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)

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

Decomposing Sample Variance of Time Series

- let $X_0, X_1, \ldots, X_{N-1}$ represent time series with N values
- let $\overline{X}$ denote sample mean of X_t's: $\overline{X} \equiv \frac{1}{N}\sum_{t=0}^{N-1} X_t$
- let $\hat{\sigma}_X^2$ denote sample variance of X_t's:

$$\hat{\sigma}_X^2 \equiv \frac{1}{N}\sum_{t=0}^{N-1}\left(X_t - \overline{X}\right)^2$$

- idea is to decompose (analyze, break up) $\hat{\sigma}_X^2$ 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

Empirical Wavelet Variance

- define empirical wavelet variance for scale $\tau_j \equiv 2^{j-1}$ as

$$\tilde{\nu}_X^2(\tau_j) \equiv \frac{1}{N}\sum_{t=0}^{N-1} \widetilde{W}_{j,t}^2, \text{ where } \widetilde{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N}$$

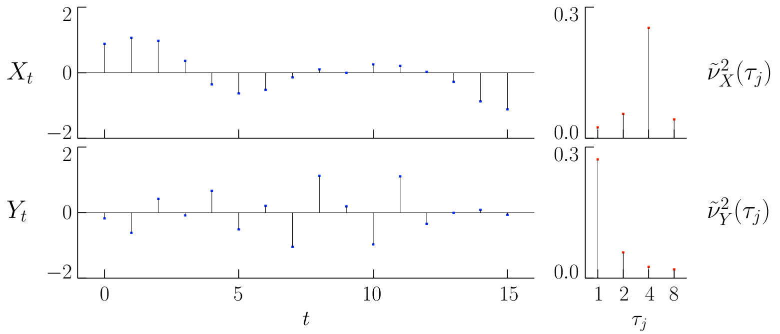
- if $N = 2^J$, obtain analysis (decomposition) of sample variance:

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{t=0}^{N-1}\left(X_t - \overline{X}\right)^2 = \sum_{j=1}^{J} \tilde{\nu}_X^2(\tau_j)$$

  (if N not a power of 2, can analyze variance to any level J_0, but need additional component involving scaling coefficients)
- interpretation: $\tilde{\nu}_X^2(\tau_j)$ is portion of $\hat{\sigma}_X^2$ due to changes in averages over scale τ_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



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 jth level equivalent MODWT filter $\{\tilde{h}_{j,l}\}$ to $\{X_t\}$ to create a new stochastic process:

$$\overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}, \quad t \in \mathbb{Z},$$

which should be contrasted with

$$\widetilde{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1$$

Theoretical Wavelet Variance: II

- if Y is any RV, let $E\{Y\}$ denote its expectation
- let var $\{Y\}$ denote its variance: var $\{Y\} \equiv E\{(Y - E\{Y\})^2\}$
- definition of time dependent wavelet variance:

$$\nu_{X,t}^2(\tau_j) \equiv \text{var}\,\{\overline{W}_{j,t}\},$$

  with conditions on X_t so that var $\{\overline{W}_{j,t}\}$ exists and is finite
- $\nu_{X,t}^2(\tau_j)$ depends on τ_j and t
- will focus on time independent wavelet variance

$$\nu_X^2(\tau_j) \equiv \text{var}\,\{\overline{W}_{j,t}\}$$

  (can adapt theory to handle time varying situation)
- $\nu_X^2(\tau_j)$ well-defined for stationary processes and certain related processes, so let's review concept of stationarity

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
  - $E\{X_t\} = \mu_X$ for all t, i.e., constant independent of t
  - $\text{cov}\{X_t, X_{t+\tau}\} = s_{X,\tau}$, i.e., depends on lag τ, but not t
- $s_{X,\tau}$, $\tau \in \mathbb{Z}$, 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

$$\sum_{j=1}^{J} \tilde{\nu}_X^2(\tau_j) = \hat{\sigma}_X^2$$

- $\nu_X^2(\tau_j)$ is thus contribution to $\text{var}\,\{X_t\}$ due to scale τ_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 τ; 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

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \text{var}\,\{X_t\} \left(\tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \cdots\right) = \text{var}\,\{X_t\},$$

as required

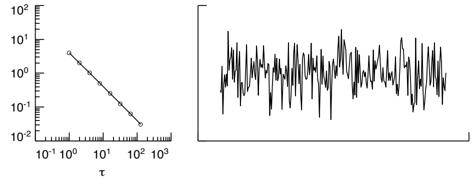
- note also that

$$\log\,(\nu_X^2(\tau_j)) \propto -\log\,(\tau_j),$$

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

$\delta = 0$
slope $= -1$



- $\nu_X^2(\tau_j)$ versus τ_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 τ_1
- note: later on, we will discuss fractionally differenced (FD) processes that are characterized by a parameter δ; 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, $\nu_X^2(\tau_j)$ 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

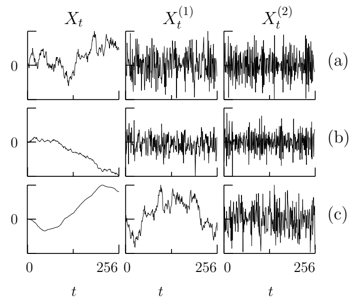
$$X_t^{(2)} = X_t^{(1)} - X_{t-1}^{(1)} = X_t - 2X_{t-1} + X_{t-2}$$

- X_t said to have dth 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)

Examples of Processes with Stationary Increments



- 1st column shows, from top to bottom, realizations from
  (a) random walk: $X_t = \sum_{u=1}^{t} \epsilon_u$, & ϵ_t is zero mean white noise
  (b) like (a), but now ϵ_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 Processes with Stationary Backward Differences: I

- let $\{X_t\}$ be nonstationary with dth order stationary differences
- if we use a Daubechies wavelet filter of width L satisfying $L \geq 2d$, then $\nu_X^2(\tau_j)$ is well-defined and finite for all τ_j, but now

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \infty$$

- works because there is a backward difference operator of order $d = L/2$ embedded within $\{\tilde{h}_{j,l}\}$, 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

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 ($L = 2$),

$$\nu_X^2(\tau_j) = \frac{\text{var}\,\{\epsilon_t\}}{6}\left(\tau_j + \frac{1}{2\tau_j}\right) \approx \frac{\text{var}\,\{\epsilon_t\}}{6}\tau_j,$$

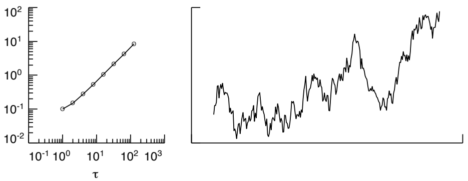
with the approximation becoming better as τ_j increases

- note that $\nu_X^2(\tau_j)$ increases as τ_j increases
- $\log\,(\nu_X^2(\tau_j)) \propto \log\,(\tau_j)$ approximately, so plot of $\log\,(\nu_X^2(\tau_j))$ vs. $\log\,(\tau_j)$ is approximately linear with a slope of $+1$
- as required, also have

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \frac{\text{var}\,\{\epsilon_t\}}{6}\left(1 + \tfrac{1}{2} + 2 + \tfrac{1}{4} + 4 + \tfrac{1}{8} + \cdots\right) = \infty$$

Wavelet Variance for Random Walk Process: II

$\delta = 1$
slope ≈ 1



- $\nu_X^2(\tau_j)$ versus τ_j 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 τ_1
- note: a fractionally differenced process with parameter $\delta = 1$ is the same as a random walk process

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(δ) process is determined by 2 parameters, namely, δ and σ_ϵ^2, where $-\infty < \delta < \infty$ and $\sigma_\epsilon^2 > 0$ (σ_ϵ^2 is less important than δ)
- 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., cov $\{X_t, X_{t+1}\} < 0$)

Fractionally Differenced (FD) Processes: II

- if $\delta \geq 1/2$, FD process $\{X_t\}$ is nonstationary with dth 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

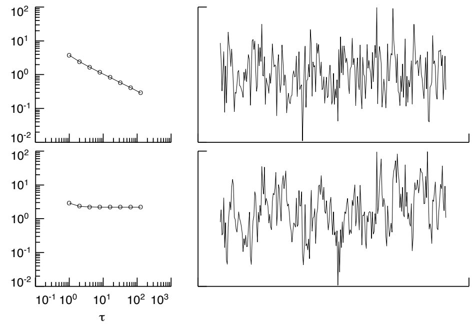
$$\nu_X^2(\tau_j) \approx C\tau_j^{2\delta-1}$$

- 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. τ_j looks approximately linear with slope $2\delta - 1$ for τ_j large enough

LA(8) Wavelet Variance for 2 FD Processes

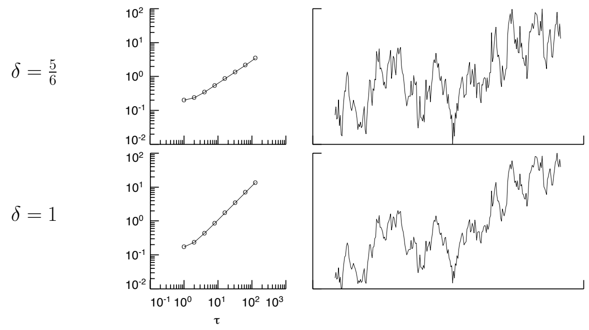
$\delta = \frac{1}{4}$

$\delta = \frac{1}{2}$



- 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 dth 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)

Wavelet Variance for Processes with Stationary Backward Differences: III

- if $\{X_t\}$ is nonstationary, then

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \infty$$

- with a suitable construction, we can take variance of nonstationary process with dth order stationary increments to be ∞
- using this construction, we have

$$\sum_{j=1}^{\infty} \nu_X^2(\tau_j) = \text{var}\,\{X_t\}$$

for both the stationary and nonstationary cases

Background on Gaussian Random Variables

- $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian (normal) RV with mean μ and variance σ^2
- will write

$$X \stackrel{\text{d}}{=} \mathcal{N}(\mu, \sigma^2)$$

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

$$\Phi(z) \equiv \mathbf{P}[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{(2\pi)}} e^{-x^2/2}\, dx$$

- inverse $\Phi^{-1}(\cdot)$ of $\Phi(\cdot)$ is such that $\mathbf{P}[Z \leq \Phi^{-1}(p)] = p$
- $\Phi^{-1}(p)$ called $p \times 100\%$ percentage point

Background on Chi-Square Random Variables

- X said to be a chi-square RV with η degrees of freedom if its probability density function (PDF) is given by

$$f_X(x;\eta) = \frac{1}{2^{\eta/2}\Gamma(\eta/2)} x^{(\eta/2)-1} e^{-x/2}, \quad x \geq 0, \ \ \eta > 0$$

- χ^2_η denotes RV with above PDF
- if $Z_1, Z_2, \ldots, Z_\eta$ are independent standard Gaussian RVs, then

$$Z_1^2 + Z_2^2 + \cdots + Z_\eta^2 \stackrel{\mathrm{d}}{=} \chi^2_\eta$$

- two important facts: $E\{\chi^2_\eta\} = \eta$ and $\operatorname{var}\{\chi^2_\eta\} = 2\eta$
- let $Q_\eta(p)$ denote the pth percentage point for the RV χ^2_η:

$$\mathbf{P}[\chi^2_\eta \leq Q_\eta(p)] = p$$

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 dth 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
  - $E\{\overline{W}_{j,t}\} = 0$ if either (i) $L > 2d$ or (ii) $L = 2d$ and $\mu_Y = 0$
  - $E\{\overline{W}_{j,t}\} \neq 0$ if $\mu_Y \neq 0$ 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 dth 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

$$\nu^2_X(\tau_j) = \operatorname{var}\{\overline{W}_{j,t}\} = E\{\overline{W}^2_{j,t}\}$$

- can base estimator on squares of

$$\widetilde{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N}, \quad t = 0, 1, \ldots, N-1$$

- recall that

$$\overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}, \qquad t \in \mathbb{Z}$$

Unbiased Estimator of Wavelet Variance: II

- comparing

$$\widetilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l \bmod N} \ \text{ with } \ \overline{W}_{j,t} \equiv \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l}$$

  says that $\widetilde{W}_{j,t} = \overline{W}_{j,t}$ 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

$$\hat{\nu}^2_X(\tau_j) \equiv \frac{1}{N - L_j + 1} \sum_{t=L_j-1}^{N-1} \widetilde{W}^2_{j,t} = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}^2_{j,t},$$

  where $M_j \equiv N - L_j + 1$

Statistical Properties of $\hat{\nu}_X^2(\tau_j)$

- assume that $\{\overline{W}_{j,t}\}$ 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.,

$$\frac{\hat{\nu}_X^2(\tau_j) - \nu_X^2(\tau_j)}{(2A_j/M_j)^{1/2}} = \frac{M_j^{1/2}(\hat{\nu}_X^2(\tau_j) - \nu_X^2(\tau_j))}{(2A_j)^{1/2}} \stackrel{\mathrm{d}}{=} \mathcal{N}(0,1)$$

approximately for large $M_j \equiv N - L_j + 1$

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

$$\hat{A}_j \equiv \frac{\left(\hat{s}_{j,0}^{(p)}\right)^2}{2} + \sum_{\tau=1}^{M_j-1} \left(\hat{s}_{j,\tau}^{(p)}\right)^2,$$

is approximately unbiased, where

$$\hat{s}_{j,\tau}^{(p)} \equiv \frac{1}{M_j} \sum_{t=L_j-1}^{N-1-|\tau|} \widetilde{W}_{j,t}\widetilde{W}_{j,t+|\tau|}, \quad 0 \le |\tau| \le M_j - 1$$

- Monte Carlo results: $\hat{A}_j$ reasonably good for $M_j \ge 128$

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

- based upon large sample theory, can form a $100(1-2p)\%$ confidence interval (CI) for $\nu_X^2(\tau_j)$:

$$\left[\hat{\nu}_X^2(\tau_j) - \Phi^{-1}(1-p)\frac{\sqrt{2A_j}}{\sqrt{M_j}}, \hat{\nu}_X^2(\tau_j) + \Phi^{-1}(1-p)\frac{\sqrt{2A_j}}{\sqrt{M_j}}\right];$$

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 $\nu_X^2(\tau_j) \ge 0$ always
- can avoid this problem by using a χ^2 approximation

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

- χ_η^2 useful for approximating distribution of sum of squared Gaussian RVs, which is what we are dealing with here:

$$\hat{\nu}_X^2(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} \overline{W}_{j,t}^2$$

- idea is to assume $\hat{\nu}_X^2(\tau_j) \stackrel{\mathrm{d}}{=} a\chi_\eta^2$, where a and η are constants to be set via moment matching
- because $E\{\chi_\eta^2\} = \eta$ and $\operatorname{var}\{\chi_\eta^2\} = 2\eta$, we have $E\{a\chi_\eta^2\} = a\eta$ and $\operatorname{var}\{a\chi_\eta^2\} = 2a^2\eta$
- can equate $E\{\hat{\nu}_X^2(\tau_j)\}$ & $\operatorname{var}\{\hat{\nu}_X^2(\tau_j)\}$ to $a\eta$ & $2a^2\eta$ to determine a & η

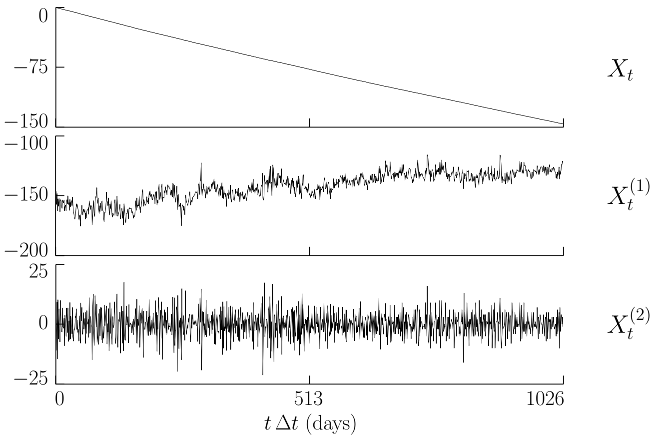
Confidence Intervals for $\nu_X^2(\tau_j)$: III

- obtain
$$\eta = \frac{2\left(E\{\hat{\nu}_X^2(\tau_j)\}\right)^2}{\mathrm{var}\,\{\hat{\nu}_X^2(\tau_j)\}} = \frac{2\nu_X^4(\tau_j)}{\mathrm{var}\,\{\hat{\nu}_X^2(\tau_j)\}} \text{ and } a = \frac{\nu_X^2(\tau_j)}{\eta}$$
- after η has been determined, can obtain a CI for $\nu_X^2(\tau_j)$: with probability $1 - 2p$, the random interval
$$\left[\frac{\eta\hat{\nu}_X^2(\tau_j)}{Q_\eta(1-p)}, \frac{\eta\hat{\nu}_X^2(\tau_j)}{Q_\eta(p)}\right]$$
traps the true unknown $\nu_X^2(\tau_j)$
- lower limit is now nonnegative
- as $N \to \infty$, above CI and Gaussian-based CI converge

Three Ways to Set η

1. use large sample theory with appropriate estimates:
$$\eta = \frac{2\nu_X^4(\tau_j)}{\mathrm{var}\,\{\hat{\nu}_X^2(\tau_j)\}} \approx \frac{2\nu_X^4(\tau_j)}{2A_j/M_j} \text{ suggests } \hat{\eta}_1 = \frac{M_j\hat{\nu}_X^4(\tau_j)}{\hat{A}_j}$$
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)

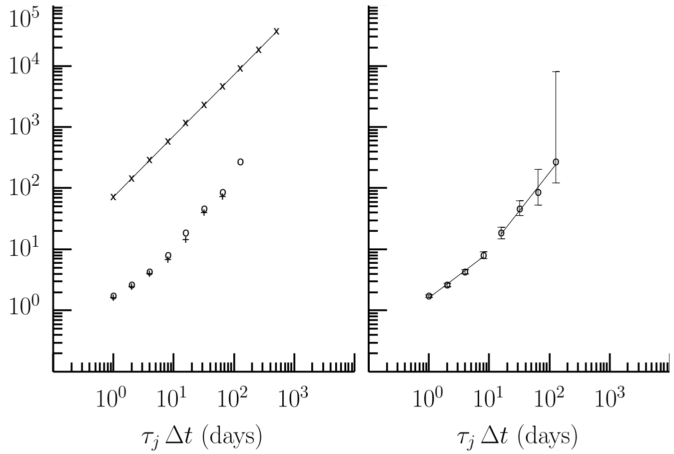
Atomic Clock Deviates: I



Atomic Clock Deviates: II

- 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 dth order stationary increments, need $L \geq 2d$, but might need $L > 2d$ to get $E\{\overline{W}_{j,t}\} = 0$
- 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 $L \geq 4$

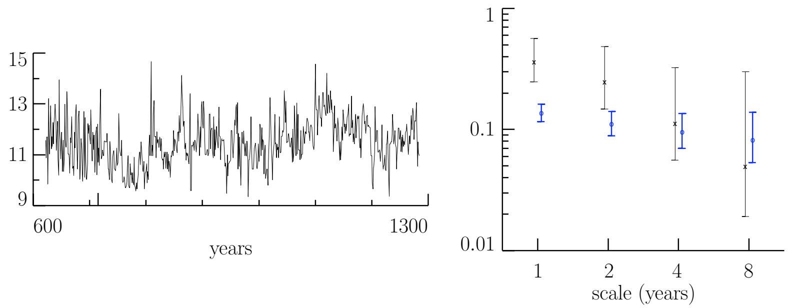
Atomic Clock Deviates: III



Atomic Clock Deviates: IV

- 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 η given by $\hat{\eta}_1$ for $j = 1, \ldots, 6$ and by η_3 for $j = 7$ & 8

Annual Minima of Nile River

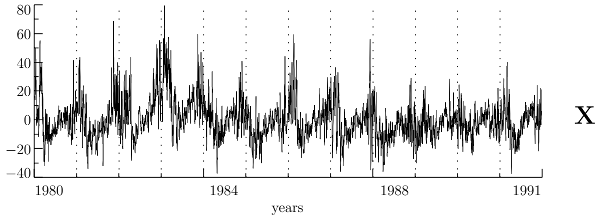


- left-hand plot: annual minima of Nile River
- right: Haar $\hat{\nu}_X^2(\tau_j)$ before (`x`'s) and after (`o`'s) year 715.5, with 95% confidence intervals based upon $\chi^2_{\eta_3}$ approximation

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

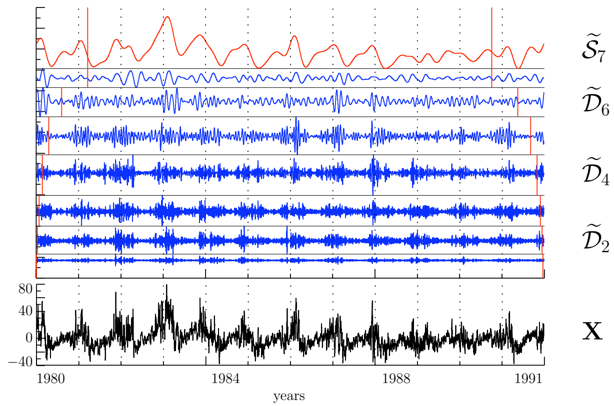
- each wavelet coefficient $\widetilde{W}_{j,t}$ formed using portion of X_t
- suppose X_t associated with actual time $t_0 + t\,\Delta t$
  - t_0 is actual time of first observation X_0
  - Δt is spacing between adjacent observations
- suppose $\tilde{h}_{j,l}$ is least asymmetric Daubechies wavelet
- can associate $\widetilde{W}_{j,t}$ with an interval of width $2\tau_j\,\Delta t$ centered at
$$t_0 + (2^j(t+1) - 1 - |\nu_j^{(H)}| \bmod N)\,\Delta t,$$
where, e.g., $|\nu_j^{(H)}| = [7(2^j - 1) + 1]/2$ for LA(8) wavelet
- can thus form 'localized' wavelet variance analysis (implicitly assumes stationarity or stationary increments locally)

Subtidal Sea Level Fluctuations: I



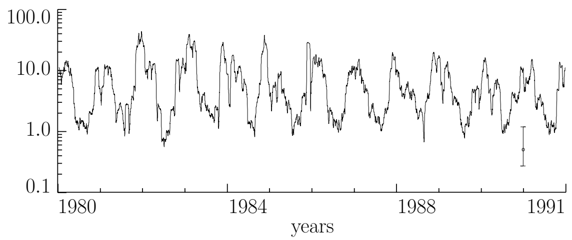
- subtidal sea level fluctuations **X** for Crescent City, CA, collected by National Ocean Service with permanent tidal gauge
- $N = 8746$ values from Jan 1980 to Dec 1991 (almost 12 years)
- one value every 12 hours, so $\Delta t = 1/2$ day
- 'subtidal' is what remains after diurnal & semidiurnal tides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

Subtidal Sea Level Fluctuations: II



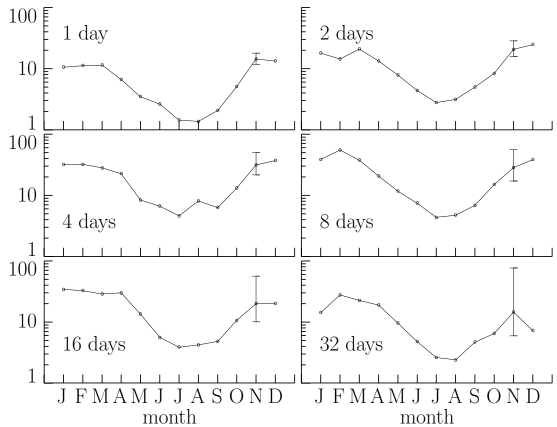
- level $J_0 = 7$ LA(8) MODWT multiresolution analysis

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 $\nu = 15.25$

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

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. $\mathbf{D}$, an N dimensional deterministic signal
  2. $\mathbf{C}$, an N dimensional stochastic signal; i.e., a vector of random variables (RVs) with covariance matrix $\Sigma_{\mathbf{C}}$
- will consider two types of noise:
  1. $\boldsymbol{\epsilon}$, an N dimensional vector of independent and identically distributed (IID) RVs with mean 0 and covariance matrix $\Sigma_{\boldsymbol{\epsilon}} = \sigma_{\epsilon}^2 I_N$
  2. $\boldsymbol{\eta}$, an N dimensional vector of non-IID RVs with mean 0 and covariance matrix $\Sigma_{\boldsymbol{\eta}}$
     - 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 $\mathbf{X}$
  1. $\mathbf{X} = \mathbf{D} + \boldsymbol{\epsilon}$
  2. $\mathbf{X} = \mathbf{D} + \boldsymbol{\eta}$
  3. $\mathbf{X} = \mathbf{C} + \boldsymbol{\epsilon}$
  4. $\mathbf{X} = \mathbf{C} + \boldsymbol{\eta}$
- in the latter two cases, the stochastic signal $\mathbf{C}$ is assumed to be independent of the associated noise

Signal Representation via Wavelets: I

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

Signal Representation via Wavelets: II

- let d_j be the jth transform coefficient in $\mathbf{d} = \mathcal{O}\mathbf{D}$
- 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 $\mathbf{d} = [-3, 1, 4, -7, 2, -1]^T$, then $d_{(0)} = d_3 = -7$, $d_{(1)} = d_2 = 4$, $d_{(2)} = d_0 = -3$ etc.
- define a normalized partial energy sequence (NPES):

$$C_{M-1} \equiv \frac{\sum_{j=0}^{M-1} |d_{(j)}|^2}{\sum_{j=0}^{N-1} |d_{(j)}|^2} = \frac{\text{energy in largest } M \text{ terms}}{\text{total energy in signal}}$$

- let $\mathcal{I}_M$ be $N \times N$ diagonal matrix whose jth diagonal term is 1 if $|d_j|$ is one of the M largest magnitudes and is 0 otherwise

Signal Representation via Wavelets: III

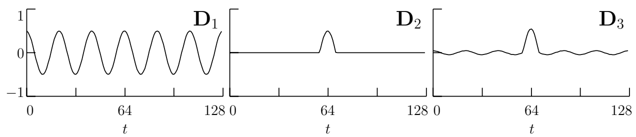
- form $\widehat{\mathbf{D}}_M \equiv \mathcal{O}^T \mathcal{I}_M \mathbf{d}$, which is an approximation to $\mathbf{D}$
- when $\mathbf{d} = [-3, 1, 4, -7, 2, -1]^T$ and $M = 3$, we have

$$\mathcal{I}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \text{ and thus } \widehat{\mathbf{D}}_M = \mathcal{O}^T \begin{bmatrix} -3 \\ 0 \\ 4 \\ -7 \\ 0 \\ 0 \end{bmatrix}$$

- one interpretation for NPES:

$$C_{M-1} = 1 - \frac{\|\mathbf{D} - \widehat{\mathbf{D}}_M\|^2}{\|\mathbf{D}\|^2} = 1 - \text{relative approximation error}$$

Signal Representation via Wavelets: IV



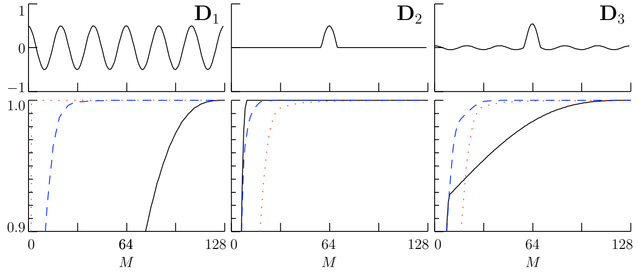
- consider three signals plotted above
- $\mathbf{D}_1$ is a sinusoid, which can be represented succinctly by the discrete Fourier transform (DFT)
- $\mathbf{D}_2$ is a bump (only a few nonzero values in the time domain)
- $\mathbf{D}_3$ is a linear combination of $\mathbf{D}_1$ and $\mathbf{D}_2$

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 $\mathcal{W}$ (wavelet)
- # of terms M needed to achieve relative error $< 1\%$:

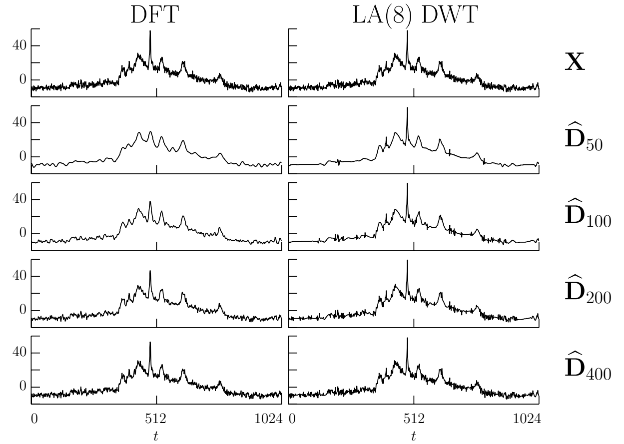
	$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$
DFT	2	29	28
identity	105	9	75
LA(8) wavelet	22	14	21

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)

Signal Representation via Wavelets: IX



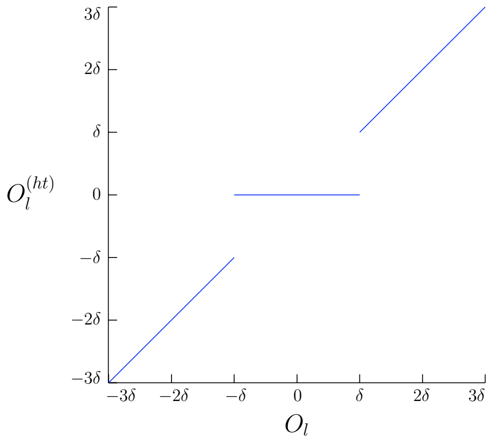
- example: DFT $\widehat{\mathbf{D}}_M$ (left-hand column) & $J_0 = 6$ LA(8) DWT $\widehat{\mathbf{D}}_M$ (right) for NMR series $\mathbf{X}$ (A. Maudsley, UCSF)

Signal Estimation via Thresholding: I

- thresholding schemes involve
  1. computing $\mathbf{O} \equiv \mathcal{O}\mathbf{X}$
  2. defining $\mathbf{O}^{(t)}$ as vector with lth element
  $$O_l^{(t)} = \begin{cases} 0, & \text{if } |O_l| \le \delta; \\ \text{some nonzero value}, & \text{otherwise}, \end{cases}$$
  where nonzero values are yet to be defined
  3. estimating $\mathbf{D}$ via $\widehat{\mathbf{D}}^{(t)} \equiv \mathcal{O}^T\mathbf{O}^{(t)}$
- simplest scheme is 'hard thresholding' ('kill/keep' strategy):
$$O_l^{(ht)} = \begin{cases} 0, & \text{if } |O_l| \le \delta; \\ O_l, & \text{otherwise.} \end{cases}$$

Hard Thresholding Function

- plot shows mapping from O_l to $O_l^{(ht)}$



Signal Estimation via Thresholding: II

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

Signal Estimation via Thresholding: III

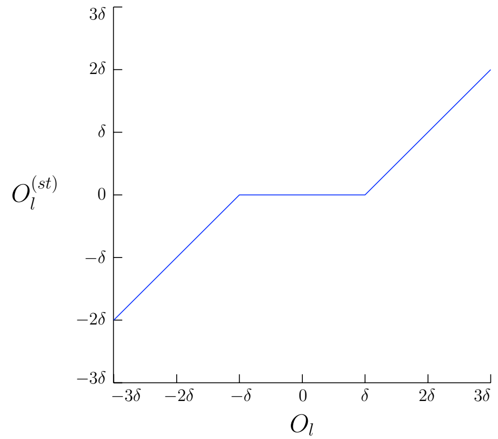
- $m\delta^2$ is a penalty for too many terms
  - rationale: heuristic says $\mathbf{d} = \mathcal{O}\mathbf{D}$ consists of just a few large coefficients
  - penalty increases as m increases
  - in minimizing γ_m, consideration of this term alone suggests that m should be small
- optimization problem: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

Signal Estimation via Thresholding: IV

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

Soft Thresholding Function

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



Signal Estimation via Thresholding: V

- third scheme is 'mid thresholding:'

$$O_l^{(mt)} = \operatorname{sign}\{O_l\}\,(|O_l| - \delta)_{++}\,,$$

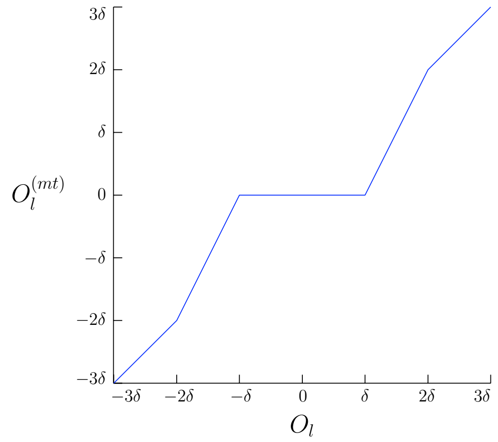
where

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

- provides compromise between hard and soft thresholding

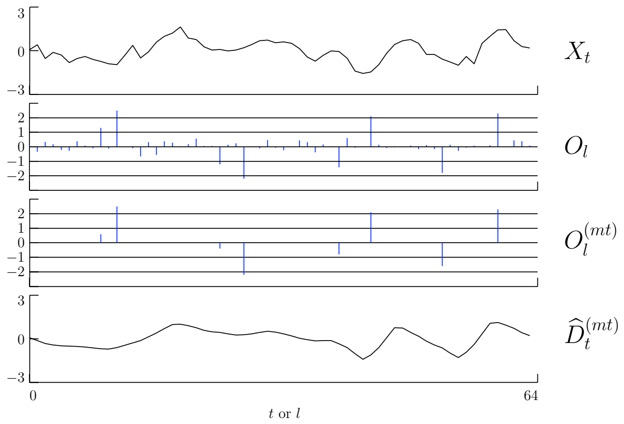
Mid Thresholding Function

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



Signal Estimation via Thresholding: VI

- example of mid thresholding with $\delta = 1$



Universal Threshold

- Q: how do we go about setting δ?
- specialize to IID Gaussian noise $\boldsymbol{\epsilon}$ with covariance $\sigma_\epsilon^2 I_N$
- can argue $\mathbf{e} \equiv \mathcal{O}\boldsymbol{\epsilon}$ is also IID Gaussian with covariance $\sigma_\epsilon^2 I_N$
- Donoho & Johnstone (1995) proposed $\delta^{(u)} \equiv \sqrt{[2\sigma_\epsilon^2 \log(N)]}$ ('log' here is 'log base e')
- rationale for $\delta^{(u)}$: because of Gaussianity, can argue that

$$\mathbf{P}\Big[\max_l\{|e_l|\} > \delta^{(u)}\Big] \leq \frac{1}{\sqrt{[4\pi \log(N)]}} \to 0 \text{ as } N \to \infty$$

and hence $\mathbf{P}\left[\max_l\{|e_l|\} \leq \delta^{(u)}\right] \to 1$ as $N \to \infty$, 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 σ_ϵ^2: $\mathbf{X} = \mathbf{D} + \boldsymbol{\epsilon}$
- using a DWT matrix $\mathcal{W}$, form $\mathbf{W} = \mathcal{W}\mathbf{X} = \mathcal{W}\mathbf{D} + \mathcal{W}\boldsymbol{\epsilon} \equiv \mathbf{d} + \mathbf{e}$
- because $\boldsymbol{\epsilon}$ 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 $\delta^{(u)} = \sqrt{[2\sigma_\epsilon^2 \log(N)]}$
  - 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 σ_ϵ is know, which is not usually the case
- if unknown, use median absolute deviation (MAD) scale estimator to estimate σ_ϵ using $\mathbf{W}_1$

$$\hat{\sigma}_{(\text{mad})} \equiv \frac{\text{median}\,\{|W_{1,0}|, |W_{1,1}|, \ldots, |W_{1,\frac{N}{2}-1}|\}}{0.6745}$$

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

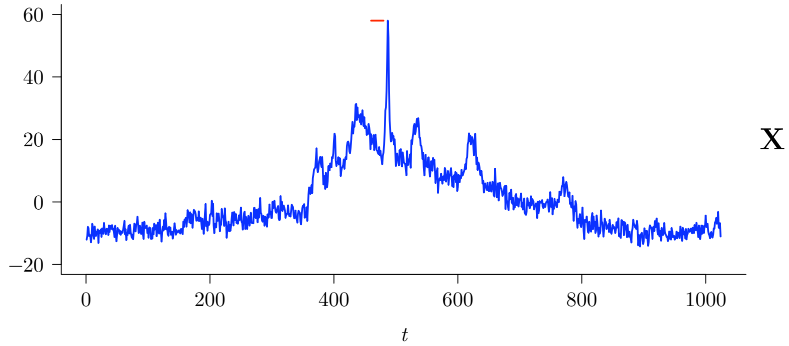
$$0.01, 0.30, 0.67, 0.80, 1.23, 1.33, 1.72, |a|, |b|$$

- median of these absolute deviations is 1.23, so

$$\hat{\sigma}_{(\text{mad})} = 1.23/0.6745 \doteq 1.82$$

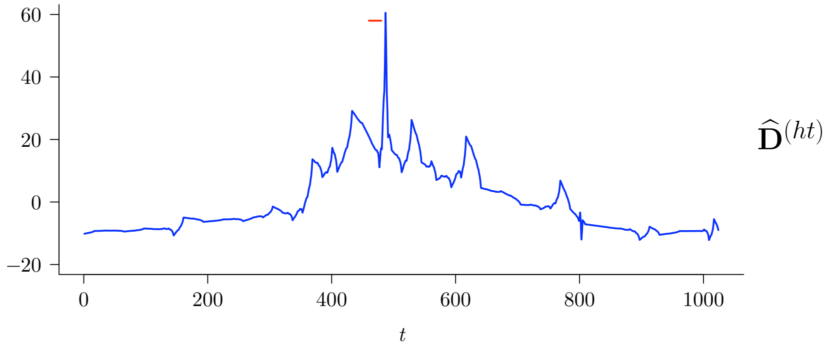
- $\hat{\sigma}_{(\text{mad})}$ not influenced adversely by a and b; i.e., scale estimate depends largely on the many small coefficients due to noise

Examples of DWT-Based Thresholding: I



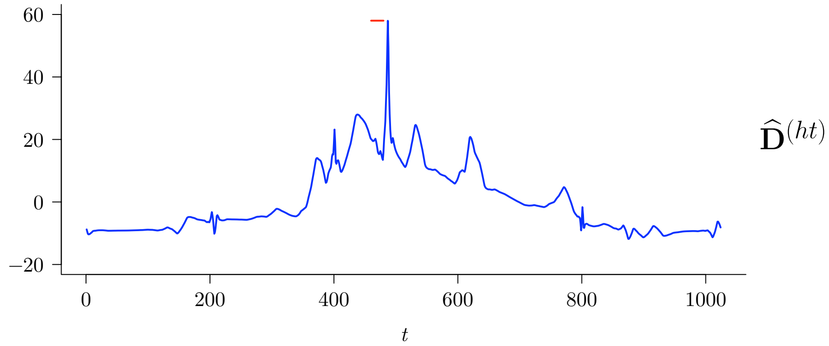
- NMR spectrum

Examples of DWT-Based Thresholding: II



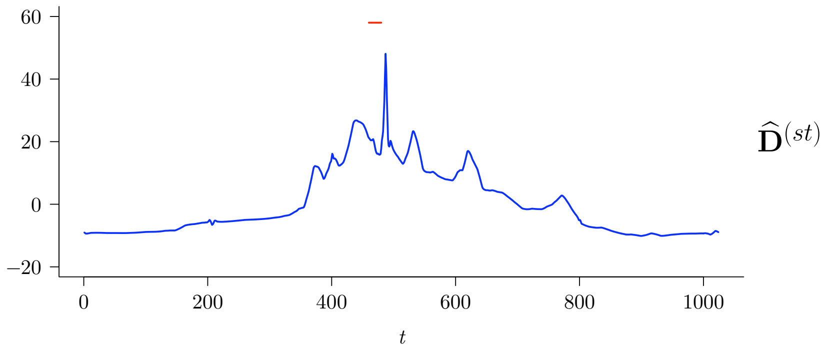
- signal estimate using $J_0 = 6$ partial D(4) DWT with hard thresholding and universal threshold level estimated by $\hat{\delta}^{(u)} = \sqrt{[2\hat{\sigma}^2_{(\text{mad})} \log(N)]} \doteq 6.49$

Examples of DWT-Based Thresholding: III



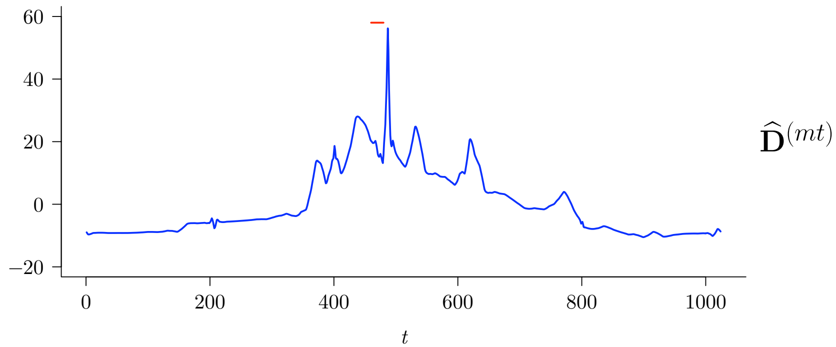
- same as before, but now using LA(8) DWT with $\hat{\delta}^{(u)} \doteq 6.13$

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

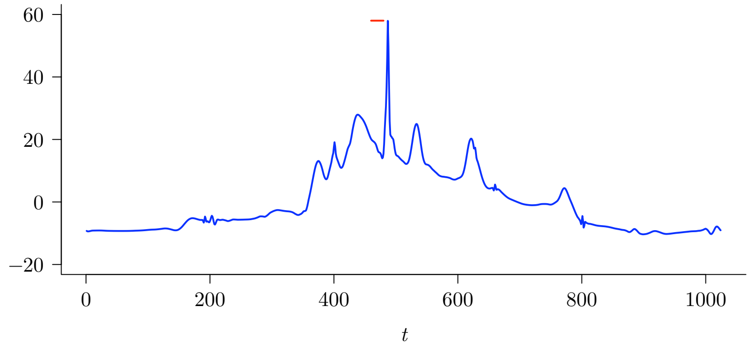
MODWT-Based Thresholding

- can base thresholding procedure on MODWT rather than DWT, yielding signal estimators $\widetilde{\mathbf{D}}^{(ht)}$, $\widetilde{\mathbf{D}}^{(st)}$ and $\widetilde{\mathbf{D}}^{(mt)}$
- because MODWT filters are normalized differently, universal threshold must be adjusted for each level:

$$\tilde{\delta}_j^{(u)} \equiv \sqrt{[2\tilde{\sigma}^2_{(\mathrm{mad})} \log{(N)}/2^j]},$$

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

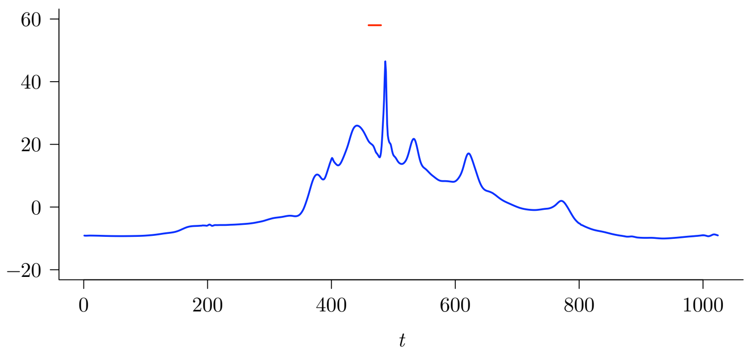
Examples of MODWT-Based Thresholding: I



$\widetilde{\mathbf{D}}^{(ht)}$

- signal estimate using $J_0 = 6$ LA(8) MODWT with hard thresholding

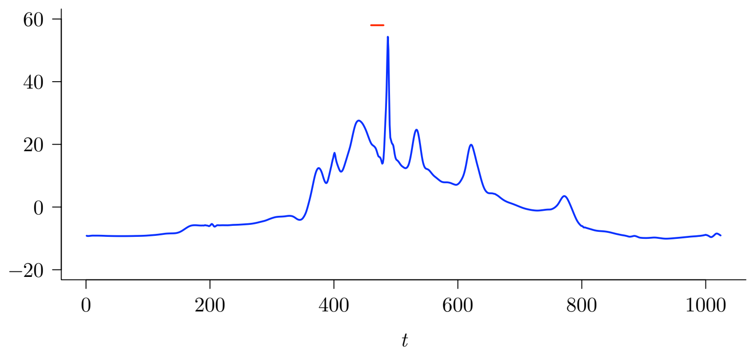
Examples of MODWT-Based Thresholding: II



$\widetilde{\mathbf{D}}^{(st)}$

- same as before, but now with soft thresholding

Examples of MODWT-Based Thresholding: III



$\widetilde{\mathbf{D}}^{(mt)}$

- 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

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

- $f_{Y|X=x}(\cdot)$ is a PDF, so its mean value is

$$E\{Y|X = x\} = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\, dy;$$

the above is called the conditional mean of Y, given X

WMTSA: 258–260

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 $U_2(X) = E\{Y|X\}$; 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 $a = E\{Y\}$)

WMTSA: 260

Conditional Mean Approach: I

- assume model of stochastic signal plus non-IID noise: $\mathbf{X} = \mathbf{C} + \boldsymbol{\eta}$ so that $\mathbf{O} = \mathcal{O}\mathbf{X} = \mathcal{O}\mathbf{C} + \mathcal{O}\boldsymbol{\eta} \equiv \mathbf{R} + \mathbf{n}$
- component-wise, have $O_l = R_l + n_l$
- because $\mathbf{C}$ and $\boldsymbol{\eta}$ are independent, $\mathbf{R}$ and $\mathbf{n}$ must be also
- suppose we approximate R_l via $\widehat{R}_l \equiv U_2(O_l)$, 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 $E\{R_l|O_l\}$, so let's work out what form this conditional mean takes
- to get $E\{R_l|O_l\}$, need the PDF of R_l given O_l, which is

$$f_{R_l|O_l=o_l}(r_l) = \frac{f_{R_l,O_l}(r_l, o_l)}{f_{O_l}(o_l)} = \frac{f_{R_l}(r_l) f_{n_l}(o_l - r_l)}{\int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(o_l - r_l)\, dr_l}$$

WMTSA: 408–409

Conditional Mean Approach: II

- mean value of $f_{R_l|O_l=o_l}(\cdot)$ yields estimator $\widehat{R}_l = E\{R_l|O_l\}$:

$$E\{R_l|O_l = o_l\} = \int_{-\infty}^{\infty} r_l f_{R_l|O_l=o_l}(r_l)\, dr_l = \frac{\int_{-\infty}^{\infty} r_l f_{R_l}(r_l) f_{n_l}(o_l - r_l) dr_l}{\int_{-\infty}^{\infty} f_{R_l}(r_l) f_{n_l}(o_l - r_l)\, dr_l}$$

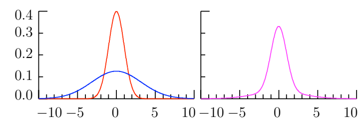
- to make further progress, we need a model for the wavelet-domain representation R_l of the signal
- heuristic that signal in the wavelet domain has a few large values and lots of small values suggests a Gaussian mixture model

Conditional Mean Approach: III

- let $\mathcal{I}_l$ be an RV such that $\mathbf{P}[\mathcal{I}_l = 1] = p_l$ & $\mathbf{P}[\mathcal{I}_l = 0] = 1 - p_l$
- under Gaussian mixture model, R_l has same distribution as

$$\mathcal{I}_l \mathcal{N}(0, \gamma_l^2 \sigma_{G_l}^2) + (1 - \mathcal{I}_l)\mathcal{N}(0, \sigma_{G_l}^2)$$

  where $\mathcal{N}(0, \sigma^2)$ is a Gaussian RV with mean 0 and variance σ^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 $p_l = 0.75$

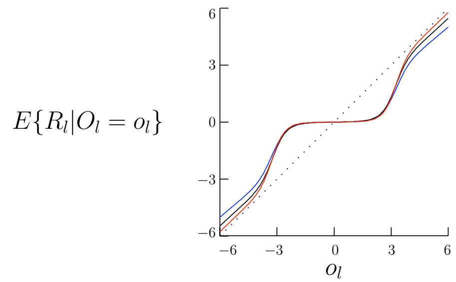


Conditional Mean Approach: VI

- 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 $(1 - \mathcal{I}_l)\mathcal{N}(0, \sigma_{G_l}^2)$
- to complete model, let n_l obey a Gaussian distribution with mean 0 and variance $\sigma_{n_l}^2$
- conditional mean estimator becomes $E\{R_l|O_l = o_l\} = \frac{b_l}{1+c_l} o_l$, where

$$c_l = \frac{p_l \sqrt{(\sigma_{G_l}^2 + \sigma_{n_l}^2)}}{(1 - p_l)\sigma_{n_l}} e^{-o_l^2 b_l / (2\sigma_{n_l}^2)}$$

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

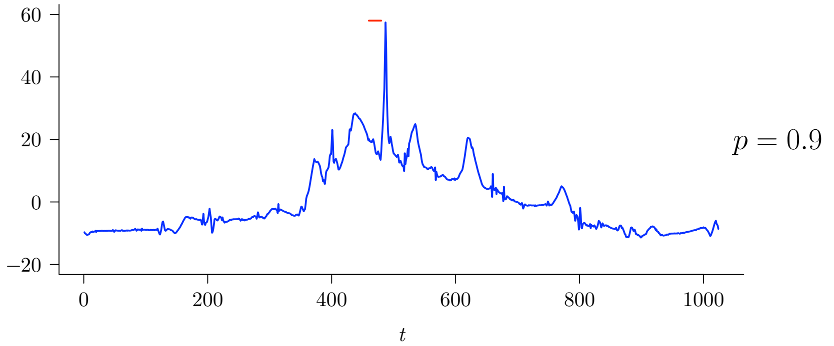
Wavelet-Based Shrinkage: I

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

Wavelet-Based Shrinkage: II

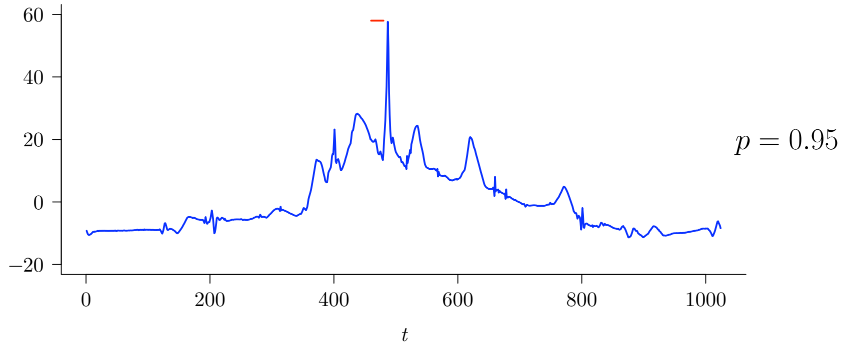
- model has three parameters σ_G^2, p and σ_ϵ^2, which need to be set
- let σ_R^2 and σ_W^2 be variances of RVs $R_{j,t}$ and $W_{j,t}$
- have relationships $\sigma_R^2 = (1-p)\sigma_G^2$ and $\sigma_W^2 = \sigma_R^2 + \sigma_\epsilon^2$
  - set $\hat{\sigma}_\epsilon^2 = \hat{\sigma}_{(\text{mad})}^2$ using $\mathbf{W}_1$
  - let $\hat{\sigma}_W^2$ be sample mean of all $W_{j,t}^2$
  - given p, let $\hat{\sigma}_G^2 = (\hat{\sigma}_W^2 - \hat{\sigma}_\epsilon^2)/(1-p)$
  - 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: I



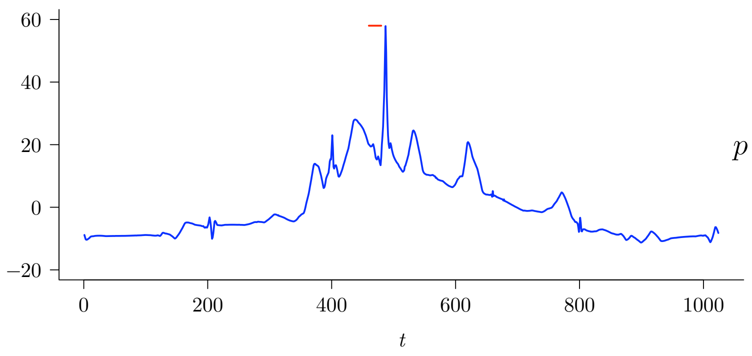
- shrinkage signal estimates of NMR spectrum based upon level $J_0 = 6$ partial LA(8) DWT and conditional mean with $p = 0.9$

Examples of Wavelet-Based Shrinkage: II



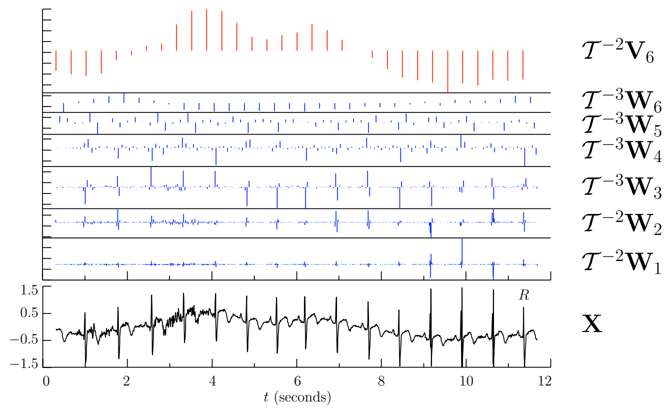
- same as before, but now with $p = 0.95$

Examples of Wavelet-Based Shrinkage: III



- 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 *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 multiscale graphical models to handle clustering and persistence
  - Cai and Silverman (2001) consider 'block' thesholding 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)

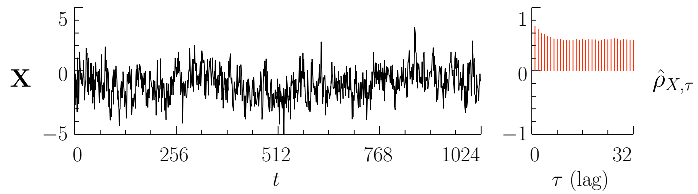
Comments on 'Next Generation' Denoising: III

- '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 *Statistica Sinica*, which has a special section entitled 'Multiscale Methods and Statistics: A Productive Marriage'

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 (will cover briefly)
  - test for trends (won't discuss – see Craigmile *et al.*, 2004, for details)

DWT of an FD Process: I

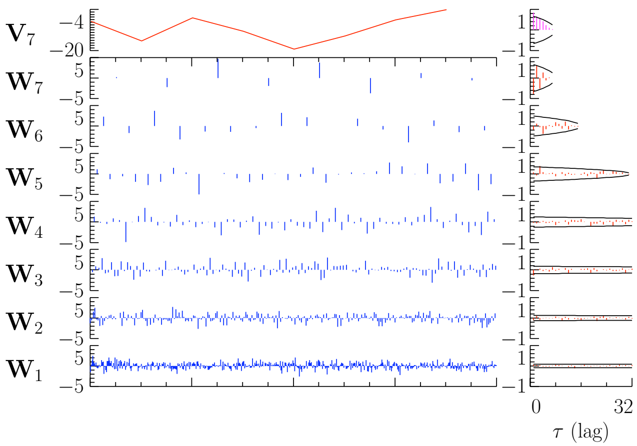


- realization of an FD(0.4) time series **X** along with its sample autocorrelation sequence (ACS): for $\tau \geq 0$,

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

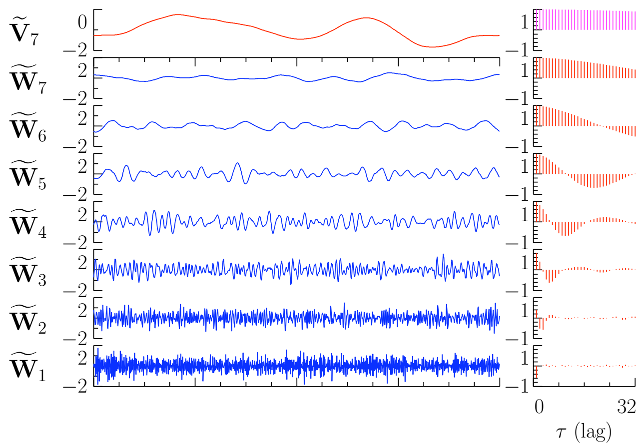
- note that ACS dies down slowly

DWT of an FD Process: II



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

MODWT of an FD Process



- 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 $\mathbf{X}$, ACSs for $\mathbf{W}_j$ consistent with white noise
- variance of RVs in $\mathbf{W}_j$ increases with j: for FD process,

$$\operatorname{var}\{W_{j,t}\} \approx c\tau_j^{2\delta} \equiv C_j,$$

  where c is a constant depending on δ but not j, and $\tau_j = 2^{j-1}$ is scale associated with $\mathbf{W}_j$
- for white noise ($\delta = 0$), $\operatorname{var}\{W_{j,t}\}$ is the same for all j
- dependence in $\mathbf{X}$ thus manifests itself in wavelet domain by different variances for wavelet coefficients at different scales

Correlations Within a Scale and Between Two Scales

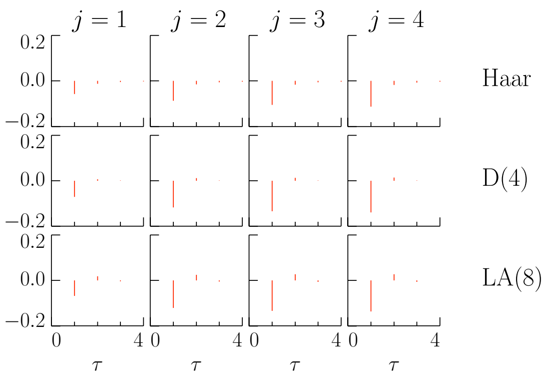
- let $\{s_{X,\tau}\}$ denote autocovariance sequence (ACVS) for $\{X_t\}$; i.e., $s_{X,\tau} = \operatorname{cov}\{X_t, X_{t+\tau}\}$
- let $\{h_{j,l}\}$ denote equivalent wavelet filter for jth level
- to quantify decorrelation, can write

$$\operatorname{cov}\{W_{j,t}, W_{j',t'}\} = \sum_{l=0}^{L_j-1} \sum_{l'=0}^{L_{j'}-1} h_{j,l} h_{j',l'} s_{X,2^j(t+1)-l-2^{j'}(t'+1)+l'},$$

  from which we can get ACVS (and hence within-scale correlations) for $\{W_{j,t}\}$:

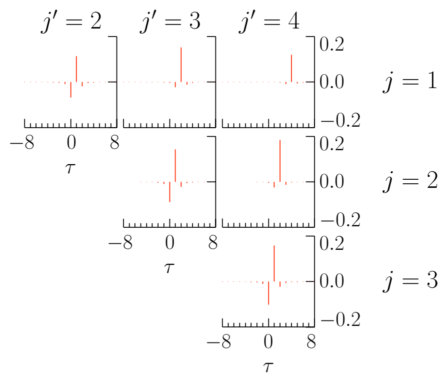
$$\operatorname{cov}\{W_{j,t}, W_{j,t+\tau}\} = \sum_{m=-(L_j-1)}^{L_j-1} s_{X,2^j\tau+m} \sum_{l=0}^{L_j-|m|-1} h_{j,l} h_{j,l+|m|}$$

Correlations Within a Scale



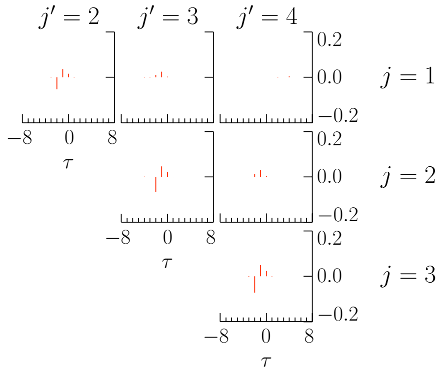
- correlations between $W_{j,t}$ and $W_{j,t+\tau}$ for an FD(0.4) process
- correlations within scale are slightly smaller for Haar
- maximum magnitude of correlation is less than 0.2

Correlations 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 $1 \le j < j' \le 4$

Correlations Between Two Scales: II



- 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 (δ and σ_ε^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 $\mathbf{X} \equiv [X_0, \ldots, X_{N-1}]^T$ with zero mean and with a multivariate Gaussian distribution
- with $N = 2^J$, recall that $\mathbf{X} = \mathcal{W}^T\mathbf{W}$, where

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_j \\ \vdots \\ \mathbf{W}_J \\ \mathbf{V}_J \end{bmatrix}$$

Basic DWT-Based Simulation Scheme

- assume $\mathbf{W}$ to contain N uncorrelated Gaussian (normal) random variables (RVs) with zero mean
- assume $\mathbf{W}_j$ to have variance $C_j = c\tau_j^{2\delta}$
- assume single RV in $\mathbf{V}_J$ to have variance C_{J+1} (see Percival and Walden, 2000, for details on how to set C_{J+1})
- approximate FD time series $\mathbf{X}$ via $\mathbf{Y} \equiv \mathcal{W}^T\Lambda^{1/2}\mathbf{Z}$, where
  - $\Lambda^{1/2}$ is $N \times N$ diagonal matrix with diagonal elements

$$\underbrace{C_1^{1/2}, \ldots, C_1^{1/2}}_{\frac{N}{2}\text{ of these}}, \underbrace{C_2^{1/2}, \ldots, C_2^{1/2}}_{\frac{N}{4}\text{ of these}}, \ldots, \underbrace{C_{J-1}^{1/2}, C_{J-1}^{1/2}}_{2\text{ of these}}, C_J^{1/2}, C_{J+1}^{1/2}$$

  - $\mathbf{Z}$ is vector of deviations drawn from a Gaussian distribution with zero mean and unit variance

Refinements to Basic Scheme: I

- covariance matrix for approximation $\mathbf{Y}$ does not correspond to that of a stationary process
- recall $\mathcal{W}$ treats $\mathbf{X}$ as if it were circular
- let $\mathcal{T}$ be $N \times N$ 'circular shift' matrix:

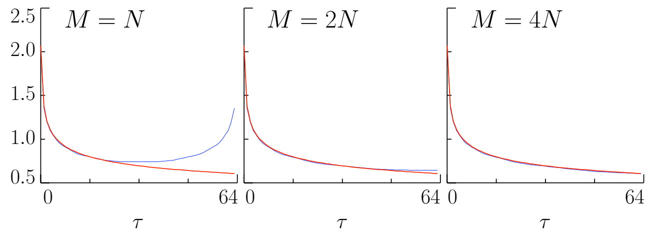
$$\mathcal{T}\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_0 \end{bmatrix}; \quad \mathcal{T}^2\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} Y_2 \\ Y_3 \\ Y_0 \\ Y_1 \end{bmatrix}; \quad \text{etc.}$$

- let κ be uniformly distributed over $0, \ldots, N-1$
- define $\widetilde{\mathbf{Y}} \equiv \mathcal{T}^{\kappa}\mathbf{Y}$
- $\widetilde{\mathbf{Y}}$ is stationary with ACVS given by, say, $s_{\widetilde{Y},\tau}$

Refinements to Basic Scheme: II

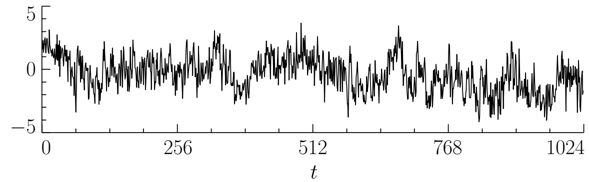
- Q: how well does $\{s_{\widetilde{Y},\tau}\}$ match $\{s_{X,\tau}\}$?
- due to circularity, find that $s_{\widetilde{Y},N-\tau} = s_{\widetilde{Y},\tau}$ for $\tau = 1, \ldots, N/2$
- implies $s_{\widetilde{Y},\tau}$ cannot approximate $s_{X,\tau}$ well for τ close to N
- can patch up by simulating $\widetilde{\mathbf{Y}}$ 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 $\{s_{\widetilde{Y},\tau}\}$ (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

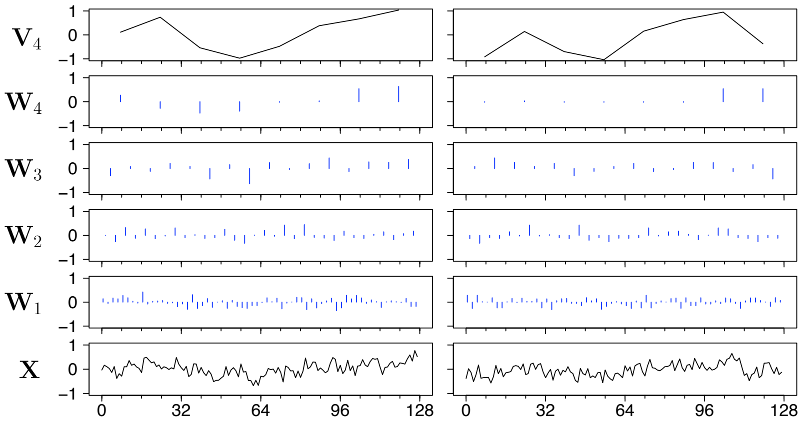
- for many (but not all!) time series, DWT acts as a decorrelating transform: to a good approximation, each $\mathbf{W}_j$ is a sample of a white noise process, and coefficients from different sub-vectors $\mathbf{W}_j$ and $\mathbf{W}_{j'}$ are also pairwise uncorrelated
- variance of coefficients in $\mathbf{W}_j$ depends on j
- scaling coefficients $\mathbf{V}_{J_0}$ 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 $\mathbf{X}$ of length $N = 2^J$, compute level J_0 DWT (the choice $J_0 = J - 3$ yields 8 coefficients in $\mathbf{W}_{J_0}$ and $\mathbf{V}_{J_0}$)
2. randomly sample with replacement from $\mathbf{W}_j$ to create bootstrapped vector $\mathbf{W}_j^{(b)}$, $j = 1, \ldots, J_0$
3. create $\mathbf{V}_{J_0}^{(b)}$ using 1st-order autoregressive parametric bootstrap
4. apply $\mathcal{W}^T$ to $\mathbf{W}_1^{(b)}, \ldots, \mathbf{W}_{J_0}^{(b)}$ and $\mathbf{V}_{J_0}^{(b)}$ to obtain bootstrapped time series $\mathbf{X}^{(b)}$ and then form $\hat{\rho}_{X,1}^{(b)}$

- repeat above many times to build up sample distribution of bootstrapped autocorrelations

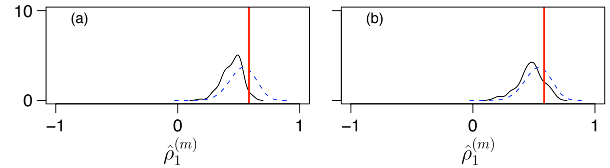
Illustration of Wavelet-Domain Bootstrapping



- Haar DWT of FD(0.45) series $\mathbf{X}$ (left-hand column) and wavelet-domain bootstrap thereof (right-hand)

Wavelet-Domain Bootstrapping of FD Series

- approximations to true PDF using (a) Haar & (b) LA(8) wavelets



vertical line indicates $\hat{\rho}_{X,1}$

- using 50 FD time series and the Haar DWT yields:

  average of 50 sample means $\doteq 0.35$ (truth $\doteq 0.53$)
  average of 50 sample SDs $\doteq 0.096$ (truth $\doteq 0.107$)

- using 50 FD time series and the LA(8) DWT yields:

  average of 50 sample means $\doteq 0.43$ (truth $\doteq 0.53$)
  average of 50 sample SDs $\doteq 0.098$ (truth $\doteq 0.107$)

MLEs of FD Parameters: I

- FD process depends on 2 parameters, namely, δ and σ_ε^2
- given $\mathbf{X} = [X_0, X_1, \ldots, X_{N-1}]^T$ with $N = 2^J$, suppose we want to estimate δ and σ_ε^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:

$$L(\delta, \sigma_\varepsilon^2 \mid \mathbf{X}) \equiv \frac{1}{(2\pi)^{N/2}|\Sigma_\mathbf{X}|^{1/2}} e^{-\mathbf{X}^T \Sigma_\mathbf{X}^{-1} \mathbf{X}/2}$$

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 δ and σ_ε^2 maximize $L(\delta, \sigma_\varepsilon^2 \mid \mathbf{X})$ or, equivalently, mininize

$$-2\log\left(L(\delta, \sigma_\varepsilon^2 \mid \mathbf{X})\right) = N \log(2\pi) + \log\left(|\Sigma_\mathbf{X}|\right) + \mathbf{X}^T \Sigma_\mathbf{X}^{-1} \mathbf{X}$$
- 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 $\mathbf{Y} = \mathcal{W}^T \Lambda^{1/2} \mathbf{Z}$, 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

    $$\mathcal{W}^T \Lambda^{1/2} I_N (\mathcal{W}^T \Lambda^{1/2})^T = \mathcal{W}^T \Lambda^{1/2} \Lambda^{1/2} \mathcal{W} = \mathcal{W}^T \Lambda \mathcal{W} \equiv \widetilde{\Sigma}_\mathbf{X},$$

    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:

$$-2\log\left(L(\delta, \sigma_\varepsilon^2 \mid \mathbf{X})\right) \approx N \log(2\pi) + \log\left(|\widetilde{\Sigma}_\mathbf{X}|\right) + \mathbf{X}^T \widetilde{\Sigma}_\mathbf{X}^{-1} \mathbf{X}$$

MLEs of FD Parameters: IV

- Q: so how does this help us?
  - easy to invert $\widetilde{\Sigma}_\mathbf{X}$:

    $$\widetilde{\Sigma}_\mathbf{X}^{-1} = \left(\mathcal{W}^T \Lambda \mathcal{W}\right)^{-1} = (\mathcal{W})^{-1} \Lambda^{-1} \left(\mathcal{W}^T\right)^{-1} = \mathcal{W}^T \Lambda^{-1} \mathcal{W},$$

    where Λ^{-1} is another diagonal matrix, leading to

    $$\mathbf{X}^T \widetilde{\Sigma}_\mathbf{X}^{-1} \mathbf{X} = \mathbf{X}^T \mathcal{W}^T \Lambda^{-1} \mathcal{W} \mathbf{X} = \mathbf{W}^T \Lambda^{-1} \mathbf{W}$$
  - easy to compute the determinant of $\widetilde{\Sigma}_\mathbf{X}$:

    $$|\widetilde{\Sigma}_\mathbf{X}| = |\mathcal{W}^T \Lambda \mathcal{W}| = |\Lambda \mathcal{W} \mathcal{W}^T| = |\Lambda I_N| = |\Lambda|,$$

    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 δ:

$$C'_j(\delta) \equiv \int_{1/2^{j+1}}^{1/2^j} \frac{2^{j+1}}{[4\sin^2(\pi f)]^\delta}\,df \approx \int_{1/2^{j+1}}^{1/2^j} \frac{2^{j+1}}{[2\pi f]^{2\delta}}\,df$$

$$C'_{J+1}(\delta) \equiv \frac{N\Gamma(1-2\delta)}{\Gamma^2(1-\delta)} - \sum_{j=1}^{J} \frac{N}{2^j} C'_j(\delta)$$

$$\sigma_\varepsilon^2(\delta) \equiv \frac{1}{N}\left(\frac{V_{J,0}^2}{C'_{J+1}(\delta)} + \sum_{j=1}^{J} \frac{1}{C'_j(\delta)} \sum_{t=0}^{\frac{N}{2^j}-1} W_{j,t}^2\right)$$

MLEs of FD Parameters: VI

- wavelet-based approximate MLE $\tilde{\delta}$ for δ is the value that minimizes the following function of δ:

$$\tilde{l}(\delta \mid \mathbf{X}) \equiv N\,\log(\sigma_\varepsilon^2(\delta)) + \log(C'_{J+1}(\delta)) + \sum_{j=1}^{J} \frac{N}{2^j}\log(C'_j(\delta))$$

- once $\tilde{\delta}$ has been determined, MLE for σ_ε^2 is given by $\sigma_\varepsilon^2(\tilde{\delta})$
- computer experiments indicate scheme works quite well

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 τ and for $\beta = 2\delta - 1$, have
    $$\log(\nu_X^2(\tau_j)) \approx \zeta + \beta\log(\tau_j)$$
  - suggests determining δ by regressing $\log(\hat{\nu}_X^2(\tau_j))$ on $\log(\tau_j)$ over range of τ_j
  - weighted LSE takes into account fact that variance of $\log(\hat{\nu}_X^2(\tau_j))$ depends upon scale τ_j (increases as τ_j increases)

Homogeneity of Variance: I

- because DWT decorrelates FD and related processes, nonboundary coefficients in $\mathbf{W}_j$ should resemble white noise; i.e.,
  $$\operatorname{cov}\{W_{j,t}, W_{j,t'}\} \approx 0$$
  when $t \neq t'$, and $\operatorname{var}\{W_{j,t}\}$ should not depend upon t
- can test for homogeneity of variance in $\mathbf{X}$ using $\mathbf{W}_j$ over a range of levels j
- suppose $U_0, \ldots, U_{N-1}$ are independent normal RVs with $E\{U_t\} = 0$ and $\operatorname{var}\{U_t\} = \sigma_t^2$
- want to test null hypothesis $H_0 : \sigma_0^2 = \sigma_1^2 = \cdots = \sigma_{N-1}^2$
- can test H_0 versus a variety of alternatives, e.g.,
  $$H_1 : \sigma_0^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_{N-1}^2$$
  using normalized cumulative sum of squares

Homogeneity of Variance: II

- to define test statistic D, start with

$$\mathcal{P}_k \equiv \frac{\sum_{j=0}^{k} U_j^2}{\sum_{j=0}^{N-1} U_j^2}, \quad k = 0, \ldots, N-2$$

and then compute $D \equiv \max(D^+, D^-)$, where

$$D^+ \equiv \max_{0 \le k \le N-2} \left(\frac{k+1}{N-1} - \mathcal{P}_k \right) \;\; \& \;\; D^- \equiv \max_{0 \le k \le N-2} \left(\mathcal{P}_k - \frac{k}{N-1} \right)$$

- 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_α such that $\mathbf{P}[D \ge x_\alpha] = \alpha$ for, e.g., $\alpha = 0.01$, 0.05 or 0.1

Homogeneity of Variance: III

- once determined, can perform α level test of H_0:
  - compute D statistic from data $U_0, \ldots, U_{N-1}$
  - reject H_0 at level α if $D \ge x_\alpha$
  - fail to reject H_0 at level α if $D < x_\alpha$
- can determine critical values x_α in two ways
  - Monte Carlo simulations
  - large sample approximation to distribution of D:

$$\mathbf{P}[(N/2)^{1/2} D \ge x] \approx 1 + 2 \sum_{l=1}^{\infty} (-1)^l e^{-2l^2 x^2}$$

(reasonable approximation for $N \ge 128$)

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):

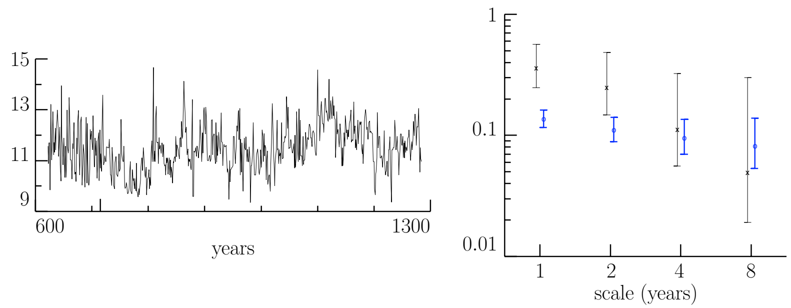
$$\mathcal{P}_k \equiv \frac{\sum_{t=L_j'}^{k} W_{j,t}^2}{\sum_{t=L_j'}^{N_j-1} W_{j,t}^2}, \quad k = L_j', \ldots, N_j - 2$$

- if null hypothesis rejected at level j, can use nonboundary MODWT coefficients to locate change point based on

$$\widetilde{\mathcal{P}}_k \equiv \frac{\sum_{t=L_j-1}^{k} \widetilde{W}_{j,t}^2}{\sum_{t=L_j-1}^{N-1} \widetilde{W}_{j,t}^2}, \quad k = L_j - 1, \ldots, N-2$$

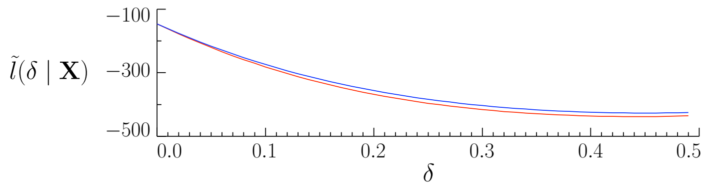
along with analogs $\widetilde{D}_k^+$ and $\widetilde{D}_k^-$ of D_k^+ and D_k^-

Example – Annual Minima of Nile River: I



- left-hand plot: annual minima of Nile River
- new measuring device introduced around year 715
- right: Haar $\hat{\nu}_X^2(\tau_j)$ before (x's) and after (o's) year 715.5, with 95% confidence intervals based upon $\chi^2_{\eta_3}$ approximation

Example – Annual Minima of Nile River: II



- based upon last 512 values (years 773 to 1284), plot shows $\tilde{l}(\delta \mid \mathbf{X})$ versus δ 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: $\tilde{\delta} \doteq 0.4532$
  - exact MLE is value minimizing lower curve: $\hat{\delta} \doteq 0.4452$

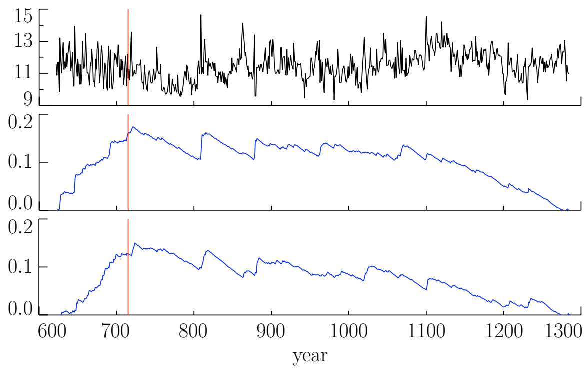
Example – Annual Minima of Nile River: III

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

			critical levels		
τ_j	M'_j	D	10%	5%	1%
1 year	331	0.1559	0.0945	0.1051	0.1262
2 years	165	0.1754	0.1320	0.1469	0.1765
4 years	82	0.1000	0.1855	0.2068	0.2474
8 years	41	0.2313	0.2572	0.2864	0.3436

- can reject null hypothesis of homogeneity of variance at level of significance 0.05 for scales τ_1 & τ_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 τ_1 & τ_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
  - C. W. J. Granger and R. Joyeux (1980), 'An Introduction to Long-Memory Time Series Models and Fractional Differencing,' *Journal of Time Series Analysis*, **1**, pp. 15–29
  - J. R. M. Hosking (1981), 'Fractional Differencing,' *Biometrika*, **68**, pp. 165–76
- wavelet cross-covariance and cross-correlation
  - B. J Whitcher, P. Guttorp and D. B. Percival (2000), 'Wavelet Analysis of Covariance with Application to Atmospheric Time Series,' *Journal of Geophysical Research*, **105**, D11, pp. 14,941–62
  - A. Serroukh and A. T. Walden (2000a), 'Wavelet Scale Analysis of Bivariate Time Series I: Motivation and Estimation,' *Journal of Nonparametric Statistics*, **13**, pp. 1–36
  - A. Serroukh and A. T. Walden (2000b), 'Wavelet Scale Analysis of Bivariate Time Series II: Statistical Properties for Linear Processes,' *Journal of Nonparametric Statistics*, **13**, pp. 37–56
- asymptotic theory for non-Gaussian processes
  - A. Serroukh, A. T. Walden and D. B. Percival (2000), 'Statistical Properties and Uses of the Wavelet Variance Estimator for the Scale Analysis of Time Series,' *Journal of the American Statistical Association*, **95**, pp. 184–96

References: II

- biased estimators of wavelet variance
  - E. M. Aldrich (2005), 'Alternative Estimators of Wavelet Variance,' Masters Thesis, Department of Statistics, University of Washington
- unbiased estimator of wavelet variance for 'gappy' time series
  - D. Mondal and D. B. Percival (2010a), 'Wavelet Variance Analysis for Gappy Time Series,' to appear in *Annals of the Institute of Statistical Mathematics*
- robust estimation
  - D. Mondal and D. B. Percival (2010b), 'M-Estimation of Wavelet Variance,' to appear in *Annals of the Institute of Statistical Mathematics*
- wavelet variance for random fields
  - D. Mondal and D. B. Percival (2010c), 'Wavelet Variance Analysis for Random Fields,' under review
- wavelet-based characteristic scales
  - M. J. Keim and D. B. Percival (2010), 'Assessing Characteristic Scales Using Wavelets,' under preparation

References: III

- wavelet-based denoising
  - A. Antoniadis (2007), 'Wavelet Methods in Statistics: Some Recent Developments and Their Applications,' *Statistical Surveys*, **1**, pp. 16–55
  - Y. Benjamini and Y. Hochberg (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,' *Journal of the Royal Statistical Society, Series B*, **57**, pp. 289–300
  - T. Cai and B. W. Silverman (2001), 'Incorporating Information on Neighboring Coefficients into Wavelet Estimation,' *Sankhya Series B*, **63**, pp. 127–48
  - M. S. Crouse, R. D. Nowak and R. G. Baraniuk (1998), 'Wavelet-Based Statistical Signal Processing Using Hidden Markov Models,' *IEEE Transactions on Signal Processing*, **46**, pp. 886–902
  - P. L. Dragotti and M. Vetterli (2003), 'Wavelet Footprints: Theory, Algorithms, and Applications,' *IEEE Transactions on Signal Processing*, **51**, pp. 1306–23
  - H.–C. Huang and N. Cressie (2000), 'Deterministic/Stochastic Wavelet Decomposition for Recovery of Signal from Noisy Data,' *Technometrics*, **42**, pp. 262–76
  - G. P. Nason (2008), *Wavelet Methods in Statistics with R*, Springer, Berlin

References: IV

  - A. M. Wink and J. B. T. M. Roerdink (2004), 'Denoising Functional MR Images: A Comparison of Wavelet Denoising and Gaussian Smoothing,' *IEEE Transactions on Medical Imaging*, **23**(3), pp. 374–87
- bootstrapping
  - A. C. Davison and D. V. Hinkley (1997), *Bootstrap Methods and their Applications*, Cambridge, England: Cambridge University Press
  - D. B. Percival, S. Sardy and A. C. Davison (2001), 'Wavestrapping Time Series: Adaptive Wavelet-Based Bootstrapping,' in *Nonlinear and Nonstationary Signal Processing*, edited by W. J. Fitzgerald, R. L. Smith, A. T. Walden and P. C. Young. Cambridge, England: Cambridge University Press, pp. 442–70
  - A. M. Sabatini (1999), 'Wavelet-Based Estimation of $1/f$-Type Signal Parameters: Confidence Intervals Using the Bootstrap,' *IEEE Transactions on Signal Processing*, **47**(12), pp. 3406–9
  - A. M. Sabatini (2006), 'A Wavelet-Based Bootstrap Method Applied to Inertial Sensor Stochastic Error Modelling Using the Allan Variance,' *Measurement Science and Technology*, **17**, pp. 2980–2988
  - B. J Whitcher (2006), 'Wavelet-Based Bootstrapping of Spatial Patterns on a Finite Lattice,' *Computational Statistics & Data Analysis*, **50**(9), pp. 2399–421

References: V

- decorrelation property of DWTs
  - P. F. Craigmile and D. B. Percival (2005), 'Asymptotic Decorrelation of Between-Scale Wavelet Coefficients', *IEEE Transactions on Information Theory*, **51**, pp. 1039–48
- parameter estimation for FD processes
  - P. F. Craigmile, P. Guttorp and D. B. Percival (2005), 'Wavelet-Based Parameter Estimation for Polynomial Contaminated Fractionally Differenced Processes,' *IEEE Transactions on Signal Processing*, **53**, pp. 3151–61
- testing for homogeneity of variance
  - B. J Whitcher, S. D. Byers, P. Guttorp and D. B. Percival (2002), 'Testing for Homogeneity of Variance in Time Series: Long Memory, Wavelets and the Nile River,' *Water Resources Research*, **38**, 10.1029/2001WR000509.
- wavelet-based trend assessment
  - P. F. Craigmile, P. Guttorp and D. B. Percival (2004), 'Trend Assessment in a Long Memory Dependence Model using the Discrete Wavelet Transform,' *Environmetrics*, **15**, pp. 313–35