Wavelet Methods for Time Series Analysis

Part II: Wavelet-Based Statistical Analysis of Time Series

- topics to covered:
  - wavelet variance (analysis phase of MODWT)
  - wavelet-based signal extraction (synthesis phase of DWT)
  - wavelet-based decorrelation of time series (analysis phase of DWT, but synthesis phase plays a role also)

Decomposing Sample Variance of Time Series

- let \( X_0, X_1, \ldots, X_{N-1} \) represent time series with \( N \) values
- let \( \bar{X} \) denote sample mean of \( X_t \)'s: \( \bar{X} \equiv \frac{1}{N} \sum_{t=0}^{N-1} X_t \)
- let \( \sigma^2_X \) denote sample variance of \( X_t \)'s:
  \[
  \sigma^2_X = \frac{1}{N} \sum_{t=0}^{N-1} (X_t - \bar{X})^2
  \]
- idea is to decompose (analyze, break up) \( \sigma^2_X \) into pieces that quantify how one time series might differ from another
- wavelet variance does analysis based upon differences between (possibly weighted) adjacent averages over scales

Wavelet Variance: Overview

- review of decomposition of sample variance using wavelets
- theoretical wavelet variance for stochastic processes
  - stationary processes
  - nonstationary processes with stationary differences
- sampling theory for Gaussian processes
- real-world examples
- extensions and summary

Empirical Wavelet Variance

- define empirical wavelet variance for scale \( \tau_j \equiv 2^j \) as
  \[
  \hat{\nu}^2_X(\tau_j) = \frac{1}{N} \sum_{t=0}^{N-1} \hat{W}_{jt,t}^2, \text{ where } \hat{W}_{jt,t} \equiv \sum_{l=0}^{L_j-1} \hat{h}_{j,t} X_{t-l \mod N}
  \]
- if \( N = 2^J \), obtain analysis (decomposition) of sample variance:
  \[
  \hat{\sigma}^2_X = \frac{1}{N} \sum_{t=0}^{N-1} (X_t - \bar{X})^2 = \sum_{j=1}^{J} \hat{\nu}^2_X(\tau_j)
  \]
  (if \( N \) not a power of 2, can analyze variance to any level \( J_0 \), but need additional component involving scaling coefficients)
- interpretation: \( \hat{\nu}^2_X(\tau_j) \) is portion of \( \hat{\sigma}^2_X \) due to changes in averages over scale \( \tau_j \); i.e., 'scale by scale' analysis of variance
Example of Empirical Wavelet Variance

- Wavelet variances for time series $X_t$ and $Y_t$ of length $N = 16$, each with zero sample mean and same sample variance

\[
\begin{align*}
X_t & = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \\
Y_t & = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T
\end{align*}
\]

Theoretical Wavelet Variance: I

- Now assume $X_t$ is a real-valued random variable (RV)
- Let $\{X_t, t \in \mathbb{Z}\}$ denote a stochastic process, i.e., collection of RVs indexed by ‘time’ $t$ (here $\mathbb{Z}$ denotes the set of all integers)
- Apply $j$th level equivalent MODWT filter $\{\tilde{h}_{j,l}\}$ to $\{X_t\}$ to create a new stochastic process:

\[
W_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}, \quad t \in \mathbb{Z},
\]

which should be contrasted with

\[
\tilde{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \mod N}, \quad t = 0, 1, \ldots, N-1
\]

Definition of a Stationary Process

- If $U$ and $V$ are two RVs, denote their covariance by

\[
\text{cov} \{U, V\} = E\{(U - E(U))(V - E(V))\}
\]

- Stochastic process $X_t$ called stationary if

\begin{itemize}
  \item $E\{X_t\} = \mu_X$ for all $t$, i.e., constant independent of $t$
  \item $\text{cov}\{X_t, X_{t+\tau}\} = s_{X,\tau}$, i.e., depends on lag $\tau$, but not $t$
\end{itemize}

- $s_{X,\tau}, \tau \in \mathbb{Z}$, is autocovariance sequence (ACVS)
- $s_{X,0} = \text{cov}\{X_t, X_t\} = \text{var}\{X_t\}$; i.e., variance same for all $t$
Wavelet Variance for Stationary Processes

- for stationary processes, wavelet variance decomposes $\text{var} \{X_t\}$:
  $$\sum_{j=1}^{\infty} \nu_{X}^2(\tau_j) = \text{var} \{X_t\},$$
  which is similar to
  $$\sum_{j=1}^{\infty} \nu_{X}^2(\tau_j) = \sigma_X^2$$

- $\nu_{X}^2(\tau_j)$ is thus contribution to $\text{var} \{X_t\}$ due to scale $\tau_j$

- note: $\nu_{X}^2(\tau_j)$ and $X_t^2$ have same units (can be important for interpretability)

White Noise Process

- simplest example of a stationary process is ‘white noise’
- process $X_t$ said to be white noise if
  - it has a constant mean $E\{X_t\} = \mu_X$
  - it has a constant variance $\text{var} \{X_t\} = \sigma_X^2$
  - $\text{cov} \{X_t, X_{t+\tau}\} = 0$ for all $t$ and nonzero $\tau$; i.e., distinct RVs in the process are uncorrelated
- ACVS for white noise takes a very simple form:
  $$s_{X,\tau} = \text{cov} \{X_t, X_{t+\tau}\} = \begin{cases} \sigma_X^2, & \tau = 0; \\ 0, & \text{otherwise}. \end{cases}$$

Wavelet Variance for White Noise Process: I

- for a white noise process, can show that
  $$\nu_{X}^2(\tau_j) = \frac{\text{var} \{X_t\}}{2^j} \propto \tau_j^{-1} \text{ since } \tau_j = 2^{j-1}$$
- note that
  $$\sum_{j=1}^{\infty} \nu_{X}^2(\tau_j) = \text{var} \{X_t\} \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \right) = \text{var} \{X_t\},$$
  as required
- note also that
  $$\log (\nu_{X}^2(\tau_j)) \propto -\log (\tau_j),$$
  so plot of $\log (\nu_{X}^2(\tau_j))$ vs. $\log (\tau_j)$ is linear with a slope of $-1$

Wavelet Variance for White Noise Process: II

- $\nu_{X}^2(\tau_j)$ versus $\tau_j$ for $j = 1, \ldots, 8$ (left-hand plot), along with sample of length $N = 256$ of Gaussian white noise
- largest contribution to $\text{var} \{X_t\}$ is at smallest scale $\tau_1$
- note: later on, we will discuss fractionally differenced (FD) processes that are characterized by a parameter $\delta$; when $\delta = 0$, an FD process is the same as a white noise process
Generalization to Certain Nonstationary Processes

- if wavelet filter is properly chosen, \( \nu^2_X(\tau_j) \) well-defined for certain processes with stationary backward differences (increments); these are also known as intrinsically stationary processes
- first order backward difference of \( X_t \) is process defined by
  \[
  X_t^{(1)} = X_t - X_{t-1}
  \]
- second order backward difference of \( X_t \) is process defined by
  \[
  X_t^{(2)} = X_t^{(1)} - X_{t-1}^{(1)} = X_t - 2X_{t-1} + X_{t-2}
  \]
- \( X_t \) said to have \( d \)th order stationary backward differences if
  \[
  Y_t \equiv \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k}
  \]
  forms a stationary process (\( d \) is a nonnegative integer)

Wavelet Variance for Processes with Stationary Backward Differences: I

- let \( \{X_t\} \) be nonstationary with \( d \)th order stationary differences
- if we use a Daubechies wavelet filter of width \( L \) satisfying \( L \geq 2d \), then \( \nu^2_X(\tau_j) \) is well-defined and finite for all \( \tau_j \), but now
  \[
  \sum_{j=1}^{\infty} \nu^2_X(\tau_j) = \infty
  \]
- works because there is a backward difference operator of order \( d = L/2 \) embedded within \( \{h_{j,t}\} \), so this filter reduces \( X_t \) to
  \[
  \sum_{k=0}^{d} \binom{d}{k} (-1)^k X_{t-k} = Y_t
  \]
  and then creates localized weighted averages of \( Y_t \)’s

Examples of Processes with Stationary Increments

- 1st column shows, from top to bottom, realizations from
  (a) random walk: \( X_t = \sum_{u=1}^{t} \epsilon_u \), & \( \epsilon_t \) is zero mean white noise
  (b) like (a), but now \( \epsilon_t \) has mean of \(-0.2\)
  (c) random run: \( X_t = \sum_{u=1}^{t} Y_u \), where \( Y_t \) is a random walk
- 2nd & 3rd columns show 1st & 2nd differences \( X_t^{(1)} \) and \( X_t^{(2)} \)

Wavelet Variance for Random Walk Process: I

- random walk process \( X_t = \sum_{u=1}^{t} \epsilon_u \) has first order (\( d = 1 \)) stationary differences since \( X_t - X_{t-1} = \epsilon_t \) (i.e., white noise)
- \( L \geq 2d \) holds for all wavelets when \( d = 1 \); for Haar \( (L = 2)\),
  \[
  \nu^2_X(\tau_j) = \frac{\var \{ \epsilon_t \}}{6} \left( \frac{\tau_j}{1 + 2\tau_j} \right) \approx \frac{\var \{ \epsilon_t \}}{6} \tau_j,
  \]
  with the approximation becoming better as \( \tau_j \) increases
- note that \( \nu^2_X(\tau_j) \) increases as \( \tau_j \) increases
- \( \log (\nu^2_X(\tau_j)) \propto \log (\tau_j) \) approximately, so plot of \( \log (\nu^2_X(\tau_j)) \) vs. \( \log (\tau_j) \) is approximately linear with a slope of \(+1\)
- as required, also have
  \[
  \sum_{j=1}^{\infty} \nu^2_X(\tau_j) = \frac{\var \{ \epsilon_t \}}{6} \left( 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \right) = \infty
  \]
Wavelet Variance for Random Walk Process: II

\[ \nu_X^2(\tau_j) \text{ versus } \tau_j \text{ for } j = 1, \ldots, 8 \] (left-hand plot), along with sample of length \( N = 256 \) of a Gaussian random walk process

- smallest contribution to var \( \{X_t\} \) is at smallest scale \( \tau_1 \)
- note: a fractionally differenced process with parameter \( \delta = 1 \) is the same as a random walk process

WMTSA: 327 II-17

Fractionally Differenced (FD) Processes: I

- can create a continuum of processes that ‘interpolate’ between white noise and random walks and ‘extrapolate’ beyond them using notion of ‘fractional differencing’ (Granger and Joyeux, 1980; Hosking, 1981)
- FD(\( \delta \)) process is determined by 2 parameters \( \delta \) and \( \sigma_e^2 \), where \(-\infty < \delta < \infty \) and \( \sigma_e^2 > 0 \) (\( \sigma_e^2 \) is less important than \( \delta \))
- if \( \delta < 1/2 \), FD process \( \{X_t\} \) is stationary, and, in particular,
  - reduces to white noise if \( \delta = 0 \)
  - has ‘long memory’ or ‘long range dependence’ if \( \delta > 0 \)
  - is ‘antipersistent’ if \( \delta < 0 \) (i.e., \( \text{cov} \{X_t, X_{t+1}\} < 0 \))

WMTSA: 281-285 II-18

Fractionally Differenced (FD) Processes: II

- if \( \delta \geq 1/2 \), FD process \( \{X_t\} \) is nonstationary with \( d \)th order stationary backward differences \( \{Y_t\} \)
  - here \( d = \lfloor \delta + 1/2 \rfloor \), where \( \lfloor x \rfloor \) is integer part of \( x \)
  - \( \{Y_t\} \) is stationary FD(\( \delta - d \)) process
- if \( \delta = 1 \), FD process is the same as a random walk process
- except possibly for two or three smallest scales, have
  \[ \nu_X^2(\tau_j) \approx C \tau_j^{2\delta - 1} \]
- thus \( \log(\nu_X^2(\tau_j)) \approx \log(C) + (2\delta - 1) \log(\tau_j) \), so a log/log plot of \( \nu_X^2(\tau_j) \) vs. \( \tau_j \) looks approximately linear with slope \( 2\delta - 1 \) for \( \tau_j \) large enough

WMTSA: 287-288, 297 II-19

LA(8) Wavelet Variance for 2 FD Processes

- see overhead 12 for \( \delta = 0 \) (white noise), which has slope \( -1 \)
- \( \delta = \frac{1}{4} \) has slope \( -\frac{1}{2} \)
- \( \delta = \frac{1}{2} \) has slope 0 (related to so-called ‘pink noise’)
LA(8) Wavelet Variance for 2 More FD Processes

- \( \delta = \frac{5}{6} \) has slope \( \frac{2}{3} \) (related to Kolmogorov turbulence)
- \( \delta = 1 \) has slope 1 (random walk)
- nonnegative slopes indicate nonstationarity, while negative slopes indicate stationarity

Wavelet Variance for Processes with Stationary Backward Differences: II

- summary: \( \nu_X^2(\tau_j) \) well-defined for process \( \{X_t\} \) that is
  - stationary
  - nonstationary with \( d \)th order stationary increments, but width of wavelet filter must satisfy \( L \geq 2d \)
- if \( \{X_t\} \) is stationary, then
  \[ \sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \text{var} \{X_t\} < \infty \]
  (recall that each RV in a stationary process must have the same finite variance)

Background on Gaussian Random Variables

- \( \mathcal{N}(\mu, \sigma^2) \) denotes a Gaussian (normal) RV with mean \( \mu \) and variance \( \sigma^2 \)
- will write
  \[ X \overset{d}{=} \mathcal{N}(\mu, \sigma^2) \]
  to mean ‘RV \( X \) has same distribution as Gaussian RV’
- RV \( \mathcal{N}(0, 1) \) often written as \( Z \) (called standard Gaussian or standard normal)
- let \( \Phi(\cdot) \) be Gaussian cumulative distribution function
  \[ \Phi(z) \equiv \text{P}[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx \]
- inverse \( \Phi^{-1}(\cdot) \) of \( \Phi(\cdot) \) is such that \( \text{P}[Z \leq \Phi^{-1}(p)] = p \)
- \( \Phi^{-1}(p) \) called \( p \times 100\% \) percentage point
Background on Chi-Square Random Variables

- $X$ said to be a chi-square RV with $\eta$ degrees of freedom if its probability density function (PDF) is given by
  \[ f_X(x; \eta) = \frac{1}{2^{\eta/2} \Gamma(\eta/2)} x^{(\eta/2)-1} e^{-x/2}, \quad x \geq 0, \quad \eta > 0 \]

- $\chi_\eta^2$ denotes RV with above PDF

- two important facts: $E\{\chi_\eta^2\} = \eta$ and $\text{var}\{\chi_\eta^2\} = 2\eta$

- let $Q_\eta(p)$ denote the $p$th percentage point for the RV $\chi_\eta^2$:
  \[ P[\chi_\eta^2 \leq Q_\eta(p)] = p \]

**Expected Value of Wavelet Coefficients**

- in preparation for considering problem of estimating $\nu^2_X(\tau_j)$ given an observed time series, need to consider $E\{\overline{W}_{j,t}\}$
- if $\{X_t\}$ is nonstationary but has $d$th order stationary increments, let $\{Y_t\}$ be stationary process obtained by differencing $\{X_t\}$ $d$ times; if $\{X_t\}$ is stationary ($d = 0$ case), let $Y_t = X_t$
- with $\mu_Y \equiv E\{Y_t\}$, have
  \[ -E\{\overline{W}_{j,t}\} = 0 \text{ if either (i) } L > 2d \text{ or (ii) } L = 2d \text{ and } \mu_Y = 0 \]
  \[ -E\{\overline{W}_{j,t}\} \neq 0 \text{ if } \mu_Y \neq 0 \text{ and } L = 2d \]

- thus have $E\{\overline{W}_{j,t}\} = 0$ if $L$ is picked large enough ($L > 2d$ is sufficient, but might not be necessary)

- knowing $E\{\overline{W}_{j,t}\} = 0$ eases job of estimating $\nu^2_X(\tau_j)$ considerably

**Unbiased Estimator of Wavelet Variance: I**

- given a realization of $X_0, X_1, \ldots, X_{N-1}$ from a process with $d$th order stationary differences, want to estimate $\nu^2_X(\tau_j)$
- for wavelet filter such that $L \geq 2d$ and $E\{\overline{W}_{j,t}\} = 0$, have
  \[ \nu^2_X(\tau_j) = \text{var}\{\overline{W}_{j,t}\} = E\{\overline{W}^2_{j,t}\} \]

- can base estimator on squares of
  \[ \overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \hat{h}_{j,l} X_{t-l} \mod N, \quad t = 0, 1, \ldots, N - 1 \]

- recall that
  \[ \overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \hat{h}_{j,l} X_{t-l}, \quad t \in \mathbb{Z} \]

**Unbiased Estimator of Wavelet Variance: II**

- comparing
  \[ \nu^2_X(\tau_j) \equiv \frac{1}{N - L_j + 1} \sum_{t=L_j-1}^{N-1} \overline{W}^2_{j,t} \]

- says that $\overline{W}_{j,t} = \overline{W}_{j,t}$ if ‘mod $N$’ not needed; this happens when $L_j - 1 \leq t < N$ (recall that $L_j = (2^j - 1)(L - 1) + 1$)

- if $N - L_j \geq 0$, unbiased estimator of $\nu^2_X(\tau_j)$ is
  \[ \nu^2_X(\tau_j) \equiv \frac{1}{N - L_j + 1} \sum_{t=L_j-1}^{N-1} \overline{W}^2_{j,t} = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}^2_{j,t} \]

where $M_j \equiv N - L_j + 1$
Statistical Properties of $\nu_X^2(\tau_j)$

- assume that $\{\hat{W}_{j,t}\}$ is Gaussian stationary process with mean zero and ACVS $\{s_{j,\tau}\}$
- suppose $\{s_{j,\tau}\}$ is such that
  $$A_j \equiv \sum_{\tau=-\infty}^{\infty} s_{j,\tau}^2 < \infty$$
  (if $A_j = \infty$, can make it finite usually by just increasing $L$)
- can show that $\hat{\nu}_X^2(\tau_j)$ is asymptotically Gaussian with mean $\nu_X^2(\tau_j)$ and large sample variance $2A_j/M_j$; i.e.,
  $$\frac{\nu_X^2(\tau_j) - \nu_X^2(\tau_j)}{(2A_j/M_j)^{1/2}} \overset{d}{\approx} N(0, 1)$$
  approximately for large $M_j \equiv N - L_j + 1$

Confidence Intervals for $\nu_X^2(\tau_j)$: I

- based upon large sample theory, can form a $100(1 - 2p)\%$ confidence interval (CI) for $\nu_X^2(\tau_j)$:
  $$\left[\hat{\nu}_X^2(\tau_j) - \Phi^{-1}(1 - p)\sqrt{2A_j/M_j}, \hat{\nu}_X^2(\tau_j) + \Phi^{-1}(1 - p)\sqrt{2A_j/M_j}\right]$$
  i.e., random interval traps unknown $\nu_X^2(\tau_j)$ with probability $1 - 2p$
- if $A_j$ replaced by $\hat{A}_j$, get approximate $100(1 - 2p)\%$ CI
- critique: lower limit of CI can very well be negative even though $\nu_X^2(\tau_j) \geq 0$ always
- can avoid this problem by using a $\chi^2$ approximation

Estimation of $A_j$

- in practical applications, need to estimate $A_j = \sum_\tau s_{j,\tau}^2$
- can argue that, for large $M_j$, the estimator
  $$\hat{A}_j \equiv \frac{(s_{j,0})^2}{2} + \sum_{\tau=1}^{M_j-1} \left(\frac{s_{j,\tau}}{M_j}\right)^2,$$
  is approximately unbiased, where
  $$s_{j,\tau} \equiv \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \frac{\hat{W}_{j,t+|\tau|}}{\hat{W}_{j,t}}$$
  $0 \leq |\tau| \leq M_j - 1$
- Monte Carlo results: $\hat{A}_j$ reasonably good for $M_j \geq 128$

Confidence Intervals for $\nu_X^2(\tau_j)$: II

- $\chi^2$ useful for approximating distribution of sum of squared Gaussian RVs, which is what we are dealing with here:
  $$\nu_X^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \hat{W}_{j,t}^2$$
- idea is to assume $\nu_X^2(\tau_j) \overset{d}{=} a\chi^2_\eta$, where $a$ and $\eta$ are constants to be set via moment matching
- because $E\{\chi^2_\eta\} = \eta$ and var $\{\chi^2_\eta\} = 2\eta$, we have $E\{a\chi^2_\eta\} = a\eta$
  and var $\{a\chi^2_\eta\} = 2a^2\eta$
- can equate $E\{\nu_X^2(\tau_j)\} \& \text{var} \{\nu_X^2(\tau_j)\}$ to $a\eta$ & $2a^2\eta$ to determine $a \& \eta$
Confidence Intervals for $\nu^2_X(\tau_j)$: III

- obtain
  \[ \eta = 2 \left( \frac{\var{\nu^2_X(\tau_j)}}{\nu} \right)^2 = \frac{2\nu^2_X(\tau_j)}{\var{\nu^2_X(\tau_j)}} \text{ and } a = \frac{\nu^2_X(\tau_j)}{\nu} \]
- after $\eta$ has been determined, can obtain a CI for $\nu^2_X(\tau_j)$: with probability $1 - 2p$, the random interval
  \[ \left[ \frac{\nu^2_X(\tau_j)}{Q\eta(p)} \right] \]
traps the true unknown $\nu^2_X(\tau_j)$
- lower limit is now nonnegative
- as $N \to \infty$, above CI and Gaussian-based CI converge

Three Ways to Set $\eta$

1. use large sample theory with appropriate estimates:
   \[ \eta = \frac{2\nu^2_X(\tau_j)}{\var{\nu^2_X(\tau_j)}} \approx \frac{2\nu^2_X(\tau_j)}{2A_j/M_j} \text{ suggests } \hat{\eta} = \frac{M_j\nu^2_X(\tau_j)}{A_j} \]
2. make an assumption about the effect of wavelet filter on $\{X_t\}$ to obtain simple approximation
   \[ \eta_3 = \max\{M_j/2^j, 1\} \]
   (this effective – but conservative – approach is valuable if there are insufficient data to reliably estimate $A_j$)
3. third way requires assuming shape of spectral density function associated with $\{X_t\}$ (questionable assumption, but common practice in, e.g., atomic clock literature)

Atomic Clock Deviates: I

- top plot: errors $\{X_t\}$ in time kept by atomic clock 571 (measured in microseconds: 1,000,000 microseconds = 1 second)
- middle: 1st backward differences $\{X_t^{(1)}\}$ in nanoseconds (1000 nanoseconds = 1 microsecond)
- bottom: 2nd backward differences $\{X_t^{(2)}\}$, also in nanoseconds

- if $\{X_t\}$ nonstationary with $d$th order stationary increments, need $L \geq 2d$, but might need $L > 2d$ to get $E\{W_{j,t}\} = 0$
- might regard $\{X_t^{(1)}\}$ as realization of stationary process, but, if so, with a mean value far from 0; $\{X_t^{(2)}\}$ resembles realization of stationary process, but mean value still might not be 0 if we believe there is a linear trend in $\{X_t^{(1)}\}$; thus might need $L \geq 6$, but could get away with $L \geq 4$

Atomic Clock Deviates: II
Atomic Clock Deviates: III

- square roots of wavelet variance estimates for atomic clock time errors \( \{X_t\} \) based upon unbiased MODWT estimator with
  - Haar wavelet (x’s in left-hand plot, with linear fit)
  - D(4) wavelet (circles in left- and right-hand plots)
  - D(6) wavelet (pluses in left-hand plot).
- Haar wavelet inappropriate
  - need \( \{X_t^{(1)}\} \) to be a realization of a stationary process with
    mean 0 (stationarity might be OK, but mean 0 is way off)
  - linear appearance can be explained in terms of nonzero mean
- 95% confidence intervals in the right-hand plot are the square roots of intervals computed using the chi-square approximation with \( \eta \) given by \( \hat{\eta}_1 \) for \( j = 1, \ldots, 6 \) and by \( \eta_3 \) for \( j = 7 \) & 8

Atomic Clock Deviates: IV

Wavelet Variance Analysis of Time Series with Time-Varying Statistical Properties

- each wavelet coefficient \( \hat{W}_{j,t} \) formed using portion of \( X_t \)
- suppose \( X_t \) associated with actual time \( t_0 + t \Delta t \)
  - \( t_0 \) is actual time of first observation \( X_0 \)
  - \( \Delta t \) is spacing between adjacent observations
- suppose \( h_{j,t} \) is least asymmetric Daubechies wavelet
- can associate \( \hat{W}_{j,t} \) with an interval of width \( 2\tau_j \Delta t \) centered at
  \[ t_0 + (2^j(t + 1) - 1 - \lfloor H_j \rfloor \text{ mod } N) \Delta t, \]
  where, e.g., \( \lfloor H_j \rfloor = [7(2^j - 1) + 1]/2 \) for LA(8) wavelet
- can thus form ‘localized’ wavelet variance analysis (implicitly assumes stationarity or stationary increments locally)

Annual Minima of Nile River

- left-hand plot: annual minima of Nile River
- right: Haar \( \hat{\mu}_2 \) before (x’s) and after (o’s) year 715.5, with
  95% confidence intervals based upon \( \chi^2_{\eta_3} \) approximation

Wavelet Variance Analysis of Time Series with Time-Varying Statistical Properties

- each wavelet coefficient \( \hat{W}_{j,t} \) formed using portion of \( X_t \)
- suppose \( X_t \) associated with actual time \( t_0 + t \Delta t \)
  - \( t_0 \) is actual time of first observation \( X_0 \)
  - \( \Delta t \) is spacing between adjacent observations
- suppose \( h_{j,t} \) is least asymmetric Daubechies wavelet
- can associate \( \hat{W}_{j,t} \) with an interval of width \( 2\tau_j \Delta t \) centered at
  \[ t_0 + (2^j(t + 1) - 1 - \lfloor H_j \rfloor \text{ mod } N) \Delta t, \]
  where, e.g., \( \lfloor H_j \rfloor = [7(2^j - 1) + 1]/2 \) for LA(8) wavelet
- can thus form ‘localized’ wavelet variance analysis (implicitly assumes stationarity or stationary increments locally)
Subtidal Sea Level Fluctuations: I

- subtidal sea level fluctuations $X$ for Crescent City, CA, collected by National Ocean Service with permanent tidal gauge
- $N = 8746$ values from Jan 1980 to Dec 1991 (almost 12 years)
- one value every 12 hours, so $\Delta t = 1/2$ day
- ‘subtidal’ is what remains after diurnal & semidiurnal tides are removed by low-pass filter (filter seriously distorts frequency band corresponding to first physical scale $\tau_1 \Delta t = 1/2$ day)

WMTSA: 185–186

Subtidal Sea Level Fluctuations: II

- level $J_0 = 7$ LA(8) MODWT multiresolution analysis

WMTSA: 186

Subtidal Sea Level Fluctuations: III

- estimated time-dependent LA(8) wavelet variances for physical scale $\tau_2 \Delta t = 1$ day based upon averages over monthly blocks (30.5 days, i.e., 61 data points)
- plot also shows a representative 95% confidence interval based upon a hypothetical wavelet variance estimate of $1/2$ and a chi-square distribution with $\nu = 15.25$

WMTSA: 324–326

Subtidal Sea Level Fluctuations: IV

- estimated LA(8) wavelet variances for physical scales $\tau_j \Delta t = 2^{j-2}$ days, $j = 2, \ldots, 7$, grouped by calendar month
Some Extensions

- wavelet cross-covariance and cross-correlation (Whitcher, Guttorp and Percival, 2000; Serroukh and Walden, 2000a, 2000b)
- asymptotic theory for non-Gaussian processes satisfying a certain ‘mixing’ condition (Serroukh, Walden and Percival, 2000)
- biased estimators of wavelet variance (Aldrich, 2005)
- unbiased estimator of wavelet variance for ‘gappy’ time series (Mondal and Percival, 2010a)
- robust estimation (Mondal and Percival, 2010b)
- wavelet variance for random fields (Mondal and Percival, 2010c)
- wavelet-based characteristic scales (Keim and Percival, 2010)

Summary

- wavelet variance gives scale-based analysis of variance
- presented statistical theory for Gaussian processes with stationary increments
- in addition to the applications we have considered, the wavelet variance has been used to analyze
  - genome sequences
  - changes in variance of soil properties
  - canopy gaps in forests
  - accumulation of snow fields in polar regions
  - boundary layer atmospheric turbulence
  - regular and semiregular variable stars

Wavelet-Based Signal Extraction: Overview

- outline key ideas behind wavelet-based approach
- description of four basic models for signal estimation
- discussion of why wavelets can help estimate certain signals
- simple thresholding & shrinkage schemes for signal estimation
- wavelet-based thresholding and shrinkage
- discuss some extensions to basic approach

Wavelet-Based Signal Estimation: I

- DWT analysis of $X$ yields $W = \mathcal{W}X$
- DWT synthesis $X = \mathcal{W}^T W$ yields multiresolution analysis by splitting $\mathcal{W}^T W$ into pieces associated with different scales
- DWT synthesis can also estimate ‘signal’ hidden in $X$ if we can modify $W$ to get rid of noise in the wavelet domain
- if $W'$ is a ‘noise reduced’ version of $W$, can form signal estimate via $\mathcal{W}^T W'$
Wavelet-Based Signal Estimation: II

- key ideas behind simple wavelet-based signal estimation
  - certain signals can be efficiently described by the DWT using
    * all of the scaling coefficients
    * a small number of 'large' wavelet coefficients
  - noise is manifested in a large number of 'small' wavelet coefficients
  - can either 'threshold' or 'shrink' wavelet coefficients to eliminate noise in the wavelet domain
- key ideas led to wavelet thresholding and shrinkage proposed by Donoho, Johnstone and coworkers in 1990s

Models for Signal Estimation: I

- will consider two types of signals:
  1. \( D \), an \( N \) dimensional deterministic signal
  2. \( C \), an \( N \) dimensional stochastic signal; i.e., a vector of random variables (RVs) with covariance matrix \( \Sigma_C \)
- will consider two types of noise:
  1. \( \epsilon \), an \( N \) dimensional vector of independent and identically distributed (IID) RVs with mean 0 and covariance matrix \( \Sigma\epsilon = \sigma^2\epsilon I_N \)
  2. \( \eta \), an \( N \) dimensional vector of non-IID RVs with mean 0 and covariance matrix \( \Sigma\eta \)
    * one form: RVs independent, but have different variances
    * another form of non-IID: RVs are correlated

Models for Signal Estimation: II

- leads to four basic 'signal + noise' models for \( X \)
  1. \( X = D + \epsilon \)
  2. \( X = D + \eta \)
  3. \( X = C + \epsilon \)
  4. \( X = C + \eta \)
- in the latter two cases, the stochastic signal \( C \) is assumed to be independent of the associated noise

Signal Representation via Wavelets: I

- consider \( X = D + \epsilon \) first
- signal estimation problem is simplified if we can assume that the important part of \( D \) is in its large values
- assumption is not usually viable in the original (i.e., time domain) representation \( D \), but might be true in another domain
- an orthonormal transform \( \mathcal{O} \) might be useful because
  - \( d = \mathcal{O}D \) is equivalent to \( D \) (since \( D = \mathcal{O}^T d \))
  - we might be able to find \( \mathcal{O} \) such that the signal is isolated in \( M \ll N \) large transform coefficients
- \( \mathcal{Q} \): how can we judge whether a particular \( \mathcal{O} \) might be useful for representing \( D \)?
Signal Representation via Wavelets: II

- let $d_j$ be the $j$th transform coefficient in $d = \mathcal{O}D$
- let $d_0, d_1, \ldots, d_{N-1}$ be the $d_j$'s reordered by magnitude:
  \[
  |d_0| \geq |d_1| \geq \cdots \geq |d_{N-1}|
  \]
- example: if $d = [-3, 1, 4, -7, 2, -1]^T$, then $d_0 = d_3 = -7, d_1 = d_2 = 4, d_2 = d_0 = -3$ etc.
- define a normalized partial energy sequence (NPES):  
  \[
  C_{M-1} = \frac{\sum_{j=0}^{M-1} |d_j|^2}{\sum_{j=0}^{N-1} |d_j|^2} = \text{energy in largest } M \text{ terms} \quad \text{total energy in signal}
  \]
- let $I_M$ be $N \times N$ diagonal matrix whose $j$th diagonal term is 1 if $|d_j|$ is one of the $M$ largest magnitudes and is 0 otherwise

Signal Representation via Wavelets: III

- form $\hat{D}_M \equiv \mathcal{O}^T I_M d$, which is an approximation to $D$
- when $d = [-3, 1, 4, -7, 2, -1]^T$ and $M = 3$, we have $I_3 = 
  \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$ and thus $\hat{D}_M = \mathcal{O}^T \begin{pmatrix}
-3 \\
0 \\
4 \\
-7 \\
0 \\
0
\end{pmatrix}$
- one interpretation for NPES:  
  \[
  C_{M-1} = 1 - \frac{||D - \hat{D}_M||^2}{||D||^2} = 1 - \text{relative approximation error}
  \]

Signal Representation via Wavelets: IV

- consider three signals plotted above
- $D_1$ is a sinusoid, which can be represented succinctly by the discrete Fourier transform (DFT)
- $D_2$ is a bump (only a few nonzero values in the time domain)
- $D_3$ is a linear combination of $D_1$ and $D_2$

Signal Representation via Wavelets: V

- consider three different orthonormal transforms
  - identity transform $I$ (time)
  - the orthonormal DFT $\mathcal{F}$ (frequency), where $\mathcal{F}$ has $(k,t)$th element $\exp(-i2\pi tk/N)/\sqrt{N}$ for $0 \leq k, t \leq N - 1$
  - the LA(8) DWT $W$ (wavelet)
- # of terms $M$ needed to achieve relative error $< 1%$:

\[
\begin{array}{ccc}
D_1 & D_2 & D_3 \\
\text{DFT} & 2 & 29 & 28 \\
\text{identity} & 105 & 9 & 75 \\
\text{LA(8) wavelet} & 22 & 14 & 21
\end{array}
\]
Signal Representation via Wavelets: VI

- use NPESs to see how well these three signals are represented in the time, frequency (DFT) and wavelet (LA(8)) domains
- time (solid curves), frequency (dotted) and wavelet (dashed)

WMTSA: 395–396 I 57

Signal Estimation via Thresholding: I

- thresholding schemes involve
  1. computing \( O \equiv O X \)
  2. defining \( O^{(t)} \) as vector with \( t \)th element
     \[
     O^{(t)}_I = \begin{cases} 
     0, & \text{if } |O_I| \leq \delta; \\
     \text{some nonzero value, otherwise,}
     \end{cases}
     \]
     where nonzero values are yet to be defined
  3. estimating \( D \) via \( \hat{D}^{(t)} \equiv O^T O^{(t)} \)
- simplest scheme is ‘hard thresholding’ (‘kill/keep’ strategy):
  \[
  O^{(ht)}_I = \begin{cases} 
  0, & \text{if } |O_I| \leq \delta; \\
  O_I, & \text{otherwise.}
  \end{cases}
  \]

WMTSA: 399 I 59

Hard Thresholding Function

- plot shows mapping from \( O_I \) to \( O^{(ht)}_I \)

WMTSA: 399 I 60
Signal Estimation via Thresholding: II

- hard thresholding is a strategy that arises from solving a simple optimization problem, namely, find $\hat{D}_M$ such that
  $$\gamma_m \equiv \|X - \hat{D}_m\|^2 + m\delta^2$$
  is minimized over all possible $\hat{D}_m = \mathcal{O}^T\mathcal{I}_m\mathcal{O}$, $m = 0, \ldots, N$
- $\delta$ is a fixed parameter that is set a priori (we assume $\delta > 0$)
- $\|X - \hat{D}_m\|^2$ is a measure of ‘fidelity’
  - rationale for this term: $\hat{D}_m$ shouldn’t stray too far from $X$
    (particularly if signal-to-noise ratio is high)
  - fidelity increases (the measure decreases) as $m$ increases
  - in minimizing $\gamma_m$, consideration of this term alone suggests that $m$ should be large

WMTSA: 398  II.61

Signal Estimation via Thresholding: III

- $m\delta^2$ is a penalty for too many terms
  - rationale: heuristic says $d = \mathcal{O}\mathcal{D}$ consists of just a few large coefficients
  - penalty increases as $m$ increases
  - in minimizing $\gamma_m$, consideration of this term alone suggests that $m$ should be small
- optimization problem: balance off fidelity & parsimony
- can show that $\gamma_m = \|X - \hat{D}_m\|^2 + m\delta^2$ is minimized when $m$ is set such that $\mathcal{I}_m$ picks out all coefficients satisfying $O_j^2 > \delta^2$

WMTSA: 398  II.62

Signal Estimation via Thresholding: IV

- alternative scheme is ‘soft thresholding’:
  $$O_l^{(st)} = \text{sign} \{O_l\} (|O_l| - \delta)_+,$$
  where
  $$\text{sign} \{O_l\} \equiv \begin{cases} +1, & \text{if } O_l > 0; \\ 0, & \text{if } O_l = 0; \\ -1, & \text{if } O_l < 0. \end{cases}$$
  $$\text{sign} \{O_l\} \equiv \begin{cases} x, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$
  - one rationale for soft thresholding: fits into Stein’s class of estimators, for which unbiased estimation of risk is possible

WMTSA: 399–400  II.61

Soft Thresholding Function

- here is the mapping from $O_l$ to $O_l^{(st)}$

WMTSA: 399–400  II.64
Signal Estimation via Thresholding: V

- third scheme is ‘mid thresholding:

\[ O_l^{(mt)} = \text{sign} \{O_l\} (|O_l| - \delta)_{++}, \]

where

\[ (|O_l| - \delta)_{++} = \begin{cases} 
2(|O_l| - \delta)_+, & \text{if } |O_l| < 2\delta; \\
|O_l|, & \text{otherwise}
\end{cases} \]

- provides compromise between hard and soft thresholding

\[ \text{Mid Thresholding Function} \]

- here is the mapping from \( O_l \) to \( O_l^{(mt)} \)

\[ O_l^{(mt)} \]

\[ O_l \]

\[ O_l \]

\[ O_l^{(mt)} \]

\[ D_l^{(mt)} \]

\[ X_l \]

\[ t = 1 \]

\[ t = 64 \]

Signal Estimation via Thresholding: VI

- example of mid thresholding with \( \delta = 1 \)

\[ X_l \]

\[ O_l \]

\[ O_l^{(mt)} \]

\[ D_l^{(mt)} \]

Universal Threshold

- Q: how do we go about setting \( \delta \)?
- specialize to IID Gaussian noise \( \epsilon \) with covariance \( \sigma^2 \epsilon I_N \)
- can argue \( \epsilon \equiv \mathcal{O} \epsilon \) is also IID Gaussian with covariance \( \sigma^2 \epsilon I_N \)
- Donoho & Johnstone (1995) proposed \( \delta^{(u)} \equiv [2\sigma^2 \epsilon \log(N)]^{1/2} \)

('log' here is 'log base \( e \'))

- rationale for \( \delta^{(u)} \): because of Gaussianity, can argue that

\[ \mathbb{P} \left[ \max_l \{ |\epsilon_l| \} > \delta^{(u)} \right] \leq \frac{1}{\sqrt{4\pi \log(N)}} \rightarrow 0 \quad \text{as} \quad N \rightarrow \infty \]

and hence \( \mathbb{P} \left[ \max_l \{ |\epsilon_l| \} \leq \delta^{(u)} \right] \rightarrow 1 \) as \( N \rightarrow \infty \), so no noise will exceed threshold in the limit.
Wavelet-Based Thresholding

- assume model of deterministic signal plus IID Gaussian noise with mean 0 and variance \( \sigma^2_e \): \( X = D + \epsilon \)
- using a DWT matrix \( \mathbf{W} \), form \( \mathbf{W} = \mathbf{WX} = \mathbf{WD} + \mathbf{W}\epsilon \equiv \mathbf{d} + \mathbf{e} \)
- because \( \epsilon \) IID Gaussian, so is \( \mathbf{e} \)
- Donoho & Johnstone (1994) advocate the following:
  - form partial DWT of level \( J_0 \): \( \mathbf{W}_1, \ldots, \mathbf{W}_{J_0} \) and \( \mathbf{V}_{J_0} \)
  - threshold \( \mathbf{W}_j \)’s but leave \( \mathbf{V}_{J_0} \) alone (i.e., administratively, all \( N/2^{j_0} \) scaling coefficients assumed to be part of \( \mathbf{d} \))
  - use universal threshold \( \delta(u) = \sqrt{2\sigma^2_e \log(N)} \)
  - use thresholding rule to form \( \mathbf{W}_{j}^{(t)} \) (hard, etc.)
  - estimate \( \mathbf{D} \) by inverse transforming \( \mathbf{W}_{1}^{(t)}, \ldots, \mathbf{W}_{J_0}^{(t)} \) and \( \mathbf{V}_{J_0} \)

MAD Scale Estimator: I

- procedure assumes \( \sigma_e \) is known, which is not usually the case
- if unknown, use median absolute deviation (MAD) scale estimator to estimate \( \sigma_e \) using \( \mathbf{W}_1 \)

\[
\hat{\sigma}_{(\text{mad})} \equiv \text{median} \{ |W_{1,0}|, |W_{1,1}|, \ldots, |W_{1,N/2^{J_0}-1}| \}
\]

- heuristic: bulk of \( W_{1,t} \)'s should be due to noise
- \( 0.6745 \) yields estimator such that \( E\{\hat{\sigma}_{(\text{mad})}\} = \sigma_e \) when \( W_{1,t} \)'s are IID Gaussian with mean 0 and variance \( \sigma^2_e \)
- designed to be robust against large \( W_{1,t} \)'s due to signal

MAD Scale Estimator: II

- example: suppose \( \mathbf{W}_1 \) has 7 small ‘noise’ coefficients & 2 large ‘signal’ coefficients (say, \( a \) & \( b \), with \( 2 \ll |a| < |b| \)):

\[
\mathbf{W}_1 = [1.23, -1.72, -0.80, -0.01, a, 0.30, 0.67, b, -1.33]^T
\]

- ordering these by their magnitudes yields

\[
0.01, 0.30, 0.67, 0.80, 1.23, 1.33, 1.72, |a|, |b|
\]

- median of these absolute deviations is 1.23, so

\[
\hat{\sigma}_{(\text{mad})} = 1.23/0.6745 \approx 1.82
\]

- \( \hat{\sigma}_{(\text{mad})} \) not influenced adversely by \( a \) and \( b \); i.e., scale estimate depends largely on the many small coefficients due to noise

Examples of DWT-Based Thresholding: I

- NMR spectrum
Examples of DWT-Based Thresholding: II

- Signal estimate using $J_0 = 6$ partial D(4) DWT with hard thresholding and universal threshold level estimated by $\hat{\delta}(u) = \sqrt[4]{2\sigma_{\text{mad}}^2 \log(N)} \approx 6.49$

Examples of DWT-Based Thresholding: III

- Same as before, but now using LA(8) DWT with $\hat{\delta}(u) = 6.13$

Examples of DWT-Based Thresholding: IV

- Signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with soft thresholding

Examples of DWT-Based Thresholding: V

- Signal estimate using $J_0 = 6$ partial LA(8) DWT, but now with mid thresholding
MODWT-Based Thresholding

• can base thresholding procedure on MODWT rather than DWT, yielding signal estimators $\tilde{D}^{(ht)}$, $\tilde{D}^{(st)}$ and $\tilde{D}^{(mt)}$
• because MODWT filters are normalized differently, universal threshold must be adjusted for each level:
  
  \[ \delta_j^{(u)} = \sqrt{2\tilde{\sigma}_j^{2 \text{mad}}} \log (N)/2^j, \]

  where now MAD scale estimator is based on unit scale MODWT wavelet coefficients
• results are identical to what ‘cycle spinning’ would yield

Examples of MODWT-Based Thresholding: I

• signal estimate using $J_0 = 6$ LA8 MODWT with hard thresholding

Examples of MODWT-Based Thresholding: II

• same as before, but now with soft thresholding

Examples of MODWT-Based Thresholding: III

• same as before, but now with mid thresholding
Signal Estimation via Shrinkage: I

- so far, we have only considered signal estimation via thresholding rules, which will map some $O_l$ to zeros
- will now consider shrinkage rules, which differ from thresholding only in that nonzero coefficients are mapped to nonzero values rather than exactly zero (but values can be very close to zero!)
- several ways in which shrinkage rules arise – will consider a conditional mean approach (identical to a Bayesian approach)

Background on Conditional PDFs: I

- let $X$ and $Y$ be RVs with marginal probability density functions (PDFs) $f_X(\cdot)$ and $f_Y(\cdot)$
- let $f_{X,Y}(x,y)$ be their joint PDF at the point $(x,y)$
- conditional PDF of $Y$ given $X = x$ is defined as
  $$f_{Y\mid X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$
- $f_{Y\mid X=x}(\cdot)$ is a PDF, so its mean value is
  $$E\{Y\mid X=x\} = \int_{-\infty}^{\infty} y f_{Y\mid X=x}(y) \, dy$$
  the above is called the conditional mean of $Y$, given $X$

Background on Conditional PDFs: II

- suppose RVs $X$ and $Y$ are related, but we can only observe $X$
- want to approximate unobservable $Y$ based on function of $X$
- example: $X$ represents a stochastic signal $Y$ buried in noise
- suppose we want our approximation to be the function of $X$, say $U_2(X)$, such that the mean square difference between $Y$ and $U_2(X)$ is as small as possible; i.e., we want
  $$E\{(Y - U_2(X))^2\}$$
to be as small as possible
- solution is to use $U_2(X) = E\{Y\mid X\}$; i.e., the conditional mean of $Y$ given $X$ is our best guess at $Y$ in the sense of minimizing the mean square error (related to fact that $E\{(Y - a)^2\}$ is smallest when $a = E\{Y\}$

Conditional Mean Approach: I

- assume model of stochastic signal plus non-IID noise:
  $$X = C + \eta$$ so that $O = OX = O(C + \eta) = R + n$
- component-wise, have $O_l = R_l + n_l$
- because $C$ and $\eta$ are independent, $R$ and $n$ must be also
- suppose we approximate $R_l$ via $\tilde{R}_l = U_2(O_l)$, where $U_2(O_l)$ is selected to minimize $E\{(R_l - U_2(O_l))^2\}$
- solution is to set $U_2(O_l)$ equal to $E\{R_l\mid O_l\}$, so let’s work out what form this conditional mean takes
- to get $E\{R_l\mid O_l\}$, need the PDF of $R_l$ given $O_l$, which is
  $$f_{R_l\mid O_l=a_l}(r_l) = \frac{f_{R_l,O_l}(r_l,a_l)}{f_{O_l}(a_l)} = \frac{f_{R_l}(r_l) f_{n_l}(a_l - r_l)}{\int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(a_l - r_l) \, dr_l}$$
Conditional Mean Approach: II

- mean value of \( f_{R_l|O_l = o_l} \) yields estimator \( \hat{R}_l = E\{R_l|O_l\} \):
  \[
  E\{R_l|O_l = o_l\} = \int_{-\infty}^{\infty} r_l f_{R_l|O_l = o_l}(r_l) \, dr_l \\
  = \int_{-\infty}^{\infty} r_l f_{R_l}(r_l) f_{n_l}(o_l - r_l) \, dr_l \\
  = \int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(o_l - r_l) \, dr_l
  \]
- to make further progress, we need a model for the wavelet-domain representation \( R_l \) of the signal
- heuristic that signal in the wavelet domain has a few large values and lots of small values suggests a Gaussian mixture model

Conditional Mean Approach: III

- let \( I_l \) be an RV such that \( P[I_l = 1] = p_l \) & \( P[I_l = 0] = 1 - p_l \)
- under Gaussian mixture model, \( R_l \) has same distribution as
  \[
  I_l \mathcal{N}(0, \gamma_l^2 \sigma_{G_l}^2) + (1 - I_l) \mathcal{N}(0, \sigma_{G_l}^2)
  \]
  where \( \mathcal{N}(0, \sigma^2) \) is a Gaussian RV with mean 0 and variance \( \sigma^2 \)
- 2nd component models small # of large signal coefficients
- 1st component models large # of small coefficients (\( \gamma_l^2 \ll 1 \))
- example: PDFs for case \( \sigma_{G_l}^2 = 10, \gamma_l^2 \sigma_{G_l}^2 = 1 \) and \( p_l = 0.75 \)

Conditional Mean Approach: VI

- let’s simplify to a ‘sparse’ signal model by setting \( \gamma_l = 0 \); i.e., large # of small coefficients are all zero
- distribution for \( R_l \) same as \( (1 - I_l) \mathcal{N}(0, \sigma_{G_l}^2) \)
- to complete model, let \( n_l \) obey a Gaussian distribution with mean 0 and variance \( \sigma_{n_l}^2 \)
- conditional mean estimator becomes \( E\{R_l|O_l = o_l\} = \frac{b_l}{1 + \gamma_l} o_l \), where
  \[
  c_l = \frac{p_l \sqrt{\sigma_{G_l}^2 + \sigma_{n_l}^2}}{(1 - p_l) \sigma_{n_l}} - \frac{\sigma_{n_l}^2 b_l}{2(\sigma_{n_l}^2)}
  \]

Conditional Mean Approach: VII

- conditional mean shrinkage rule for \( p_l = 0.95 \) (i.e., \( \approx 95\% \) of signal coefficients are 0); \( \sigma_{n_l}^2 = 1 \); and \( \sigma_{G_l}^2 = 5 \) (curve furthest from dotted diagonal), 10 and 25 (curve nearest to diagonal)
- as \( \sigma_{G_l}^2 \) gets large (i.e., large signal coefficients increase in size), shrinkage rule starts to resemble mid thresholding rule
Wavelet-Based Shrinkage: I

- assume model of stochastic signal plus Gaussian IID noise: \( X = C + \epsilon \) so that \( W = WX = WC + W\epsilon \equiv R + e \)
- component-wise, have \( W_{j,t} = R_{j,t} + e_{j,t} \)
- form partial DWT of level \( J_0 \), shrink \( W_j \)'s, but leave \( V_{j_0} \) alone
- assume \( E\{R_{j,t}\} = 0 \) (reasonable for \( W_j \), but not for \( V_{j_0} \))
- use a conditional mean approach with the sparse signal model
  - \( R_{j,t} \) has distribution dictated by \((1 - I_{j,t})N(0, \sigma^2_C)\), where \( P[I_{j,t} = 1] = p \) and \( P[I_{j,t} = 0] = 1 - p \)
  - \( R_{j,t} \)'s are assumed to be IID
  - model for \( e_{j,t} \) is Gaussian with mean 0 and variance \( \sigma^2_\epsilon \)
  - note: parameters do not vary with \( j \) or \( t \)

Examples of Wavelet-Based Shrinkage: I

- shrinkage signal estimates of NMR spectrum based upon level \( J_0 = 6 \) partial LA(8) DWT and conditional mean with \( p = 0.9 \)

Wavelet-Based Shrinkage: II

- model has three parameters \( \sigma^2_C, p \) and \( \sigma^2_\epsilon \), which need to be set
- let \( \sigma^2_R \) and \( \sigma^2_W \) be variances of RVs \( R_{j,t} \) and \( W_{j,t} \)
- have relationships \( \sigma^2_R = (1 - p)\sigma^2_C \) and \( \sigma^2_W = \sigma^2_R + \sigma^2_\epsilon \)
  - set \( \hat{\sigma}^2_\epsilon = \hat{\sigma}^2_{\text{med}} \) using \( W_1 \)
  - let \( \hat{\sigma}^2_W \) be sample mean of all \( W_{j,t}^2 \)
  - given \( p \), let \( \hat{\sigma}^2_C = (\hat{\sigma}^2_W - \hat{\sigma}^2_\epsilon)/(1 - p) \)
  - \( p \) usually chosen subjectively, keeping in mind that \( p \) is proportion of noise-dominated coefficients (can set based on rough estimate of proportion of ‘small’ coefficients)

Examples of Wavelet-Based Shrinkage: II

- same as before, but now with \( p = 0.95 \)
Examples of Wavelet-Based Shrinkage: III

\[ p = 0.99 \]

- same as before, but now with \( p = 0.99 \) (as \( p \to 1 \), we declare there are proportionately fewer significant signal coefficients, implying need for heavier shrinkage)

Comments on ‘Next Generation’ Denoising: I

- ‘classical’ denoising looks at each \( W_{j,t} \) alone; for ‘real world’ signals, coefficients often cluster within a given level and persist across adjacent levels (ECG series offers an example)

Comments on ‘Next Generation’ Denoising: II

- here are some ‘next generation’ approaches that exploit these ‘real world’ properties:
  - Crouse \textit{et al.} (1998) use hidden Markov models for stochastic signal DWT coefficients to handle clustering, persistence and non-Gaussianity
  - Huang and Cressie (2000) consider scale-dependent multi-scale graphical models to handle clustering and persistence
  - Cai and Silverman (2001) consider ‘block’ thresholding in which coefficients are thresholded in blocks rather than individually (handles clustering)
  - Dragotti and Vetterli (2003) introduce the notion of ‘wavelet footprints’ to track discontinuities in a signal across different scales (handles persistence)

Comments on ‘Next Generation’ Denoising: III

- ‘classical’ denoising also suffers from problem of overall significance of multiple hypothesis tests
- ‘next generation’ work integrates idea of ‘false discovery rate’ (Benjamini and Hochberg, 1995) into denoising (see Wink and Roerdink, 2004, for an applications-oriented discussion)
- for more recent developments (there are a lot!!!), see
  - review article by Antoniadis (2007)
  - Chapters 3 and 4 of book by Nason (2008)
  - October 2009 issue of \textit{Statistica Sinica}, which has a special section entitled ‘Multiscale Methods and Statistics: A Productive Marriage’

WMTSA: 425  II–91

WMTSA: 453–455 II–93

WMTSA: 490  II–94

WMTSA: 495–497 II–96
Wavelet-Based Decorrelation of Time Series: Overview

- DWT well-suited for decorrelating certain time series, including ones generated from a fractionally differenced (FD) process
- on synthesis side, leads to
  - DWT-based simulation of FD processes
  - wavelet-based bootstrapping
- on analysis side, leads to
  - wavelet-based estimators for FD parameters
  - test for homogeneity of variance
  - test for trends (won’t discuss – see Craigmyle et al., 2004, for details)

\[ W_{\tilde{w}} \text{ well-suited for decorrelating certain time series including ones generated from a fractionally differenced } \text{LnlM process.} \]

- on synthesis side, leads to
  - \( \tilde{W}_{\tilde{w}} \)-based simulation of \( \text{nlLTRYM processes} \)
  - \( \tilde{W}_{\tilde{w}} \)-based bootstrapping
- on analysis side, leads to
  - \( \tilde{W}_{\tilde{w}} \)-based estimators for \( \text{nl} \) parameters
  - test for homogeneity of variance
  - test for trends (\( \text{don’t discuss – see Craigmyle et al., 2004, for details} \))

\[ \hat{\rho}_{X, \tau} \equiv \frac{\sum_{t=0}^{N-1-\tau} X_t X_{t+\tau}}{\sum_{t=0}^{N-1} X_t^2} \]

- note that ACS dies down slowly

- \( \text{LA(8) DWT of FD(0.4) series and sample ACSs for each } W_j \) & \( \tilde{V}_7 \), along with 95% confidence intervals for white noise

- \( \text{LA(8) MODWT of FD(0.4) series & sample ACSs for MODWT coefficients, none of which are approximately uncorrelated} \)
DWT of an FD Process: III

- in contrast to \( X \), ACSs for \( W_j \) consistent with white noise
- variance of RVs in \( W_j \) increases with \( j \) for FD process, 
  \[ \text{var} \{ W_{j,t} \} \approx c \tau_j^{2\delta} \equiv C_j, \]
  where \( c \) is a constant depending on \( \delta \) but not \( j \), and \( \tau_j = 2^{j-1} \)
  is scale associated with \( W_j \)
- for white noise \( (\delta = 0) \), \( \text{var} \{ W_{j,t} \} \) is the same for all \( j \)
- dependence in \( X \) thus manifests itself in wavelet domain by different variances for wavelet coefficients at different scales

Correlations Within a Scale

\[
\begin{array}{cccc}
  j = 1 & j = 2 & j = 3 & j = 4 \\
  \begin{array}{cccc}
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
    \cdot & \cdot & \cdot & \cdot \\
  \end{array} \\
  \text{Haar} & \text{D(4)} & \text{LA(8)}
\end{array}
\]

- correlations between \( W_{j,t} \) and \( W_{j,t+\tau} \) for an FD(0.4) process
- correlations within scale are slightly smaller for Haar
- maximum magnitude of correlation is less than 0.2

Correlations Within a Scale and Between Two Scales

- let \( \{ s_{X,\tau} \} \) denote autocovariance sequence (ACVS) for \( \{ X_t \} \); 
  i.e., \( s_{X,\tau} = \text{cov} \{ X_t, X_{t+\tau} \} \)
- let \( \{ h_{j,l} \} \) denote equivalent wavelet filter for \( j \)th level
- to quantify decorrelation, can write
  \[
  \text{cov} \{ W_{j,t}, W_{j',t'} \} = \sum_{l=0}^{L_j-1} \sum_{l'=-l}^{L_j-1} h_{j,l}h_{j',l'} s_{X,2^j(t+1)-l-2^{j'+1}+l'},
  \]
  from which we can get ACVS (and hence within-scale correlations) for \( \{ W_{j,t} \} \):
  \[
  \text{cov} \{ W_{j,t}, W_{j,t+\tau} \} = \sum_{m=-(L_j-1)}^{L_j-1} s_{X,2^j\tau+m} \sum_{l=0}^{L_j-|m|-1} h_{j,l}h_{j,l+|m|}
  \]

Correlations Between Two Scales: I

- correlation between Haar wavelet coefficients \( W_{j,t} \) and \( W_{j',t'} \) from FD(0.4) process and for levels satisfying \( 1 \leq j < j' \leq 4 \)
Correlations Between Two Scales: II

\[
\begin{array}{cccc}
  j' = 2 & j' = 3 & j' = 4 & j = 1 \\
  0.0 & 0.0 & 0.2 & 0.2 \\
  0.0 & 0.0 & 0.2 & 0.2 \\
  0.0 & 0.0 & 0.2 & 0.2 \\
  \cdots & \cdots & \cdots & \cdots \\
  \tau & \tau & \tau & \tau \\
  -8 & 0 & 8 & \cdots \\
  -8 & 0 & 8 & \cdots \\
  -8 & 0 & 8 & \cdots \\

\end{array}
\]

- same as before, but now for LA(8) wavelet coefficients
- correlations between scales decrease as \( L \) increases

Wavelet Domain Description of FD Process

- DWT acts as a decorrelating transform for FD processes and other (but not all!) intrinsically stationary processes
- wavelet domain description is simple
  - wavelet coefficients within a given scale approximately uncorrelated (refinement: assume 1st order autoregressive model)
  - wavelet coefficients have scale-dependent variance controlled by the two FD parameters \( \delta \) and \( \sigma^2 \)
  - wavelet coefficients between scales also approximately uncorrelated (approximation improves as filter width \( L \) increases)

DWT-Based Simulation

- properties of DWT of FD processes lead to schemes for simulating time series \( X \equiv [X_0, \ldots, X_{N-1}]^T \) with zero mean and with a multivariate Gaussian distribution
- with \( N = 2^J \), recall that \( X = W^T W \), where

\[
W = \begin{bmatrix}
W_1 \\
W_2 \\
\vdots \\
W_J \\
V_J
\end{bmatrix}
\]

Basic DWT-Based Simulation Scheme

- assume \( W \) to contain \( N \) uncorrelated Gaussian (normal) random variables (RVs) with zero mean
- assume \( W_j \) to have variance \( C_j = c_j^2 \delta \)
- assume single RV in \( V_j \) to have variance \( C_{J+1} \) (see Percival and Walden, 2000, for details on how to set \( C_{J+1} \))
- approximate FD time series \( X \) via \( Y \equiv W^T \Lambda^{1/2} Z \), where
  - \( \Lambda^{1/2} \) is \( N \times N \) diagonal matrix with diagonal elements
  \[
  \begin{array}{c}
  C_1^{1/2}, \ldots, C_1^{1/2}, C_2^{1/2}, \ldots, C_2^{1/2}, \ldots, C_{J-1}^{1/2}, C_J^{1/2}, C_J^{1/2}, C_J^{1/2}
  \end{array}
  \]
  - \( \frac{N}{2} \) of these \( \frac{N}{4} \) of these \( \frac{N}{8} \) of these \( \frac{N}{16} \) of these
- \( Z \) is vector of deviations drawn from a Gaussian distribution with zero mean and unit variance
Refinements to Basic Scheme: I

- covariance matrix for approximation $Y$ does not correspond to that of a stationary process
- recall $W$ treats $X$ as if it were circular
- let $T$ be $N \times N$ ‘circular shift’ matrix:
  \[
  T \begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_0 \end{bmatrix}; \quad T^2 \begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} Y_2 \\ Y_0 \\ Y_1 \\ Y_3 \end{bmatrix}; \quad \text{etc.}
  \]
- let $\kappa$ be uniformly distributed over $0, \ldots, N - 1$
- define $\tilde{Y} \equiv T^\kappa Y$
- $\tilde{Y}$ is stationary with ACVS given by, say, $s_{\tilde{Y},\tau}$

Refinements to Basic Scheme: II

- Q: how well does $\{s_{\tilde{Y},\tau}\}$ match $\{s_{X,\tau}\}$?
- due to circularity, find that $s_{\tilde{Y},N-\tau} = s_{\tilde{Y},\tau}$ for $\tau = 1, \ldots, N/2$
- implies $s_{\tilde{Y},\tau}$ cannot approximate $s_{X,\tau}$ well for $\tau$ close to $N$
- can patch up by simulating $\tilde{Y}$ with $M > N$ elements and then extracting first $N$ deviates ($M = 4N$ works well)

Refinements to Basic Scheme: III

- plot shows true ACVS $\{s_{X,\tau}\}$ (thick curves) for FD(0.4) process and wavelet-based approximate ACVSs $\{s_{\tilde{Y},\tau}\}$ (thin curves) based on an LA(8) DWT in which an $N = 64$ series is extracted from $M = N$, $M = 2N$ and $M = 4N$ series

Example and Some Notes

- simulated FD(0.4) series (LA(8), $N = 1024$ and $M = 4N$)
- notes:
  - can form realizations faster than best exact method
  - can efficiently simulate extremely long time series in ‘real-time’ (e.g., $N = 2^{30} = 1,073,741,824$ or even longer!)
  - effect of random circular shifting is to render time series slightly non-Gaussian (a Gaussian mixture model)
Wavelet-Domain Bootstrapping

- for many (but not all!) time series, DWT acts as a decorrelating transform: to a good approximation, each $W_j$ is a sample of a white noise process, and coefficients from different sub-vectors $W_j$ and $W_j'$ are also pairwise uncorrelated
- variance of coefficients in $W_j$ depends on $j$
- scaling coefficients $V_{j_0}$ are still autocorrelated, but there will be just a few of them if $J_0$ is selected to be large
- decorrelating property holds particularly well for FD and other processes with long-range dependence
- above suggests the following recipe for wavelet-domain bootstrapping of a statistic of interest, e.g., sample autocorrelation sequence $\hat{\rho}_{X,\tau}$ at unit lag $\tau = 1$

Recipe for Wavelet-Domain Bootstrapping

1. given $X$ of length $N = 2^J$, compute level $J_0$ DWT (the choice $J_0 = J - 3$ yields 8 coefficients in $W_{J_0}$ and $V_{J_0}$)
2. randomly sample with replacement from $W_j$ to create bootstrapped vector $W_j^{(b)}$, $j = 1, \ldots, J_0$
3. create $V_{J_0}^{(b)}$ using 1st-order autoregressive parametric bootstrap
4. apply $W_j^{T}$ to $W_j^{(b)}$, $W_{J_0}^{(b)}$ and $V_{J_0}^{(b)}$ to obtain bootstrapped time series $X^{(b)}$ and then form $\hat{\rho}_{X,1}^{(b)}$
- repeat above many times to build up sample distribution of bootstrapped autocorrelations

Illustration of Wavelet-Domain Bootstrapping

- approximations to true PDF using (a) Haar & (b) LA(8) wavelets
- using 50 FD time series and the Haar DWT yields:
  - average of 50 sample means $\hat{\rho}_{1} \approx 0.35$ (truth $= 0.53$)
  - average of 50 sample SDs $\hat{\rho}_{1} \approx 0.096$ (truth $= 0.107$)
- using 50 FD time series and the LA(8) DWT yields:
  - average of 50 sample means $\hat{\rho}_{1} \approx 0.43$ (truth $= 0.53$)
  - average of 50 sample SDs $\hat{\rho}_{1} \approx 0.098$ (truth $= 0.107$)

Wavelet-Domain Bootstrapping of FD Series

- Haar DWT of FD(0.45) series $X$ (left-hand column) and wavelet-domain bootstrap thereof (right-hand)
MLEs of FD Parameters: I

- FD process depends on 2 parameters, namely, $\delta$ and $\sigma_\varepsilon^2$
- given $\mathbf{X} = [X_0, X_1, \ldots, X_{N-1}]^T$ with $N = 2^j$, suppose we want to estimate $\delta$ and $\sigma_\varepsilon^2$
- if $\mathbf{X}$ is stationary (i.e. $\delta < 1/2$) and multivariate Gaussian, can use the maximum likelihood (ML) method

MLEs of FD Parameters: II

- definition of Gaussian likelihood function:
  $$L(\delta, \sigma_\varepsilon^2 | \mathbf{X}) \equiv \frac{1}{(2\pi)^{N/2}} |\Sigma_\mathbf{X}|^{-1/2} e^{-\mathbf{X}^T \Sigma_\mathbf{X}^{-1} \mathbf{X}/2}$$
  where $\Sigma_\mathbf{X}$ is covariance matrix for $\mathbf{X}$, with $(s,t)$th element given by $s_{X,s-t}$, and $|\Sigma_\mathbf{X}|$ & $\Sigma_\mathbf{X}^{-1}$ denote determinant & inverse
- ML estimators of $\delta$ and $\sigma_\varepsilon^2$ maximize $L(\delta, \sigma_\varepsilon^2 | \mathbf{X})$ or, equivalently, minimize
  $$-2 \log (L(\delta, \sigma_\varepsilon^2 | \mathbf{X})) = N \log (2\pi) + \log (|\Sigma_\mathbf{X}|) + \mathbf{X}^T \Sigma_\mathbf{X}^{-1} \mathbf{X}$$
- exact MLEs computationally intensive, mainly because of the need to deal with $|\Sigma_\mathbf{X}|$ and $\Sigma_\mathbf{X}^{-1}$
- good approximate MLEs of considerable interest

MLEs of FD Parameters: III

- key ideas behind first wavelet-based approximate MLEs
  - have seen that we can approximate FD time series $\mathbf{X}$ by $\mathbf{Y} = \mathbf{W}^T \Lambda^{1/2} \mathbf{Z}$, where $\Lambda^{1/2}$ is a diagonal matrix, all of whose diagonal elements are positive
  - since covariance matrix for $\mathbf{Z}$ is $I_N$, the one for $\mathbf{Y}$ is
    $$\mathbf{W}^T \Lambda^{1/2} I_N \mathbf{W}^T \Lambda^{1/2} \mathbf{W} = \mathbf{W}^T \Lambda \mathbf{W} \equiv \widetilde{\Sigma}_\mathbf{X},$$
    where $\Lambda \equiv \Lambda^{1/2} \Lambda^{1/2}$ is also diagonal
  - can consider $\widetilde{\Sigma}_\mathbf{X}$ to be an approximation to $\Sigma_\mathbf{X}$
  - leads to approximation of log likelihood:
    $$-2 \log (L(\delta, \sigma_\varepsilon^2 | \mathbf{X})) \approx N \log (2\pi) + \log (|\widetilde{\Sigma}_\mathbf{X}|) + \mathbf{X}^T \widetilde{\Sigma}_\mathbf{X}^{-1} \mathbf{X}$$

MLEs of FD Parameters: IV

- Q: so how does this help us?
  - easy to invert $\widetilde{\Sigma}_\mathbf{X}$:
    $$\widetilde{\Sigma}_\mathbf{X}^{-1} = (\mathbf{W}^T \Lambda \mathbf{W})^{-1} = (\Lambda)^{-1} (\mathbf{W}^T)^{-1} = \mathbf{W}^T \Lambda^{-1} \mathbf{W},$$
    where $\Lambda^{-1}$ is another diagonal matrix, leading to
    $$\mathbf{X}^T \widetilde{\Sigma}_\mathbf{X}^{-1} \mathbf{X} = \mathbf{X}^T \mathbf{W}^T \Lambda^{-1} \mathbf{W} \mathbf{X} = \mathbf{W}^T \Lambda^{-1} \mathbf{W}$$
  - easy to compute the determinant of $\widetilde{\Sigma}_\mathbf{X}$:
    $$|\widetilde{\Sigma}_\mathbf{X}| = |\mathbf{W}^T \Lambda \mathbf{W}| = |\Lambda \mathbf{W}^T \mathbf{W}| = |\Lambda I_N| = |\Lambda|,$$
    and the determinant of a diagonal matrix is just the product of its diagonal elements
MLEs of FD Parameters: V

- define the following three functions of \( \delta \):
  \[
  C'_j(\delta) \equiv \int_{1/2^{j+1}}^{1/2^j} \frac{2^{j+1}}{2\pi f} \frac{df}{\delta} \approx \int_{1/2^{j+1}}^{1/2^j} \frac{2^{j+1}}{2\pi f} \frac{df}{2^{2\delta}} \\
  C'_{J+1}(\delta) \equiv \frac{N\Gamma(1-2\delta)}{\Gamma^2(1-\delta)} - \sum_{j=1}^{J} \frac{N}{2} C'_j(\delta) \\
  \sigma^2_\delta(\delta) \equiv \frac{1}{N} \left( \frac{V^2_{j,0}}{C'_{J+1}(\delta)} + \sum_{j=1}^{J} \frac{1}{2} \frac{N}{2} \sum_{t=0}^{N-1} W^2_{j,t} \right)
  \]

Other Wavelet-Based Estimators of FD Parameters

- second MLE approach: formulate likelihood directly in terms of nonboundary wavelet coefficients
  - handles stationary or nonstationary FD processes
    (i.e., need not assume \( \delta < 1/2 \))
  - handles certain deterministic trends
- alternative to MLEs are least square estimators (LSEs)
  - recall that, for large \( \tau \) and for \( \beta = 2\delta - 1 \), have
    \[
    \log (\nu^2_X(\tau_j)) \approx \zeta + \beta \log (\tau_j)
    \]
  - suggests determining \( \delta \) by regressing \( \log (\nu^2_X(\tau_j)) \) on \( \log (\tau_j) \)
    over range of \( \tau_j \)
  - weighted LSE takes into account fact that variance of \( \log (\nu^2_X(\tau_j)) \)
    depends upon scale \( \tau_j \) (increases as \( \tau_j \) increases)

MLEs of FD Parameters: VI

- wavelet-based approximate MLE \( \tilde{\delta} \) for \( \delta \) is the value that minimizes the following function of \( \delta \):
  \[
  \tilde{I}(\delta | X) \equiv N \log(\sigma^2_\delta(\delta)) + \log(C'_{J+1}(\delta)) + \sum_{j=1}^{J} \frac{N}{2} \log(C'_j(\delta))
  \]
- once \( \tilde{\delta} \) has been determined, MLE for \( \sigma^2_\epsilon \) is given by \( \sigma^2_\epsilon(\tilde{\delta}) \)
- computer experiments indicate scheme works quite well

Homogeneity of Variance: I

- because DWT decorrelates FD and related processes, nonboundary coefficients in \( W_j \) should resemble white noise; i.e.,
  \[
  \text{cov} \{ W_{j,t}, W_{j',t'} \} \approx 0
  \]
  when \( t \neq t' \), and \( \{ W_{j,t} \} \) should not depend upon \( t \)
- can test for homogeneity of variance in \( X \) using \( W_j \) over a range of levels \( j \)
- suppose \( U_0, \ldots, U_{N-1} \) are independent normal RVs with \( E\{U_t\} = 0 \) and \( \text{var} \{U_t\} = \sigma^2_i \)
- want to test null hypothesis \( H_0 : \sigma^2_0 = \sigma^2_1 = \cdots = \sigma^2_{N-1} \)
- can test \( H_0 \) versus a variety of alternatives, e.g.,
  \[
  H_1 : \sigma^2_0 = \cdots = \sigma^2_k \neq \sigma^2_{k+1} = \cdots = \sigma^2_{N-1}
  \]
  using normalized cumulative sum of squares
Homogeneity of Variance: II

- to define test statistic $D$, start with
  $$P_k \equiv \frac{\sum_{j=0}^{k} U_j^2}{\sum_{j=0}^{N-1} U_j^2}, \quad k = 0, \ldots, N - 2$$
  and then compute $D : = \max (D^+, D^-)$, where
  $$D^+ \equiv \max_{0 \leq k \leq N-2} \left( \frac{k+1}{N-1} - P_k \right) \quad \& \quad D^- \equiv \max_{0 \leq k \leq N-2} \left( P_k - \frac{k}{N-1} \right)$$

- can reject $H_0$ if observed $D$ is ‘too large,’ where ‘too large’ is quantified by considering distribution of $D$ under $H_0$
- need to find critical value $x_\alpha$ such that $P[D \geq x_\alpha] = \alpha$ for, e.g., $\alpha = 0.01, 0.05$ or $0.1$

Homogeneity of Variance: III

- once determined, can perform $\alpha$ level test of $H_0$:
  - compute $D$ statistic from data $U_0, \ldots, U_{N-1}$
  - reject $H_0$ at level $\alpha$ if $D \geq x_\alpha$
  - fail to reject $H_0$ at level $\alpha$ if $D < x_\alpha$
- can determine critical values $x_\alpha$ in two ways
  - Monte Carlo simulations
  - large sample approximation to distribution of $D$:
    $$\Pr[(N/2)^{1/2}D \geq x] \approx 1 + 2 \sum_{l=1}^{\infty} (-1)^l e^{-2l^2x^2}$$
    (reasonable approximation for $N \geq 128$)

Homogeneity of Variance: IV

- idea: given time series $\{X_t\}$, compute $D$ using nonboundary wavelet coefficients $W_{j,t}$ (there are $M_j' \equiv N_j - L_j'$ of these):
  $$P_k \equiv \frac{\sum_{t=L_j'} W_{j,t}^2}{\sum_{t=L_j'}^{N_j-1} W_{j,t}^2}, \quad k = L_j', \ldots, N_j - 2$$

- if null hypothesis rejected at level $j$, can use nonboundary MODWT coefficients to locate change point based on
  $$\tilde{P}_k \equiv \frac{\sum_{t=L_j-1}^{k} \tilde{W}_{j,t}^2}{\sum_{t=L_j-1}^{N_j-1} \tilde{W}_{j,t}^2}, \quad k = L_j - 1, \ldots, N - 2$$
  along with analogs $\tilde{D}_k^+$ and $\tilde{D}_k^-$ of $D_k^+$ and $D_k^-$

Example – Annual Minima of Nile River: I

- left-hand plot: annual minima of Nile River
- new measuring device introduced around year 715
- right: Haar $\tilde{\psi}_\chi^2(\tau_j)$ before (x’s) and after (o’s) year 715.5, with 95% confidence intervals based upon $\chi^2_{121}$ approximation
Example – Annual Minima of Nile River: II

- based upon last 512 values (years 773 to 1284), plot shows $\hat{l}(\delta \mid \mathbf{X})$ versus $\delta$ for the first wavelet-based approximate MLE using the LA(8) wavelet (upper curve) and corresponding curve for exact MLE (lower)
  - wavelet-based approximate MLE is value minimizing upper curve: $\hat{\delta} = 0.4532$
  - exact MLE is value minimizing lower curve: $\hat{\delta} = 0.4452$

Example – Annual Minima of Nile River: III

- results of testing all Nile River minima for homogeneity of variance using the Haar wavelet filter with critical values determined by computer simulations

<table>
<thead>
<tr>
<th>$\tau_j$</th>
<th>$M_j'$</th>
<th>$D$</th>
<th>10%</th>
<th>5%</th>
<th>1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year</td>
<td>331</td>
<td>0.1559</td>
<td>0.0945</td>
<td>0.1051</td>
<td>0.1262</td>
</tr>
<tr>
<td>2 years</td>
<td>165</td>
<td>0.1754</td>
<td>0.1320</td>
<td>0.1469</td>
<td>0.1765</td>
</tr>
<tr>
<td>4 years</td>
<td>82</td>
<td>0.1000</td>
<td>0.1855</td>
<td>0.2068</td>
<td>0.2474</td>
</tr>
<tr>
<td>8 years</td>
<td>41</td>
<td>0.2313</td>
<td>0.2572</td>
<td>0.2864</td>
<td>0.3436</td>
</tr>
</tbody>
</table>

- can reject null hypothesis of homogeneity of variance at level of significance 0.05 for scales $\tau_1$ & $\tau_2$, but not at larger scales

Example – Annual Minima of Nile River: IV

- Nile River minima (top plot) along with curves (constructed per Equation (382)) for scales $\tau_1$ & $\tau_2$ (middle & bottom) to identify change point via time of maximum deviation (vertical lines denote year 715)

Summary

- DWT approximately decorrelate certain time series, including ones coming from FD and related processes
- leads to schemes for simulating time series and bootstrapping
- also leads to schemes for estimating parameters of FD process
  - approximate maximum likelihood estimators (two varieties)
  - weighted least squares estimator
- can also devise wavelet-based tests for
  - homogeneity of variance
  - trends (see Craigmile et al., 2004, for details)
References: I

• fractionally differenced processes
  – wavelet cross-covariance and cross-correlation
• asymptotic theory for non-Gaussian processes

References: II

• biased estimators of wavelet variance
• unbiased estimator of wavelet variance for ‘gappy’ time series
  – robust estimation
  – wavelet variance for random fields
  – wavelet-based characteristic scales

References: III

• wavelet-based denoising

References: IV

• bootstrapping
References: V

- decorrelation property of DWTs

- parameter estimation for FD processes

- testing for homogeneity of variance

- wavelet-based trend assessment