Interactions Between Modern Harmonic Analysis and Statistical Estimation

Emmanuel Candès, California Institute of Technology

IDR Meeting, Singapore, August 2004

Outline of Lectures

- Lecture I. Classical estimation
- Lecture II. Nonlinear estimation
- Lecture III. Efficient representations and efficient estimation
- Lecture IV. Estimation in ill-posed linear inverse problems

Thanks...

To Carl de Boor for making corrections and suggesting changes

Lecture 2: Nonlinear estimation

- Stein's phenomenon
- Thresholding rules
- Oracle inequalities
- Wavelet shrinkage
- Statistical theory

Linear Estimation

Well-suited

- for estimating/compressing Gaussian processes;
- for estimating function classes that are not ellipsoids.

Limitations

- Many stochastic models of interest are non-Gaussian.
- Many function classes of interest are not ellipsoids.

The Ramp Process

Example of a non-Gaussian process

- $T = [0, 1)$
- Draw τ uniform over $[0, 1)$
- Define path by

$$X(t, \tau) = t - 1_{\{t \geq \tau\}}$$

Very simple process and very easy to code accurately; e.g.

- Extract τ by locating the jump
- Quantization to the required fidelity

The Ramp Process and PCA

Covariance $\Sigma(s, t)$ of the ramp process:

$$\Sigma(s, t) = \min(s, t) - st.$$

Same as that of the Brownian bridge (Brownian motion for W(0) = W(1) = 0)!

Ignore non-Gaussianity and apply the same coder as for the Brownian bridge.

- Coefficients z_i in the KL basis are random
- Typical size in mean square $Ez_i^2 = (4\pi i^2)^{-1}$
- To achieve a distortion D, we must code about $1/D$ coefficients
- Many coefficients needed to represent a typical realization of Ramp.

The Ramp Process and Wavelets

Yves Meyer observed that the wavelet coefficients of Ramp decay very rapidly, essentially exponentially.

Scalar quantization of wavelet coefficients can capture realization of Ramp with 1% accuracy using a few dozens rather than tens of thousands of coefficients.

By abandoning the KL transform, in this case, we get a transform in which scalar quantization works very well.

Other Functional Classes

- $T = [0, 1)$
- $\|f\|^2_{W_p^m} = \int_0^1 |f(t)|^p + |f^{(m)}(t)|^p \, dt.$
- $\mathcal{F} = \{f : \|f\|_{W_p^m} \leq C\}.$

Data

$$Y(dt) = f(t)dt + \epsilon W(dt), \quad t \in [0, 1)$$

Mean squared error

$$MSE(\hat{f}, f) = E \int_{[0,1)} (\hat{f}(t) - f(t))^2 dt$$

Asymptotically Minimax Solution (Small MSE)

$$R_\epsilon(\hat{f}, \mathcal{F}) \sim \inf_{\hat{f}} \sup_{\mathcal{F}} R(\hat{f}, f)$$

- $p = 2$. The solution is a Linear estimator with simple expression in the Fourier Domain.
- $p \neq 2$. The solution must be a Nonlinear estimator

The James Stein-Estimator

- Problem: estimate $\theta \in \mathbf{R}^d$ from $y \sim N(\theta, I)$
- Mean-squared error

$$MSE(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|^2$$

- Maximum-likelihood estimate $\hat{\theta}^{MLE} = y$

$$MSE(\hat{\theta}^{MLE}, \theta) = d$$

- James-Stein estimate

$$\hat{\theta}^{JS} = w(y)y, \quad w(y) = 1 - (d-2)/\|y\|^2$$

- Surprising phenomenon

$$MSE(\hat{\theta}^{JS}, \theta) < MSE(\hat{\theta}^{MLE}, \theta), \quad \forall \theta \in \mathbf{R}^d$$

- Shrinking improves performance!

Ideal Linear Shrinkage Estimator and Oracle Inequalities

- Problem: estimate $\theta \in \mathbf{R}^d$ from $y \sim N(\theta, I)$
- Family of diagonal estimators

$$\hat{\theta}^c = cy$$

- Mean squared error

$$MSE(\hat{\theta}^c, \theta) = (1-c)^2\|\theta\|^2 + c^2 d$$

  Why? Assume, $y_i = \theta_i + z_i$

$$\begin{aligned} E(\hat{\theta}_i - \theta_i)^2 &= \text{bias}^2 + \text{Variance} \\ &= [(1-c)\theta_i]^2 + c^2 \end{aligned}$$

- Ideal shrinkage

$$c^* = \text{argmin}(1-c)^2\|\theta\|^2 + c^2 d = \frac{\|\theta\|^2}{\|\theta\|^2 + d}$$

- Oracle inequality (Donoho and Johnstone)

$$MSE(\hat{\theta}^{JS}, \theta) \leq 2 + \inf MSE(\hat{\theta}^c, \theta)$$

Adaptive Estimation

Previous lecture

$$Y(dt) = f(t)dt + \epsilon W(dt) \quad t \in [0,1)$$

and

$$f \in \mathcal{F}^m(C) = \{f : \|f\|_{W_2^m} \leq C\}$$

- Pinsker's solution requires knowledge of m and C
- In practice, m and C are unknown
- How to achieve asymptotic minimaxity over $\mathcal{F}^m(C)$, simultaneously for each value of m and $C > 0$?

Adaptation by Oracle Inequalities

Sequence model

$$y_k = \theta_k + \epsilon z_k$$

and Sobolev ellipsoids

$$\Theta(C) =: \{\theta : \sum_j \sum_{k \in B_j} |\theta_k|^2 \leq C\}$$

B_j dyadic subbands

- Take data in the frequency domain
- Apply the James-Stein estimator to the blocks B_j

$$\hat{\theta}_j^{BJS}(y) = \begin{cases} y_j & j < J_0 \\ \left(1 - \frac{(n_j - 2)\epsilon^2}{\|y_j\|^2}\right)_+ y_j & J_0 \leq j < J_\epsilon \\ 0 & j \geq J_\epsilon \end{cases}$$

- Return to the original domain

- Key result (Efromovich, 1986)

$$\sup_{\theta \in \Theta} MSE(\hat{\theta}^{BJS}, \theta) \sim \inf_{\hat{\theta}} \sup_{\theta \in \Theta} MSE(\hat{\theta}, \theta)$$

Ideal Shrinkage

$$y_i = \theta_i + \sigma z_i$$

- Family of shrinkage estimators $\hat{\theta}_i = w y_i$
- Ideal estimator $\inf_w MSE(\hat{\theta}_i, \theta)$
- Solution

$$\hat{\theta}_i^* = w_i \theta_i, \quad w_i^* = \frac{\theta_i^2}{\theta_i^2 + \sigma^2}, \quad E(\hat{\theta}_i^* - \theta) = \frac{\theta_i^2 \sigma^2}{\theta_i^2 + \sigma^2}$$

- Ideal projection $w_i \in \{0, 1\}$

$$\hat{\theta}_i^I = w_i \theta_i, \quad w_i = \begin{cases} 0 & |\theta_i| < \sigma \\ 1 & |\theta_i| \geq \sigma \end{cases}$$

- Risk of ideal projection

$$E(\hat{\theta}_i^I - \theta) = \min(\theta_i^2, \sigma^2)$$

- Comparison

$$MSE(\hat{\theta}^I, \theta) \leq 2MSE(\hat{\theta}^*, \theta)$$

- Conclusion: good compressibility $\asymp$ small ideal risk

Ideal Risk

$$y_i = \theta_i + z_i$$

For each coordinate, "omniscient" estimator would choose either $\hat{\theta}_i = 0$ or $\hat{\theta}_i = y_i$ so as to minimize the MSE.

Risk (Mean-squared error)

$$MSE(\hat{\theta}^I, \theta) = \sum_i \min(\theta_i^2, \sigma^2)$$

Interpretation

- Rearrange the coefficient in decreasing order $|\theta|^2_{(1)} \geq |\theta|^2_{(2)} \geq \ldots \geq |\theta|^2_{(n)}$
- $N(\sigma)$: number of those θ_i's s.t. $\theta_i^2 \geq \sigma^2$

$$\begin{aligned} MSE(\hat{\theta}^I, \theta) &= N\sigma^2 + \sum_{i>N} |\theta|^2_{(i)} \\ &= N\sigma^2 + \|\theta - \theta_N\|^2 \\ &= \text{Number ot terms} \times \text{noise level} + \text{ Approx Error} \end{aligned}$$

Thresholding Rules

- Hard-thresholding

$$\hat{\theta}_i = \begin{cases} y_i & |y_i| \geq \lambda \\ 0 & |y_i| < \lambda \end{cases}$$

- Soft-thresholding

$$\hat{\theta}_i = \begin{cases} y_i - \lambda & y_i \geq \lambda \\ 0 & |y_i| < \lambda \\ -y_i + \lambda & y_i < -\lambda \end{cases}$$

- Many others

Thresholding Rules and Complexity Penalized Estimation

- Hard-thresholding

$$\hat{\theta}_H(y) = \text{argmin}(y-\theta)^2 + \lambda^2 1_{\{\theta \neq 0\}}$$

- Soft-thresholding

$$\hat{\theta}_S(y) = \text{argmin}(y-\theta)^2 + 2\lambda|\theta|$$

- For n-dimensional data

$$\hat{\theta}_H(y) = \text{argmin} \sum_i (y_i-\theta_i)^2 + \lambda^2 \#\{\theta_i \neq 0\}$$

and

$$\hat{\theta}_S(y) = \text{argmin} \sum_i (y-\theta)^2 + 2\lambda \sum_i |\theta_i|$$

Oracle Inequalities

Donoho and Johnstone

Thresholding at $\lambda = \sigma\sqrt{2\log n}$

$$MSE(\hat{\theta},\theta) \leq (2\log n + 1)\cdot(\sigma^2 + \sum_i \min(\theta_i^2, \sigma^2))$$

- Thresholding comes close to the ideal risk.
- Ideal risk is a proxy

Sparsity implies good compressibility which in turn implies good estimation.

Toy Example, I

Basic statistical model

$$y_i = \theta_i + z_i, \quad z_i \text{ i.i.d. } N(0,1), \quad i = 1,2,\ldots,n.$$

- Suppose all the coefficients are zero except for two spikes, each of size $\sqrt{n}/2$.
- Signal to noise ratio is 1/2.

Two naive estimators

(i) $\hat{\theta}_i = y_i$, $MSE = \sum_i \sigma_n^2 = n$,

(ii) $\hat{\theta}_i = 0$, $MSE = \sum_i \theta_i^2 = n/2$.

Toy Example, II

Hard thresholding rule at $\sqrt{2\log n}$

- Data from the two spikes pass the threshold unchanged, so are essentially unchanged; estimator of type (i).
- In all other coordinates, the estimator sets all data to zero except for the small fraction of noise that exceeds the threshold.

Mean-squared error

$$MSE(\hat{\theta},\theta) \asymp 2 + n\,E(Z^2, Z^2 \geq 2\log n) \asymp 2 + \sqrt{2\log n}$$

Much better!

Consequences

- Identify problems of scientific interest
- Find efficient representations (bases) for those classes
- Rotate data into those bases
- Apply thresholding
- Invert

Significant interaction with the agenda of modern harmonic analysis

Quantifying the Performance

- Quality of estimation is linked to the sparsity (decay) of the coordinates θ_i
- One measure of the sparsity if the weak ℓ_p norm (Marcinkiewicz)

$$w\ell_p(C) = \{\theta : |\theta|_{(k)} \leq Ck^{-1/p}, \forall k\}$$

- Note

$$\ell_p(C) \subset w\ell_p(C), \qquad \ell_p(C) := \{\theta, \sum_i |\theta_i|^p \leq C^p\}$$

Ideal Risk and Weak ℓ_p balls

Assume that $\Theta \subset w\ell_p(C)$, then

$$\sum_i \min(\theta_i^2, \epsilon^2) = O((\epsilon^2)^{\frac{2\alpha}{2\alpha+1}}), \quad 1/p =: \alpha + 1/2$$

Lower Bounds

Suppose

$$\ell_{p,+}(C) \subset \Theta$$

meaning that Θ contains n-dimensional hyperrectangles of the form $[0, Cn^{-1/p}]^n$ for arbitrary large n.

Then

$$\begin{aligned} \inf_{\hat{f}} \sup_{\Theta} E\|\hat{\theta} - \theta\|^2 &\geq \inf_{\hat{f}} \sup_{\ell_{p,+}(C)} E\|\hat{\theta} - \theta\|^2 \\ &\geq c \cdot (\epsilon^2)^{\frac{2\alpha}{2\alpha+1}} \end{aligned}$$

Lower Bounds and Bayes Risk

- Prior π on Θ (normed measure)

$$\int_{\Theta} \pi(d\theta) = 1$$

- Bayes risk (averaged risk)

$$B(\pi) := \inf_{\hat{\theta}} \int E\|\hat{\theta} - \theta\|^2 \, \pi(d\theta)$$

attained for

$$\hat{\theta}_\pi = E(\theta|y)$$

Key result of statistical decision theory

$$\inf_{\hat{\theta}} \sup_{\Theta} E\|\hat{\theta} - \theta\|^2 \geq B(\pi), \qquad \forall \pi$$

Significance of Orthosymmetry

- Hyperrectangle

$$R := \prod_i [0, \tau_i] \subset \Theta$$

and

$$\inf_{\hat{f}} \sup_{\Theta} E\|\hat{\theta} - \theta\|^2 \geq \inf_{\hat{f}} \sup_{R} E\|\hat{\theta} - \theta\|^2$$

- Prior π supported on vertices of rectangle R

$$\theta_i \begin{cases} 0 & \text{w.p. } 1/2 \\ \tau_i & \text{w.p. } 1/2 \end{cases}, \quad \tau_i's \text{ independent}$$

- Coordinates are independent
  - Any given coordinate does not give any information about any other
  - Good procedures treat each coordinate individually
- Bayes' estimator

$$\hat{\theta}_{\pi,i} = E(\theta_i|y_i)$$

Unconditional bases

- Orthonormal basis $(\phi_i)_i$
- Function space norm $\| \ \|_{\mathbf{F}}$
- Sequence space norm $\| \ \|_{\mathbf{f}}$
- Coefficients $\theta_i(f) = \langle f, \phi_i \rangle$

$$\|f\|_{\mathbf{F}} \sim \|\theta(f)\|_{\mathbf{f}}$$

- Sequence space norm has

$$\|(\pm_i \theta_i)_i\|_{\mathbf{f}} = \|\theta\|_{\mathbf{f}}$$

for all choices of signs.

Optimality of Unconditional Bases

Define

$$\Theta = \{\theta(f) : \ f \in \mathcal{F}\}$$

and critical exponent

$$p^*(\Theta) = \inf\{p : \ \Theta \subset w\ell_p\}$$

Then for any orthogonal transform U

$$p^*(U\Theta) \geq p^*(\Theta)$$

Interpretation:

- Among all orthobases, the unconditional basis provides the sparsest coefficient sequence.
- Optimality for nonlinear approximation
- Optimality for diagonal estimation

Examples

- Sinusoids, L_2 Sobolev
- Wavelets, L_p Sobolev
- Wavelets, Besov, Triebel Spaces

Estimation over Besov/Triebel Bodies

$\mathcal{F} = \{f : \|f\|_{B^{\sigma}_{p,q}} \leq C\}$

Data

$$Y(dt) = f(t)dt + \epsilon W(dt), \quad t \in T$$

Seek minimax estimator $\hat{f}$ so that

$$\sup_{\mathcal{F}} E\|\hat{f} - f\|_2^2 = \min$$

Answer (asymptotic, $p \leq 2$)

$$\hat{f} = \sum_I \eta_I(\langle y, \psi_I \rangle_n)\psi_I$$

$\eta_I(\cdot; \alpha, p, q, C)$ is a scalar nonlinearity (hard, soft thresholding)

Interpretation:

1. Go into wavelet basis
2. Throw out small coefficients

Example

Class $\mathcal{F} = \{f : TV(f) \leq C\}$.

With $\mathcal{F}$ as above, then

$$M^*(\epsilon, \mathcal{F}) := \inf_{\hat{f}} \sup_{\mathcal{F}} MSE(f, \hat{f}) \asymp \epsilon$$

- Get truncated noisy wavelet series

$$Y = \sum_{0 \leq j \leq J} \sum_{k=0}^{2^j - 1} y_{j,k}\psi_{j,k}$$

- Apply thresholding

$$\hat{\theta}_{j,k} = \eta(y_{j,k})$$

- Invert wavelet transform

$$\hat{f} = \sum_{0 \leq j \leq J} \sum_{k=0}^{2^j - 1} \hat{\theta}_{j,k}\psi_{j,k}$$

$$MSE(f, \hat{f}) = O(\log \epsilon^{-1})\, M^*(\epsilon, \mathcal{F})$$

Adaptive Minimaxity

Donoho, Johnstone, Kerkyacharian, Picard

- α, p, q and C are unknown
- How to nearly achieve the asymptotic minimax risk for each value of $\alpha \in [\alpha_0, \alpha_1]$, p, q and $C > 0$.

Adaptation by Oracle Inequality

- Take data into wavelet domain
- Ignore small scales
- Apply hard thresholding, $\lambda = \sqrt{2 \log n} \cdot \epsilon$
- Return to original domain

Summary

- Efficient representations lead to efficient estimations
- Certain representations emerge as optimal
- Same representation “solves” many estimation problems (adaptivity)
- Challenge: find optimal representations for models of scientific interest.
- For those models, unconditional bases are unlikely...