LONG MEMORY MODELS

Peter M. Robinson

Department of Economics, London School of Economics

Summary

Long memory models are statistical models that describe strong correlation or dependence across time series data.   This kind of phenomenon is often referred to as "long memory" or "long-range dependence".   It refers to persisting correlation between distant observations in a time series.   For scalar time series observed at equal intervals of time that are covariance stationary, so that the mean, variance, and autocovariances (between observations separated by a lag j) do not vary over time, it typically implies that the autocovariances decay so slowly, as j increases, as not to be absolutely summable. However, it can also refer to certain nonstationary time series, including ones with an autoregressive unit root, which exhibit even stronger

correlation at long lags. Evidence of long memory has often been been found in economic and financial time series, where the noted extension to possible nonstationarity can cover many macroeconomic time series, as well as in such fields as astronomy, agriculture, geophysics and chemistry.

As long memory is now a technically well developed topic formal definitions are needed, and we defer these to the paper itself. But by way of partial motivation, long memory models can be thought of as complementary to the very well known and widely applied stationary and invertible autoregressive and moving average (ARMA) models, whose autocovariances are not only summable but decay exponentially fast as a function of lag j. Such models are often referred to as "short memory" models, becuse there is negligible correlation across distant time intervals. However these models are often combined with the most basic long memory ones since together they offer the ability to describe both short and long memory feartures in many time series.

Keywords: long memory; parametric models; semiparametric models; volatility models; nonstationary models

Introductory Definitions and Discussion

We begin by introducing some basic notation. Let x_t, $t = 0, \pm 1, ...,$ be an equally spaced, real valued time series. We suppose initially that x_t is covariance stationary, so that the mean

$$\mu = E(x_t)$$

and lag-j autocovariances (or variance when $j = 0$)

$$\gamma(j) = Cov(x_t, x_{t+j})$$

do not depend on t. We further suppose that x_t has a spectral density, denoted

$$f(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j)e^{-ij\lambda}, \quad -\pi \leq \lambda \leq \pi,$$

where λ denotes "frequency". Note that $f(\lambda)$ is a non-negative, even function. We might then say that x_t has "long memory" if

$$f(0) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) = \infty,$$

so that $f(\lambda)$ diverges at frequency zero. The extreme alternative that

$$f(0) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) = 0$$

is on the other hand possible; this phenomenon is sometimes referred to as "negative dependence" or "anti-persistence". The intermediate situation is

$$0 < f(0) < \infty,$$

when we say that x_t has "short memory". It is also possible that $f(\lambda)$ might diverge or be zero at one or more frequencies λ in $(0, \pi]$, possibly indicating seasonal or cyclic behaviour. We shall discuss the modelling of such phenomena, but mainly focus on behaviour at zero frequency, which empirically seems the most interesting. An excellent textbook reference to theory and methods for long memory is Giraitis, Koul and Surgailis (2012).

Nonparametric estimates of $f(\lambda)$ have been found to be heavily peaked around zero frequency in case of many economic time series, going back to Adelman (1965), lending support for the presence of long memory. Moreover empirical evidence of long memory in various fields, such as astronomy, chemistry, agriculture and geophysics, dates from much earlier times, see for example Fairfield Smith (1938), Hurst (1951) .

One feature of interest in early work was behaviour of the sample mean,

$$\bar{x} = n^{-1} \sum_{t=1}^{n} x_t.$$

If $f(\lambda)$ is continuous and positive at $\lambda = 0$,

$$Var(\bar{x}) = \frac{1}{n} \sum_{j=1-n}^{n-1} \left(1 - \frac{|j|}{n}\right) \gamma(j) \sim \frac{2\pi f(0)}{n}, \quad \text{as } n \to \infty,$$

3

But, for example, Fairfield Smith (1938) fitted a law $n^{-\alpha}$, $0 < \alpha < 1$ to spatial agricultural data, disputing the above n^{-1} law. At this point it is convenient to switch notation, to $d = (1 - \alpha)/2$, because d, referred to as the "differencing" parameter, features more commonly in econometric modelling. Fairfield Smith's (1938) law for the variance of the sample mean is thus n^{2d-1}, which from (1.5) arises if

$$\gamma(j) \sim c_1 j^{2d-1}, \quad \text{as } j \to \infty, \tag{1.1}$$

for $c_1 > 0$. Under additional conditions (see Yong, 1974), (1.1) is equivalent to a corresponding power law for $f(\lambda)$ near zero frequency,

$$f(\lambda) \sim c_2 |\lambda|^{-2d}, \quad \text{as } \lambda \to 0, \tag{1.2}$$

for $c_2 > 0$. The behaviour of the sample mean under such circumstances, and the form and behaviour of the best linear unbiased estimate of the population mean, was discussed by Adenstedt (1974). He anticipated the practical usefulness of (1.2) in the long memory range $0 < d < \frac{1}{2}$, but also treated the anti-persistent case $-\frac{1}{2} < d < 0$. The sample mean tends to be highly statistically inefficient under anti-persistence, but for long memory Samarov and Taqqu (1988) found it to have remarkably good efficiency.

A number of explanations of how long memory behaviour might arise have been proposed. Macroeconomic time series, in particular, can be thought of as aggregating across micro-units. Consider the random-parameter autoregressive model of order 1 $(AR(1))$,

$$X_t(\alpha) = A(\alpha)X_{t-1}(\alpha) + \varepsilon_t(\alpha),$$

where α indexes micro-units, the $\varepsilon_t(\alpha)$ are independent and homoscedastic with zero mean across α, t, and $A(\alpha)$ is a random variable with support $(-1, 1)$ or $[0, 1)$. Then, conditional on α, $X_t(\alpha)$ is a stationary $AR(1)$ sequence. Robinson (1978a) showed

4

that the "unconditional autovariance" which we again denote by $\gamma(j)$, is given by

$$\gamma(j) = Cov\left\{X_t(\alpha), X_{t+j}(\alpha)\right\} = \sum_{u=0}^{\infty} E\left\{A(\alpha)^{j+2u}\right\}, \tag{1.3}$$

and that the "unconditional spectrum" $f(\lambda)$ at $\lambda = 0$ is proportional to $E\left\{(1 - A(\alpha))^{-2}\right\}$, and thus infinite, if $A(\alpha)$ has a probability density with a zero at 1 of order less than or equal to 1. One class with this property considered by Robinson (1978a) was the (possibly translated) Beta distribution, for which Granger (1980) explicitly derived the corresponding power law behaviour of the spectral density of cross-sectional aggregates $x_t = N^{-\frac{1}{2}} \sum_{i=1}^{N} X_t(\alpha_i)$, where the α_i are independent drawings: clearly $Cov(x_t, x_{t+j})$ is $\gamma(j)$ due to the independence properties. Indeed, if $A(\alpha)$ has a Beta $(c, 2 - 2d)$ distribution on $(0, 1)$, for $c > 0$, $0 < d < \frac{1}{2}$, $E\left\{A(\alpha)^k\right\}$ decays like k^{2d-2}, so (1.3) decays like j^{2d-1}, as in (1.1). Intuitively, a sufficient density of individuals with close-to-unit-root behaviour produces the aggregate long memory. For further developments, in relation to more general models see e.g. Lippi and Zaffaroni (1997).

Parametric Models

The differencing parameter, d, introduced in the previous section, concisely describes long memory properties, and so much interest in the possibility of long memory or anti-persistence focusses on the question of its value. In practice d is typically regarded as unknown, and so its estimation has been the focus of much research. Indeed, as discussion of the previous section indicates an estimate of d is useful even in estimating the variance of the sample mean.

In order to estimate d we need to consider the modelling of dependence in more detail. The simplest possible realistic model for a covariance stationary series is a parametric one that expresses $\gamma(j)$ for all j, or $f(\lambda)$ for all λ, as a parametric function of just two parameters, d and an unknown scale factor. The earliest such model is "fractional noise", which arises from considerations of self-similarity. A

5

stochastic process $\{y(t); -\infty < t < \infty\}$ is self-similar with "self-similarity parameter" $H \in (0,1)$ if, for any $a > 0$, $\{y(at); -\infty < t < \infty\}$ has the same distribution as $\{a^H y(t); -\infty < t < \infty\}$. If the differences $x = y(t) - y(t-1)$, for integer t, are covariance stationary, we obtain

$$\gamma(j) = \frac{\gamma(0)}{2} \left\{ |j+1|^{2H} - 2|j|^{2H} + |j-1|^{2H} \right\}.$$

This decays like j^{2H-2} as $j \to \infty$, so on taking $H = d + \frac{1}{2}$ we have again the asymptotic law (1.1); $\gamma(0)$ is the unknown scale parameter in this model.

This model was studied by Mandelbrot and Van Ness (1968) and others, but, it extends less naturally to richer stationary series, and nonstationary series, and has an unpleasant spectral form (see the later discussion of Whittle estimates), so it has received less attention in recent years than another two-parameter model, the "fractional differencing" model proposed by Adenstedt (1974):

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| 1 - e^{i\lambda} \right|^{-2d}, \quad -\pi \leq \lambda \leq \pi. \tag{2.1}$$

When $d = 0$, this is just the spectral density of a white noise series (with variance σ^2), while for $0 < |d| < \frac{1}{2}$ both properties (1.1) and (1.2) hold, Adenstedt (1974) giving a formula for $\gamma(j)$ as well as other properties. Note that $d < \frac{1}{2}$ is necessary for integrability of $f(\lambda)$, that is for x_t to have finite variance; this restriction is sometimes called the stationarity condition on d. Another mathematically important restriction is that of invertibility, $d > -\frac{1}{2}$. We shall discuss estimation of models such as (2.1).

The "typical spectral shape of an economic variable" was identified by Granger (1966) as not only entailing spectral divergence at zero frequency, but monotonic decay with frequency. Both "fractional differencing" and "fractional noise" models have this simple property. But even if monotonicity holds, as it may, at least approximately, in case of deseasonalized series, the notion that the entire autocorrelation structure can be explained by a single parameter, d, is highly questionable. Though

6

d determines the long-run or low-frequency behaviour of $f(\lambda)$, greater flexibility in modelling short-run, high-frequency, behaviour may be desired. The model (2.1) was referred to as "fractional differencing" because it is the spectral density of x_t generated by

$$(1 - L)^d x_t = e_t, \tag{2.2}$$

where $\{e_t\}$ is a sequence of uncorrelated variables with zero mean and variance σ^2, L is the lag operator, $Lx_t = x_{t-1}$ and

$$(1 - L)^d = \sum_{j=0}^{\infty} \frac{\Gamma(j - d)}{\Gamma(-d)\Gamma(j + 1)} L^j,$$

where $\Gamma(.)$ denotes the gamma function. With $d = 1$ (and a suitable initial condition), (2.2) would describe a random walk model. The model

$$(1 - L)^d a(L) x_t = b(L) e_t. \tag{2.3}$$

was stressed by Box and Jenkins (1971), d here being a positive integer, $a(L)$ and $b(L)$ being the polynomials

$$a(L) = 1 - \sum_{j=1}^{p} a_j L^j, \quad b(L) = 1 + \sum_{j=1}^{q} b_j L^j,$$

all of whose zeros are outside the unit circle, with $a(L)$ and $b(L)$ having no zero in common to ensure identifiability of the autoregressive (AR) order p and the moving average (MA) order q. Granger and Joyeux (1980) considered instead fractional $d \in (-\frac{1}{2}, \frac{1}{2})$ in (2.3), giving a fractional autoregressive integrated moving average model of orders p, d, q (often abbreviated as $FARIMA(p, d, q)$ or $ARFIMA(p, d, q)$). It has spectral density

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \left|\frac{b(e^{i\lambda})}{a(e^{i\lambda})}\right|^2, \quad -\pi \leq \lambda \leq \pi. \tag{2.4}$$

Granger and Joyeux (1980) principally discussed the simple $FARIMA(0, d, 0)$ case (2.1) of Adenstedt (1974), but they also considered estimation of d, prediction, and

simulation of long memory series. Further discussion of $FARIMA(p, d, q)$ models was provided by Hosking (1981), much of it based on Adenstedt's (1974) model (2.1), but he also gave results for the general case (2.4), especially the $FARIMA(1, d, 0)$.

An enduringly popular proposal for estimating d, or H, used the adjusted rescaled range (R/S) statistic

$$R/S = \frac{\max_{1 \leq j \leq n} \sum_{t=1}^{j} (x_t - \bar{x}) - \min_{1 \leq j \leq n} \sum_{t=1}^{j} (x_t - \bar{x})}{\left\{ \frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2 \right\}^{\frac{1}{2}}}$$

of Hurst (1951), Mandelbrot and Wallis (1969). Large sample statistical properties of the R/S statistic were studied by Mandelbrot and Taqqu (1979), Taqqu (1979), and it was considered in an economic context by Mandelbrot (1972). But its limit distribution is nonstandard and difficult to use in statistical inference, while it has no known optimal efficiency properties with respect to any known family of distributions.

Despite the distinctive features of long memory series, there is no over-riding reason why traditional approaches to parametric estimation in time series should be abandoned in favour of rather special approaches like R/S. In fact, if x_t is assumed Gaussian, the Gaussian maximum likelihood estimate (MLE) might be expected to have optimal asymptotic statistical properties, and unlile R/S, can be tailored to the particular parametric model assumed.

The literature on the Gaussian MLE developed first with short memory processes in mind (see e.g. Whittle, 1951, Hannan, 1973). One important finding was that the Gaussian likelihood can be replaced by various approximations without affecting first order limit distributional behaviour. Under suitable conditions, estimates maximizing such approximations, called "Whittle estimates" are all $\sqrt{n}$-consistent and have the same limit normal distribution as the Gaussian MLE.

One particular Whittle estimate which seems particularly computationally advantageous is the discrete-frequency form. Suppose the parametric spectral density has

form $f(\lambda; \theta, \sigma^2) = (\sigma^2/2\pi)h(\lambda; \theta)$, where θ is an r-dimensional unknown parameter vector and σ^2 is a scalar as in (2.1). If σ^2 is regarded as varying freely from θ, and $\int_{-\pi}^{\pi} \log h(\lambda; \theta)d\lambda = 0$ for all admissible values of θ, then we have what might be called a "standard parameterization". For example, we have a standard parameterization in (2.1) with $\theta = d$, and in (2.4) with θ determining the a_j, $1 \le j \le p$ and b_j, $1 \le j \le q$. Define also the periodogram

$$I(\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} x_t e^{it\lambda} \right|^2$$

and the Fourier frequencies $\lambda_j = 2\pi jn$. Denoting by θ_0 the true value of θ, then the discrete frequency Whittle estimate of θ_0 minimizes the following approximation to a constant minus the Gaussian log likelihood,

$$\sum_{j=1}^{n-1} \frac{I(\lambda_j)}{h(\lambda_j; \theta)}. \tag{2.5}$$

Hannan (1973) stressed this estimate. It has the advantages of using directly the form of h, which is readily written down in case of autoregressive moving average (ARMA) models, Bloomfeld's (1972) spectral model, and others; on the other hand, autocovariances, partial autocovariances, AR coefficients and MA coefficients, which variously occur in other types of Whittle estimate, tend to be more complicated except in special cases, indeed for (2.4) the form of autocovariances, for example, can depend on the question of multiplicity of zeros of $a(L)$. Another advantage of (2.5) is that it makes direct use of the fast Fourier transform, which enables the periodograms $I(\lambda_j)$ to be rapidly computed even when n is very large. A third advantage is that mean-correction of x_t is dealt with simply by omission of the frequency $\lambda_0 = 0$.

A notable feature of Whittle estimates of θ_0, first established in case of short memory series, is that while they are only asymptotically efficient when x_t is Gaussian, their limit distribution (in case of "standard parameterizations") is unchanged by many departures from Gaussianity. Thus the same, relatively convenient, statistical methods (hypothesis testing, interval estimation) can be used without worrying

9

too much about the question of Gaussianity. Hannan established asymptotic statistical properties for several Whittle forms in case x_t has a linear representation in homoscedastic stationary martingale differences having finite variance.

It is worth noting that Hannan (1973) established first consistency under only ergodicity of x_t, so that long memory was actually included here. However, for his asymptotic normality result, with $\sqrt{n}$-convergence, which is crucial for developing statistical inference, his conditions excluded long memory, and clearly (2.5) appears easier to handle technically in the presence of a smooth h than of one with a singularity. Robinson (1978b) developed extensions to cover "nonstandard parameterizations" his treatment hinting at how a modest degree of long memory might be covered. He reduced the problem to a central limit theorem for finitely many sample autocovariances, whose asymptotic normality had been shown by Hannan (1976) to rest crucially on square integrability of the spectral density; note that (2.1) and (2.4) are square integrable only for $d < \frac{1}{4}$. In fact for some forms of Whittle estimate, Yajima (1985) established the central limit theorem, again with $\sqrt{n}$-rate, in case of model (2.1) with $0 < d < \frac{1}{4}$.

Fox and Taqqu (1986) provided the major breakthrough in justifying Whittle estimation in long memory models. Their objective function was not (2.5) but the continuous frequency form

$$\int_{-\pi}^{\pi} \frac{I(\lambda)}{h(\lambda; \theta)} d\lambda, \tag{2.6}$$

but their basic insight applies to (2.5) also. Because the periodogram $I(\lambda)$ is an asymptotically unbiased estimate of the spectral density only at continuity points it can be expected to "blow up" as $\lambda \to 0$. However, since $h(\lambda; \theta)$ also "blows up" as $\lambda \to 0$ and appears in the denominator, some "compensation" can be expected. Actually, limiting distributional behaviour depends on the "score" (the derivative in θ of (2.6) or (2.5)) at θ_0 being asymptotically normal; Fox and Taqqu (1987) gave general conditions for such quadratic forms to be asymptotically normal, which then

apply to Whittle estimates with long memory such that $0 < d < \frac{1}{2}$.

Gaussianity of x_t was assumed by Fox and Taqqu (1986), and by Dahlhaus (1989), who considered the actual Gaussian MLE and discrete-frequency Whittle estimate, and established asymptotic efficiency. For (2.6) Giraitis and Surgailis (1990) relaxed Gaussianity to a linear process in independent and identically distributed (iid) innovations, thus providing a partial extension of Hannan's (1973) work to long memory. The bulk of this asymptotic theory has not directly concerned the discrete frequency form (2.5), and has focussed mainly on the continuous frequency form (2.6), though the former benefits from the neat form of the spectral density in case of the popular $FARIMA(p, d, q)$ class (2.4); on evaluating the integral in (2.6), we have a quadratic form involving the Fourier coefficients of $h(\lambda; \theta)^{-1}$, which are generally rather complicated for long memory models. Also, in (2.6) and the Gaussian MLE, correction for an unknown mean must be explicitly carried out, not dealt with merely by dropping zero frequency.

Other estimates have been considered. While Whittle estimation of the models (2.2) and (2.3) requires numerical optimization, Kashyap and Eom (1988) proposed a closed-form estimate of d in (2.2) by a log periodogram regression (across λ_j, $j = 1, ..., n-1$). This idea does not extend nicely to $FARIMA(p, d, q)$ models (2.3) with $p > 0$ or $q > 0$, but it does to

$$f(\lambda) = \frac{1}{2\pi} \left|1 - e^{i\lambda}\right|^{-2d} \exp\left\{\sum_{k=1}^{p-1} \beta_k \cos((k-1)\lambda)\right\}, \quad -\pi \le \lambda \le \pi \qquad (2.7)$$

(see Robinson (1994a)) which combines (2.2) with Bloomfeld's (1972) short memory exponential model; Moulines and Soulier (1999) provided asymptotic theory for log peridogram regression estimation of (2.7). They assumed Gaussianity, which for technical reasons is harder to avoid when a nonlinear function of the periodogram, such as the log, is involved, than in Whittle estimation, despite this being originally motivated by Gaussianity. Whittle estimation is also feasible with (2.7), indeed Robinson

(1994a) noted that it can be reparameterized as

$$f(\lambda) = \frac{\exp}{2\pi}\left\{\sum_{k=1}^{p-1}\theta_k\cos\{(k-1)\lambda\} - 2d\sum_{k=p-1}^{\infty}\frac{\cos(k\lambda)}{k}\right\},$$

taking $\theta_1 = \beta_1$, $\theta_k = \beta_k - 2/(k-1)$, $2 \le k \le p-1$, from which it can be deduced that the limiting covariance matrix of Whittle estimates is desirably diagonal.

In econometrics generalized method of moments (GMM) has beeen proposed for estimating many models, including long memory models. But GMM objective functions seem in general to be less computationally attractive than (2.5), require stronger regularity conditions in asymptotic theory, and do not deal so nicely with an unknown mean. Also, unless a suitable weighting is employed they will be less efficient than Whittle estimates in the Gaussian case, have a relatively cumbersome limiting covariance matrix, and are not even asymptotically normal under $d > \frac{1}{4}$. But note that $\sqrt{n}$-consistency and asymptotic normality of Whittle estimates cannot even be taken for granted, having been shown not to hold over some or all of the range $d \in (0, \frac{1}{2})$ for certain nonlinear functions x_t of a underlying Gaussian long memory process (see e.g Giraitis and Taqqu (1999)).

Even assuming Gaussianity of x_t, nonstandard limit distributional behaviour for Whittle estimates can arise in certain models. As observed in the previous section, a spectral pole (or zero) could arise at a non-zero frequency, to explain a form of cyclic behaviour. Gray, Zhang and Woodward (1989) proposed the "Gegenbauer" model

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|1 - 2e^{i\lambda}\cos\omega + e^{2i\lambda}\right|^{-2d}\left|\frac{b(e^{i\lambda})}{a(e^{i\lambda})}\right|^2, \quad -\pi \le \lambda \le \pi, \quad\quad (2.8)$$

for $\omega \in (0, \pi]$. To compare with (2.1), $f(\lambda)$ diverges at frequency ω if $d > 0$. When ω is known our previous discussion of estimation and asymptotic theory applies. If ω is unknown, then Whittle procedures can be adapted, but it seems that such estimates of ω (but not of the other parameters) will be n-consistent with a nonstandard limit distribution. Giraitis, Hidalgo and Robinson (2001) established n-consistency for an estimate of ω that, after being suitably standardized, cannot converge in distribution.

12

Semiparametric Models

"Semiparametric" models for long memory retain the differencing parameter d but treat the short memory component nonparametrically. Correct specification of p and q is very important in parametric $FARIMA(p, d, q)$ models. In particular, under-specification of p or q leads to inconsistent estimation of AR and MA coefficients, but also of d, as does over-specification of both, due to a loss of identifiability. Procedures of order-determination developed for short memory models, such as AIC, have been adapted to FARIMA models but there is no guarantee that the underlying model belongs to the finite-parameter class proposed. That an attempt to seriously model short-run features can lead to inconsistent estimation of long-run properties seems very unfortunate, especially if the latter happen to be the aspect of most interest.

Short-run modelling is seen from (1.1) and (1.2) to be almost irrelevant at very low frequencies and very long lags, where d dominates. This suggests that estimates of d can be based on information arising from only one or other of these domains, and that such estimates should have validity across a wide range of short memory behaviour. Because this robustness requires estimates to essentially be based on only a vanishingly small fraction of the data as sample size increases, one expects slower rates of convergence than for estimates based on a correct finite-parameter model. But in very long series, such as arise in finance, the degrees of freedom available may be sufficient to provide adequate precision. These estimates are usually referred to as "semiparametric", though their slow convergence rates make them more akin to "nonparametric" estimates in other areas of statistics, indeed some are closely related to the smoothed nonparametric spectrum estimates familiar from short memory time series analysis.

It is worth stressing that not just point estimation of d is of interest, but also interval estimation and hypothesis testing. Probably the test of most interest to practitioners is a test of long memory, or rather, a test of short memory $d = 0$ against long memory

13

alternatives $d > 0$, or anti-persistent alternatives $d < 0$, or both, $d \neq 0$. For this we need a statistic with a distribution that can be satisfactorily approximated, and computed, under $d = 0$, and that has good power. In a parametric setting, tests of $d = 0$ - perhaps of Wald, Lagrange multiplier or likelihood-ratio type - can be based on Whittle functions such as (2.5) and the $FARIMA(p, d, q)$ family. Actually, much of the limit distribution theory for Whittle estimation primarily concerned with stationary long memory, $0 < d < \frac{1}{2}$, does not cover $d = 0$, or $d < 0$, but other earlier short memory theory, such as Hannan's (1973), can provide null limit theory for testing $d = 0$. Because the test statistic is based on assumed p and q, the null limit distribution developed on this basis is generally invalid if p and q are misspecified, as discussed earlier; this can lead, for example, to mistaking unaccounted-for short memory behaviour for long memory, and rejecting the null too often. The invalidity of tests for $d = 0$ for the R/S statistic introduced in the previous section in the presence of unanticipated short memory autocorrelation was observed by Lo (1991), who proposed a corrected statistic (using smoothed nonparametric spectral estimation at frequency zero) and developed its limit distribution under $d = 0$ in the presence of a wide range of short memory dependence (described by mixing conditions), and tested stock returns for long memory.

The null limit theory of Lo's (1991) modified R/S statistic is nonstandard. Any number of possible statistics has sensitivity to long memory. Of these, some have the character of "method-of-moments" estimates, minimizing a "distance" between population and sample properties. Robinson (1994b) proposed an "averaged periodogram" estimate of d, employing what would be a consistent estimate of $f(0)$ under $d = 0$, establishing consistency under finiteness of only second moments and allowing for the presence of an unknown slowly varying factor $L(\lambda)$ in $f(\lambda)$, so that (1.2) is relaxed to

$$f(\lambda) \sim L(\lambda) \left| \lambda \right|^{-2d}, \quad \text{as } \lambda \to 0. \tag{3.1}$$

14

In this setting, Delgado and Robinson (1996) proposed data-dependent choices of the bandwidth number (analogous to the one discussed later in relation to log periodogram estimation, for example) that is required in the estimation, and Lobato and Robinson (1996) established limit distribution theory, which is complicated: the estimate is asymptotically normal for $0 \leq d < \frac{1}{4}$, but non-normal for $d \geq \frac{1}{4}$. Various other semiparametric estimates of d share this latter property, which is due to $f(\lambda)$ not being square-integrable for $d \geq \frac{1}{4}$.

The traditional statistical practice of regression turns out to be fruitful. The asymptotic law (1.1) suggests two approaches, nonlinearly regressing sample autocovariances on cj^{2d-1}, and ordinary linear regression (OLS) of logged sample autocovariances on $\log j$ and an intercept, as proposed by Robinson (1994a). But the limit distributional properties of these estimates are as complicated as those for the averaged periodogram estimate, intuitively because OLS is a very *ad hoc* procedure in this setting, the implied "disturbances" in the "regression model" being far from uncorrelated or homoscedastic.

We can only expect OLS to yield nice results if the disturbances are suitably "whitened". In case at least of short memory series the (Toeplitz) covariance matrix of $x_1, ..., x_n$ is approximately diagonalized by a unitary transformation, such that normalized periodograms $u_j = \log \{I(\lambda_j)/f(\lambda_j)\}$ (cf (2.4)), sufficiently resemble a zero-mean, uncorrelated, homoscedastic sequence. In case of long memory series, (1.2) suggests consideration of

$$\log I(\lambda_j) \simeq \log c - 2d \log \lambda_j + u_j, \tag{3.1}$$

for a positive constant c and λ_j close to zero, as pursued by Geweke and Porter-Hudak (1983), though they instead employed a narrow band version of the "fractional differencing" model (2.1), specifically replacing $\log \lambda_j$ by $\log |1 - e^{i\lambda_j}|$. They carried out OLS regression over $j = 1, ..., m$, where m, a bandwidth or smoothing number,

15

is much less than n but is regarded as increasing slowly with n in asymptotic theory. (Geweke and Porter-Hudak's (1983) approach was anticipated by a remark of Granger and Joyeux (1980)). Geweke and Porter-Hudak argued, in effect, that as $n \to \infty$ their estimate $\tilde{d}$ satisfies

$$m^{\frac{1}{2}}(\tilde{d} - d) \to_d N\left(0, \frac{\pi^2}{24}\right), \tag{3.2}$$

giving rise to extremely simple inferential procedures. But the heuristics underlying their argument are defective, and they, and some subsequent authors, did not come close to providing a rigorous proof of (3.2). One problem with their heuristics is that for long memory (and anti-persistent) series the u_j are not actually asymptotically uncorrelated or homoscedastic for fixed j with $n \to \infty$, as shown by Kűnsch (1986), and elaborated upon by Hurvich and Beltrao (1993), Robinson (1995a). Robinson (1995a) showed that this in itself invalidates Geweke and Porter-Hudak's (1983) argument. Even for j increasing with n, the approximation of the u_j by an uncorrelated, homoscedastic sequence is not very good, and this, and the nonlinearly-involved periodogram, makes a proof of (3.2) non-trivial.

In Robinson (1995a), (3.2) was established, explicitly in case of the approximation (3.1) rather than Geweke and Porter-Hudak's version, though indicating that the same result holds there. His result applies to the range $|d| < \frac{1}{2}$, providing simple interval estimates as well as a simple test of short memory, $d = 0$. Robinson (1995a) assumed Gaussianity, but Velasco (2000) gave an extension to linear processes x_t, both authors employing Künsch's (1986) suggestion of trimming out the lowest λ_j to avoid the anomalous behaviour of periodograms there, but Hurvich, Deo and Brodsky (1998) showed that this was unnecessary for (3.2) to hold, under suitable conditions. These authors also addressed the issue of choice of the bandwidth, m, providing optimal asymptotic minimum mean-squared error theory. If $f(\lambda)\lambda^{2d}$ is twice differentiable at $\lambda = 0$, the optimal bandwidth is of order $n^{4/5}$, but the multiplying constant depends on unknown population quantities. A consistent estimate of this constant

16

was proposed by Hurvich and Deo (1999), and hence a feasible, data-dependent choice of m. Hurvich and Beltrao (1994) had related mean squared error to integrated mean squared error in spectral density estimation, and thence proposed cross-validation procedures for choosing both m and the trimming constant. The "log-periodogram estimates" just discussed have been greatly used empirically, deservedly so in view of their nice asymptotic properties and strong intuitive appeal. But in view of the limited information it employs there is a concern about precision, and it is worth asking at least whether the information can be used more efficiently. In fact Robinson (1995a) showed that indeed the asymptotic variance in (3.2) can be reduced by "pooling" adjacent periodograms, prior to logging.

A proposal of Künsch (1987), however, leads to an alternative frequency-domain estimate that does even better. He suggested a narrow-band discrete-frequency Whittle estimate (cf (2.5)). This essentially involves Whittle estimation of the "model" $f(\lambda) = C\lambda^{-2d}$ over frequencies $\lambda = \lambda_j$, $j = 1, ..., m$, where m plays a similar role as in log periodogram estimation. After that, C can be eliminated by a side calculation (much as the innovation variance is eliminated in getting (2.5)), and d is estimated by $\hat{d}$ which minimizes

$$\log\left\{\frac{1}{m}\sum_{j=1}^{m}\lambda_j^{2d}I(\lambda_j)\right\} - \frac{2d}{m}\sum_{j=1}^{m}\log\lambda_j. \tag{3.3}$$

There is no closed-form solution to (3.3) but it is easy to handle numerically. Robinson (1995b) established that

$$m^{\frac{1}{2}}(\hat{d} - d) \to_d N(0, \frac{1}{4}). \tag{3.4}$$

For the same m sequence, $\hat{d}$ is then more efficient than the log periodogram estimate $\tilde{d}$ (cf (3.2)), while the pooled log periodogram estimate of Robinson (1995a) has asymptotic variance that converges to $\frac{1}{4}$ from above as the degree of pooling increases. While $\hat{d}$ is only implicitly defined,it is nevertheless easy to locate, and the linear involvement of the periodogram in (3.3) makes it possible to establish (3.4) under

simpler and milder conditions than needed for (3.2), Robinson employing a linear process for x_t in martingale difference innovations. This, and the coverage of all $d \in (-\frac{1}{2}, \frac{1}{2})$, may have implications also for further development of the asymptotic theory of parametric Whittle estimates discussed in the previous section. An additional feature of the asymptotic theory of Robinson (1995a), and that of Robinson (1995b), is the purely local nature of the assumptions on $f(\lambda)$; and the way in which the theory fits in with earlier work on smoothed nonparametric spectral estimation for short memory series; (1.2) is refined to

$$f(\lambda) = C\,|\lambda|^{-2d}\left(1 + O\,|\lambda|^{\beta}\right), \quad \text{as } \lambda \to 0,$$

where $\beta \in (0, 2]$ is analogous to the local smoothness parameter involved in the spectral estimation work, and no smoothness, or even boundedness, is imposed on f away from zero frequency. Note that the parameter β also enters into rules for optimal choice of m; see Henry and Robinson (1996). Lobato and Robinson (1998) provided a Lagrange multiplier test of the short memory hypothesis $d = 0$ based on (3.3) that avoids estimation of d.

Various refinements to the semiparametric estimates $\tilde{d}$ and $\hat{d}$, and their asymptotic theory, have been developed. Hurvich and Beltrao (1994), Hurvich and Deo (1999) have proposed bias-reduced estimates, while Andrews and Guggenberger (2003), Robinson and Henry (2000) have developed estimates that can further reduce the bias, and have smaller asymptotic minimum mean squared error, using respectively an extended regression and higher-order kernels, Robinson and Henry (2000) at the same time introducing a unified M-estimate class that includes $\tilde{d}$ and $\hat{d}$ as special cases. Giraitis and Robinson's (2003) development of an Edgeworth expansion for a modified version of $\hat{d}$ also leads to bias reduction, and a rule for bandwidth choice. An alternative refinement of $\hat{d}$ was developed by Andrews and Sun (2004). Additionally, Moulines and Soulier (1999, 2000) and Hurvich and Brodsky (2001) considered a broad-band

version of $\tilde{d}$ originally proposed by Janacek (1982), effectively extending the regression in (3.1) over all Fourier frequencies after including cosinusoidal terms, corresponding to the model (2.7) with p, now a bandwidth number, increasing slowly with n. These authors showed that if $f(\lambda)\lambda^{2d}$ is analytic over all frequencies, an asymptotic mean squared error of order $\log n/n$ can thereby be obtained, which is not achievable by the refinements to $\tilde{d}$ and $\hat{d}$ we have discussed, though the latter require only local-to-zero assumptions on $f(\lambda)$.

Volatility Models

For financial time series, "long memory" has been found not so much in raw time series x_t as in nonlinear instantaneous functions such as their squares, x_t^2. Thus, whereas we have so far presented long memory as purely a second-order property of a time series, referring to autocovariances or spectral structure, these do not completely describe non-Gaussian processes, where "memory" might usefully take on a rather different meaning. Passing a process through a nonlinear filter can change asymptotic autocovariance structure, and as Rosenblatt (1961) showed, if x_t is a stationary long memory Gaussian process satisfying (1.1), then x_t^2 has autocovariance decaying like j^{4d-2}, so has "long memory" only when $\frac{1}{4} \le d < \frac{1}{2}$, and even here, since $4d-2 < 2d-1$, x_t^2 has "less memory" than x_t.

Financial time series frequently suggest a reverse kind of behaviour. In particular, asset returns, or logged asset returns, may exhibit little autocorrelation, as is consistent with the efficient markets hypothesis, whereas their squares are noticeably correlated. Whereas our previous focus on second order moments led to linear time series models, we must now consider nonlinear ones. There is any number of possibilities, but Engle (1982) proposed to model this phenomenon by the autoregressive conditionally heteroscedastic model of order p $(ARCH(p))$, such that

$$x_t = \varepsilon_t \sigma_t,$$

19

where

$$\sigma_t^2 = E\left(x_t^2 \,|x_{t-1}, x_{t-2}, ...\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j x_{t-j}^2, \qquad (4.1)$$

with $\alpha_0 > 0$, $\alpha_j \geq 0$, $1 \leq j \leq p$, and ε_t is a sequence of iid random variables (possibly Gaussian). Under suitable conditions on the α_j, it follows that the x_t are martingale differences (and thus uncorrelated), whereas the x_t^2 have an $AR(p)$ representation, in terms of martingale difference (but not conditionally homoscedastic) innovations. The model was extended by Bollerslev (1986) to the generalized autoregressive conditionally heteroscedastic model of index p, q $(GARCH(p, q))$ which implies that the x_t^2 have an $ARMA(\max(p, q), q)$ representation in a similar sense.

The ARCH and GARCH models have found considerable use in finance. But they imply that the autocorrelations of the squares x_t^2 either eventually cut off completely or decay exponentially, whereas empirical evidence of slower decay perhaps consistent with long memory, has accumulated, see e.g. Whistler (1990), Ding, Granger and Engle (1993). Robinson (1991) had already suggested ARCH-type models capable of explaining greater autocorrelation in squares, so that (4.1) is extended to

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j x_{t-j}^2, \qquad (4.2)$$

or replaced by

$$\sigma_t^2 = \left(\alpha_0 + \sum_{j=1}^{\infty} \alpha_j x_{t-j}\right)^2. \qquad (4.3)$$

In case of both models, and related situations, Robinson (1991) developed Lagrange multiplier or score tests of "no-ARCH" (which is consistent with $\alpha_j = 0$, $j \geq 1$) against general parameterizations in (4.2) and (4.3); such tests should be better at detecting autocorrelation in x_t^2 that falls off more slowly than ones based on the $ARCH(p)$, (4.2), say.

We can formally rewrite (4.2) as

$$x_t^2 - \sum_{j=1}^{\infty} \alpha_j x_{t-j}^2 = \alpha_0 + \nu_t, \qquad (4.4)$$

20

where the $\nu_t = x_t^2 - \sigma_t^2$ are martingale differences. Robinson (1991) suggested the possibility of using for α_j in (4.4) the AR weights from the $FARIMA(0, d, 0)$ model (see (2.1)), taking $\alpha_0 = 0$, and Whistler (1990) applied this version of his test to test $d = 0$ in exchange rate series. This $FARIMA(0, d, 0)$ case was further considered by Ding and Granger (1996), along with other possibilities, but sufficient conditions of Giraitis, Kokoszka and Leipus (2000) for existence of a covariance stationary solution of (4.4) rule out long memory, though they do permit strong autocorrelation in x_t^2 that very closely approaches it, and Giraitis and Robinson (2001) have established asymptotic properties of Whittle estimates based on squares for this model. For $FARIMA(p, d, q)$ AR weights α_j in (4.2), x_t^2 is not covariance stationary when $d > 0$, $\alpha_0 > 0$, and Baillie, Bollerslev and Mikkelsen (1996) called this FIGARCH, a model that has since been widely applied in finance.

For model (4.3), Giraitis, Robinson and Surgailis (2000) have shown that if the weights α_j decay like j^{d-1}, $0 < d < \frac{1}{2}$, then any integral power x_t^k, such as the square, has long memory autocorrelation, satisfying (1.1) irrespective of k. This model also has the advantage over (4.2) of avoiding the non-negativity constraints on the α_j, and an ability to explain leverage.

An alternative approach to modelling autocorrelation in squares, and other nonlinear functions, alongside possible lack of autocorrelation in x_t, expresses σ_t^2 directly in terms of past ε_t, rather than past x_t, leading to a nonlinear MA form. Nelson (1991) proposed the exponential GARCH (EGARCH) model, where we take

$$\ln \sigma_t^2 = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j g(\varepsilon_{t-j}),$$

g being a user-chosen nonlinear function, for example Nelson stressed $g(z) = \theta z + \gamma(|z| - E|z|)$, which is useful in describing a leverage effect. Nelson (1991) noted the potential for choosing the α_j to imply long memory in σ_t^2, but stressed short memory, ARMA, weights α_j. On the other hand, Robinson and Zaffaroni (1997) proposed

nonlinear MA models, such as

$$x_t = \varepsilon_t \left(\alpha_0 + \sum_{j=1}^{\infty} \alpha_j \varepsilon_{t-j} \right), \tag{4.5}$$

where the ε_t are an iid sequence. They showed the ability to choose the α_j such that x_t^2 has long memory autocorrelation, and proposed use of Whittle estimation based on the x_t^2.

Another model, closely related to (4.5), proposed by Robinson and Zaffaroni (1998), replaces the first factor ε_t by η_t, where the η_t are iid and independent of the ε_t, again long memory potential was shown. This model is a special case of

$$x_t = \eta_t h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...), \tag{4.6}$$

of which the short memory stochastic volatility model of Taylor (1986) is also a special case. Long memory versions of Taylor's model were studied by Breidt, Crato and de Lima (1998), choosing

$$h(\varepsilon_{t-1}, \varepsilon_{t-2}, ...) = \exp \left(\alpha_0 + \sum_{j=1}^{\infty} \alpha_j \varepsilon_{t-j} \right), \tag{4.7}$$

the α_j being MA weights in the $FARIMA(p, d, q)$. They considered Whittle estimation based on squares, discussing its consistency, and applying the model to stock price data.

Asymptotic theory for ML estimates of models such as (4.5), (4.6) and (4.7) is considerably more difficult to derive, indeed it is hard to write down the likelihood, given, say, Gaussian assumptions on ε_t and η_t. In order to ease mathematical tractability in view of the nonlinearity in (4.7), Gaussianity of ε_t was stressed by Breidt et al. In that case, we can write the exponent of h in (4.7) as $\alpha_0 + z_t$, where z_t is a stationary Gaussian, possibly long memory, process, and likewise the second factor in (4.5). Such models are all covered by modelling x_t as a general nonlinear function of a vector

unobservable Gaussian process ξ_t. Starting from an asymptotic expansion for the covariance of functions of multivariate normal vectors, Robinson (2001) indicated how long memory in nonlinear functions of x_t depends on the long memory in ξ_t and the nature of the nonlinearity involved, with application also to cyclic behaviour, cross-sectional and temporal aggregation, and multivariate models. Allowance for quite general nonlinearity means that relatively little generality is lost by the Gaussianity assumption on ξ_t, while the scope for studying autocorrelation structure of functions such as $|x_t|$ can avoid the assumption of a finite fourth moment in x_t, which has been controversial.

Semiparametric models and methods for long memory in volatility have also been considered. In particular, Hurvich, Moulines and Soulier (2005) investigated properties of the narrow-band discrete-frequency Whittle estimate (discussed in the previous section) based on the $\log x_t^2$ series.

Nonstationary Models

In time series econometrics, unit root models have been a major focus over the past 30 years. Previously to this, modelling of economic time series typically involved a combination of short memory, $I(0)$, series and ones that are nonstochastic, either in the sense of sequences such as dummy variables or polynomial time trends, or of conditioning on predetermined economic variables. On the other hand, unit root modelling starts from the random walk model, i.e. (2.2) for $t \geq 1$ with $d = 1, e_t$ white noise and $x_0 = 0$, and then generalizes e_t to be a more general $I(0)$ process, modelled either parametrically or nonparametrically; x_t is then said to be an $I(1)$ process. Such models, often with the involvement also of nonstationary time trends, have been successfully used in macroeconometrics, frequently in connection with cointegration analysis.

One essential preliminary step is the testing of the unit root hypothesis. Numerous such tests have been proposed, often directed against $I(0)$ alternatives, and using classical Wald, Lagrange multiple and likelihood-ratio procedures, see e.g. Dickey and Fuller (1979). In classical situations, these lead to a null χ^2 limit distribution, a non-central local χ^2 limit distribution, Pitman efficiency, and a considerable degree of scope for robustness to the precise implementation of the test statistics, for example to the estimate of the asymptotic variance matrix that is employed. The unit root tests against $I(0)$ alternatives lose such properties, for example the null limit distribution is nonstandard. This nonstandard behaviour arises essentially because the unit root is nested unsmoothly in an AR system: in the $AR(1)$ case, the process is stationary with exponentially decaying autocovariance structure when the AR coefficient α lies between -1 and 1, has unit root nonstationarity at $\alpha = 1$, and is "explosive" for $|\alpha| > 1$. The tests directed against AR alternatives seem not to have very good powers against fractional alternatives, as Monte Carlo investigation of Diebold and Rudebusch (1991) suggests.

Any number of models can potentially nest a unit root, and the fractional class turns out to have the "smooth" properties that lead classically to the standard, optimal asymptotic behaviour referred to earlier. Robinson (1994c) considered the model

$$\varphi(L)x_t = u_t, \quad t \geq 1; \quad x_t = 0, \quad t \leq 0, \tag{5.1}$$

where u_t is an $I(0)$ process with parametric autocorrelation and

$$\varphi(L) = (1 - L)^{d_1}(1 + L)^{d_2} \prod_{j=3}^{h}(1 - 2\cos\omega_j L + L^2)^{d_j}, \tag{5.2}$$

where the ω_j are given distinct real numbers in $(0, \pi)$, and the d_j, $1 \leq j \leq h$, are arbitrary real numbers. The initial condition in (5.1) avoids an unbounded variance, the main interest being in nonstationary x_t. Robinson (1994c) proposed tests for specified values of the d_j against, fractional, alternatives in the class (5.2). For example, in

24

the simplest case the unit root hypothesis $d = 1$ can be tested, but against fractional alternatives $(1 - L)^d$ for $d > 1$, $d < 1$ or $d \neq 1$. Some other null d may be of interest, e.g. $d = \frac{1}{2}$, this being the boundary between stationarity and nonstationarity in the fractional domain. The region $d \in [\frac{1}{2}, 1)$ has been referred to as mean-reverting, MA coefficients of x_t decaying, albeit more slowly than under stationary, $d < \frac{1}{2}$. Note that the models (5.1) and (5.2) also cover seasonal and cyclical components (cf the Gegenbauer model (2.8)) as well as stationary and overdifferenced ones. Robinson (1994c) showed that his Lagrange multiplier tests enjoy the classical large-sample properties of such tests, described above.

To intuitively explain this outcome, note that unlike in unit root tests against AR alternatives, the test statistics are based on the null differenced x_t, which are $I(0)$ under the null hypothesis. This would suggest that estimates of memory parameters d_j in (5.1) and (5.2) and of parameters describing u_t, such as Whittle estimates, will also continue to possess the kind of standard asymptotic properties - $\sqrt{n}$-consistency and asymptotic normality - under nonstationarity as we have encountered in stationary circumstances. Beran (1995), in case $\varphi(L) = (1 - L)^d$ and u_t white noise, indicated this, though the initial consistency proof he provides, an essential preliminary to asymptotic distribution theory for his implicitly-defined estimate, appears to assume that the estimate lies in a neighbourhood of the true d, itself a consequence of consistency; over a suitably wide range of d-values, the objective function does not converge uniformly.

Velasco and Robinson (2000) adopted a somewhat different approach, employing instead the model

$$(1 - L)^s x_t \;=\; v_t, \quad t \geq 1; \; x_t = 0, \quad t \leq 0, \tag{5.3}$$

$$(1 - L)^{d-s} v_t \;=\; u_t, \quad t = 0, \pm 1, ..., \tag{5.4}$$

where s denotes the integer part of $d + \frac{1}{2}$ and u_t is a parametric $I(0)$ process such

as white noise;v_t is a stationary $I(d-s)$ process, invertible also unless $d = -\frac{1}{2}$. The distinction between the two definitions of nonstationary $I(d)$ processes, in (5.1) on the one hand and (5.3) and (5.4) on the other, was discussed by Marinucci and Robinson (1999); this entails, for example, convergence to different forms of fractional Brownian motion. Velasco and Robinson (2000) considered a version of discrete-frequency Whittle estimation (cf. (2.5)) but nonstationarity tends to bias periodogram ordinates, and to sufficiently reduce this they in general (for $d \geq \frac{3}{4}$ and with (5.4) modified so that v_t has an unknown mean) found it necessary to suitably "taper" the data, and then, in order to overcome the undesirable dependence this produces between neighbouring periodograms, to use only Fourier frequencies λ_j, such that j is a multiple of p: p is the "order" of the taper, such that $p \geq [d + \frac{1}{2}] + 1$ is required for asymptotic normality of the estimates, with $\sqrt{n}$ rate of convergence; since d is unknown a large p can be chosen for safety's sake, but the asymptotic variance is inflated by a factor varying directly with p. The theory is invariant to an additive polynomial trend of degree up to p.

For parametric versions of the form (5.1), such as the $FARIMA(p, d, q)$, Hualde and Robinson (2011), considered the conditional sum of squares type of estimate used by Box and Jenkins (1971) in an $ARMA$ context. This particular form, rather than other versions of Whittle estimate, turns out to be crucial in establishing asymptotic statistical properties without resorting to trimming, indeed as a result asymptotic effciency is achieved. A notable feature of Hualde and Robinson (2011) is that it is not necessary to know in advance whether d lies in the stationary, nonstationary, anti-persistent or non-invertible regions.

In case of semiparametric models, tapering has played a larger role in covering nonstationarity. In a similar model to (5.3) and (5.4), originated in Hurvich and Ray (1995), who required $s = 1$ in (5.3), while in (5.4) they allowed $-\infty < d < \frac{1}{2}$, so that (unlike Beran (1995) and Velasco and Robinson (2000)) they covered non-

stationarity only up to $d < 3/2$ (though this probably fits many applications) and on the other hand covered any degree of noninvertibility. Hurvich and Ray's (1995) concern, however, was not with asymptotic theory for parameter estimates, rather they found that asymptotic bias in $I(\lambda_j)$, for fixed j as $n \to \infty$, could be notably reduced by use of a cosine bell taper, leading them to recommend use of tapering (and omission of frequency λ_1) in the log periodogram estimation of d discussed earlier in the paper, in case nonstationarity is feared. Limit distribution theory was established by Velasco (1999a), analogous to that described above for log periodogram estimates in case $d \geq \frac{1}{2}$, in a semiparametric version of (5.3) and (5.4) (so u_t has nonparametric autocorrelation), using a general class of tapers. Further, Velasco (1999b) established analogous results for local Whittle estimates, cf (3.4). Once again, there is invariance to polynomial trends, but tapering, imposed for asymptotic normality when $d \geq \frac{3}{4}$ and for consistency when $d > 1$, entails skipping frequencies and/or an efficiency loss. However, Hurvich and Chen (2000) proposed a taper, applied to first differences when $d < 3/2$, that, with no skipping, loses less efficiency. Shimotsu and Phillips (2005) considered an "exact" form of local Whittle function, without tapering, based on (5.1), establishing asymptotic properties when the optimization covers an interval of width less than $9/2$. However, tapering can be a wise precaution when nonstationarity is believed possible, and can be useful in theoretical refinements even under stationarity, see e.g. Giraitis and Robinson (2003).

Final Comments

We have introduced various ways of modelling long memory in economic and financial time series, along with methods for estimating them, theoretical results that are useful in justifying the estimates, and methods for drawing statistical inferences on them, such as testing hypotheses and setting confidence intervals. The literature on

long memory is now quite mature, and so we have chosen to focus on some basic literature for scalar time series. Space does not permit a full treatment of developments even in this scalar case, let alone in other settings where long memory can arise and has been studied, which are beyond the scope of the paper and would oveburden it with notation. These include models for multivariate data, including regression and cointegration models, spatial models, panel data models, and functional time series models.

REFERENCES

[1] ADELMAN, I. (1965), 'Long Cycles: Fact or Artefact?', *American Economic Review*, 55: 444-63.

[2] ADENSTEDT, R.K. (1974), 'On Large-sample Estimation of the Mean of a Stationary Random Sequence', *Annals of Statistics*, 2: 1095-107.

[3] ANDREWS, D.W.K., and K. GUGGENBERGER (2003), 'A Bias-Reduced Log-periodogram Estimator for the Long Memory Parameter', *Econometrica*, 71: 675-712.

[4] ANDREWS, D.W.K. and Y. SUN (2004), 'Adaptive Local Polynomial Whittle Estimation of Long-range Dependence', *Econometrica* 72:, 569-614.

[5] BAILLIE, R.T., T. BOLLERSLEV and H.O. MIKKELSEN (1996), 'Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity; *Journal of Econometrics*, 74: 3-30.

[6] BERAN, J. (1995), 'Maximum Likelihood Estimation of the Differencing Parameter for Invertible Short- and Long-memory ARIMA Models', *Journal of the Royal Statistical Society, Series B*, 57: 659-72.

[7] BLOOMFIELD, P. (1972), 'An Exponential Model for the Spectrum of a Scalar Time Series', *Biometrika*, 60: 217-26.

[8] BOLLERSLEV, T. (1986), 'Generalized Autoregressive Conditional Heteroskedasticity', *Journal of Econometrics*, 31: 307-27.

[9] BOX, G.E.P., and G.M. JENKINS (1971), *Time Series Analysis, Forecasting and Control*, San Francisco: Holden-Day.

[10] BREIDT, F.J., N. CRATO, and P. de LIMA (1998), 'The Detection and Estimation of Long Memory in Stochastic Volatility', *Journal of Econometrics*, 83: 325-34.

[11] DAHLHAUS, R. (1989), 'Efficient Parameter Estimation for Self-similar Processes', *Annals of Statistics*, 17: 1749-66.

[12] DELGADO, M.J., and P.M. ROBINSON (1996), 'Optimal Spectral Bandwidth for Long Memory', *Statistica Sinica*, 6: 97-112.

[13] DICKEY, D.A., and W.A. FULLER (1979), 'Distribution of the Estimators for Autoregressive Time Series with a Unit Root', *Journal of the American Statistical Association*, 74: 427-31.

[14] DIEBOLD, F.X., and G.D. RUDEBUSCH (1991), 'On the Power of Dickey-Fuller Tests Against Fractional Alternatives', *Economic Letters*, 35: 155-60.

[15] DING, Z., and C.W.J. GRANGER (1996), 'Modelling Volatility Persistence of Speculative Returns: A New Approach', *Journal of Econometrics*, 73: 185-215.

[16] _____, C.W.J. GRANGER, and R.F. ENGLE (1993), 'A Long Memory Property of Stock Market Returns and a New Model', *Journal of Empirical Finance*, 1: 83-106.

[17] ENGLE, R.F. (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica*, 50: 987-1007.

[18] FAIRFIELD SMITH, H. (1938), 'An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops', *Journal of Agricultural Science*, 28: 1-23.

[19] FOX, R., and M.S. TAQQU (1986), 'Large Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series', *Annals of Statistics*, 14: 517-32.

[20] _____ (1987), 'Central Limit Theorems for Quadratic Forms in Random Variables Having Long-range Dependence', *Probability Theory and Related Fields*, 74: 213-40.

[21] GEWEKE, J., and S. PORTER-HUDAK (1983), 'The Estimation and Application of Long-Memory Time Series Models', *Journal of Time Series Analysis*, 4: 221-38.

[22] GIRAITIS, L., F.J. HIDALGO, and P.M. ROBINSON (2001), 'Gaussian Estimation of Parametric Spectral Density with Unknown Pole', *Annals of Statistics*, 29: 987-1023.

[23] _____, P. KOKOSZKA, and R. LEIPUS (2000), 'Stationary ARCH Models: Dependence Structure and Central Limit Theorem', *Econometric Theory*, 16:3-22.

[24] _____, H. KOUL, and D. SURGAILIS (2012), *Large Sample Inference for Long Memory Processes,* London, Imperial College Press.

[25] _ _ _ _ _ _ _ _, and P.M. ROBINSON (2001), 'Whittle Estimation of ARCH Models', *Econometric Theory*, 17: 608-31.

[26] _ _ _ _ _ _ _ _, and P.M. ROBINSON (2003), 'Edgeworth Expansions for Semiparametric Whittle Estimation of Long Memory', *Annals of Statistics*, 31: 987-1023.

[27] _ _ _ _ _ _ _ _, P.M. ROBINSON, and D. SURGAILIS (2000), 'A Model for Long Memory Conditional Heteroscedasticity', *Annals of Applied Probability*, 10: 1002-24.

[28] _ _ _ _ _ _ _ _, and D. SURGAILIS (1990), 'A Central Limit Theorem for Quadratic Forms in Strongly Dependent Random Variables and its Application to Asymptotical Normality of Whittle's Estimate', *Probability Theory and Related Fields*, 86: 87-104.

[29] _ _ _ _ _ _ _ _, and M.S. TAQQU (1999), 'Whittle Estimator for Finite - Variance Non-Gaussian Time Series with Long Memory', *Annals of Statistics*, 27: 178-203.

[30] GRANGER, C.W.J. (1966), 'The Typical Spectral Shape of an Economic Variable', *Econometrica*, 34: 150-61.

[31] _ _ _ _ _ _ _ _ _ (1980), 'Long Memory Relationships and the Aggregation of Dynamic Models', *Journal of Econometrics*, 14: 227-38.

[32] _ _ _ _ _ _ _ _ _, and R. JOYEUX (1980), 'An Introduction to Long-memory Time Series and Fractional Differencing', *Journal of Time Series Analysis*, 1: 15-29.

[33] GRAY, H.L., N.I. ZHANG, and W.A. WOODWARD (1989), 'On Generalized Fractional Processes', *Journal of Time Series Analysis*, 10: 233-57.

[34] HANNAN, E.J. ((1970), 'The Asymptotic Theory of Linear Time Series Models', *Journal of Applied Probability*, 10: 130-45.

[35] _____, 'The Asymptotic Distribution of Serial Covariances', *Annals of Statistics*, 4: 396-9.

[36] HENRY, M., and P.M. ROBINSON (1996), 'Bandwidth Choice in Gaussian Semiparametric Estimation of Long-range Dependence', in P.M. Robinson and M. Rosenblatt, eds., *Athens Conference on Applied Probability and Time Series Analysis*, vol.II: Time Series Analysis. In Memory of E.J. Hannan, New York: Springer-Verlag, pp.220-32..

[37] HOSKING, J.R.M. (1981), 'Fractional Differencing', *Biometrika*, 68: 165-76.

[38] HUALDE, J., and P.M.ROBINSON (2011), 'Gaussian Pseudo-maximum Likelihood Estimation of Fractional Time Series Models', *Annals of Statistics*, 39: 3152-81.

[39] HURST, H. (1951), 'Long Term Storage Capacity of Reservoirs', *Transactions of the American Society of Civil Engineers*, 116: 770-99.

[40] HURVICH, C.M., and K.I. BELTRAO (1993), 'Asymptotics for the Low-Frequency Estimates of the Periodogram of a Long Memory Time Series', *Journal of Time Series Analysis*, 14: 455-72.

[41] _____, and K.I. BELTRAO (1994), 'Automatic Semiparametric Estimation of the Memory Parameter of a Long-memory Time Series', *Journal of Time Series Analysis*, 15: 285-302.

[42] _____, and J. BRODSKY (2001), 'Broadband Semiparametric Estimation of the Memory Parameter of a Long-memory Time Series using Fractional Exponential Model', *Journal of Time Series Analysis*, 22: 221-49.

[43] _____, and W.W. CHEN (2000), 'An Efficient Taper for Potentially Overdifferenced Long-memory Time Series', *Journal of Time Series Analysis*, 21: 155-80.

[44] _____, and R.S. DEO (1999), 'Plug-in Selection of the Number of Frequencies in Regression Estimates of the Memory Parameter of a Long-memory Time Series', *Journal of Time Series Analysis*, 20: 331-41.

[45] _____, R. DEO, and J. BRODSKY (1998), 'The Mean Squared Error of Geweke and Porter-Hudak's Estimates of the Memory Parameter of a Long Memory Time Series', *Journal of Time Series Analysis*, 19: 19-46.

[46] _____, E. MOULINES, E., and P. SOULIER (2005) 'Estimating Long Memory in Volatility', *Econometrica, 73*, 1283-1328.

[47] _____, and B.K. RAY (1995), 'Estimation of the Memory Parameter for Nonstationary or Noninvertible Fractionally Integrated Processes', *Journal of Time Series Analysis*, 16: 17-41.

[48] JANACEK, G.J. (1982), 'Determining the Degree of Differencing for Time Series via Log Spectrum', *Journal of Time Series Analysis*, 3: 177-83.

[49] KASHYAP, R., and K. EOM (1988), 'Estimation in Long-memory Time Series Model', *Journal of Time Series Analysis*, 9: 35-41.

[50] KÜNSCH, H.R. (1986), 'Discrimination Between Monotonic Trends and Long-range Dependence', *Journal of Applied Probability*, 23: 1025-30.

[51] _____ (1987), 'Statistical Aspects of Self-similar Processes', *Proceedings of the First World Congress of the Bernoulli Society*, VNU Science Press, 1: 67-74.

[52] LIPPI, M., and P. ZAFFARONI (1998), 'Aggregation and Simple Dynamics: Exact Asymptotic Results', preprint.

[53] LO, A.W. (1991), 'Long-term Memory in Stock Market Prices', *Econometrica*, 59: 1279-313.

[54] LOBATO, I.G., and P.M. ROBINSON (1996), 'Averaged Periodogram Estimation of Long Memory', *Journal of Econometrics*, 73: 303-24.

[55] _____, and P.M. ROBINSON (1998), 'A Nonparametric Test for $I(0)$', *Review of Economic Studies*, 65: 475-95..

[56] MANDELBROT, B.B. (1972), 'Statistical Methodology for Non-periodic Cycles: From the Covariance to R/S Analysis', *Annals of Economic and Social Measurement*, 1: 259-90.

[57] _____, and M.S. TAQQU (1979), 'Robust R/S Analysis of Long-run Serial Correlations', *Bulletin of the International Statistical Institute*, 48,2: 69-104.

[58] _____, and J.W. VAN NESS (1968), 'Fractional Brownian Motions, Fractional Noises and Applications', *SIAM Review*, 10: 422-37.

[59] _____, and T.R. WALLIS (1969), 'Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long Run Statistical Dependence', *Water Resources Research*, 5: 967-88.

[60] MARINUCCI, D., and P.M. ROBINSON (1999), 'Alternative Forms of Fractional Brownian Motion', *Journal of Statistical Planning and Inference*, 80: 111-22.

[61] MOULINES, E., and P. SOULIER (1999), 'Broadband Log Periodogram Regression of Time Series with Long Range Dependence', *Annals of Statistics*, 27: 1415-39.

[62] _____, and P. SOULIER (2000), 'Data Driven Order Selection for Projection Estimates of the Spectral Density of Time Series with Long Range Dependence', *Journal of Time Series Analysis*, 21: 193-218.

[63] NELSON, D. (1991), 'Conditional Heteroskedasticity in Asset Returns: A New Approach', *Econometrica*, 59: 347-70.

[64] ROBINSON, P.M. (1978a), 'Statistical Inference for a Random Coefficient Autoregressive Model', *Scandinavian Journal of Statistics*, 5: 163-8.

[65] _____ (1978b), 'Alternative Models for Stationary Stochastic Processes', *Stochastic Processes and their Applications*, 8: 151-52.

[66] _____ (1991), 'Testing for Strong Serial Correlation and Dynamic Conditional Heteroskedasticity in Multiple Regression', *Journal of Econometrics*, 47: 67-84.

[67] _____ (1994a), 'Time Series with Strong Dependence', in C.A. Sims, ed., *Advances in Econometrics*, vol.1, Cambridge: Cambridge University Press, pp.47-95.

[68] _____ (1994b), 'Semiparametric Analysis of Long-memory Time Series', *Annals of Statistics*, 22: 515-39.

[69] _____ (1994c), 'Efficient Tests of Nonstationary Hypotheses', *Journal of the American Statistical Association*, 89: 1420-37.

[70] _____ (1995a), 'Log-periodogram Regression of Time Series with Long Range Dependence', *Annals of Statistics*, 23: 1048-72.

[71] _____ (1995b), 'Gaussian Semiparametric Estimation of Long-range Dependence', *Annals of Statistics*, 23: 1630-61.

[72] _____ (2001), 'The Memory of Stochastic Volatility Models', *Journal of Econometrics*, 101: 195-218.

[73] _____, and M. HENRY (2000), 'Higher-order Kernel Semiparametric M-Estimation of Long Memory', forthcoming *Journal of Econometrics.*_____

[74] _____, and P. ZAFFARONI (1997), 'Modelling Nonlinearity and Long Memory in Time Series', *Fields Institute Communications*, 11: 161-70.

[75] _____, and P. ZAFFARONI (1998), 'Nonlinear Time Series with Long Memory: A Model for Stochastic Volatility', *Journal of Statistical Planning and Inference*, 68: 359-71.

[76] ROSENBLATT, M. (1961), 'Independence and Dependence', in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp.411-43.

[77] SAMAROV, A., and M.S. TAQQU (1988), 'On the Efficiency of the Sample Mean in Long Memory Noise', *Journal of Time Series Analysis*, 9: 191-200.

[78] SHIMOTSU, K., and P. PHILLIPS (2005), 'Exact Local Whittle Estimation of Fractional Integration', *Annals of Statistics*, 33: 1890-933.

[79] TAQQU, M.S. (1975), 'Weak Convergence to Fractional Brownian Motion and to the Rosenblatt Process', *Zeitschrift für Wahrscheinlichkeitstheorie*, 31: 287-302.

[80] TAYLOR, S.J. (1986), *Modelling Financial Time Series*, Chichester, UK.

[81] VELASCO, C. (1999a), 'Non-stationary Log-periodogram Regression', *Journal of Econometrics*, 91: 325-71.

[82] _____ (1999b), 'Gaussian Semiparametric Estimation of Nonstationary Time Series', *Journal of Time Series Analysis*. 20: 87-127.

[83] _____ (2000), 'Non-Gaussian Log-periodogram Regression', *Econometric Theory*, 16: 44-79.

[84] _____, and P.M. ROBINSON (2000), 'Whittle Pseudo-maximum Likelihood Estimation for Nonstationary Time Series', *Journal of the American Statistical Association*, 95: 1229-43.

[85] WHISTLER, D. (1990), 'Semiparametric Models of Daily and Intra-daily Exchange Rate Volatility', Ph.D. thesis, University of London.

[86] WHITTLE, R. (1951), *Hypothesis Testing in Time Series Analysis*, Uppsala: Almqvist.

[87] YAJIMA, Y. (1985), 'On Estimation of Long-memory Time Series Models', *Australian Journal of Statistics*, 27: 303-20.

[88] YONG, C.H. (1974), *Asymptotic Behaviour of Trigonometric Series*, Hong Kong: Chinese University.