E2C2 : Extreme Events, Causes and Consequences

Spectral Analysis of Stochastic Processes

Henning Rust
hrust@uni-potsdam.de

Nonlinear Dynamics Group, University of Potsdam

Lecture Notes for the
E2C2 / GIACS Summer School,
Comorova, Romania
September 2007

Chapter 1

Time Series Analysis and Stochastic Modelling

A time series is a sequence of data points measured at successive time intervals. Understanding the mechanism that generated this time series or making predictions are the essence of time series analysis. Applications range from physiology (e.g., **?**) and systems biology (e.g., **?**), over economy (e.g., **?**) and geophysical problems as daily weather forecasts (e.g., **?**) or detection and prediction of climate change (e.g., **?**) to astronomy (e.g., **?**).

This chapter introduces some concepts of linear time series analysis and stochastic modelling. Starting with random variables, we briefly introduce spectral analysis and discuss some special stochastic processes. An emphasis is made on the difference between short-range and long-range dependence, a feature especially relevant for trend detection and uncertainty analysis. Equipped with a canon of stochastic processes, we present and discuss ways of estimating optimal process parameters from empirical data.

1.1 Basic Concepts of Time Series Analysis

1.1.1 Random Variables

A random variable X is a mapping $X : \Omega \to \mathbb{R}$ from a sample space Ω onto the real axis. Given a random variable one can define probabilities of an event. As a simple example we consider a die with a sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. The probability P for the event $d \in \Omega$, $d =$"the number on a die is smaller than 4" is denoted as $P(X < 4)$. Such probabilities can be expressed using the cumulative probability distribution function

$$F_X(x) = P(X \leq x). \tag{1.1}$$

For a continuous random variable X, $F_X(x)$ is continuous. A discrete random variable X leads to $F_X(x)$ being a step function.

With miniscules x we denote a realisation $x \in \mathbb{R}$, a possible outcome of a random variable X. A sample is a set of N realisations $\{x_i\}_{i=1,\ldots,N}$.

Dependence

Two random variables X and Y are *independent* if their joint probability distribution $P(X < x, Y < y)$ can be written as a product of the individual distributions: $P(X < x, Y < y) =$

$P(X < x)P(Y < y)$, or, using the conditional distribution, $P(X < x \mid Y < y) = P(X < x)$. They are called *dependent* otherwise. A particular measure of dependence is the covariance which specifies the linear part of the dependence:

$$\mathrm{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])(Y - \mathrm{E}\left[Y\right])\right], \tag{1.2}$$

where $\mathrm{E}\left[X\right] = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x$ denotes the expectation value with $f_X(x) = \mathrm{d}F_X(x)/\mathrm{d}x$ being the probability density function (**?**).

A normalised measure quantifying the strength and direction[1] of the linear relationship of two random variables X and Y is the correlation

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}, \tag{1.3}$$

with $\mathrm{var}(X) = \mathrm{cov}(X, X)$ being the variance of X. It is a normalised measure taking values in the interval $-1 \leq \mathrm{cor}(X, Y) \leq 1$. An absolute correlation of unity, i.e. $|\mathrm{cor}(X, Y)| = 1$, implies Y being a linear function of X.

The variables X and Y are said to be uncorrelated if their covariance function, and thus their correlation, vanishes:

$$\mathrm{cov}(X, Y) = 0. \tag{1.4}$$

Correlated random variables are also dependent. The opposite statement is not necessarily true, because the correlation captures only the linear part of the dependence. Independent random variables X_i with identical distribution function are referred to as independent and identically-distributed random variables (IID).

For a set of random variables $X_1, X_2, \ldots, X_M$ the covariance matrix $\mathbf{\Sigma}$ with elements Σ_{ij} is defined as

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j). \tag{1.5}$$

1.1.2 Stochastic Processes

The word *stochastic* originates from the Greek *stochazesthai* ($\sigma\tau\omega\chi\acute{\alpha}\zeta\epsilon\sigma\theta\alpha\iota$) meaning "to aim at" or "to guess at" (**?**). It is used in the sense of random in contrast to deterministic. While in a deterministic model the outcome is completely determined by the equations and the input (initial conditions), in a stochastic model no exact values are determined but probability distributions. In that sense, a stochastic model can be understood as a means to guess at something.

The choice between a deterministic and a stochastic model is basically one of what information is to be included in the equations describing the system. On the one hand information can be limited simply by the lack of knowledge. On the other hand it might not be benefitting the modelling objective to include certain information.

A stochastic process $X(t)$ or X_t is an indexed collection of random variables with the indices specifying a time ordering. The index set can either be discrete ($t \in \mathbb{N}$) or continuous ($t \in \mathbb{R}$). In the latter case the collection consists of an uncountable infinite number of random variables.

With $\mathbf{x}_{t=1,\ldots,N} = (x_1, x_2, \ldots, x_N)$ we denote a realisation of the stochastic process X_t for $1 \leq t \leq N$. Where unambiguous, we omit the index. Depending on the context, $\mathbf{x}$ also denotes an empirical data record which is considered as a realisation of a possibly unknown process.

[1] Direction specifies the sign of the correlation. The direction is positive if X increases when Y increases and negative otherwise. It is not meant in the sense of variable X influencing variable Y or vice versa.

Stationarity

If the joint probability distribution $P(X_{t_1} < x_1, X_{t_2} < x_2, \ldots, X_{t_n} < x_n)$, with $t_i \in \mathbb{N}$, is identical with a displaced one $P(X_{t_1+k} < x_1, X_{t_2+k} < x_2, \ldots, X_{t_n+k} < x_n)$ for any admissible $t_1, t_2, \ldots, t_n$ and any k, the process is called *completely stationary*.

An alleviated concept is *stationarity up to an order m* with the first m moments of the joint probability distribution existing and being equal to those of the displaced one. For practical applications, frequently stationarity up to order $m = 2$ or *weak stationarity* is required. Testing for higher orders or for complete stationarity is usually not feasible.

Autocovariance

The covariance of two instances X_{t_1} and X_{t_2} at two different times t_1 and t_2 of a stochastic process X_t is called *autocovariance*

$$\operatorname{cov}(X_{t_i}, X_{t_j}) = \mathrm{E}\left[\left(X_{t_i} - \mathrm{E}\left[X_{t_i}\right]\right)\left(X_{t_j} - \mathrm{E}\left[X_{t_j}\right]\right)\right]. \tag{1.6}$$

For processes which are stationary at least up to order $m = 2$ the autocovariance depends only on the differences $\tau = t_i - t_j$ and can be written as

$$\operatorname{cov}(\tau) = \mathrm{E}\left[\left(X_t - \mathrm{E}\left[X_t\right]\right)\left(X_{t+\tau} - \mathrm{E}\left[X_t\right]\right)\right]. \tag{1.7}$$

Analogously, the *autocorrelation* can be defined as

$$\rho(X_{t_i}, X_{t_j}) = \frac{\operatorname{cov}(X_{t_i}, X_{t_j})}{\sqrt{\operatorname{var}(X_{t_i})}\sqrt{\operatorname{var}(X_{t_j})}}, \tag{1.8}$$

and in case of at least weak stationarity it can be written as

$$\rho(\tau) = \rho(X_t, X_{t+\tau}). \tag{1.9}$$

In the following, we refer to Equation (1.9) as the *autocorrelation function* or ACF.

Given a sample of an unknown process, we can estimate process characteristics as the ACF. An estimator for a characteristic T is a function of the sample and will be denoted by $\widehat{T}$. A nonparametric estimate for the autocovariance of a zero mean record $\mathbf{x}_{t=1,\ldots,N}$ is the sample autocovariance function or autocovariance sequence (**?**)

$$\widehat{\operatorname{cov}}(\tau) = \frac{1}{N}\sum_{t=1}^{N-\tau} x_t x_{t+\tau}, \quad 0 \le \tau < N. \tag{1.10}$$

Using the divisor N instead of $N - \tau$ ensures that the estimate for the autocovariance matrix $\mathbf{\Sigma}$ is non-negative definite. Consequently an estimate for the autocorrelation function can be formulated as

$$\widehat{\rho}(\tau) = \frac{\widehat{\operatorname{cov}}(\tau)}{\widehat{\sigma}_x^2}, \tag{1.11}$$

with $\widehat{\sigma}_x^2 = \widehat{\operatorname{cov}}(0)$ being an estimate of the record's variance.

The estimator is asymptotically normal and the variance can be approximated by $\widehat{\rho}(\tau) = n^{-1}$, a 95% confidence interval is thus given by $[-1.96n^{-1}, 1.96n^{-1}]$ (**?**).

For a more detailed description of these basic concepts refer to, e.g., **??**, or **?**.

1.1.3 Spectral Analysis

While the time domain approach to time series analysis originates from mathematical statistics, the spectral or frequency domain approach has its root in communication engineering. Whenever a signal fluctuates around a certain stable state we might use periodic functions to describe its behaviour. Spectral analysis aims at splitting the total variability of a stationary stochastic process into contributions related to oscillations with a certain frequency.

For the definition of the spectral density of a continuous stationary process $X(t)$ consider initially the process

$$X_T(t) = \begin{cases} X(t), & \text{if} \quad -T \leq t \leq T \\ 0, & \text{otherwise} \end{cases} . \tag{1.12}$$

This process is absolutely integrable due to the compact support $[-T, T]$. We now can define a Fourier integral as

$$G_T(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-T}^{T} X_T(t) \mathrm{e}^{-i\omega t} \mathrm{d}t. \tag{1.13}$$

and a power spectral density

$$\widetilde{S}(\omega) = \lim_{T \to \infty} \frac{|G_T(\omega)|^2}{2T}. \tag{1.14}$$

We use the notion of power as energy per time to obtain a finite spectral density. The aim is to represent the stochastic process and not only a single realisation. We thus have to average over multiple realisations. This leads us to a definition of a *power spectral density* for stochastic processes

$$S(\omega) = \lim_{T \to \infty} \mathrm{E}\left[\frac{|G_T(\omega)|^2}{2T} \right]. \tag{1.15}$$

In the following, we refer to the power spectral density simply as the spectral density. A detailed derivation of this concept is given by **?**.

Relation to the Autocovariance

The relation between the spectral density $S(\omega)$ and the autocovariance function $\mathrm{cov}(\tau)$ of a zero mean process is surprising at first sight. Using some basic properties of the Fourier transform, we can establish that the spectral density can be expressed as the Fourier transform of the autocovariance function:

$$S(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{cov}(\tau) \mathrm{e}^{-i\omega\tau} \mathrm{d}\tau. \tag{1.16}$$

In a more general setting, when the spectral density function does not exist but the integrated spectrum $F(\omega)$ does, this result manifests in the Wiener-Khinchin theorem using the more general Fourier-Stieltjes form of the integral (**?**)

$$\rho(\tau) = \int_{-\infty}^{\infty} \mathrm{e}^{i\omega\tau} \mathrm{d}F(\omega), \tag{1.17}$$

with $\rho(\tau)$ being the ACF. Thus, the spectral density and the ACF are equivalent descriptions of the linear dynamic properties of a process.

Estimation of the Spectral Density

Given a zero mean series **x** sampled at discrete time points $t = 1, \ldots, N$, an estimator for the spectral density function (1.15) can be formulated using the periodogram $I(\omega_j)$

$$\widehat{S}(\omega_j) = I(\omega_j) = \frac{1}{2\pi N} |\sum_{t=1}^{N} x_t \mathrm{e}^{-it\omega_j}|^2, \tag{1.18}$$

with the Fourier frequencies $\omega_j = 2\pi j/N$ and $j = 1, \ldots, [(N-1)/2]$, where $[.]$ denotes the integer part. The largest frequency to be resolved is the Nyquist frequency $f_{Ny} = \frac{1}{2\pi}\omega_{Ny} = \frac{2\pi}{2\Delta}$, where Δ is the sampling interval.

The periodogram is not a consistent[2] estimator for the spectral density because its variance does not decrease with increasing length N of the sample. To obtain a consistent estimator we can locally smooth the periodogram with, e.g., a suitable rectangular or bell-shaped window (e.g., **?**).

For Gaussian stochastic processes, the periodogram $I(\omega_j)$ at a certain angular frequency ω_j is χ^2-distributed with two degrees of freedom: $I(\omega_j) \sim \chi^2_2$. This result allows to derive confidence intervals for the power at a certain frequency (**?**).

Instead of the angular frequency ω, it is convenient to use the frequency $f = \omega/2\pi$ in practical applications. In this way, we can specify frequencies as reciprocal periods, which are expressed in units of the sampling interval, e.g., 1/days for data with a daily resolution.

1.1.4 Long-Range Dependence

For some stochastic processes, such as the popular autoregressive and moving average type models that will be introduced in Section 1.2, the ACF, denoted as $\rho_{\mathrm{SRD}}(\tau)$, decays exponentially and is thus summable:

$$\sum_{\tau=-\infty}^{\infty} \rho_{\mathrm{SRD}}(\tau) = \mathrm{const} < \infty. \tag{1.19}$$

These processes are called *short-range dependent* (SRD) or short-range correlated.

This characteristic contradicted other findings from, e.g., a spatial analysis of agricultural data by **?** or the famous Nile River flow minima studied by **?**. The behaviour they observed for the ACF for large lags – often referred to as Hurst phenomenon – was not consistent with hitherto existing models describing dependence. It was well represented assuming an algebraic decay of the ACF: $\rho_{\mathrm{LRD}}(\tau) \propto \tau^{-\gamma}$. This type of decay leads to a diverging sum

$$\sum_{\tau=-\infty}^{\infty} \rho_{\mathrm{LRD}}(\tau) = \infty. \tag{1.20}$$

Processes with an ACF following (1.20) are called *long-range dependent* (LRD), long-range correlated or long-memory processes.

Alternative Definitions of Long-Range Dependence

0 Today, it is common to use equation (1.20) as definition for a long-range dependent process (**?**). The following alternative formulations in the time and spectral domain, respectively, are consistent with (1.20).

[2] An estimator $\widehat{T}_N$ for the parameter T based on N observations is consistent iff for all $\epsilon > 0$ $\lim_{N\to\infty} P((\widehat{T}_N - T) < \epsilon) = 1$.

Time Domain Let X_t be a stationary process. If there exists a real number $\gamma \in (0,1)$ and a constant $c_\rho > 0$ such that

$$\lim_{\tau \to \infty} \frac{\rho(\tau)}{\tau^{-\gamma}} = c_\rho \tag{1.21}$$

holds, then X_t is called a process with long-range dependence or long memory.

Spectral Domain If there exists a real number $\beta \in (0,1)$ and a constant $c_f > 0$ such that

$$\lim_{\omega \to 0} \frac{S(\omega)}{|\omega|^{-\beta}} = c_S \tag{1.22}$$

holds then X_t is called a process with long-range dependence.

The concept of long-memory refers to non-periodic stochastic processes. Therefore the recurrence due to periodicities, such as the Milankovitch cycles[3] in the climate system, is not to be considered as long-range dependence, even if their (deterministic) behaviour causes dependence for infinite time lags.

1.2 Some Stationary Stochastic Processes

This section introduces some examples of stationary stochastic processs, SRD, as well as LRD processes. They can be used to describe e.g. temperaure anomalies or run-off records.

Gaussian Processes

A stochastic process X_t with joint probability distribution $P(X_{t_1} < x_1, X_{t_2} < x_2, \ldots, X_{t_k} < x_k)$ for all $t_k \in \mathbb{N}$ being multivariate normal is called a Gaussian process. It is completely defined by its first two moments, the expectation value $\mathrm{E}[X_t]$ and the autocovariance function $\mathrm{cov}(X_{t_i}, X_{t_j})$ or autocovariance matrix $\mathbf{\Sigma}$. This implies that a weakly stationary Gaussian process is also completely stationary (cf. Section 1.1.2). Furthermore, a Gaussian process is thus a linear process, because the autocovariance matrix specifies only linear relationships between the different instances X_{t_k}. Simple examples for discrete time Gaussian stochastic processes are the Gaussian white noise and autoregressive processes described in the following.

1.2.1 Processes with Short-Range Dependence

Gaussian White Noise Process

Let $\{X_t\}_{t=0,\pm1,\pm2,\ldots}$ be a sequence of uncorrelated[4] Gaussian random variables, then X_t is called a Gaussian white noise process or a purely random process. It possesses "no memory" in the sense that the value at time t is not correlated with a value at any other time s: $\mathrm{cov}(X_t, X_s) = 0, \forall t \neq s$. Hence, the spectrum is flat. A plot of a realisation and the corresponding periodogram of a zero mean Gaussian white noise process is given in Figure 1.1. In the following, we frequently refer to a Gaussian white noise process simply as white noise. To state that X_t is such a process with mean μ and variance σ^2, we use the notation $X_t \sim \mathcal{WN}(0,1)$.

[3] Milankovitch cycles are quasi-periodic changes in the earth's orbital and rotational properties resulting from being exposed to the gravitational force of multiple planets. The change in the orbital parameters effect the earth's climate.

[4] or, more generally, independent

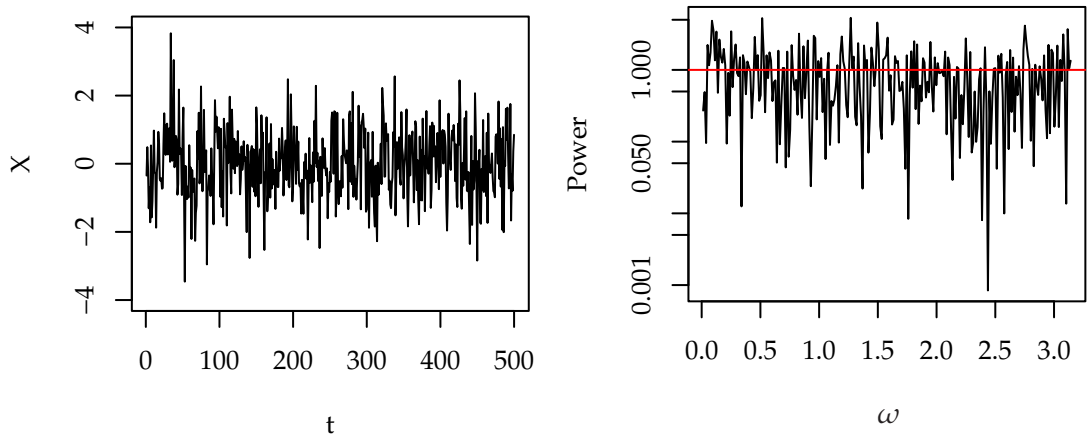

Figure 1.1: Realisation of a white noise process with 500 points (left) and the corresponding periodogram with the spectral density as red solid line (right).

First Order Autoregressive Process (AR[1])

An instance X_t at time t of an autoregressive process depends on its predecessors in a linear way. A first order autoregressive process (AR[1]), depends thus only on its last predecessor and it includes a *stochastic noise* or *innovations* term η_t in the following way:

$$X_t = aX_{t-1} + \eta_t. \tag{1.23}$$

If not explicitely stated otherwise, we assume throughout this thesis the driving noise term η_t to be a zero mean Gaussian white noise process with variance σ_η^2: $\eta_t = \mathcal{WN}(0, \sigma_\eta^2)$. A discussion for non-zero mean noise terms can be found in **?**. If η_t is Gaussian then the AR process is completely specified by its second order properties. Those are in turn determined by the propagator a and the variance σ_η^2 of η. Figure 1.2 shows a realisation and the corresponding periodogram of an AR[1] process.

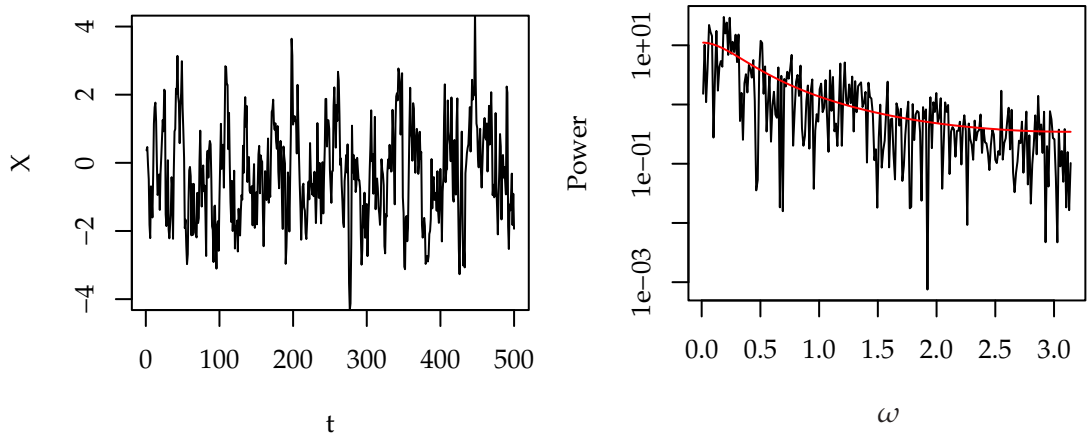

Figure 1.2: Realisation of an AR[1] process with 500 points (left) and the corresponding periodogram with the spectral density as red solid line (right). The autoregressive parameter is set to $a = 0.7$.

Mean and Autocovariance function Assuming $X_0 = 0$ the derivation of the mean and autocovariance function for this process is straightforward. Iteratively using (1.23) yields

$$\begin{aligned} X_t &= aX_{t-1} + \eta_t \\ &= a(aX_{t-2} + \eta_{t-1}) + \eta_t \\ &\vdots \\ &= a^{t-1}\eta_1 + a^{t-2}\eta_2 + \cdots + a^2\eta_{t-2} + a\eta_{t-1} + \eta_t. \end{aligned} \tag{1.24}$$

Taking expectation values on both sides and using $\mathrm{E}[\eta_t] = 0$ yields $\mathrm{E}[X_t] = 0$. Using this result, the autocovariance function for $\tau \geq 0$ is

$$\begin{aligned} \mathrm{cov}(X_t, X_{t+\tau}) &= \mathrm{E}\left[(a^{t-1}\eta_1 + a^{t-2}\eta_2 + \cdots + \eta_t)(a^{t+\tau-1}\eta_1 + a^{t+\tau-2}\eta_2 + \cdots + \eta_{t+\tau})\right] \\ &= \sigma_\eta^2(a^{2(t-1)+\tau} + a^{2(t-2)+\tau} + \cdots + a^\tau) \\ &= \begin{cases} \sigma_\eta^2 t, & \text{if} |a| = 1 \\ \sigma_\eta^2 a^\tau \left(\frac{1-a^{2t}}{1-a^2}\right), & \text{else.} \end{cases} \end{aligned} \tag{1.25}$$

Stationarity Setting $|a| < 1$ leads to $\mathrm{cov}(X_t, X_{t+\tau})$ being asymptotically (i.e. for large t) independent of t and thus the process is called asymptotically stationary. For $a = 1$ we obtain a non-stationary process, i.e. with $X_0 = 0$

$$X_t = X_{t-1} + \eta_t = X_{t-2} + \eta_{t-1} + \eta_t = \cdots = \sum_{i=0}^{t} \eta_i \,. \tag{1.26}$$

This process is known as the *random walk* (**?**). In the limit of small step sizes, the random walk approximates *Brownian motion*.

Higher Order Autoregressive Processes (AR$[p]$)

The autoregression in (1.23) can be straightforwardly extended to include more regressors with larger time lags. A general definition of an AR$[p]$ process with p regressors is

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \eta_t, \tag{1.27}$$

with $\eta_t \sim \mathcal{WN}(0, \sigma_\eta)$ being white noise. Rearranging the terms and using the back-shift operator $BX_t = X_{t-1}$, (1.27) reads

$$(1 - a_1 B - a_2 B^2 - \cdots - a_p B^p) X_t = \eta_t. \tag{1.28}$$

It is convenient to define the autoregressive polynomial of order p

$$\Phi(z) = (1 - a_1 z - a_2 z^2 - \cdots - a_p z^p). \tag{1.29}$$

Now, (1.27) reads simply

$$\Phi(B) X_t = \eta_t. \tag{1.30}$$

Deriving the condition for asymptotic stationarity of an AR$[p]$ process is somewhat more intricate. It leads to investigating whether the roots of $\Phi(z)$, with $z \in \mathbb{C}$, lying outside the unit circle of the complex plane (cf. **??**).

The general form of the autocorrelation function is given by

$$\rho(\tau) = A_1\phi_1^{|\tau|} + A_2\phi_2^{|\tau|} + \cdots + A_p\phi_p^{|\tau|}, \tag{1.31}$$

with ϕ_i being the reciprocals of the roots of $\Phi(z)$:$\{\phi_i \in \mathbb{C} \mid \Phi(1/\phi_i) = 0\}$. For real values $\phi_i \in \mathbb{R}$ the terms in (1.31) decay exponentially and the asymptotic behaviour for large τ is thus also an exponential decay. Complex values $\phi_i \in \mathbb{C}$ with non-zero imaginary part lead to damped oscillatory terms indicating a "pseudo-periodic" behaviour of X_t. The latter can be illustrated with "Yules Pendulum", a pendulum swinging in a resistive medium randomly being kicked to sustain the motion. Since the oscillation dies away due to the damping but gets randomly actuated, an exact periodic behaviour is not recovered.

A realisation of an AR[2] process with $a_1 = 0.4$ and $a_2 = -0.8$ is depicted in Figure 1.3 together with the corresponding periodogram. The two reciprocal roots of the AR poly-

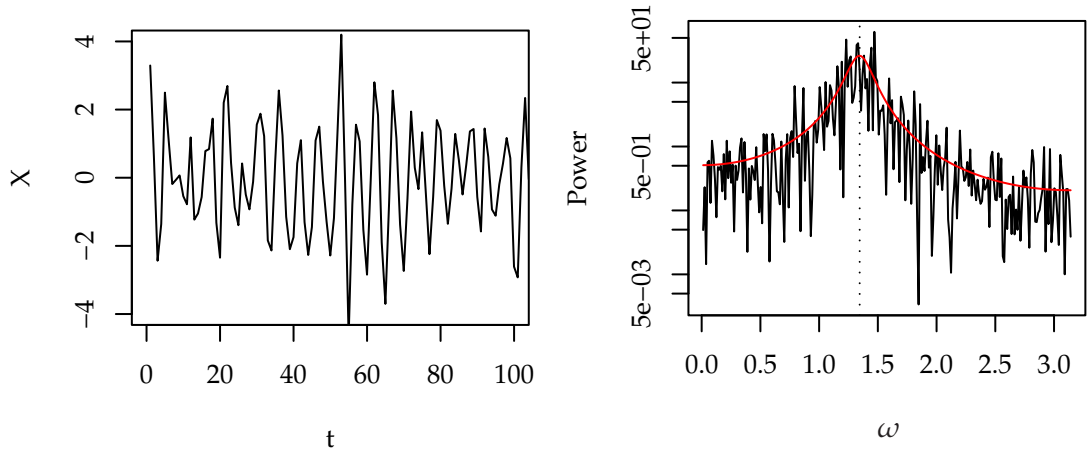

Figure 1.3: Realisation (left) and spectral density and periodogram (right) of an AR[2] process with a pseudo-periodic behaviour. The dotted vertical line at $\omega = 1.35$ marks the predominant frequency of the oscillating behaviour .

nomial are $\phi_1 = 0.2 - 0.87i$ and $\phi_2 = 0.2 + 0.87i$. The argument of the complex number $\arg(\phi_1) = 1.35$ gives the dominant frequency of the oscillating behaviour. The absolute value $|\phi_1| = 0.89$ is related to the damping of the oscillation. A characteristic time scale is given by $T_1 = -1/\ln|\phi_1| = 8.96$.

Moving Average Processes (MA$[q]$)

A different class of linear processes is based on the idea of averaging random shocks η_t which occur independently in time, e.g., $\eta_t \sim \mathcal{WN}(0, \sigma_\eta)$. Such processes can be written as

$$X_t = \eta_t + b_1\eta_{t-1} + \cdots + b_q\eta_{t-q} \tag{1.32}$$

and are called moving average processes of order q or MA$[q]$ processes. Analogously to AR$[p]$ process we can use the back-shift operator B for a convenient notation:

$$X_t = \Psi(B)\eta_t, \tag{1.33}$$

involving the moving average polynomial

$$\Psi(z) = (1 + b_1 z + b_2 z^2 + \cdots + b_q z^q). \tag{1.34}$$

The MA process (1.32) is thus a special case of the general linear process

$$X_t = \sum_{i=-\infty}^{\infty} b_i B^i \eta_t, \tag{1.35}$$

which allows also for future shocks to be included. In contrast, the MA$[q]$ process defined in (1.32) includes only those terms involving past and present shocks η_{t-i} with $i \geq 0$, leading to a *causal process*.

Autocovariance Different from AR processes, where X_t is a linear combination of its predecessors and a random shock, a variable X_t of a MA process depends only on a finite amount of past and present random shocks. This leads to a cut-off in the autocorrelation function for MA$[q]$ processes for lags larger than q, whereas the ACF of AR$[p]$ processes gradually fades to zero (1.31).

Stationarity and Invertibility Being a linear combination of independent identically distributed random variables η_t, the MA process is trivially stationary for any choice of the parameters b_i. A more interesting characteristic of these processes is the invertibility. In case (1.33) can be written as

$$\eta_t = \Psi^{-1}(B) X_t, \tag{1.36}$$

the process is called *invertible*. Invertibility requires the complex zeros of $\Psi(z)$ lying outside the unit circle (cf. **?**).

ARMA and ARIMA Processes

The combination of AR$[p]$ and MA$[q]$ processes leads to the concept of ARMA$[p, q]$ processes

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{j=0}^{q} b_j \eta_{t-j}. \tag{1.37}$$

A convenient notation using the back-shift operator is

$$\Phi(B) X_t = \Psi(B) \eta_t \tag{1.38}$$

with $\Phi(z)$ and $\Psi(z)$ being the AR and MA polynomials, respectively. If $\Phi(z)$ and $\Psi(z)$ have no common zeros, we can write (1.38) as

$$X_t = \sum_{i=0}^{\infty} \gamma_i \eta_{t-i}, \tag{1.39}$$

with

$$\Gamma(z) = \sum_{i=1}^{\infty} \gamma_i z^i = \Psi(z)/\Phi(z) \tag{1.40}$$

being the quotient of the MA and AR polynomial. Such ARMA$[p, q]$ processes are *invertible* if the complex roots of $\Psi(z)$ lie outside the unit circle. Analogously to AR processes, they are stationary if we find the roots of $\Phi(z)$ outside the unit circle. As a linear stationary stochastic process, an ARMA$[p, q]$ process is completely defined by its autocovariance function $\mathrm{cov}(\tau)$ or, equivalently, in terms of the spectral density.

Despite the general nature of this formulation, ARMA$[p, q]$ models cannot describe all stationary linear processes. Their generality can be compared to that of a rational

function. This analogy arises from their spectral density function being essentially the quotient of $\Psi(e^{i\omega})$ and $\Phi(e^{i\omega})$:

$$S(\omega) = \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(e^{i\omega})|^2}{|\Phi(e^{i\omega})|^2}, \tag{1.41}$$

and thus a rational function in the spectral domain.

Integrated Processes **?** included a class of non-stationary processes Y_t in the framework, namely those being partial sums of stationary ARMA$[p,q]$ processes X_t: $Y_t = \sum_{i=0}^{t} X_i$. Using the difference operator[5] $(1-B)^d$, we can also obtain X_t as the *increment process* of Y_t writing

$$X_t = (1-B)^d Y_t, \tag{1.42}$$

with d being at first an integer value, later we allow $d \in \mathbb{R}$. Now, equation (1.38) can be generalised to

$$\Phi(B)(1-B)^d Y_t = \Psi(B)\eta_t \tag{1.43}$$

including now also integrated processes. Since Y_t is the outcome of the sum or integration of X_t such processes are called integrated ARMA$[p,q]$ or ARIMA$[p,d,q]$ processes, with the integer d specifying the degree of integration.

1.2.2 Processes with Long-Range Dependence

In the following, we introduce two representatives of long-range dependent processes. The first example, the fractional Gaussian noise (fGn), has its roots in the fractal community and has been made popular in the 1960s by the French mathematician Benoît Mandelbrot. It was the first mathematical model to describe long-range dependence. About two decades later **?** and **?** proposed a different way to describe this phenomenon. They introduced the concept of fractional differencing leading to fractional differenced processes (FD) and fractional ARIMA processes. Further examples of models for long-range dependence can be found in, e.g., **?** or **?**.

Increments of Self-Similar Processes

In the 1960s Mandelbrot introduced self-similar stochastic processes, actually dating back to **?**, in the statistics community (**?**). A process Y_t is called self-similar if $Y_t \overset{d}{=} c^{-H} Y_{ct}$, where $a \overset{d}{=} b$ means a and b are equal in distribution. H is called the self-similarity parameter or Hurst exponent and c is a positive constant. To model data with a stationary appearance one considers the stationary increment process $X_t = Y_t - Y_{t-1}$ of the self-similar process Y_t. Following **?**, its autocorrelation function reads

$$\rho(\tau) = \frac{\sigma}{2}\left[(\tau+1)^{2H} - 2\tau^{2H} + (\tau-1)^{2H}\right], \tag{1.44}$$

with asymptotic behaviour ($\tau \to \infty$) resulting from a Taylor expansion

$$\rho(\tau) \to H(2H-1)\tau^{2H-2}, \quad \text{for } \tau \to \infty. \tag{1.45}$$

Thus, for $0.5 < H < 1$ the variance decays algebraically and the increment process X_t is long-range dependent. For $0 < H < 0.5$ the ACF is summable; it even sums up to zero,

[5] For $d = 1$, $(1-B)X_t = X_t - X_{t-1}$. This can be regarded as the discrete counterpart of the derivative.

resulting in SRD. This "pathologic" situation is rarely encountered in practice; it is mostly the result of over-differencing (**?**).

The increments of self-similar processes are thus capable of reproducing LRD and offered an early description of the Hurst phenomenon. The well-known *fractional Brownian motion* is a Gaussian representative of a self-similar process. Its increment series is a long-range dependent stationary Gaussian process with asymptotic properties as given in (1.45) and is referred to as *fractional Gaussian noise* (fGn).

Fractional Difference Processes (FD)

After Mandelbrot promoted self-similar processes, it were **?** and **?** who formulated a different model to describe long-range dependence. This model fits well into the framework of the linear processes discussed in Section 1.2.1. They allowed the difference parameter d in (1.42) to take real values. This results in a fractional difference filter

$$(1-B)^d X_t = \eta_t, \tag{1.46}$$

with $d \in \{d \in \mathbb{R} \mid -1/2 < d < 1/2\}$ and η_t being white noise. The fractional difference operator is defined using an infinite power series

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k, \tag{1.47}$$

with the binomial coefficient

$$\binom{d}{k} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}, \tag{1.48}$$

and $\Gamma(x)$ denoting the gamma function. The process X_t has the asymptotic properties of a long-range dependent process as given in (1.21) and is referred to as *fractional difference process* (FD). For $d > 1/2$ the second moment is not finite anymore. In such cases the increment process $X = (1-B)Y$ can be described as a stationary FD process.

Fractional ARIMA Processes (FARIMA$[p,d,q]$)

The FD process has been formulated in the framework of linear models. Although the power series expansion of $(1-B)^d$ is infinite, it can be included into the concept of ARIMA models leading to fractional ARIMA or FARIMA[6] processes. Using the notation of (1.43), we get

$$\Phi(B)(1-B)^d X_t = \Psi(B)\eta_t, \tag{1.49}$$

with $\Phi(z)$ and $\Psi(z)$ being again the autoregressive and moving average polynomials defined in (1.29) and (1.34). Now, we allow d being a real number with $d \in \{d \in \mathbb{R} \mid -1/2 < d < 1/2\}$. The spectral density of FARIMA processes can be obtained from the corresponding result for ARMA processes (1.41), multiplied by $|1-\mathrm{e}^{i\omega}|^{-2d}$:

$$S(\omega) = \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(\mathrm{e}^{i\omega})|^2}{|\Phi(\mathrm{e}^{i\omega})|^2} |1-\mathrm{e}^{i\omega}|^{-2d}. \tag{1.50}$$

For small frequencies $\omega \to 0$ the limiting behaviour for the spectral density is given by

$$S(\omega) \approx \frac{\sigma_\eta^2}{2\pi} \frac{|\Psi(1)|^2}{|\Phi(1)|^2} |\omega|^{-2d} = c_f |\omega|^{-2d}. \tag{1.51}$$

[6]Sometimes referred to as ARFIMA processes

With $2d = \beta$ (1.51) recovers the asymptotic behaviour required for a long-range dependent process given by (1.22). The relation to the Hurst exponent is $2d = \beta = 2H - 1$.

1.2.3 Motivation for Autoregressive Moving Average Models

The ARMA$[p, q]$ model class presented in the preceeding sections has a flexibility or generality which can be compared to that of a rational function as can be seen from (1.41). Including fractional differencing (1.50), we obtain the FARIMA$[p, d, q]$ class and this rational function is supplemented by a term converging for small frequencies to a power-law, cf. (1.51). This allows for flexible modelling of the spectral density (or ACF) including LRD. Because here the main goal is a suitable description of the ACF and especially the detection of LRD, the latter is a motivation for using FARIMA$[p, d, q]$ processes.

Besides flexibility and feasibility, there are physically based arguments for autoregressive moving average models. Quite generally autoregressive processes can be regarded as discretised linear ordinary differential equations including a stochastic noise term. This allows to describe relaxations and oscillations and offers thus a framework for modelling many physical systems.

1.3 Parameter Estimation for Stochastic Processes

The model building process consists of two parts: the inference of a proper model structure and the estimation of the unknown model parameters. A thorough discussion of the model selection is presented in the subsequent chapter. Here we assume that a suitable model is known and present parameter estimation strategies for the stochastic processes (Section 1.2.2), particularly a maximum likelihood approach to FARIMA$[p, d, q]$ processes.

Besides the likelihood approach there are also other ways to estimate the fractional difference parameter. We discuss the detrended fluctuation analysis (DFA) as a frequently used heuristic method[7] in the following, while the description of the rescaled range analysis and the the semi-parametric[8] log-periodogram regression is deferred to Appendices B.3.1 and B.3.2, respectively. These semi-parametric and heuristic approaches are formulated on the basis of the asymptotic behaviour and focus on estimating the Hurst exponent H or the fractional difference parameter d to quantify long-range dependence. It is not intended to include the high frequency behaviour in the description, it is rather attempted to reduce its influence on the estimation of H or d.

A section on generating realisations of a FARIMA$[p, d, q]$ process concludes this chapter.

Estimation Based on the Autocovariance Function

The ARMA$[p, q]$ processes, as formulated in section 1.2.1, can be considered as a parametric model for the ACF. Such a model completely determines the ACF and vice versa. Therefore, one way to obtain estimates of the model parameters is to use the empirical

[7] As heuristic we denote an estimator without an established limiting distribution. The specification of confidence intervals and thus statistical inference is not possible.

[8] We call an approach semi-parametric if only the behaviour for large scales (small frequencies) is described. Contrary to a heuristic approach, the limiting distribution for the estimator is available.

autocovariance series

$$\hat{\rho}(\tau) = \frac{1}{c_N} \sum_{t=1}^{N-\tau} (X_t - \mu)(X_{t+\tau} - \mu) \tag{1.52}$$

to determine the model parameters.

For autoregressive processes the relation between the model parameters and the ACF is given by the so called Yule-Walker equations; for moving average processes the model parameters can be obtained recursively from the ACF using the innovations algorithm (**?**). Because we can write causal ARMA$[p, q]$ processes as moving average processes of infinite order, we can use this algorithm also to obtain estimates for ARMA$[p, q]$ parameters on the basis of the autocovariance sequence. These estimators are consistent in the sense defined in Section 1.1.3: they converge in probability ($X_n \xrightarrow{P} Y : P(|X_n - Y| > 0) \to 0$ for $n \to \infty$ to the true values but they are not efficient, i.e. they do not have the smallest possible variance. Efficient estimates can be obtained using the maximum likelihood approach presented in the following. The preliminary estimates based on the Yule-Walker equations or the innovations algorithm can be used as initial guesses in the likelihood approach described in the following (**?**).

1.3.1 Maximum Likelihood for FARIMA$[p, d, q]$ Processes

Consider a zero mean stationary FARIMA$[p, d, q]$ process. As a Gaussian process, it is fully specified by the autocovariance matrix $\mathbf{\Sigma}(\boldsymbol{\theta})$, which in turn depends on the process parameters $\boldsymbol{\theta} = (d, a_1, \ldots, a_p, b_1, \ldots, b_p, \sigma_\eta)$. We can thus formulate the probability density for obtaining a realisation $\mathbf{x}$ as

$$p(\mathbf{x}|\boldsymbol{\theta}) = (2\pi)^{-\frac{N}{2}} |\mathbf{\Sigma}(\boldsymbol{\theta})|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\mathbf{x}^\dagger \mathbf{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x}}, \tag{1.53}$$

with $\mathbf{x}^\dagger$ denoting the transposed of $\mathbf{x}$. For a given realisation $\mathbf{x}'$, the same expression can be interpreted as the likelihood for the parameter vector $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}') = p(\mathbf{x}'|\boldsymbol{\theta}). \tag{1.54}$$

This suggests the formulation of an estimator for $\boldsymbol{\theta}$ based on a realisation $\mathbf{x}'$: the estimate $\widehat{\boldsymbol{\theta}}$ is chosen such that (1.54) is maximised. It is convenient to use the logarithm of the likelihood, $l(\boldsymbol{\theta}|\mathbf{x}') = \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}')$. Due to its monotonicity, the estimator can then be formulated as the argument of the maximum log-likelihood

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x}'). \tag{1.55}$$

Asymptotic Properties of the Maximum Likelihood Estimator

For causal and invertible Gaussian processes and an increasing sample size $N \to \infty$ the maximum-likelihood estimator (MLE) is unbiased and converges to the true parameter vector $\boldsymbol{\theta}^0$ almost surely (**??**). Furthermore, the estimator converges in distribution

$$N^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \xi \tag{1.56}$$

to a Gaussian random variable $\xi \sim \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\theta}^0})$, with $\mathbf{V}_{\boldsymbol{\theta}^0} = 2\mathbf{D}^{-1}(\boldsymbol{\theta}^0)$ being the covariance matrix for the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)^\dagger$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. The $M \times M$ matrix $\mathbf{D} = D_{ij}(\boldsymbol{\theta}^0)$ is defined by

$$D_{ij}(\boldsymbol{\theta}^0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_i} \log S(\omega; \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log S(\omega; \boldsymbol{\theta}) \mathrm{d}\omega |_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}. \tag{1.57}$$

The rate of convergence of $\widehat{\boldsymbol{\theta}}$ is $N^{-1/2}$ as expected also for the short-range dependent case. This is particularly interesting since some estimates exhibit a slower rate of $N^{-\alpha}$ in the presence of long-range dependence as, e.g., regression parameters or the mean value as shown in (B.6).

? established asymptotic efficiency of the MLE for long-range dependent processes [9]. This implies minimal variance (as specified by the Cramér-Rao bound), or, equivalently, the Fisher information matrix

$$\boldsymbol{\Gamma}_n(\boldsymbol{\theta}^0) = \mathrm{E}\left[[\mathbf{l}'(\widehat{\boldsymbol{\theta}}|\mathbf{x})][\mathbf{l}'(\widehat{\boldsymbol{\theta}}|\mathbf{x})]^\dagger\right] \tag{1.58}$$

converges to the inverse of the estimators covariance matrix

$$\lim_{n\to\infty} \boldsymbol{\Gamma}_n(\boldsymbol{\theta}^0) = \mathbf{V}_{\boldsymbol{\theta}^0}^{-1}. \tag{1.59}$$

The maximum likelihood estimation (1.55) is formulated as a nonlinear optimisation problem which can be solved numerically. This requires inverting the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ for every optimisation step in the parameter space. This is a costly and potentially unstable procedure and thus not feasible for long records. A convenient approximation to the MLE has been proposed by **?** and is described in the following.

1.3.2 Whittle Estimator

The main idea of **?** was to give suitable approximations for the terms in the log-likelihood

$$l(\boldsymbol{\theta}|\mathbf{x}') = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{x}^\dagger\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x} \tag{1.60}$$

which are dependent on $\boldsymbol{\theta}$: the determinant $\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and $\mathbf{x}^\dagger\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{x}$. For both terms the approximation involves an integral over the spectral density $S(\omega;\boldsymbol{\theta})$ of the process which is in a successive step approximated by a Riemann sum. Finally, an appropriate rescaling of the spectral density $S(\omega;\boldsymbol{\theta}) = \theta_1 S(\omega;\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = (1,\theta_2,\ldots,\theta_M)^\dagger$ and $\theta_1 = 2\pi\sigma_\eta^{-2}$ yields a discrete version of the Whittle estimator:

1. Minimise
$$Q(\boldsymbol{\theta}^*) = \sum_{j=1}^{[(N-1)/2]} \frac{I(\omega_j)}{S(\omega_j;\boldsymbol{\theta}^*)} \tag{1.61}$$
with respect to $\boldsymbol{\theta}^*$.

2. Set
$$\hat{\sigma}_\eta^2 = 2\pi\hat{\theta}_1 = \frac{4\pi}{N}Q(\boldsymbol{\theta}^*). \tag{1.62}$$

$I(\omega_j)$ denotes the periodogram of the realisation $\mathbf{x}'$ at the Fourier frequencies $\omega_j = \frac{2\pi j}{N}$ with $j = 1,\ldots,\left[\frac{N-1}{2}\right]$ where $[.]$ denotes the integer part. A detailed derivation of the discrete Whittle estimator for LRD processes has been given by **?**.

[9] See also the correction notes at
`http://math.uni-heidelberg.de/stat/people/dahlhaus/ExtendedCorrNote.pdf`

Asymptotic Properties of the Whittle Estimator

It can be shown, that the Whittle aproximation has the same asymptotic distribution as the exact ML estimator (**?**). Therefore, it is asymptotically efficient for Gaussian processes. The assumption of a zero mean process has been made for simplification of the representation, the asymptotic result does not change if the mean is consistently estimated and substracted (**?**). Besides simplifying the optimisation, the choice of the scale factor brings about a further convenience: the estimate of the innovations variance $\sigma_\eta^2 = 2\pi\theta_1$ is asymptotically independent (i.e. for $N \to \infty$) of the other parameter estimates.

Confidence Intervals

The asymptotic distribution (1.56) can be used to obtain approximate α100% confidence regions for the parameter estimate $\widehat{\boldsymbol{\theta}}$ as

$$\mathrm{CI}_\alpha = \{\boldsymbol{\theta} \in \mathbb{R}^{(p+q+1)} \mid (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\intercal \mathbf{V}^{-1}(\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \leq N^{-1}\chi^2_{(p+q+1),\alpha}\}, \tag{1.63}$$

with $\chi^2_{m,\alpha}$ specifying the α-quantile of the χ^2 distribution with m degrees of freedom, where m is the number of parameters (**?**). For a single parameter estimate $\widehat{\theta}_j$ we can also obtain approximate confidence intervals using the standard normal distribution. Writing the jth diagonal element of the estimator's covariance matrix $\mathbf{V}(\widehat{\theta})$ as v_{jj} we obtain

$$\mathrm{CI}_\alpha = \{\theta \in \mathbb{R} \mid |\theta - \widehat{\theta}_j| \leq N^{-1/2}\Phi_{((1+\alpha)/2)}v_{jj}^{1/2}\}, \tag{1.64}$$

with $\Phi_{((1+\alpha)/2)}$ being the $(1+\alpha)/2$-quantile of the standard normal distribution (**?**).

An alternative variant is to use a bootstrap approach. This approach consists of generating an ensemble of time series of original length using the model obtained for the empirical series and a subsequent parameter estimation from all ensemble members. Confidence intervals can then be obtained from the empirical frequency distribution of the estimates following **?**. A variant based on resampling the residuals can be found in **?**. Bootstrap approaches are especially useful if the asymptotic confidence intervals are not reliable, e.g., if the residual distribution is not Gaussian.

Non-stationary Long-Range Dependence

So far, we have assumed a fractional difference parameter in the stationary range $-1/2 < d < 1/2$. **?** showed that the Whittle estimation presented here can be extended to non-stationary processes with $1/2 \leq d < 1$. It is consistent for $d < 1$ and preserves its asymptotic normality for $d < 3/4$. For larger d the asymptotic normality can be recovered using a cosine bell taper for the calculation of the periodogram. For such a non-stationary process we cannot easily express the likelihood as in (1.54). We can, however, calculate an approximate log-likelihood using the ML-estimate of the residuals variance $\widehat{\sigma}_\epsilon^2$ as

$$l(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \approx -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma}_\epsilon^2. \tag{1.65}$$

Performance

? performed an extensive simulation study on bias and variance of the fractional integration parameter and the autoregressive and moving average parameters in FARIMA$[p, d, q]$ models using ML estimation. Given the correct model the ML estimation outperforms the

log-periodogram regression. Mis-specifying the model orders results in biased estimates of d, as well as of the autoregressive parameters. Similar results have been obtained by **?** investigating three types of Whittle estimators. The estimator proposed here is the most accurate of those under investigation provided the correct model is chosen. This stresses the need of reliable model selection (Chapter 2).

Numerical Aspects

In contrast to moment based estimators, such as the Yule-Walker equations or the innovations algorithms (e.g., **?**) the optimisation problem in (1.55) is generally carried out numerically. It is thus burdened with problems inherent to nonlinear optimisation, for instance, getting trapped in local minima. This problem can be partially surmounted by initialising the algorithm with preliminary estimates. For the AR or ARMA components, these estimates can be obtained using the Yule-Walker equations, the innovations algorithm, or an estimate obtained by minimising the conditional sum of squares (**??**).

The optimisation routines used in the following are those implemented in R's[10] `optim` routine, namely the simplex method (**?**) or, alternatively, a quasi-Newton algorithm (**?**). While the latter is faster, the simplex method is more robust. For box-constrained numerical optimisation, the algorithm by **?** is available.

Implementation

The numerical algorithm given in `S-Plus` by **?** and translated to `R` by Martin Mächler was used in a modified and supplemented form available as the `R`-package `farisma`[11]. Extensions are made with respect to the initial guesses of the parameter values and tools for visualisation, simulation and model selection were supplemented.

1.3.3 Detrended Fluctuation Analysis – A Heuristic Approach

The first attempts to describe the long-range dependent phenomenon were made before suitable stochastic models had been developed. Starting with the rescaled range statistic (cf. Appendix B.3.1) proposed by Hurst in 1951, several heuristic methods have emerged to estimate the Hurst coefficient H. The approaches are based, for example, on a direct calculation of the autocovariance sequence or on studying the behaviour of the variance of the mean of subsamples with increasing length in a log-log plot. A relatively recent approach investigates the variance of residuals of a regression in subsamples and is thus to a certain extend robust to instationarities. This approach has become widely used in the physics community and is known as detrended fluctuation analysis (DFA) or residuals of regression. Other heuristic methods are discussed by, e.g., **?** or **?**. These heuristic methods are not suitable for statistical inference about the long-range dependence parameter, but are rather diagnostic tools to start with. Confidence intervals cannot be obtained straightforwardly which renders the interpretation of the result difficult. For statistical inference one can either consider the log-periodogram regression (Appendix B.3.2) or the full parametric modelling approach presented above (**?**).

Residuals of regression or detrended fluctuation analysis (DFA) was developed by **??** while studying DNA nucleotides and later heartbeat intervals (e.g., **?**). Their aim was to investigate for long-range dependent processes underlying these records excluding the

[10] `R` is a software package for statistical computing and freely available from `http://www.r-project.org/` (**?**)

[11] `http://www.pik-potsdam.de:~hrust/tools/`

influence of a possible trend. Direct estimation of the autocorrelation function from empirical data is limited to rather small time lags s and is affected by observational noise and instationarities like trends. DFA received a lot of attention in recent years and became a frequently used tool for time series analysis with respect to LRD (e.g., **?**). Many works investigating for LRD in various fields of research are based on this method. Some examples are empirical temperature records (e.g., **?????**), or temperature series from climate models (e.g., **??**), run-off (e.g., **?**) and wind speed records (e.g., **?**). Also with respect to extreme events DFA has been used (e.g., **?**). This popularity is probably due to the simplicity of the method. There are, however, difficulties and pitfalls which are easily overlooked when applying DFA. The latter are described in more detail in Appendix A.2.

For a description of the method, consider again the time series $\{X_i\}_{i=1,\ldots,N}$. Similar to Hurst's rescaled range analysis (cf. Appendix B.3.1), the detrended fluctuation analysis is based on the aggregated time series $Y(t) = \sum_{i=1}^{t} X_i$. First, divide $Y(t)$ into M non-overlapping segments of length s. Then, for DFA of order n (DFAn), in each segment $m = 1, \ldots, M$ fit a polynomial of order n. This polynomial trend $p_{s,m}^{(n)}$ is subtracted from the aggregated series:

$$Y_{s,m}(t) = Y(t) - p_{s,m}^{(n)}(t). \tag{1.66}$$

For every segment m the squared fluctuation is calculated as

$$F_m^2(s) = \frac{1}{s} \sum_{t=(m-1)s+1}^{ms} Y_{s,m}(t)^2. \tag{1.67}$$

Averaging over all segments $m = 1, \ldots, M$ yields the squared fluctuation function for the time scale s:

$$F^2(s) = \frac{1}{M} \sum_{m=1}^{M} F_m^2(s). \tag{1.68}$$

This procedure is repeated for several scales s. The time scales s are limited at the lower bound by the order n of DFA and at the upper bound by the length of the record. Due to the increasing variability of $F^2(s)$ with s, a reasonable choice for the maximum scale is $s_{\max} \approx N/10$. The fluctuation function $F(s)$ is then investigated in a double logarithmic plot.

Interpretation of the Slope

Using the asymptotic behaviour of fractional Gaussian noise (fGn) and FARIMA processes, **?** showed that the resulting fluctuation function obtained with DFA is asymptotically proportional to s^H, with H being the Hurst exponent. Thus the asymptotic slope in the log-log plot yields an estimate $\widehat{H}_{\mathrm{DFA}n}$ (or, equivalently, $\widehat{d}_{\mathrm{DFA}n}$ (1.49)) of the Hurst exponent (or fractional difference parameter)(**?**). For increments of self-similar processes, as fGn, the asymptotic behaviour is already reached for very moderate sample size. For those processes, one can choose to fit a straight line in the range $1 < \log s < \log N/10$. Values for $\log s > \log N/10$ are frequently excluded due to a large variability.

Asymptotically, we can relate the Hurst exponent H to the exponent γ quantifying the algebraic decay of the ACF (1.21) by

$$H = 1 - \gamma/2\,, \quad \text{with} \quad 0.5 < H < 1. \tag{1.69}$$

It can be as well related to the fractional difference parameter (1.49) by $d = H - 0.5$ (cf. Section 1.2.2; **?**). For an uncorrelated process we get a squared fluctuation function

$F^2(s) \propto s$ (i.e. $F(s) \propto s^{0.5}$) which reflects the linear increase of the variance of the aggregated series $Y(t)$ with t (**?**).

More details on the basics of DFA can be found in (**?**). **?** and **?** investigate extensively the effects of various types of instationarities on DFA using simulation studies with realisations of the self-similar increment process fGn.

Limiting Behaviour, Uncertainty Analysis

Studies regarding the variability of the estimate $\widehat{H}_{\mathrm{DFA}n}$ are rare. Analytical approaches towards a derivation of the limiting distribution could not be found in the literature. While a first impression of bias and variance can be obtained from the simulations in **?**, a more detailed study is given in Appendix A.1. Furthermore, **?** performed a systematic Monte Carlo study to quantify the variability of $\widehat{H}_{\mathrm{DFA}n}$ based on realisations of Gaussian white noise. He derived empirical confidence intervals which can be used for testing the null hypothesis of a Gaussian white noise process. Weron estimated the slope of the logarithmic fluctuation function for scales $s > 10$ and $s > 50$. For many practical applications the asymptotic behaviour has not been reached for such small scales. This implies that suitable conditions to transfer these confidence intervals are rare. Furthermore, neither the upper bound for the straight line fit has been specified, nor the number of sampling points s for which the fluctuation function has been calculated. Although Weron's confidence bands might therefore not be suitable for a direct transfer to many practical applications, the idea he presented leads to a promising approach of obtaining confidence intervals using a simulation approach. If it is possible to specify a parametric model for the observed record, such as a FARIMA$[p, d, q]$, confidence intervals can be easily estimated using a parametric bootstrap (**?**). If, however, such a model has been identified, an estimate including asymptotic confidence intervals for the Hurst coefficient can be derived directly (Section 1.3.1).

1.4 Simulations from Long-Range Dependent Processes

There are several ways of obtaining a realisation of a LRD model. An extensive survey of generators has been conducted by **?**. Here, we focus on the description of the direct spectral method which was implemented and used in this work. It makes use of some essentials of spectral analysis: a) the relation between the spectrum $S(\omega)$ and the ACF (1.16), b) the periodogram $I(\omega)$ (1.18) as an estimate for the spectral density and c) the sampling properties of $I(\omega)$. The latter are such that the Fourier coefficients a_j and b_j of a Gaussian process $\{X_t\}_{t=1,\ldots,N}$, defined by

$$a_j = N^{-1/2} \sum_{t=1}^{N} X_t \cos(\omega_j t), \quad b_j = N^{-1/2} \sum_{t=1}^{N} X_t \sin(\omega_j t), \tag{1.70}$$

are Gaussian random variables. This implies that the periodogram

$$I(\omega_j) = \frac{1}{2\pi} \begin{cases} a_j^2 + b_j^2, & j = 1, \ldots, (N/2) - 1 \\ a_j^2, & j = 0, N/2 \end{cases} \tag{1.71}$$

is basically the sum of two squared normal variables and thus follows a scaled χ-squared distribution with two degrees of freedom and an expectation value $S(\omega_j; \boldsymbol{\theta})$. The periodogram can be written as a scaled χ-squared distributed random variable with two

degrees of freedom (**?**)

$$I(\omega_j) = \frac{1}{2} S(\omega_j; \boldsymbol{\theta}) \zeta \,, \text{ with } \quad \zeta \sim \chi^2_2. \tag{1.72}$$

Exploiting these properties, we can specify Gaussian random numbers a_j and b_j such that (1.72) holds, i.e.

$$a_j, b_j = \frac{1}{2} S(\omega_j; \boldsymbol{\theta}) \xi \,, \qquad \text{for} \quad j = 1, \ldots, (N/2) - 1, \tag{1.73}$$

and

$$a_j = S(\omega_j; \boldsymbol{\theta}) \xi \,, \qquad \text{for} \quad j = 0, N/2, \tag{1.74}$$

where $\xi \sim \mathcal{N}(0, 1)$ is a Gaussian random variable. Using further a property of Fourier series $f(\omega_j)$ from real valued records $\{x_t\}_{t=1,\ldots,N}$: $f(-\omega_j) = f^*(\omega_j)$, we obtain a realisation of the process with spectral density $S(\omega_j; \boldsymbol{\theta})$ by the inverse Fourier series of a realisation of $Z_j = a_j + ib_j$ for $j = -N, \ldots, N$. Using the Fast Fourier Transform (FFT), this algorithm is fast at least for N being dyadic (order $\mathcal{O}(N \log N)$) and thus it is particularly interesting for simulating long records.

Besides the effect of aliasing (**?**), the periodicity of the realisation is a problem: the end of the record is highly correlated with the beginning. It is thus necessary to generate much longer records and extract a series of needed length from it.

Chapter 2

Model Selection

Entia non sunt multiplicanda praeter necessitatem.
William of Ockham (1295-1349)

The most crucial and delicate step in model building is the choice of an adequate model which satisfactorily describes the data. The notion of a satisfactory description is subjective and depends on the research questions. If physical understanding or explanation is the goal of the modelling efforts, *lex parsimoniae* – the principle of parameter parsimony – may be a guiding principle. It states that among models with the same explanatory power the one with fewer parameters is to be preferred. The principle dates back to the English Franciscan friar William of Ockham and is also known as Ockhams razor. If, on the other hand, prediction is the focus of the modelling effort, useful models in the sense of **?** are those with a high predictive power regardless of a physical interpretation; model complexity – in terms of number of parameters – is secondary in this respect.

The final aim of our modelling effort is the derivation of statistical quantities from time series such as trend and quantile estimates. For a reliable inference, we need the information about the dependence structure. Therefore, it is not the prediction of future observations which is in the focus, it is rather the reliable identification of the ACF. In this respect, we use lex parsimoniae as a guiding principle.

For model choice or model selection, we can identify two different concepts: goodness-of-fit tests and model comparison strategies. With a goodness-of-fit test, the hypothesis is tested that the observed record is a realisation of the model proposed. The second concept compares the fitness of two or more models in order to choose one or a few models out of a set of suitable models. Commonly, model selection involves goodness-of-fit tests and model comparison strategies. A plausible strategy is to base the model comparison on a set of models which pass the goodness-of-fit test. Within the family of the FARIMA$[p, d, q]$ processes, model selection reduces to choosing appropriate model orders p and q, and to deciding whether a fractional difference parameter d is needed or not.

This chapter presents a goodness-of-fit test particularly suitable within the framework of FARIMA$[p, d, q]$ models and Whittle maximum likelihood estimation. Further, we discuss standard model comparison strategies. These are approaches within a test-theoretical setting, such as the likelihood-ratio test, and criteria from information theory, such as the Akaike information criterion. Being confronted with non-nested models[1],

[1]The model g is nested in model f if it constrains one or more parameters of f, typically to zero. This notion is different from nested in the context of climate models.

non-Gaussian processes or simply situations where asymptotic results are not applicable, other than the standard approaches have to be considered. In this context, we take up an approach proposed by **?** and **?** and develop a model selection strategy for non-nested FARIMA$[p, d, q]$ models.

2.1 Goodness-of-Fit Tests

It seems plausible to expect from a "good" model a description of the data such that the difference between data and model output (the residuals) does not exhibit any structure of interest with respect to the modelling task. If the aim is to describe the correlation structure, it is plausible to demand for independent (or at least uncorrelated) residuals. This implies that there is no evidence for more (linear) structure in the underlying dynamics. Consequently many goodness-of-fit tests are based on testing the residuals for compatibility with a white noise process. A fundamental test in this respect is the Portmanteau test. It is based on the sum over the squared autocovariance series of the residuals. Several modifications have been proposed to improve the finite sample properties of the Portmanteau test. A comprehensive discussion can be found in **?**. Here, we restrict ourselves to a brief description of the basic Portmanteau statistic and a spectral variant which fits well in the framework of Whittle-based parameter estimation. A short note on hypothesis testing precedes the discussion.

2.1.1 Hypothesis Testing

Hypothesis testing is an algorithm to decide for or against an uncertain hypothesis minimising a certain risk. An example for a hypothesis is the "goodness-of-fit", i.e. H_0 "the observed data is a plausible realisation of the model". The information relevant for this hypothesis is summarised in a test statistic, e.g., the sum of the squared residual autocovariance series. Knowing the distribution of the test statistic under this *null hypothesis* H_0, we choose a *critical region* including those values of the test statistic which we consider as extreme and as evidence against the hypothesis. The probability of finding the test statistic under the assumption of H_0 in this critical region is called the *α-value* or *size* of the test. Common α-values are 0.01 and 0.05 corresponding to a 1% and 5%-level of significance.

The probability of the test statistic falling into the critical region under an alternative hypothesis H_A is called the *power* (pow) of the test; $1 - \text{pow}$ is referred to as the *β-value*. If the observed value falls inside the critical region, we reject the null hypothesis. If it is found outside we conclude that there is not enough evidence to reject H_0. Note, that this lack of evidence against H_0 does not imply evidence for the null hypothesis (**?**).

A test with low power cannot discriminate H_0 and H_A and is said to be not sensitive to the alternative hypothesis. For large α-values the test is referred to as not being specific; H_0 is frequently rejected even if it is true. The optimal test would be both, sensitive and specific.

If not stated otherwise, we use a 5%-level of significance in the following.

2.1.2 Portmanteau Test

Let $\{x_t\}_{t=1,\ldots,N}$ be the record under consideration and $f_t(\boldsymbol{\theta})$ the model output at index t. The parameter vector $\boldsymbol{\theta}$ has elements $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\intercal$. The residual series is denoted as

$r_t = x_t - f_t(\boldsymbol{\theta})$. The Portmanteau or Box-Pierce statistic is

$$Q_l = N \sum_{\tau=1}^{l} \widehat{\rho}^2_{\text{res}}(\tau), \tag{2.1}$$

with $\widehat{\rho}_{\text{res}}(\tau)$ being the autocorrelation sequence of the residuals r_t. Under the null hypothesis H_0 "the residuals are uncorrelated", Q_N is asymptotically χ-squared distributed with $l - m$ degrees of freedom (**??**).

2.1.3 Spectral Variant of the Portmanteau Test

? suggested a goodness-of-fit test for long-memory processes which is equivalent to the Portmanteau test but formulated in the spectral domain. The test statistic involves quantities which are necessarily calculated during the Whittle parameter estimation, thus it integrates smoothly into the framework. The test is a generalisation of a result from **?** for short memory processes. Define

$$A_N = \frac{4\pi}{N} \sum_j \left(\frac{I(\omega_j)}{\mathcal{S}(\omega_j; \boldsymbol{\theta})} \right)^2 \quad \text{and} \quad B_N = \frac{4\pi}{N} \sum_j \frac{I(\omega_j)}{\mathcal{S}(\omega_j; \boldsymbol{\theta})}, \tag{2.2}$$

with $S(\omega_j; \boldsymbol{\theta})$ being the spectral density of the model under consideration. $I(\omega_j)$ denotes the periodogram and the sums extend over all Fourier frequencies ω_j. The test statistic is now defined as

$$T_N(\widehat{\boldsymbol{\theta}}) = \frac{A_N(\widehat{\boldsymbol{\theta}})}{B^2_N(\widehat{\boldsymbol{\theta}})}. \tag{2.3}$$

It can be shown that $N^{1/2}(A_N(\widehat{\boldsymbol{\theta}}), B_N(\widehat{\boldsymbol{\theta}}))$ converges to a bivariate Gaussian random variable. We can simplify the test considering that $T_N(\widehat{\boldsymbol{\theta}})$ itself is asymptotically normal with mean π^{-1} and variance $2\pi^{-2}N^{-1}$:

$$P(T_N \leq c) \approx \Phi(\pi\sqrt{N/2}(c - (1/\pi))), \tag{2.4}$$

with Φ denoting the standard normal distribution function. **?** showed numerically that this approximation is acceptable already for moderate sample size of $N \approx 128$.

2.2 Model Comparison

A central idea in model selection is the comparison of the model residual variances, residual sums of squares or likelihoods. Intuitively, it is desirable to obtain a small residual variance or a large likelihood. However, increasing the number of parameters trivially reduces the variance. Consequently, one has to test whether a reduction in residual variance due to an additional parameter is random or a significant improvement. An improvement is called random, if the effect could have been induced by an arbitrary additional parameter. An effect is called significant if a special parameter improves the quality of the fit in a way which is not compatible with a random effect.

In the following, we present standard procedures from the test theoretical framework, such as the likelihood-ratio test, and criteria from information theory, such as the criteria of Akaike, Hannan and Quinn or Schwarz.

2.2.1 Likelihood-Ratio Test

The likelihood-ratio test (LRT) is based on the ratio of the two likelihood functions for two nested models f and g evaluated at the maximum. We assume that g is nested in f and constrains k-parameters. According to lex parsimoniae, we ask if the models perform equally well and take then the simpler one. The null hypothesis is now H_0 "g is an admissible simplification of f". A convenient asymptotic distribution for the test statistic is obtained using the log-likelihood functions l_f and l_g. The log-likelihood-ratio is then expressed as the difference

$$\mathrm{lr} = -2(l_g - l_f). \tag{2.5}$$

Under H_0 this test statistic is asymptotically χ-squared distributed with k degrees of freedom. Increasing the α-values with the sample size N leads to a consistent model selection strategy, i.e. the probability of identifying the right model approaches one with increasing sample size.

2.2.2 Information Criteria

A different approach was carried forward with a paper by **?**. It had tremendous influence on the way model comparison was carried out in the 1970s. Hitherto existing strategies relied on eye-balling or heuristic criteria. Due to increasing computational power it became possible to estimate parameters in models involving considerably more than two or three parameters. Therfore, it was now necessary to find a reliable procedure to choose among them.

Akaike Information Criterion

Akaike suggested a simple and easy to apply criterion. It is basically an estimate of the relative Kullback-Leibler distance to the "true" model. The Kullback-Leibler distance, or negentropy, is a measure of distance between a density function $p(\mathbf{x}|\boldsymbol{\theta})$ and the true probability density. It is thus different from the test theoretical approach described above. This measure was initially called "An Information Criterion" which later changed into *Akaike Information Criterion* (AIC). It is defined as

$$\mathrm{AIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + 2m, \tag{2.6}$$

with $l(\boldsymbol{\theta}|\mathbf{x})$ being the log-likelihood and m the number of parameters of the model under consideration. The term $2m$, frequently referred to as "penalty term", is indeed a bias correction. The model with the smallest AIC, i.e. the smallest distance to the true model, is typically chosen as the best. Within the framework of a multi-model approach, one might also retain all models within 2 of the minimum (cf. **?**).

In the decades after it has been proposed AIC enjoyed great popularity in many branches of science. One major reason why it was much more used than, e.g., Mallows' C_p criterion (**?**) developed at the same time might be its simplicity. Simulation studies, and later **?**, showed that the AIC systematically selects too complex models, which led to modified approaches.

Hannan-Quinn Information Criterion

A consistent criterion was later proposed by **?**. The difference to the Akaike criterion is a modified "penalty term" including the number of data points N

$$\mathrm{HIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + 2mc\log\log N, \tag{2.7}$$

with $c > 1$. This criterion is known as the *Hannan-Quinn information criterion* (HIC).

Bayesian-Schwarz Information Criterion

Another consistent criterion was formulated within the framework of Bayesian modelling by **?**:

$$\text{BIC} = -2l(\boldsymbol{\theta}|\mathbf{x}) + m \log N, \tag{2.8}$$

termed Schwarz information criterion or *Bayesian information criterion* (BIC).

2.2.3 Information Criteria and FARIMA$[p, d, q]$ Models

? investigated these three information criteria regarding order selection in FARIMA$[p, d, 0]$ processes and proved consistency for HIC and BIC regarding the selection of the autoregressive order p. An extensive simulation study was conducted by **?** also in the context of FARIMA$[p, d, q]$ but with nontrivial moving average order q. As a result **?** obtained that HIC and BIC can be quite successful when choosing among the class of FARIMA$[p, d, q]$ processes with $0 \leq p, q \leq 2$ for realisations coming from a FARIMA$[1, d, 1]$ process with $d = 0.3$. The rate of success, i.e. the fraction of correctly identified models, depends, however, heavily on the AR and MA components of the true model. It varies between 0.5% ($N = 500, a_1 = -0.5$ and $b_1 = -0.3$) and 99% ($N = 500, a_1 = 0.9$ and $b_1 = 0$). Although in many cases the rate of success is larger than 70%, in specific cases it can be unacceptable small.

Unfortunately, the information criteria do not provide any measure of performance by themselves. It is, however, desirable to identify situations with a success rate as low as 0.5%. In these situations one might either consult a different selection strategy or, at least, one wishes to be informed about the uncertainty in the selection. In situations which can be related to **?**'s simulation studies one might refer to those for a measure of performance. In other situations one has to conduct such studies for the specific situation at hand.

A further problem is that Akaike derived his information criterion for nested models (**??**) which makes it inappropriate when asking for a choice between, e.g., a FARIMA$[1, d, 0]$ and an ARMA$[3, 2]$. Reverting to a common model where both variants are nested in (here FARIMA$[3, d, 2]$) and testing the two original models for being admissible simplifications might result in a loss of power (**?**).

Confronted with these difficulties, we develop a simulation-based model selection approach for FARIMA$[p, d, q]$ processes which is particularly suitable for non-nested models.

2.3 Simulation-Based Model Selection

In the previous section, we have discussed some standard techniques to discriminate between nested models. In many practical data analysis problems one is often confronted with the task of choosing between non-nested models. Simple examples are linear regression problems, where we might have to decide between two different sets of explanatory variables (e.g., **?**). More complex examples include discriminating different mechanisms in cellular signal transduction pathways (e.g., **?**).

? was the first to consider the problem of testing two non-nested hypotheses within the framework of the likelihood approach and to derive an asymptotic result. Later a simulation approach was suggested by **?**. It is based on the idea to obtain a distribution

of the test statistic on the basis of the two fitted models. We take up this idea and develop and test a framework for discriminating between non-nested FARIMA$[p, d, q]$ processes.

2.3.1 Non-Nested Model Selection

? suggested a test statistic for discriminating non-nested models based on the log-likelihood. Consider $l_f(\widehat{\boldsymbol{\theta}})$ and $l_g(\widehat{\boldsymbol{\Xi}})$, the log-likelihood of two models f and g with parameter estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$, respectively. The test statistic is defined as

$$T_f = \left(l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x}) \right) - \mathrm{E}_{\widehat{\boldsymbol{\theta}}} \left[l_f - l_g \right], \tag{2.9}$$

where $\mathrm{E}_{\widehat{\boldsymbol{\theta}}} \left[l_f - l_g \right]$ denotes the expected likelihood ratio under the hypothesis H_f "model f is true". A large positive value of T_f can be taken as evidence against H_g "model g is true", and a large negative value as evidence against H_f. Note that the expression in (2.9) is not symmetric in f and g due to the expectation value in the last term. It can be shown that T_f is asymptotically normal under the null hypothesis H_f. The variance of the distribution of T_f is not known in general, which renders testing impossible. For specific situations, however, it is possible to calculate the variance; for some examples, see also **?**.

In the following section, we develop a simulation-based approach for discriminating models coming from the FARIMA$[p, d, q]$ class based on this test statistic.

2.3.2 Simulation-Based Approach for FARIMA$[p, d, q]$

Instead of (2.9) we consider only the difference $\mathrm{lr}_{\mathrm{obs}} = l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x})$ and compare it to the distributions of lr_f and lr_g (**?**). These two distributions are obtained similarly to the value for $\mathrm{lr}_{\mathrm{obs}}$ but from records $\mathbf{x}_f$ and $\mathbf{x}_g$ simulated with the respective model. If the two distributions of lr_f and lr_g are well separated, we can discriminate the two models. Depending on the observed record, we favour one or the other model or possibly reject both. A large overlap of the distributions indicates that the two models are hard to distinguish.

Schematically this procedure can be described as follows:

1. Estimate parameters $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Xi}}$ for models f and g from the observed record, calculate the log-likelihood-ratio $\mathrm{lr}_{\mathrm{obs}} = l_f(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - l_g(\widehat{\boldsymbol{\Xi}}|\mathbf{x})$.

2. Simulate R datasets $\mathbf{x}_{f,r}, r = 1, \ldots, R$ with model f and parameters $\widehat{\boldsymbol{\theta}}$. Estimate parameters $\widehat{\boldsymbol{\theta}}_{f,r}$ for model f and $\widehat{\boldsymbol{\Xi}}_{f,r}$ for model g for each ensemble member $\mathbf{x}_{f,r}$. Calculate $\mathrm{lr}_{f,r} = l_f(\widehat{\boldsymbol{\theta}}_{f,r}|\mathbf{x}_{f,r}) - l_g(\widehat{\boldsymbol{\Xi}}_{f,r}|\mathbf{x}_{f,r})$. This yields the distribution of the test statistic under the hypothesis H_f.

3. Simulate R datasets $\mathbf{x}_{g,r}, r = 1, \ldots, R$ with model g and parameters $\widehat{\boldsymbol{\Xi}}$. Estimate parameters $\widehat{\boldsymbol{\theta}}_{g,r}$ for model f and $\widehat{\boldsymbol{\Xi}}_{g,r}$ for model g for each ensemble member $\mathbf{x}_{g,r}$. Calculate $\mathrm{lr}_{g,r} = l_f(\widehat{\boldsymbol{\theta}}_{g,r}|\mathbf{x}_{g,r}) - l_g(\widehat{\boldsymbol{\Xi}}_{g,r}|\mathbf{x}_{g,r})$. This yields the distribution of the test statistic under the hypothesis H_g.

4. Compare the observed ratio $\mathrm{lr}_{\mathrm{obs}}$ to the distribution of $\mathrm{lr}_{f,r}$ and $\mathrm{lr}_{g,r}$. In case the distributions are well separated (Figure 2.1), this might yield support for the one or the other model (Figure 2.1, (a), solid line), or evidence against both of them (Figure 2.1, (b), dashed line). If the two distributions show a large overlap (Figure 2.2) models

f and g cannot be discriminated. Depending on the observed value, we find situations where we can still reject one model (Figure 2.2, b) and situations where no model can be rejected (Figure 2.2, a).

The representation of the distributions as density estimates or histograms (Figures 2.1 and 2.2, left) appear more intuitive because the separation or overlap of the distributions is directly visible. Density estimates or histograms require, however, additional parameters, such as the smoothing band width or the bin size. Therefore we prefer in the following the representation using the empirical cumulative distribution function (ECDF) (Figures 2.1 and 2.2, right).

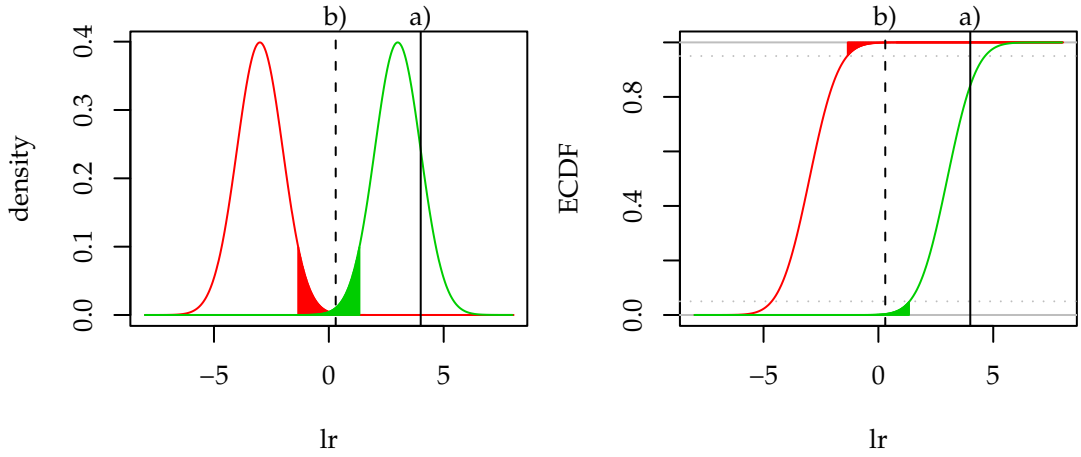

Figure 2.1: Possible outcomes of the simulation-based model selection. The two distributions are well separated. A one-sided 5% critical region is marked by the filled areas. The solid and dashed vertical line exemplify two possible values for an observed log-likelihood-ratio. The left plot shows the density function and the right plot the empirical cumulative distribution function.

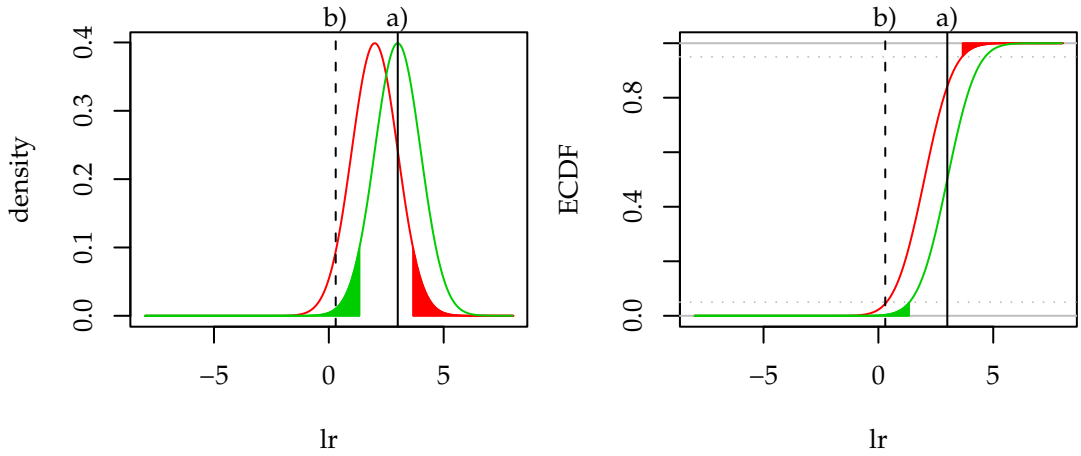

Figure 2.2: Possible outcomes of the simulation-based model selection. The two distributions are not well separated. The colour and line-coding is the same as in Figure 2.1.

The diagram in Figure 2.3 depicts the procedure in a flow-chart-like way. Besides a visual inspection of the result in a plot of the densities (histograms) or the empirical cumulative distribution functions, we can estimate critical regions, p-values and the power

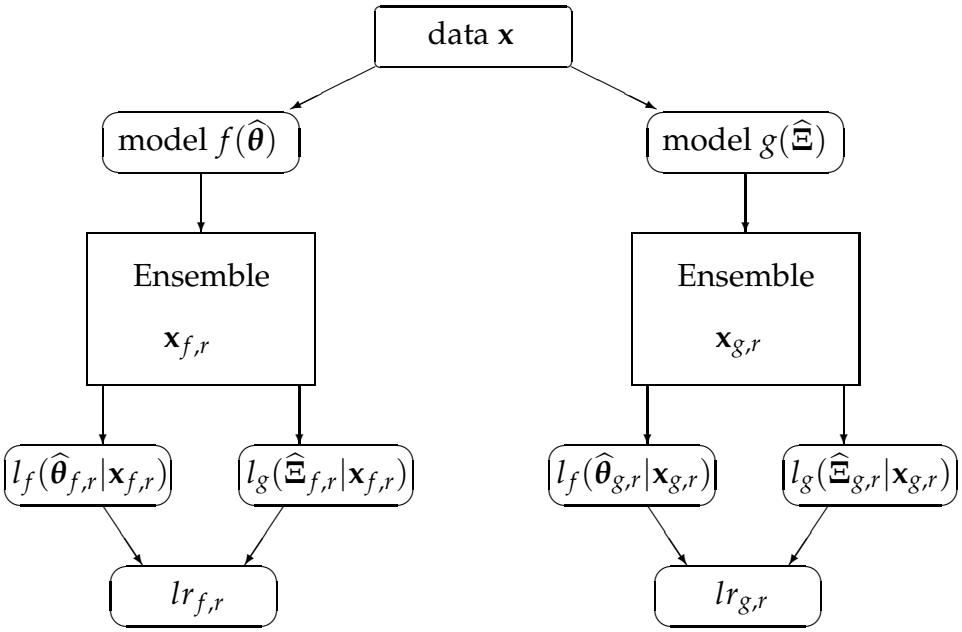

Figure 2.3: Schematic representation of the simulation-based model selection.

of the tests. Examples of this simulation-based model selection strategy in different settings are given by, e.g., **?** or **?**.

Critical Regions, p-values and Power

Critical Regions If we follow the specification (2.9) of the test statistic for testing the hypothesis H_f against H_g, we expect a deviation towards negative values if H_f is not true. To estimate the limit of a critical region, i.e. a critical value, at a nominal level α for R runs, we use

$$\widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,\alpha} = \mathrm{lr}_{f,([\alpha R])}\,. \tag{2.10}$$

$\mathrm{lr}_{f,(r)}$ denotes the r-th largest value and $[\alpha R]$ the integer part of αR.

Reversing the hypothesis and testing H_g against H_f, we expect a deviation towards positive values if H_g is not true and use $\mathrm{lr}_{g,([(1-\alpha)R])}$ as an estimate for the critical value.

p-values Furthermore, we might estimate p-values in a similar way. A one-sided p-value for testing hypothesis H_f against H_g can be estimated as

$$\widehat{p}_f(\mathrm{lr}_{\mathrm{obs}}) = \frac{\#(\mathrm{lr}_{f,r} < \mathrm{lr}_{\mathrm{obs}})}{R}, \tag{2.11}$$

where $\#(\mathrm{lr}_{f,r} < \mathrm{lr}_{\mathrm{obs}})$ denotes the number of likelihood ratios $\mathrm{lr}_{f,r}$ smaller than the observed value $\mathrm{lr}_{\mathrm{obs}}$.

Power The power $\mathrm{pow}_f(\alpha, g)$ of testing H_f against H_g associated with a specified level α is also straightforwardly estimated. The power is defined as the probability of finding the statistic under the alternative hypothesis in the critical region (**?**). An estimate for the power is thus given by

$$\widehat{\mathrm{pow}}_f(\alpha, g) = \frac{\#(\mathrm{lr}_{g,r} < \widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,\alpha})}{R}. \tag{2.12}$$

2.3.3 An Illustrating Example

As an example consider the problem of discriminating between an AR[1] and a FD process on the basis of a time series of length $N = 400$. We start with a realisation of an AR[1] process and consider this series as an observed one. This time series is depicted in Figure 2.4 in the time domain as well as in the spectral domain. We suppose that the un-

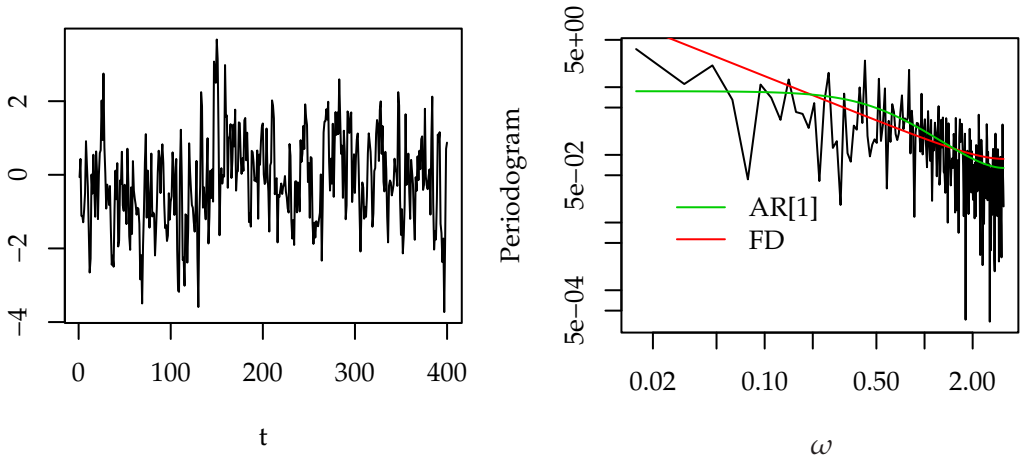

Figure 2.4: Realisation of an AR[1] process with $a_1 = 0.6, \sigma_\eta^2 = 1$ and length $N = 400$ in the time domain (left) and spectral domain (right) together with the spectral densities of the AR[1] (green) and FD (red) processes.

derlying process is either AR[1] or FD and we use the simulation-based model selection to discriminate the two models.

Fitting the AR[1] and the FD Model

Following the procedure given above, we use the Whittle estimator (Section 1.3.2) to obtain the following parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{a}_1, \widehat{\sigma}_{\eta,f})^\intercal$ and $\widehat{\boldsymbol{\Xi}} = (\widehat{d}, \widehat{\sigma}_{\eta,g})^\intercal$ for model f and g, respectively. This results in $\widehat{a}_1 = 0.57(0.04)$, with the value in parentheses denoting the standard deviation, and $\widehat{\sigma}^2_{\eta,f} = 0.995$ for the AR[1] process (model f); $\widehat{d} = 0.47(0.04)$ and $\widehat{\sigma}^2_{\eta,g} = 1.049$ for the FD process (model g).

In both cases the goodness-of-fit test (2.4) does not reject the model on the 5%-level of significance: the p-values obtained are $\widehat{p}_f = 0.551$ and $\widehat{p}_g = 0.063$ for the AR[1] and FD process, respectively, and thus both are larger than 0.05. This implies that we are not able to discriminate the processes solely on the basis of this test. The periodogram of the original series is compared to the spectral density of the fits in Figure 2.4 (right).

The Simulation-Based Model Selection

Following the scheme depicted in Section 2.3.2, we are left with two sets of log-likelihood-ratios: $\mathrm{lr}_{f,r}$ obtained from the ensemble $\mathbf{x}_{f,r}$ with the underlying process being AR[1] and $\mathrm{lr}_{g,r}$ obtained from the ensemble $\mathbf{x}_{g,r}$ with the underlying process being FD. The two hypotheses are shown in Figure 2.5 as histograms (left) and cumulative distribution functions (right). The likelihood ratio calculated for the original series $\mathrm{lr}_{\mathrm{obs}}$ (black vertical line) is located close to the centre of H_f and in the 5% critical region of H_g. This thus supports the AR[1] in favour of the FD process.

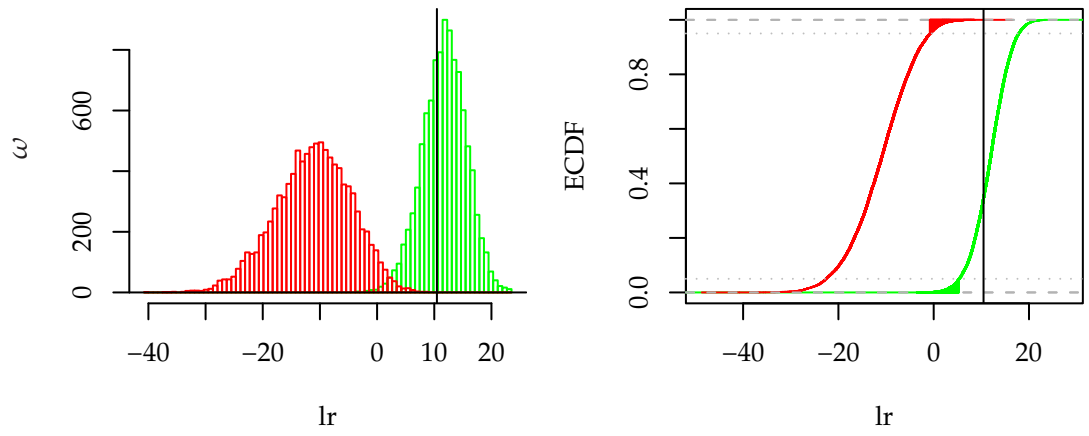

Figure 2.5: Distributions of the log-likelihood-ratio for the AR[1] (lr_f, green) and the FD (lr_g, red) model as histogram (left) and cumulative distributions function (right). The critical regions are indicated as filled regions and the log-likelihood-ratio for the original series is shown as black vertical line. The distributions are obtained from $R = 10\,000$ runs.

The visualisation of the hypotheses in Figure 2.5 provides us with an idea to what extent the two models are distinguishable and which hypothesis to reject. In this example the distributions are well separated and the observed value is close to the centre of the AR-hypothesis but at the edge of the distribution resulting from the FD model. Additional to this visual impression, we estimate also critical regions, the power of the tests and p-values in the way described above.

Estimating Critical Regions Consider the AR[1] model as the hypothesis under test and the FD as the alternative. For the given example using (2.10) we find a critical value at $\alpha = 0.05$ of $\widehat{\mathrm{lr}}^{\mathrm{crit}}_{f,0.05} = \mathrm{lr}_{f,([500])} = 5.393$. The critical region extends towards smaller values and thus the observed value $\mathrm{lr}_{\mathrm{obs}} = 10.463$ lies outside. Thus the hypothesis H_f "the realisation stems from an AR[1] model", cannot be rejected

For the inverse situation of testing FD against the alternative AR[1], we find a lower limit of a critical region at $\widehat{\mathrm{lr}}^{\mathrm{crit}}_{g,0.05} = \mathrm{lr}_{g,([500])} = 1.115$. The observed value is thus inside and we can reject the hypothesis that the realisation stems from an FD model.

Estimating the Power of the Test Considering again the AR[1] model as the hypothesis under test, with (2.12) we obtain an estimate $\widehat{\mathrm{pow}}_f(\alpha = 0.05, g) = 0.9978$. Simultaneously, we have obtained an estimate $\beta = 1 - \widehat{\mathrm{pow}}_f(\alpha = 0.05, g) = 0.0022$. This estimate of the power being close to one is a result of the distributions being well separated. This high power implies that H_f (AR[1]) is very likely rejected for realisations from an FD process (H_g).

The corresponding power estimate for the inverse situation amounts to $\widehat{\mathrm{pow}}_g(\alpha = 0.05, f) = 0.999$.

Estimating p-Values For the AR[1] model we estimate a p-value $\widehat{p}_f = 0.3533$ using (2.11). We thus have no evidence to reject this hypothesis. Since in our example, no $\mathrm{lr}_{g,r}$ exceeds

the observed likelihood ratio, we find $\widehat{p}_g = 0$ and thus strong evidence against the FD hypothesis.

On the basis of the above procedure we can thus very confidently reject the hypothesis of the data being generated from an FD model in favour of the autoregressive alternative. In the given example, we know the underlying process and find that it was correctly identified.

In order to investigate whether the simulation-based model selection strategy can be referred to as a statistical test, we have to evaluate whether the rate of rejection corresponds to the nominal level α.

2.3.4 Testing the Test

In the setting presented above, we straightforwardly estimated critical regions, p-values and the power, for testing simple hypotheses, i.e. we assume that the null as well as the alternative hypotheses are models with known parameters, namely the ML estimates for the specified models. This is different from the situation where the parameters are not known or, in other words, from testing composite hypotheses, i.e. the record is either compatible with an AR[1] or with an FD process. In the latter situation the parameters are not known but have to be estimated. This is the question we actually have to address. Thus we test the size of such a test in a simulation study. In the current setting, this implies to investigate whether the rejection rate according to the estimated p-values correspond to the nominal rate α. We furthermore investigate to what extend the estimate of the power is a useful approximation.

Testing the Size

We use again the AR[1] example process and perform an extensive simulation study using different values for the parameter $a_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and various sample sizes $N \in \{100, 200, 400, 800, 2\,000, 4\,000\}$[2]. For a_1 and N fixed, we generate $M = 1\,000$ realisations. For each such realisation we use the simulation-based model selection (again with 1 000 runs) to estimate p-values for a test against a FD model. The fractional difference parameter d of this alternative hypothesis is estimated from the realisations. On the basis of this p-value for a given realisation, we can either reject or accept the null hypothesis on a prescribed nominal level of significance α. The fraction of the number of rejections of the total number of realisations M is called the rejection rate r_α. For a statistical test this rate must agree with the nominal level α for large M, or, equivalently, $r_\alpha/\alpha \to 1$ for $M \to \infty$.

Figure 2.6 shows the relative rate of rejection r_α/α for three nominal levels of α. In the four upper panels (a-d) the dependence on the sample size N is given for four values of a_1. The two bottom panels (e,f) depict the dependence on the parameter a_1 for a small and a large sample size. The largest deviations from unity in all panels can be observed for a nominal level of $\alpha = 0.01$. For this level, we expect only ten of the 1 000 realisations to be rejected. Thus a deviation of 50% implies five instead of ten rejected realisation. In this range the statistic is very limited.

The plot for the N-dependence with a small autoregressive parameter (a_1=0.1, panel a), suggests that an increasing sample size reduces the difference between nominal level

[2]Obtaining results for larger N, e.g., $N = 8\,000$ could not be accomplished within one week of computation time for a given a_1 and $M = 1\,000$ runs on one node of the IBM p655 Cluster equipped with IBM Power4 CPUs with 1.1 GHz.

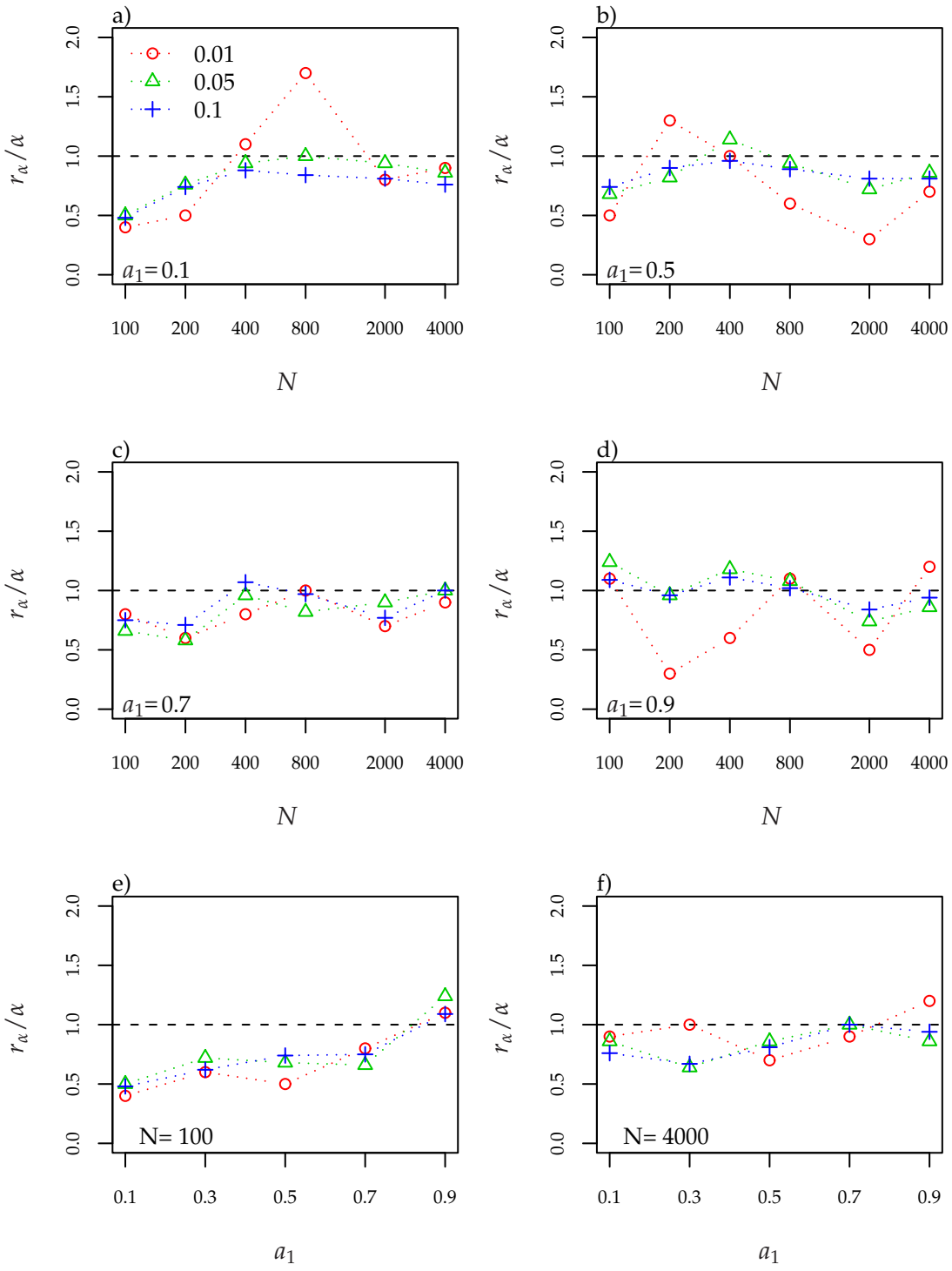

Figure 2.6: Relative rate of rejection at the 1% (circles), 5% (triangles) and 10% (crosses) nominal levels of significance. The four top panels show the dependence on the length N for $a_1 = 0.1$ (a), $a_1 = 0.5$ (b), $a_1 = 0.7$ (c) and $a_1 = 0.9$ (d). The two bottom panels show the dependence on the AR-parameter a_1 for $N = 100$ (e) and $N = 4\,000$ (f).

and calculated rejection rate. For larger a_1, the sample size does not have a clear influence anymore (panels b-d). Except for $a_1 = 0.9$, the rejection rates are predominantly smaller than the corresponding nominal level. The a_1-dependence for small sample size ($N = 100$) (panel e) shows that the rejection rate increases with the autoregressive parameter a_1. For $a_1 = 0.1$, we find the rejection rate being about 50% smaller than the nominal level. Only for $a_1 = 0.9$ more realisations of the AR[1] process are rejected than expected from the value of α. For a larger sample size ($N = 4\,000$), the mismatch between nominal level and rejection rate reduces, especially for small a_1 (panel f). In this case, we observe a relative rejection rate $r_\alpha/\alpha < 1$ for $a_1 = 0.9$ and $\alpha \in \{0.05, 0.1\}$.

In situation where we find a rejection rate smaller than the nominal level, we speak of a conservative test; the rejection under the null hypothesis is not as likely as expected. In the opposite case of a rejection rate larger than the nominal value, one is more likely to reject a true null hypothesis than expected. For the present example, this implies that the AR hypothesis is rejected in favour of an FD process. The consequence is falsely detecting LRD. Thus in the case of discriminating between an AR[1] and an FD process, special attention has to be paid to processes with large autoregressive parameters.

For the situation of the example under consideration ($a_1 = 0.6$ and $N = 400$), the relative rejection rates are close to one for $N = 400$ and $a_1 = 0.5$ ($r_{0.05}/0.05 = 1.14$) or $a_1 = 0.7$ ($r_{0.05}/0.05 = 0.96$). We can thus expect that the selection procedure is close to a statistical test in this case.

Estimating the Power

With further simulations we study the estimates of the power for the simulation-based model selection. The power is the rate of rejection for records from the alternative hypothesis. Thus we need to study realisations of a suitable process serving as such. In this setting this is an FD process with a difference parameter d chosen adequately to the corresponding null hypothesis. This parameter changes with the length N and the parameter a_1 of the AR[1] process. Thus, we estimate d from 20 realisations of the null hypothesis AR[1] for given a_1 and N and take the average value for the alternative hypothesis. With this alternative hypothesis we generate 1 000 realisations and obtain an estimate for the power according to (2.12) for each run. This sample of power estimates is then compared to the rate of rejection obtained from the ensemble by means of a box-plot. The results for various a_1 and N is shown in Figure 2.7. The blue crosses mark the difference parameter used for the alternative hypothesis. The power increases with N but in a different manner for different a_1. The most rapid increase can be observed for $a_1 = 0.5$ and $a_1 = 0.7$. In general the power estimates are very variable in regions of low power and especially for parameters $a_1 = 0.1$ and $a_1 = 0.9$ where the AR[1] and the FD process have quite similar spectral densities. For the other two parameter values ($a_1 = 0.5$ and $a_1 = 0.7$) the estimate of the power is on average good. Particularly, for the setting of the example studied above ($a_1 = 0.6$, $N = 400$), the power estimate close to one should be reliable.

In some cases, we find a fractional difference parameter $d > 0.5$ which corresponds to the non-stationary domain. We are, however, still able to obtain estimates and an approximate likelihood as discussed in Section 1.3.2.

In case we cannot distinguish the two models, for example due to a low power, we can obtain a rough estimate of the time series length needed for a satisfactorily power on the basis of simulations, as described in the following.

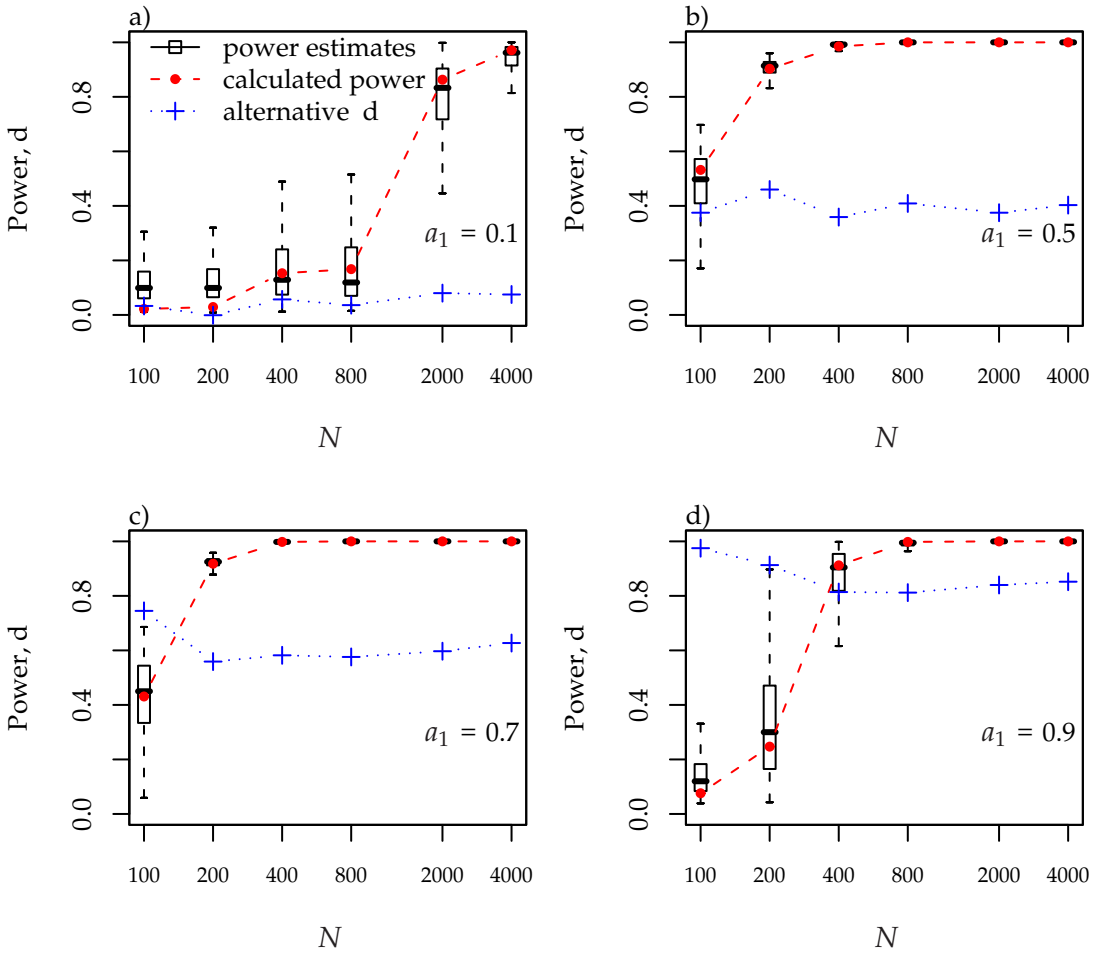

Figure 2.7: Box-plot of the power estimates for 1 000 runs dependent on the record length N. The power obtained from the simulation is marked by the red dots and the fractional difference parameter used for the alternative hypothesis as blue crosses. Different values of a_1 are shown: 0.1 (a), 0.5 (b), 0.7 (c) and 0.9 (d).

2.3.5 Estimating the Required Sample Size

The power of the test changes with the length of the series under consideration. This is depicted in Figure 2.8. The model parameters used in the simulations are the ML estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\Xi}$ obtained for the original time series length.

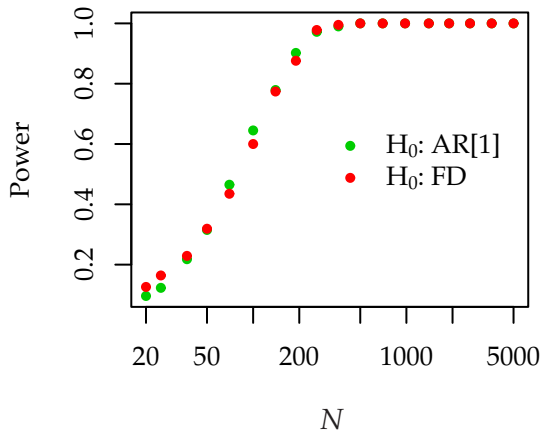

Figure 2.8: Power of testing the AR[1] hypothesis (green) and FD hypothesis (red) calculated for various lengths.

In cases where it is not possible to discriminate the models, i.e. the length of the observed record does not lead to satisfactory power of the test, we might repeat the procedure with increasing length of the simulated series. We then get an idea what record length is needed to distinguish the two models. Interpretation of this result has to be treated with care, because on the one hand, we use the estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\Xi}$ instead of the "true" parameter values. These estimates depend on the sample size N. On the other hand, every hypothesis is trivially falsified for a sufficient length of the time series. An increasing power does thus not mean that we are able to decide for one model; we might end up with rejecting both. However, for a moderate increase of sample length, we might still get an appropriate idea of the power.

In the context of geophysical records the lack of data basically implies to postpone the analysis until sufficient data has been collected. In other settings, with the possibility of influencing the size of the sample beforehand, this approach might be useful to design the experiment. This ensures that a sufficient size of the sample can be taken. Then the research question can be answered while an unnecessary surplus of data collection can be avoided, e.g., to reduce costs.

2.3.6 Bootstrapping the Residuals

Within the framework of FARIMA$[p, d, q]$ models, the ensemble members $\mathbf{x}_{f,r}$ or $\mathbf{x}_{g,r}$ can be generated from a white noise process η_t using a linear filter (cf. Section 1.2.1)

$$x_t = \sum_{i=0}^{M} \gamma_i \eta_{t-i}. \qquad [1.39]$$

In general the sum (1.39) extends over infinitively many elements M. In practical application, the filter is truncated at a suitable length. The γ_i can be derived from the model parameters (**??**).

For invertible models (1.39) can be inverted to obtain the residual series $\{r_t\}$ from the time series $\{x_t\}$ given the model. In situations where the distribution of the residual series is not close to a Gaussian, we might consider replacing the realisation $\{\eta_t\}$ of the white noise process with a realisation of a bootstrap resampling of the residuals $\{r_i\}$ (**??**). In this way the distribution of the residuals is conserved.

2.4 Summary

We started the discussion on model selection with an example of a goodness-of-fit test: a spectral variant of the Portmanteau Test. This test is particularly suitable for the kind of models used and parameter estimation strategy we pursue here. For model comparison, we suggested the standard likelihood-ratio test and the Akaike-type information criteria (AIC, BIC, HIC). These methods can be used for choosing between nested models.

Starting from a general likelihood-ratio approach, we developed a simulation-based strategy for the discrimination of non-nested FARIMA$[p, d, q]$ models. The formulation of this approach is within the framework of statistical hypothesis testing, i.e. the concepts of critical regions, p-values, and power are used. We illustrated this method with a simulated example series from an AR$[1]$ process. On the basis of one realisation, the AR$[1]$ model could correctly be identified as the underlying process. The FD model as alternative hypothesis could be rejected, with a power close to one.

In order to study to what extend the estimates provided for the p-values and the power can be interpreted in the sense of a statistical test, we performed an extensive simulation study related to the above example. The results showed a dependence of the relative rejection rate on the parameter value a_1 and on the sample size N. The quality of the power estimates depend as well on the sample length and on the parameter values. We find useful estimates of the power in regions of high power and especially when the spectral densities of the AR$[1]$ and the FD process are not similar (here $a_1 = 0.5$ and $a_1 = 0.7$). In regions of low power and especially towards the boundaries of the interval $(0, 1)$ for a_1, estimates become highly variable and are to be interpreted with care.

If we cannot discriminate between the two models, a simulation study with an increased record length gives an idea how many data points are needed for a reliable decision. Because the simulations with increased length will be based on the parameters estimated for the observed length, the range of this analysis is limited.

Given the idea of generating ensembles from the two hypotheses under test, one can think of other statistics than the likelihood ratio. Especially in the context of detecting LRD, statistics focusing on the low frequency behaviour might be suitable. Two statistics, one based on the log-periodogram regression and one based on the DFA are discussed in Appendix B.4. In the example presented here, the likelihood ratio as test statistic results, however, in the more powerful test.

In the following chapter, we develop a strategy to detect LRD on the basis of parametric modelling and the model selection strategies presented and developed here.

Chapter 3

Detection of Long-Range Dependence

"Nothing is what it seems!"
Morpheus to Neo in *The Matrix*, 1999

The phenomenon of long-range dependence (LRD, Section 1.1.4) has received attention since the 1950s, the time the British hydrologist H. E. Hurst studied the Nile River minimum flows, until today (**???????**). Its discussion is not limited to hydrology but is prevailing in other fields of the geosciences as well as in economics (e.g., **?????**). With this variety of applications different ways of detecting and describing the phenomenon arose. We can basically distinguish two strategies for the detection of LRD.

1. The historically older one is the search for a certain asymptotic behaviour in the autocorrelation function, in the spectrum, or similar statistics (Section 1.1.4). These approaches consist of a simple parametric description of the asymptotic behaviour of the statistic under consideration in regions where these asymptotics are assumed to hold. Such strategies are referred to as semi-parametric. If the asymptotic behaviour is sufficiently assured to hold and if the properties of the estimator are known, as it is the case for the log-periodogram regression (Appendix. B.3.2), inference about LRD can be made (**?**). For estimators with unknown limiting distributions, as, e.g., DFA-based estimation (Section 1.3.3), inference in not possible straightforwardly. Such approaches are referred to as heuristic; they are useful for a preliminary analysis but not for statistical inference (**?**).

2. An alternative strategy is a full-parametric modelling approach including model selection as outlined in Chapter 2. The difference to semi-parametric or heuristic approaches is that non-asymptotic behaviour is not disregarded but explicitly modelled. Within the likelihood framework a limiting distribution for the estimator is available, and thus inference about the parameter values – in particular about the fractional difference parameter d – is possible (Section 1.3). More subtle is the model selection discussed in Chapter 2. The inference about model parameters, such as the fractional difference parameter, is only meaningful if the proper model is used. The problem of selecting a suitable model corresponds to selecting an adequate region for the log-periodogram regression (**?**): a small region as well as a complex model lead to a smaller bias but brings along a larger variance for the estimate, and vice versa.

In this chapter, we develop a strategy for the detection of LRD based on the full-parametric modelling approach and the model selection strategy proposed in Chapter 2. In the first section, we generate two example series: one from a LRD and one from a SRD process. The latter is constructed such that its realisations with a given length can be easily mistaken for realisations of the LRD process. In the subsequent sections, these series are analysed with DFA and the log-periodogram regression. Finally, we develop the full-parametric strategy and illustrate its abilities to discriminate SRD and LRD with these example series.

3.1 Constructing the Example Process

Contrary to the example studied in Section 2.3.3, we consider two processes which are similar in their high frequency behaviour and differ only in the low frequency region, i.e. one is LRD, the other SRD. Realisations of them are then taken as "observed" records. Based on these "observations" we aim to decide whether the underlying process is SRD or the LRD.

In order to strongly challenge the various model selection strategies, we construct a SRD process such that it mimics a given LRD process as closely as possible. To this end, we generate a realisation of a LRD model and search for an SRD model representing this realisation reasonably well. Within the framework of the FARIMA$[p, d, q]$ family, a simple LRD process is the FARIMA$[1, d, 0]$ (or, for short, FAR$[1]$) process. A realisation of this process can be reasonably well modelled with an ARMA$[3, 2]$ process (for details, cf. Appendix B.5). We expect a realisation of this SRD model to be easily mistaken for stemming from the FAR$[1]$ process. It is thus an ideal candidate to illustrate and test detection strategies for LRD.

Figure 3.1 shows the realisation of the ARMA$[3, 2]$ (green) and the FAR$[1]$ process (red) in the time and frequency domain. Both realisations have $N = 2^{15} = 32\,768$ data points.

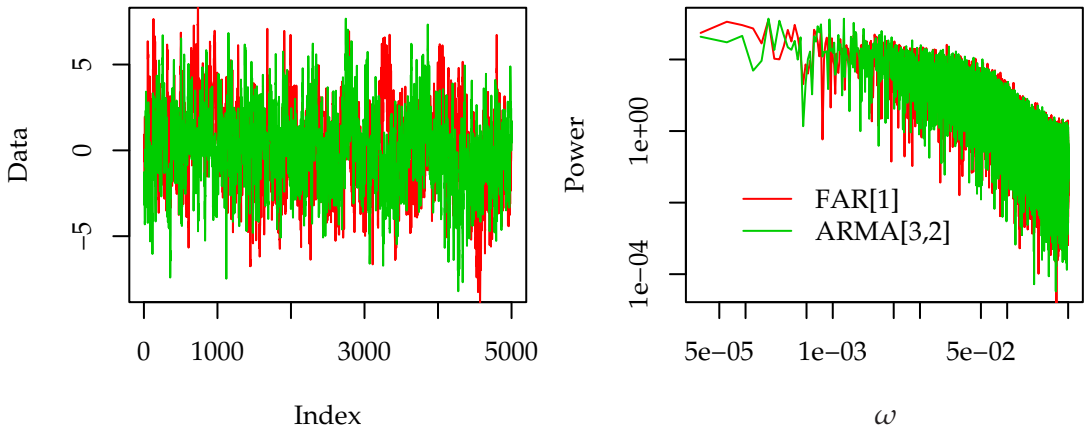

Figure 3.1: Time domain (left) and frequency domain (right) representation of the ARMA$[3, 2]$ realisation (green, $a_1 = 2.178, a_2 = -1.488, a_3 = 0.309, b_1 = -1.184, b_2 = 0.218$) compared to the FAR$[1]$ realisation (red, $d = 0.3, a_1 = 0.7$).

For the analyses in the following sections, we consider these two records as "observed" data and try to infer the underlying processes.

3.2 Detrended Fluctuation Analysis

For the detection of LRD it has become a widely used practice to compare the DFA fluctuation function (Section 1.3.3) in a double-logarithmic plot to an expected asymptotic slope for SRD or uncorrelated processes (e.g., **????**). For the "observed" records, obtained from the ARMA$[3,2]$ and the FAR$[1]$ process, we depict these fluctuation function in Figure 3.2. From a visual inspection of the left panel one might be tempted to consider the slopes H

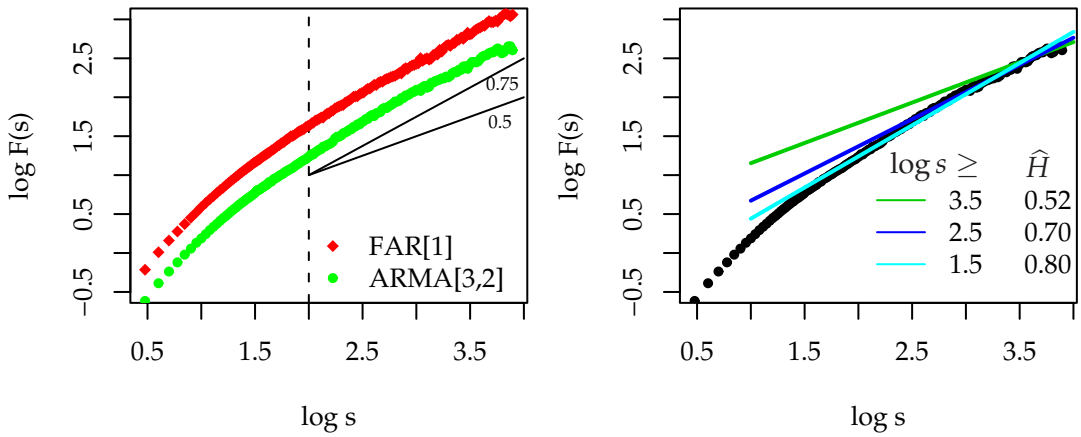

Figure 3.2: DFA1 fluctuation functions in double logarithmic representation for the ARMA$[3,2]$ (green) and the FAR$[1]$ (red) realisation (left). The two solid black lines mark the asymptotic behaviour expected for SRD, or uncorrelated, processes ($H = 0.5$) and for a LRD process with $H = 0.75$. The dashed vertical line marks the lower bound of scales considered. Straight line fits for various ranges of scales to the DFA1 fluctuation function of the ARMA$[3,2]$ realisation are shown in the right panel.

of both fluctuation functions being approximately constant with $H > 0.5$ for large scales s. For $\log s > 2$, for example, the plot suggests that in both cases the slopes are rather compatible with $H = 0.75$ than with $H = 0.5$. Thus, this provides no evidence for an SRD process in neither case. Following this line of argument, we would falsely infer a LRD process with $H = 0.75$ underlying both realisations. This demonstrates that a reliable discrimination of realisations from LRD or SRD processes is not easily achieved on the basis of DFA. Consequently such results should be interpreted with care.

A more detailed analysis of the fluctuation function of the ARMA$[3,2]$ realisation in the right panel of Figure 3.2 shows already that the slope systematically increases with decreasing lower bound of scales for the fit (i.e. an increasing range of scales considered). This should raise awareness and provides first evidence that the asymptotic behaviour has not been reached. If it had been reached, the slope should fluctuate around a constant value but should not exhibit a systematic variation. A critical review on the use of DFA to infer LRD explicitely addressing this point is given in Appendix A together with a detailed analysis of bias and variance for Hurst exponent estimation.

3.3 Log-Periodogram Regression

Contrary to a DFA-based estimator, a limiting distribution does exist for the estimator $\widehat{d}_{\text{lp}}$ based on the log-periodogram regression (Appendix B.3.2). We can thus provide confidence intervals and test whether the estimate for the fractional difference parameter is

compatible with zero. For both realisations, we perform the log-periodogram regression for various bandwidths and give a 95% confidence interval for the estimate $\widehat{d}_{\text{lp}}$. Figure 3.3 illustrates the regression for the ARMA$[3,2]$ realisation (left) and the FAR$[1]$ realisation (right). The estimates $\widehat{d}_{\text{lp}}$ with the corresponding asymptotic 95% confidence intervals

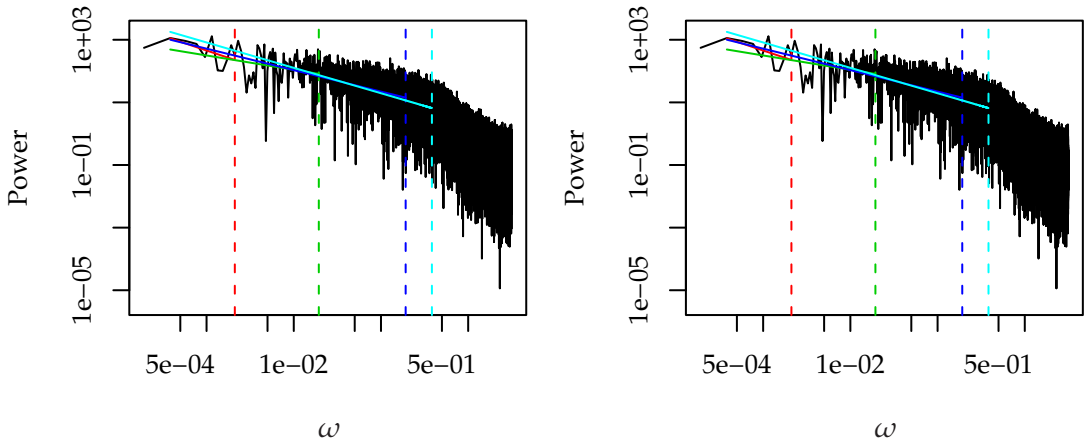

Figure 3.3: Log-periodogram regression for various bandwidths (different colours) for the ARMA$[3,2]$ (left) and the FAR$[1]$ realisation (right).

are shown in Figure 3.4. For the ARMA$[3,2]$ realisation we observe an increase of the

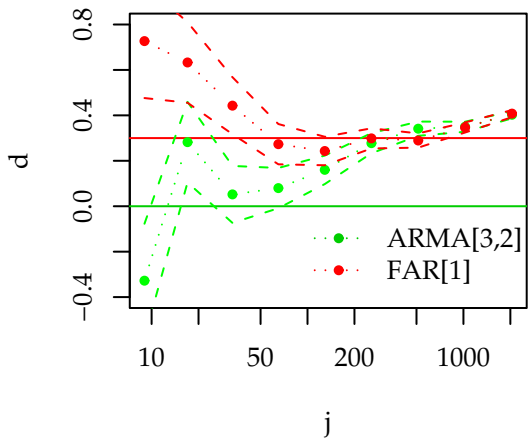

Figure 3.4: Log-periodogram regression estimates d_{lp} of the fractional difference parameter d for different bandwidth obtained from the ARMA$[3,2]$ realisation (green) and the FAR$[1]$ realisation (red). The abscissa specifies the bandwidths as index j of the largest Fourier frequency ω_j included in the fit. The green and red dashed lines specify the asymptotic 95% confidence intervals around the corresponding estimates. The corresponding solid lines mark the true value of the fractional difference parameter.

estimate $\widehat{d}_{\text{lp}}$ changing from negative to positive values. It thus passes $d = 0$. Furthermore, the 95% confidence interval for three of the smallest four bandwidths encloses the value $d = 0$ expected for SRD processes. This is different for the FAR$[1]$ realisation: the estimates are all strictly positive and not compatible with $d = 0$ within a 95% confidence interval. In this case, we find, as expected, no evidence of SRD. Depending on the choice

of the bandwidth we get ambiguous results for the ARMA$[3,2]$ realisation. Here, the inference of the decay of the ACF (i.e. SRD or LRD) depends on the choice of the bandwidth. As mentioned before, the problem of choosing an optimal bandwidth is comparable to the model selection problem. Optimal bandwidth choice is discussed in, e.g., **?**.

3.4 Full-Parametric Modelling Approach

Instead of describing only the asymptotic behaviour for large time-scales/low frequencies from an empirical data set, a different approach for the detection of LRD is proposed in the following. We rephrase the question as a model selection problem: *Is the empirical record best represented by a LRD or a SRD model?* This implies that we need to impose some kind of principle to make the selection of a "best" model well defined. The principle we follow can be ascribed to *lex parsimoniae* – the principle of parameter parsimony. In a modelling framework, it can be expressed as follows: a model should be as complex as necessary but not more complex (Chapter 2). We thus have to decide about the necessary model complexity. This is accomplished by model selection strategies described in Chapter 2.

In order to discriminate a more complex SRD (ARMA$[p,q]$) from a simpler LRD (FARIMA$[p',d,q']$) model, we might require non-nested model selection, for example FARIMA$[1,d,0]$ and ARMA$[3,2]$. Alternatively, it is possible to use the simplest common model (here FARIMA$[3,d,2]$) and test with the standard likelihood-ratio test whether one or the other model is an admissible simplification of the common model. According to **?** those tests have potentially lower power than a direct comparison of the two non-nested models. It is thus convenient to have a strategy for non-nested model selection at hand.

For both "observed" records, we estimate parameters for an ARMA$[3,2]$, a FAR$[1]$, and a FARIMA$[3,d,2]$ process. The two simple models are then tested for being an admissible simplification of the FARIMA$[3,d,2]$ using the standard likelihood-ratio test (Section 2.2.1). A direct comparison of the ARMA$[3,2]$ and the FAR$[1]$ is finally realised with the simulation-based model selection (Section 2.3).

3.4.1 Modelling the ARMA$[3,2]$ Realisation

We begin with estimating model parameters for the three processes considered as potential models for the ARMA$[3,2]$ realisation. The resulting parameters and p-values for the goodness-of-fit test (Section 2.1.3) are listed in Table 3.1. The parameter estimates

Parameter	ARMA$[3,2]$	FAR$[1]$	FARIMA$[3,d,2]$
d	-	$0.300(0.015)$	$0.316(0.022)$
a_1	$2.182(0.041)$	$0.705(0.014)$	$1.750(0.531)$
a_2	$-1.493(0.078)$	-	$-1.094(0.883)$
a_3	$0.309(0.036)$	-	$0.248(0.387)$
b_1	$-1.179(0.043)$	-	$-1.062(0.528)$
b_2	$0.210(0.042)$	-	$0.372(0.552)$
p-val	0.946	0.931	0.935

Table 3.1: Parameters and asymptotic standard deviation in parentheses for the ARMA$[3,2]$, FAR$[1]$ and the FARIMA$[3,d,2]$ model estimated from the ARMA$[3,2]$ realisation.

for the FAR$[1]$ model are, within one standard deviation, compatible with the parameters of the original FAR$[1]$ process used to motivate the ARMA$[3,2]$ model. Likewise,

the ARMA$[3,2]$ estimates are compatible with the parameters of the underlying process. The standard deviation of the FARIMA$[3,d,2]$ parameter estimates is about one order of magnitude larger than for the other two models, except for d. Taking these standard deviations into account, the AR and MA parameters are compatible with those from the ARMA$[3,2]$. Similarly, the estimate for the fractional difference parameter d is compatible with the corresponding parameter of the original FAR$[1]$ process. None of the three models can be rejected on any reasonable level of significance according to the goodness-of-fit test. Furthermore, $d = 0$ is not within the asymptotic 95% confidence interval for the estimate $\widehat{d}_{\text{lp}}$ for both fractional models.

In Figure 3.5 (left), we compare the spectral densities of the two competing models, FAR$[1]$ and the ARMA$[3,2]$, to the periodogram of the "observed" series. The spectral

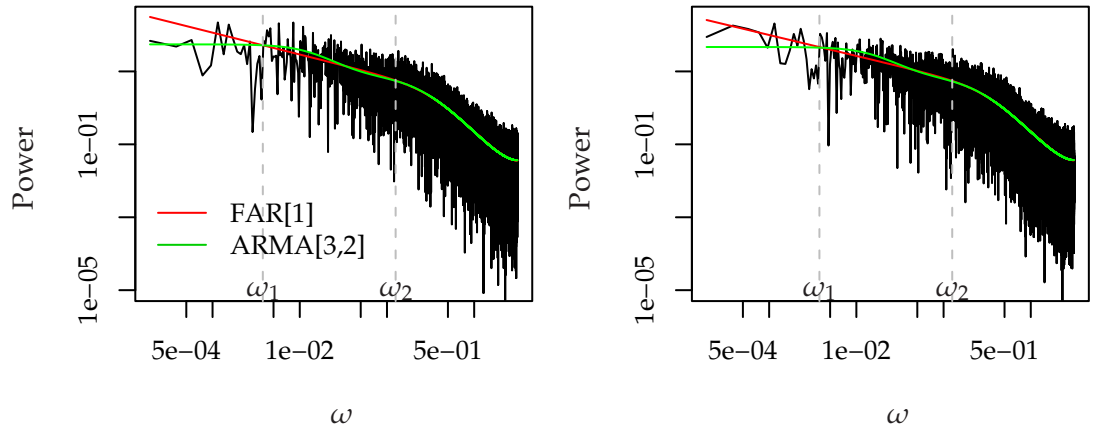

Figure 3.5: Periodogram of the ARMA$[3,2]$ (left) and FAR$[1]$ (right) realisation together with the spectral densities of the ARMA$[3,2]$ (green) and FAR$[1]$ (red) model fitted to these realisation. The dashed vertical lines mark regions on the frequency axis where the spectral densities diverge (left), are similar (middle) and are quasi identical (right).

densities are quasi indistinguishable for frequencies $\omega \gtrsim \omega_2 = 0.13$. Differences are visible in the range $4 \cdot 10^{-3} = \omega_1 \gtrsim \omega \gtrsim \omega_2$. Here, the ARMA$[3,2]$ spectral density does not exactly reproduce the power-law like behaviour of the FAR$[1]$ process. However, it mimics it to a certain extend. Divergence of the spectral densities can be observed for $\omega \lesssim \omega_1$: in this range the difference between the SRD and the LRD model manifests.

Likelihood-Ratio Test

Standard likelihood-ratio tests (Section 2.2.1) of the FARIMA$[3,d,2]$ against the two simpler models yield p-values of $p = 1$ and $p = 0.793$ for the ARMA$[3,2]$ and the FAR$[1]$, respectively. We can thus not reject neither model as being an admissible simplification of the FARIMA$[3,d,2]$ on any reasonable level of significance. The power of testing FARIMA$[3,d,2]$ against FAR$[1]$ is not large enough to reject the false simplification. We are thus in a situation where we cannot decide about the SRD or LRD on the basis of a likelihood-ratio test. Furthermore, the fractional difference parameter estimate for the FARIMA$[3,d,2]$ model is also not compatible with zero, giving no indication for SRD. We consequently need a more specific test for a reliable decision and we consider the simulation-based approach introduced in Section 2.3.

Simulation-Based Model Selection

Analogously to the procedure outlined in Section 2.3.2, we now perform the simulation-based model selection with the likelihood ratio as test statistic. The two models we consider as candidates for the underlying process are the FAR[1] and the ARMA[3, 2]. To facilitate the presentation, we denote in the following the ARMA[3, 2] and FAR[1] process as model f and g, respectively.

Figure 3.6 shows the cumulative distribution functions of the two log-likelihood-ratios. We find the value obtained for the "observed" record $\mathrm{lr}_{\mathrm{obs}}$ well outside the critical

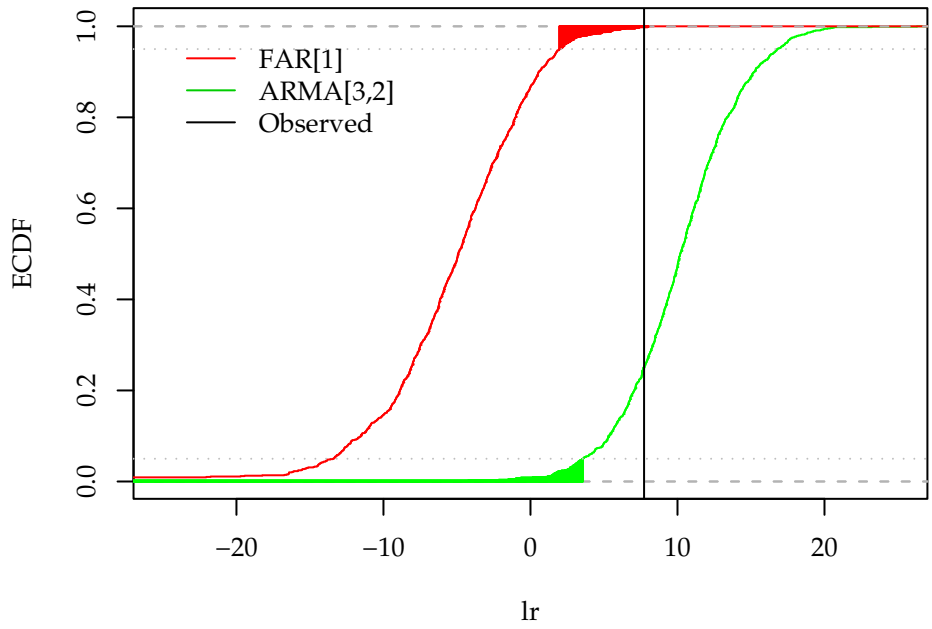

Figure 3.6: Empirical cumulative distribution function of log-likelihood-ratios $\mathrm{lr}_{f,r}$ (green) and $\mathrm{lr}_{g,r}$ (red) obtained from 1 100 simulations. The critical regions upto α-levels of 0.05 are filled with the respective color. The log-likelihood-ratio $\mathrm{lr}_{\mathrm{obs}}$ for the observed series is shown as a vertical line.

region of model f and inside the critical region of model g. The estimate of the power of testing H_f against H_g for the specified critical level $\alpha = 0.05$ yields $\widehat{\mathrm{pow}}_f(0.05, g) = 0.979$. For the inverse situation of testing H_g against H_f we find $\widehat{\mathrm{pow}}_g(0.05, f) = 0.977$. We can thus very confidently reject H_g (FAR[1]) in favour of H_f (ARMA[3, 2]) and correctly identify the underlying process.

3.4.2 Modelling the FAR[1] Realisation

Next, we estimate the model parameters obtained from the FAR[1] realisation. The resulting paramters and p-values for the goodness-of-fit test (Section 2.1.3) are listed in Table 3.2. Again, no model can be rejected on the basis of the goodness-of-fit test and the original FAR[1] parameters are recovered within one standard deviation. Also for the FARMA[3, d, 2] model, the estimates are compatible with the parameters of the original FAR[1] process.

Figure 3.5 (right) compares the ARMA[3, 2] and the FAR[1] spectral densities to the periodogram of the FAR[1] realisation. The picture is very similar to the left panel in the

Parameter	ARMA$[3,2]$	FAR$[1]$	FARIMA$[3,d,2]$
d	-	$0.287(0.015)$	$0.296(0.019)$
a_1	$2.178(0.044)$	$0.710(0.014)$	$0.704(3.582)$
a_2	$-1.488(0.083)$	-	$0.001(1.298)$
a_3	$0.309(0.039)$	-	$-0.002(1.216)$
b_1	$-1.184(0.045)$	-	$-0.007(3.588)$
b_2	$0.218(0.045)$	-	$0.002(1.707)$
p-val	0.333	0.356	0.359

Table 3.2: Parameters and asymptotic standard deviation in parentheses for the ARMA$[3,2]$, FAR$[1]$ and the FARIMA$[3,d,2]$ model estimated from the FAR$[1]$ realisation.

same figure. The two spectral densities are quasi identical in the high frequency region but differ for low frequencies. Especially for $\omega < \omega_1$ the difference between the LRD and SRD asymptotic behaviour is clearly visible.

Likelihood-Ratio Test

A standard likelihood-ratio test of the ARMA$[3,2]$ model being an admissible simplification of FARIMA$[3,d,2]$ is rejected on a 5%-level with a p-value of $p = 0.006$. The same test for FAR$[1]$ as simplified model cannot be rejected ($p = 0.889$). In this case, with the underlying process being LRD, we can discriminate ARMA$[3,2]$ and FAR$[1]$ already on the basis of standard nested-model selection. The simulation-based selection criterion for non-nested models is not needed. Thus, also for the second "observed" record, a correct identification of the underlying process could be achieved on the basis of a full-parametric model.

3.5 Summary

We illustrated three different strategies for the detection of LRD along the lines of a challenging example: one realisation of a FAR$[1]$ and one from an ARMA$[3,2]$ process with parameters chosen such that realisations with a specific length have very similar spectral characteristics.

With the heuristic DFA, the two realisations were not distinguishable. Furthermore, with the log-log plot as the central argument, we falsely "inferred" a LRD processes underlying both records. The log-periodogram regression revealed differences in the two series. For the FAR$[1]$ realisation, the asymptotic 95% confidence intervals did not enclose $d = 0$ for all the bandwidths used. Some did enclose this value for the ARMA$[3,2]$ realisation. Here, a suitable strategy for selecting the proper bandwidth is required.

The third approach to detect LRD is based on the full-parametric modelling and model selection strategies presented previously. We fitted FAR$[1]$, ARMA$[3,2]$, and the smallest common model, FARIMA$[3,d,2]$, to both realisations. For the ARMA$[3,2]$ realisation, we could not reject neither the FAR$[1]$ nor the ARMA$[3,2]$ as admissible simplifications of the FARIMA$[3,d,2]$ on the basis of a standard likelihood-ratio test. Here, we had to use the simulation-based approach to directly compare the two alternatives FAR$[1]$ and ARMA$[3,2]$. The likelihood-ratio test was, however, sufficient in the case of the FAR$[1]$ realisation. For both "observed" records the proper model and, in particular, the correct dependence structure could be identified.

This example illustrated the reliability of the full-parametric modelling approach with respect to the detection of LRD. Furthermore, it turned out that the standard model selection, in some cases, such as the example at hand, does not have enough power to discriminate the two alternatives under consideration. In such cases, a suitable strategy for the discrimination of non-nested models, such as the simulation-based model selection, is required

The approach presented here, i.e. rephrasing the problem of detecting LRD as a model selection problem, is different from other approaches mainly because of the direct comparison of the most suitable SRD and the most suitable LRD process. A pivotal element in this model comparison is the model selection strategy based on the log-likelihood-ratio.

Appendices

Appendix A

Review of Detrended Fluctuation Analysis

Detrended fluctuation analysis (DFA, Section 1.3.3) has been widely used to infer a LRD process from empirical time series and also to quantify the strength of the dependence, i.e. estimating the Hurst exponent H, or, equivalently, the fractional difference parameter $d = H - 0.5$ (e.g. **????**). Little work has been done on characterising the inferential power of DFA and the properties of the estimator for the Hurst exponent, e.g., quantifying bias and variance. In this chapter, we present a simulation study which provides estimates of bias and variance for power-law noise processes. Furthermore, we critically review the limits of the method to infer LRD, suggest refinements and point to typical pitfalls. Finally, we exemplify the discussion with an analysis of the Prague daily temperature anomalies.

A.1 Bias and Variance for Self-Similar Increment Processes

Analogously to the work of **?** and **?**, we use Monte Carlo (MC) simulations to study the bias and variance of the estimator $\widehat{H}_{\mathrm{DFA}n}$ for DFA1 and DFA2 ($n = 1, 2$) introduced in Section 1.3.3. In addition to a white noise process, we consider a process with a spectral density $S(\omega) = |\omega|^{-\beta}$, also referred to as power-law noise. The spectral density of this process is similar to the one of fGn or FD processes; especially the asymptotic behaviour for low frequencies ω is the same. Realisations of power-law noise with a prescribed exponent $\beta = 2H - 1$ can be obtained using the inverse Fourier transform (Section 1.4; **?**).

Bias

We estimate the bias and variance from MC ensembles of realisations from power-law noise with different length N and prescribed Hurst exponents $0 < H < 1$. The estimate $\widehat{H}_{\mathrm{DFA}n}$ is obtained as the slope of a straight line fit to $\log F(s)$ versus $\log s$ for $s > 10$. We calculate the bias using

$$\mathrm{bias}(\widehat{H}_{\mathrm{DFA}n}) = \overline{\widehat{H}_{\mathrm{DFA}n} - H}, \tag{A.1}$$

with the bar denoting the ensemble mean. For $\widehat{H}_{\mathrm{DFA1}}$ and $\widehat{H}_{\mathrm{DFA2}}$ and an ensemble size of 2 500 runs the bias is shown in Figure A.1. The estimator based on DFA1 exhibits a positive bias for power-law noise with $0 < H < 0.5$, increasing with decreasing H. The bias reduces with an increasing record length. This is in line with an analytical result derived by **?**. In contrast, we find a negative bias for $0.5 < H < 1$, increasing in magnitude with

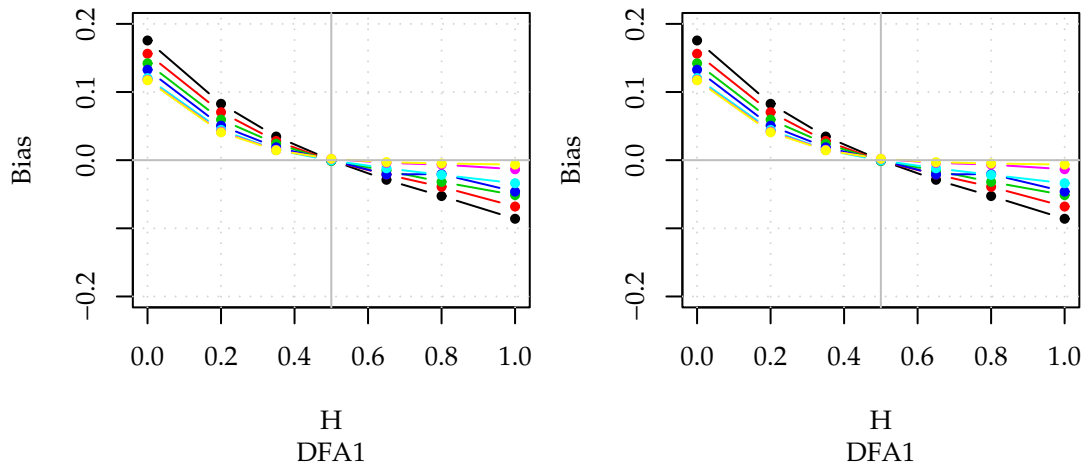

Figure A.1: Bias of $\widehat{H}_{\mathrm{DFA1}}$ (left) and $\widehat{H}_{\mathrm{DFA2}}$ (right). The bias is estimated from a MC ensemble of power-law noise realisations with a prescribed Hurst exponent H and various sample lengths N.

H. An increasing record length reduces the bias faster than for $0 < H < 0.5$. For the white noise process ($H = 0.5$) no bias is observed for all length of the records. This "neutral point" moves towards $H \approx 0.6$ for the DFA2 based estimator. This is accompanied with an overall increase of the bias.

Variance

The variance of $\widehat{H}_{\mathrm{DFA}n}$ estimated from the same ensemble is shown in Figure A.2. A striking difference to the previous plot is the asymmetry. The variance $\mathrm{var}(\widehat{H}_{\mathrm{DFA1}})$ increases monotonically with $0 < H < 1$. This implies that for a given length N the Hurst exponent can be estimated more precisely for $0 < H < 0.5$ then for a white noise process ($H = 0.5$). Furthermore, Figure A.2 depicts a decreases in variance with the length N of the series.

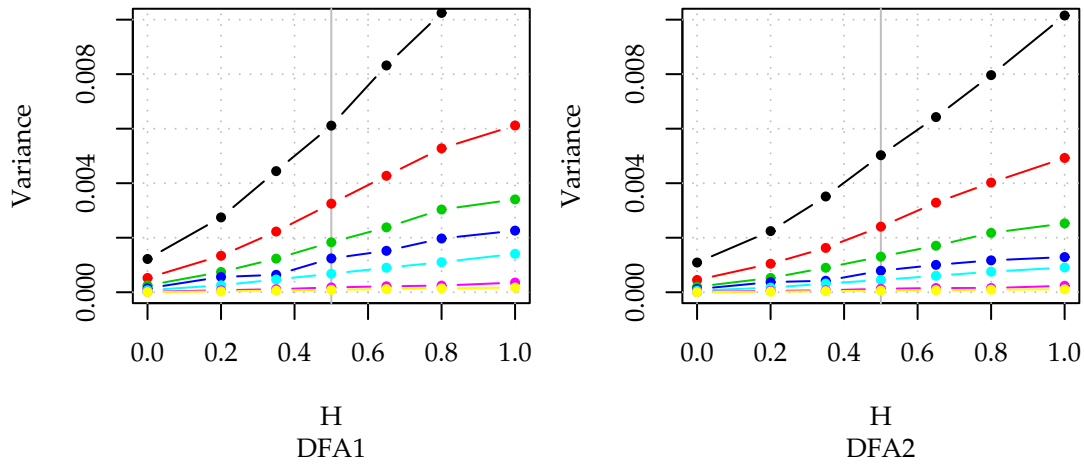

Figure A.2: Variance of $\widehat{H}_{\mathrm{DFA1}}$ (left) and $\widehat{H}_{\mathrm{DFA2}}$ (right). The variance is estimated from a MC ensemble of power-law noise realisations with a prescribed Hurst exponent H. The colour coding of the record length is analogue to Fig. A.1.

Figures A.1 and A.2 allow for a quick and rough estimate of bias and variance for

increments of self-similar processes with $0 < H < 1$. For a more precise DFA-bases Hurst-exponent estimation from empirical data, one would have to adjust this simulation studies to the situation at hand, i.e. adjust the record length and the fit interval for the straight line and find a suitable parametric model to represent the data. This leads to a parametric bootstrap approach suitable to estimate confidence intervals. A parametric ansatz using one of the models described in Section 1.2.2 is suitable to accomplish this task. A different approach was suggested by **?**, they split the series under consideration in m records and performed the analysis m times to obtain an estimate for the variability. This, however, is only feasible if very long records, e.g., from simulation runs, are available.

A far more serious shortcoming of standard DFA analyses, as it is commonly used, is the *a priori* assumption of scaling. Straight lines are frequently fit to $\log F(s)$ versus $\log s$ without evaluating any criteria supporting this simple model (cf. Appendix A.2 and **?**).

A.2 DFA and the Detection of Long-Range Dependence

Detrended fluctuation analysis including a subsequent straight line fit to the logarithmic fluctuation function $\log F(s)$ with $\log s$ as regressor yields an asymptotically unbiased estimator for the Hurst exponent H and thus for the fractional difference parameter d. This has been shown by **?** for processes with the same asymptotic behaviour as fractional Gaussian noise or FARIMA processes. The limiting distribution of this estimator has not been studied so far. This makes statistical inference for the Hurst exponent impossible (cf. Section 1.3.3).

We first investigate the non-asymptotic behaviour of this estimator for power-law noise. Subsequently, we discuss the need to infer that the asymptotic behaviour has been reached for the series under consideration before Hurst exponent estimation is meaningful.

A.2.1 Inference of Scaling

Prior to the estimation of a scaling exponent, the existence of a scaling region has to be inferred from an empirical record (e.g., temperature anomalies, run-off, financial time series, etc.). Besides in **?**, it has not been studied, if DFA can be used to infer such a scaling region. It is thus not clear if LRD can be inferred from realisations of a process when it is not *a priori* clear, that this process exhibits scaling. It is therefore still an open question how *sensitive* and *specific*[1] DFA behaves when investigating processes of unknown correlation structure.

A necessary condition for the existence of LRD is the scaling of the fluctuation function $F(s)$ in the asymptotic region according to (1.21). Thus, in the remainder of this section, we address the following questions:

1. How to conclude scaling from the DFA fluctuation function?
2. Does a region of scaling necessarily imply LRD?

[1] A procedure, that detects compatibility with LRD with a high probability, whensoever present, is called sensitive. An algorithm that with a high probability rejects LRD, when not present, is said to be specific. The optimal algorithm would be sensitive and specific. A sensitive but unspecific algorithm, however, would produce many false positive results, i.e. one would frequently detect LRD. This algorithm would not be suitable for a reliable inference. On the other hand, an un-sensitive but specific algorithm would be very conservative and would often reject the existence of LRD.

Two Example Processes

To illustrate the line of argument, we consider a SRD as well as a LRD process and apply DFA to both. As an example for a LRD process we simulate power-law noise according to the method given in **?** with a Hurst exponent of $H = 0.6$. This process shows power-law scaling in the ACF for a wide range of scales. For the SRD process we choose a superposition of three AR[1]-processes (cf. Section 1.2.1),

$$x(i) = \sum_{j=1}^{3} A_j y_j(i)\,, \quad y_j(i) = a_j y_j(i-1) + \eta_j(i)\,, \tag{A.2}$$

with $\eta_j(i)$ being Gaussian white noise of zero mean and variance $1 - a_j^2$. The latter ensures $\mathrm{var}(y_j) = 1$. We choose $A_1 = 0.913$, $A_2 = 0.396$, $A_3 = 0.098$, $a_1 = 0.717$, $a_2 = 0.953$ and $a_3 = 0.998$. Using $a_j = e^{-1/\tau_j}$ we find the following characteristic time scales for the individual AR[1] processes: $\tau_1 = 3$ days, $\tau_2 = 21$ days and $\tau_3 \approx 1.5$ years. The choice of the model parameters is motivated by the example of the Prague temperature record studied in Section A.3.1 and will become clear during the discussion.

Establish Scaling

Figure A.3 shows the fluctuation functions for a realisation of each of the two example processes with length $N = 70\,492$ corresponding to the Prague temperature record (cf. Section A.3.1). For every order of magnitude, 50 values of equal distance in logarithmic scale are calculated.

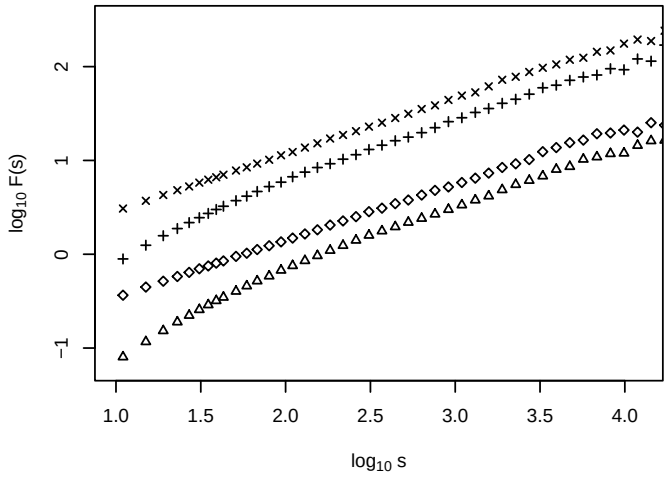

Figure A.3: Fluctuation functions calculated for an artificial LRD process with exponent $H = 0.6$ ($\times$ DFA1, $\diamond$ DFA2) and a superposition of three AR-processes ($+$ DFA1, $\triangle$ DFA2) as defined in Section A.2.1. For every order of magnitude, approx. 50 values are calculated. For clarity, only every third value is plotted.

To reliably infer power-law scaling of the fluctuation function, a straight line in the log-log plot has to be established. Since a straight line is tantamount to a constant slope, local estimates of the slope $\widehat{H}_{\mathrm{DFA}n}$ of $\log F(s)$ versus $\log s$ have to be evaluated for constancy in a sufficient range (e.g., **???**). The extend of a sufficient range is still a matter of debate (e.g., **?** and references therein). This concept has been introduced in the context of

estimating correlation dimensions (**??**) and, in a different setting, has also been suggested for DFA (**?**).

Calculating Local Slopes

For a finite amount of data the estimation of the local slopes brings along a certain variability. Even for a process like the power-law noise, the local slopes of the empirical fluctuation function show variations around a constant H. This has two consequences for the calculating and interpreting the local slopes:

1. estimating the local slopes by finite differences results in a large variability and
2. this variability has to be quantified to allow for statistical inference.

Regarding the first point, the variability can be reduced fitting a straight line to $\log F(s)$ versus $\log s$ within a small window. The window is then shifted successively over all calculated scales s. Figure A.4 shows the local slopes of a realisation of the SRD model for different window sizes. Choosing the optimal window size, one has to trade bias for variance: for small windows, the bias is small, but the variability renders the interpretation difficult, whereas for large windows, the variance is reduced at the cost of a biased estimate of H. Thus, the extreme case of a single straight line fit to the whole range of scales considered is maximally biased. Since only one value of for the estimate $\widehat{H}_{\mathrm{DFA}n}$ is calculated, this does not allow to evaluate constancy.

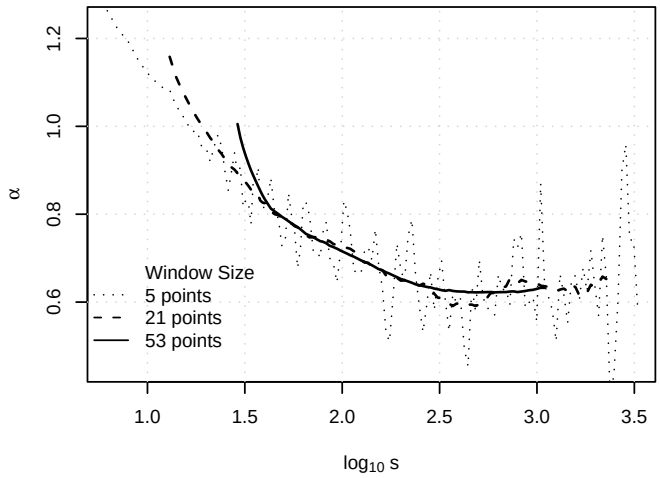

Figure A.4: Local slopes α (or H) of a realisation of the short-memory model for different window sizes. 50 points per order of magnitude are calculated, but for clarity reasons only every second value is plotted for the smallest window size. For small windows, the bias is very low, but the variability renders the interpretation difficult, whereas for large windows, the variance is reduced at the cost of a biased estimator $\widehat{H}_{\mathrm{DFA1}}$.

The second problem of quantifying the variability for a given length of the record is not straightforward. Since vicinal local slopes are not independent, confidence regions cannot be estimated straightforwardly from the procedure described in Section A.2.1 (**?**). Instead, we perform Monte Carlo simulations: for the two example processes, we simulate 1 000 realisations to estimate mean and standard deviation for the estimator $\widehat{H}_{\mathrm{DFA1}}$ for the scales considered. The distribution of the estimates is approximately Gaussian.

Thus, we employ the interval $[\overline{\widehat{H}_{\mathrm{DFA1}}} - 1.96\widehat{\sigma}, \overline{\widehat{H}_{\mathrm{DFA1}}} + 1.96\widehat{\sigma}]$ as estimates of the 95% confidence region, with the bar denoting again the ensemble mean. $\widehat{\sigma}^2$ is the variance of $\widehat{H}_{\mathrm{DFA1}}$ estimated from the ensemble.

Inferring Scaling for the Example Records

Figure A.5 displays the local slopes of the DFA1 (a) and DFA2 (b) fluctuation functions, estimated from one realisation of each of the example models. The realisation of the LRD

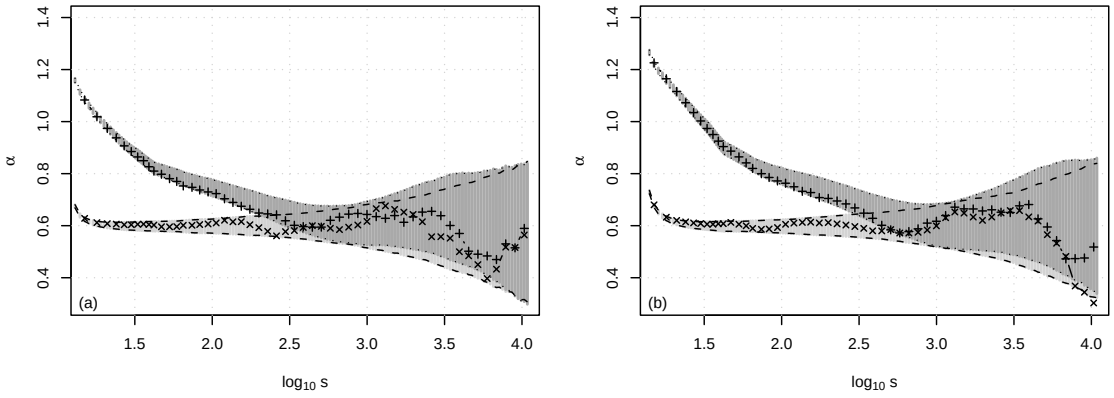

Figure A.5: Local slopes α (or H) of the DFA1 (a) and DFA2 (b) fluctuation function calculated for an artificial LRD process with exponent $H = 0.6$ ($\times$) and a superposition of three AR-processes ($+$) as defined in Section A.2.1 (window size 21 points). The dashed and the dotted lines border the shadowed $1.96\widehat{\sigma}$ intervals obtained from 1 000 realisations of the two processes, respectively.

process shows fluctuations around a constant H within the corresponding $1.96\widehat{\sigma}$ interval, increasing like $\widehat{\sigma} \propto \sqrt{s}$ (**?**). The local slope $\widehat{H}_{\mathrm{DFA1}}(s)$ of the SRD realisation, however, decreases constantly in the beginning and basically follows the local slope of the LRD realisation for scales larger than $\log s \approx 2.5$. Thus, for a certain choice of parameters, a SRD model can mimic scaling in a finite range. Due to the principle of variance superposition for DFA (**?**) a suitable superposition of three AR[1] processes produces this effect in the fluctuation function analogously to the same effect in the spectral domain described in **?**.

Investigating for LRD one studies primarily the behaviour on large scales s or small frequencies assuming that influences from low frequencies are negligible here and do not bias the estimation of the LRD parameter. In our example, the $1.96\widehat{\sigma}$-cones from the LRD and SRD process are virtually indistinguishable in this range. Thus, based on the given record length and only considering large s, one cannot distinguish the realisations of the two models by means of DFA. For longer time series, the cones would shrink and the region of overlapping would become smaller.

Inference of SRD

However, a general dilemma related to the inference of LRD emerges: For a finite time series, one will always find a SRD model to describe the data (**?**). Thus, considering the inference of LRD, DFA is sensitive, but not specific. An alternative ansatz is to investigate if the underlying process is SRD. This requires

1. to show compatibility with a SRD model and
2. to exclude possible LRD models.

The first condition is always fulfilled, since one will always find a SRD model to describe a finite data set. The second condition is not fulfilled for the given example, because the record length is not sufficient to detect the SRD character $H = 0.5$ for large s of the AR-model by means of DFA. For longer time series as shown in Figure A.6, when a plateau of $H = 0.5$ is identifiable, LRD can be excluded and the specificity of DFA to infer SRD increases.

A.2.2 Pitfalls

We discuss typical difficulties for DFA and similar heuristic approaches. Problems arise for example when investigating the results only in a double logarithmic plot, or, if the observed behaviour is not consistent with the asymptotic behaviour. Also the transfer from a scaling exponent of the fluctuation function to the ACF is only possible in the asymptotic limit.

The Double Logarithmic Plot

Investigating only the double logarithmic plot of the fluctuation function, one is tempted to rashly conclude for LRD. Due to properties of the logarithm, fluctuations are suppressed in a double logarithmic plot and the deviation from a straight line is not easily visible (**?**). Also, restricting the analysis to a straight line in the log-log plot forces $F(s)$ in the *procrustean bed* of power-laws. It will always yield some value for the slope but the suitability of the linear description is not evaluated. For the inference of LRD, this procedure would be sensitive but not specific. This results in attributing LRD to all processes with an estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ for the largest scales observed. Such a result would trivialise the concept of LRD and provide no insight into the process. Thus, to reliably infer a power-law, a straight line may not be assumed *a priori* but has to be established, as discussed in Section A.2.1. Even if scaling is present, it is difficult to determine the beginning and ending of the scaling region in the log-log plot. However, the estimate $\widehat{H}_{\mathrm{DFA}n}$ derived from a straight line fit strongly depends on the fit boundaries if the realisation does not stem from a scale free process.

Finite scaling of short-memory processes

According to Section 1.1.4, the autocorrelations of SRD processes decay exponentially for large s and are negligible on scales large compared to the decorrelation time

$$\begin{aligned} \tau_D &= 1 + 2\sum_{s=1}^{\infty} \rho(s) = 1 + 2\sum_{s=1}^{\infty} e^{-s/\tau} \\ &\approx 2\tau \text{ for } \tau \gg 1. \end{aligned} \tag{A.3}$$

Consequently, for scales large enough, the slope of the fluctuation function of such a process converges to $H = 0.5$. However, for a finite set of data one cannot be *a priori* sure that the series is long enough to observe this. For a record of the SRD model defined in Section A.2.1 (length 70 492 points) the local slopes of the fluctuation function for the largest observed scales are compatible with power-law scaling. A plateau about $H = 0.5$ is not observed (Figure A.5). Thus, one might be tempted to conclude an underlying LRD process. However, analysing a much longer record (1 000 000 points) of the same model yields such a plateau about $H = 0.5$ for large s (Figure A.6). Therefore, for a process with unknown correlation structure it is misleading to use solely the estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ as

argument for LRD. It might very well be that the record is too short to observe the plateau about $H = 0.5$ characteristic for SRD. This stresses the need for confidence intervals for

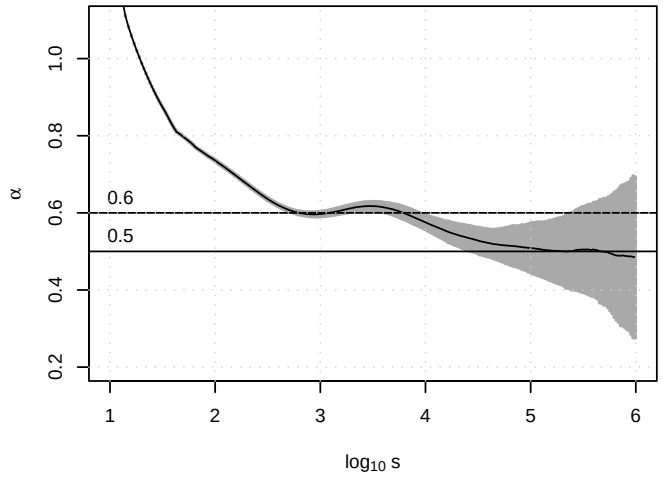

Figure A.6: Local slopes α (or H) from the empirical fluctuation function of the short-memory model (A.2), estimated from 200 realisations of length N=1 000 000 (solid line). A region of approximatively constant slope occurs between $\log s \approx 2.8$ ($s \approx 600$) and $\log \approx 3.8$ ($s \approx 6\,000$, corresponds to approximately 16 years). On larger scales, the slope reduces to $\widehat{H}_{\mathrm{DFA1}} = 0.5$, revealing the SRD nature of the model.

the estimate and for means to reliably discriminate between SRD and LRD processes.

DFA to ACF transfer of a finite "scaling" region

As shown in Sect. A.2.1, under certain conditions also SRD processes can mimic a finite "scaling" region. Thus, the question arises, if such a scaling region derived from the fluctuation function corresponds to a power-law like behaviour in the same region in the ACF. Such an assumption was used by, e.g., **?**. To challenge this hypothesis, we compare the properties of the fluctuation function (Figure A.6) with the analytical ACF (Figure A.7). The dashed lines depict the ACF of the single AR[1] processes with the largest and the smallest time scale. The ACF of the superposition of the three AR[1] processes is given by the dashed-dotted line. The solid line represents a power-law with exponent $\gamma = 0.8$. This value can be expected when applying $H = 1 - \gamma/2$ (1.69) to the estimate $\widehat{H}_{\mathrm{DFA1}} \sim 0.6$ derived from the fluctuation function. The range of scales where the power-law-like behaviour holds for the fluctuation function is $\log s \gtrsim 2.5$ (Figure A.5). We find that the region of almost constant slope of the ACF is located on smaller scales between $s \approx 1$ and maximally $s \approx 1\,000$ ($\simeq 3$ years). Thus, based on a finite scaling region found in the fluctuation function of a SRD process, it is in general not valid to conclude that an equal scaling region exists also for the ACF.

A.3 Investigating the Prague Temperature Anomalies

In the following, we analyse the temperature anomalies from Prague.To challenge both strategies, we furthermore impose the task of discriminating the Prague record from an artificial SRD time series. The latter is a realisation of a superposition of three AR[1] processes. The superposition was constructed such that its realisations mimic the Prague

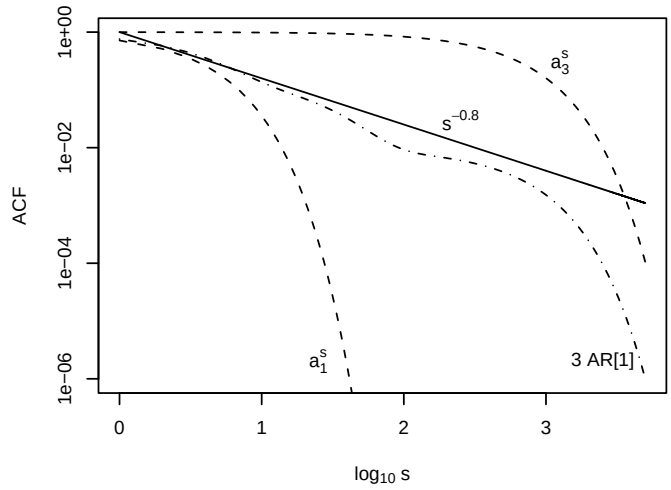

Figure A.7: Analytical ACF of the SRD model (A.2) (dashed dotted line). The dashed lines depict the ACF of the single AR[1] processes with shortest and longest time constant. In a range of scales from $\log s \approx 0$ ($s \approx 1$) to maximally $\log s \approx 3$ ($s \approx 1\,000$) the ACF approximately follows the power law with $\gamma = 0.8$ (solid line). For larger scales, it turns into an exponential decay determined by the AR[1] process with the largest time constant $\tau \approx 1.5$years.

temperature residuals as closely as possible. The structure of this process is given in (A.2), the corresponding parameters in Section A.2.1. By definition this process is SRD.

We start with a DFA for the Prague normalised temperature anomalies and for a realisation of the AR-superposition. This is followed by the full-parametric modelling approach applied only to the AR-superposition.

A.3.1 Detecting Long-Range Dependence using DFA

DFA is reported to eliminate the influences of a possible trend in the mean on the analysis (e.g., **?**), we thus use the normalised temperature anomalies and study DFA1 and DFA2. An investigation of higher orders of DFA does not significantly affect the discussion presented while for DFA1 the effect of a trend might be suspected. The fluctuation function $F(s)$ (1.68) is calculated for approximately 50 points per order of magnitude upto $s_{\max} = N/4$ and is shown in Figure A.8 in double logarithmic representation. The behaviour is qualitatively different from white noise. To depict the variability of this estimate of the fluctuation function, we performed simulations with the AR model (A.2) and estimate its fluctuation function from 1 000 realisations with the same length as the observed record. We approximate a 95% confidence region by plotting $1.96\widehat{\sigma}$ intervals as grey shadows. As expected, the variability of the fluctuation function estimate increases wit the time scale s. Because the observed fluctuation functions for DFA1 and DFA2 are both well inside the 95% confidence interval, we consider the Prague temperature anomalies compatible with the AR model from the viewpoint of DFA.

Following the discussion in Section A.2.1, we estimate the local slopes to investigate for power-law scaling. From the fluctuation function of the Prague record, we estimate the local slopes using a straight line fit in a small window of 21 points. Figure A.9 shows the result and compares it to 1 000 runs from the AR model and from power-law noise with exponent $H = 0.6$. Accounting for the variability, we find the observed slopes consistent with a constant slope as expected from the power-law noise process (light grey) for $\log s \gtrsim 2.5$. In the same range of scales, however, the observed slopes also agree with

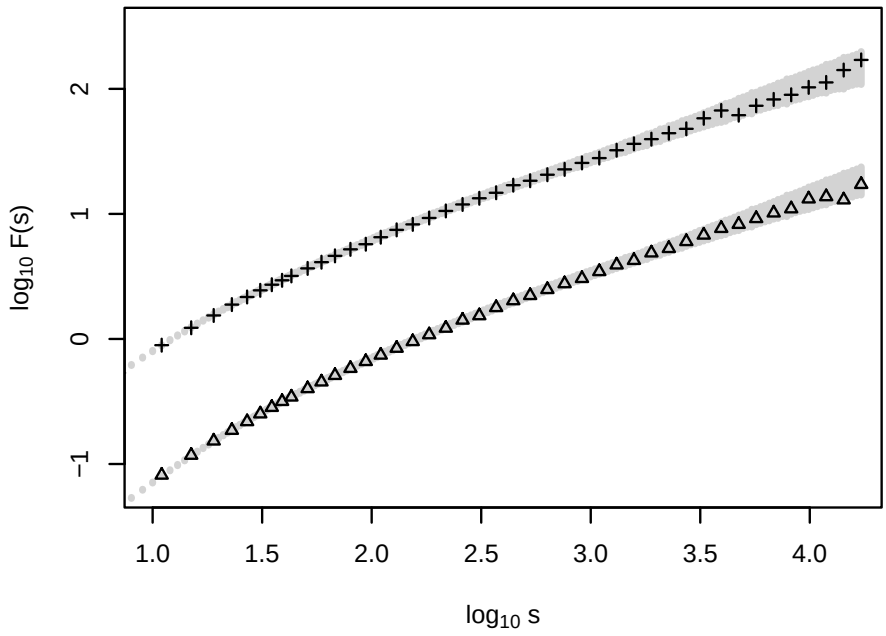

Figure A.8: DFA1 (+) and DFA2 (△) fluctuation function of Prague daily temperature data calculated for approximately 50 points per order of magnitude. Only every 4th point is shown to enhanced clarity. The shadows mark $1.96\widehat{\sigma}$ confidence regions derived from 1 000 runs of the AR model.

the AR model. Since we are looking for a scaling region, we consider at the most $\log s \gtrsim 2.5$. Thus, from the given data, one cannot decide whether the Prague temperature time series is a realization of a SRD (dark grey) or a LRD process (light grey). This plot shows furthermore, that considering the scales $\log s \lesssim 2.5$, the power-law noise is not a sufficient model for the observed series.

A.3.2 Detecting Long-Range Dependence using Parametric Models

From a previous analysis (not shown), we find the FARIMA$[2, d, 2]$ being the most suitable model for the Prague maximum temperature residuals with parameters given in Table A.1. The fractional difference parameter was estimated to be $\widehat{d} = 0.109(0.017)$ for this model. Employing the Gaussian limiting distribution, we find a 95% confidence interval for the difference parameter of $[0.076, 0.142]$ and thus $d = 0$ for SRD processes is not included. We thus infer the underlying process to be LRD.

It remains to investigate whether we can correctly identify the realisation of the AR-superposition as coming from an SRD process. This is pursued in the following.

Stochastic Modelling of the AR$[1]$ Superposition

The AR model (A.2) is constructed such that realisations generated from it mimic the characteristics of the Prague temperature anomalies. We thus do not have to account for deterministic components, such as the seasonal cycle or a trend. Starting with the set of FARIMA$[p, d, q]$ and ARMA$[p, q]$ models compiled in a previous analysis and reduce these with the model selection strategies: the goodness-of-fit test, the HIC-based model selection and, finally, the likelihood-ratio-based model selection.

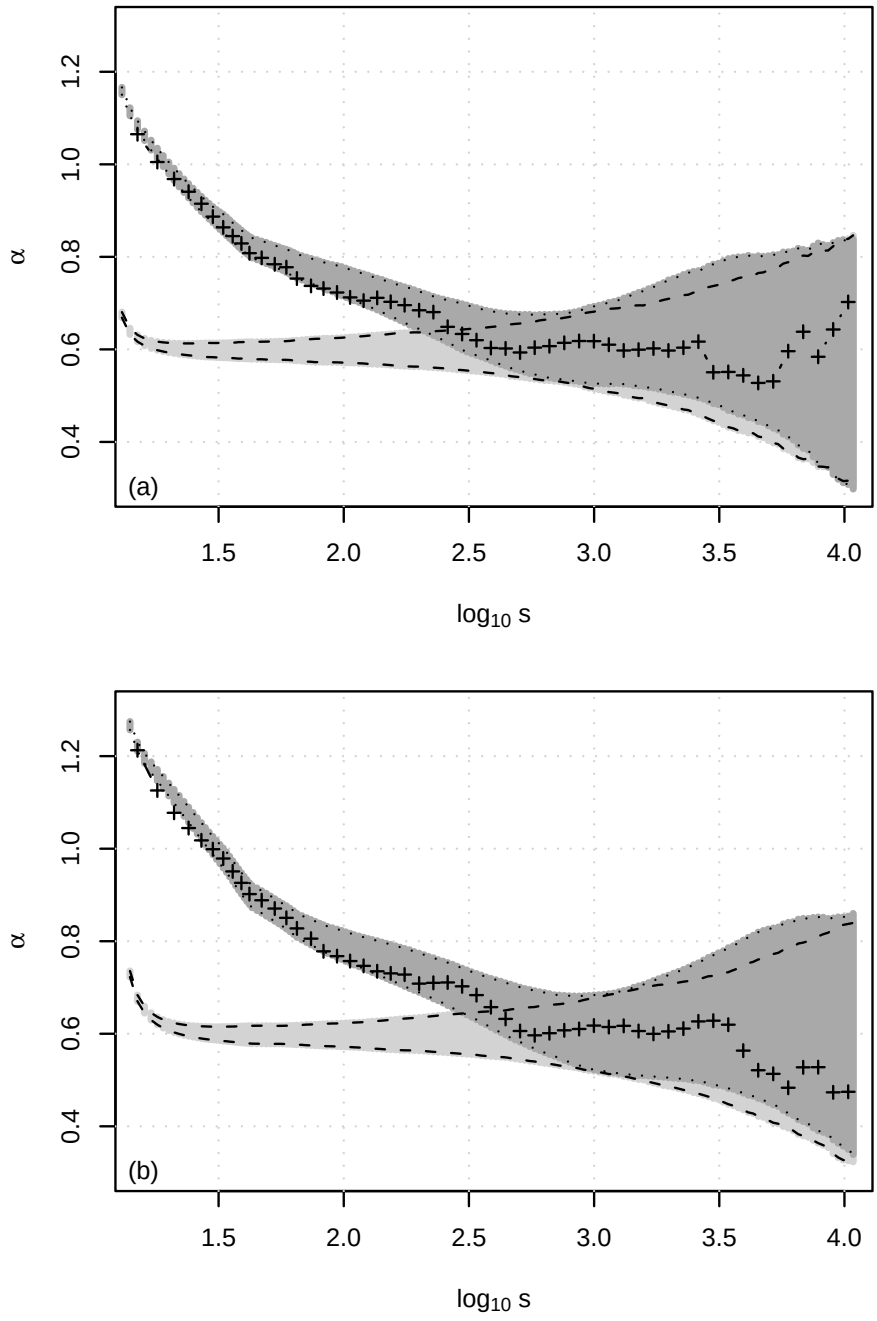

Figure A.9: Local slopes α (or H) of the fluctuation functions plotted in Figure A.8 for DFA1 (a) and DFA2 (b) of the Prague daily temperature data. The dotted lines border the $1.96\widehat{\sigma}$ confidence regions of the short-range correlated model (A.2) (dark shadow), the dashed lines those of the long-memory model with $H = 0.6$ (light shadow).

Parameter	FARIMA$[2,d,2]$	ARMA$[5,3]$	ARMA$[8,1]$
d	$0.109(0.017)$	–	–
a_1	$1.210(0.066)$	$1.548(0.275)$	$1.762(0.012)$
a_2	$-0.332(0.042)$	$-0.139(0.544)$	$-0.879(0.012)$
a_3	–	$-0.679(0.368)$	$0.145(0.008)$
a_4	–	$0.314(0.116)$	$-0.035(0.008)$
a_5	–	$-0.050(0.019)$	$0.005(0.008)$
a_6	–	–	$-0.003(0.008)$
a_7	–	–	$0.002(0.008)$
a_8	–	–	$-0.005(0.004)$
b_1	$-0.513(0.057)$	$-0.741(0.275)$	$-0.955(0.011)$
b_2	$-0.106(0.007)$	$-0.567(0.326)$	–
b_3	–	$0.342(0.095)$	–
p-val	0.074	0.071	0.075

Table A.1: Maximum likelihood parameter estimates and asymptotic standard deviation in parentheses for the FARIMA$[2,d,2]$, ARMA$[5,3]$ and ARMA$[8,1]$ process obtained from the Prague temperature residuals. The last line gives the p-values of the goodness-of-fit test.

Goodness-of-fit On the basis of the goodness-of-fit test (2.4) applied to the before mentioned set of models we reject only the FD (FARIMA$[0,d,0]$) model on a 5%-level of significance.

HIC Model Selection Figure A.10 depicts the HIC values for the models considered. It is striking that for the most cases the ARMA$[p,q]$ model has a smaller HIC than the corresponding FARIMA$[p,d,q]$ model. Investigating the fractional difference parameter for the latter models (Table A.2) reveals that besides FARIMA$[1,d,0]$, FARIMA$[1,d,1]$ and FARIMA$[2,d,0]$ all estimates are compatible with the $d = 0$ within one (or, in one case, 1.96) standard deviation. We retain these models, as well as ARMA$[p,q]$ models with orders $(p,q) \in \{(2,1),(2,2),(3,1),(3,2)\}$. Keeping other FARIMA$[p,d,q]$ models is not meaningful, because their fractional difference parameter is compatible with zero and thus they are equivalent to the corresponding ARMA$[p,q]$ processes.

Likelihood-Ratio Test With a likelihood-ratio test, we investigate whether the three models with a non-trivial fractional difference parameter are admissible simplifications of the FARIMA$[2,d,1]$, the simplest models with trivial difference parameter. The p-values in Table A.3 (right) reveal that we can reject this hypothesis for all three cases on a 5%-level (or even 1%-level) of significance. This implies, that there is no evidence for a LRD process underlying this data series.

Among the ARMA$[p,q]$ models, we find the ARMA$[2,2]$ and ARMA$[3,1]$ as admissible simplifications of the ARMA$[3,2]$ (Table A.3, left, lines 1 and 2). Both models can be further simplified by ARMA$[2,1]$ (lines 3 and 4). The latter is the most suitable model out of the canon we started with for the realisation of the AR-superposition (A.2).

Using the full-parametric modelling approach we are thus able to correctly identify the realisation of a superposition of three AR$[1]$ processes as coming from an SRD process. This was not possible by means of DFA. The original model, a superposition of AR$[1]$ processes (A.2), was not included in the model canon we used to describe the example series with. However, such a superposition can be well approximated with ARMA$[p,q]$ processes.

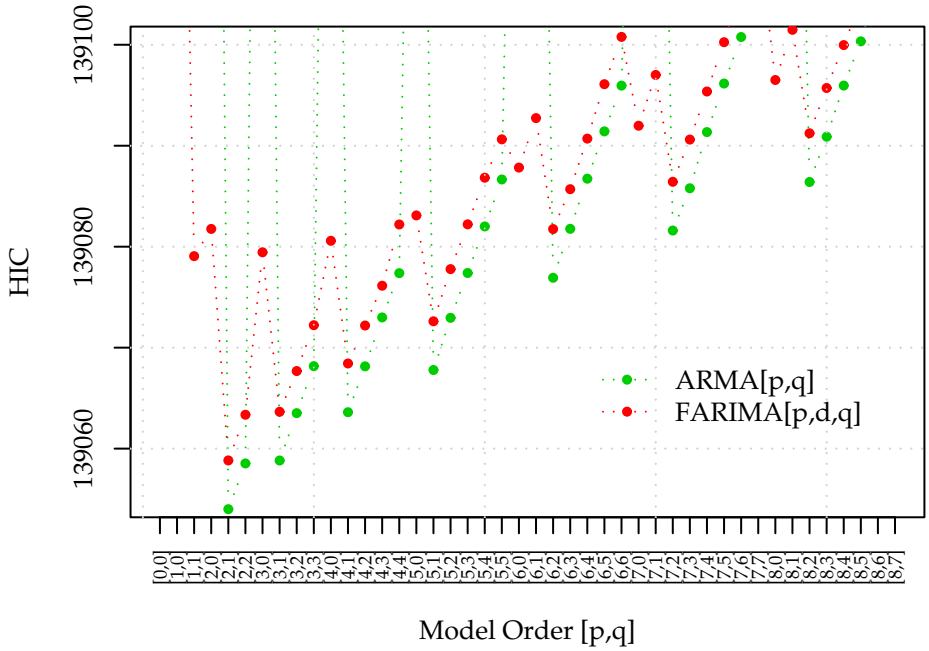

Figure A.10: HIC for various ARMA$[p,q]$ (green) and FARIMA$[p,d,q]$ (red) models fitted to realisation of the AR model (A.2) The model orders $[p,q]$ are plotted along the abscissa.

Model	$\widehat{d}(\sigma_{\widehat{d}})$
$[1,d,0]$	$0.148(0.010)$
$[1,d,1]$	$0.158(0.011)$
$[2,d,0]$	$0.160(0.011)$
$[2,d,1]$	$0.000(0.015)$
$[2,d,2]$	$0.000(0.037)$
$[3,d,1]$	$0.000(0.037)$
$[3,d,2]$	$0.012(0.017)$
$[4,d,1]$	$0.000(0.040)$
$[3,d,3]$	$0.028(0.038)$
$[4,d,2]$	$0.027(0.038)$
$[5,d,1]$	$0.000(0.042)$
$[4,d,3]$	$0.023(0.022)$
$[5,d,2]$	$0.000(0.041)$

Table A.2: Estimates and asymptotic standard deviation of the fractional difference parameter obtained for the FARIMA$[p,d,q]$ models with smallest HIC.

ARMA$[p,q]$

	Model f	Model g	p-val
1	$[3,2]$	$[2,2]$	1
2	$[3,2]$	$[3,1]$	0.713
3	$[2,2]$	$[2,1]$	0.584
4	$[3,1]$	$[2,1]$	0.939

FARIMA$[p,d,q]$

	Model f	Model g	p-val
1	$[2,d,1]$	$[2,d,0]$	<0.001
2	$[2,d,1]$	$[1,d,1]$	<0.001
3	$[2,d,1]$	$[1,d,0]$	<0.001

Table A.3: p-values for a likelihood-ratio test of model g being an admissible simplification of model f for ARMA$[p,q]$ and FARIMA$[p,d,q]$ models of the realisation stemming from the AR[1]-superposition (A.2).

A.4 Summary

In a simulation study, we investigated bias and variance for a DFA-based estimator for the Hurst exponent using realisations of power-law noise. These properties of the estimator are not only influenced by the time series length but also by the Hurst exponent of the underlying process. Without knowledge about bias and variance, inference about the Hurst exponent cannot be made.

We further considered the inference of LRD by means of DFA with respect to the notions of sensitivity and specificity. The inference of a LRD process underlying a given time series requires not only to show compatibility of the data with a LRD process in a certain range of scales. Furthermore, other possible correlation structures, especially SRD, have to be excluded.

Power-law like behaviour in some range of scales of the DFA fluctuation function alone is frequently taken as evidence for LRD. To reliably infer power-law scaling, it must not be assumed but has to be established. This can be done by estimating local slopes and investigating them for constancy in a sufficient range. However, finite data sets bring along natural variability. To decide, if a fluctuating estimation of the slope has to be considered as being constant, we calculated empirical confidence intervals for a LRD and a simple SRD model.

Discussing typical difficulties of interpreting DFA results, we note that scaling cannot be concluded from a straight line fit to the fluctuation function in a log-log representation. Additionally, we show that a local slope estimate $\widehat{H}_{\mathrm{DFA}n} > 0.5$ for large scales does not necessarily imply long-memory. If the length of the time series is not sufficiently large compared to the time scales involved, also for SRD processes $\widehat{H}_{\mathrm{DFA}n} = 0.5$ may not be reached. Finally, we demonstrated, that it is not valid to conclude from a finite scaling region of the fluctuation function to an equivalent scaling region of the ACF.

With the Prague temperature anomalies and a corresponding artificial series from a SRD process, we exemplify the problems and pitfalls discussed. Using DFA we could not discriminate the Prague record from the artificial series. Furthermore, by means of a log-log plot, we classify the underlying process of both series as LRD, leading to a false positive result for the SRD process. A reliable identification was possible with the full-parametric ansatz (Section 3.4).

Because it is always possible to find and SRD process to describe a finite set of data, some criterion is needed to evaluate the performance of the description taking the complexity of the model into account. We thus approach the problem of distinguishing SRD and LRD processes on the basis of a realisation of a parametric modelling point of view developed in Chapter 3.

Appendix B

Long-Range Dependence – Effects, Methods, Mechanisms

A simple example to illustrate the effect of long-range dependence on the estimation of statistical quantities is illustrated in the first section. This is followed by some physical explanations of the phenomenon. Further, we describe the log-periodogram regression and the rescaled-range statistic as methods to quantify LRD. This is followed by introducing two LRD specific statistics which can be used as further criteria for the simulation-based model selection described in Section 2.3.2. These statistics are based on the log-periodogram regression and on DFA. Subsequently, we describe the generation of an artificial series, which is used as a challenging example to illustrate the LRD detection strategies described in Chapter 3.

B.1 Effects of Long-Range Dependence

We illustrate the effect of LRD on statistical quantities estimated from a realisation of a stochastic process with a simple example: consider the arithmetic mean $\bar{X} = N^{-1} \sum_{t=1}^{N} X_t$ as an estimator for the the expectation value μ, i.e. $\widehat{\mu} = \bar{X}$. A basic result in statistics is the decrease in variance of the arithmetic mean with an increasing sample size N. For random variables X_t, identically distributed with expectation value μ and variance σ^2, the variance of the arithmetic mean $\bar{X}$ can be calculated as (**?**)

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N^2} \sum_{s,t=1}^{N} \mathrm{cor}(X_s, X_t). \tag{B.1}$$

For independent X_t with $\mathrm{cor}(X_s, X_t) = 0$ for $s \neq t$ this reduces to the well known result $\mathrm{var}(\bar{X}) = \sigma^2/N$. In case of a dependent process X_t, the calculation of $\mathrm{var}(\bar{X})$ requires to consider the ACF $\rho(\tau)$. For a weakly stationary process, we find

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N} \left[1 + 2 \sum_{\tau=1}^{N-1} \left(1 - \frac{\tau}{N} \right) \rho(\tau) \right]. \tag{B.2}$$

In case of a more specific description of the stationary process equation (B.2) can be further simplified. It is instructive to consider two examples: an autoregressive process (AR[1], cf. Section 1.2.1), as a prominent representative of a SRD processes, and a LRD process, specified only by the limiting behaviour (1.21).

Example 1: Short-Range Dependence

With the autocorrelation function for an AR[1] process $\rho(\tau) = a^\tau$ with $|a| < 1$ as described in Section 1.2.1, equation (B.2) reads

$$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{N}\left[1 + 2\sum_{\tau=1}^{N-1}\left(1 - \frac{k}{N}\right)a^\tau\right], \tag{B.3}$$

which in the limit of large N reduces to

$$\mathrm{var}(\bar{X}) \approx \frac{\sigma^2}{N}\left[1 + \frac{2a}{1-a}\right] = \frac{\sigma^2}{N}c_a, \tag{B.4}$$

with $c_a = 1 + \frac{2a}{1-a}$ being a constant depending on the AR coefficient. In this case the variance of the mean decreases at the same rate as for uncorrelated data, namely $\mathrm{var}(\bar{X}) \propto N^{-1}$, but with a different constant.

Example 2: Long-Range Dependence

For a process with an algebraically decaying ACF, $\rho(\tau) \propto \tau^{-\gamma}$ with $\gamma \in (0, 1)$, for large lags the sum in (1.20) does not converge to a constant value. It behaves instead as

$$\sum_{\tau=-(N-1)}^{N-1} \rho(\tau) \propto N^{1-\gamma}, \tag{B.5}$$

which in turn leads to

$$\mathrm{var}(\bar{X}) \propto \frac{\sigma^2}{N^\gamma}. \tag{B.6}$$

Because $\gamma < 1$, this algebraic decay of the variance of the mean is slower than the exponential decay for uncorrelated or short-range correlated processes (**?**).

These examples point towards the pivotal difference in records from LRD and SRD processes: given a sample of length N, for SRD processes the mean value can be estimated with the variance decreasing as $1/N$, while for a LRD process the variance of this estimator decreases only with $1/N^\gamma$ and is thus in general larger. Consequently, the uncertainty of statistical measures inferred from records of LRD processes is larger than in the case of SRD. Statistical measures which are affected are, e.g., regression coefficients (**?**), semi-parametric trend estimates (**??**) or quantiles estimated from the frequency distribution (**?**). The latter effect is of tremendous importance in extreme value statistics, e.g., when setting design values for hydraulic constructions as dams, bridges or dykes (**?**). The effect on regression coefficients is relevant with respect to trend detection, a highly debated topic in the early years of climate change research: assuming a LRD process underlying temperature records allows small trends to "hide" in the low frequency variability and renders detection more difficult (**?**). It is thus of considerable importance to reliably detect an underlying LRD process. Having discovered that a frequently used heuristic approach is prone to falsely detect LRD (Section A.2; **?**), the development of a different approach to the detection of LRD based on parametric modelling and model selection is suggested in Chapter 3.

B.2 Physical Explanations of the Phenomenon

In the time after **?**, various explanations or descriptions of the Hurst phenomenon emerged. In the 1960s Mandelbrot suggested increments of self-similar processes to describe this effect (cf. Section 1.2.2). The aggregation of processes with different time scales was suggested by **?**. Certain types of instationarities being held responsible for the phenomenon were discussed by **?**. The latter two approaches are discussed in brevity in the following. An extensive overview about the Hurst phenomenon from a hydrological perspective has been given by **??**.

Aggregation of Processes with Different Characteristic Time Scales

? showed that it is possible to preserve the Hurst phenomenon (in particular the rescaled range statistic, Section B.3.1) for a finite time series with conventional linear stochastic models (cf. Section 1.2.1). **?** showed that a superposition of n first order autoregressive processes (AR[1], Section 1.2.1) lead to a long-range dependent process in the limit of large n. The parameters a and σ_η of these processes (1.23) have to be independent random variables following a beta distribution. Physically, the parameter a is related to the relaxation time of the process. This implies that a suitable superposition of relaxations with different characteristic time scales can evoke the Hurst phenomenon.

Instationarities

Instationarities in the empirical series, such as a trend in the mean, have been held responsible for the Hurst phenomenon (e.g., **?**). The latter motivated the development of methods supposed to eliminate the influence of trends, like the detrended fluctuation analysis (DFA) described in Section 1.3.3.

Other Explanations

Further variants of models showing the Hurst phenomenon are proposed for example by **??** and **?**. Especially interesting is the idea of **?**. He suggests that heteroscedasticity, i.e. a non-homogeneous variance, might also account for the phenomenon. Heteroscedasticity can also be found in run-off records (**?**).

B.3 Further Heuristic and Semi-Parametric Methods to Quantify LRD

B.3.1 Rescaled Adjusted Range

? studied the flow of the Nile using the *rescaled adjusted range statistic* (R/S). Consider the inflow X_t at time t, the cumulative inflow $Y_t = \sum_{i=1}^{t} X_i$ and the mean $\bar{X}_{t,s} = 1/s \sum_{i=t+1}^{t+s} X_t$. The rescaled adjusted range is then

$$R/S = \mathrm{E}\left[\frac{R(t,s)}{S(t,s)}\right]_t, \tag{B.7}$$

with

$$R(t,s) = \max_{0\le i\le s}[Y_{t+i} - Y_t - \frac{i}{s}(Y_{t+s} - Y_t)] - \min_{0\le i\le s}[Y_{t+i} - Y_t - \frac{i}{s}(Y_{t+s} - Y_t)],$$

and

$$S(t,s) = \sqrt{s^{-1} \sum_{i=t+1}^{t+s} (X_i - \bar{X}_{t,s})^2}.$$

For large s, a plot of $\log R/S$ versus $\log s$ is approximately scattered around a straight line with slope H (**?**).

The R/S statistic was refined by Mandelbrot and others (e.g., **???**). Later **?** derived the asymptotic behaviour for R/S for short-range and long-range dependent processes. The latter can be described following **?**:

For X_t such that X_t^2 is ergodic and $t^{-\frac{1}{2}} \sum_{k=1}^{t} X_k$ converges weakly to *Brownian motion* as $t \to \infty$ then for $s \to \infty$ and $Q(t,s) = R(t,s)/S(t,s)$

$$s^{-\frac{1}{2}} Q(t,s) \xrightarrow{d} \xi, \tag{B.8}$$

with ξ being a nondegenerate random variable.

It remains open for which range of s the described asymptotic behaviour starts and what the confidence interval of the estimate is. For a finite sample the distribution of H is neither normal nor symmetric. All this makes inference about the value of H difficult.

In fact, in the presence of unanticipated high frequency behaviour, **?** noticed the invalidity of a test for $H = 0.5$, i.e. uncorrelated series. Based on the asymptotic theory for the R/S statistic by **?** this null hypothesis is rejected too often. **?** developed a corrected statistic accounting for a wide range of high frequency components.

The R/S statistic was not designed to discriminate between short-range and long-range dependent processes, it is rather a tool to measure the Hurst exponent H in situations where the region of asymptotic behaviour is known. This is a severe shortcoming encountered also in the description of further heuristic approach such as DFA (cf. Appendix A). In many fields where long-range dependence is to be taken into account, such as geophysics or econometrics, also high frequency components are present. They need to be accounted for in order to obtain reliable results.

B.3.2 Log-Periodogram Regression

In case a full-parametric description of the spectral density (Chapter 3) cannot be achieve or is not pursued. The fractional difference parameter can be estimated using a semi-parametric approach in the spectral domain. Referring to a definition of LRD (1.22), the spectral density $S(\omega)$ close to the origin behaves as

$$S(\omega) \propto c_f |\omega|^{-\beta}, \; |\omega| \to 0. \tag{1.22}$$

The exponent β can be related to the fractional difference parameter d by $\beta = 2d$. Taking the logarithm on both sides, the equation motivates a regression approach to estimate β or, equivalently, the fractional difference parameter d. Consider

$$\log I(\omega_j) \approx \log c_f - \beta \log |\omega_j| + u_j, \tag{B.9}$$

with a positive constant c_f and $j = l, \ldots, m$.

The GPH Estimator

? (GPH) were the first to suggest the estimation of a parameter quantifying the long-range dependence (either β or d) using this kind of regression. However, instead of $\log \omega_j$ they

considered $\log|1 - e^{i\omega_j}|$ as regressor motivated by the exact form of the spectral density for fractional ARIMA processes (1.50). Furthermore, they included frequencies ω_j in the regression starting with $j = 0$ to obtain the estimate $\widehat{d}_{\text{lp}} = \widehat{\beta}/2$. **?** argued for the following Gaussian limiting distribution of this estimator:

$$m^{1/2}(\widehat{d}_{\text{lp}} - d) \xrightarrow{d} N\left(0, \frac{\pi^2}{24}\right). \tag{B.10}$$

With this result, they provided a simple way to approximate confidence intervals and carry out hypothesis testing (cf. Section 2.1.1).

However, **?** showed that their arguments do not hold. Mainly because the residuals u_j cannot be considered as asymptotically independent or homoscedastic. Despite this, the desired limiting distribution (B.10) could be established by **?** for Gaussian processes using $\log \omega_j$ as regressor as suggested in (B.9) and by trimming out the smallest frequency ω_j (**?**). This result was generalised to linear processes by **?**.

Semi-Parametric Whittle-Estimation

An even more efficient estimator was suggested by **?** with the limiting distribution provided by **?**. It is essentially based on Whittle estimation of the model $S(\omega) = C\omega^{-2d}$ for the first m frequencies:

$$\widehat{d}_K = \arg\min_d \left\{ \log\left[\frac{1}{m}\sum_{j=1}^{m} \frac{I(\omega_j)}{\omega_j^{-2d}}\right] - \frac{2d}{m}\sum_{j=1}^{m}\log\omega_j \right\}. \tag{B.11}$$

Under milder regularity conditions than those required for (B.9), the limiting distribution is given by

$$m^{1/2}(\widehat{d}_k - d) \xrightarrow{d} N\left(0, \frac{1}{4}\right), \tag{B.12}$$

having a smaller variance and leading thus to a more efficient estimator. Different from (B.9) no closed form for $\widehat{d}_k$ can be given. The estimate has to be obtained numerically in the same way as described for the full-parametric Whittle estimator in Section 1.3.2.

Bandwidth Choice and Related Approaches

For $\widehat{d}_K$, as well as for $\widehat{d}_{\text{lp}}$ a bandwidth m has to be chosen for the estimation. A brief review on various approaches on the choice of an optimal bandwidth is given in **?**. **?** discuss an optimal choice of the bandwidth in the sense of a bias-variance trade-off.

The log-periodogram regression for the non-stationary domain $d > 0.5$ has been discussed by **?**. A similar approach to estimate the long-range dependence parameter using wavelet analysis is described by **?** and **?**.

Combining Log-Periodogram Regression and ARMA$[p, q]$ Parameter Estimation

It is conceivable to estimate the (fractional) difference parameter d and the autoregressive moving average parameters a_i and b_j separately. In the early days of ARIMA$[p, d, q]$ processes (with $d \in \mathbb{N}$), the estimation for integer differences has been carried out by heuristic methods and for the following inference about the AR and MA parameters d was assumed to be known (**?**). For fractional differences $d \in \mathbb{R}$, the first step can be

carried out using, e.g., the log-periodogram regression; the inference about ARMA parameters, for example, by maximum likelihood. For the log-periodogram regression it is possible to specify a confidence interval for the estimate of d in contrast to the heuristic methods of estimating the integer difference paramters. However, the influence of estimating d on the estimation of the autoregressive and moving average parameters is not accounted for. In a combined approach, i.e. simultaneous estimation of the parameters d, a_i and b_j, this influence can be quantified via the covariance matrix $\mathbf{V}$ of the parameter vector $\boldsymbol{\theta}$ (Section 1.3.1).

B.4 Specific Test Statistics for Model Selection

In this section, we evaluate two statistics which can be used in the simulation-based model selection described in Section 2.3.2. Being especially interested in the low frequency behaviour, we investigate the performance of the log-periodogram regression (Section B.3.2) and the DFA (Section 1.3.3) as further model selection criteria. For both approaches one has to choose an appropriate bandwidth for the straight line fit. We use this free parameter to optimise the power (Section 2.3.3), i.e. the separation of the two distributions resulting from the simulations. For both criteria, we calculate p-values and power and try to identify an optimal bandwidth. It turned out to be also instructive to examine the dependence of the estimates on the bandwidth.

As an example, we consider the setting introduced in Section 3: a realisation of an ARMA$[3,2]$ process (Section 3.1; Appendix B.5) and models FAR$[1]$ and ARMA$[3,2]$. We estimate the model parameters from the realisation and obtain thus two fully specified models. These two models are used to generate two ensembles of realisations. For all ensemble members the estimates $\widehat{d}_{\mathrm{lp}}$ (or $\widehat{d}_{\mathrm{DFA}}$, Section 1.3.3) are calculated. The estimate from the observed series is then compared to the distribution obtained from the ensembles. The p-values and power can be obtained in the same way as described for the likelihood-ratio (cf. Section 2.3.2). For better readability, we denote the ARMA$[3,2]$ model as f and the corresponding hypothesis as H_f, analogously g represents the FAR$[1]$ model.

B.4.1 Log-Periodogram Regression

Bandwidth-Dependence of the Power

The power for a critical nominal level of $\alpha = 0.05$ as a function of the bandwidth of the linear fit is shown in Figure B.1 (left) for testing H_f against H_g (green) and vice versa (red). In both cases a maximum power is achieved for a bandwidths $0 < \omega < \omega_1 = 0.0021$: $\widehat{\mathrm{pow}}_f(g, 0.05) = 0.851$ and $\widehat{\mathrm{pow}}_g(f, 0.05) = 0.836$. Thus this test is not as powerful as the one based on the likelihood-ratio which exhibits a power $\widehat{\mathrm{pow}} > 0.9$ for both test situations (Section 3.4.1). The distribution functions for the ensembles of fractional difference parameter estimates for the two hypothesis are shown in Figure B.1 (right) for the bandwidth with maximal power. The estimate obtained for the "observed" series is depicted as a vertical bar and is located in the gap between the two critical regions. Thus, neither hypothesis can be rejected in this example.

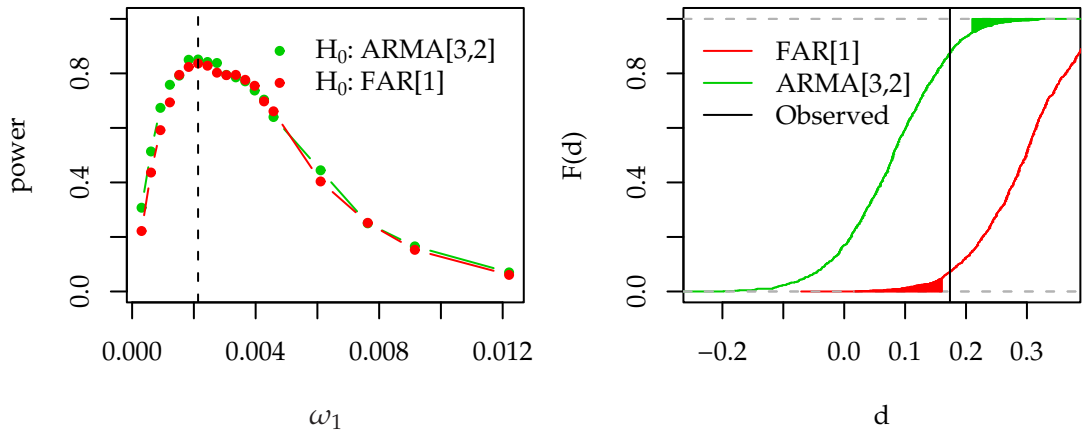

Figure B.1: Power of testing H_f against H_g (green) and vice versa (red) at level $\alpha = 0.05$ for various bandwidths $0 < \omega < \omega_1$ in the spectral domain (left). The lower bound ω_1 where maximal power is achieved is marked with a dashed vertical line. Distributions of the estimated fractional difference parameters obtained from the bootstrap ensembles using the log-periodogram regression and the bandwidths of maximal power (right).

Calculating p-Values

We further calculate p-values analogously to Section 2.3.2. For the bandwidth of maximal power, testing H_f yields a p-value of $\widehat{p}_f = 0.126$. Testing H_g, we find $\widehat{p}_g = 0.071$. Figure B.2 shows the p-values for different bandwidths together with the power for the two tests. The left plot refers to testing H_f, the right plot H_g. In regions of high power,

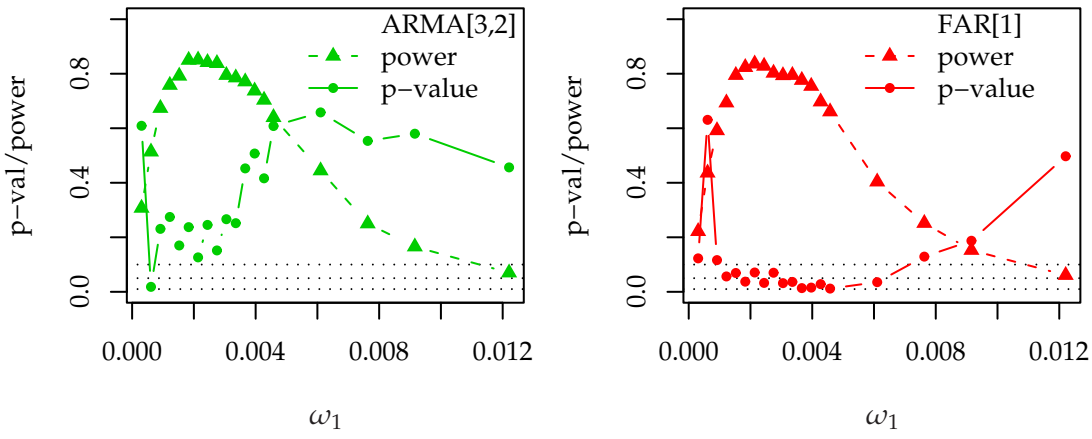

Figure B.2: p-values for different bandwidths together with the power for testing H_f (left) and H_g (right) based on the log-periodogram regression. The dotted horizontal lines mark the 1%, 5% and 10%-level (bottom to top).

we find p-values for H_f fluctuating above any reasonably level of significance in the left plot. Whereas testing H_g p-values close to, or even below the 5%-level of significance can be observed in this region. This can be taken as indications for rejecting the FAR[1] in favour of the ARMA$[3, 2]$.

Bandwidth-Dependence of the Estimate

We confront the estimates $\widehat{d}_{\mathrm{lp}}$ for different bandwidth with the corresponding mean values and standard deviations of the ensembles. This might provide further indications for the model selection (Figure B.3). For a small bandwidth the fluctuation of the estimate $\widehat{d}_{\mathrm{lp}}$

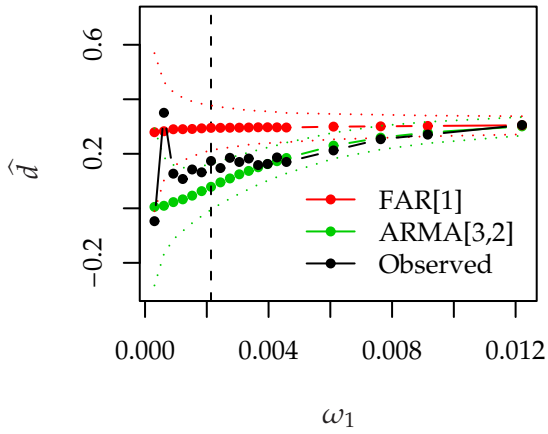

Figure B.3: Dependence of the mean and standard deviation of the distributions and the estimate on the bandwidth. The vertical line marks the bandwidth where maximal power is attained.

obtained from the observed series is large. It falters between the mean values of both hypothesis. With an increasing bandwidth this estimate approaches the course of the mean of H_f and stays close to it while the two distributions merge. This may be taken as further indication for the ARMA$[3,2]$ hypothesis.

B.4.2 Detrended Fluctuation Analysis

Bandwidth-Dependence of the Power

We calculate the power as a function of the range of scales (or bandwidth) for the straight line fit to the log-fluctuation function obtained with DFA (Figure B.4, left). For testing H_f, we find a maximum power at the standard 5%-level of $\widehat{\mathrm{pow}}_f(g, 0.05) = 0.861$ for a range of scales $\log s > \log s_1 = 2.6$. The maximum power $\widehat{\mathrm{pow}}_g(g, 0.05) = 0.851$ for H_g is attained for a range of scales $\log s > \log s_1 = 2.7$. Also using this statistic does not yield a test as powerful as the one based on the likelihood ratio (Section 3.4.1). The two distribution functions for the fractional difference parameter estimates are shown in Figure B.4 (right) together with the estimate of the "observed" series. Similar to the log-periodogram regression, we find the observed value $\mathrm{lr}_{\mathrm{obs}} = 0.190$ in a gap between the two critical regions. We can thus not reject neither hypothesis.

Calculating p-Values

We calculate the p-values for the range of scales with maximum power. This yields $\widehat{p}_f = 0.132$ for a test of compatibility with H_f and $\widehat{p}_g = 0.057$ for H_g. Figure B.5 shows the p-values together with the respective power for a variety of ranges of scales used for the straight line fit. Here, we also find the p-values in regions of high power predominantly above a 10%-level for H_f and frequently below the 5%-level for H_g. We note that the p-values do not fluctuate as much as they do in the spectral domain. However, the

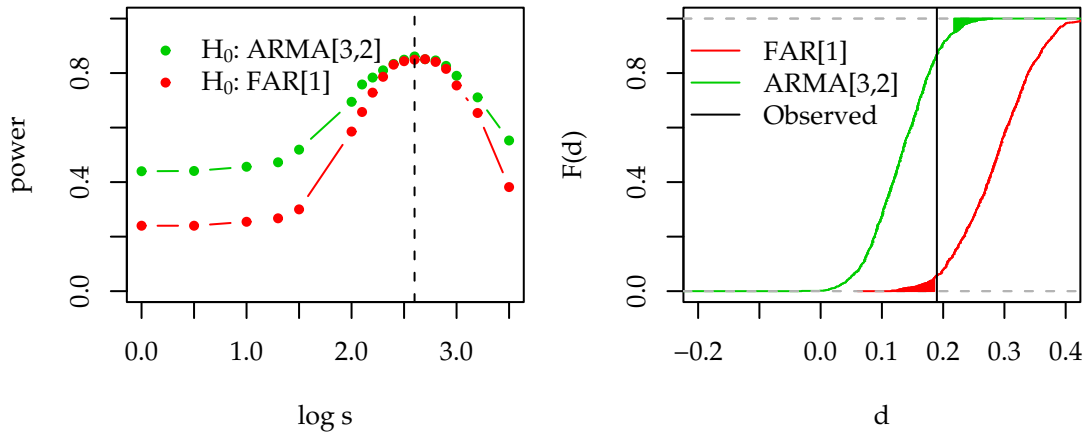

Figure B.4: Power of testing H_f hypothesis against H_g (green) and vice versa (red) for various ranges of scales $\log s > \log s_1$ in the time domain (left). Distributions of the estimated fractional difference parameters (right) obtained from the bootstrap ensembles using DFA in the time domain and the range of scales where maximal power for testing hypothesis f is achieved.

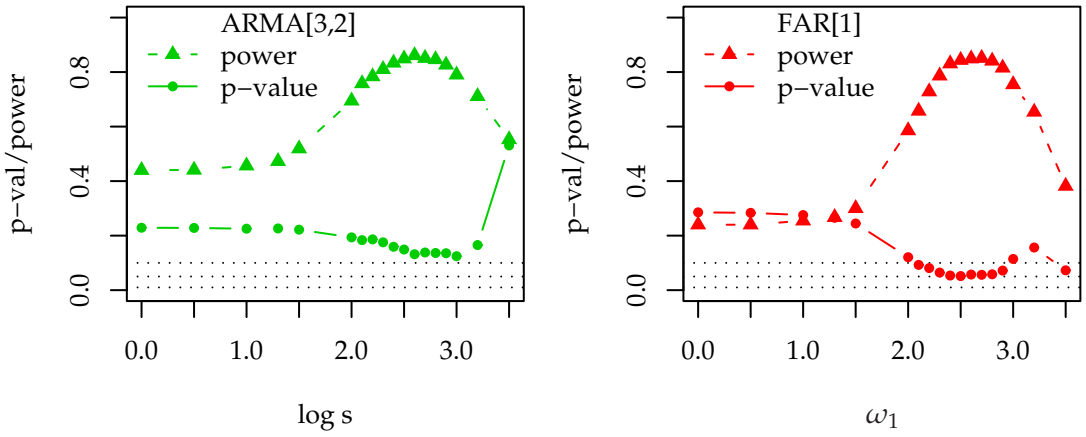

Figure B.5: Power and p-value for the testing H_f (left) and H_g (right) based on the DFA. The horizontal lines mark the 1%, 5% and 10%-levels (bottom to top).

difference to the 10%-level in the case of model f and the difference to the 5%-level in case of model g are also less pronounced.

Bandwidth-Dependence of the Estimate

The p-values changing with the range of scales used for the fit (Figure B.6) indicate that the estimates $\widehat{d}_{\text{DFA}}$ rather follows the mean of H_f. The distinction is not as clear as it is in the case of the log-periodogram regression in Figure B.3. This, as well as the p-values

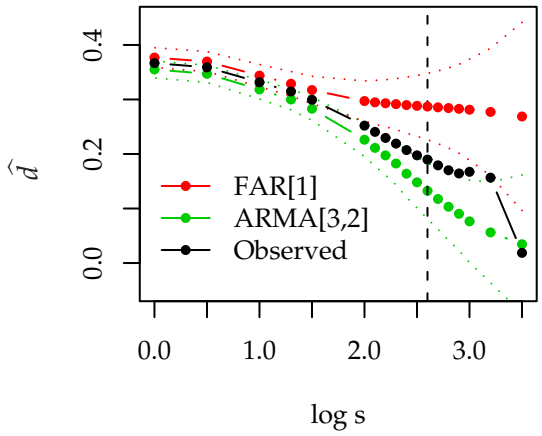

Figure B.6: Development of the mean and standard deviation of the hypothesis with varying range of scales for DFA. The vertical line marks the range where maximal power is attained.

obtained in the previous paragraph, may be taken as a hint to favour the ARMA$[3, 2]$.

B.5 Constructing the Example Process – Detailed Description

The aim of this section is to generate challenging example series for testing the strategies to detect LRD (Chapter 3). We start with a realisation $\{x_i\}_{i=1,\ldots,N}$ of an FARIMA$[1, d, 0]$ (or, for short, FAR[1]) process with parameters $a_1 = 0.7$, $d = 0.3$ and length $N = 2^{15} = 32\,768$. The length and the parameters are chosen such that they are plausible for river run-off, the context which we are going to apply this detection strategy in. A realisation of this process is shown in the time and spectral domain in Figure B.7. Now, we aim to find a SRD model which describes this realisation well. We expect such a SRD model to produce realisations which will be difficult to discriminate from realisations of the original LRD process.

Selecting an ARMA$[p, q]$ model for the FAR$[1]$ Realisation

We fit various ARMA$[p, q]$ models with $p < 4$ and $q < \min(p, 2)$ and compare the values for the Hannan-Quinn criterion (HIC) (cf. Section 2.2.2) in Figure B.8. The smallest value is attained for the ARMA$[3, 2]$ process. Very close to this value, separated by less than 2 units, we find the ARMA$[4, 1]$ model. About 4 units more yields the HIC for ARMA$[4, 2]$. Note, that at this stage we do not necessarily intent to find one "best model". We are rather looking for some ARMA$[p, q]$ model representing the given FAR$[1]$ realisation well enough, such that realisations can be easily mistaken for coming from an LRD process.

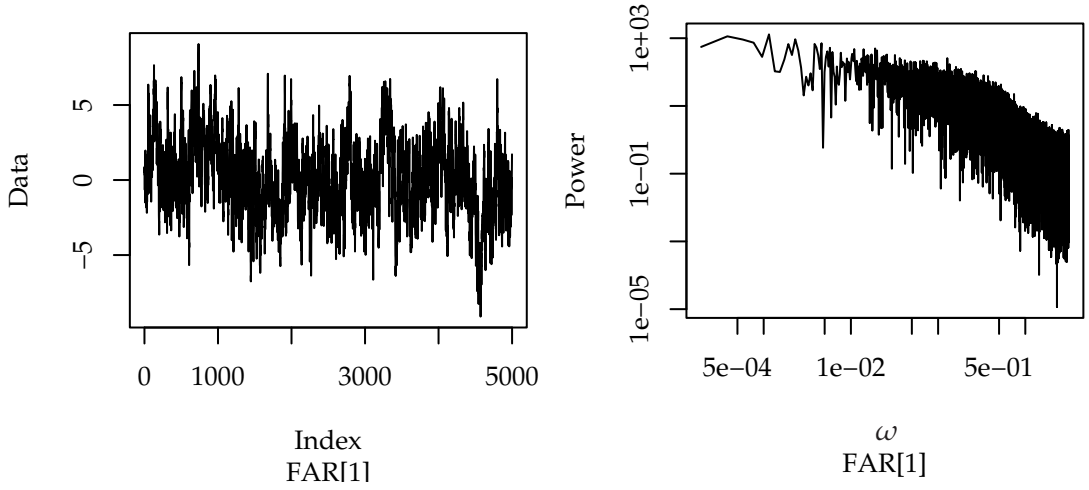

Figure B.7: Realisation of a FAR[1] process with parameters $a_1 = 0.7$, $d = 0.3$ and length $N = 32\,768$, in the time domain (left, only the first 5 000 data points) and in the frequency domain (right).

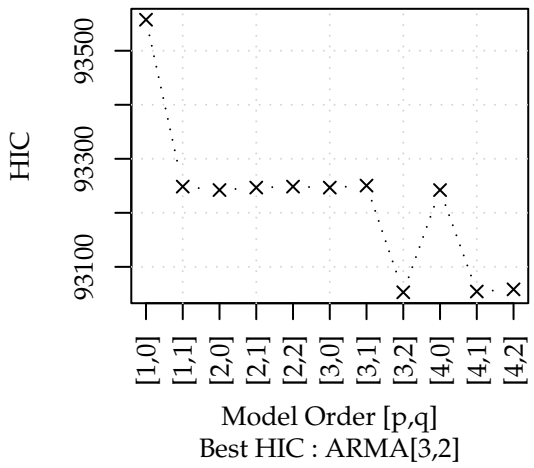

Figure B.8: Multiple ARMA$[p, q]$ fits to the realisation of the FAR$[1]$ process shown in Figure B.7. The plot shows the values for the Hannan-Quinn criterion (HIC) for models with the order $1 < p < 4$ and $0 < q < \min(p, 2)$.

There is no need to discriminate between these three similar performing models. For our upcoming example, we choose the ARMA$[3,2]$ process with parameters $a_1 = 2.178, a_2 = -1.448, a_3 = 0.309, b_1 = -1.184$ and $b_2 = 0.218$. The model is completely specified and we can now generate a realisation, which are used to study the detection strategies for LRD in Chapter 3.