of $(x_1, \ldots, x_s)$ and $(y_1, \ldots, y_s)$ is Gaussian, all the conditional and marginal conditional distributions we are interested in are Gaussian, and therefore they are completely determined by their first two moments. Obtaining the mean and variance of Ψ_s using Kalman filter is straightforward, see Harrison and Stevens (1976), West and Harrison (1997). On the other hand, the mean and variance of $\Psi_{j|j+1}$ can be found by considering the observation equation

$$\begin{bmatrix} y_j \\ x_{j+1} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{H} \end{bmatrix} x_j + \begin{bmatrix} v_j \\ w_{j+1} \end{bmatrix}, \tag{2.38}$$

using as prior mean and variance of x_j the one obtained from the Kalman filter.

Some numerical problems may arise during the computations of the inverse matrices needed by the Kalman filter algorithm. In fact the high correlation between neighbouring g_j's can cause most standard numerical linear algebra packages, such as LAPACK and IMSL, to detect a singularity in some of the matrices involved. As a matter of fact, all the matrices involved in the computations are nonsingular, from a theoretical point of view, and one way to overcome this difficulty would be to perform the computations with a higher degree of accuracy. In our implementation of the Markov chain we use, for this portion of the program, the package MIRACL, available by ftp at `ftp.compapp.dcu.ie/pub/crypto/`, which supports a real number arithmetic with any prespecified degree of accuracy.

2.5 A note about the convergence of the Markov chain.

We want to prove that the Markov chain defined in this chapter converge to the posterior distribution. Each full conditional distribution, or the Metropolis step used instead of it, defines a Markov transition kernel. The product of all these kernels gives another Markov transition kernel, which is the one we must consider when we are interested in the convergence issue. More specifically, let Θ be the parameter space, with points $\theta = (m, \sigma_G^2, \gamma, d, K_1, \ldots, K_s, g_1, \ldots, g_s)$ and σ-algebra $\mathcal{S}_\Theta$, and let G_m [G_σ, G_K, G_g] be the transition kernel corresponding to the full conditional distribution of m [σ_G^2, $\mathbf{K}$, $\mathbf{G}$] and M_γ [M_d] the transition kernel corresponding to the Metropolis step for γ [d]. All these kernels are defined on Θ, even if they 'move' only one coordinate. Now define a global Markov transition kernel $\boldsymbol{K}$ by putting, for each $\theta \in \Theta$ and $A \in \mathcal{S}_\Theta$:

$$\boldsymbol{K}(\theta, A) = \int \cdots \int G_g(\xi_1, A) G_K(\xi_2, d\xi_1) M_d(\xi_3, d\xi_2) M_\gamma(\xi_4, d\xi_3) G_\sigma(\xi_5, d\xi_4) G_m(\theta, d\xi_5) \tag{2.39}$$

Since the posterior distribution is an invariant measure for each kernel in the product, the same is true for $\boldsymbol{K}$. According to Theorem 1 in Tierney (1994), to prove that the Markov chain converges in total variation norm to the posterior distribution, we need to show that $\boldsymbol{K}$ is irreducible and aperiodic. Since each Gibbs or Metropolis step can move the relevant coordinate of the parameter without restrictions in the support of the corresponding marginal distribution of the posterior, from each point the chain can move in one step to any other point in the support of the posterior. This implies both irreducibility and aperiodicity.

Chapter 3

Regression

3.1 Modelling Nonstationarity in the Mean.

In the previous chapter we presented a model for a stationary process with a possible long-memory behaviour. Now we want to relax the hypothesis of stationarity. More precisely, we will allow the mean of the process to vary over time, in a linear fashion with respect to a fixed vector parameter. Another way of looking at this model is to think of it as a linear regression model with stationary errors. Let

$$Z_t = F_t\beta + X_t, \quad t \in \mathbb{Z} \tag{3.1}$$

be the process we are interested in. Here, $\mathbf{X} = (X_t)$ is a centered stationary process which can be described by the model introduced in the previous chapter, β is a r-dimensional parameter, and F_t is , for each t, a $1 \times r$ matrix of covariates. The simplest use of this extended model is to consider a linear trend for the mean of the process:

$$\mathbb{E}(Z_t) = \beta_0 + \beta_1 t. \tag{3.2}$$

The introduction of a regression term in the model is clearly a big step forward from the point of view of the applications. To the best of our knowledge, no Bayesian analysis of model (3.1) has been developed so far, although within a frequentist

framework quite a few works exist, dealing with the problem of linear regression with long-memory errors. See Yajima (1988), Yajima (1991), Koul and Mukherjee (1993), Smith and Chen (1996). A good review can be found in Beran (1994b, Chap. 9). For an application to some climatological data sets see Smith (1993).

3.2 The Likelihood.

The likelihood can be obtained in the following way. First, on can notice that, from (3.1), $Z_t - F_t\beta$ is a centered stationary process. Therefore, using Whittle approximation (1.9), the likelihood can be written approximately, up to a multiplicative constant, as

$$\prod_{j=1}^{N} f(\omega_j)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{N}\frac{I_N^X(\omega_j)}{f(\omega_j)}\right\}, \tag{3.3}$$

where I_N^X is the periodogram of the process $\mathbf{X}$, i.e.

$$I_N^X(\omega) = \frac{1}{N\pi}\left|\sum_{t=1}^{N}(Z_t - F_t\beta)e^{-it\omega}\right|^2. \tag{3.4}$$

The parameters of the model are now all the ones that describe the spectral density of $\mathbf{X}$, plus the regression parameter β.

3.3 Prior Distribution for β.

Let us make the following assumptions about the joint prior distribution of the parameters:

(1) the parameter β is, a priori, independent of all the parameters which describe the spectral density of $\mathbf{X}$;

(2) the prior distribution for the spectral density is the one described in Chapter 2.

Then we are left to specify the marginal prior distribution of β. We will assume a flat improper distribution. If one does not want to use improper distributions, an alternative would be to use a Gaussian distribution. We will consider in detail only the first case, the modifications needed in the Markov chain Monte Carlo sampling scheme in the second case being straightforward.

3.4 Sampling from the posterior distribution.

The posterior distribution for the regression model (3.1), with a flat prior distribution for the regression parameter β, has exactly the form (2.9), with the only difference that now the y_j's are the periodogram ordinates at the Fourier frequencies of the centered process, which are given by formula (3.4). If an analytical exam of the posterior distribution was very difficult for the model without regression introduced in last chapter, now, with the addition of the regression parameter in the model, things are even worst. Fortunately, we can use also for this extended model a Markov chain Monte Carlo method to draw a sample from the posterior distribution of the parameters. We propose to use a successive substitution sampling scheme. In the rest of the section we describe the full conditional distributions that we need.

Full conditional distributions of the spectral parameters. First of all, note that for any fixed value of β one can sample from all the other parameters by centering the observation process, computing the periodogram ordinates at the Fourier frequencies and applying the techniques described in Section 2.3. In fact, with respect to the spectral parameters, the posterior distribution has the same analytical form (2.9) as in the case without regression described in last chapter. Therefore, to perform a successive substitution sampling, one only needs to find the full conditional distribution of β.

Full conditional distributions of β. Assuming a flat prior, the full conditional distribution of β is proportional to the likelihood (3.3). One can observe that β only appears in a quadratic form at the exponent: therefore its full conditional distribution is Gaussian. Let us now compute in some detail what the mean and variance of that distribution are. Define

$$a(\omega) := \frac{1}{\sqrt{N\pi}} \sum_{t=1}^{N} Z_t e^{-it\omega}, \tag{3.5}$$

$$b(\omega) := \frac{1}{\sqrt{N\pi}} \sum_{t=1}^{N} F_t e^{-it\omega}; \tag{3.6}$$

so that the periodogram can be written in terms of a and b as

$$I_N^X(\omega) = \left|a(\omega) - b(\omega)\beta\right|^2. \tag{3.7}$$

Then the logarithm of the density of the full conditional distribution of β is given by:

$$\begin{aligned}
&-\sum_{j=1}^{N} \frac{\left|a(\omega_j) - b(\omega_j)\beta\right|^2}{2f(\omega_j)} \\
&= \ldots - \sum_{j=1}^{N} \left\{ \frac{\beta^t b(\omega_j) b^*(\omega_j)\beta}{2f(\omega_j)} - 2\frac{\mathrm{Re}\big(a^*(\omega_j)b(\omega_j)\big)}{2f(\omega_j)} \right\} \\
&= \ldots - \frac{1}{2}\beta^t \mathrm{Re}\left(\sum_{j=1}^{N} \frac{b(\omega_j)b^*(\omega_j)}{f(\omega_j)}\right)\beta + \mathrm{Re}\left(\sum_{j=1}^{N} \frac{\bar{a}(\omega_j)b^t(\omega_j)}{f(\omega_j)}\right)\beta,
\end{aligned}$$

where " * " denotes the transpose conjugate, " $\bar{}$ " the conjugate, " Re() " the real part, and the dots stand for something which does not depend on β. Therefore, indicating with μ and Σ respectively the mean and the variance matrix of the full conditional distribution of β, we have

$$\Sigma^{-1} = \mathrm{Re}\left(\sum_{j=1}^{N} \frac{b(\omega_j)b^*(\omega_j)}{f(\omega_j)}\right), \tag{3.8}$$

and

$$\Sigma^{-1}\mu = \mathrm{Re}\left(\sum_{j=1}^{N} \frac{\bar{a}(\omega_j)b'(\omega_j)}{f(\omega_j)}\right). \tag{3.9}$$

From the above relations Σ and μ can be easily computed. Note that the $a(\omega_j)$'s and $b(\omega_j)$'s can be computed once and for all before starting the Markov chain, since they depend only on known quantities. An highly efficient way of performing the computation is to use the fast Fourier transform.

Chapter 4

Forecasting

4.1 Including Future Observations in the Model.

In the Bayesian approach one defines a joint probability distribution for the parameters of interest and the data, thought of as observable random variables, before being observed. The customary way to proceed is to define the conditional distribution of the observables given the parameters (the 'model') and the marginal distribution of the parameters (the 'prior'). This uniquely defines the joint distribution. Then, after seeing the data, and assuming the parameter space is sufficiently regular (e.g. a Borel space), one bases the inference about the parameters on their conditional distribution, given the data. In many practical applications, being given the observations $X_1, \ldots, X_N$, it is of interest to make inference also on the future behaviour of the series. From a Bayesian point of view, the way of doing that is to define a joint probability distribution for the observations, the parameters and the future realisations of the series, and then consider the conditional distribution of the future realisations and the parameters, given the data. Assume one is interested in the future observations $X_{N+1}, \ldots, X_{N+M}$. This chapter is devoted to explaining how one can obtain a sample from the conditional distribution of the future observations

and the parameters, given the data, using a Markov chain Monte Carlo approach. Throughout this chapter we will make the additional distributional assumption that the process $(X_t)_{t\in\mathbb{Z}}$ be conditionally Gaussian, given the parameters.

The idea is again that of successive substitution sampling. Assume for the time being that, in addition to the data $X_1, \ldots, X_N$, one also knows the value of $X_{N+1}, \ldots, X_{N+M}$. Then he or she can draw from the conditional distribution of the parameters, given $X_1, \ldots, X_{N+M}$, using the technique described in the previous chapters. At this point, to conclude the cycle, one has to draw from the distribution of $X_{N+1}, \ldots, X_{N+M}$, given the parameters and $X_1, \ldots, X_N$.

4.2 Sampling from the future observations.

To sample from the full conditional distribution of the future observations $X_{N+1}, \ldots, X_{N+M}$, assume that the data and the parameters of the model are given. To be precise, we assume that we are given, in addition to $X_1, \ldots, X_N$, the regression parameter β, γ, σ_ϵ^2, d, and $g_1, \ldots, g_s$, where now $s = \lfloor (N+M)/2 \rfloor$. Since the covariance function of the process is given by (1.4), we can numerically evaluate it at a finite number of t's. For a reason to be explained later, in order to sample from the future observations we need $R(t)$ for $t = 0, \ldots, N+M$. Below we give some details on how the computations can be carried out. To take advantage of the fast Fourier transform algorithm, one needs to have the spectral density at $N+M$ points instead of s. The values of g at the additional points are obtained by local polynomial interpolation, based on four neighbouring points. To obtain a good approximation in the computation of the covariances, some care must be taken near the zero frequency when $d \neq 0$. In fact, in this case, we can write

$$R(t) = \sum_{j=1}^{N+M} \int_{\lambda_{j-1}}^{\lambda_j} \cos(k\lambda)\lambda^{-d} g(\lambda)\, d\lambda, \tag{4.1}$$

where $\lambda_j = 2\pi j/(N+M)$. Setting aside the first few values of j, if $N+M$ is sufficiently large, the approximation

$$\int_{\lambda_{j-1}}^{\lambda_r} \cos(k\lambda)\lambda^{-d}g(\lambda)\,d\lambda \approx \frac{2\pi}{N+M}\cos(k\lambda_j)\lambda_j^{-d}g(\lambda_r) \tag{4.2}$$

is usually very accurate. On the other hand, for the very low frequencies one can obtain an improved approximation as follows. Let $\nu_j = \int_{\lambda_{j-1}}^{\lambda_j} \lambda^{-d}\,d\lambda$, i.e. ν_j is the mean value of the function $(\,\cdot\,)^{-d}$ on the interval $(\lambda_{j-1}, \lambda_j)$. Then, taking into account that $\cos(k\,\cdot\,)$ and g are continuous functions,

$$\int_{\lambda_{j-1}}^{\lambda_j} \cos(k\lambda)\lambda^{-d}g(\lambda)\,d\lambda \approx \frac{2\pi}{N+M}\cos(k\lambda_j)g(\lambda_j)\nu_j \tag{4.3}$$

gives better results for small values of j. From the covariances, obtained using (4.2), (4.3) and the fast Fourier transform, we are then able to directly compute the conditional means and variances of each of the conditional normal distributions for the X_{N+r}. We use the efficient Durbin-Levinson recursions to do this (see Ramsey (1974)). Let ρ_j be the correlation between X_1 and X_{j+1}, given the current value of the parameters. If one sequentially solves, for $k = 1, \ldots, N+M-2$, the recurrence relations

$$\Phi_1 = \alpha_1^{(1)} = \rho_1 \tag{4.4a}$$

$$\sigma_1^2 = 1 - \Phi_1^2 \tag{4.4b}$$

$$\Phi_{k+1} = \alpha_{k+1}^{(k+1)} = \frac{\left(\rho_{k+1} - \sum_{j=1}^{k}\alpha_j^{(k)}\rho_{k+1-j}\right)}{\sigma_k^2} \tag{4.4c}$$

$$\alpha_j^{(k+1)} = \alpha_j^{(k)} - \Phi_{k+1}\alpha_{k+1-j}^{(k)}, \qquad (j = 1, \ldots, k) \tag{4.4d}$$

$$\sigma_{k+1}^2 = \sigma_k^2(1 - \Phi_{k+1}^2), \tag{4.4e}$$

then the conditional mean and variances of $X_{n+1}, \ldots, X_{n+m}$ are given by the expressions:

$$\mathbb{E}(x_{k+1}|x_1, \ldots, x_k) = h'_{k+1}\beta + \sum_{j=1}^{k} \alpha_j^{(k)}(x_{k+1-j} - h'_{k+1-j}\beta), \tag{4.5}$$

$$\mathrm{Var}(x_{k+1}|x_1, \ldots, x_k) = v_{1,1}(1 - \Phi_1^2) \cdots (1 - \Phi_k^2), \tag{4.6}$$

$k = N, \ldots, N + M - 1$. Based on (4.5) and (4.6), one can easily generate sequentially each X_{N+r}, $r = 1, \ldots, M$, from its conditional distribution, given the parameters and the past up to time $N + r - 1$. Observe that these are all Gaussian distributions.

From the output of the successive substitution sampling one can compute summaries of the future of the series or look at plots of individual samples from the distribution of the future.

Chapter 5

Applications

5.1 A global temperature data set.

We have investigated various applications of the methodology and simulation techniques described in the previous chapters to the analysis of a climatological data set. The data, coming from a report of the Intergovernmental Panel on Climate Change, consist 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 1,628, is plotted in Figure 5.1, together with a five-year moving average (solid line).

5.2 Stationary model.

First, we want to perform an analysis using the basic model without trend introduced in Chapter 2. The focus of the analysis here is on detecting a long-memory behaviour in the data and in estimating the long-memory parameter d. A comparison with different estimates of d which can be obtained using alternative methods will be presented at the end of the section. The prior for the memory parameter d has been

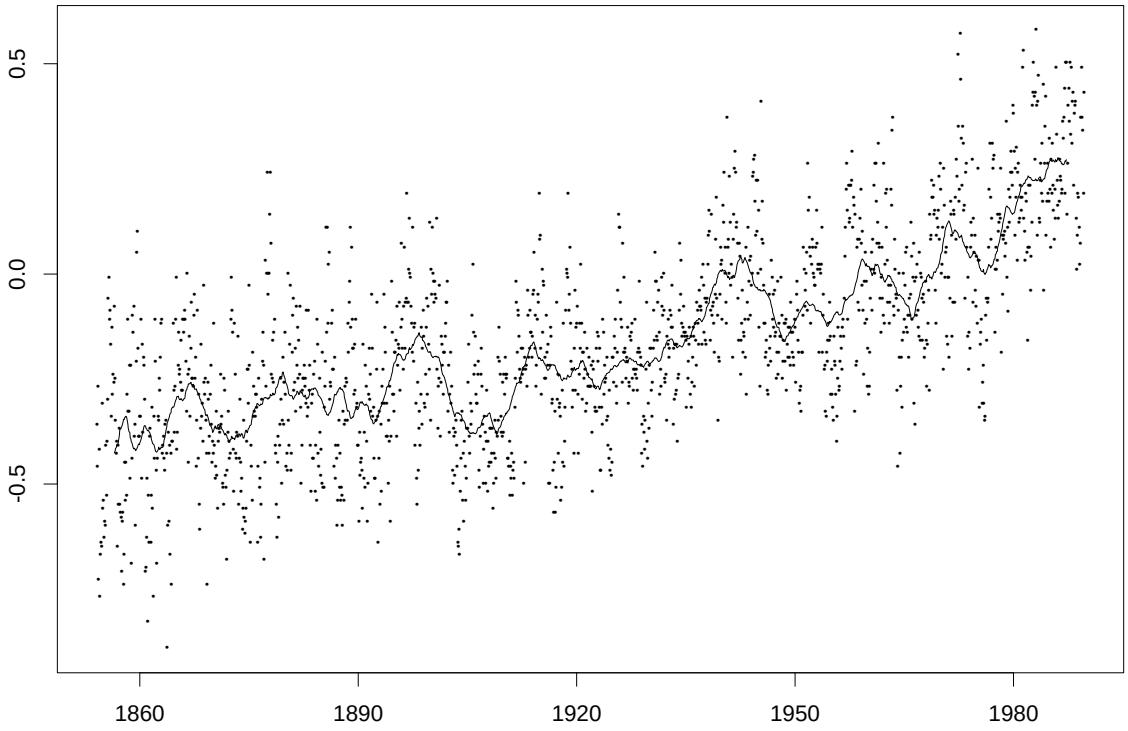

Figure 5.1: Temperature data.

taken of the form given in (2.8), with $a_d = b_d = 1$ and $\alpha = 0.4$. It follows that, a priori, Prob$(d = 0) = 0.4$ and E$(d|d > 0) = 0.5$. 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 $a_\gamma = 10^9$, $b_\gamma = 10^{-7}$, $a_G = b_G = 10$. All the posterior estimates are based on 3,500 iterations of the Markov chain. The chain seems to reach the stationary state very fast, but we use nonetheless a burn in period of 1,500 iterations. We tried to run the chain with different starting points, obtaining essentially the same results. In Figure 5.2 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

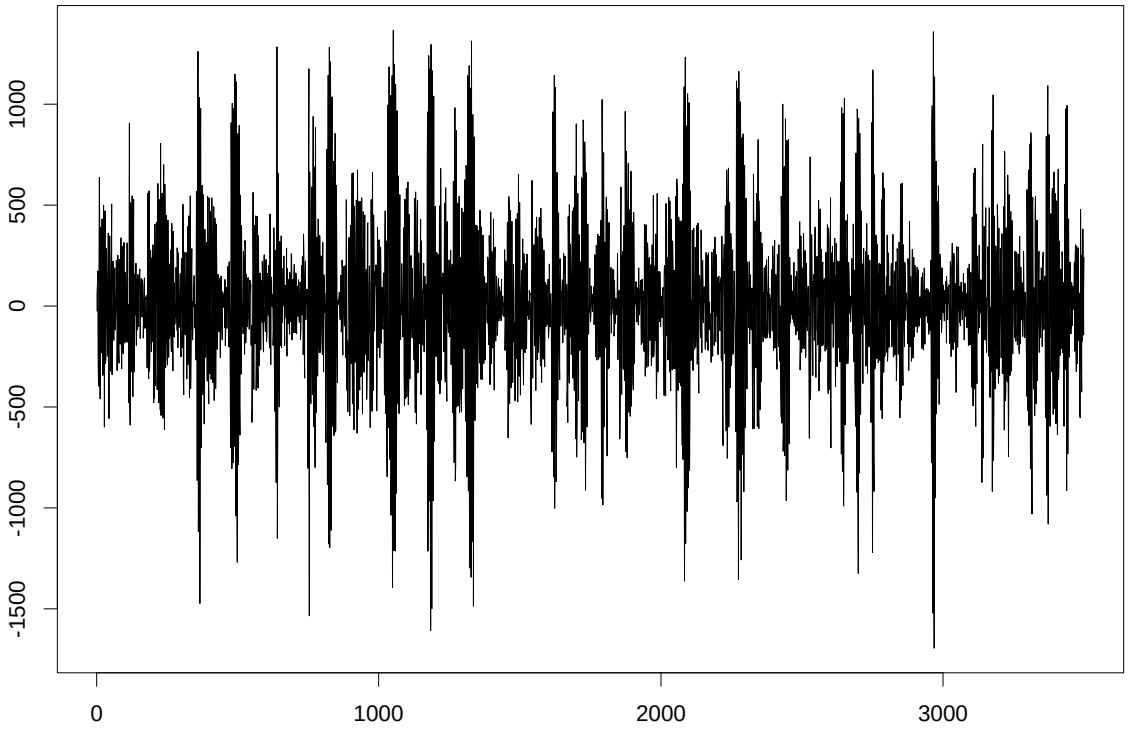

Figure 5.2: Log posterior density.

are not in contrast with the assumed approximate stationarity of the chain. The acceptance rate for the Metropolis steps, for the parameters d and γ, were about 90%.

Figure 5.3 shows, on a logarithmic scale, the posterior estimate of the spectral density f (solid line) together with posterior estimate of its continuous part g (dashed line). It can be seen that the high weight given to the low frequencies is explained not only by the term λ^{-d} in the spectral density, but also by an increase in the function g as λ goes to zero. Samples from the simulated values of the parameter $\mathbf{G}$ are shown in Figure 5.4. The plots should give a rough idea about the variability in the posterior distribution of $\mathbf{G}$. The smoothness of the realisations of $\mathbf{G}$ can be

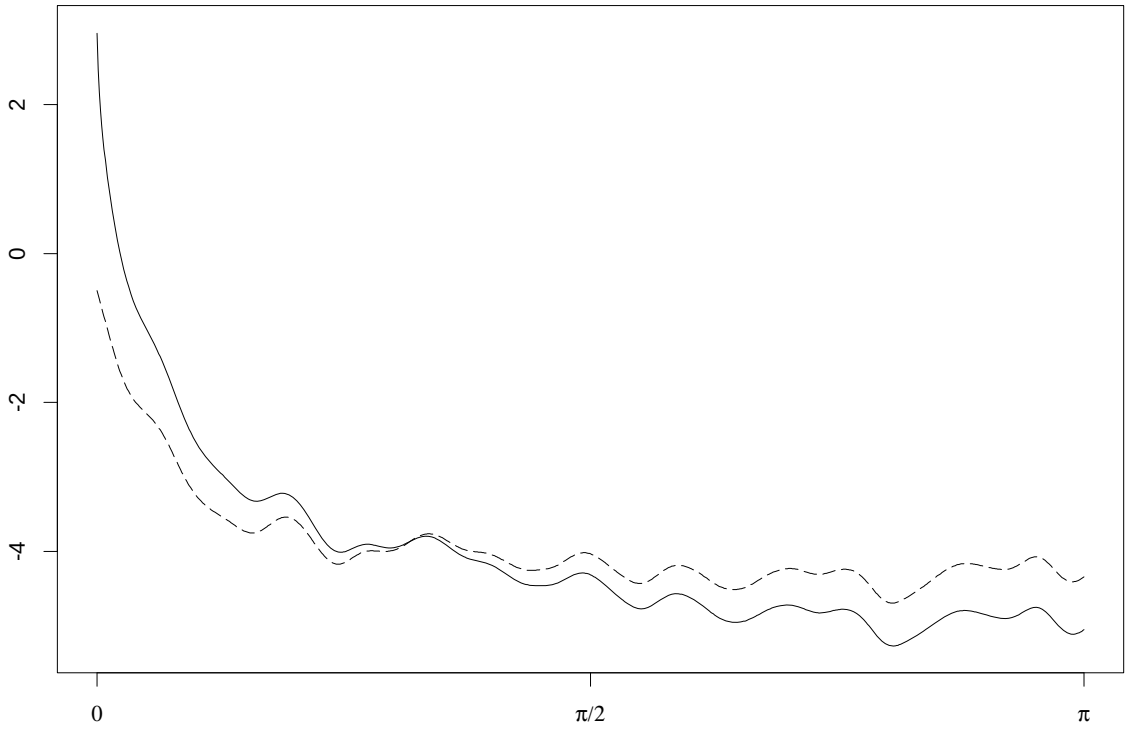

Figure 5.3: Estimated spectral density (log scale).

controlled by acting on the prior distribution of the parameter γ.

Before going to the long-memory parameter d, let us show some plots from the Markov chain Monte Carlo simulation, relative to the 'hyperparameters' of our model. Figure 5.5 shows the simulated values of the parameter m. Its estimated mean and standard deviation are -3.6 and 0.4, respectively.

In Figure 5.6 a plot of the simulated values of the parameter σ_G^2 is provided, on a logarithmic scale. Posterior mean and standard deviation of the posterior distribution of this parameter, estimated from the Markov chain Monte Carlo simulation, are 1.58 and 2.81, respectively. From Figure 5.6 one can see that the distribution is skewed to the right and has a relatively havy tail. The central 90% probability interval is

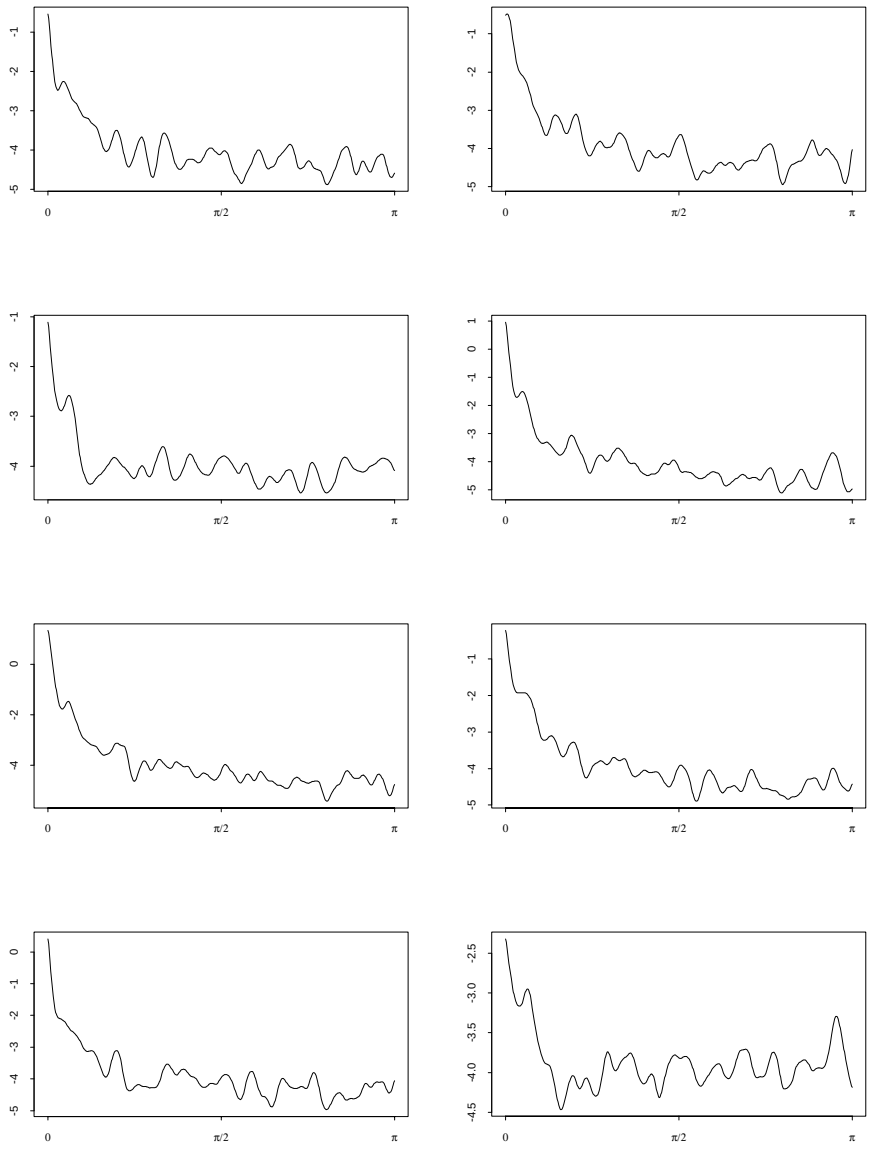

Figure 5.4: Simulated values of the g_j's.

estimated to be (0.25,4.15).

Figure 5.7 shows the simulated values of the smoothness parameter γ. Mean and standard deviation of its posterior distribution are 0.96 and 0.01 respectively.

Let us turn now to the analysis of the long-memory parameter d. Figure 5.8 shows the simulated values of d. Since its posterior distribution has a positive mass

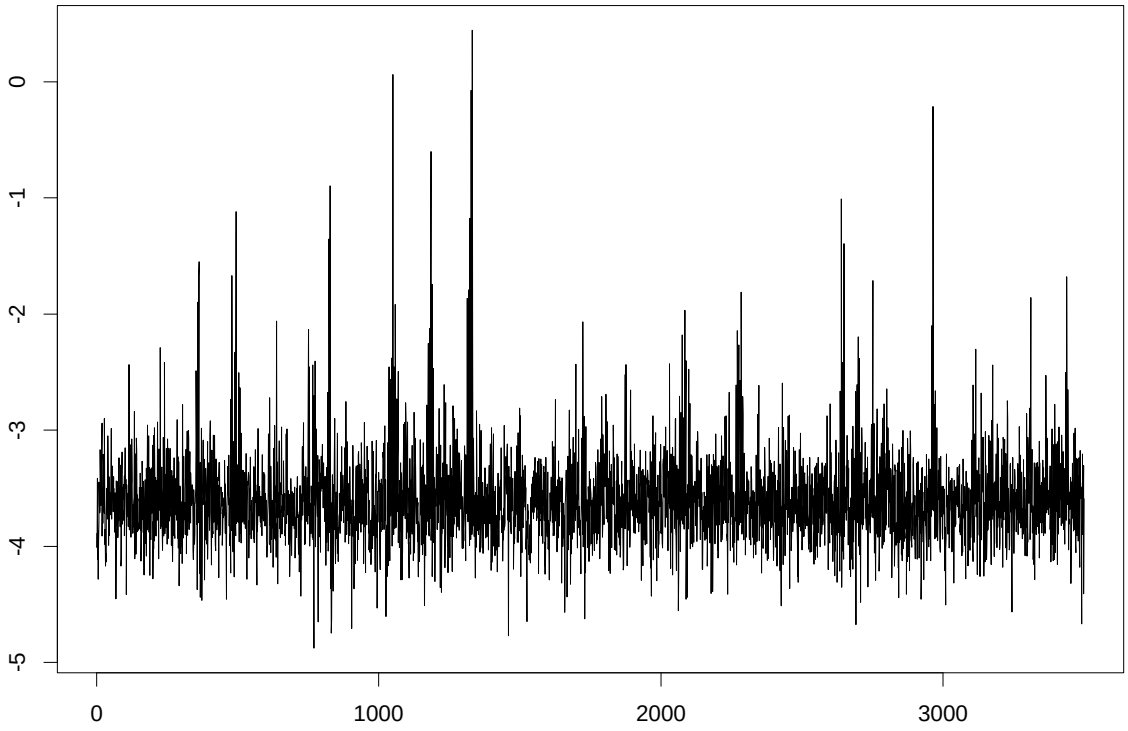

Figure 5.5: Simulated values of m.

on the point zero, it is useful in the analysis to keep separate the realisations of d which are zero from the other. The histogram of the positive values of d sampled by the Markov chain (Figure 5.9) should give an idea of the posterior distribution of d, given $\{d > 0\}$. A five-number summary of the same distribution is provided in Table 5.2. A posteriori, $\text{Prob}(d = 0) = 0.05$ and $\mathbb{E}(d|d > 0) = 0.63$. It follows that the posterior mean of d is 0.60. Let us recall that the Bayes factor in favor of an hypothesis is the ratio of the posterior odds to the prior odds, and it is customarily interpreted as the evidence provided by the data in favor of the hypothesis (see Kass and Raftery (1995)). It is quite easy to find the Bayes factor in favor of $\{d > 0\}$ from the output of the simulation and the prior specification: in fact the prior odds ratio

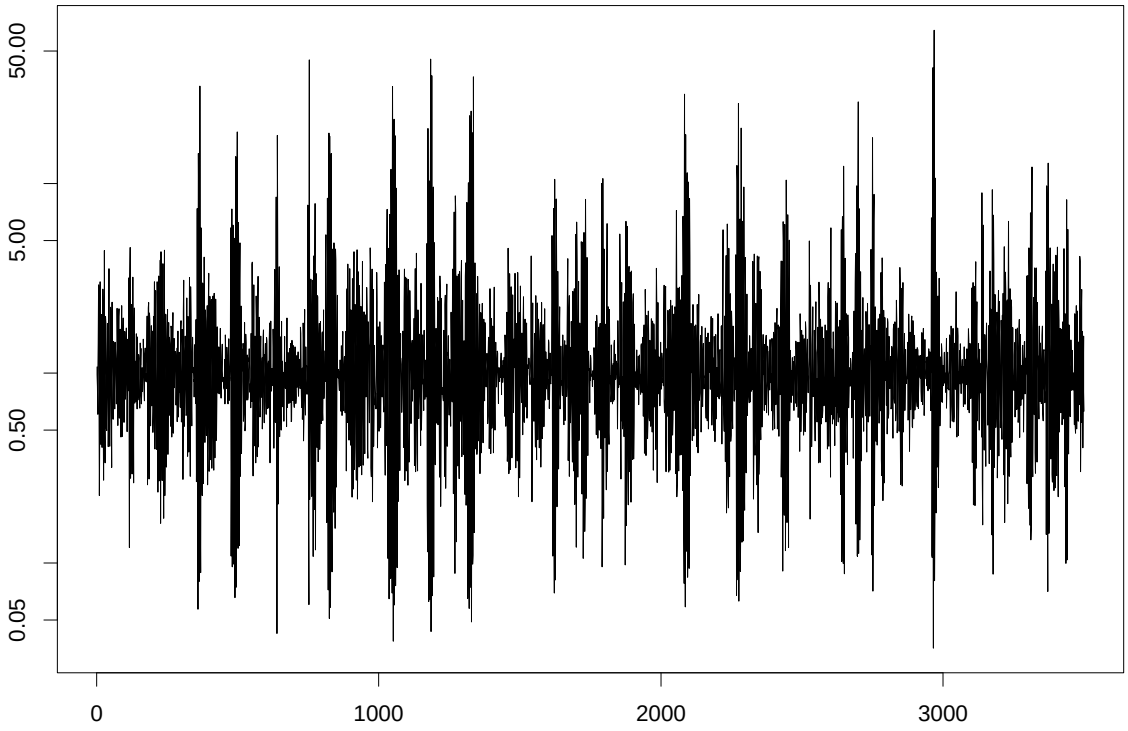

Figure 5.6: Simulated values of σ_G^2 (log scale).

is just $(1-\alpha)/\alpha$, while the Monte Carlo estimate of the posterior odds is given by

$$\frac{\#\{d^{(r)} : d^{(r)} > 0\}}{\#\{d^{(r)} : d^{(r)} = 0\}}, \tag{5.1}$$

where $d^{(1)}, \ldots, d^{(3500)}$ are the values of d sampled by the Markov chain. The numerical value of the Bayes factor turns out to be 12.6. This means precisely that a posteriori the odds in favor of the hypothesis $\{d > 0\}$ are 12.6 times the prior odds. It appears also from the posterior probability of $\{d > 0\}$ that this hypothesis is strongly favoured

Min.	1st Qu.	Median	3rd Qu.	Max.
0.02	0.46	0.61	0.75	0.99

Table 5.1: Summary of the posterior distribution of d.

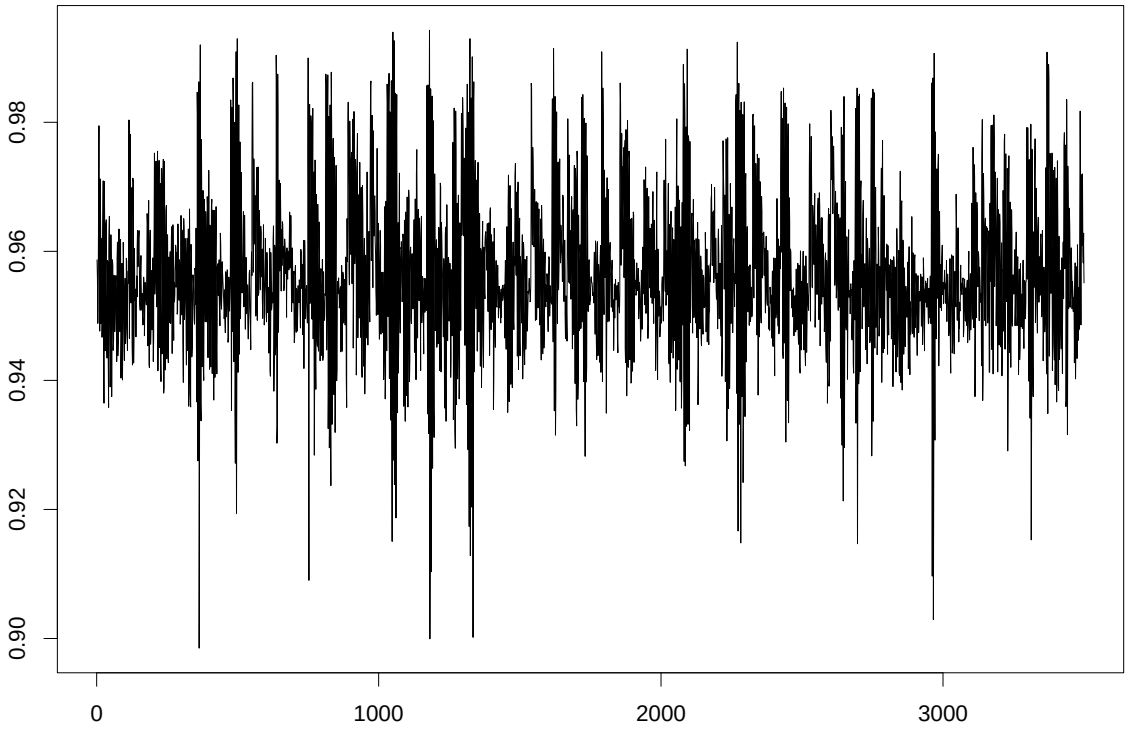

Figure 5.7: Simulated values of γ.

by the data.

We can compare now our Bayesian estimate of d with the estimate one can obtain using other methods that can be found in the literature. The different estimates considered here are reported in Table 5.2. The first one is obtained by the method proposed in Geweke and Porter-Hudak (1983). To use their estimator we must choose the number m of Fourier frequencies to include in the computation of the estimate. Taking $m = \lfloor \sqrt{N} \rfloor$, as suggested in the original paper, we obtain $\hat{d}_m = 0.56$. More recent studies show that the optimal choice for m is of order $O(N^{4/5})$, see Hurvich, Rohit and Brodsky (1996). For our data, with $m = \lfloor N^{4/5} \rfloor$, we get $\hat{d}_m = 0.57$. Disregarding also the first l Fourier frequencies in the computations, as suggested in

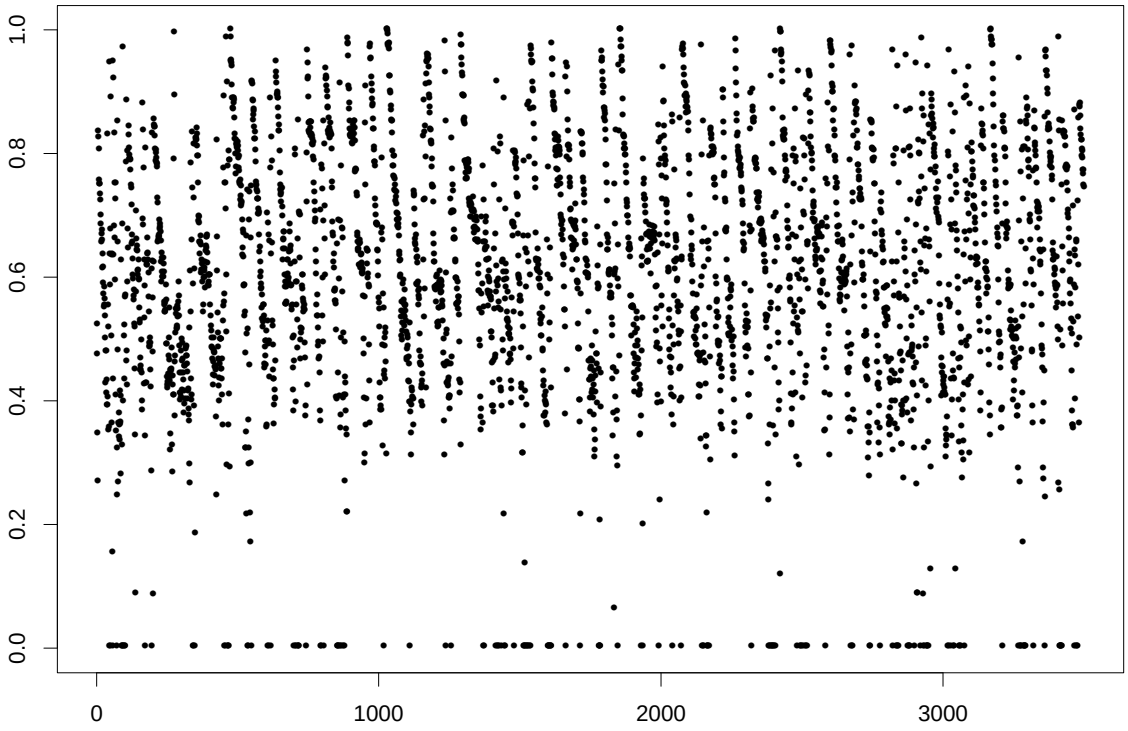

Figure 5.8: Simulated values of d.

Robinson (1995b), results in a small difference in the estimate of d. For example, keeping $m = 370$, for $l = 5$, 10, 20, 50, we get respectively $\hat{d}_{l,m} = 0.56$, 0.56, 0.60, 0.54. An other alternative approach to the estimation of the memory parameter is contained in Robinson (1995a). Using the method proposed in that paper we get the estimate $\hat{d_s} = 0.48$. All the estimates, except for the last one, are pretty close to one another, and indicate that there is in the series a long-memory behaviour with long-memory parameter around 0.6.

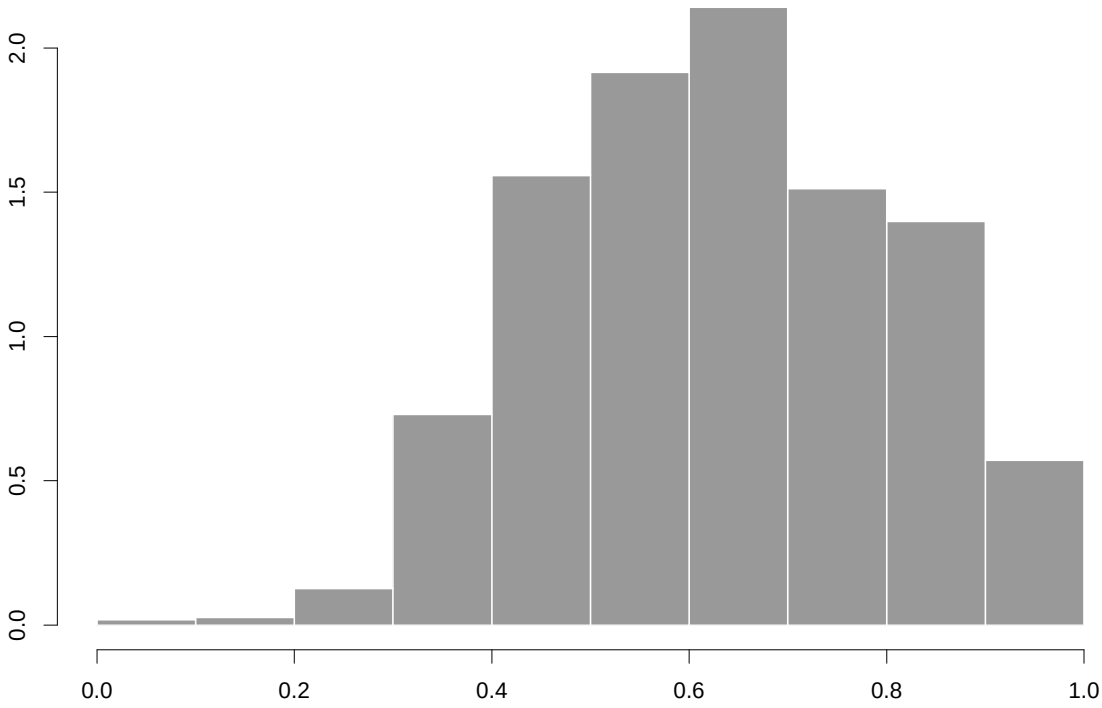

Figure 5.9: Posterior distribution of d, given $\{d > 0\}$.

5.3 Analysis of the trend

In the previous section we assumed that our data come from a series with constant zero mean. Now we include in the model also a linear trend, i.e. we assume that

$$\mathbb{E}(X_t) = \beta_0 + \beta_1 t. \tag{5.2}$$

More elaborate forms of linear regression could be used by including higher order powers of t, or different functions, e.g., spline functions. Clearly, the inclusion of a trend in the model may have an effect on the inference about the other parameters, and in particular the long-memory parameter d. In fact, for example, in the data we are using, the increasing behaviour at the end of the series can be explained by

a nonstationary mean or by a long-term fluctuation of the random component, or by a combination of the two. The choice of the model should be dictated by the scientific underlying problem. Here, we will use, for illustrative purposes, the simple linear regression (5.2). The prior for the regression parameter is the one described in Section 3.1, and the specification of the prior distribution for the spectral density is the same of the previous section. The Monte Carlo estimates are based, also in this section, on a run of length 3500 of the Markov chain, following 1500 burn in iterations. Table 5.3 reports a five-number summary of the posterior distributions of β_0 and $120\,\beta_1$. Note that $120\beta_1$ is the increase in the average temperature in centigrade degrees per decade. Figure 5.10 reproduce Figure 5.1, with the addition of the posterior estimate of the linear trend. The regression line obtained with the methodology developed in the present work is very close to the one that can be obtained by the method of least squares. What is very different is the uncertainty about the estimates, which is summarised in Table 5.4. From the table one can infer for example, assuming normality of the posterior distribution of β, that, with 95% probability, the temperature increase per decade lies in the interval (0.035, 0.049).

0.60	Bayesian estimate
0.56	GPH $m = \lfloor \sqrt{N} \rfloor$
0.57	GPH $m = \lfloor N^{4/5} \rfloor$
0.56	GPH $m = 370$ $l = 10$
0.60	GPH $m = 370$ $l = 20$
0.48	P. Robinson

Table 5.2: Different estimates of the long-memory parameter.

	Min.	1st Qu.	Median	3rd Qu.	Max.
β_0	-1.23	-0.46	-0.44	-0.41	1.39
$120\,\beta_1$	0.028	0.040	0.042	0.044	0.079

Table 5.3: Summary of the posterior distributions of β_0 and $120\,\beta_1$.

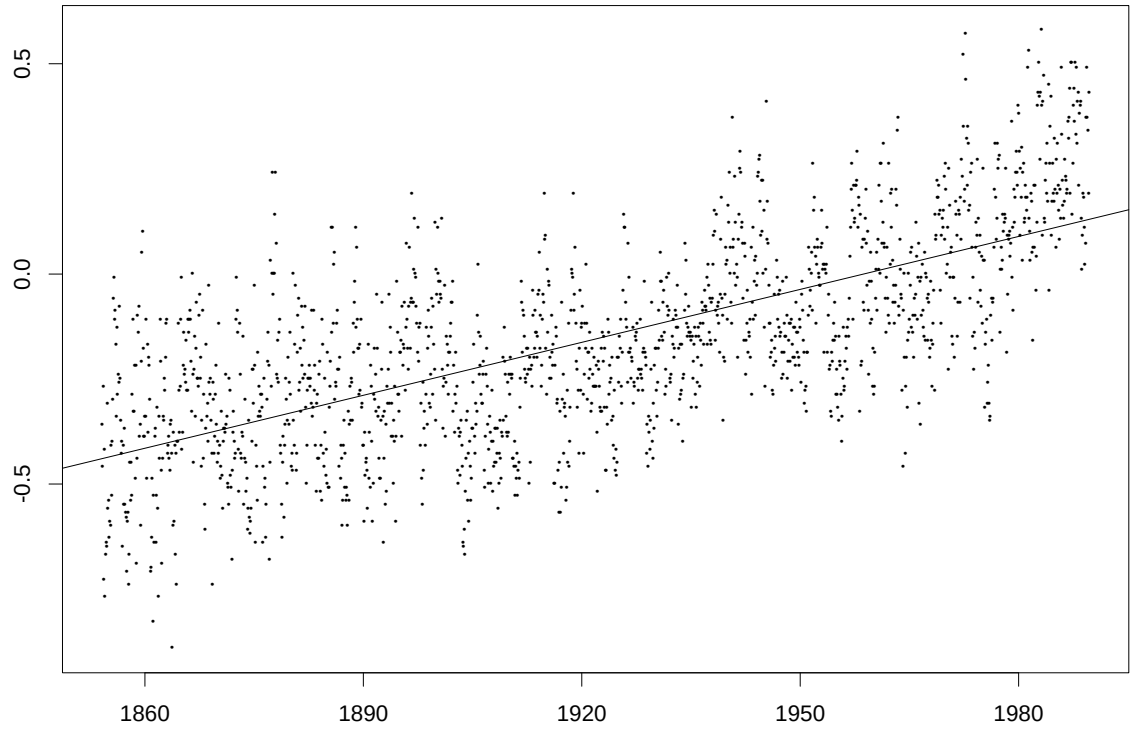

Figure 5.10: Posterior estimate of the trend

On the other hand, using least square theory, a 95% confidence interval for the same quantity is given by (0.040, 0.044).

It is interesting also to compare the inference that can be made about the long-memory parameter with the results obtained in the previous section. Now, from the simulation, we have that $\text{Prob}(d = 0) = 0.24$ and $\mathbb{E}(d|d > 0) = 0.36$, which implies

	Mean	St. Dev.	St. Dev. of LSE
β_0	-0.44	0.092	0.0087
120 β_1	0.042	0.0036	0.0011

Table 5.4: Estimate of the regression parameter

a posterior mean of d of 0.27. The Bayes factor in favor of $\{d > 0\}$ is only 2.1, quite a bit lower than for the model without trend. In fact this less strong support of the data for a long-memory behaviour and the shift of the center of the posterior distribution of d towards a lower value can be explained along the following lines. The data in Figure 5.1 clearly show some sort of increasing pattern. If one sticks to the model with zero mean, the only possible explanation for that kind of behaviour is given by a long term fluctuation in the series, which can be obtained, within the given model, only by allowing for a spectral density which puts a lot of weight on the very low frequencies. Including a trend in the model, on the other hand, amounts to allowing for a different explanation of the behaviour observed in the data, namely that the mean of the underlying process follows an increasing pattern. What happens in practice is that both these explanations are included in the posterior distribution: the posterior 'explains' the increasing pattern observed in the data by an increasing linear trend and a spectral density which puts a considerable fraction of the total variability of the series on the low frequencies.

5.4 Forecasting

We can include in our model also a segment of the future of the series, as explained in Chapter 4. We run the Markov chain again, including in the model the next 200 values of the series, corresponding to about 17 years. The Markov chain Monte Carlo approximation to the marginal posterior distributions of all the parameters did not seem to change significantly; of course, theoretically they are the same. Let us focus now on the predictive distribution. Figure 5.11 shows eight randomly selected samples from the predictive distribution of the next 200 values of the series. The first part of each plot represents the last 200 observations from the series. Observed values and simulated future values are separated by a dashed line. The plots should give an

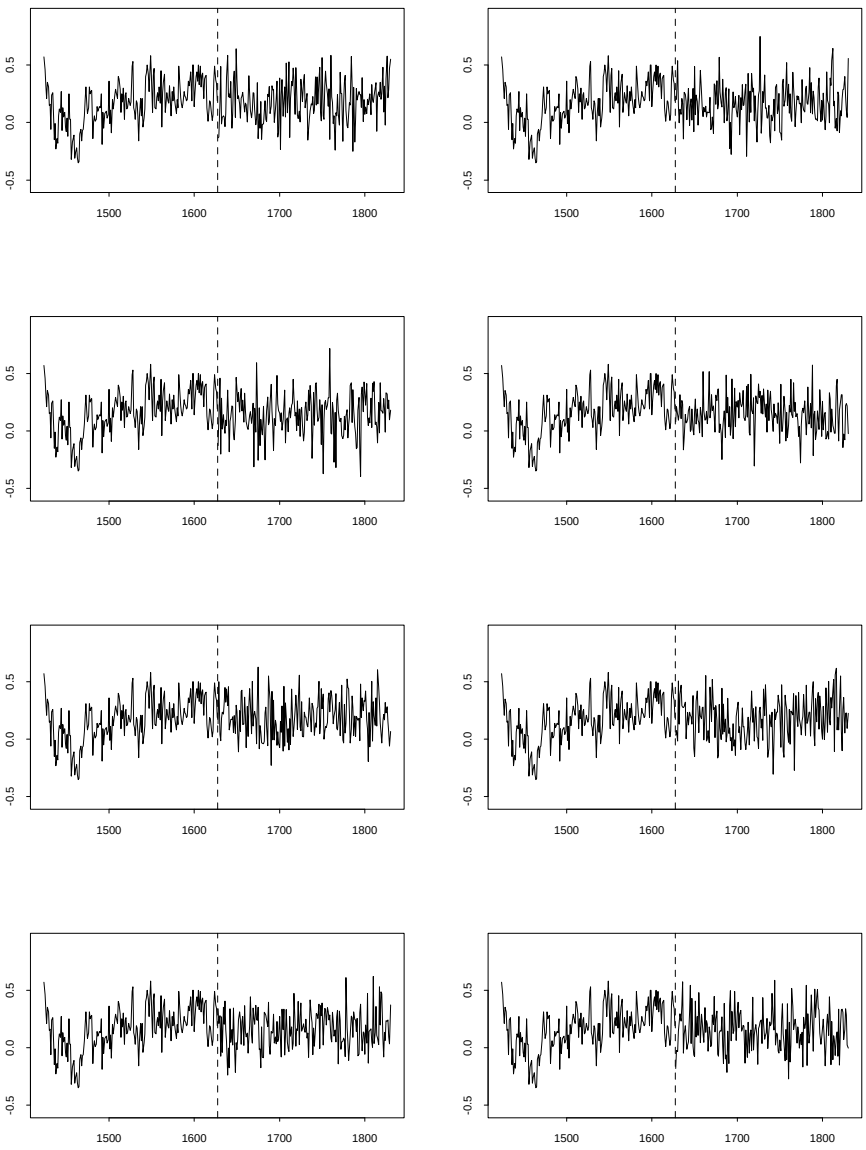

Figure 5.11: Simulated futures.

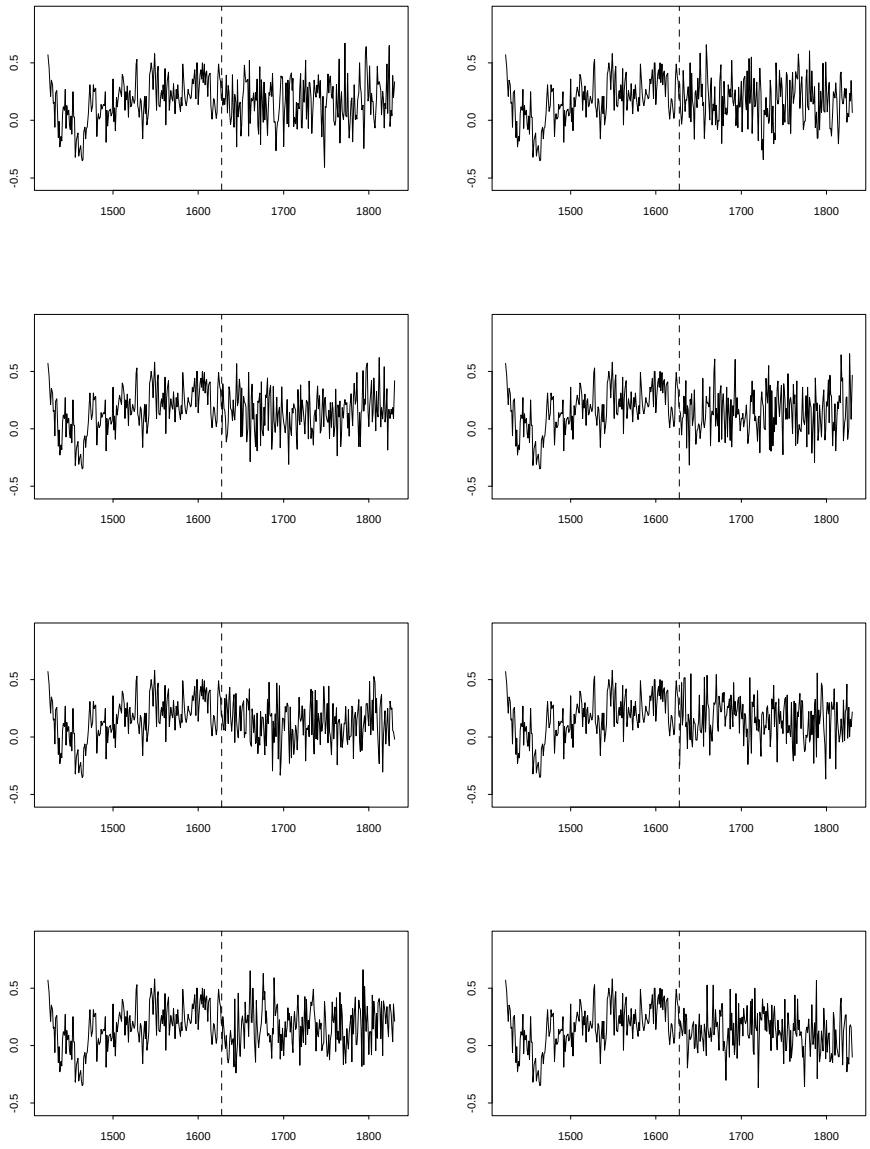

Figure 5.12: Simulated futures, from estimated spectral density.

idea of the behaviour of the series, according to the model, in the next 17 years. The impression that one can get from the plots is that, loosely speaking, the volatility of the future is higher than the volatility of the observed series. A possible explanation for that is that the model we are using, together with the prior, implies a high

volatility of the series of the observables, and the posterior/predictive distribution is not able to 'learn' about this particular aspect of the data, at least with the sample size at hand. We also tried to simulate the futures keeping the spectral density and regression parameters fixed at the value of their posterior estimate (Figure 5.12), but the results, in term of the volatility did not change substantially.

Chapter 6

Concluding remarks

We would like, in this final chapter, to discuss the main contributions to the field of time series analysis contained in this dissertation. This will also serve to highlight the difficulties encountered during the research which led to the final form of this thesis, as well as to point out some issues that remain open. In doing that, we will hint to possible directions for future research.

As far as we know, the analysis contained in Chapters 2, 3 and 4, represents the first fully Bayesian treatment of long-memory time series. The key features of our approach are contained in the words *Bayesian* and *nonparametric*. We see the Bayesian approach to Statistics not just as a mere tool to produce inferences on the parameters of a model, but rather as a way of *explicitly* incorporating a subjective prior opinion into the statistical analysis. The way we defined the class of priors in Chapter 2 reflects this attitude. We split the prior distribution into two components. Apart from mathematical convenience, this should help the statistician to incorporate his own opinion in the prior in a meaningful way. In fact, the long-memory parameter d reflects features of the observations process about which it is relatively easy to elicit an opinion. On the other hand, the elicitation of a prior distribution for the continuous part of the spectral density can be quite difficult. For

that reason we want to be able to express at least a qualitative opinion about the smoothness of the function. This is precisely the aim of a smoothness prior of the form (2.5). The difficulty of eliciting a prior distribution for the continuous part of the spectral density is also accounted for by the essentially nonparametric nature of the smoothness prior. If little can be said a priori about the spectral density, then one should try to keep the support of its prior distribution as large as possible, so that the data could complement the scarce prior information without inappropriate limitations. This is the attitude that prompted the nonparametric approach and the resulting priors, whose support is essentially the set of all continuous functions on the torus. Many generalisations of this prior can be envisaged, the most straightforward, but probably not the most useful, being the substitution of the AR(p) form (2.5) with a more general ARMA form. In this case the Markov chain Monte Carlo simulation could be carried over with only minor modifications, since ARMA models too can be written as Dynamic Linear Models; see West and Harrison (1997). More interesting extensions might be obtained incorporating a less vague opinion in the smoothness prior. For example, how could one incorporate the opinion that the weight put by $\mathbf{G}$ on the high frequencies is less than the weight put on the low frequencies? We saw in Section 5.4 that the predictions coming out from our Markov chain Monte Carlo simulation were more volatile than the observed series. Next time we have to analyse an analogous data set, we would like to incorporate in our prior a smaller weight for the high frequencies. One way of doing so is to work on the parameter m, using a sequence $m_1, \ldots, m_s$ of values, one for every Fourier frequency, to express the mean of the function $\mathbf{G}$, instead of one single m. Another way is to model $\mathbf{G}$ as a nonstationary process, such as, for example, an integrated AR(p) model.

The ability of making predictions is a very important aspect of time series analysis. We tried to give an answer to this problem, in the context of long-range dependence,

from a Bayesian viewpoint. An essential feature of our approach to the prediction problem is that it provides samples from the predictive distribution, as opposed to estimates. Estimates and summary statistics can always be trivially computed from the output of the simulation. The advantage of having a set of possible future realisations, from which one can get a precise feeling of how the series will look like in the future, is clear. Although we feel that this is the right direction to go, much work remains to be done. One direction just noted is in devising priors that are better able to adapt to the observed behaviour of the series at different frequency levels. Also, it would be nice to take into account that many real world time series may be better modelled with non-Gaussian distributions. Although this is not relevant for the posterior distribution of the spectral density, as long as we use the Whittle approximation to the likelihood function, it may be crucial when dealing with predictions.

While the prediction side of our model needs further investigation, we are quite happy with the part concerning the spectral inference, with and without trend: we are able to give posterior estimates, as well as descriptive statistics of the posterior distribution, of the spectral density, the long-memory parameter and the regression parameters. The Markov chain Monte Carlo simulation scheme that we proposed to sample from approximately the posterior distribution apparently behaves well, in terms of convergence to the stationary distribution and mixing properties, as it can be seen from the plots of the trajectories of the different parameters in Chapter 5. We had some numerical problems with the inversion of nearly singular matrices when drawing from the full conditional distribution of $\mathbf{G}$. Problems of this sort are likely to arise in any model which incorporates a smoothness prior, at least when making use of the technique based on a Dynamic Linear Model representation to sample from the full conditional distribution. The higher the degree of smoothness one wants to

incorporate in the model, the higher the risk of facing this kind of numerical problem. We resort to a special software to solve it by carrying out the relevant computations in extended precision arithmetic. This solution may be beneficial also to researchers making use of smoothness priors in different areas.

In conclusion, we regard the research that led to this dissertation as fruitful, successfully introducing the Bayesian point of view in the area of long-memory time series, and we deem it sets the stage for promising future developments in that field.

Bibliography

Ansley, C. F. and Kohn, R. J. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions, *The Annals of Statistics* **13**: 1286–1316.

Baillie, R. T. and King, M. L., eds. (1996). Fractional differencing and long memory processes, *Journal of Econometrics* **73**: 1–324.

Beran, J. (1992). Statistical methods for data with long-range dependence, *Statistical Science* **7**: 404–427.

Beran, J. (1994a). On a class of m-estimators for Gaussian long-memory models, *Biometrika* **81**: 755–766.

Beran, J. (1994b). *Statistical Methods for Long Memory Processes*, Chapman & Hall, New York.

Box, G. E. P. and Jenkins, G. M. (1970). *Time series analysis: forecasting and control*, Holden Day, San Francisco.

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

Carlin, J. B. and Dempster, A. P. (1989). Sensitivity analysis of seasonal adjustments: empirical case studies, *Journal of the American Statistical Association* **84**: 6–20.

Carlin, J. B., Dempster, A. P. and Jonas, A. B. (1985). On models and methods for Bayesian time series analysis, *Journal of Econometrics* **30**: 67–90.

Carter, C. and Kohn, R. (1994). On Gibbs sampling for state space models, *Biometrika* **81**: 541–53.

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

de Finetti, B. (1937). La prévision, ses lois logiques, ses sources subjectives, *Ann. Inst. H. Poincaré* **7**: 1–68. [English translation in *Studies in Subjective Probability* (1980) (H. E. Kyburg and H. E. Smokler, eds.) 53–118. Krieger, Malabar, FL].

Devroye, L. (1986). *Non-uniform random variate generation*, Springer-Verlag, New York.

Doob, J. L. (1953). *Stochastic processes*, Wiley, New York.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models, *Journal of Time Series Analysis* **15**: 183–102.

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.

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

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting, *Journal of the Royal Statistical Society, Ser. B* **38**: 205–247.

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

Hurst, H. E. (1951). Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers* **116**: 770–779.

Hurvich, C. M. and Beltrao, K. I. (1993). Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series, *Journal of Time Series Analysis* **14**: 455–472.

Hurvich, C. M. and Beltrao, K. I. (1994). Aknowledgement of priority for "Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series", *Journal of Time Series Analysis* **15**: 64.

Hurvich, C. M., Rohit, D. and Brodsky, J. (1996). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long memory time series, *Technical report*, Statistics and Operations Research Department, Leonard N. Stern School of Business, New York University.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**: 773–795.

Kolmogorov, A. N. (1940). Wienersche Spiralen und einige andere interessante Kurven in Hilbertschen Raum, *Comptes Rendus (Doklady) Acad. Sci. URSS (N.S.)* **26**: 115–118.

Koul, H. L. and Mukherjee, K. (1993). Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors, *Probability Theory and Related Fields* **95**: 535–553.

Mandelbrot, B. B. and van Ness, J. V. (1968). Fractional brownian motions, fractional noises and applications, *SIAM Review* **10**: 422–437.

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

Mandelbrot, B. B. and Wallis, J. R. (1969). Computer experiments with fractional gaussian noises, *Water Resources Research* **5**: 228–267.

Olshen, R. A. (1967). Asymptotic properties of the periodogram of a discrete stationary process, *Journal of Applied Probability* **4**: 508–528.

Ramsey, F. L. (1974). Characterization of the partial autocorrelation function, *The Annals of Statistics* **2**: 1296–1301.

Regazzini, E. and Petris, G. (1992). Some critical aspects of the use of exchangeability in statistics, *Journal of the Italian Statistical Society* **1**: 103–130.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long range dependence, *The Annals of Statistics* **23**: 1630–1661.

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

S-plus. Statistical Sciences, Inc. Seattle.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes. Stochastic Models With Infinite Variance*, Chapman & Hall, New York.

Schervish, M. J. (1995). *Theory of Statistics*, Springer-Verlag, New York.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops, *Journal of Agricultural Science* **28**: 1–23.

Smith, R. L. (1993). Long-range dependence and global warming, *in* V. Barnett and F. Turkman (eds), *Statistics for the Environment*, Wiley, Chichester, pp. 141–161.

Smith, R. L. and Chen, F.-L. (1996). Regression in long-memory time series, *Technical Report 2341*, Department of Statistics, University of North Carolina, Chapel Hill.

Taqqu, M. S. (1986). A bibliographical guide to self-similar processes and long-range dependence, *in* E. Eberlein and M. S. Taqqu (eds), *Dependence in Probability and Statistics*, Birkhäuser, Boston.

Tardella, L. (1996). Personal communication.

Terrin, N. and Hurvich, C. M. (1994). An asymptotic Wiener-itô representation for the low frequency ordinates of the periodogram of a long memory time series, *Stochastic Processes and their Applications* **54**: 297–307.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *The Annals of Statistics* **22**: 1701–1728.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression, *Journal of the Royal Statistical Society, Ser. B* **40**: 364–372.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*, 2nd edn, Springer-Verlag, New York.

Whittle, P. (1953). Estimation and information in stationary time series, *Ark. Mat.* **2**: 423–434.

Yajima, Y. (1988). On estimation of a regression model with long-memory stationary errors, *The Annals of Statistics* **16**: 791–807.

Yajima, Y. (1991). Asymptotic properties of the Lse in a regression model with long-memory stationary errors, *The Annals of Statistics* **19**: 158–177.

Zygmund, A. (1959). *Trigonometric series*, 2nd edn, Cambridge University Press, Cambridge.

Biography

Giovanni Petris was born in Milan, Italy in 1963. He received his B.S. in Mathematics from the University of Milan in 1989. He is a member of the American Statistical Association, the Institute of Mathematical Statistics and the American Mathematical Society.