Wavelets

Christof Seiler

Stanford University, Spring 2016, Stats 205

Introduction

What we've seen so far

- Nonparametric regression using smoothers
- Different types of smoothers: e.g. kernel and local polynomial
- Penalized regression

Today

- Construct basis functions that are
  - Multiscale
  - Adaptive
- Find sparse set of coefficients for a given basis

Introduction

- In nonparametric regression we estimated the unkonwn function f directly
- With wavelets we use a orthogonal series representation of f
- This shifts the estimation problem
  - from directly estimating f
  - to estimating a set of scalar coefficients that represents f
- Similar to penalized regression but regularization will be replaced by thresholding
- Wavelets are used in the image file format JPEG 2000 to compress data

Assumptions

- Obervations

$$Y_i = f(x_i) + \epsilon_i \qquad i = 1, \ldots, n$$

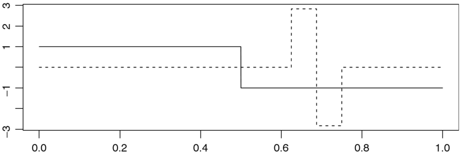
- The ϵ_i are iid
- The function f is square integrable $\int f^2 < \infty$
- Defined on a close interval $[a, b]$

Basis Function

- A set of functions $\Psi = \{\psi_1, \psi_2, \dots\}$ is called a basis for a class of functions $\mathcal{F}$
- If any function $f \in \mathcal{F}$ can be represented as a linear combination of the basis functions ψ_i
- Written as
$$f(x) = \sum_{i=1}^{\infty} \theta_i \psi_i(x)$$
with θ_i are scalar constants refered to as coefficients
- The constants θ_i are inner products of the function f and the basis functions ψ_i
$$\theta_i = \langle f, \psi_i \rangle = \int f(x)\psi_i(x)\, dx$$
- The basis is orthogonal if $\langle \psi_i, \psi_j \rangle = 0$ for $i \neq j$
- The basis is orthonormal if orthogonal and $\langle \psi_i, \psi_j \rangle = 1$

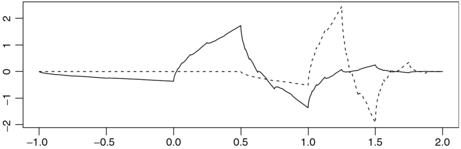
Basis Function

- Many sets of basis functions
- We consider orthonomal wavelet bases
- A simple wavelet function was first introduced by Haar in 1910



Source: Hollander, Wolfe, and Chicken (2013)

- More flexible and powerful wavelets were developed by Daubechies in 1992 and many others



Source: Hollander, Wolfe, and Chicken (2013)

Multiresolution Analysis

- We consider wavelet functions ψ

$$\Psi = \{\psi_{jk} : j, k \text{ integers}\}$$

  with

$$\psi_{jk} = 2^{j/2}\psi(2^j x - k)$$

  that form a basis for square-integrable functions
- Ψ is a collection of translations and dilations of ψ
- The ψ is constructed to ensure the the set Ψ is orthonormal
- The property $\int \psi_i^2 = 1$ implies that the value of ψ is near 0 except over a small range
- This property combined with the constrution above means that as j increases ψ_{jk} becomes increasingly localized

Multiresolution Analysis

- A careful construction of ψ leads to a multiresolution analysis
- It provides an interpretation of the wavelet representation f in terms of location and scale by rewritting

$$f(x) = \sum_{i=1}^{\infty} \theta_i \psi_i(x)$$

in terms of translation k and scaling j as ($\mathbb{Z}$ is set of integers)

$$f(x) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x)$$

- This can be intepreted as approximation at different scale j
- Here scale j is the same as frequency
- For a fixed j the index k represents behavior of f at resolution j and a particular location

Multiresolution Analysis

- Consider the approximation

$$f_J(x) = \sum_{j<J} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x)$$

- As J increases f_J is able to model smaller scales (higher frequncy) behavior of f
- Corresponds to changes that occur over smaller interval of the x-axis
- As J deceases f_J models larger scale (lower frequency) behavior of f
- Adding gloabl scaling term (think of it as the intercept)

$$f_J(x) = \sum_{k \in \mathbb{Z}} \xi_{j_0 k} \phi_{j_0 k}(x) + \sum_{j_0<j<J} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x)$$

Multiresolution Analysis

- Consider a simple example

$$f(x) = x, \quad x \in [0, 1)$$

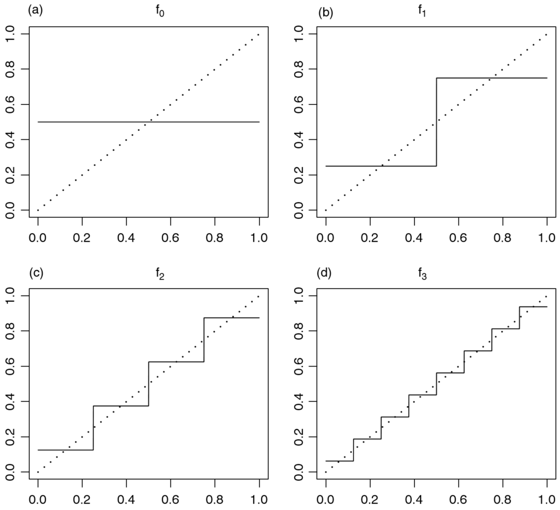
- The Haar wavelet functions are defined as

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2), \\ -1, & x \in [1/2, 1) \end{cases}$$

- and

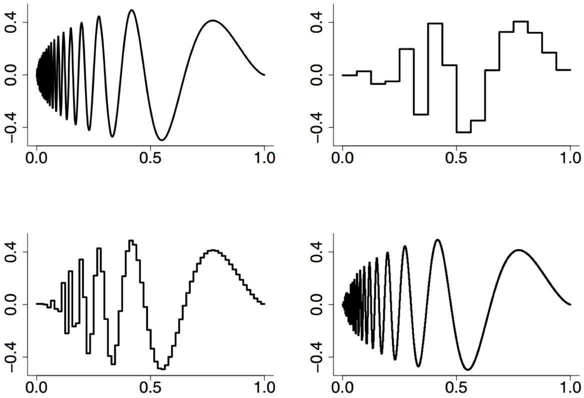
$$\phi(x) = 1, \quad x \in [0, 1)$$

Linear Example



Source: Hollander, Wolfe, and Chicken (2013)

Doppler Example



Source: Wasserman (2006)

Discrete Wavelet Transform

- The simple linear function example has exact solution to determine coefficients
- Usually this is not the case and numerical approximations are necessary to estimate coefficients
- One numerical methods is called the **cascade algorithm** by Mallat 1989
- It works if the sample size is a power of 2

  $$n = 2^J$$

  for some positive integer J
- Using this algorithm restricts the upper level of summation to $J - 1$ with

  $$J = \log_2(n)$$

Sparsity

- Wavelet methods are closely related to the concept of sparity
- A function
$$f(x) = \sum_j \theta_j \psi_j(x)$$
is sparse in a basis $\psi_1, \psi_2, \ldots$ if most of the θ_j are zero (or close it zero)
- Sparsity is not captured well by the L_2 norm but it is capture by the L_1 norm

Sparsity

- For example,

$$a = (1, 0, \ldots, 0) \qquad b = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$$

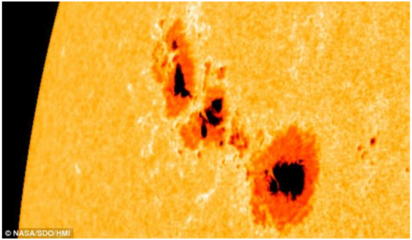
- then both have the same L_2 norm

$$\|a\|_2 = \sqrt{1 + 0 + \cdots + 0} = 1$$
$$\|b\|_2 = \sqrt{1/n + \cdots + 1/n} = \sqrt{n \times 1/n} = 1$$

- but with L_1 norm

$$\|a\|_1 = 1 + 0 + \cdots + 0 = 1$$
$$\|b\|_1 = 1/\sqrt{n} + \cdots + 1/\sqrt{n} = n \times 1/\sqrt{n} = \sqrt{n}$$

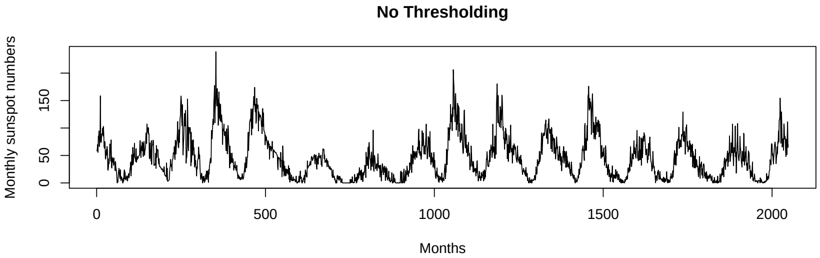
Wavelet Thresholding



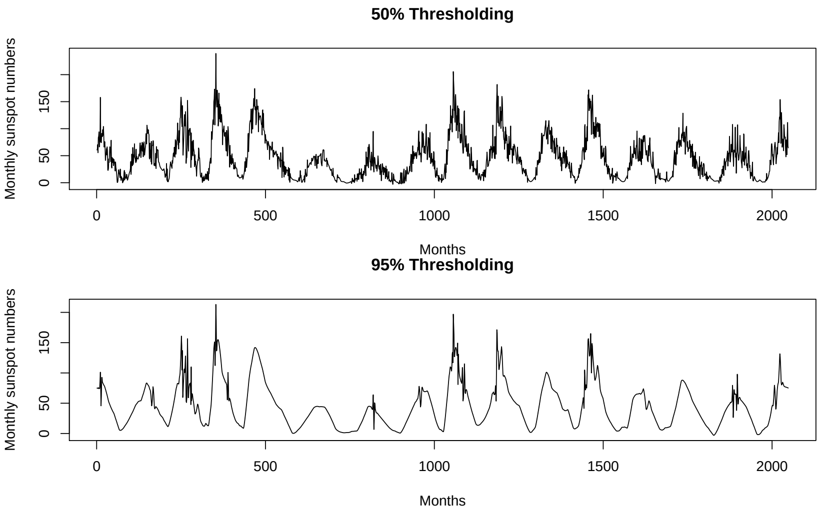
- Monthly sunspot numbers from 1749 to 1983
- Collected at Swiss Federal Observatory, Zurich until 1960, then Tokyo Astronomical Observatory
- Sunspots are temporary phenomena on the photosphere of the sun that appear visibly as dark spots compared to surrounding regions
- They correspond to concentrations of magnetic field flux that inhibit convection and result in reduced surface temperature compared to the surrounding photosphere

Wavelet Thresholding

- The original data has length 2820, but only the first 2048 are used here to make it a dyadic number
- So the modified data is now from January 1749 to July 1919



Wavelet Thresholding



Wavelet Thresholding

- The drawback of manual thresholding is the **subjective** choice of the threshold
- One might mistakenly chose to threshold all but few coefficients and oversmooth f
- Other methods are based on **theoretical** or data-driven considerations
- Many such methods are based on the **assumption** that the erros are **normally distributed**
- For instance: Donoho and Johnstone (1994). Ideal spatial adaptation via wavelet shrinkage

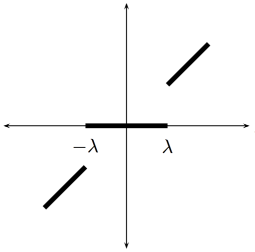
Wavelet Thresholding

- Hard thresholding (wavelet coefficient θ, threshold λ)

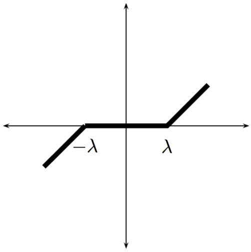
$$\eta_H(\theta, \lambda) = \theta \cdot I(|\theta| > \lambda)$$

- Soft thresholding

$$\eta_S(\theta, \lambda) = \mathrm{sgn}(\theta)(|\theta| - \lambda)_+ = \begin{cases} \theta + \lambda & \theta < -\lambda \\ 0 & -\lambda \leq \theta \leq \lambda \\ \theta - \lambda & \theta > \lambda \end{cases}$$



Hard thresholding



Soft thresholding

Source: Wasserman (2006)

Wavelet Thresholding

- The discrete wavelet transform operation may be represented in matrix form
$$\tilde{\theta} = Wy = Wf + W\epsilon$$
- Writing the unobserved coefficients as $\theta = Wf$ and the error coefficients as $\tilde{\epsilon} = W\epsilon$, we have
$$\tilde{\theta} = \theta + \tilde{\epsilon}$$
- The matrix W is orthogonal by design, so the $\tilde{\epsilon}$ are still normally distributed (under the normal error assumption)
- Unless the noise is excessive, the $\tilde{\epsilon}$ are generally smaller in magnitude than θ
- Which means that under the sparsity property, error coefficients may be ignored

Wavelet Thresholding

- Donoho and Johnstone make use of this and define soft thresholding rule to $\tilde{\theta}$ using the threshold

$$\lambda_v = \sqrt{2\sigma^2 \log(n)}$$

with σ^2 being the variance of the errors ϵ
- The variance is usually not known and needs to be estimated
- They propose the "VisuShrink" algorithm using thresholding η_S

$$\widehat{\theta} = \eta_S(\tilde{\theta}, \lambda_v)$$

- and the inverse discrete wavelet tranform W^{-1}

$$\widehat{f}_v = W^{-1}\widehat{\theta}$$

Wavelet Thresholding

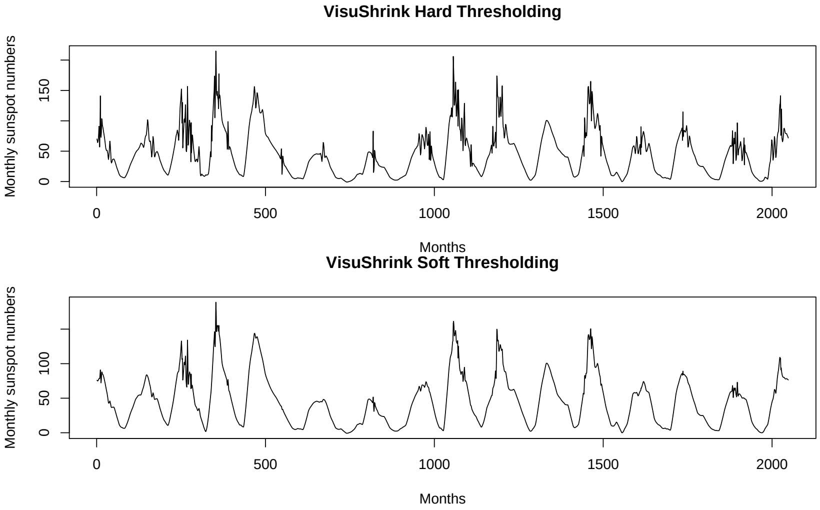
- In general, thresholding procedure:
  - decompose the data via discrete wavelet transform
  - apply some method of thresholding
  - reconstruct using the inverse wavelet tranform on the thresholded coefficients

$$\widehat{f} = W^{-1}\eta(Wy, \lambda)$$

- The threshold rule can be hard or soft threshold without affecting the asymptotic mean squared error

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}((f(x_i) - \widehat{f}_v(x_i))^2\right)$$

Wavelet Thresholding



Other Important Topics

- Different thresholding per level (Donhoho and Johnstone 1995) called “SureShrink”
- Thresholding without strong distributional assumptions on the errors using cross-validation (Nason 1996)
- Practical, simultaneous confidence bands for wavelet estimators are not available (Wasserman 2006)
- Standard wavelet basis functions are not invariant to translation and rotations
- Recent work by Mallat (2012) and Bruna & Mallat (2013) extend wavelets to handle these kind of invariances
- This provides a promising new direction for the theory of convolutional neural network

References

- Hollander, Wolfe, and Chicken (2013). Nonparametric Statistical Methods
- Wasserman (2006). All of Nonparametric Statistics
- Donoho and Johnstone (1994). Ideal Spatial Adaptation via Wavelet Shrinkage
- Donoho Johnstone (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage
- Nason (1996). Wavelet Shrinkage using Cross-Validation
- Mallat (2012). Group Invariant Scattering
- Bruna and Mallat (2013). Invariant Scattering Convolution Networks