Nonparametric Estimation: Part II

Regression in Transformed Space

Yanjun Han

Department of Electrical Engineering
Stanford University

yjhan@stanford.edu

November 20, 2015

Outline

Gaussian white noise model

The problem

Recover the function $f \in \mathcal{F}$ via noisy observation $(Y_t : t \in [0,1])$, where

$$dY_t = f(t)dt + \epsilon dB_t$$

where $(B_t : t \in [0,1])$ is the standard Brownian motion.

Relationship with the regression model: $\epsilon = \sigma/\sqrt{n}$.

Gaussian sequence estimation

Last time: estimation in function space

- Linear estimates: kernel estimator, local polynomial approximation estimator, etc...
- Nonlinear estimates: Lepski's adaptive estimator

Gaussian sequence estimation

Suppose that $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis in $L^2([0,1])$, then

$$y_j \triangleq \int_0^1 \phi_j(t) dY_t = (f, \phi_j) + \int_0^1 \phi_j(t) dB_t \triangleq \theta_j + \epsilon z_j$$

where $\{z_j\}_{j=1}^{\infty}$ are iid $\mathcal{N}(0,1)$ random noises.

If we adopt the L_2 risk, two estimation problems are equivalent due to L_2 isometry.

Projection estimator

The simplest estimate for $\{\theta_j\}_{j=1}^{\infty}$ is the projection estimator:

$$\hat{\theta}_j = \begin{cases} y_j, & 1 \le j \le m \\ 0, & j > m \end{cases}$$

MSE of the projection estimator

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 = \underbrace{\sum_{j>m} \theta_j^2}_{\text{Squared Bias}} + \underbrace{m\epsilon^2}_{\text{Variance}}$$

- The bias-variance tradeoff: bandwidth $m \uparrow$, bias $\downarrow$, variance $\uparrow$
- Choice of the basis is important: $\{\phi_j\}_{j=1}^{m}$ should correspond to the rate of the Kolmogorov m-width

Fourier transform and Sobolev ellipsoid

Consider the Sobolev ball $\mathcal{S}_1^{k,2}(L)$ given by

$$\mathcal{S}_1^{k,2}(L) = \left\{ f \in C[0,1] : \int_0^1 |f^{(k)}(x)|^2 dx \leq L^2 \right\}$$

and consider the following Fourier basis:

$$\phi_0(t) = 1, \phi_{2j-1}(t) = \sqrt{2}\cos(2\pi jt), \phi_{2j}(t) = \sqrt{2}\sin(2\pi jt)$$

Lemma

$f \in \mathcal{S}_1^{k,2}(L)$ if and only if $\theta_j = (f, \phi_j)$ satisfies

$$\sum_{j=1}^{\infty} (2\pi j)^{2k} (|\theta_{2j-1}|^2 + |\theta_{2j}|^2) \leq L^2.$$

Estimation in ellipsoids

Consider the Gaussian mean estimation problem with parameter set

$$\Theta(L) = \left\{ \theta : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq L^2 \right\}$$

where $0 < a_1 \leq a_2 \leq \cdots$ and $a_j \to \infty$.

- The linear estimator: $\hat{\theta} = \{\hat{\theta}_j\}_{j=1}^{\infty}$ with $\hat{\theta}_j = c_j y_j, j \geq 1$
- The MSE:

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 = \sum_{j=1}^{\infty} (1 - c_j)^2 \theta_j^2 + c_j^2 \epsilon^2$$

Minimax linear estimator

Minimax linear estimator

The optimal weight sequence $\{c_j^*\}_{j=1}^{\infty}$ is the solution to the optimization problem: $L(\{c_j^*\}_{j=1}^{\infty}) = \min L(\{c_j\}_{j=1}^{\infty})$ where

$$L(\{c_j\}_{j=1}^{\infty}) = \max_{\sum_{j=1}^{\infty} a_j^2\theta_j^2 \le L^2} \sum_{j=1}^{\infty}(1-c_j)^2\theta_j^2 + c_j^2\epsilon^2$$

- Intuitively, we have $\frac{1-c_j^*}{a_j} = \lambda$ for small j, and $c_j^* = 0$ for large j.
- Minimax theorem can be used here to swap the order of min and max

Lemma

The minimax linear estimator $\hat{\theta}^* = \{\hat{\theta}_j^*\}_{j=1}^{\infty}$ is given by $\hat{\theta}_j^* = c_j^* y_j$, where $c_j^* = (1-\lambda a_j)_+$ and λ is the solution to the following equation

$$\frac{\epsilon^2}{\lambda}\sum_{j=1}^{\infty} a_j(1-\lambda a_j)_+ = L^2.$$

Pinsker's theorem

Pinsker's theorem

If $0 < a_1 \leq a_2 \leq \cdots$ and $a_j \to \infty$ in the Gaussian mean estimation problem, the linear estimator $\hat{\theta}^*$ is asymptotically minimax:

$$\sup_{\theta \in \Theta(L)} \mathbb{E}_\theta \|\hat{\theta}^* - \theta\|^2 = (1 + o(1)) \inf_{\hat{\theta}} \sup_{\theta \in \Theta(L)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

where $o(1) \to 0$ as $\epsilon \to 0$.

- The estimator is asymptotically minimax even in constants!

For estimation in Sobolev ellipsoids, we have the following corollary:

Corollary (Sobolev ellipsoid estimation)

$$\inf_{\hat{f}} \sup_{f \in \mathcal{S}_1^{k,2}(L)} \mathbb{E}_f \|\hat{f} - f\|_2^2 = (1 + o(1)) C^* \epsilon^{\frac{4k}{2k+1}}$$

where $C^* = (L^2(2k+1))^{\frac{1}{2k+1}} (\frac{k}{1+k})^{\frac{2k}{2k+1}}$ is Pinsker's constant.

James-Stein estimator

Pinsker's theorem requires perfect knowledge of the coefficients $\{a_n\}$ and the radius L

- Adaptive estimation: find an estimator which performs nearly as well as the best linear estimates

Consider the estimation problem of $\theta \in \mathbb{R}^d$ in the model $y \sim \mathcal{N}(\theta, \epsilon^2 I_d), d \geq 3$, the James-Stein estimator is defined as

$$\hat{\theta}^{\text{JS}} = \left(1 - \frac{(d-2)\epsilon^2}{\|y\|^2}\right)_+ y$$

Theorem (Oracle inequality for James-Stein estimator)

$$\inf_{c \in \mathbb{R}} \mathbb{E}_\theta \|cy - \theta\|^2 \leq \mathbb{E}_\theta \|\hat{\theta}^{JS} - \theta\|^2 \leq 2\epsilon^2 + \inf_{c \in \mathbb{R}} \mathbb{E}_\theta \|cy - \theta\|^2$$

Block James-Stein estimator

Divide the parameter vector θ into dyadic blocks $\{B_i\}$ with $B_i = \{j : \ell_{i-1} \le j < \ell_i\}, \ell_i/\ell_{i-1} = r > 1$, and write

$$\Theta(L) = \left\{ \theta : \sum_{i=1}^{\infty} \sum_{j \in B_i} a_j^2 \theta_j^2 \le L^2 \right\}.$$

In the sequel we assume that $a_j = (1 + o(1))j^k$, e.g., $f \in \mathcal{S}_1^{k,2}(2\pi L)$ expressed in Fourier basis.

Block James-Stein estimator

$$\hat{\theta}_i^{\mathrm{BJS}}(y) = \begin{cases} y_i & i \le I_0 \\ \left(1 - \frac{(\ell_i - 2)\epsilon^2}{\|y_i\|^2}\right)_+ y_i & I_0 < i \le I_\epsilon \\ 0 & i > I_\epsilon \end{cases}$$

where y_i is the observation vector in block B_i.

Choice of the threshold

Distinguish three cases:

- $i \le I_0$: $\mathbb{E}_\theta \|\hat{\theta}_i^{\mathrm{BJS}} - \theta_i\|^2 = (\ell_{i+1} - \ell_i)\epsilon^2$
- $I_0 < i \le I_\epsilon$: $\mathbb{E}_\theta \|\hat{\theta}_i^{\mathrm{BJS}} - \theta_i\|^2 \le 2\epsilon^2 + \inf_{c_i \in \mathbb{R}} \mathbb{E}_\theta \|c_i y_i - \theta_i\|^2$ by oracle inequality
- $i > I_\epsilon$: $\mathbb{E}_\theta \|\hat{\theta}_i^{\mathrm{BJS}} - \theta_i\|^2 = \|\theta_i\|^2$

Add them together:

$$\mathbb{E}_\theta \|\hat{\theta}^{\mathrm{BJS}} - \theta\|^2 \le (2I_\epsilon - 2I_0 + \ell_{I_0})\epsilon^2 + \sum_{j > \ell_{I_\epsilon}} \theta_j^2 + \sum_{i=1}^{\infty} \inf_{c_i \in \mathbb{R}} \mathbb{E}_\theta \|c_i y_i - \theta_i\|^2$$

Since when $a_j = (1 + o(1))j^k$, the optimal MSE is $\asymp \epsilon^{\frac{4k}{1+2k}}$, we can choose

- I_0 any fixed constant
- $L^2 / a_{\ell_{I_\epsilon}}^2 \le \epsilon^2 \Leftrightarrow I_\epsilon \asymp \ln(1/\epsilon)$

to make the first two terms negligible.

Performance of block James-Stein estimator

The previous argument shows that

$$\mathbb{E}_\theta \|\hat{\theta}^{\mathrm{BJS}} - \theta\|^2 \le (1 + o(1)) \sum_{i=1}^{\infty} \inf_{c_i \in \mathbb{R}} \mathbb{E}_\theta \|c_i y_i - \theta_i\|^2$$

Note that $\sup_{j,j' \in B_i} a_j / a_{j'} \sim (\ell_i / \ell_{i-1})^k = r^k \to 1$ as $r \to 1$, the best blockwise linear estimator performs similarly to the best coordinatewise linear estimator. A closer inspection of the linear minimax estimator shows that the multiplicative gap is at most r^{2k}.

Theorem (Efroimovich-Pinsker'84)

If $a_j \asymp j^k$ for some $k > 0$,

$$\sup_{\theta \in \Theta(L)} \mathbb{E}_\theta \|\hat{\theta}^{BJS} - \theta\|^2 \le (r^{2k} + o(1)) \cdot \inf_{\hat{\theta}} \sup_{\theta \in \Theta(L)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

and block James-Stein estimator is adaptively minimax if we choose $r = \ell_{i+1}/\ell_i \to 1$.

Linear estimators in function space

Kernel estimator:

$$\hat{f}(x) = \int K_h(x-t)dY_t$$

- Write $\hat{f} = K_h * Y$, by Fourier transform we have $\hat{\theta}_j = \hat{K}_h(j) \cdot y_j$

Smoothed spline estimator:

$$\hat{f} = \arg\min_g \int (dY_t - g(t)dt)^2 + \lambda \int |g''(t)|^2 dt$$

- Parseval's inequality yields

$$\hat{\theta}_j = \arg\min_{x_j} (y_j - x_j)^2 + \lambda j^2 x_j^2$$

which yields a linear estimator again: $\hat{\theta}_j = \frac{y_j}{1+\lambda j^2}$.

More on Fourier transform and adaptive scheme

Question 1: Is the Fourier basis the right basis in other spaces?

- In general, no!
- We will show that, in Besov space, the wavelet basis is the right one.

Question 2: in adaptive estimation, what if a_j does not increase polynomially with j?

- Block James-Stein estimator fails: oscillation within blocks $\sup_{j,j' \in B_i} a_j / a'_j$ may be really large
- However, under mild conditions on $\{a_j\}$, aggregation using projection estimates can still yield adaptive minimax estimator (next lecture)

Multiresolution analysis: father wavelets

Fix some function $\varphi(\cdot) \in L^2(\mathbb{R})$ such that $\{\varphi(\cdot - k), k \in \mathbb{Z}\}$ forms an orthonormal system, i.e.,

$$\int \varphi(x-k)\varphi(x-l)dx = \delta_{kl}, \qquad k, l \in \mathbb{Z}$$

Nested sequence of linear spaces:

- Define $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$, and the linear space

$$V_j = \{f : f(x) = \sum_k c_k \varphi_{jk}(x), \{c_k\} \in \ell_2\}$$

- We hope that $\{V_j\}$ is nested: $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$
- $\varphi(\cdot)$ is called the father wavelet, level j corresponds to resolution.

Example: $\varphi(x) = \chi_{[0,1]}(x)$ is called the Haar basis.

Multiresolution analysis: mother wavelets

Denote by W_j the orthogonal complement of V_j in V_{j+1}, i.e., $W_j = V_{j+1} \ominus V_j$, then

$$V_j = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_{j-1}$$

- If $\cup_{j=1}^{\infty} V_j$ is dense in $L^2(\mathbb{R})$, and denote by $\{\psi_{jk}(x)\}$ the orthonormal basis of W_j, then any $f \in L^2$ can be written as

$$f(x) = \sum_k \alpha_k \varphi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_k \beta_{jk} \psi_{jk}(x)$$

  where $j_0 \in \mathbb{Z}$ is arbitrary initial resolution level.
- If there exists some function $\psi(\cdot) \in L^2(\mathbb{R})$ such that $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, we call $\psi(\cdot)$ the mother wavelet.

Example: for Haar basis, mother wavelet is $\psi(x) = \chi_{[0,1/2]}(x) - \chi_{[1/2,1]}(x)$

Conditions for father and mother wavelets

Question 1: when $\{\varphi(\cdot - k), k \in \mathbb{Z}\}$ forms an orthonormal system?

- Answer: $\sum_k |\hat{\varphi}(\omega + 2\pi k)|^2 = 1$ a.e.

Question 2: when $\{V_j\}$ is a nested sequence?

- Answer: there exists a 2π-periodic function $m_0(\omega)$ such that $\hat{\varphi}(\omega) = m_0(\omega/2)\hat{\varphi}(\omega/2)$

Question 3: when ψ_{jk} can be expressed as $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$?

- Answer: $\hat{\psi}(\omega) = m_1(\omega/2)\hat{\varphi}(\omega/2)$, where $m_1(\omega) = m_0^*(\omega + \pi)e^{-i\omega}$

Question 4: when $\cup_{j=1}^{\infty} V_j$ is dense in $L_2(\mathbb{R})$?

- Answer: it is sufficient to let $\varphi(x)$ satisfy the previous two conditions and

$$|\varphi(u)| \leq \Phi(|u|), \qquad u \in \mathbb{R}$$

where $\Phi(\cdot)$ is a bounded nonincreasing function on $[0, \infty)$ and $\int \Phi(|u|)du < \infty$.

Compactly supported wavelets

To obtain a finite summation over k, we hope that both wavelets $\varphi(\cdot)$ and $\psi(\cdot)$ have compact supports.

Theorem (Daubechies' construction)

For each integer $N > 0$, there exists father wavelet $\varphi(\cdot)$ supported on $[0, 2N-1]$ and the corresponding mother wavelet $\psi(\cdot)$ supported on $[-N+1, N]$ such that

$$\int \psi(x) x^l = 0, \qquad l = 0, 1, \cdots, N-1$$

- There also exist so-called coiflets and symmlets whose father wavelet $\varphi(\cdot)$ also have first vanishing $N-1$ moments.

Projection operator

Denote by P_V the orthogonal projection operator to linear space V.

Lemma

For compactly supported mother wavelet $\psi(\cdot)$ with vanishing first $N-1$ moments, we have

$$P_{V_0} p = p, \qquad \forall p \in \mathcal{P}_1^{N-1}$$

Projection operator as a kernel:

$$\begin{aligned} P_{V_0} f(x) &= \sum_k \left(\int f(y)\varphi(y-k)dy \right) \varphi(x-k) \\ &= \int \left(\sum_k \varphi(x-k)\varphi(y-k) \right) f(y)dy \equiv \int K_0(x,y) f(y) dy \end{aligned}$$

where $K_0(x,y) = \sum_k \varphi(x-k)\varphi(y-k)$ is the projection kernel.

- Similarly, $K_j(x,y) = 2^j K_0(2^j x, 2^j y) = \sum_k \varphi_{jk}(x)\varphi_{jk}(y)$

Approximation by wavelets

Theorem (Wavelet approximation in Sobolev space)

For compactly supported mother wavelet $\psi(\cdot)$ with vanishing first $N-1$ moments, and for $f \in \mathcal{S}_1^{N,p}$ with $p \in [1,\infty]$, we have

$$\|f - K_j f\|_p \lesssim 2^{-jN} \|f^{(N)}\|_p, \qquad j \to \infty$$

Proof:

- To bound $|f(x) - K_j f(x)|$, denote by $g_x(\cdot)$ the Taylor polynomial of degree $N-1$ of $f(\cdot)$ at x
- By the previous lemma, $K_j[g_x](x) = g_x(x) = f(x)$
- Hence, $|f(x) - K_j f(x)| = |K_j[f - g_x](x)| = |K_0[\tilde{f}_j - \tilde{g}_{x,j}](2^j x)|$, where $\tilde{f}_j(y) = f(y/2^j)$
- Then by homogeneity and Taylor's formula, it is easy to show that

$$\|f - K_j f\|_p \lesssim \|\tilde{f}_j^{(N)}(2^j x)\|_p = 2^{-jN} \|f^{(N)}\|_p$$

Modulus of smoothness

Definition (Modulus of smoothness)

Define the r-th symmetric difference operator Δ_h^r by

$$\Delta_h^r f(x) = \Delta_h(\Delta_h^{r-1} f)(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f(x + \left(k - \frac{r}{2}\right) h)$$

and the r-th order modulus of smoothness $\omega^r(f,t)_p$ by

$$\omega^r(f,t)_p = \sup_{0<h\leq t} \|\Delta_h^r f\|_p.$$

Some properties:

- $\omega^r(f,t)_p \to 0$ as $t \to 0$ if $f \in L^p$ for $1 \leq p < \infty$, or $f \in C$ for $p = \infty$
- $\omega^r(f+g,t)_p \leq \omega^r(f,t)_p + \omega^r(g,t)_p$, $\omega^r(f,\lambda t)_p \leq (\lambda+1)^r \omega^r(f,t)_p$, $\omega^{r+1}(f,t)_p \leq 2\omega^r(f,t)_p \lesssim t^r \int_t^\infty \frac{\omega^{r+1}(f,u)_p}{u^{r+1}} du$
- $f \in \mathcal{S}_1^{r,p} \Rightarrow \omega^r(f,t)_p \leq t^r \|f^{(r)}\|_p$, $\omega^{r+k}(f,t)_p \leq t^r \omega^k(f^{(r)},t)_p$
- If $\liminf_{t\to 0} \omega^r(f,t)_p/t^r = 0$, we have $f \in \mathcal{P}_1^{r-1}$

Besov space

Definition (Besov space)

Define $f \in \mathcal{B}^s_{p,q}$ if and only if, for $r = \lfloor s \rfloor + 1$,

$$\|f\|_{\mathcal{B}^s_{p,q}} = \|f\|_p + \begin{cases} \left[\int_0^\infty \left(\frac{\omega^r(f,t)_p}{t^s}\right)^q \cdot \frac{dt}{t}\right]^{\frac{1}{q}} & 1 \le q < \infty \\ \sup_{t>0} \frac{\omega^r(f,t)_p}{t^s} & q = \infty \end{cases}$$

exists and is finite. Define $f \in \mathcal{B}^s_{p,q}(L)$ iff $\|f\|_{\mathcal{B}^s_{p,q}} \le L$.

Properties:

- Monotonicity in q: $q \le q' \Rightarrow \mathcal{B}^s_{p,q} \subset \mathcal{B}^s_{p,q'}$
- Sobolev space: $\mathcal{B}^k_{p,1} \subset \mathcal{S}^{k,p}_1 \subset \mathcal{B}^k_{p,\infty}$
- Embedding theorem: $p \le p' \Rightarrow \mathcal{B}^s_{p,q} \subset \mathcal{B}^{s-1/p+1/p'}_{p',q}$
- Continuous embedding: $s > 1/p \Rightarrow \mathcal{B}^s_{p,q} \subset C$

Equivalent characterization I: discrete norm

Breaking the integral $\int_0^\infty$ into $\sum_j \int_{2^j}^{2^{j+1}}$ and applying the inequality $\omega^r(f,t)_p \le \omega^r(f,2t)_p \le 2^r\omega^r(f,t)_p$, we have the following equivalent characterization.

Theorem (Discrete characterization of Besov space)

The Besov norm $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ *is equivalent to*

$$\|f\|_{\tilde{\mathcal{B}}^s_{p,q}} = \|f\|_p + \|\{2^{js}\omega^r(f,2^{-j})_p\}\|_{\ell_q}$$

Corollary

$f \in \mathcal{B}^s_{p,q}$ if and only if $f \in L_p$ and $\{2^{js}\omega^r(f,2^{-j})_p\} \in \ell_q$, $r = \lfloor s \rfloor + 1$.

- Motivate us to apply multiresolution analysis!

Equivalent characterization II: Paley-Littlewood decomposition

Theorem (Paley-Littlewood decomposition)

$f \in \mathcal{B}^s_{p,q}$ if and only if there exist functions $\{f_j\}_{j=0}^{\infty}$ such that $f = \sum_{j=0}^{\infty} f_j$ (weakly), and

$$\|f_j\|_p \le 2^{-js}\epsilon_j, \quad \|f_j^{(r)}\|_p \le 2^{j(r-s)}\epsilon_j'$$

where $\{\epsilon_j\} \in \ell_q, \{\epsilon_j'\} \in \ell_q$.

Proof of necessity:

- Define f_j as the result by passing f through a bandpass filter which picks out frequency $[2^j, 2^{j+1}]$. It's to easy to have $f = \sum_{j=0}^{\infty} f_j$.
- The time width of filter j is $\asymp 2^{-j}$, thus $\|f_j\|_p \lesssim \omega^r(f, 2^{-j})_p$
- Since the frequency of f_j cannot exceed 2^{j+1}, $\|f^{(r)}\|_p \lesssim 2^{jr}\|f\|_p$

Proof of sufficiency: apply the previous decomposition with localized frequencies to each f_j

Equivalent characterization III: kernel approximation

Theorem (Kernel approximation in Besov space)

Fix a kernel $K(x,y)$ which maps all polynomials of degree $\lfloor s \rfloor$ to themselves and satisfies $|K(x,y)| \le F(x-y)$, $\int |x|^r F(x) < \infty$. Define $K_j f(x) = \int 2^j K(2^j x, 2^j y) f(y) dy$, then $f \in \mathcal{B}^s_{p,q}$ if and only if

$$f \in L_p, \qquad \epsilon_j = 2^{js} \|K_j f - f\|_p \in \ell_q$$

- Insights: kernel estimator with bandwidth h has bias $O(h^s)$
- Can be proved via Paley-Littlewood decomposition, but more directly by introducing the K-functional

K-functional and modulus of smoothness

Definition (K-functional)

The K-functional $K_r(f,t)_p$ is defined as

$$K_r(f,t)_p = \inf_{g^{(r-1)} \in \text{A.C.}_{\text{loc}}} \|f-g\|_p + t\|g^{(r)}\|_p$$

Theorem (Equivalence of K-functional and modulus of smoothness)

$$\omega^r(f,t)_p \asymp K_r(f,t^r)_p$$

Proof of necessity:

- $\|K_j f - f\|_p \le \|K_j(f-g)\|_p + \|f-g\|_p + \|K_j g - g\|_p \lesssim \|f-g\|_p + 2^{-jr}\|g^{(r)}\|_p$
- Minimize over g and apply the equivalence to get $\|K_j f - f\|_p \lesssim K_r(f,2^{-jr})_p \lesssim \omega^r(f,2^{-j})_p$

Proof of sufficiency: substitute $g = K_j f$ in the definition of K-functional and apply the opposite inequality

Equivalent characterization IV: wavelet approximation

Consider the wavelet basis generated by a compactly supported mother wavelet $\psi(\cdot)$ with first $\lfloor s \rfloor$ vanishing moments, the previous theorem entails that $f \in \mathcal{B}^s_{p,q}$ if and only if

$$\|P_{V_j} f - f\|_p = 2^{-js}\epsilon_j, \qquad \{\epsilon_j\} \in \ell_q$$

This condition is further equivalent to

$$\|P_{W_j} f\|_p = \|P_{V_{j+1}} f - P_{V_j} f\|_p = 2^{-js}\epsilon'_j, \qquad \{\epsilon'_j\} \in \ell_q$$

Theorem (Wavelet approximation of Besov spaces)

Fix the wavelet basis described above. Then $f \in \mathcal{B}^s_{p,q}$ if and only if

$$P_{V_0} f \in L^p, \qquad \{2^{js}\|P_{W_j} f\|_p\} \in \ell_q$$

Inequality for wavelet coefficients

Lemma

If $\sum_k |\varphi(x-k)| \leq M$ and $\{\varphi(x-k) : k \in \mathbb{Z}\}$ constitutes an orthonormal system, there exists constants C_1, C_2 such that

$$C_1 \|\lambda\|_{\ell_p} \leq \|\sum_k \lambda_k \varphi(x-k)\|_p \leq C_2 \|\lambda\|_{\ell_p}$$

Corollary

By homogeneity, we have

$$C_1 \|\lambda\|_{\ell_p} 2^{\frac{j}{2}-\frac{j}{p}} \leq \|\sum_k \lambda_k \varphi_{jk}(x)\|_p \leq C_2 \|\lambda\|_{\ell_p} 2^{\frac{j}{2}-\frac{j}{p}}$$

Wavelet coefficients for function in Besov space

Combining them together yields:

Theorem (Parameter set for wavelet coefficients)

Consider the wavelet basis with compactly supported mother wavelet $\psi(\cdot)$ with vanishing first $R \geq \lfloor s \rfloor$ moments, and

$$f(x) = \sum_k \alpha_{j_0 k} \varphi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_k \beta_{jk} \psi_{jk}(x)$$

then the Besov norm $\|\cdot\|_{B^s_{p,q}}$ is equivalent to the norm $\|\cdot\|_{b^s_{p,q}}$, where

$$\begin{aligned} \|f\|_{b^s_{p,q}} &= \|\alpha_{j_0}\|_{\ell_p} + \|2^{j(s+\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_{\ell_p}\|_{\ell_q} \\ &= \left(\sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0 k}|^p \right)^{\frac{1}{p}} + \left[\sum_{j=j_0}^{\infty} \left(2^{j(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{k=0}^{2^j-1} |\beta_{jk}|^p \right)^{\frac{1}{p}} \right)^q \right]^{\frac{1}{q}} \end{aligned}$$

Ideal truncated estimate

Consider the Gaussian mean estimation model $y_k = \theta_k + \epsilon z_k$ with known parameter θ, but we constrain our estimator to be either $\hat{\theta}_k = y_k$ or $\hat{\theta}_k = 0$

- It is easy to show that the ideal truncated estimator is

$$\hat{\theta}_k = y_k \mathbb{1}(\theta_k \geq \epsilon)$$

- The corresponding MSE is $R_T(\theta) = \sum_k \min\{\theta_k^2, \epsilon^2\}$

Theorem (Donoho-Liu-MacGibbon'90)

If the parameter set is a hyperrectangle $\Theta(\tau) = \prod_{i=1}^{\infty}[-\tau_i, \tau_i]$, *we have*

$$\sup_{\theta \in \Theta(\tau)} R_T(\theta) \leq 2.22 \times \inf_{\hat{\theta}} \sup_{\theta \in \Theta(\tau)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

Solid orthosymmetric parameter set

Definition (Solid orthosymmetric parameter set)

The parameter set Θ is called solid and orthosymmetric if and only if: $\theta = \{\theta_i\}_{i \in I} \in \Theta$ implies $\{\lambda_i \theta_i\}_{i \in I} \in \Theta$ for any $\lambda_i \in [-1, 1], i \in I$.

If Θ is solid and orthosymmetric, the minimax L_2 risk over Θ can be decomposed into

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 = \inf_{\hat{\theta}} \sup_{\Theta(\tau) \subset \Theta} \sup_{\theta \in \Theta(\tau)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

$$\geq \frac{1}{2.22} \sup_{\Theta(\tau) \subset \Theta} \sup_{\theta \in \Theta(\tau)} R_T(\theta) = \frac{1}{2.22} \sup_{\theta \in \Theta} R_T(\theta)$$

Lemma

If Θ is solid and orthosymmetric, we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \geq \frac{1}{2.22} \sup_{\theta \in \Theta} R_T(\theta) = \frac{1}{2.22} \sup_{\theta \in \Theta} \sum_{k \in I} \min\{\theta_k^2, \epsilon^2\}$$

Thresholding estimator

Definition (Thresholding estimators)

The soft- and hard-thresholding estimators $\eta_t^s(\cdot)$ and $\eta_t^h(\cdot)$ with threshold t are defined as

$$\eta_t^s(y) = \mathrm{sgn}(y)(|y| - t)_+, \qquad \eta_t^h(y) = y\mathbb{1}(|y| \geq t)$$

Theorem (Donoho-Johnstone'94)

For Gaussian mean estimation $y_i = \theta_i + \epsilon z_i, 1 \leq i \leq n$, $t = \epsilon\sqrt{2\ln n}$ yields

$$\mathbb{E}_\theta \|\eta_t^s(y) - \theta\|^2 \leq (2\ln n + 1)\left(\sum_{i=1}^n \min\{\theta_i^2, \epsilon^2\} + \epsilon^2\right)$$

and similar results hold for hard-thresholding with $t = \epsilon\sqrt{2\ln n + \ln\ln n}$.

- Choice of the threshold: $\mathbb{P}(\max_{1\leq i\leq n} |z_i| \geq \sqrt{2\ln n}) \to 0$

VisuShrink estimator

Consider the Besov ball $\mathcal{B}^s_{p,q}(L)$ and the wavelet model:

$$y_{j,k} = \theta_{j,k} + \epsilon z_{j,k}, \quad j \geq j_0, 0 \leq k \leq 2^j - 1, \theta \in \Theta$$

where Θ is solid and orthosymmetric.

VisuShrink estimator

$$\hat{\theta}_{j,k}(y) = \begin{cases} y_{j,k} & j = j_0 \\ \eta^s_t(y_{j,k}) & j_0 < j \leq j_\epsilon \\ 0 & j > j_\epsilon \end{cases}$$

where $t = \epsilon\sqrt{2 \ln n_\epsilon}$, and n_ϵ is the number of observations to which nonlinearity applies.

Choice of the parameters:

- j_0 can be any fixed constant, e.g., $j_0 = 2$
- j_ϵ is chosen s.t. $\sup_{\theta \in \Theta} \sum_{j > j_\epsilon} \sum_k |\theta_{j,k}|^2 \leq \epsilon^2$, yielding $j_\epsilon \asymp \ln(1/\epsilon)$
- As a result, $\ln n_\epsilon \asymp \ln(1/\epsilon)$

Adaptive optimality of VisuShrink

Performance of VisuShrink estimator:

$$\mathbb{E}_\theta \|\hat{\theta}^{VISU} - \theta\|^2 \le 2^{j_0}\epsilon^2 + (2\ln n_\epsilon + 1)\left(\sum_{j_0<j\le j_\epsilon}\sum_k \min\{\theta_{j,k}^2, \epsilon^2\} + \epsilon^2\right) + \epsilon^2$$
$$\lesssim \ln(1/\epsilon)\sum_{j_0<j\le j_\epsilon}\sum_k \min\{\theta_{j,k}^2, \epsilon^2\} + O(\epsilon^2\ln(1/\epsilon))$$

Since $\Theta^s_{p,q}(L)$ is solid and orthosymmetric, we have:

Theorem (Donoho-Johnstone'94)

The VisuShrink estimator is near optimal and adaptive:

$$\sup_{\theta\in\Theta^s_{p,q}(L)} \mathbb{E}_\theta\|\hat{\theta}^{VISU} - \theta\|^2 \lesssim \ln(1/\epsilon)\cdot \inf_{\hat{\theta}} \sup_{\theta\in\Theta^s_{p,q}(L)} \mathbb{E}_\theta\|\hat{\theta} - \theta\|^2$$

Unconditional basis is the best basis

Definition (Unconditional basis)

$\{f_i\}_{i\in I}$ is an unconditional basis for $(\mathcal{F}, \|\cdot\|)$ if and only if there exists a universal constant C such that for any $J \subset I$ and $\lambda_i \in [-1, 1]$,

$$\|\sum_{i\in J} \lambda_i f_i\| \leq C\|\sum_{i\in J} f_i\|$$

- In transformed space Θ, unconditional basis is equivalent to $c_1\Theta' \subset \Theta \subset c_2\Theta'$ for some $c_1, c_2 > 0$ and solid orthosymmetric Θ'
- Wavelet basis is an unconditional basis for Besov space

Optimality in terms of the ideal truncated estimator: for unconditional basis $\{f_i\}$ (resp. Θ) and any other basis $\{f_i'\}$ (resp. Θ')

$$\sup_{\theta\in\Theta} R_T(\theta) \lesssim \inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_f\|\hat{f}-f\|^2 \leq \sup_{\theta'\in\Theta'} \mathbb{E}_\theta\|\hat{\theta}^{VISU}-\theta\|^2 \lesssim \ln(1/\epsilon)\cdot \sup_{\theta'\in\Theta'} R_T(\theta')$$

Block thresholding

Instead of individual truncation, length-L block truncation can also be implemented, with ideal risk

$$R_{BT}(\theta) = \sum_{j=1}^{n/L} \min\{\|\theta_{B_j}\|^2, L\epsilon^2\}$$

- Use James-Stein estimator $\hat{\theta}_j^\lambda = (1 - \frac{\lambda L\epsilon^2}{\|y_j\|^2})_+ y_j$ in each block

Theorem (Oracle inequality, Cai'99)

$$\mathbb{E}_\theta \|\hat{\theta}^\lambda - \theta\|^2 \leq \lambda R_{BT}(\theta) + 4n\epsilon^2 \mathbb{P}(\chi_L^2 \geq \lambda L)$$

- VisuShrink corresponds to $L = 1, \lambda \sim 2 \ln n$
- It can be better to use $L = \ln n, \lambda = 4.50524$

Minimax Bayes estimation

Consider the Gaussian mean estimation model in $\Theta^s_{p,q}(L)$ with

$$\|\theta\|_{b^s_{p,q}} = \|\alpha_{j_0}\|_{\ell_p} + \|2^{j(s+\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_{\ell_p}\|_{\ell_q} \le L$$

Minimax Bayes estimation:

- Replace the hard constraint $\theta \in \Theta^s_{p,q}(L)$ with an "in mean" constraint, i.e.

$$\tau \in \Theta^s_{p,q}(L), \quad \tau_{j,k} = (\mathbb{E}_\pi|\theta_{j,k}|^{p\wedge q})^{1/(p\wedge q)}$$

- The minimax Bayes estimation:

$$\inf_{\hat\theta} \sup_{\pi:\tau\in\Theta^s_{p,q}(L)} \mathbb{E}_\pi\mathbb{E}_\theta\|\hat\theta-\theta\|^2 = \sup_{\pi:\tau\in\Theta^s_{p,q}(L)} \inf_{\hat\theta} \mathbb{E}_\pi\mathbb{E}_\theta\|\hat\theta-\theta\|^2$$

  where the supremum is taken with respect to all prior π.

Solution to the minimax Bayes estimation

Due to convexity we have the following lemma.

Lemma

Separable rules are minimax, i.e., it suffices to consider independent priors to different $\theta_{j,k}$. Moreover, $\tau_{j,k}$ does not depend on k.

Denote by t_j the identical value shared by all $\tau_{j,k}$, the minimax Bayes estimation problem reduces to

$$\max \sum_{j=j_0}^{\infty} 2^j \rho_{p\wedge q}(t_j, \epsilon) \qquad s.t. \sum_{j=j_0}^{\infty} (2^{j(s+1/2)} t_j)^q \le L^q$$

where

$$\rho_p(\tau, \epsilon) = \inf_{\hat{\theta}} \sup_{\pi:\mathbb{E}|\theta|^p \le \tau^p} \mathbb{E}_\pi \mathbb{E}_\theta (\hat{\theta} - \theta)^2$$

is the minimax Bayes risk in the univariate model $y = \theta + \epsilon z$.

Thresholding estimators are near minimax!

Theorem (Minimax estimation over ℓ_q balls)

In the Gaussian mean estimation over ℓ_q balls, the soft- and hard-thresholding estimators with proper thresholds are near minimax:

$$\inf_{\{t_j\}} \sup_{\theta:\|\theta\|_q \le L} \mathbb{E}_\theta \|\{\eta^x_{t_j}(y_j)\} - \theta\|^2 \lesssim \inf_{\hat{\theta}} \sup_{\theta:\|\theta\|_q \le L} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

where $x = s, h$.

Applying to the previous problem and using the equivalence of minimax estimation and minimax Bayes estimation, we have the following theorem.

Theorem (Donoho-Johnstone'98)

The thresholding estimators with proper thresholds depending only on resolution level j are near minimax over $\Theta^s_{p,q}(L)$ ($x = s, h$):

$$\inf_{\{t_j\}} \sup_{\theta \in \Theta^s_{p,q}(L)} \mathbb{E}_\theta \|\{\eta^x_{t_j}(y_{j,k})\} - \theta\|^2 \lesssim \inf_{\hat{\theta}} \sup_{\theta \in \Theta^s_{p,q}(L)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

Choice of the threshold

Compare to the VisuShrink:

- By using $t_j = \epsilon\sqrt{2 \ln n_\epsilon}$ in each resolution level, the resulting VisuShrink estimator is near optimal with a logarithmic gap
- Can remove the logarithmic gap by choosing a better threshold!

Theorem (Cai'12)

Choosing j_0, j_ϵ *with* $2^{j_0} \asymp \epsilon^{-\frac{2}{2s+1}}, 2^{j_\epsilon} \asymp \epsilon^{-2}$, *the thresholding estimator*

$$\hat{\theta}_{j,k}(y) = \begin{cases} y_{j,k} & j = j_0 \\ \eta^s_{t_j}(y_{j,k}) & j_0 < j \leq j_\epsilon \\ 0 & j > j_\epsilon \end{cases}$$

with $t_j = \epsilon\sqrt{2(j - j_0) \ln 2}$ *is near minimax over* $\Theta^s_{p,q}(L)$ *within constants.*

- Not adaptive (j_0 depends on s)!

SURE: Stein's unbiased risk estimator

Idea: estimate the risk of the thresholding estimator, and then choose a threshold to minimize the estimated risk

Definition

Stein's unbiased risk estimator In Gaussian mean estimation model $y_i = \theta_i + \epsilon z_i, 1 \le i \le d$, if $g(y) \triangleq \hat{\theta}(y) - y$ is weakly differentiable, then

$$\hat{r}(y) = (d + 2\nabla \cdot g(y))\epsilon^2 + \|g(y)\|^2$$

satisfies that $\mathbb{E}_\theta \hat{r}(y) = \mathbb{E}_\theta(\hat{\theta}(y) - \theta)^2$ for any $\theta \in \Theta$.

Soft thresholding: for each resolution level j, divide all $\{y_{j,k}, 0 \le k < 2^j\}$ randomly into two half samples I, I', and

$$t_I = \arg\min_{t \ge 0} \sum_{k \in I'} (1 - 2 \cdot \mathbb{1}(|y_{j,k}| \le t))\epsilon^2 + (|y_{j,k}| \wedge t)^2.$$

Use t_I as the thresholding for half sample I. $t_{I'}$ is obtained similarly.

Performance of SureShrink estimator

Stein's unbiased risk estimator for soft-thresholding:

- Bias: zero by definition
- Variance: small by measure concentration, for SURE can be expressed as a sum of independent random variables

Theorem (Performance of SureShrink estimator)

If $s > 1/p - 1/2$,

$$\sup_{\theta \in \Theta^s_{p,q}(L)} \mathbb{E}_\theta \|\hat{\theta}^{SURE} - \theta\|^2 \leq (1 + o(1)) \cdot \inf_{\{t_j\}} \sup_{\theta \in \Theta^s_{p,q}(L)} \mathbb{E}_\theta \|\{\eta^s_{t_j}(y_{j,k})\} - \theta\|^2$$

$$\lesssim \inf_{\hat{\theta}} \sup_{\theta \in \Theta^s_{p,q}(L)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

Practical implementation:

- do not split samples
- when $\sum_k (y^2_{j,k} - \epsilon^2)$ is small, use the usual $\epsilon\sqrt{2j \ln 2}$ threshold to sufficiently filter out the noise

Minimax L_2 risk over Besov balls

Theorem (Minimax L_2 risk)

The minimax L_2 risk of estimating function from Besov ball $\mathcal{B}^s_{p,q}(L)$, $s > 1/p$ (to ensure continuous embedding), is

$$\inf_{\hat{f}} \sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f \left\| \hat{f} - f \right\|^2 \asymp \epsilon^{\frac{4s}{2s+1}}$$

Minimax linear L_2 risk over Besov balls

Define $QHull(\Theta) = \{\eta : \eta^2 \in conv(\theta^2, \theta \in \Theta)\}$

Lemma (Donoho-Liu-MacGibbon'90)

$$\inf_{\hat{\theta}^{lin}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta}^{lin} - \theta\|^2 = \inf_{\hat{\theta}^{lin}} \sup_{\theta \in QHull(\Theta)} \mathbb{E}_\theta \|\hat{\theta}^{lin} - \theta\|^2$$

Furthermore, if $QHull(\Theta)$ is solid and orthosymmetric,

$$\inf_{\hat{\theta}^{lin}} \sup_{\theta \in QHull(\Theta)} \mathbb{E}_\theta \|\hat{\theta}^{lin} - \theta\|^2 \leq \frac{5}{4} \inf_{\hat{\theta}} \sup_{\theta \in QHull(\Theta)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2$$

Theorem (Minimax linear L_2 risk)

$$\inf_{\hat{f}^{lin}} \sup_{f \in \mathcal{B}^s_{p,q}(L)} \mathbb{E}_f \|\hat{f}^{lin} - f\|^2 \asymp \epsilon^{\frac{4(s-1/p+1/(p\vee 2))}{2(s-1/p+1/(p\vee 2))+1}}$$

- $QHull(\Theta^s_{p,q}) = \Theta^{s-1/p+1/(p\vee 2)}_{p\vee 2, q\vee 2}$
- Linear estimator is strictly suboptimal when $p < 2$

Estimation with L_r risk

Target: the normalized minimax risk over Besov balls

$$R_r^*(\mathcal{B}_{p,q}^s(L), \epsilon) = \left(\inf_{\hat{f}} \sup_{f \in \mathcal{B}_{p,q}^s(L)} \mathbb{E}_f \|\hat{f} - f\|_r^r \right)^{\frac{1}{r}}$$

for $1 \leq r < \infty$, and standard extension for $r = \infty$.

- For general $r \neq 2$, the estimation in function space $\mathcal{B}_{p,q}^s(L)$ is no longer equivalent to that in sequence space $\Theta_{p,q}^s(L)$
- Some phenomena never occur when $r = 2$: for Sobolev ball estimation, the phase transition point between dense regime and sparse regime is $r = \frac{p(2k+d)}{d} > 2$ due to $p > d$ (see previous lecture)

Minimax L_r risk

Throughout we assume $s > 1/p - 1/r$ and $r < \infty$.

Theorem (Minimax L_r risk)

$$R_r^*(\mathcal{B}_{p,q}^s(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+1}} & r < (2s+1)p \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{s}{2s+1}} (\ln(1/\epsilon))^{(\frac{1}{2} - \frac{p}{qr})_+} & r = (2s+1)p \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{s-1/p+1/r}{2(s-1/p)+1}} & r > (2s+1)p \end{cases}$$

Theorem (Minimax linear L_r risk)

$$R_r^{lin}(\mathcal{B}_{p,q}^s(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+1}} & r \leq p \\ (\epsilon^2)^{\frac{s-1/p+1/r}{2(s-1/p+1/r)+1}} & r > p \end{cases}$$

Three different zones

Homogeneous zone: $r \leq p$

- Optimal rate is $(\epsilon^2)^{\frac{s}{2s+1}}$
- Linear estimator attains the optimal rate

Intermediate zone: $p < r < (2s+1)p$

- Optimal rate is $(\epsilon^2)^{\frac{s}{2s+1}}$
- Linear estimator cannot attain the optimal rate

Sparse zone: $r \geq (2s+1)p$ (implies $r > 2$!)

- Optimal rate is worse that $(\epsilon^2)^{\frac{s}{2s+1}}$
- Linear estimator cannot attain the optimal rate

Wavelet thresholding estimate for $p \leq r$

The estimator

Denote by $\hat{\alpha}_{jk}, \hat{\beta}_{jk}$ empirical wavelet coefficients, consider the following estimator:

$$\tilde{\beta}_{jk} = \begin{cases} \eta_{t_j}^h(\hat{\beta}_{jk}) & j_0 \leq j \leq j_\epsilon \\ 0 & j > j_\epsilon \end{cases}$$

and $\hat{f} = \sum_k \hat{\alpha}_{j_0 k} \varphi_{j_0 k} + \sum_{j=j_0}^{\infty} \sum_k \tilde{\beta}_{jk} \psi_{jk}$.

The choice of parameters:

- Gross level j_0: variance not too large, i.e., $\epsilon 2^{j_0/2} \asymp R_r^*(\mathcal{B}_{p,q}^s(L), \epsilon)$
- Detailed level j_ϵ: bias not too large, by $\mathcal{B}_{p,q}^s \subset \mathcal{B}_{r,q}^{s-1/p+1/r}$ we have $2^{-j_\epsilon(s-1/p+1/r)} \asymp R_r^*(\mathcal{B}_{p,q}^s(L), \epsilon)$
- Threshold: $t_j = K\epsilon\sqrt{j - j_0}$ for some constant K

Theorem (Donoho-Johnstone-Kerkyacharian-Picard'96)

This estimator is minimax order-optimal!

Adaptive estimation

Suppose that the modulus of smoothness satisfies that $s \leq s_{\max}$

- Gross level j_0: $2^{j_0} \asymp \epsilon^{\frac{2}{2s_{\max}+1}}$
- Detailed level j_ϵ: $2^{j_\epsilon} \asymp n/\ln n$
- Threshold: still use $t_j = \epsilon\sqrt{j - j_0}$

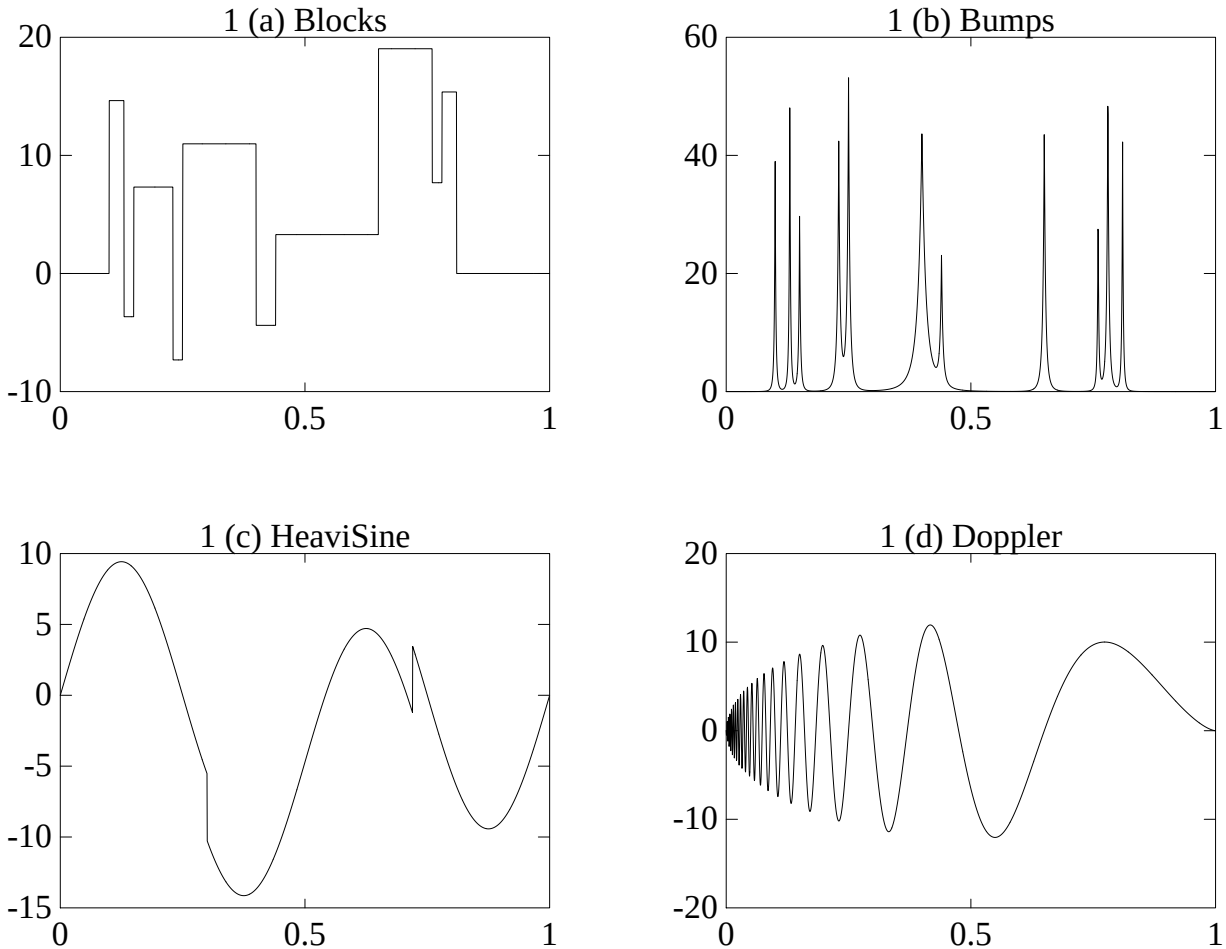
Theorem (Adaptation results)

The adaptive estimator achieves the minimax risk order when $r \geq (2s+1)p$, *and possesses a logarithmic gap* $(\ln(1/\epsilon))^{\frac{s}{2s+1}}$ *when* $r < (2s+1)p$.

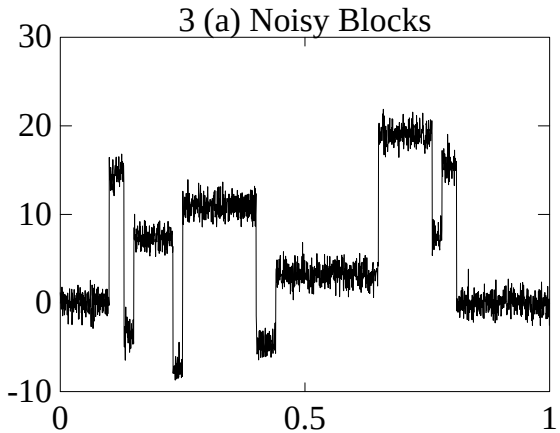
Extension: apply Lepski's trick to choose the threshold t_j!

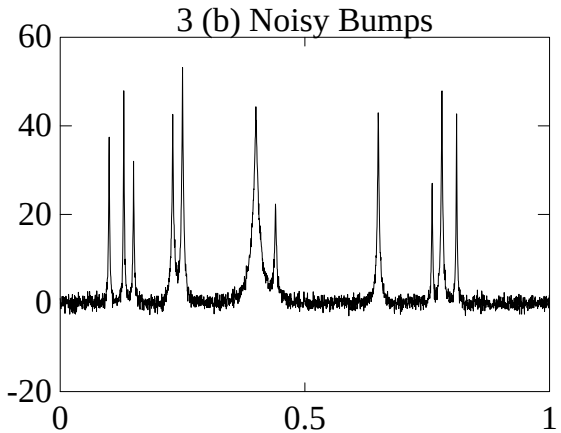
- Fully adaptive without logarithmic gap! (Juditsky'97)

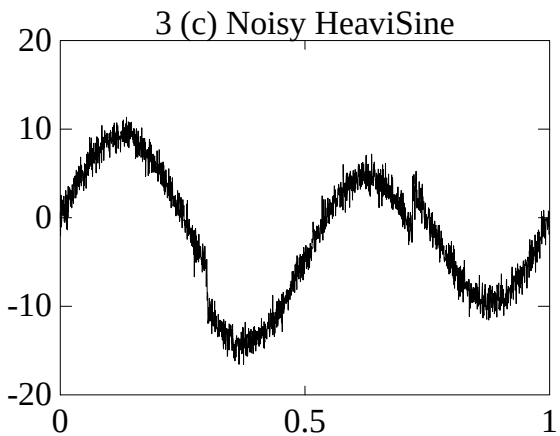
Original signal

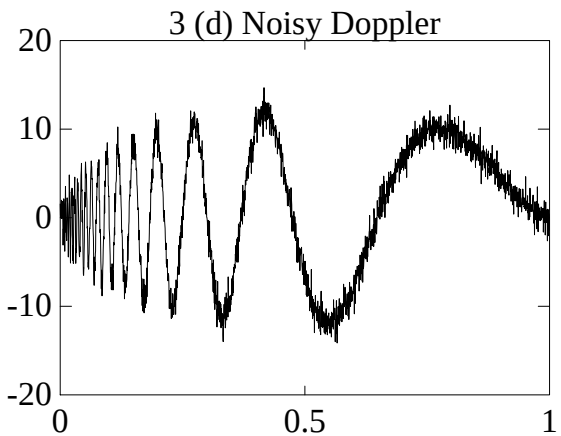


Noisy signal

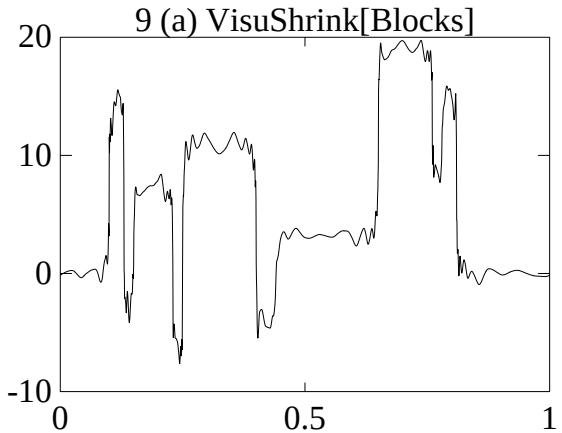


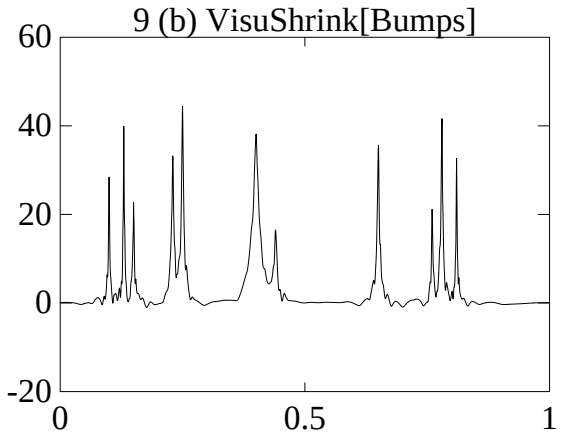


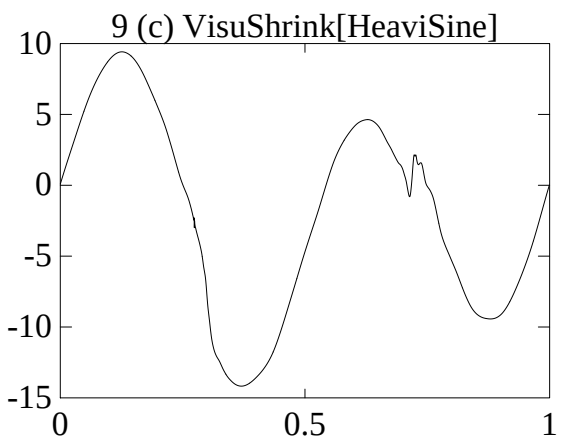


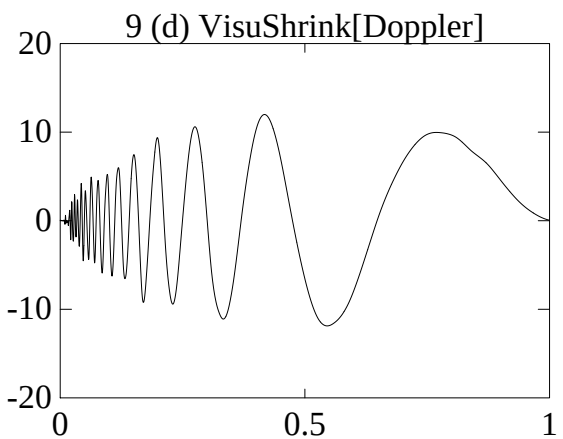


Reconstructed signal by VisuShrink









Estimation over special function spaces

Hölder ball: $\mathcal{H}_1^s(L) = \mathcal{B}_{\infty,\infty}^s(L)$

Estimation over Hölder ball

$$R_r^*(\mathcal{H}_1^s(L), \epsilon) \asymp (\epsilon^2)^{\frac{s}{2s+1}}$$

Sobolev ball: $\mathcal{B}_{p,1}^k \subset \mathcal{S}_1^{k,p} \subset \mathcal{B}_{p,\infty}^k$

Estimation over Sobolev ball

$$R_r^*(\mathcal{S}_1^{k,p}(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{k}{2k+1}} & r < (2k+1)p \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{k-1/p+1/r}{2(k-1/p)+1}} & r > (2k+1)p \end{cases}$$

Functions with bounded variation: $\mathcal{B}_{1,1}^1 \subset BV \subset \mathcal{B}_{1,\infty}^1$

Estimation of functions with bounded variation

$$R_r^*(BV(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{1}{3}} & r < 3 \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{1}{r}} & r > 3 \end{cases}$$

Other settings

Non-equidistant grid in fixed design:

- Suppose $x_i = H^{-1}(i/n)$ in regression model (or transform the noise to Standard Brownian motion via time change in the Gaussian white noise model)
- Cai'98: if $H(\cdot)$ is Lipschitz, can estimate $f \circ H(\cdot)$ and then recover f
- Unknown H can be estimated by interpolation

Non-Gaussian noise:

- Suffice to impose tail conditions for the noise in the sequential model
- Juditsky'97: for any $\lambda \in [\epsilon, c_1\epsilon\sqrt{\ln(1/\epsilon)}]$, it holds

$$\mathbb{E}[|z|^p \mathbb{1}(|z| \geq \frac{\lambda}{2})] \leq c_2 \lambda^p \exp(-\frac{\lambda^2}{c_3 \epsilon^2})$$

Unknown noise level:

- Donoho-Johnstone'94: estimate the noise level using the median of the empirical wavelet coefficient at the finest level

High-dimensional results

Consider the isotropic Besov ball $\mathcal{B}^s_{p,q,d}(L)$ in d dimensional space. Assume $s > d/p - d/r$ and $r < \infty$.

Theorem (Minimax L_r risk)

$$R_r^*(\mathcal{B}^s_{p,q,d}(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+d}} & r < (1+2s/d)p \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{s}{2s+d}} (\ln(1/\epsilon))^{(\frac{1}{2} - \frac{p}{qr})_+} & r = (1+2s/d)p \\ (\epsilon^2 \ln(1/\epsilon))^{\frac{s-d/p+d/r}{2(s-d/p)+d}} & r > (1+2s/d)p \end{cases}$$

- Minimax risk achieved by product wavelet basis

Theorem (Minimax linear L_r risk)

$$R_r^{lin}(\mathcal{B}^s_{p,q,d}(L), \epsilon) \asymp \begin{cases} (\epsilon^2)^{\frac{s}{2s+d}} & r \le p \\ (\epsilon^2)^{\frac{s-d/p+d/r}{2(s-d/p+d/r)+d}} & r > p \end{cases}$$

Properties of wavelet thresholding estimator

Properties of wavelet thresholding estimator $\hat{f}_n$:

- As smooth as the truth: unconditional basis property yields

$$\lim_{n\to\infty} \mathbb{P}_f\{\|\hat{f}_n\|_{\mathcal{B}^s_{p,q}} \leq C\|f\|_{\mathcal{B}^s_{p,q}}\} = 1$$

  for some constant C.
- Near optimal for spatial adaptation
- Near optimal for estimating the function at a point
- Near optimal for estimating the function under global loss

Optimality of unconditional basis

Three norms:

- Asymptotics of compression: $\|\theta\|_{c,m} = \sup_n n^m \sum_{k>n} |\theta_{(k)}|^2$
- Asymptotics of estimation: $\|\theta\|_{e,r} = \sqrt{\sup_\delta \delta^{-2r} \sum_k \min\{\delta^2, \theta_k^2\}}$
- Weak ℓ^p ball: $\|\theta\|_{w\ell^p} = \sup_k k^{1/p} |\theta_{(k)}|$

Lemma

Three norms are equivalent when $p = \frac{2}{2m+1}, r = \frac{2m}{2m+1}$.

Critical exponent of Θ: $p^*(\Theta) = \inf\{p : \|\theta\|_{w\ell^p} < \infty, \forall \theta \in \Theta\}$

Theorem (Optimality of unconditional basis)

If Θ is ℓ^2 bounded, solid and orthosymmetric, then for any orthogonal transformation $U : \ell^2 \to \ell^2$, we have $p^*(U\Theta) \geq p^*(\Theta)$.

- Example: $p^*(\Theta_{BV}) = 2/3 < 1 = p^*(U_{WF}\Theta_{BV})$, where U_{WF} transforms wavelet basis to Fourier basis.

Wavelet transform on bounded interval

The Daubechies' father wavelet $\varphi(\cdot)$ has support $[0, 2N-1]$, what if we are only interested in $[0,1]$?

- Boundary adjustment: for each resolution level j, set

$$\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$$

  for $0 \le k < 2^j - 2N$ as usual, and suitably add $2N$ wavelets to ensure $\text{span}(\varphi_{jk}, 0 \le k < 2^j)$ contain polynomials of degree no more than $N-1$. Same applies to mother wavelets.
- Discrete wavelet transform: implemented by a sequence of finite filtering steps instead of matrix multiplication, with complexity $\mathcal{O}(n)$

Local property of wavelets

If $f \in \mathcal{H}_1^s$ and $\psi(\cdot)$ has $\lfloor s \rfloor$ vanishing moments, we have

$$\left| \int f\psi_{jk} - 2^{-j/2} f(k/2^j) \right| \lesssim 2^{-j(s+1/2)}$$

Multiresolution analysis

Asymptotic equivalence of models under mild conditions:

- Gaussian white noise model:

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}}dB_t, \qquad t \in [0,1]$$

- Density estimation model: generate n iid samples from common density g with support $[0,1]$ ($g = f^2, \sigma = 1/2$)

Proof depends heavily on multiresolution analysis (see Brown et al'04)!

General modulus of smoothness

Definition (Ditzian-Totik modulus of smoothness)

$$\omega_\varphi^r(f,t)_p = \sup_{0<h\le t} \|\Delta_{h\varphi(x)}^r f\|_p$$

Trigonometric approximation on $[0,1]$ ($\varphi \equiv 1$):

- Denote by $E_n^T(f)_p$ the best approximation error in L_p norm using trigonometric series of degree no more than n
- Direct inequality: $E_n^T(f)_p \le C_{r,1}\omega^r(f,n^{-1})_p$
- Converse: $\omega^r(f,n^{-1})_p \le C_{r,2}n^{-r}\sum_{k=1}^n k^{r-1}E_k^T(f)_p$

Polynomial approximation on $[0,1]$ ($\varphi = \sqrt{x(1-x)}$):

- Denote by $E_n(f)_p$ the best approximation error in L_p norm using algebraic polynomials of degree no more than n
- Direct inequality: $E_n(f)_p \le D_{r,1}\omega_\varphi^r(f,n^{-1})_p$
- Converse: $\omega_\varphi^r(f,n^{-1})_p \le D_{r,2}n^{-r}\sum_{k=1}^n k^{r-1}E_k(f)_p$

General K-functionals

Definition (General K-functional)

The general K-functional $K_{r,\varphi}(f,t)_p$ is defined as

$$K_{r,\varphi}(f,t)_p = \inf_{g^{(r-1)} \in \mathrm{A.C.}_{\mathrm{loc}}} \|f-g\|_p + t\|\varphi^r g^{(r)}\|_p$$

Theorem (Equivalence of K-functional and modulus of smoothness)

Under mild conditions on φ, we have

$$\omega^r_\varphi(f,t)_p \asymp K_{r,\varphi}(f,t^r)_p$$

Bias analysis of plug-in estimator: for $n\hat{p} \sim \mathrm{B}(n,p)$ and any f,

$$\begin{aligned}|\mathbb{E}_p f(\hat{p}) - f(p)| &\le \inf_{g\in C^2} |\mathbb{E}_p g(\hat{p}) - g(p)| + 2\|f-g\|_\infty \\ &\lesssim \inf_{g\in C^2} n^{-1}\|p(1-p)g''(p)\|_\infty + \|f-g\|_\infty \lesssim \omega^2_\varphi(f, n^{-1/2})_\infty\end{aligned}$$

where $\varphi(x) = \sqrt{x(1-x)}$.

References

Nonparametric regression:

- A series of paper by Donoho, Johnstone, etc. in 1990s.
- I. Johnstone. *Gaussian estimation: sequence and wavelet models.* Manuscript (available online at `http://statweb.stanford.edu/~imj/GE06-11-13.pdf`), 2013.

Approximation theory:

- R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer Science & Business Media, 1993.
- Z. Ditzian and V. Totik, *Moduli of smoothness*. Vol. 9. Springer Science & Business Media, 2012.

Wavelet and Besov space:

- W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications.* Vol. 129. Springer Science & Business Media, 2012.