Chapter 5
Introduction to Long-Range Dependence

5.1 The Hurst Phenomenon

The history of long-range dependence as a concrete phenomenon believed to be important in its own right should be regarded as beginning in the 1960s with a series of papers by Benoit Mandelbrot and his coworkers, such as Mandelbrot (1965) and Mandelbrot and Wallis (1968). The cause was a need to explain an empirical finding by Hurst (1951, 1956) that studied the flow of water in the Nile. A particular data set studied by Hurst appears in Figure 5.1.

Many features of this data set are interesting (one of which is how long ago the data were collected). Harold Hurst, who was interested in the design of dams, looked at these data through a particular statistic. Given a sequence of n observations $X_1, X_2, \ldots, X_n$, define the partial sum sequence $S_m = X_1 + \ldots + X_m$ for $m = 0, 1, \ldots$ (with $S_0 = 0$). The statistic Hurst calculated is

$$\frac{R}{S}(X_1, \ldots, X_n) = \frac{\max_{0 \le i \le n}(S_i - \frac{i}{n}S_n) - \min_{0 \le i \le n}(S_i - \frac{i}{n}S_n)}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{n}S_n)^2)^{1/2}}. \tag{5.1}$$

Note that S_n/n is the sample mean of the data. Therefore, $\max_{0 \le i \le n}(S_i - \frac{i}{n}S_n)$, for example, measures how far the partial sums rise above the straight line they would follow if all observations were equal (to the sample mean), and the difference between the maximum and the minimum of the numerator in (5.1) is the difference between the highest and lowest positions of the partial sums with respect to the straight line of uniform growth. It is referred to as the range of observations. The denominator of (5.1) is, of course, the sample standard deviation. The entire statistic in (5.1) has then been called the *rescaled range* or R/S *statistic*.

When Harold Hurst calculated the R/S statistic on the Nile data in Figure 5.1, he found that it grew as a function of the number n of observations, approximately as $n^{0.74}$.

G. Samorodnitsky, *Stochastic Processes and Long Range Dependence*,
Springer Series in Operations Research and Financial Engineering,
DOI 10.1007/978-3-319-45575-4_5

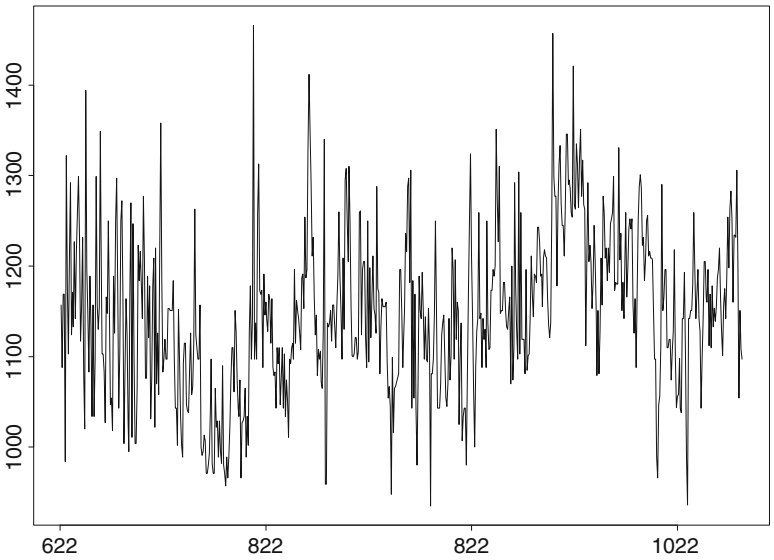

Fig. 5.1 Annual minima of the water level in the Nile for the years 622 to 1281, measured at the Roda gauge near Cairo

To see that this observation is interesting, let us suppose that $X_1, X_2, \ldots$ is a sequence of random variables. If we apply the R/S statistic to the first n observations $X_1, X_2, \ldots, X_n$ for increasing values of n, what would we expect the resulting sequence of values of the R/S statistic to be like, for the "usual" models of $X_1, X_2, \ldots$?

Example 5.1.1. Suppose that $X_1, X_2, \ldots$ is, in fact, a stationary sequence of random variables with a finite variance and a common mean μ. Define the centered partial sum process by

$$S^{(n)}(t) = S_{[nt]} - [nt]\mu, \ 0 \le t \le 1\,. \tag{5.2}$$

The classical functional central limit theorem (Donsker's theorem, invariance principle) says that if $X_1, X_2, \ldots$ are i.i.d., then

$$\frac{1}{\sqrt{n}} S^{(n)} \Rightarrow \sigma_* B \quad \text{weakly in } D[0,1], \tag{5.3}$$

where σ_*^2 is equal to the common variance σ^2 of the observations, and B is the standard Brownian motion on $[0,1]$ (Theorem 14.1 in Billingsley (1999)). Here $D[0,1]$ is the space of right continuous functions on $[0,1]$ having left limits equipped with the Skorokhod J_1 topology. In fact, the functional central limit theorem is known to hold for stationary processes with a finite variance that are much more

general than an i.i.d. sequence (with the limiting standard deviation σ_* not equal, in general, to the standard deviation of the X_i); see a survey by Merlevéde et al. (2006).

The function $f : D[0, 1] \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \sup_{0\le t\le 1} (x(t) - tx(1)) - \inf_{0\le t\le 1} (x(t) - tx(1)) ,$$

$\mathbf{x} = (x(t),\ 0 \le t \le 1) \in D[0, 1]$, is easily seen to be continuous. It is straightforward to check that the range of the first n observations (the numerator in the R/S statistic) is equal to $f(S^{(n)})$. Therefore, if the invariance principle (5.3) holds, then by the continuous mapping theorem, Theorem 10.2.4,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}(\text{the range of the first } n \text{ observations}) = f\left(\frac{1}{\sqrt{n}} S^{(n)}\right) \\
&\Rightarrow f(\sigma_* B) = \sigma_* \Big[\sup_{0\le t\le 1} (B(t) - tB(1)) - \inf_{0\le t\le 1} (B(t) - tB(1)) \Big] \\
&\qquad := \sigma_* \big[\sup_{0\le t\le 1} B_0(t) - \inf_{0\le t\le 1} B_0(t) \big],
\end{aligned}$$

where B_0 is a Brownian bridge on $[0, 1]$. Further, if the stationary process $X_1, X_2, \ldots$ (or its bilateral extension in Proposition 1.1.2) is ergodic, then by the pointwise ergodic theorem, Theorem 2.1.1 (or (2.8)), we have, with probability 1,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (X_i - \frac{1}{n} S_n)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&\to E\big(X_1^2\big) - \Big(E\big(X_1\big)\Big)^2 = \sigma^2 .
\end{aligned}$$

Assuming, therefore, that the functional central limit theorem holds, and that the observations form an ergodic process, we see that

$$\frac{1}{\sqrt{n}} \frac{R}{S}(X_1, \ldots, X_n) \Rightarrow \frac{\sigma_*}{\sigma} \big[\sup_{0\le t\le 1} B_0(t) - \inf_{0\le t\le 1} B_0(t) \big]. \tag{5.4}$$

That is, the R/S statistic grows, distributionally, as the square root of the sample size.

The distributional $n^{0.5}$ rate of growth of the values of the R/S statistic obtained under, apparently quite reasonable, assumptions of Example 5.1.1 looks incompatible with the empirical $n^{0.74}$ rate of growth Hurst observed on the Nile data. Therefore, if one wants to construct a stochastic model of observations with a similar behavior of the R/S statistic to the one observed by Hurst, some of the "reasonable" assumptions of the example must be dropped.

The following example looks at what happens when the assumption of finite variance of the observations is dropped and replaced by an assumption of appropriately heavy tails.

Example 5.1.2. In order to be able to concentrate better on the effect of heavy tails, we will assume that the observations $X_1, X_2, \ldots$ are i.i.d. Assume that the balanced regular variation property of Definition 4.2.7 holds, and $0 < \alpha < 2$; this guarantees that the variance of the observations is infinite. In this case, we will apply the Poisson convergence result in Theorem 4.4.1 to understand the "size" of the R/S statistic.

We begin with a typical "truncation" step, needed because various sums of points are not continuous functionals of point processes in the topology of vague convergence. For $\epsilon > 0$, let

$$S_m^{(\epsilon)} = \sum_{j=1}^{m} X_j \mathbf{1}\big(|X_j| > \epsilon a_n\big), \; m = 0, 1, 2 \ldots ,$$

where, as usual, $a_n = \inf\big\{x > 0 : P(|X_1| > x) \le 1/n\big\}$, $n = 1, 2, \ldots$. Consider a modified version of the R/S statistic defined by

$$RS_n(\epsilon) = \frac{\max_{0\le i\le n}(S_i^{(\epsilon)} - \frac{i}{n}S_n^{(\epsilon)}) - \min_{0\le i\le n}(S_i^{(\epsilon)} - \frac{i}{n}S_n^{(\epsilon)})}{(\sum_{i=1}^{n} X_i^2 \mathbf{1}\big(|X_j| > \epsilon a_n\big))^{1/2}}. \tag{5.5}$$

Note that $RS_n(\epsilon) = g_\epsilon(N_n)$, where N_n is the point process in (4.52) and $g_\epsilon : M_+^{\mathrm{R}}\big([0,1] \times \overline{\mathbb{R}}_0^d\big) \to (0, \infty)$ is defined by

$$g_\epsilon(K) = \frac{R_\epsilon(K)}{(\int_{[0,1]\times(\mathbb{R}\setminus[-\epsilon,\epsilon])} y^2 K(ds, dy))^{1/2}},$$

with

$$R_\epsilon(K) = \sup_{0\le t\le 1}\left[\int_{[0,t]\times(\mathbb{R}\setminus[-\epsilon,\epsilon])} yK(ds, dy) - t\int_{[0,1]\times(\mathbb{R}\setminus[-\epsilon,\epsilon])} yK(ds, dy)\right]$$
$$- \inf_{0\le t\le 1}\left[\int_{[0,t]\times(\mathbb{R}\setminus[-\epsilon,\epsilon])} yK(ds, dy) - t\int_{[0,1]\times(\mathbb{R}\setminus[-\epsilon,\epsilon])} yK(ds, dy)\right].$$

According to Exercise 4.6.8, the law of the limiting Poisson process in Theorem 4.4.1 does not charge the set of the discontinuities of the function g_ϵ; see Exercise 5.5.1. Therefore, by the continuous mapping theorem (Theorem 10.2.4),

$$g_\epsilon(N_n) \Rightarrow g_\epsilon(N) \quad \text{in } [0, \infty) \text{ as } n \to \infty,$$

and we can represent the limit distributionally as

$$g_\epsilon(N) = \frac{\sup_{0\le t\le 1} Y_\epsilon(t) - \inf_{0\le t\le 1} Y_\epsilon(t)}{(\sum_{j=1}^{\infty} \Gamma_j^{-2/\alpha} \mathbf{1}(\Gamma_j < \epsilon^{-\alpha}))^{1/2}}\,.$$

Here

$$Y_\epsilon(t) = \sum_{j=1}^{\infty} \big(\mathbf{1}(U_j \le t) - t\big)\theta_j \Gamma_j^{-1/\alpha} \mathbf{1}(\Gamma_j < \epsilon^{-\alpha}),\ \ 0 \le t \le 1\,. \tag{5.6}$$

Recall that (U_i) is a sequence of i.i.d. standard uniform random variables, (θ_i) is a sequence of i.i.d. random variables taking the value 1 with probability p, and the value -1 with probability $q = 1 - p$, and (Γ_i) is a sequence of standard Poisson arrivals on $(0, \infty)$, with all three sequences being independent.

Corollary 3.4.2 says that Y_ϵ is an infinitely divisible process. Furthermore, if we take $0 < \epsilon_1 < \epsilon_2$ and use the same random ingredients in (5.6) for the two processes, Y_{ϵ_1} and Y_{ϵ_2}, then the difference $Y_{\epsilon_1} - Y_{\epsilon_2}$ can be written in the form

$$Y_{\epsilon_1}(t) - Y_{\epsilon_2}(t) = L_{\epsilon_1,\epsilon_2}(t) - tL_{\epsilon_1,\epsilon_2}(1),\ \ 0 \le t \le 1,$$

where

$$L_{\epsilon_1,\epsilon_2}(t) = \sum_{j=1}^{\infty} \mathbf{1}(U_j \le t)\theta_j \Gamma_j^{-1/\alpha} \mathbf{1}\big(\epsilon_2^{-\alpha} \le \Gamma_j < \epsilon_1^{-\alpha}\big),\ \ 0 \le t \le 1\,.$$

By Corollary 3.4.2, $L_{\epsilon_1,\epsilon_2}$ is a Lévy process without a Gaussian component, whose one-dimensional Lévy measure $\rho_{\epsilon_1,\epsilon_2}$ is the measure

$$m(dx) = \big(p\mathbf{1}(x > 0) + q\mathbf{1}(x < 0)\big)\alpha|x|^{-(\alpha+1)}\,dx$$

restricted to the set $(-\epsilon_2, -\epsilon_1) \cup (\epsilon_1, \epsilon_2)$, and whose local shift is

$$b = (2p - 1)\int_{\epsilon_2^{-\alpha}}^{\epsilon_1^{-\alpha}} [\![x^{-1/\alpha}]\!]\,dx\,;$$

see Example 3.2.3. It is, in fact, a compound Poisson Lévy process. It follows from the general properties of Lévy processes that $\sup_{0\le t\le 1} |L_{\epsilon_1,\epsilon_2}(t)| \to 0$ in probability as $\epsilon_2 \to 0$, uniformly in $0 < \epsilon_1 < \epsilon_2$; see Kallenberg (1974). Therefore, the same is true for $\sup_{0\le t\le 1} |Y_{\epsilon_1}(t) - Y_{\epsilon_2}(t)|$. We conclude that

$$Y_\epsilon(t) \to Y(t) := \sum_{j=1}^{\infty} \big(\mathbf{1}(U_j \le t) - t\big)\theta_j \Gamma_j^{-1/\alpha},\ \ 0 \le t \le 1,\ \ \text{as } \epsilon \to 0$$

in probability in the uniform topology in the space $D([0,1])$. Therefore, as $\epsilon \to 0$,

$$g_\epsilon(N) \to \frac{\sup_{0\le t\le 1}\sum_{j=1}^{\infty}\big(\mathbf{1}(U_j \le t) - t\big)\theta_j \Gamma_j^{-1/\alpha} - \inf_{0\le t\le 1}\sum_{j=1}^{\infty}\big(\mathbf{1}(U_j \le t) - t\big)\theta_j \Gamma_j^{-1/\alpha}}{(\sum_{j=1}^{\infty}\Gamma_j^{-2/\alpha})^{1/2}} \tag{5.7}$$

in probability. Notice that by the strong law of large numbers, $\Gamma_j/j \to 1$ as $j \to \infty$ with probability 1. This and the fact that $0 < \alpha < 2$ imply that the dominator on the right-hand side is finite and justifies the convergence.

Let $g(N)$ denote the random variable on the right-hand side of (5.7). If one shows that the statement

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty} P\left(\left|\frac{1}{\sqrt{n}}\frac{R}{S}(X_1,\ldots,X_n) - RS_n(\epsilon)\right| > \lambda\right) = 0 \tag{5.8}$$

for every $\lambda > 0$ is true, then a standard weak convergence argument, e.g., in Theorem 3.2 in Billingsley (1999), allows us to conclude that

$$\frac{1}{\sqrt{n}}\frac{R}{S}(X_1,\ldots,X_n) \Rightarrow g(N)\,. \tag{5.9}$$

Note that (5.9) means that even in the heavy-tailed case, the R/S statistic grows as the square root of the sample size.

The validity of (5.8) is verified in Exercise 5.5.2.

We conclude, therefore, as was done in Mandelbrot and Taqqu (1979), that infinite variance alone cannot explain the Hurst phenomenon. A different drastic departure from the assumptions leading to the square root of the sample size rate of growth of the R/S statistic was suggested in Mandelbrot (1965), and it had nothing to do with heavy tails. The idea was, instead, to take as a model a stationary process with a finite variance, but with correlations decaying so slowly as to invalidate the functional central limit theorem (5.3). The simplest model of that sort is *fractional Gaussian noise*, which is the increment process of fractional Brownian motion.

Let us begin with a fractional Brownian motion, or *FBM*, constructed in Example 3.5.1. This is a zero-mean Gaussian process $(B_H(t),\ t \ge 0)$ that is self-similar with exponent of self-similarity $H \in (0,1)$ and stationary increments. These properties imply that $B_H(0) = 0$ and $E(B_H(t) - B_H(s))^2 = \sigma^2|t-s|^{2H}$ for some $\sigma > 0$; see Section 8.2. Taking an appropriately high moment of the increment and using the Kolmogorov criterion in Theorem 10.7.7 allows us to conclude that a fractional Brownian motion has a continuous version, and we always assume that we are working with such a version.

A fractional Gaussian noise, or *FGN*, is a discrete step increment process of a fractional Brownian motion defined by $X_j = B_H(j) - B_H(j-1)$ for $j = 1, 2, \ldots$. The stationarity of the increments of the FBM implies that this is a stationary Gaussian

process. Using the fact $ab = (a^2 + b^2 - (a-b)^2)/2$ and the incremental variance of the FBM, we easily see that

$$\mathrm{Cov}(X_{j+n}, X_j) = \frac{\sigma^2}{2}\Big[(n+1)^{2H} + |n-1|^{2H} - 2n^{2H}\Big] \tag{5.10}$$

for $j \geq 1$, $n \geq 0$. That is,

$$\rho_n := \mathrm{Corr}(X_{j+n}, X_j) \sim H(2H-1)n^{-2(1-H)} \quad \text{as } n \to \infty. \tag{5.11}$$

In particular, $\rho_n \to 0$ as $n \to \infty$. This implies that the FGN is a mixing, hence ergodic, process; see Example 2.2.8. Furthermore, by the self-similarity of the fractional Brownian motion, for every n,

$$\mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}B_H(n) = \sigma n^{2H}\,. \tag{5.12}$$

Suppose now that a set of observations $X_1, X_2, \ldots$ forms a fractional Gaussian noise as defined above, and let us consider the behavior of the R/S statistic on these observations. The ergodicity of the FGN implies that the denominator of the statistic converges a.s. to the standard deviation of the observations, σ; see Example 2.1.5. For the numerator of the R/S statistic, we notice that $S_i = B_H(i)$ for every i, and the self-similarity of the FBM gives us

$$\begin{aligned}
&\max_{0\leq i\leq n}(S_i - \frac{i}{n}S_n) - \min_{0\leq i\leq n}(S_i - \frac{i}{n}S_n)\\
&= \max_{0\leq i\leq n}(B_H(i) - \frac{i}{n}B_H(n)) - \min_{0\leq i\leq n}(B_H(i) - \frac{i}{n}B_H(n))\\
&\stackrel{d}{=} n^H\left[\max_{0\leq i\leq n}(B_H(\frac{i}{n}) - \frac{i}{n}B_H(1)) - \min_{0\leq i\leq n}(B_H(\frac{i}{n}) - \frac{i}{n}B_H(1))\right].
\end{aligned}$$

By the continuity of the sample paths of the fractional Brownian motion, we have

$$\begin{aligned}
&\max_{0\leq i\leq n}(B_H(\frac{i}{n}) - \frac{i}{n}B_H(1)) - \min_{0\leq i\leq n}(B_H(\frac{i}{n}) - \frac{i}{n}B_H(1))\\
&\to \sup_{0\leq t\leq 1}(B_H(t) - tB_H(1)) - \inf_{0\leq t\leq 1}(B_H(t) - tB_H(1))
\end{aligned}$$

with probability 1. That is, for the FGN,

$$n^{-H}\frac{R}{S}(X_1, \ldots, X_n) \Rightarrow \sup_{0\leq t\leq 1}(B_H(t) - tB_H(1)) - \inf_{0\leq t\leq 1}(B_H(t) - tB_H(1))\,,$$

and so the R/S statistic grows distributionally at the rate n^H as a function of the sample size. Therefore, selecting an appropriate H in the model will, finally, explain

the Hurst phenomenon. In particular, the exponent H of self-similarity of fractional Brownian motion is often referred to as a *Hurst parameter*.

This success of the fractional Gaussian noise model in explaining the Hurst phenomenon is striking. We have used the self-similarity of the fractional Brownian motion in the above computation, but it is not hard to see that a very important property of the fractional Gaussian noise is the unusually slow decay of correlations in (5.11), especially for high values of H (i.e., close to 1). For these values of H, the variance of the partial sums in (5.12) also increases unusually fast. Unlike the previous unsuccessful attempt to explain the Hurst phenomenon by introducing in the model unusually heavy tails (infinite variance in this case), the FGN model succeeds here by introducing unusually long memory. Particularly vivid terminology was introduced in Mandelbrot and Wallis (1968), in the context of weather and precipitation: unusually heavy tails were called the *Noah effect*, referring to the biblical story of Noah and extreme incidents of precipitation, while unusually long memory was called the *Joseph effect*, referring to the biblical story of Joseph and long stretches (seven years) of time greater than average and less than average precipitation. This success of the FGN brought the fact that memory of a certain length can make a big difference to the attention of many. The terms "long-range dependent process" and "long memory" came into being; they can already be found in the early papers by Mandelbrot and coauthors.

5.2 The Joseph Effect and Nonstationarity

The Joseph effect is clearly visible in Figure 5.2: in the left plot, where the observations are those of fractional Gaussian noise with Hurst parameter $H = 0.8$, there are long stretches of time (hundreds of observations) during which the observations tend to be on one side of the true mean 0. This is clearly not the case on the right plot of i.i.d. normal observations. Returning momentarily to the Nile data in Figure 5.1, we see evidence of the Joseph effect there as well.

Such behavior of the observations obviously seems to indicate lack of stationarity, and in general, the relationship between long-range dependence and nonstationarity is delicate in a number of ways. We have seen that the Joseph effect involves long stretches of time when the process tends to be above the mean, and long stretches of time when the process tends to be below the mean. Quoting a description in Mandelbrot (1983), page 251, of a fractional Gaussian noise with $H > 1/2$: "Nearly every sample looks like a 'random noise' superimposed upon a background that performs several cycles, whichever the sample's duration. However, these cycles are *not* periodic, that is, *cannot* be extrapolated as the sample lengthens."

This discussion shows that in application to real data, either stationary long memory models or appropriate nonstationary models can be used in similar situations. There is, obviously, no "right" or "wrong" way to go here, beyond the principle of parsimony.

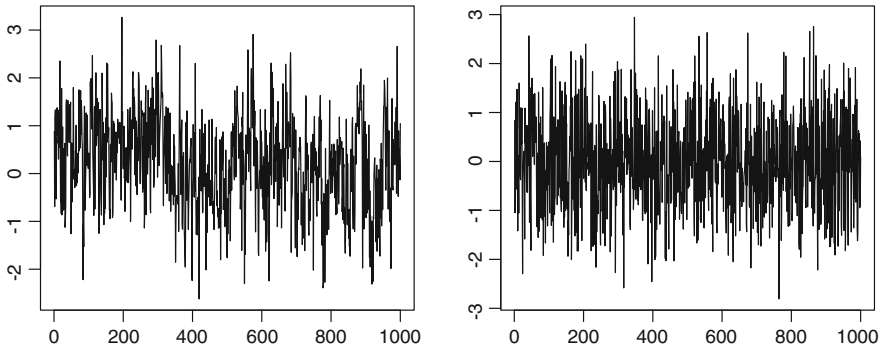

Fig. 5.2 Fractional Gaussian noise with $H = 0.8$ (left plot) and i.i.d. standard Gaussian random variables (right plot)

Among the first to demonstrate the difficulty of distinguishing between stationary long memory models and certain nonstationary models was the paper Bhattacharya et al. (1983), in which it was suggested that instead of fractional Gaussian noise or another model with long memory, the Hurst phenomenon can be explained by a simple nonstationary model as follows. Let $Y_1, Y_2, \ldots$ be a sequence of independent identically distributed random variables with a finite variance σ^2. Let $0 < \beta < 1/2$, choose $a \geq 0$, and consider the model

$$X_i = Y_i + (a+i)^{-\beta},\ i = 1, 2, \ldots . \tag{5.13}$$

Clearly, the stochastic process $X_1, X_2, \ldots$ is nonstationary, for it contains a nontrivial drift. However, it is asymptotically stationary (as the time increases), and the drift can be taken to be very small to start with (by taking a to be large). This process has no memory at all, since the sequence $Y_1, Y_2, \ldots$ is i.i.d. It does, however, cause the R/S statistic to behave in the same way as if the sequence $X_1, X_2, \ldots$ were a fractional Gaussian noise, or another long-range dependent process.

To see why this is true, note that for this model, the numerator of the R/S statistic is bounded between

$$r_n - R_n^Y \leq \max_{0\leq i\leq n}(S_i - \frac{i}{n}S_n) - \min_{0\leq i\leq n}(S_i - \frac{i}{n}S_n) \leq r_n + R_n^Y\,,$$

where

$$r_n = \max_{0\leq i\leq n}(s_i - \frac{i}{n}s_n) - \min_{0\leq i\leq n}(s_i - \frac{i}{n}s_n)\,,$$

$$R_n^Y = \max_{0\leq i\leq n}(S_i^Y - \frac{i}{n}S_n^Y) - \min_{0\leq i\leq n}(S_i^Y - \frac{i}{n}S_n^Y)\,,$$

and $S_m^Y = Y_1 + \ldots + Y_m$, $s_m = \sum_{j=1}^m (a+j)^{-\beta}$ for $m = 0, 1, 2, \ldots$.

Since s_m is a sum of a decreasing sequence of numbers, we see that $\min_{0\le i\le n}(s_i - \frac{i}{n}s_n) = 0$. On the other hand, by Theorem 10.5.6,

$$s_n \sim \frac{1}{1-\beta} n^{1-\beta} \quad \text{as } n \to \infty.$$

If we denote by i_n^* the value of i over which the maximum is achieved in $\max_{0\le i\le n}(s_i - \frac{i}{n}s_n)$, then we see that

$$i_n^* = \lfloor (s_n/n)^{-1/\beta} - a \rfloor \sim (1-\beta)^{1/\beta} n$$

as $n \to \infty$. Therefore,

$$\max_{0\le i\le n} (s_i - \frac{i}{n}s_n) = s_{i_n^*} - \frac{i_n^*}{n} s_n \sim \beta(1-\beta)^{1/\beta - 2}\, n^{1-\beta}\,,$$

so that

$$r_n \sim C_\beta\, n^{1-\beta},\ C_\beta = \beta(1-\beta)^{1/\beta-2}$$

as $n \to \infty$.

Recall that $Y_1, Y_2, \ldots$ are i.i.d. random variables with a finite variance. Therefore, the range R_n^Y of the first n observations from this sequence grows distributionally as $n^{1/2}$. We immediately conclude that

$$\frac{1}{n^{1-\beta}} \left[\max_{0\le i\le n} (S_i - \frac{i}{n}S_n) - \min_{0\le i\le n} (S_i - \frac{i}{n}S_n) \right] \to C_\beta$$

in probability as $n \to \infty$.

Similarly, for the denominator of the R/S statistic, we have a bound

$$D_n^Y - d_n \le \Big(\sum_{i=1}^n (X_i - \frac{1}{n}S_n)^2 \Big)^{1/2} \le D_n^Y + d_n\,,$$

where

$$D_n^Y = \Big(\sum_{i=1}^n (Y_i - \frac{1}{n}S_n^Y)^2 \Big)^{1/2}, \quad d_n = \Big(\sum_{i=1}^n \big((a+i)^{-\beta} - \frac{1}{n}s_n\big)^2 \Big)^{1/2}.$$

We know that $D_n^Y/n^{1/2} \to \sigma$ a.s. as $n \to \infty$, while an elementary computation using, for example, Theorem 10.5.6, leads to $d_n \sim C'_\beta\, n^{1/2-\beta}$ as $n \to \infty$ for some $0 < C'_\beta < \infty$. Therefore,

$$n^{-1/2} \Big(\sum_{i=1}^n (X_i - \frac{1}{n}S_n)^2 \Big)^{1/2} \to \sigma$$

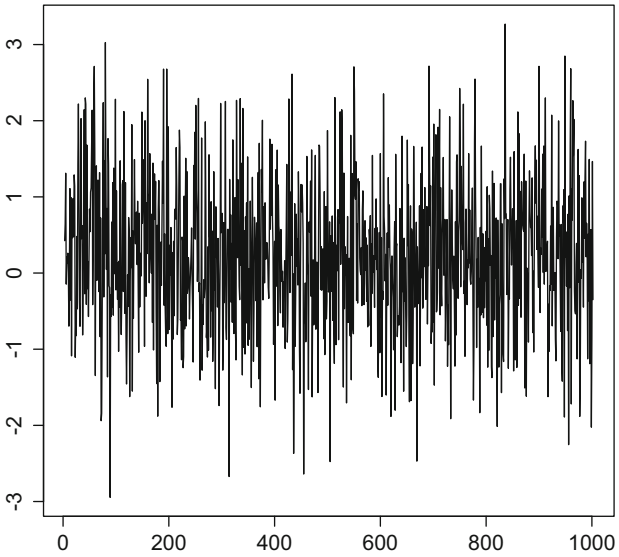

Fig. 5.3 Observations from the model (5.13) with standard normal noise, $a = 2$ and $\beta = 1/4$. No Joseph effect is visible

a.s., and we conclude that

$$\frac{1}{n^{1-\beta}}\frac{R}{S}(X_1, \ldots, X_n) \to \frac{C_\beta}{\sigma}$$

in probability as $n \to \infty$.

Therefore, for the i.i.d. model with small drift as in (5.13), the R/S statistic grows as $n^{1-\beta}$, the same rate as for the FGN with $H = 1 - \beta$, and so the R/S statistic cannot distinguish between these two models. Apart from fooling the R/S statistic, however, the model (5.13) is not difficult to tell apart from a stationary process with correlations decaying as in (5.11). Even visually, the observations from the model (5.13) do not appear to exhibit the Joseph effect, as the plot in Figure 5.3 indicates.

A very important class of nonstationary models that empirically resemble long-memory stationary models is that of *regime-switching models*. The name is descriptive, and it makes it clear where the lack of stationarity comes from. The fractional Gaussian noise also appears to exhibit different "regimes" (the Joseph effect), but the nonstationary regime-switching models are usually those with breakpoints, whose location changes with the sample size, in either a random or nonrandom manner.

One class of regime-switching models is obtained by taking a parametric model that would be stationary if its parameters were kept constant and then changing the parameters along a sequence of nonrandom time points, again chosen relative to the

sample size. Such a change can affect the mean and the variance (among many other things) of the process after breakpoints, and to many sample statistics this will look like long memory.

To see what might happen, consider a sample $X_1, \ldots, X_n$, where the observations come from $r \geq 2$ subsamples of lengths proportional to the overall sample size. That is, given fixed proportions $0 < p_i < 1$, $i = 1, \ldots, r$ with $p_1 + \ldots + p_r = 1$, the sample has the form

$$X_1^{(1)}, \ldots, X_{[np_1]}^{(1)}, X_{[np_1]+1}^{(2)}, \ldots, X_{[n(p_1+p_2)]}^{(2)}, \ldots, X_{[n(1-p_r)]+1}^{(r)}, \ldots, X_n^{(r)}, \tag{5.14}$$

where the ith subsample forms a stationary ergodic process with a finite variance, $i = 1, \ldots, r$. Since one of the common ways to try to detect long-range dependence is by looking for a slow decay of covariances and correlations, let us check the behavior of the sample covariance of the sample (5.14). Fix a time lag m and denote by $\hat{R}_m(n)$ the sample covariance at that lag based on the n observations in (5.14). Note that

$$\hat{R}_m(n) = \frac{1}{n}\sum_{j=1}^{n-m}(X_j - \bar{X})(X_{j+m} - \bar{X}) = A_m(n) + B_m(n)\,,$$

where $\bar{X} = (X_1 + \ldots + X_n)/n$ is the overall sample mean,

$$A_m(n) = \frac{1}{n}\sum_{j=1}^{n-m} X_j X_{j+m} - (\bar{X})^2\,,$$

and

$$B_m(n) = \frac{1}{n}\bar{X}\left(\sum_{j=1}^{m} X_j + \sum_{j=n-m+1}^{n} X_j\right) - \frac{m}{n}(\bar{X})^2\,.$$

By ergodicity, $\bar{X} \to \sum_{i=1}^{r} p_i\mu_i$, where μ_i is the mean of the ith subsample, $i = 1, \ldots, r$. Further, since m is fixed, $B_m(n) \to 0$ in probability as $n \to \infty$.

Finally, if I_i denotes the set of indices within $\{1, \ldots, n\}$ corresponding to the ith subsample, $i = 1, \ldots, r$, then by ergodicity,

$$\frac{1}{n}\sum_{j=1}^{n-m} X_j X_{j+m}$$

$$= \sum_{i=1}^{r} \frac{\mathrm{Card}(I_i \cap (I_i - m))}{n} \frac{1}{\mathrm{Card}(I_i \cap (I_i - m))} \sum_{j \in I_i \cap (I_i - m)} X_j^{(i)} X_{j+m}^{(i)}$$

$$+\frac{1}{n}\sum_{i=1}^{r}\sum_{\substack{j\in\{1,\ldots,n-m\}\\ j\in I_i, j+m\in I_{i+1}}} X_j X_{j+m}$$

$$\to \sum_{i=1}^{r} p_i E\left(X_1^{(i)} X_{1+m}^{(i)}\right) = \sum_{i=1}^{r} p_i (R_m^{(i)} + \mu_i^2)\,,$$

where $R_m^{(i)}$ is the covariance at lag m of the ith subsample. We conclude that

$$\hat{R}_m(n) \to \sum_{i=1}^{r} p_i \left(R_m^{(i)} + \mu_i^2\right) - \left(\sum_{i=1}^{r} p_i \mu_i\right)^2 \tag{5.15}$$

$$= \sum_{i=1}^{r} p_i R_m^{(i)} + \frac{1}{2}\sum_{i_1=1}^{r}\sum_{i_2=1}^{r} p_{i_1} p_{i_2} \left(\mu_{i_1} - \mu_{i_2}\right)^2$$

in probability as $n \to \infty$. What (5.15) indicates is that if there is regime-switching as we have described, and (some of) the mean values in different regimes are different, then the sample covariance function will tend to stabilize, at large, but fixed, lags at a positive value.

This is what is often observed in practice, and long memory is suspected. Of course, this regime-switching model is simply a deterministic way of mimicking the Joseph effect (recall Figure 5.2), and an example of this phenomenon can be seen in Figure 5.4, where $r = 4$, $p_1 = p_2 = p_3 = p_4 = 1/4$, and the four different stationary ergodic processes are all autoregressive processes of order 1, with normal innovations with the mean and the standard deviation both equal to 1. The autoregressive coefficients are $\psi_1 = 0.7$, $\psi_2 = 0.75$, $\psi_3 = 0.65$, $\psi_4 = 0.8$.

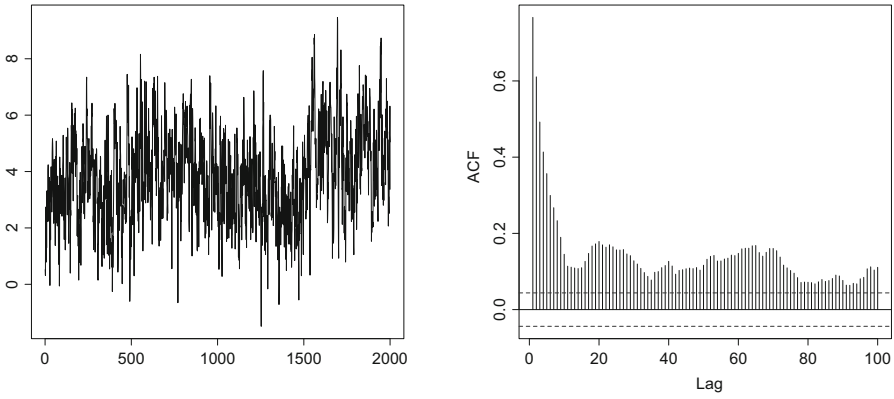

Fig. 5.4 Observations from a regime-switching AR(1) model (left plot) and their sample autocorrelation function (right plot)

5.3 Long Memory, Mixing, and Strong Mixing

The notion of memory in a stationary stochastic process is by definition related to the connections between certain observations and those occurring after an amount of time has passed. If $X_1, X_2, \ldots$ is the process, then the passage of time corresponds to the shifted process, $X_{k+1}, X_{k+2}, \ldots$, for a time shift k. In other words, the notion of memory is related to the connections between the process and its shifts. This makes the language of the ergodic theory of stationary processes, elements of which are outlined in Chapter 2, an attractive language for describing the memory of a stationary process.

We begin by observing that it is very natural to say that a nonergodic stationary process $\mathbf{X}$ *has infinite memory*. Indeed, a nonergodic process has the structure given in Proposition 2.1.6. That is, it is a mixture of the type

$$\big(X_n,\, n \in \mathbb{Z}\big) \stackrel{d}{=} \begin{cases} \big(Y_n,\, n \in \mathbb{Z}\big) \text{ with probability } p, \\ \big(Z_n,\, n \in \mathbb{Z}\big) \text{ with probability } 1-p, \end{cases}$$

where stationary processes $\big(Y_n,\, n \in \mathbb{Z}\big)$ and $\big(Z_n,\, n \in \mathbb{Z}\big)$ have different finite-dimensional distributions, and the choice with probability $0 < p < 1$ is made independently of the two stationary processes. This means that the result of a single "coin toss" (with probabilities p and $1-p$) will be "remembered forever." Therefore, it certainly makes sense to call stationary ergodic processes "processes with finite memory," and stationary nonergodic processes "processes with infinite memory."

It is very tempting to try to use another ergodic theoretical notion, stronger than ergodicity, such as weak mixing or mixing, for example, to define finite and short memory in a stationary process. Then ergodic stationary processes that lack this stronger property will be naturally called processes with long memory. If the property of mixing were used for this purpose, for example, then a long-range dependent process would be an ergodic but nonmixing process.

Such definitions of long-range dependence are possible, but they have not become standard, for reasons that will be discussed below. Before we do that, however, it is important to note that the approaches to memory of a stationary process via the ergodic theoretical properties of the corresponding shift transformation are very attractive from the following point of view. Let $\mathbf{X}$ be a stationary process, and let the process $\mathbf{Y}$ be derived from the process $\mathbf{X}$ by means of a point transformation $Y_n = g(X_n)$ for all n, where $g : \mathbb{R} \to \mathbb{R}$ is a measurable function. Clearly, $\mathbf{Y}$ is also a stationary process. It is intuitively clear that the process $\mathbf{X}$ "remembers at least as much" as the process $\mathbf{Y}$ does. If, in particular, g is a one-to-one map, and g^{-1} is also measurable, then this intuition says that the processes $\mathbf{X}$ and $\mathbf{Y}$ should have the "same length of memory": if one of them has long memory, then so should the other one.

This, apparently very natural, requirement has proved to be difficult to satisfy by many of the proposed definitions of long-range dependence. It is, however, automatic with ergodic theoretically based definitions. Indeed, it follows from

Corollary 2.2.5 and Proposition 2.2.14 that if $\mathbf{X}$ is mixing (respectively weak mixing), then the process $\mathbf{Y}$ with $Y_n = g(X_n)$ for all n is also mixing (respectively weak mixing). This would imply that short memory was preserved under a measurable map, and if the map is one-to-one, with a measurable inverse, then the map must preserve long memory as well.

It is instructive to record what the ergodic theoretically based notions of memory discussed above mean for stationary Gaussian processes. Let $\mathbf{X}$ be a (real-valued) stationary Gaussian process with covariance function R_k, $k \geq 0$ and spectral measure F on $(-\pi, \pi]$. That is, $R_k = \int_{(-\pi,\pi]} \cos(kx)\, F(dx)$ for $k \geq 0$. Then

- the process $\mathbf{X}$ is ergodic if and only if the spectral measure F is atomless;
- the process $\mathbf{X}$ is mixing if and only if $R_k \to 0$ as $k \to \infty$;

see Examples 2.2.8 and 2.2.18. The requirement that the covariance function vanish as the time lag increases, however, proved to be insufficient in dealing with long memory for Gaussian processes. Indeed, many "unusual" phenomena have been observed for Gaussian processes whose covariance functions vanish in the limit, but sufficiently slowly, as we have already seen in the example of fractional Gaussian noise. Therefore, the mixing property is not believed to be sufficiently strong to say that a stationary process with this property has short memory. A stronger requirement is needed.

For this purpose, strong mixing conditions, some of which are discussed in Section 2.3, have been used. A possible connection between strong mixing properties and lack of long memory (i.e., presence of short memory) has been observed, beginning with Rosenblatt (1956). We discuss results in this spirit in Comments to Chapter 9. Such results explain why the absence of one or another strong mixing condition (as opposed to ergodic-theoretical mixing) is sometimes taken as the definition of long-range dependence.

The strong mixing properties share with the ergodic-theoretical notions of ergodicity and mixing the following very desirable feature: if a process $\mathbf{Y}$ is derived from a process $\mathbf{X}$ by means of a one-to-one point transformation $Y_n = g(X_n)$ for all n, where $g : \mathbb{R} \to \mathbb{R}$ is a one-to-one function such that both g and g^{-1} are measurable, then the process $\mathbf{X}$ has long memory in the sense of lacking one of the strong mixing properties if and only if the process $\mathbf{Y}$ does; see Exercise 2.6.11.

The role that the strong mixing conditions play in eliminating the possibility of long-range dependence is real, but limited. Its effects are felt more in the behavior of the partial sums of a process than in, say, the behavior of the partial maxima.

Overall, the strong mixing conditions have not become standard definitions of absence of long-range dependence, i.e., of short memory. To some extent, this is due to the fact that the effect of strong mixing conditions is limited. More importantly, the strong mixing conditions are not easily related to the natural building blocks of many stochastic models and are difficult to verify, with the possible exception of Gaussian processes and Markov chains. Even in the latter cases, necessary and sufficient conditions are not always available, particularly for more complicated types of strong mixing.

5.4 Comments on Chapter 5

Comments on Section 5.2

The fact that the i.i.d. model with small drift as in (5.13) can be easily distinguished from the fractional Gaussian noise with $H > 1/2$ was shown in Künsch (1986) using the *periodogram*.

In Mikosch and Stărică (2016) and Mikosch and Stărică (2000a), the procedure of changing the parameters of an otherwise stationary model was applied to the short-memory GARCH(p, q) model, resulting in behavior resembling long-range dependence.

Various other regime-switching models mimicking long-range dependence are suggested in Diebold and Inoue (2001).

5.5 Exercises to Chapter 5

Exercise 5.5.1. *Show that the function g_ϵ in Example 5.1.2 is a function on $M_+^{\mathrm{R}}\left([0,1] \times \overline{\mathbb{R}}_0^d\right)$ that is continuous at all points at which the denominator in its definition does not vanish.*

Exercise 5.5.2. *In this exercise, we will check the validity of the statement (5.8). Write*

$$\frac{1}{\sqrt{n}} \frac{R}{S}(X_1, \ldots, X_n) = \frac{M_n - m_n}{D_n}, \quad RS_n(\epsilon) = \frac{M_n(\epsilon) - m_n(\epsilon)}{D_n(\epsilon)} .$$

(i) Use the maximal inequality in Theorem 10.7.4 to show that for some finite positive constant c,

$$P\left(\frac{1}{a_n} |M_n - M_n(\epsilon)| > \lambda\right) \le \frac{c}{\lambda} \frac{n^{1/2}}{a_n} \left(E\left(X_1^2 \mathbf{1}(|X_1| \le \epsilon a_n)\right)\right)^{1/2}$$

for each $\lambda > 0$ and $n = 1, 2, \ldots$. Next, use the estimate on the moments of truncated random variables in Proposition 4.2.3 and the fact that $0 < \alpha < 2$ to show that

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} P\left(\frac{1}{a_n} |M_n - M_n(\epsilon)| > \lambda\right) = 0$$

for every $\lambda > 0$. A similar argument proves that

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} P\left(\frac{1}{a_n} |m_n - m_n(\epsilon)| > \lambda\right) = 0$$

for every $\lambda > 0$.

(ii) Check that

$$|D_n - D_n(\epsilon)| \le \frac{|S_n|}{n^{1/2}} + \left(\sum_{i=1}^{n} X_i^2 \mathbf{1}\big(|X_i| \le \epsilon a_n\big) \right)^{1/2}$$

and conclude that for every $\lambda > 0$,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} P\left(\frac{1}{a_n} |D_n - D_n(\epsilon)| > \lambda \right) = 0 \,.$$

(iii) Use the truncation argument used in Example 5.1.2 and the already checked part of (5.8) to conclude that $a_n^{-1}(M_n - m_n)$ *converges weakly to the numerator of (5.7) and so the corresponding sequence of laws is tight.*

(iv) Show that

$$\lim_{M \to \infty} \limsup_{n \to \infty} \sup_{0 < \epsilon < 1} P\left(\frac{a_n}{D_n(\epsilon)} > M \right) = 0$$

and

$$\lim_{M \to \infty} \limsup_{n \to \infty} \sup_{0 < \epsilon < 1} P\left(\frac{a_n^2}{D_n D_n(\epsilon)} > M \right) = 0 \,.$$

(v) Put together the previous parts of the exercise to obtain (5.8).