Wavestrapping Time Series: Adaptive Wavelet-Based Bootstrapping

D. B. Percival, S. Sardy and A. C. Davison

1 Introduction

Suppose we observe a time series that can be regarded as a realization of a portion $X_0, X_1, \ldots, X_{N-1}$ of a real-valued zero mean Gaussian stationary process $\{X_t\}$ with autocovariance sequence (ACVS) $s_{X,\tau} \equiv \operatorname{cov}\{X_t, X_{t+\tau}\}$. Suppose also that we compute a statistic based upon our time series, e.g., the sample autocorrelation for unit lag:

$$\hat{\rho}_{X,1} \equiv \frac{\sum_{t=0}^{N-2} X_t X_{t+1}}{\sum_{t=0}^{N-1} X_t^2}. \tag{1.1}$$

To thoroughly assess the quality of $\hat{\rho}_{X,1}$ as an estimator of the corresponding population quantity $\rho_{X,1} \equiv s_{X,1}/s_{X,0}$, we need to know the distribution of $\hat{\rho}_{X,1}$; however, calculating the exact distribution of a statistic of a time series can be very difficult, so it is of interest to find reasonable approximations. If our time series were a white noise process (i.e., a sample of uncorrelated random variables (RVs), which — because of the Gaussian assumption — yields independent and identically distributed (IID) RVs), we could make use of two quite different approximations. The first approximation is based on large sample theory, which says that, as $N \to \infty$, $\hat{\rho}_{X,1}$ is approximately normally distributed with mean zero and variance $1/N$ (Bartlett, 1946; Priestley, 1981, Equation (5.3.39)). The second approximation is based on bootstrapping (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). Here we randomly sample with replacement from the original time series to create a new series of N values, for which we then compute the unit lag sample autocorrelation, say, $\hat{\rho}^{(1)}_{X,1}$. If we repeat this procedure M times to obtain $\hat{\rho}^{(1)}_{X,1}, \hat{\rho}^{(2)}_{X,1}, \ldots, \hat{\rho}^{(M)}_{X,1}$, we can use the sample distribution of these M bootstrap estimates as an approximation to the unknown distribution of $\hat{\rho}_{X,1}$. While the large sample distribution is obviously faster to compute than the bootstrap distribution for the case of $\hat{\rho}_{X,1}$, a major advantage of the bootstrap approximation is its adaptability to other statistics of interest, for which a significant amount of research might be required to work out the large sample distribution.

More generally, if $\{X_t\}$ is not necessarily white noise, we must reconsider both the large sample and bootstrap approximations to the distribution of $\hat{\rho}_{X,1}$. Under an assumption that the ACVS damps down to zero 'rapidly,' large sample approximations to the distribution of $\hat{\rho}_{X,1}$ have been worked out, but are unappealing. In particular, if we let $\rho_{X,\tau} \equiv s_{X,\tau}/s_{X,0}$ denote the τth element of the autocorrelation sequence (ACS), then $N^{1/2}(\hat{\rho}_{X,\tau} - \rho_{X,\tau})$ converges in distribution to a Gaussian distribution with mean zero and variance (Priestley, 1981, (5.3.37))

$$\sum_{\tau=-\infty}^{\infty} \left\{ \rho_{X,\tau}^2 (1 + 2\rho_{X,1}^2) + \rho_{X,\tau+1}\rho_{X,\tau-1} - 4\rho_{X,1}\rho_{X,\tau}\rho_{X,\tau-1} \right\}.$$

This expression depends upon the entire ACS, which is typically unknown in practice. Under the same assumption of rapid decorrelation, several variations on the bootstrap procedure have been proposed that provide good approximations to the distribution of $\hat{\rho}_{X,1}$ and related statistics (see Davison and Hinkley, 1997, Chapter 8, and §3 below). On the other hand, if the ACVS does not damp down rapidly but rather exhibits 'long memory' (see §2), then the large sample theory for $\hat{\rho}_{X,1}$ is currently incomplete, and standard bootstrapping procedures are known not to work very well. Stationary long memory processes (LMPs) are becoming increasingly important as models for a wide range of time series (Beran, 1994), so it is of interest to have decent approximations for the distribution of $\hat{\rho}_{X,1}$ and related statistics that allow for such processes.

In this paper, we propose 'wavestrapping,' an adaptive wavelet-based scheme for bootstrapping certain statistics for time series that can be modeled by stationary processes with either rapidly decaying ('short memory') or long memory ACVSs. The basis for this methodology is the work of Flandrin (1992), Wornell (1995) and McCoy and Walden (1996), who show that the discrete wavelet transform (DWT) approximately decorrelates long memory processes. We demonstrate that, by applying the bootstrap in the wavelet domain, we can approximate the distribution of $\hat{\rho}_{X,1}$ reasonably well for long memory processes. When applied to certain short memory processes, this DWT-based scheme is not as successful, a result that can be attributed to the fact that the DWT need not be an adequate decorrelating transform for such processes; however, in such cases, a generalization of the DWT based on discrete wavelet packet transforms (DWPTs) can yield an acceptable decorrelating transform. We propose a procedure for adaptively selecting a decorrelating transform for a given time series that involves a 'top-down' search of a collection of DWPTs with the help of white noise tests.

The remainder of this paper is organized as follows. We first review short and long memory models for time series (§2) and current approaches for bootstrapping time series (§3), after which we discuss the basic ideas behind the DWT (§4). Because the DWT acts a decorrelating transform for long memory processes, we can use it to define a bootstrapping scheme. We demonstrate the effectiveness of this scheme via Monte Carlo experiments (§5). We then consider why DWT-based bootstrapping does not work well for certain short memory processes and why wavestrapping can correct this deficiency (§6). We demonstrate via Monte Carlo experiments that wavestrapping works reasonably well for both short and long memory processes (§7). We then give examples of wavestrapping (§8), including one involving two time series of interest in atmospheric science, and we conclude with a discussion of directions for future research (§9).

2 Models for stationary time series

In this paper we concentrate on time series that can be modeled as a stationary process $\{X_t\}$ with an ACVS $s_{X,\tau}$ and spectral density function (SDF) $S_X(\cdot)$ related by

$$s_{X,\tau} = \int_{-1/2}^{1/2} e^{i2\pi f\tau} S_X(f)\, df, \quad \tau = \ldots, -1, 0, 1, \ldots.$$

Let $\{\epsilon_t\}$ be a Gaussian white noise process with mean zero and variance σ_ϵ^2. Two simple models that fit into the above framework are the first order autoregressive model (AR(1)) $X_t = \phi X_{t-1} + \epsilon_t$ with $|\phi| < 1$, for which

$$s_{X,\tau} = \frac{\phi^{|\tau|}\sigma_\epsilon^2}{1-\phi^2} \text{ and } S_X(f) = \frac{\sigma_\epsilon^2}{|1-\phi e^{-i2\pi f}|^2},$$

and the first order moving average model (MA(1)) $X_t = \epsilon_t - \theta\epsilon_{t-1}$, for which

$$s_{X,\tau} = \begin{cases} (1+\theta^2)\sigma_\epsilon^2, & \tau = 0; \\ -\theta\sigma_\epsilon^2, & \tau = \pm 1; \\ 0, & \text{otherwise.} \end{cases} \quad \text{and } S_X(f) = \sigma_\epsilon^2 |1-\theta e^{-i2\pi f}|^2.$$

Both these models have ACVSs that rapidly decay to zero: in the case of the AR(1) model, the rate of decay is exponential, whereas the MA(1) ACVS is identically zero for all lags $|\tau| \geq 2$. Because of this rapid decorrelation with increasing τ, the AR(1) and MA(1) models are sometimes said to have 'short memory.'

As an example of a simple model exhibiting long memory, let us consider a stationary Gaussian fractionally differenced (FD) process $\{X_t\}$ (Granger and Joyeux, 1980; Hosking, 1981; Beran, 1994). In terms of the white noise process $\{\epsilon_t\}$, we can represent an FD process as an infinite order MA process, namely,

$$X_t = \sum_{k=0}^{\infty} \frac{\Gamma(k+\delta)}{\Gamma(k+1)\Gamma(\delta)} \epsilon_{t-k},$$

where $-\frac{1}{2} < \delta < \frac{1}{2}$. The SDF for this process is given by

$$S_X(f) = \sigma_\epsilon^2 |2\sin(\pi f)|^{-2\delta},$$

while its ACVS can be obtained using

$$s_{X,\tau} = s_{X,\tau-1}\frac{\tau+\delta-1}{\tau-\delta}, \ \ \tau = 1, 2, \ldots, \ \text{ with } \ s_{X,0} = \frac{\sigma_\epsilon^2\Gamma(1-2\delta)}{\Gamma^2(1-\delta)}$$

(for $\tau < 0$, we have $s_{X,\tau} = s_{X,-\tau}$). When $0 < \delta < \frac{1}{2}$, the SDF has a pole at zero, in which case the process exhibits slowly decaying autocovariances because we have, for some $C_s > 0$,

$$\lim_{\tau \to \infty} \frac{s_{X,\tau}}{C_s \tau^{2\delta-1}} = 1;$$

i.e., the ACVS decays at a slower (hyperbolic) rate than for the AR(1) and MA(1) models.

3 Current approaches for bootstrapping time series

Existing procedures for bootstrapping time series can be divided into those which resample in the time and the frequency domains. In this section we review them; see Davison and Hinkley (1997, Chapter 8) and Bühlmann (1999) for details and further references.

3.1 Time domain

3.1.1 Residual bootstrap

When it is credible that $X_0, \ldots, X_{N-1}$ result from a model for which residuals can be identified, a form of model-based resampling may be applied. For example, if the series has AR(p) representation

$$X_t = \sum_{u=1}^{p} \phi_u X_{t-u} + \epsilon_t, \tag{3.1}$$

where $\{\epsilon_t\}$ is a white noise process, we can use the estimated coefficients $\hat{\phi}_u$ to determine residuals

$$r_t = X_t - \sum_{u=1}^{p} \hat{\phi}_u X_{t-u}.$$

We then generate bootstrap series according to (3.1), but with the ϕ_u replaced by their estimates, and with $\{\epsilon_t\}$ replaced with a white noise process generated by sampling independently with replacement from the residuals r_t, ideally centered and scaled to have the same mean and variance as the $\{\epsilon_t\}$. Under suitable conditions the properties of statistics constructed from the bootstrap series will mimic repeated sampling properties of statistics constructed from the original series. This procedure has the drawback that a specific model must be fitted and used for the resampling, and it will fail if that model is incorrect.

In practice the model fitted is generally selected from the data. For example, the p in (3.1) is often selected by minimizing a model selection criterion such as AIC. This corresponds to the sieve bootstrap, whose philosophy is that a wide class of models should be compared, with the best-fitting model chosen for the bootstrap. This can greatly improve on the simplistic approach in which the model is fixed, but its performance depends heavily on the adequacy of the model class chosen.

3.1.2 Block bootstrap

A nonparametric time domain approach is block resampling (Künsch, 1989). The motivation is that for many purposes the dominant property of a time series is its short-range dependence, which may (largely) be preserved by resampling blocks of consecutive observations. The simplest example is the sample autocorrelation for unit lag, $\hat{\rho}_{X,1}$, which is (almost) the solution of

$$\sum_{t=0}^{N-2} X_t(X_{t+1} - \rho X_t) = 0$$

and depends only on the marginal distribution of successive pairs of observations. The idea is to rewrite the original series as the bivariate time series

$$Y_0, Y_1, \ldots, Y_{N-2} = \begin{pmatrix} X_0 \\ X_1 \end{pmatrix}, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \ldots, \begin{pmatrix} X_{N-2} \\ X_{N-1} \end{pmatrix},$$

rewrite the algorithm that computes $\hat{\rho}_{X,1}$ as a function of the Ys, resample b blocks of l consecutive Ys, where $bl = N$, and compute the statistic from the resampled data. This preserves dependence between the Xs from which the statistic is calculated and can give excellent results for short-range dependent series. Its main drawbacks are twofold. First, it is not automatic because consideration has to be given to the rewriting of the statistic. Second, not every statistic can be written as a function of short blocks of data. Both drawbacks are non-trivial, and a simplified approach is usually applied in practice.

The simplification is to construct a new series by concatenating resampled blocks of $X_0, \ldots, X_{N-1}$, but this generally yields resampled series that are much less dependent than the original data. A drawback of this simple approach is the choice of block length, about which little is known of practical use; somewhat discouragingly, this is analogous to the choice of bandwidth in smoothing problems. Moreover, the method requires that certain sums of ACVS elements be bounded and so fails entirely for LMPs.

3.2 Bootstrapping in the frequency domain

A quite different approach is based on the Fourier transform for stationary processes (Priestley, 1981). Let

$$\widetilde{X}_k = \sum_{t=0}^{N-1} X_t e^{-i2\pi kt/N}, \quad k = 0, \ldots, N-1,$$

be the discrete Fourier transform (DFT) of the time series. The sequence $\widetilde{X}_0, \ldots, \widetilde{X}_{N-1}$ comprises the empirical Fourier transform of the data; the periodogram has elements $N^{-1}|\widetilde{X}_k|^2$, which summarize frequency information in the series. Under suitable conditions and as $N \to \infty$, the real and imaginary parts of the $\widetilde{X}_k$ are distributed like a sample of independent normal variables, with means zero and variances $NS_X(k/N)/2$. One implication of this is that the phase and modulus of each $\widetilde{X}_k$ are independent. A second is that the $N^{-1}|\widetilde{X}_k|^2$ are asymptotically independent with scaled chi-squared distributions. Both properties suggest possible resampling schemes.

3.2.1 Phase scrambling

The independence of the phase and modulus of the $\widetilde{X}_k$ suggests that a resampled series with the same periodogram can be made by generating phases but keeping moduli fixed. To be specific, let $U_0, \ldots, U_{N-1}$ be independent variables uniform on $(0, 2\pi)$ and set

$$\widetilde{X}_k^* = 2^{-1/2}\left\{e^{-iU_k}\widetilde{X}_k + \overline{e^{-iU_{N-k}}\widetilde{X}_{N-k}}\right\}, \quad k = 0, \ldots, N-1,$$

where the overbar denotes complex conjugate. Then the inverse Fourier transform of $\widetilde{X}_0^*, \ldots, \widetilde{X}_{N-1}^*$ is a series with the same periodogram as $X_0, \ldots, X_{N-1}$ but randomized phases. Unfortunately this resampling scheme and its variants apply to a very limited range of statistics, because they mimic only second-order properties of the original data. Moreover variability is underestimated because this resampling scheme fixes the periodogram, unlike for the original series whose periodogram is random, and statistics that can be computed from the periodogram, such as $\hat{\rho}_{X,1}$, display no variation across resamples.

3.2.2 Bootstrapping the periodogram

Another frequency domain approach potentially suitable for statistics that can be computed from the periodogram stems from the observation that the $N^{-1}|\widetilde{X}_k|^2$, $k = 1, \ldots, N-1$, have independent exponential distributions with means $S_X(k/N)$ as $N \to \infty$. This suggests using an estimate $\hat{S}_X(k/N)$ of the SDF to make residuals $r'_t = N^{-1}|\widetilde{X}_k|^2/\hat{S}_X(k/N)$, $k = 1, \ldots, N-1$, which are then resampled and merged with the estimate to give a new periodogram with elements $S_X(k/N)r_t'^*$, where the $r_t'^*$ are a random sample taken with replacement from the r'_t. The motivation is that the $N^{-1}|\widetilde{X}_k|^2/S_X(k/N)$ form a random sample from the exponential distribution, and the hope is that the r'_t are (almost) such a sample also.

If the SDF $S_X(f)$ is known apart from the values of a few parameters, this approach is essentially model-based, and will share the good and bad aspects of the schemes discussed in §3.1.1. If $S_X(f)$ is estimated nonparametrically, for example by a kernel method, then a bandwidth must be chosen. It turns out that three bandwidths are needed if the bootstrap is to work, one for the original estimate, a smaller one to estimate the residuals consistently and a larger one to give an estimate to which the resampled residuals should be added. Unfortunately the literature contains little theoretical guidance about how they should be chosen, while the numerical evidence is scant and equivocal. Hence although this method has the appeal of not involving the construction of a bootstrap series, it cannot yet be recommended for general use, even for statistics that depend only on the periodogram. In any case it may not be applied to other statistics.

4 The discrete wavelet transform

The discrete wavelet transform (DWT) is an orthonormal transform $\mathcal{W}$ that takes a time series $\mathbf{X} = [X_0, X_1, \ldots, X_{N-1}]^T$ and yields a vector of N DWT coefficients $\mathbf{W} \equiv \mathcal{W}\mathbf{X}$. The orthonormality condition $\mathcal{W}^T\mathcal{W} = I_N$ implies that we can reconstruct the time series from its DWT coefficients via $\mathbf{X} = \mathcal{W}^T\mathbf{W}$, so $\mathbf{W}$ is fully equivalent to $\mathbf{X}$. Under the assumption that N is an integer multiple of 2^{J_0}, where J_0 is an integer denoting the number of levels in the DWT, we can partition the DWT coefficient vector into subvectors:

$$\mathbf{W} = [\mathbf{W}_1^T, \mathbf{W}_2^T, \ldots, \mathbf{W}_{J_0}^T, \mathbf{V}_{J_0}^T]^T.$$

The subvector $\mathbf{W}_j$ contains $N_j \equiv N/2^j$ wavelet coefficients associated with scale $\tau_j \equiv 2^{j-1}$, whereas $\mathbf{V}_{J_0}$ contains $N/2^{J_0}$ scaling coefficients associated with scale $\lambda_{J_0} \equiv 2^{J_0}$. To see what we mean by scale and what the wavelet and scaling coefficients are telling us about the time series, let us define an average of λ contiguous time series values ending with index t as

$$\overline{X}_t(\lambda) \equiv \frac{1}{\lambda}\sum_{l=0}^{\lambda-1} X_{t-l}.$$

We define the scale associated with this average to be λ. With this definition, let us consider the special case of the Haar DWT, for which the DWT coefficients have the form

$$W_{j,n} \propto \overline{X}_{\lambda_j(n+1)-1}(\tau_j) - \overline{X}_{\lambda_j(n+1)-1-\tau_j}(\tau_j) \text{ and } V_{J_0,n} \propto \overline{X}_{\lambda_{J_0}(n+1)-1}(\lambda_j),$$

where $W_{j,n}$ and $V_{J_0,n}$ are the nth elements of, respectively, $\mathbf{W}_j$ and $\mathbf{V}_{J_0}$. Note that the Haar wavelet coefficients $W_{j,n}$ are proportional to first differences of adjacent averages over scale τ_j, whereas the Haar scaling coefficients $V_{J_0,n}$ are proportional to averages over scale λ_{J_0}. This same pattern holds for DWTs other than the Haar, in that we can regard the wavelet coefficients as being proportional to (higher order) differences of (weighted) averages over scale τ_j, and the scaling coefficients as being proportional to (weighted) averages over scale λ_{J_0}.

We can formally describe the DWT in terms of wavelet and scaling filters as follows. Let $\mathbf{h}_1 = [h_{1,0}, \ldots, h_{1,L-1}, 0 \ldots, 0]^T$ be a vector of length N whose first $L < N$ elements are the unit level wavelet filter coefficients for a Daubechies compactly supported wavelet (see Daubechies, 1992, Chapter 6). For example, the Haar wavelet filter has $L = 2$ coefficients, namely, $h_{1,0} = \frac{1}{\sqrt{2}}$ and $h_{1,1} = -\frac{1}{\sqrt{2}}$. Let $H_{1,k}, k = 0, \ldots, N-1$, be the DFT of $\mathbf{h}_1$. In the Haar case, we have $H_{1,k} = (1 - e^{-i2\pi k/N})/\sqrt{2}$. Let $\mathbf{g}_1 = [g_{1,0}, \ldots, g_{1,L-1}, 0, \ldots, 0]^T$ be a vector of length N containing the zero padded scaling filter coefficients for unit level, defined via $g_{1,l} = (-1)^{l+1} h_{1,L-1-l}$ for $l = 0, \ldots, L-1$, and let $G_{1,k}$ denote its DFT. Like the Haar wavelet filter, the Haar scaling filter has two nonzero elements, namely, $g_{1,0} = g_{1,1} = \frac{1}{\sqrt{2}}$, and its DFT is $G_{1,k} = (1 + e^{-i2\pi k/N})/\sqrt{2}$. The level j wavelet filter is given by the elements of the vector $\mathbf{h}_j$, which is the inverse DFT of

$$H_{j,k} = H_{1,2^{j-1}k \bmod N} \prod_{l=0}^{j-2} G_{1,2^l k \bmod N}, \quad k = 0, \ldots, N-1.$$

When $N > L_j = (2^j - 1)(L - 1) + 1$, the last $N - L_j$ elements of $\mathbf{h}_j$ are zero, so the jth wavelet filter $\mathbf{h}_j$ has no more than L_j non-zero elements. In the Haar case, we have $L_j = 2^j$, and, when $N > 2^j$, the elements of $\mathbf{h}_j$ are

$$h_{j,l} = \begin{cases} 1/2^{j/2}, & l = 0, \ldots, 2^{j-1} - 1; \\ -1/2^{j/2}, & l = 2^{j-1}, \ldots, 2^j - 1; \text{ and} \\ 0, & l = 2^j, \ldots, N-1. \end{cases}$$

Similarly, the level J_0 scaling filter is contained in $\mathbf{g}_{J_0}$, whose elements are the inverse DFT of

$$G_{J_0,k} = \prod_{l=0}^{J_0-1} G_{1,2^l k \bmod N}, \quad k = 0, \ldots, N-1.$$

The elements of the Haar $\mathbf{g}_{J_0}$ are

$$g_{J_0,l} = \begin{cases} 1/2^{J_0/2}, & l = 0, \ldots, 2^{J_0} - 1; \text{ and} \\ 0, & l = 2^{J_0}, \ldots, N-1. \end{cases}$$

To obtain the jth level wavelet coefficients, we filter $\mathbf{X}$ using $\mathbf{h}_j$ and subsample every 2^jth value from the filter output:

$$W_{j,n} = \sum_{l=0}^{\min(L_j,N)-1} h_{j,l} X_{2^j(n+1)-1-l \bmod N}, \quad n = 0, \ldots, N_j - 1; \tag{4.1}$$

an analogous expression yields the J_0th level scaling coefficients. In the Haar case we can write

$$W_{j,n} = \frac{1}{2^{j/2}} \sum_{l=0}^{2^{j-1}-1} X_{2^j(n+1)-1-l} - \frac{1}{2^{j/2}} \sum_{l=2^{j-1}}^{2^j-1} X_{2^j(n+1)-1-l}.$$

This example is atypical in that we do not need to use the 'modulo N' operation. For wavelet filters such that $L > 2$, the wavelet coefficients are obtained by treating the time series as if it were circular (i.e., as if it were a periodic sequence with period N). This assumption is problematic and yields a certain number of 'boundary' coefficients whose statistical properties differ from coefficients unaffected by circularity (the number of boundary coefficients on any given level is no more than $L - 2$, which is consistent with there being no such coefficients in the Haar case). In practice the DWT coefficients are not computed directly via (4.1) but rather via an elegant pyramid algorithm (Mallat, 1989) that filters $\mathbf{X}$ using $\mathbf{h}_1$ and $\mathbf{g}_1$, retains the odd-indexed values of the wavelet filter output as the unit level wavelet coefficients and then repeats this process with $\mathbf{X}$ replaced by the odd-indexed values of the scaling filter output to obtain the level $j = 2$ wavelet coefficients and so forth.

5 DWT-based bootstrapping

The idea behind DWT-based bootstrapping is to make use of the fact that, for FD and certain other stationary processes, the DWT acts as a decorrelating transform for a time series; i.e., whereas the time series itself can exhibit a high degree of autocorrelation, its DWT coefficients can — to a reasonable approximation — be regarded as uncorrelated. To quantify this decorrelation effect, we first note that, if we ignore boundary coefficients, then within a given level we have

$$\operatorname{cov}\{W_{j,n}, W_{j,n+\tau}\} = \sum_{m=-(L_j-1)}^{L_j-1} s_{X,2^j\tau+m} \sum_{l=0}^{L_j-|m|-1} h_{j,l}h_{j,l+|m|}. \tag{5.1}$$

We can use the above to compute the unit lag correlations $\operatorname{corr}\{W_{j,t}, W_{j,t+1}\}$ for, e.g., an FD process with $\delta = 0.45$. Table 5.1 lists these correlations for the Haar, D(4) and LA(8) wavelet filters and scales 1, 2, 4 and 8; here 'D(4)' and 'LA(8)' refer to the Daubechies extremal phase filter with four nonzero coefficients and to her least asymmetric filter with eight coefficients (Daubechies, 1992). Note that these correlations are all negative, with departures from zero increasing somewhat as j increases. For larger lags, computations indicate that the autocorrelation damps down roughly as dictated by an AR(1) model, i.e., $\operatorname{corr}\{W_{j,t}, W_{j,t+\tau}\} \approx (\operatorname{corr}\{W_{j,t}, W_{j,t+1}\})^{|\tau|}$. To quantify correlation between different levels, we note that (again ignoring boundary coefficients)

$$\operatorname{cov}\{W_{j,n}, W_{j',n'}\} = \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_{j'}-1} h_{j,l}h_{j',l'}s_{X,2^j(n+1)-l-2^{j'}(n'+1)+l'}. \tag{5.2}$$

For the same FD process as before, Table 5.2 lists $\max_{n,n'} |\operatorname{corr}\{W_{j,n}, W_{j',n'}\}|$ for $1 \leq j < j' \leq 4$. We can deduce from these two tables that, while the

Scale	Haar	D(4)	LA(8)
1	−0.0626	−0.0797	−0.0767
2	−0.0947	−0.1320	−0.1356
4	−0.1133	−0.1511	−0.1501
8	−0.1211	−0.1559	−0.1535

Table 5.1: Lag one autocorrelations for wavelet coefficients of scales 1, 2, 4, and 8 for an FD process with $\delta = 0.45$ using the Haar, D(4) and LA(8) wavelet filters.

	Haar			D(4)			LA(8)		
Scale	2	4	8	2	4	8	2	4	8
1	0.13	0.17	0.14	0.09	0.09	0.04	0.06	0.03	0.00
2		0.17	0.21		0.12	0.11		0.08	0.03
4			0.18			0.13			0.08

Table 5.2: Maximum absolute cross-correlations for wavelet coefficients between scales for an FD process with $\delta = 0.45$ using the Haar, D(4) and LA(8) wavelet filters.

unit lag correlations within levels are somewhat larger for the D(4) and LA(8) wavelets than for the Haar, wavelets of greater width than the Haar lead to a decrease in maximum absolute correlation between levels.

We can gain additional insight into the decorrelation properties of the DWT by noting the frequency domain equivalent of (5.2), namely,

$$\operatorname{cov}\{W_{j,n}, W_{j',n'}\} = \int_{-1/2}^{1/2} e^{i2\pi f(2^j(n+1)-2^{j'}(n'+1))} H_j(f) H_{j'}^*(f) S_X(f)\, df, \quad (5.3)$$

where $H_j(\cdot)$ is the transfer function for the jth level wavelet filter:

$$H_j(f) \equiv \sum_{l=0}^{L_j-1} h_{j,l} e^{-i2\pi fl}.$$

When $j = j'$ and $n' = n + \tau$, we obtain the frequency domain equivalent of (5.1):

$$\operatorname{cov}\{W_{j,n}, W_{j,n+\tau}\} = \int_{-1/2}^{1/2} e^{i2^{j+1}\pi f\tau} \mathcal{H}_j(f) S_X(f)\, df, \quad (5.4)$$

where $\mathcal{H}_j(f) \equiv |H_j(f)|^2$. A jth level wavelet filter has a nominal pass-band given by $|f| \in [\frac{1}{2^{j+1}}, \frac{1}{2^j}]$. We can thus argue that the above should be approximately zero for $\tau \neq 0$ when $S_X(\cdot)$ is approximately constant over this

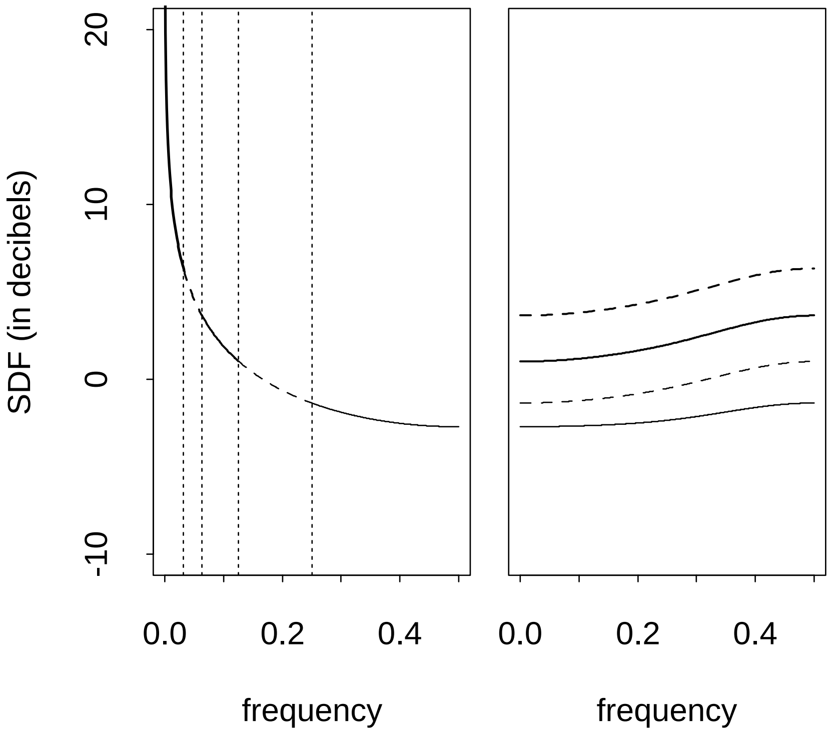


Figure 5.1: SDFs for an FD process with $\delta = 0.45$ (left-hand plot) and for the corresponding nonboundary LA(8) wavelet coefficients in $\mathbf{W}_j$, $j = 1, 2, 3, 4$ (bottom to top curves in the right-hand plot). The vertical dotted lines mark the beginning of the nominal pass-bands $[\frac{1}{2^{j+1}}, \frac{1}{2^j}]$ for $\mathbf{W}_j$.

pass-band. An alternative formulation is to note that

$$\operatorname{cov}\{W_{j,n}, W_{j,n+\tau}\} = \int_{-1/2}^{1/2} e^{i2\pi f\tau} S_j(f)\, df,$$

where

$$S_j(f) \equiv \frac{1}{2^j} \sum_{k=0}^{2^j-1} \mathcal{H}_j(\tfrac{f+k}{2^j}) S_X(\tfrac{f+k}{2^j}); \tag{5.5}$$

i.e., if we ignore boundary coefficients, then $W_{j,n}$ can be regarded as a portion of a stationary process with SDF $S_j(\cdot)$, so $W_{j,n}$ will be approximately white noise if $S_j(\cdot)$ is approximately constant. As an example, the left-hand panel of Figure 5.1 shows the SDF $S_X(\cdot)$ on a decibel (dB) scale for an FD process with $\delta = 0.45$. The right-hand panel shows, from bottom to top, $S_j(f), j = 1, \ldots, 4$, based upon an LA(8) wavelet. We see that, in contrast to $S_X(\cdot)$, the SDFs for the wavelet coefficients have a quite limited range of variation (less than 3 dB, i.e., a factor of two). We can also see why the DWT is so well-suited for SDFs for FD processes: as $S_X(f)$ diverges to infinity with decreasing f, the widths of the nominal pass-bands decrease commensurately so that $S_X(\cdot)$ does not vary much over any given pass-band.

Finally, let us consider the exact covariance matrix $\Sigma_{\mathbf{W}}$ for all the DWT coefficients $\mathbf{W}$ (this allows us to examine covariances involving the scaling and boundary wavelet coefficients). Let $\Sigma_{\mathbf{X}}$ be the covariance matrix for $\mathbf{X}$ (because of stationarity, its (j,k)th element is $s_{X,j-k}$). Since $\mathbf{W} = \mathcal{W}\mathbf{X}$, we have $\Sigma_{\mathbf{W}} = \mathcal{W}\Sigma_{\mathbf{X}}\mathcal{W}^T$ (note that the elements of $\mathcal{W}$ can be deduced from (4.1)). The top row of Figure 5.2 depicts the corresponding correlation matrix for level $J_0 = 6$ Haar, D(4) and LA(8) DWTs when $\mathbf{X}$ consists of a portion of size $N = 256$ from an FD process with $\delta = 0.45$. These plots show a grey-scale coding of the magnitudes of the elements of the correlation matrices after setting the diagonal elements to zero (these elements are unity by definition and dominate the off-diagonal elements). Let us focus first on the Haar case (upper left-hand corner). The dotted vertical and horizontal lines delineate the portions of the correlation matrix involving the DWT coefficients of different scales. As we go from the upper left-hand to lower right-hand corners, we pass along the diagonals of square submatrices that involve correlations within a given scale. The faint diagonal within each of these submatrices is primarily due to the lag one autocorrelations (see the values in the second column of Table 5.1, which can be used to gauge the magnitudes depicted in the plot). The faint lines going between opposite corners of the off-diagonal (nonsquare) submatrices are due to correlations between scales (Table 5.2 lists the largest such magnitudes). If we compare this plot with the corresponding plots for the D(4) and LA(8) DWTs, we see that the square submatrices are roughly the same (indicating that the autocorrelations within a scale are similar for the three wavelets), but that the lines going between opposite corners of the nonsquare submatrices are fainter (indicating a decrease in correlations between scales as L increases). Additionally, there are some dark points in these latter two plots that tend to line up horizontally and vertically. These are attributable to the scaling and boundary wavelet coefficients and are seen to be relatively few in number (the Haar DWT is free of boundary coefficients). Finally the dark spot in the lower right-hand corner of all three plots is due to the four scaling coefficients, which are highly autocorrelated for an FD process. The overall impression that the top row of plots gives is that the three DWTs do a credible job of decorrelating the highly autocorrelated FD process.

We can thus bootstrap a time series via its DWT using the following steps.

1. Given a time series $\mathbf{X}$ of length 2^J, compute a level $J_0 = J - 2$ DWT to obtain the wavelet coefficient vectors $\mathbf{W}_1, \ldots, \mathbf{W}_{J_0}$ and the scaling coefficient vector $\mathbf{V}_{J_0}$. This recipe for setting J_0 yields four coefficients each in $\mathbf{W}_{J_0}$ and $\mathbf{V}_{J_0}$ — decreasing J_0 has the effect of giving us more

3. Apply the inverse DWT to $\mathbf{W}_1^{(b)}, \ldots, \mathbf{W}_{J_0}^{(b)}$ and $\mathbf{V}_{J_0}^{(b)}$ to obtain the bootstrapped time series $\mathbf{X}^{(b)}$, for which we can then compute our statistic of interest, i.e., the unit lag sample autocorrelation $\hat{\rho}_{X,1}^{(b)}$ obtained from (1.1) with X_t replaced by $X_t^{(b)}$.

By repeating the above over and over again, we can build up a sample distribution of bootstrapped autocorrelations, which we use as a surrogate for the distribution of the actual sample autocorrelation.

Let us comment on a variation of the above scheme. As noted before, the DWT treats a time series as if it were circular. This aspect of the DWT can be problematic for a long memory series, for which there can be a large discrepancy between X_0 and X_{N-1}. Greenhall *et al.* (1999) provide evidence that an effective way to get around this difficulty for long memory processes is to replace $\mathbf{X}$ by a series of length $2N$ created by tacking on a time-reversed version of $\mathbf{X}$ to itself:

$$\mathbf{X}_{(c)} \equiv [X_0, X_1, \ldots, X_{N-2}, X_{N-1}, X_{N-1}, X_{N-2}, \ldots, X_1, X_0]^T.$$

We then use the DWT of this circularized series to form our bootstrapped series $\mathbf{X}_{(c)}^{(b)}$, from which we extract the first N elements to compute the sample autocorrelation. We refer to using the DWT on $\mathbf{X}_{(c)}$ rather than $\mathbf{X}$ as using reflection — rather than periodic — boundary conditions.

To see how well DWT-based bootstrapping works, see Table 7.1, which reports the results of a Monte Carlo study described in detail in §7. The bottom quarter of this table shows how well the standard deviations of the DWT-based bootstrapped $\hat{\rho}_{X,1}^{(b)}$ (under the column labeled 'DWT') compare with the actual standard deviation for $\hat{\rho}_{X,1}$ (under 'True'). Here we looked at time series of length $N = 128$ and 1024 that are realizations of an FD process with $\delta = 0.45$, and we used the LA(8) DWT with both periodic and reflection boundary conditions. We also report results for the block bootstrap (under the 'Block' column). We see that, while the DWT-based bootstrap tends to underestimate the true standard deviation by about 15% and 10% using, respectively, periodic and reflection boundary conditions, it is an improvement on the block bootstrap, which underestimates by about 30%.

6 Wavestrapping time series

While DWT-based bootstrapping works reasonably well for long memory FD processes, the question arises as to whether we can expect it to be useful for other processes. As simple examples, let us consider realizations of the AR(1) process $X_t = 0.9X_{t-1} + \epsilon_t$ and the MA(1) process $X_t = \epsilon_t + 0.99\epsilon_{t-1}$. The correlation matrices for the DWT coefficients $\mathbf{W}$ are shown in the middle and bottom rows of Figure 5.2 for, respectively, AR(1) and MA(1) series of length

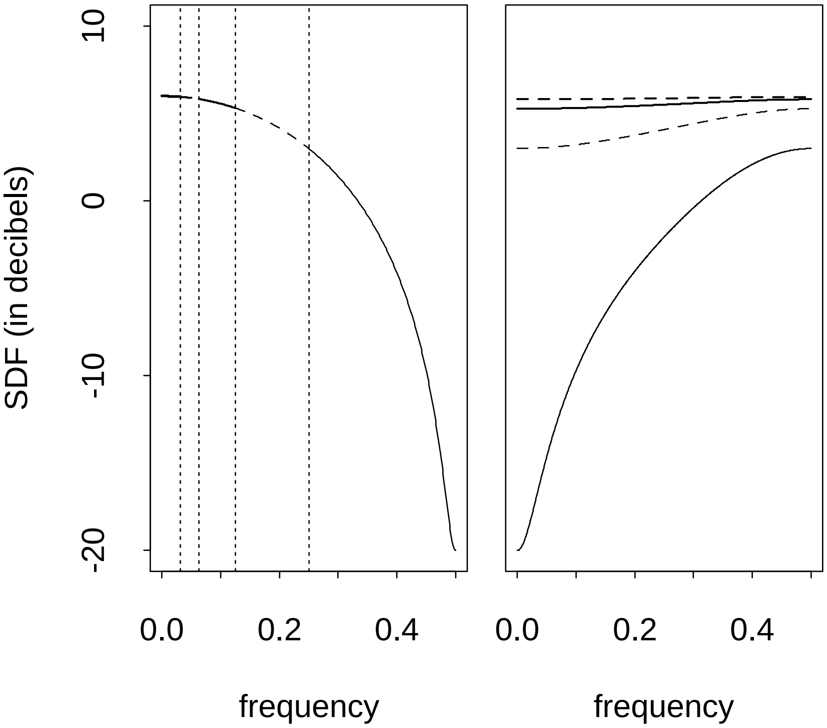


Figure 6.1: As in Figure 5.1, but now for an MA(1) process with $\theta = -0.99$.

$N = 256$. When compared to the FD case in the first row, we see higher levels of correlation, particularly within scale $j = 1$ for the MA(1) process. Figure 6.1 shows the SDF for this process, along with the SDFs $S_j(\cdot)$ for the nonboundary LA(8) wavelet coefficients in $\mathbf{W}_1, \ldots, \mathbf{W}_4$. The MA(1) SDF has considerable variation within the nominal pass-band $[\frac{1}{4}, \frac{1}{2}]$ for $\mathbf{W}_1$, which leads to $S_1(\cdot)$ being a poor approximation to white noise, thus explaining the high levels of correlation within scale $j = 1$.

Let us attempt to correct the poor decorrelating properties of the DWT in cases like the MA(1) process by considering a generalization of the DWT based upon adaptively picking out a transform from a level J_0 wavelet packet (WP) table (details on how to compute WP tables can be found in, e.g., Wickerhauser, 1994, Bruce and Gao, 1996, and Percival and Walden, 2000). Figure 6.2 shows an example of such a table. The jth row of the table is composed of 2^j vectors $\mathbf{W}_{j,n}$ $n = 0, \ldots, 2^j - 1$. Each vector has $N_j = N/2^j$ elements, and collectively all 2^j vectors form the coefficients for a jth level discrete wavelet packet transform (DWPT). Like the DWT, a DWPT is an orthonormal transform from which we can recover $\mathbf{X}$; moreover, the transform

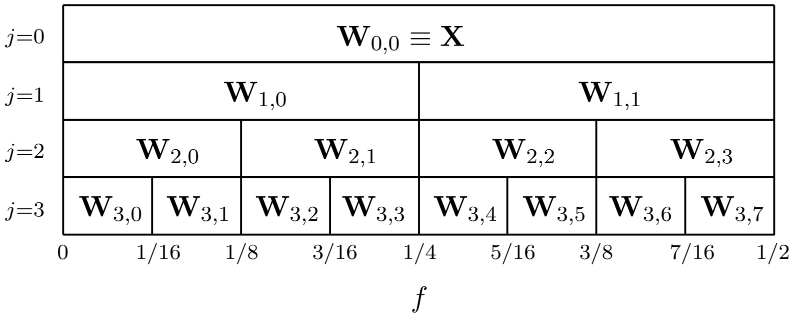


Figure 6.2: Wavelet packet table of order $J_0 = 3$ and associated pass-bands. The jth row of the table contains the subvectors $\mathbf{W}_{j,n}$ of a jth level DWPT.

can be formulated as filtering operations involving a wavelet and a scaling filter. The filter that yields the nth subvector has a nominal pass-band given by $[\frac{n}{2^{j+1}}, \frac{n+1}{2^{j+1}}]$. When taken together, the 2^j pass-bands partition the interval $[0, \frac{1}{2}]$ into 2^j intervals of equal length. Figure 6.2 shows the subvectors $\mathbf{W}_{j,n}$ for levels $j = 1, 2$ and 3 enclosed by rectangles spanning the nominal pass-bands. For convenience, we define $\mathbf{W}_{0,0} = \mathbf{X}$ so that the time series itself is associated with a 'zeroth level' DWPT (i.e., the identity transform) covering the entire frequency band.

The collection of DWPT coefficients for levels $j = 0, \ldots, J_0$ forms a level J_0 WP table. The vertical stacking of coefficients in the figure tells us how coefficients from DWPTs of different levels are related: given a subvector $\mathbf{W}_{j,n}$ of level j, we obtain the subvectors $\mathbf{W}_{j+1,2n}$ and $\mathbf{W}_{j+1,2n+1}$ of level $j + 1$ via an orthonormal transform (one subvector is formed using the wavelet filter, and the other, the scaling filter, but the order in which these get used depends upon n). Thus, as depicted in the table, we can obtain $\mathbf{W}_{3,4}$ and $\mathbf{W}_{3,5}$ via an orthonormal transform of $\mathbf{W}_{2,2}$. We can extract a large number of different orthonormal transforms from a WP table. For example, if we start with a jth level DWPT, we can obtain 2^{2^j} different transforms by choosing either to keep each $\mathbf{W}_{j,n}$ or to transform it into the two subvectors $\mathbf{W}_{j+1,2n}$ and $\mathbf{W}_{j+1,2n+1}$. We can obtain even more transforms by keeping or splitting across more than two levels. In fact a DWT of level J_0 is one such transform, consisting of $\mathbf{W}_1 = \mathbf{W}_{1,1}$, $\mathbf{W}_2 = \mathbf{W}_{2,1}$, $\ldots$, $\mathbf{W}_{J_0} = \mathbf{W}_{J_0,1}$ and $\mathbf{V}_{J_0} = \mathbf{W}_{J_0,0}$.

With so many different transforms at our disposal, a careful selection of coefficients from the WP table can lead to an orthonormal transform that partitions the frequency interval $[0, \frac{1}{2}]$ into subintervals such that, within each subinterval, the SDF for $\mathbf{X}$ does not vary much. Given knowledge of the SDF $S_X(\cdot)$ and a stopping level J_0, we can adaptively select a transform by starting

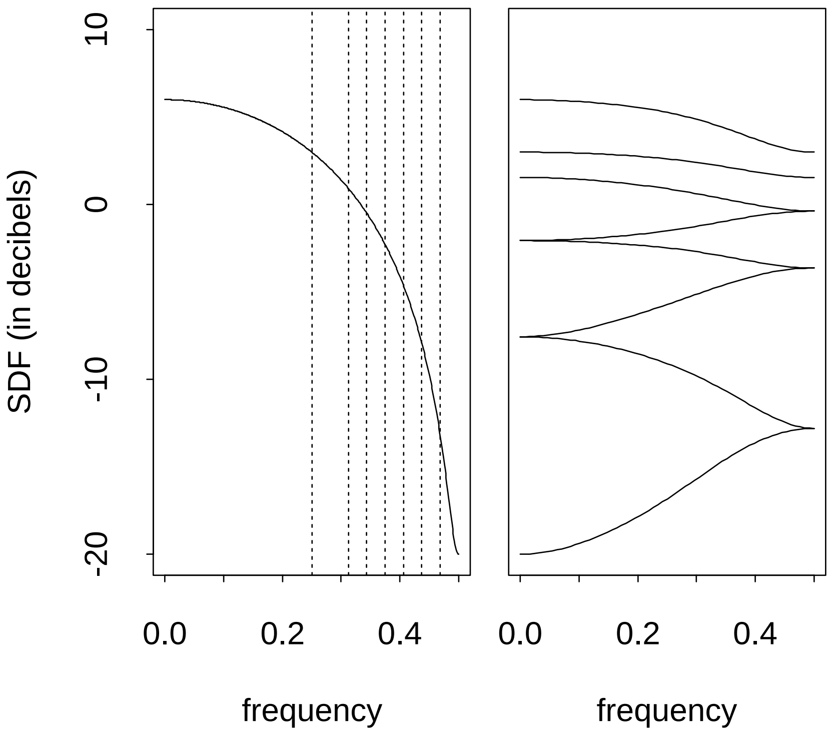


Figure 6.3: Transform selected from an LA(8) WP table of level $J_0 = 4$ that converts the MA(1) process with $\theta = -0.99$ into approximately uncorrelated coefficients. The transform consists of the eight subvectors $\mathbf{W}_{1,0}$, $\mathbf{W}_{3,4}$, $\mathbf{W}_{4,10}, \ldots, \mathbf{W}_{4,15}$, which partition $[0, \frac{1}{2}]$ into the nominal pass-bands shown by the vertical lines in the left-hand plot (the solid curve is the SDF for the MA(1) process). The corresponding SDFs for the subvectors are shown from top to bottom in the right-hand plot. The first five of these SDFs have variations less than 3 dB, while the SDFs for $\mathbf{W}_{4,13}$, $\mathbf{W}_{4,14}$ and $\mathbf{W}_{4,15}$ vary by, respectively, 3.9, 5.3 and 7.2 dB. If we were to increase the level to $J_0 = 6$, these three subvectors would be replaced by three $j = 5$ level subvectors $\mathbf{W}_{5,26}, \mathbf{W}_{5,27}, \mathbf{W}_{5,28}$ and six $j = 6$ level subvectors $\mathbf{W}_{6,58}, \ldots, \mathbf{W}_{6,63}$, all of whose SDFs vary by less than 3 dB.

with $\mathbf{W}_{0,0} = \mathbf{X}$ and recursively applying the following simple rule. If the level of $\mathbf{W}_{j,n}$ is J_0, we retain it; otherwise, we consider the SDF associated with the nonboundary coefficients in $\mathbf{W}_{j,n}$ (this SDF can be computed via an equation analogous to (5.5)). If the SDF varies no more than, say, 3 dB, then we retain $\mathbf{W}_{j,n}$; otherwise, we replace $\mathbf{W}_{j,n}$ by $\mathbf{W}_{j+1,2n}$ and $\mathbf{W}_{j+1,2n+1}$, and then apply the simple rule to both of these vectors. Figure 6.3 shows a transform from a fourth level WP table that is adapted to the MA(1) process.

In practice, of course, we do not know $S_X(\cdot)$, so we propose to replace the 3 dB criterion with a statistical test for the null hypothesis that the values in $\mathbf{W}_{j,n}$ are a sample from a white noise process (there are a number of appropriate test statistics in the literature, two of which we describe briefly in §6.1 below). We can now outline the steps needed to create 'wavestrap' samples of a time series.

1. Given a time series $\mathbf{X}$ of length 2^J, compute a level $J_0 = J - 2$ WP table. Enter step 2 with starting values $j = n = 0$ and $\mathbf{W}_{0,0} \equiv \mathbf{X}$.

2. If $j = J_0$, retain $\mathbf{W}_{j,n}$; if $j < J_0$, perform a white noise test on $\mathbf{W}_{j,n}$ using one of the test statistics given in §6.1. If we fail to reject the null hypothesis, then retain $\mathbf{W}_{j,n}$; if we reject, then discard $\mathbf{W}_{j,n}$ (after transforming it into $\mathbf{W}_{j+1,2n}$ and $\mathbf{W}_{j+1,2n+1}$), and repeat this step twice again, once on $\mathbf{W}_{j+1,2n}$, and once on $\mathbf{W}_{j+1,2n+1}$.

3. The desired adaptively chosen transform consists of all the subvectors that are retained after step 2 has been applied as many times as needed. Randomly sample (with replacement) from each of the subvectors in the transform to create the similarly dimensioned wavestrapped subvectors.

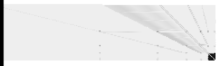
4. Apply the inverse of the adaptively chosen transform to the wavestrapped subvectors to obtain the wavestrapped time series, for which we can then compute, e.g., a unit lag sample autocorrelation.

As was the case for the DWT-based bootstrap, we repeat the last two steps above over and over again to build up a sample distribution of wavestrapped autocorrelations.

Figure 6.4 shows the correlation matrices for the wavestrap transforms picked out for a single realization of the same three processes considered in Figure 5.2 for DWT-based bootstrapping. A comparison of the largest correlations in the corresponding plots of these two figures shows that, by this measure, wavestrapping does better than the DWT-based procedure for the MA(1) process, is about the same for the AR(1) process, and, not unexpectedly, is worse for the FD process, which is well-matched to the DWT. The dark squares in the lower right-hand corners in the top row are due to $\mathbf{W}_{j,0}$, which are highly correlated for an FD process.

6.1 White Noise Tests

Here we briefly describe two well-known test statistics that can be used to assess the null hypothesis that the WP coefficients $\mathbf{W}_{j,n}$ are a sample from a white noise process.



the literature). For $0 < \tau < N_j$, we define the sample autocorrelation to be

$$\hat{\rho}_{j,n,\tau} = \frac{\sum_{t=0}^{N_j-1-\tau} W_{j,n,t} W_{j,n,t+\tau}}{\sum_{t=0}^{N_j-1} W_{j,n,t}^2}.$$

There are three variations on the portmanteau test in the literature. The Box–Pierce test statistic and Ljung–Box–Pierce test statistic are respectively

$$Q_{j,n} = N_j \sum_{\tau=1}^{K} \hat{\rho}_{j,n,\tau}^2 \text{ and } \widetilde{Q}_{j,n} = N_j(N_j+2) \sum_{\tau=1}^{K} \frac{\hat{\rho}_{j,n,\tau}^2}{N_j - \tau}$$

(Box and Pierce, 1970; Ljung and Box, 1978). For either test statistic, we reject the null hypothesis of white noise at significance level α when the statistic exceeds the $(1-\alpha) \times 100\%$ percentage point $Q_K(1-\alpha)$ for the chi-square distribution with K degrees of freedom. The third variation (McLeod and Li, 1983; Brockwell and Davis, 1991, §9.4) is to use the Ljung–Box–Pierce test on the sample autocorrelations for the squares of $W_{j,n,t}$, namely,

$$\hat{\rho}_{j,n,\tau}^{[2]} = \frac{\sum_{t=0}^{N_j-1-\tau} (W_{j,n,t}^2 - \overline{W^2}_{j,n})(W_{j,n,t+\tau}^2 - \overline{W^2}_{j,n})}{\sum_{t=0}^{N_j-1} (W_{j,n,t}^2 - \overline{W^2}_{j,n})^2},$$

where $\overline{W^2}_{j,n}$ is the sample mean of the squares of the elements of $\mathbf{W}_{j,n}$.

6.1.2 Cumulative Periodogram Test

Let $|\widetilde{W}_{j,n,k}|^2$ be the squared modulus of the DFT of $\mathbf{W}_{j,n}$ at the Fourier frequency $f_k \equiv k/N_j$. Based upon the $M_j \equiv \frac{N_j}{2} - 1$ frequencies satisfying $0 < f_k < 1/2$, we form the normalized cumulative periodogram

$$\mathcal{P}_l \equiv \frac{\sum_{k=1}^{l} |\widetilde{W}_{j,n,k}|^2}{\sum_{k=1}^{M_j} |\widetilde{W}_{j,n,k}|^2}, \qquad l = 1, \ldots, M_j.$$

We then compute the test statistic $D \equiv \max\{D^+, D^-\}$, where

$$D^+ \equiv \max_{1 \le l \le M_j - 1} \left(\frac{l}{M_j - 1} - \mathcal{P}_l \right) \text{ and } D^- \equiv \max_{1 \le l \le M_j - 1} \left(\mathcal{P}_l - \frac{l-1}{M_j - 1} \right).$$

We reject the null hypothesis of white noise at the α level of significance if D exceeds the upper $\alpha \times 100\%$ percentage point $D(\alpha)$ for D under the null hypothesis. A simple approximation for $D(\alpha)$ is given by

$$\widetilde{D}(\alpha) \equiv \frac{C(\alpha)}{(M_j-1)^{1/2} + 0.12 + \frac{0.11}{(M_j-1)^{1/2}}},$$

where $C(0.05) = 1.358$ (Stephens, 1974).

7 Simulation study

Here we report on a Monte Carlo experiment that we conducted to see how well the DWT-based bootstrap, wavestrapping and the block bootstrap do at assessing the standard deviation of the unit lag sample autocorrelation $\hat{\rho}_{X,1}$ for Gaussian white noise and the three nonwhite processes considered in previous sections, namely, an FD process with $\delta = 0.45$, an AR(1) process with $\phi = 0.9$ and an MA(1) process with $\theta = -0.99$. We used a pseudo-random number generator of uncorrelated Gaussian deviates ϵ_t with mean zero and unit variance to simulate the white noise process. Using the same generator, we can simulate

- AR(1) time series by setting $X_0 = (\frac{1}{1-0.9^2})^{1/2}\epsilon_0$ and $X_t = 0.9X_{t-1} + \epsilon_t$, $t = 1, \ldots, N-1$ (Kay, 1981);

- MA(1) series by using the process definition $X_t = \epsilon_t + 0.99\epsilon_t$

		Wavestrap				
Process	Boundary	DWT	Port	Pgrm	Block	True
WN						
$N = 128$	periodic	8.2	8.7	8.8	8.1	8.7
	reflection	8.3	8.6	8.7		
$N = 1024$	periodic	3.1	3.1	3.1	3.0	3.1
	reflection	3.2	3.2	3.1		
AR(1)						
$N = 128$	periodic	5.7	5.2	5.1	5.4	4.8
	reflection	5.5	5.1	5.4		
$N = 1024$	periodic	1.6	1.5	1.5	1.5	1.4
	reflection	1.6	1.5	1.5		
MA(1)						
$N = 128$	periodic	7.1	6.8	6.8	6.5	6.3
	reflection	7.0	6.8	6.6		
$N = 1024$	periodic	2.6	2.4	2.3	2.2	2.2
	reflection	2.6	2.4	2.4		
FD						
$N = 128$	periodic	9.4	8.3	8.5	7.7	10.7
	reflection	9.9	8.8	9.6		
$N = 1024$	periodic	4.4	4.2	4.2	3.4	5.3
	reflection	4.7	4.5	4.7		

Table 7.1: Standard deviations ($\times 100$) of unit lag sample autocorrelations as assessed by DWT-based bootstrapping, wavestrapping with the Ljung–Box–Pierce portmanteau test statistic, wavestrapping with the cumulative periodogram test statistic, and the block bootstrap, along with the 'true' standard deviation as determined by 10,000 simulated series. Independent replications indicate that the standard error for all numbers reported above is roughly 0.1.

is quite close to the true value. On other hand, block bootstrapping is inferior to the other techniques for the FD process. Reflection boundary conditions work better with both DWT-based bootstrapping and wavestrapping, and the cumulative periodogram test statistic is better with the latter than the portmanteau statistic. While DWT-based bootstrapping and wavestrapping yield similar results, they both underestimate the true standard deviation by about 10%.

When we decrease the sample size to $N = 128$, there is more disparity among the four methods. Wavestrapping outperforms the DWT-based bootstrap for the three short memory processes, but the converse is true for the FD process. With the exception of the MA(1) process, wavestrapping also does better than

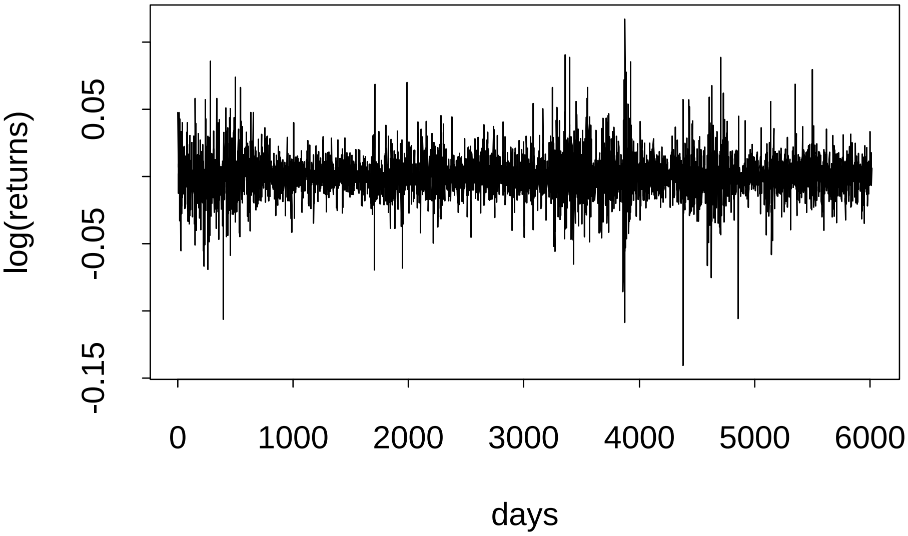


Figure 8.1: Log of daily returns on BMW share prices (1973–96).

the block bootstrap. Finally, we note that the cumulative periodogram test and reflection boundary conditions generally do better with wavestrapping than the portmanteau test and periodic boundary conditions.

8 Applications

We now apply our methodology to a series X_t of $N = 6016$ daily log returns on BMW share prices between 1973 and 1996; see Figure 8.1. This time series is actually irregularly sampled because no trading takes place on weekends and holidays, but we ignore these gaps and treat the data as a regularly sampled series. The unit lag sample autocorrelation is small, $\hat{\rho}_{X,1} \doteq 0.081$. If we apply the standard large sample theory appropriate for Gaussian white noise, we would attach to this estimate a standard error of $1/\sqrt{N} \doteq 0.013$. In fact the Gaussian assumption is suspect, and the data are better modeled by a t distribution with 3.9 degrees of freedom. Taken at face value, however, the standard error tells us that although small, $\hat{\rho}_{X,1}$ is significantly different from zero; this could presumably be exploited by traders. When we apply the block bootstrap with blocks of length 30, 50, 100, 200 and 500, the standard errors are 0.012, 0.012, 0.014, 0.016 and 0.015, while the DWT-based bootstrap and the wavestrap give 0.023 and 0.020. Though all are larger than the value 0.013 for Gaussian white noise, these confirm the presence of autocorrelation. Simulation using blocks of t_4 innovations with variances 1, 4, 9 and 16, to give the type of stochastic volatility seen in Figure 8.1, gives values of $\hat{\rho}_{X,1}$ whose

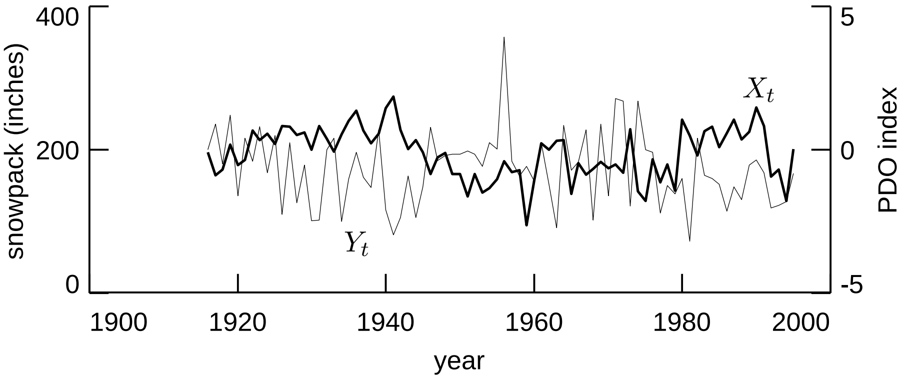


Figure 8.2: Pacific decadal oscillation (PDO) index X_t (thick curve, right-hand axis) and March 15 snow depth at Paradise Ranger Station (1600 meters above sea level) on Mt. Rainer Y_t (thin curve, left-hand axis). Both time series have one value per year from 1916 to 1995.

standard error is 0.02. It seems that the DWT and the wavestrap are able to reproduce this, but that the block bootstrap is not.

As a second example, let us show how wavestrapping can help assess the significance of the sample cross-correlation between two time series. This statistic is often used in the physical sciences as a first step in investigating possible relationships between two series. Figure 8.2 shows two annual time series of interest in atmospheric science, namely, the Pacific decadal oscillation (PDO) index X_t (thick curve) and springtime snow depth Y_t at a location in the Washington Cascade Mountains (thin). The PDO index (Mantua *et al.*, 1997) is the leading principal component of sea-surface temperature over the extratropical north Pacific ocean and has been implicated as a major source of interannual variability in temperature and precipitation in western North America. When the PDO index is positive (corresponding to cold sea-surface temperatures in the central Pacific Ocean and warm temperatures off the west coast of North America), an observational record of nearly a century suggests that the mean winter time temperature tends to be high, while precipitation tends to be low. The data shown in Figure 8.2 support this statement, as does the fact that the sample correlation coefficient between X_t and Y_t is negative:

$$\hat{\rho}_{XY} \equiv \frac{\sum_{t=0}^{N-1}(X_t - \overline{X})(Y_t - \overline{Y})}{\left[\sum_{t=0}^{N-1}(X_t - \overline{X})^2 \sum_{t=0}^{N-1}(Y_t - \overline{Y})^2\right]^{1/2}} \doteq -0.27.$$

Were we to assume that all 80 observations were independent and normally distributed, confidence limits based upon large sample statistical theory would declare $\hat{\rho}_{XY}$ to be significantly different from zero at more than the 95% level.

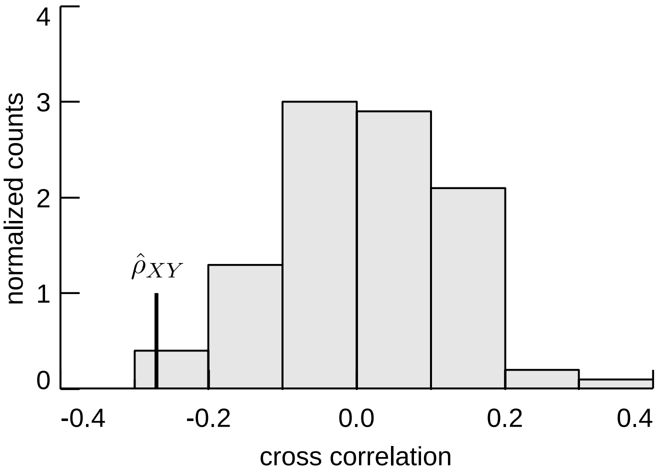


Figure 8.3: Histogram of wavestrapped cross-correlations to assess significance of $\hat{\rho}_{XY}$ computed for the two time series in Figure 8.2.

However, since both the PDO index and (to a lesser extent) snowpack have considerable year-to-year correlation and since both time series are short, we need another way of ascertaining if the sample cross-correlation is significantly different from zero.

We can address the question of the significance of $\hat{\rho}_{XY}$ by wavestrapping X_t and Y_t separately. The resulting wavestrapped series should be approximately pairwise uncorrelated because any relationship between the two series will be destroyed by resampling separately. The values of $\hat{\rho}_{XY}$ over many wavestrapped series will have a distribution reflecting a null ρdWya3.13ol u2g0i4s *W* . *s* . . *be* . *f* . *ske* . *f* ory processes and offers an improvement for long memory processes. While these results are promising, there is considerable work to be done to be put wavestrapping on a sound theoretical foundation. Questions that need to be addressed include the following.

1. For what kinds of statistics and processes can we expect wavestrapping to yield either a reasonable approximate distribution or a reasonable approximation to certain aspects of that distribution? For example, Monte Carlo experiments indicate that, whereas the standard deviation of the wavestrap distribution for $\hat{\rho}_{X,1}$ is a good approximation to the actual standard deviation, the same cannot be said for the bias.

2. What are the asymptotic properties of wavestrapping? Although we are really interested in small to moderate sample sizes, it would be of interest to know what conditions are needed for wavestrapping to be a consistent estimator.

3. Can wavestrapping handle non-Gaussian and/or nonlinear processes? The Gaussian assumption is a convenient starting point, but real-world applications dictate that we move beyond it.

4. Can we offer better guidance on the subjective aspects of wavestrap, namely, choice of wavelet and level J_0? The LA(8) wavelet and picking $J_0 = J - 2$ gave good results in our Monte Carlo study, but it is not clear if these would be good choices for other statistics and processes.

Even for statistics such as the sample ACS, there is room for improving the performance of wavestrap, particularly for long memory processes, where it tends to underestimate the variability in the sample autocorrelation. Two possible improvements to our work would be to combine wavestrapping with a parametric approach and to use a different procedure for picking out a decorrelating transform from a WP table. Let us close by briefly commenting on why we feel these to be worth studying.

If we compare the wavestrapping results in Table 7.1 for the AR and FD processes when $N = 1024$, we see somewhat better estimation of the true standard deviation for the AR case (a 7% overestimate as compared to an 11% underestimate in the FD case). If we focus on the DWT, we find that the between/across scale correlations of the wavelet coefficients for the AR and FD processes are quite similar to each other (in fact the correlations in the AR case are a little larger in magnitude). There is a big difference, however, in the properties of their scaling coefficients: for the AR process, the scaling coefficients are reasonably close to white noise (because the SDF flattens out as $f \to 0$), whereas they have a long memory structure for the FD process. This suggests that the underestimation of variability in the FD case is attributable to the scaling coefficients (note that any orthonormal transform we pick from a

WP table must include one subvector corresponding to the scaling coefficients from a DWT of some level $J' \leq J_0$). One way to account for the correlation in the scaling coefficients would be to use a parametric bootstrap. If we set J_0 to, say, $J-4$ rather than our standard choice of $J-2$, we would have at least sixteen scaling coefficients, which would be enough to entertain an AR(1) model. Although an AR(1) model is not a perfect match to the correlation properties of the scaling coefficients for an FD process, this simple model is capable of approximating the correlation structure over limited number of lags, which is really all we require. In addition, a study of the SDFs on the right-hand sides of Figures 5.1 and 6.3 suggests that the remaining correlation structure in wavelet and WP coefficients might be well-modeled by fitting an AR(1) process to each set of coefficients and then bootstrapping with respect to the fitted models. Limited tests suggest that this is a promising idea.

Finally, with regard to picking a decorrelating transform from a WP table, wavestrapping does its search through the table in a 'top-down' manner, so the obvious alternative to consider is a 'bottom-up' approach. A well-known example of such an approach is the 'best basis' algorithm (Coifman and Wickerhauser, 1992), which selects between a 'parent' node $\mathbf{W}_{j,n}$ and its 'children' $\mathbf{W}_{j+1,2n}$ and $\mathbf{W}_{j+1,2n+1}$ based upon a cost functional. To see why this algorithm leads to a decorrelating transform, let $j=1$ and $n=0$ for simplicity, and suppose that the nonboundary WP coefficients in the parent node have the following SDF:

$$S_{1,0}(f) = \begin{cases} \sigma^2_{2,0}, & 0 \leq |f| \leq 1/4; \\ \sigma^2_{2,1}, & 1/4 \leq |f| \leq 1/2. \end{cases}$$

The variance for a process with this SDF is $\sigma^2_{1,0} = \frac{1}{2}(\sigma^2_{2,0} + \sigma^2_{2,1})$. If we assume for simplicity that the scaling and wavelet filters are perfect high- and low-pass filters, the SDFs of the nonboundary WP coefficients in $\mathbf{W}_{2,0}$ and $\mathbf{W}_{2,1}$ are given by, respectively, $S_{2,0}(f) = \sigma^2_{2,0}$ and $S_{2,1}(f) = \sigma^2_{2,1}$ for $-1/2 \leq f \leq 1/2$; i.e., both are white noise processes with variances given by, respectively, $\sigma^2_{2,0}$ and $\sigma^2_{2,1}$ (note that $S_{1,0}(\cdot)$ is not a white noise SDF unless $\sigma^2_{2,0} = \sigma^2_{2,1}$). If we assume Gaussianity and, e.g., the L_1 cost functional, then the costs of each coefficient $W_{1,0,t}$ in $\mathbf{W}_{1,0}$ and of each coefficient $W_{2,m,t}$ in $\mathbf{W}_{2,m}$ are, respectively,

$$E\{|W_{1,0,t}|\} = \left(\frac{\sigma^2_{2,0} + \sigma^2_{2,1}}{2\pi}\right)^{1/2} \text{ and } E\{|W_{2,m,t}|\} = \left(\frac{2\sigma^2_{2,m}}{\pi}\right)^{1/2}.$$

Since there are N_1 coefficients in $\mathbf{W}_{1,0}$ and N_2 coefficients in each of $\mathbf{W}_{2,0}$ and $\mathbf{W}_{2,1}$, the total expected costs of the parent node and its children are thus, respectively,

$$C_1 \equiv N_1 \left(\frac{\sigma^2_{2,0} + \sigma^2_{2,1}}{2\pi}\right)^{1/2} \text{ versus } C_2 \equiv N_2 \left[\left(\frac{2\sigma^2_{2,0}}{\pi}\right)^{1/2} + \left(\frac{2\sigma^2_{2,1}}{\pi}\right)^{1/2}\right].$$

It is an easy exercise to verify that $C_2 \leq C_1$ always, with equality occurring if and only if $\sigma^2_{2,0} = \sigma^2_{2,1}$ (i.e., the nonboundary coefficients in $\mathbf{W}_{1,0}$ are white noise). Since the best basis algorithm works by making a comparison such as the above on each node, we can argue that this algorithm will tend to pick out a decorrelating transform. Tests to date, however, indicate that best basis picks out too many small groups of coefficients (not ideal for bootstrapping), so we are currently exploring ways of 'pruning' back the best basis transform.

10 Acknowledgments

We would like to thank Bill Fitzgerald, Richard Smith, Andrew Walden and Peter Young for the considerable efforts that they put into organizing the program on Nonlinear and Nonstationary Signal Processing at the Isaac Newton Institute. We would also like to thank Chris Bretherton (Department of Atmospheric Sciences, University of Washington) for providing us with the snowpack/PDO time series and for discussions concerning them. This work was supported in part by the Swiss National Science Foundation (Sardy and Davison) and by the U.S. National Science Foundation (Percival).

References

Bartlett, M.S. (1946) 'On the theoretical specification of sampling properties of autocorrelated time series,' *Supplement to the Journal of the Royal Statistical Society* **8**, 27–41.

Beran, J. (1994) *Statistics for Long-Memory Processes*, New York: Chapman & Hall.

Box, G.E.P., Pierce, D.A. (1970) 'Distribution of residual autocorrelations in autoregressive-integrated moving average time series models,' *Journal of the American Statistical Association* **65**, 1509–1526.

Bretherton, C.S. (1998) 'Effective Degrees of Freedom for Spatially and Temporally Correlated Data,' *Seventh International Meeting on Statistical Climatology*, Whistler, British Columbia.

Brockwell, P.J., Davis, R.A. (1991) *Time Series: Theory and Methods* (Second Edition), New York: Springer.

Bruce, A.G., Gao, H.–Y. (1996) *Applied Wavelet Analysis with S-PLUS*, New York: Springer–Verlag.

Bühlmann, P. (1999) 'Bootstrapping time series,' *Bulletin of the 52nd Session of the International Statistical Institute* **1**, 201–204.

Coifman, R.R., Wickerhauser, M.V. (1992) 'Entropy-based algorithms for best basis selection,' *IEEE Transactions on Information Theory* **38**, 713–718.

Daubechies, I. (1992) *Ten Lectures on Wavelets*, Philadelphia: SIAM.

Davies, R.B., Harte, D.S. (1987) 'Tests for Hurst Effect,' *Biometrika* **74**, 95–101.

Davison, A.C., Hinkley, D.V. (1997) *Bootstrap Methods and their Application*, Cambridge, UK: Cambridge University Press.

Efron, B., Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Flandrin, P. (1992) 'Wavelet analysis and synthesis of fractional Brownian motion,' *IEEE Transactions on Information Theory* **38**, 910–917.

Granger, C.W.J., Joyeux, R. (1980) 'An introduction to long-memory time series models and fractional differencing,' *Journal of Time Series Analysis* **1**, 15–30.

Greenhall, C.A., Howe, D.A., Percival, D.B. (1999) 'Total Variance, an Estimator of Long-Term Frequency Stability.' *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **46**, 1183–1191.

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

Kay, S.M. (1981) 'Efficient Generation of Colored Noise,' *Proceedings of the IEEE* **69**, 480–481.

Künsch, H.R. (1989) 'The jackknife and the bootstrap for general stationary observations,' *Annals of Statistics* **17**, 1217–1241.

Ljung, G.M., Box, G.E.P. (1978) 'On a measure of lack of fit in time series models,' *Biometrika* **65**, 297–303.

Mallat, S.G. (1989) 'A theory for multiresolution signal decomposition: the wavelet representation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M., Francis, R.C. (1997) 'A Pacific interdecadal climate oscillation with impacts on salmon production,' *Bulletin of the American Meteorological Society* **78**, 1069–1079.

McCoy, E.J., Walden, A.T. (1996) 'Wavelet analysis and synthesis of stationary long-memory processes,' *Journal of Computational and Graphical Statistics* **5**, 26–56.

McLeod, A.I., Li, W.K. (1983) 'Diagnostic checking ARMA time series models using squared-residual autocorrelations,' *Journal of Time Series Analysis* **4**, 269–273.

Percival, D.B., Walden, A.T. (2000) *Wavelet Methods for Time Series Analysis*, Cambridge, UK: Cambridge University Press.

Priestley, M.B. (1981) *Spectral Analysis and Time Series*, London: Academic Press.

Stephens, M.A. (1974) 'EDF Statistics for Goodness of Fit and Some Comparisions,' *Journal of the American Statistical Association* **69**, 730–737.

Wickerhauser, M.V. (1994) *Adapted Wavelet Analysis from Theory to Software*, Wellesley, MA: AK Peters.

Wood, A.T.A., Chan, G. (1996) 'Simulation of Stationary Gaussian Processes in $[0,1]^d$,' *Journal of Computational and Graphical Statistics* **3**, 3, 409–432.

Wornell, G.W. (1995) *Signal Processing with Fractals: A Wavelet Based Approach*, Englewood Cliffs, NJ: Prentice Hall.