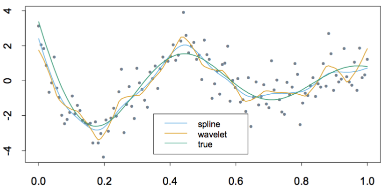
Wavelet Theory for Statistical Signal Estimation

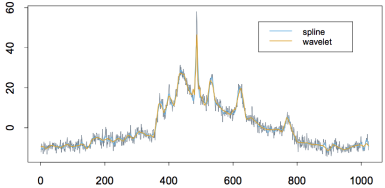
Adam Quinn Jaffe

March 19, 2020

Motivation 1

- For some regression problems, we want to use a basis expansion whch is not smooth like piece-wise polynomials, splines, etc. This is particularly important in signal processing and image processing.

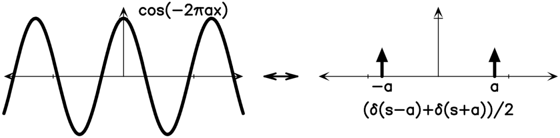




- Important distinction when making modeling decisions

Motivation 2

- The standard tools for signal processing (and image processing, to a degree) are Fourier methods which use the complex exponentials as a basis. Why not just use these?
- Complex exponentials are "localized in frequency but not localized in time".

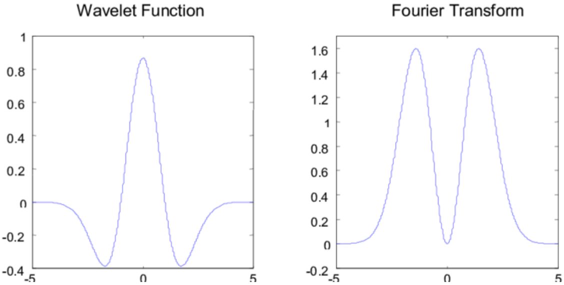


time domain frequency domain

- Can we localize in both frequency and time? By the uncertainty principle, there are fundamental limits to how tightly you can do these simultaneously.

Motivation 2 (cont'd)

However, some clever constructions can balance this trade-off:



Introduction

- A wavelet is a wave-like (smooth) function with small (compact) support. Usually we consider a large (infinite) system of wavelets which form a basis for a suitable function space.
- The localization in both time and frequency leads to the concept of *multi-resolution analysis*
- The process of *wavelet smoothing* is just a least-squares projection onto a function subspace spanned by a small number of the possible wavelets (cf. low-pass filtering in signal processing)
- There are important differences between continuous-time and discrete-time wavelet theory
- Lots of different choices of wavelet systems, each with their own advantages and disadvantages.

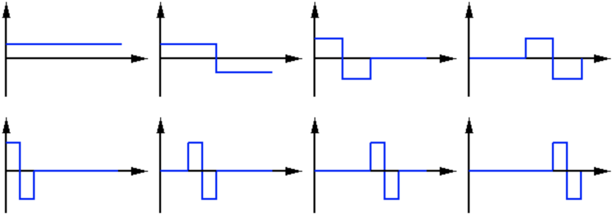
Mathematical Background

Some Functional Analysis

- Let $L^2(S)$ denote the hilbert space of real-valued, square-integrable functions from a set S to $\mathbb{R}$ (or $\mathbb{C}$)
- An *orthonormal basis* for $L^2(S)$ is a collection $V = \{v_i\}_{i=1}^{\infty} \subseteq L^2(S)$ such that $\operatorname{span}(V)$ is dense in $L^2(S)$, and with $\langle v_i, v_j \rangle = 0$ when $i \neq j$ and 1 when $i = j$. In particular, any vector $x \in L^2(S)$ can be written uniquely as $f = \sum_{i=1}^{\infty} f_i v_i$ for some constants $\{f_i\}_{i=1}^{\infty} \subseteq \mathbb{R}$ called the *coordinates* of f with respect to V.
- One classical orthonormal basis for $L^2([0,1])$ is the set of complex exponentials $\mathcal{E} = \{\exp(2\pi inx)\}_{n=0}^{\infty}$; the coordinates of a function $f \in L^2([0,1])$ in the basis of $\mathcal{E}$ is called the *Fourier series* of f.

Some Functional Analysis

- Another classical orthonormal basis for $L^2([0,1])$ is $\mathcal{H}$, the triangular array of *Haar functions*, along with the constant function:



…

Multi-Resolution Analysis

Definition

A *multi-resolution analysis (MRA) of* $L^2(\mathbb{R})$ is a collection of closed subspaces $\{V_j\}_{j=-\infty}^{\infty}$ or $L^2(\mathbb{R})$ satisfying the following properties:

(1) $\{0\} \subseteq \cdots \subseteq V_2 \subseteq V_1 \subseteq V_0 \subseteq V_{-1} \subseteq V_{-2} \subseteq \cdots \subseteq L^2(\mathbb{R})$

(2) V_0 is the closed linear span of a compactly-supported function $\phi \in L^2(\mathbb{R})$ and its integer-translates $\phi_m(x) = \phi(x-m)$

(3) If $f \in V_k$, then $g(x) = f(x - m2^k)$ has $g \in V_k$ for all $m \in \mathbb{Z}$.

(4) Writing $g(x) = f(2x)$, we have $f \in V_k$ if and only if $g \in V_{k+1}$

- There is some disagreement on the indexing convention for $\{V_j\}$
- There is a very simple MRA related to the Haar system $\mathcal{H}$, but its generators have some undesirable properties
- Can we find an MRA whose generators are "nice"?

Multi-Resolution Analysis (cont'd)

- If $\{V_j\}$ is any MRA, we have

$$L^2(\mathbb{R}) = \overline{\bigcup_{j=\infty}^{-\infty} V_j} \tag{1}$$

- For each j, we have $V_j \subseteq V_{j-1}$, so there exists a closed subspace $W_j \subseteq V_{j-1}$ such that $V_j \oplus W_j = V_{j-1}$.
- Note that $\{W_j\}_j$ are orthogonal subspaces, since for any $k \geq j+1$ we have

$$W_j \perp V_j \supseteq V_{k-1} \supseteq W_k \tag{2}$$

- Now by induction we have the following for any J:

$$L^2(\mathbb{R}) = V_J \oplus \bigoplus_{j=J}^{-\infty} W_j = \bigoplus_{j=\infty}^{-\infty} W_j \tag{3}$$

Multi-Resolution Analysis (cont'd)

- Suppose that V_0 is generated by ϕ and its translates. Then $\phi(x/2) \in V_1 \subseteq V_0$, so we can find coefficients $\{a_j\}$ with

$$\phi\left(\frac{x}{2}\right) = \sum_{j=\infty}^{-\infty} a_j \phi(x-j) \quad \Rightarrow \quad \phi(x) = \sum_{j=\infty}^{-\infty} a_j \phi(2x-j) \tag{4}$$

  The second part is called the *refinability equation*.

- Using these coefficients, define

$$\psi(x) = \sum_{j=\infty}^{-\infty} (-1)^j a_{1-j} \phi(2x-j) \tag{5}$$

- We call ϕ and ψ the *father wavelet* and *mother wavelet* respectively. It turns out that ψ and its translates form an orthonormal basis for W_0, even though ϕ and its translates need not be an orthonormal basis for V_0.

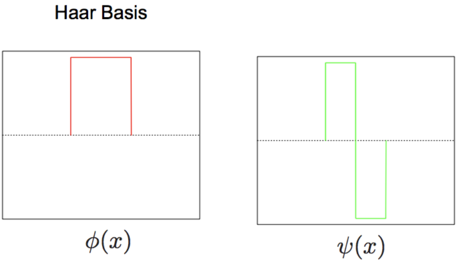
Multi-Resolution Analysis (cont'd)

- Let $\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k)$ denote the father wavelet with altered scaling and centering, called a *son*. Similarly let $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$ be a *daughter*.
- Now for any $f \in L^2(\mathbb{R})$ and any J, we can write

$$f(x) = \sum_{k=-\infty}^{\infty} a_{J,k}\phi_{j,k}(x) + \sum_{j=J}^{-\infty}\sum_{k=-\infty}^{\infty} b_{j,k}\psi_{j,k}(x) \qquad (6)$$

  The coefficients $\{b_{j,k}\}$ of the daughters are called the *detail* and they are uniquely determined by f. (They can be found via orthogonal projection of f onto the closure of $\bigoplus_{j=J}^{-\infty} W_j$.) However, the remaining coefficients $\{a_{J,k}\}$, called the *gross*, are not uniquely determined in general.
- There is a degree in freedom in choosing J, and this will be useful for applying these ideas to discrete-time signals later.

Example: Haar MRA

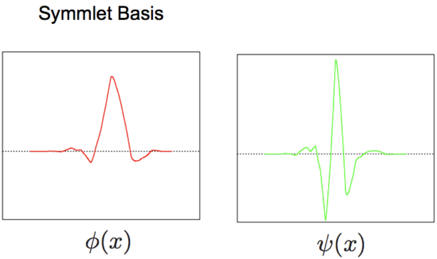


$$\int_{-\infty}^{\infty} \psi(x) dx = 0$$

Multi-Resolution Analysis (cont'd)

- Often we are interested in systems of wavelets whose father and mother are continuous and compactly supported. Does such a system exist? The answer is yes but it was highly non-trivial to confirm.
- Suppose that the support of ψ is fixed as a compact set $K \subseteq \mathbb{R}$. There are two reasonable questions to ask:
  - Which mother ψ with $\text{supp}(\psi) = K$ is the smoothest?
  - Which mother ψ with $\text{supp}(\psi) = K$ has the most vanishing moments?
- If ψ has p vanishing moments, then $\langle f, \psi \rangle = 0$ for any degree-$(p-1)$ polyomial f. In particular, V_0 must then contain all degree-$(p-1)$ polyomials.

Example: Debauchies' Symmlet MRA



$$\int_{-\infty}^{\infty} x^j \psi(x) dx = 0, \qquad \text{for all } j = 0, 1, \ldots 7$$

Wavelet Transforms

Continuous-Time Wavelet Transform

- Let ϕ and ψ be compactly-supported father and mother wavelets, and let Ψ_J be the basis of $\bigoplus_{j=J}^{-\infty} W_j$ consisting of the daughters of ψ whose resolutions are at least as low as 2^J. Let $\Psi = \bigcup_{J=-\infty}^{\infty} \Psi_J$ be the basis of $\bigoplus_{j=\infty}^{-\infty} W_j$ consisting of all daughters of ψ.
- Let $\mathbf{W} : L^2([0,1]) \to \ell_2(\mathbb{R})$ denote the linear operator which expresses f in the coordinates of Ψ, called the *complete wavelet transform*
- For any J, let $\mathbf{W}_J : L^2([0,1]) \to \ell_2(\mathbb{R})$ denote the linear operator which first projects orthogonally onto $\bigoplus_{j=J}^{-\infty} W_j$, then expresses the resultant vector in the coordinates of Ψ. Call this the *partial wavelet transform for level J*.

Discrete-Time Wavelet Transform

- In discrete-time, father and mother wavelets ϕ and ψ are any functions $[N] \to \mathbb{R}$ which satisfy the refinement equations. Assume $N = 2^n$ for simplicity.
- The lowest possible resolution for such functions is 1 and the highest possible resolution is $N = 2^n$. Hence, the spanning criterion reduces to

$$L^2([N]) = V_J \oplus \bigoplus_{j=J}^{0} W_j = \bigoplus_{j=n}^{0} W_j \tag{7}$$

  for any $0 \leq J \leq n$.
- Again, let Ψ_J be the basis of $\bigoplus_{j=J}^{0} W_j$ along with $\Psi = \bigcup_{J=-\infty}^{0} \Psi_J$. Let $\mathbf{W} : L^2([N]) \to L^2([N])$ denote the linear operator which expresses f in the coordinates of Ψ; in discrete-time this is just a matrix. For any J, let $\mathbf{W}_J$ be defined analogously.

Fast Wavelet Transform (FWT)

- The transformation $\mathbf{W}$ is an $N \times N$ matrix, but $\mathbf{W}f$ can be computed from f in $O(N)$ time (cf. $\Omega(N^2)$ for general matrix-vector multiplication, or $O(N \log N)$ for FFT).
- For the Haar basis, this is easy to see by example; for general MRA, this is possible because of the refinement equations
- The strategy is a class of algorithms called *pyramid algorithms* which are common in image processing.

Isometric Propteries

- In either continuous-time or discrete, the wavelet transform enjoys many nice properties:
- (Orthogonality) For any $\theta_1, \theta_2 \in L^2([N])$, we have $\langle \mathbf{W}\theta_1, \mathbf{W}\theta_2 \rangle = \langle \theta_1, \theta_2 \rangle$
- (Isometry) For any $\theta \in L^2([N])$, we have $||\mathbf{W}\theta|| = ||\theta||$
- (Isomorphism) The map **W** is a bijection

Signal Estimation

Problem Statement

- Suppose that $f : [0,1] \to \mathbb{R}$ is an unknown function and that the vector $y_i \in \mathbb{R}^N$ is observed, where

$$y_i = f(t_i) + z_i, \qquad t_i = (i-1)/N, \qquad z_i \sim N(0, \sigma^2). \tag{8}$$

  How should one estimate the vector $(f(t_i))_i \in \mathbb{R}^N$ from y?

- Generally, we assume that f comes from some class of functions $\mathcal{F}$ which imposes smoothness conditions, and then we study the minimax risk over the whole class.
- Often we assume that $\mathcal{F}$ is a *Sobolev ball*

$$W_2^m(C) = \left\{ f : [0,1] \to \mathbb{R}; \sum_{j=0}^{m} \left\| \frac{d^k f}{dt^k} \right\|_{L^2([0,1])}^2 \leq C \right\} \tag{9}$$

  for some $m \in \{0, 1, 2, \dots\}$ and some $C > 0$.

Simple Approaches

- One way to phrase this problem is via an optimization in the wavelet domain. Importantly, this is only possible because **W** being orthogonal implies that the transformed noise is also independent gaussian!
- However, we have $\text{Im}(\mathbf{W}) = L^2([N])$, so the problem

$$\begin{aligned} \text{minimize } ||y - \mathbf{W}\theta||_2^2 \\ \text{over } \theta \in L^2([N]) \end{aligned}$$

  experiences too much over-fitting. The solution is always $\theta = \mathbf{W}^\mathsf{T} y$.
- So, add some regulalization! An ℓ_2 penalty will just shrink the coefficients, but this is usually not a sufficient reduction. An ℓ_1-penalty will force some coefficients to go to 0, so the effect is similar to low-pass filtering in order to remove noise.

Regularization

- If we set-up the optimization problem,

$$\begin{aligned} \text{minimize } & ||y - \mathbf{W}\theta||_2^2 + 2\lambda||\theta||_1 \\ \text{over } & \theta \in L^2([N]) \end{aligned}$$

  then the explicit solution is given by the soft-threshholding operator:

$$\hat{\theta}_j = \eta_\lambda(y_j^*) = \mathrm{sign}(y_j^*)(|y_j^*| - \lambda)_+ \tag{10}$$

  where $y^* = \mathbf{W}^\mathsf{T} y$ is the inverse FWT of the observed signal y.
- But how should we choose λ? One idea is to use the expectation of the maxima of the independent Gaussian noise terms, $\lambda = \sigma\sqrt{2\log N}$.
- Another idea is to choose λ adaptively, using Stein's unbiased risk estimator (SURE).

Stein's Unbiased Risk Estimator

Lemma (Stein, 1981)

If $X \sim N(\mu, \sigma^2 I_N)$ is a multivariate gaussian with $\mu \in \mathbb{R}^N$, and $\delta : \mathbb{R}^N \to \mathbb{R}^N$ is a C^2 estimator of μ, then

$$\mathbb{E}\left[||\delta(X) - \mu||^2\right] = N + \mathbb{E}\left[||\delta(X) - X||^2 + 2\nabla \cdot (\delta(X) - X)\right]. \tag{11}$$

In other words,

$$R(X) = N + ||\delta(X) - X||^2 + 2\nabla \cdot (\delta(X) - X) \tag{12}$$

is an unbiased estimator of the risk of δ, called Stein's unbiased risk estimator (SURE).

- If δ_λ is a parameterized family of estimators, then Stein's lemma suggests that a method for choosing λ is to minimize R_λ with respect to the observed data.

Adaptive Threshholding

- In the setting above, we have a family of soft-threshholding operators, so we can (heuristically) choose λ by minimizing SURE.
- SURE amounts to

$$R_\lambda(X) = N - 2\#\{1 \leq i \leq d : |X_i| \leq \lambda\} + \sum_{i=1}^{N} (|X_i| \wedge \lambda)^2. \tag{13}$$

- Computing the minimizer λ^* is easy if the X_i are sorted, so it can be found in $O(N \log N)$ time. The process of soft threshholding already takes $O(N)$ time, so we have only increased the running time by a factor of $\log N$.
- In addition to choosing λ adaptively, we can choose a different λ_j for each level of detail j. That is, we should choose λ_j^* optimally according to SURE, but only with respect to the wavelet coefficients that correspond to details at level j.

SureShrink

Let $X \in \mathbb{R}^N$ denote the observed signal, and let $w = \mathbf{W}^\mathsf{T} X$ denote its wavelet transform. For each j, write $\mathbf{w}_j$ for the vector $(w_{k+2^j})_{k=1}^{2^j}$.

Definition

For each level j, let $\mu_j : \mathbb{R}^{2^j} \to \mathbb{R}^{2^j}$ denote the following: Partition the indices $\{1, 2, \ldots 2^j\}$ into two sets I, I' uniformly at random, then let $\lambda_j^*(\mathbf{w}_I)$ and $\lambda_j^*(\mathbf{w}_{I'})$ denote the SURE-optimal threshholds with respect to the given subset of the data. Then set:

$$(\mu_j(\mathbf{w}_j))_k = \begin{cases} \eta_{\sqrt{2\sigma \log N}}(w_k) & \text{if } \frac{\sigma^2}{\sqrt{N}} \sum_{i=1}^N (w_i^2 - 1) \leq (\log N)^{\frac{3}{2}} \\ \eta_{\lambda_j^*(\mathbf{w}_I)}(w_k) & \text{else, if } k \in I' \\ \eta_{\lambda_j^*(\mathbf{w}_{I'})}(w_k) & \text{else, if } k \in I \end{cases} \tag{14}$$

Finally, let $\hat{\mathbf{f}}$ denote the inverse wavelet transform of the vector $(\mu_0(\mathbf{w}_0), \mu_1(\mathbf{w}_1), \ldots \mu_{\log N}(\mathbf{w}_{\log N}))$.

SureShrink

- The estimator $\hat{\mathbf{f}}$ defined on the previous slide is called *SureShrink*. Once a certain wavelet transform / inverse transform has been specified, the estimator only depends on the observed signal.
- Intuitively, the idea is as follows: If there is not overwhelming evidence that the signal is non-trivial, then we set the threshhold to be the expected maximum value of the noise. Else, we adaptively choose a soft threshhold for each level, but we learn this parameter from half of the data and apply it to the other half.
- In practice, the data-splitting scheme does change the estimator too much. But this scheme is essential for the analysis of the risk of the estimator.
- This procedure takes $O(N \log N)$ time to compute.

Optimality

Theorem (Donoho and Johnstone, 1994)

Suppose that the wavelet transform $\mathbf{W}$ *corresponds to a multi-resolution analysis whose mother wavelet has* r *vanishing moments and* r *continuous derivatives. If* $r > \sigma$*, then SureShrink is asymptotically minimax in the following sense: For any* $m \in \mathbb{Z}^+$, $C > 0$*, we have*

$$\inf_{\delta} \sup_{f \in W_2^m(C)} R(\delta, f) \asymp \sup_{f \in W_2^m(C)} R(\hat{\mathbf{f}}, f) \tag{15}$$

as $N \to \infty$*, where the infimum is taken over all estimators* δ *of* f*.*

The most surprising thing above this result is that *SureShrink* is minimax over a huge smoothness class for the possible underlying signal. That is, it is able to estimate the smoothness adaptively without relying on external information. (In fact, it is minimax over an even larger smoothness class but we have focused on the Sobolev ball for simplicity.)

Thank you!