

Ordered processes and oracle inequalities related to high dimensional linear models

Yuri Golubev

CNRS, Université Aix-Marseille 1

Rennes, 18–20 June 2007

Outline of the talk

- 1 Statistical problem
- 2 Ordered processes: definition, properties, examples
- 3 Linear models and spectral regularization
 - Basic spectral regularization methods
 - Principle of the Empirical Risk Minimization
 - Excess risk and oracle inequalities
- 4 Oracle inequalities
 - Main theorem
 - Unbiased Risk Estimation
- 5 A numerical example

The talk deals with recovering $\theta = (\theta(1), \dots, \theta(n))^T \in \mathbb{R}^n$ from the noisy data

$$Y = A\theta + \epsilon,$$

where

- A is a $m \times n$ - matrix with $m \geq n$
- $\epsilon \in \mathbb{R}^m$ is a white Gaussian noise with a known variance

$$\mathbf{E}\epsilon(k)\epsilon(l) = \sigma^2\delta_{kl}, \quad k, l = 1, \dots, m$$

- n may be very large (infinity).

Example: our linear model can be viewed as a discrete approximation of the integral equation

$$y(u) = \int A(u, v)\theta(v) dv.$$

Ordered processes: an example

Let $W(t)$, $t \geq 0$ be a standard Wiener process. It is well known that for any $\mu > 0$

$$\mathbf{E} \max_{t \geq 0} \left[W(t) - \frac{\mu}{2} \mathbf{E} W^2(t) \right] = \frac{1}{\mu}.$$

Consider $\xi_0(t) = \xi \times t$, $t \geq 0$, where ξ is $\mathcal{N}(0, 1)$. Then

$$\mathbf{E} \max_{t \geq 0} \left[\xi_0(t) - \frac{\mu}{2} \mathbf{E} \xi_0^2(t) \right] = \frac{1}{2\mu}.$$

What is $\xi(t)$ satisfying

$$\mathbf{E} \max_{t \geq 0} \left[\xi(t) - \frac{\mu}{2} \mathbf{E} \xi^2(t) \right] \leq \frac{C}{\mu} \quad \text{for some } C > 0?$$

Definition

A separable process $\xi(t)$, $t \geq 0$ with $\mathbf{E}\xi(t) = 0$ is called **ordered** if

$$\mathbf{E}\xi(t_1)\xi(t_2) \geq \min\{\mathbf{E}\xi^2(t_1), \mathbf{E}\xi^2(t_2)\}$$

Some examples:

- fractional Wiener processes $W_H(t)$ with

$$\mathbf{E}W_H(t_1)W_H(t_2) = \frac{1}{2} \left[t_1^{2H} + t_2^{2H} - |t_1 - t_2|^{2H} \right]$$

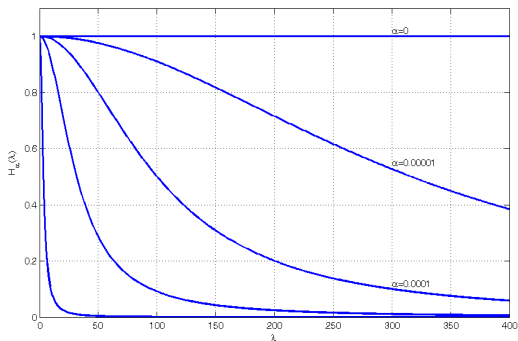
are ordered if $H \geq 1/2$.

- ordered processes related to the so-called ordered smoothers / Kneip (1994)/.

Ordered Smoother

The family of functions $H_\alpha(\lambda)$, $\alpha, \lambda \in \mathbb{R}^+$ is called *ordered* if

- $0 \leq H_\alpha(\lambda) \leq 1$
- for all $\lambda \in \mathbb{R}^+$ $H_{\alpha_1}(\lambda) \geq H_{\alpha_2}(\lambda)$, $\alpha_1 \leq \alpha_2$.



Basic processes

Let $\xi(k)$ be i.i.d. $\mathcal{N}(0, 1)$ and $\lambda_1 \leq \lambda_2 \leq \dots$. Then

$$B(t) = \sum_{k=1}^{\infty} [1 - H_t(\lambda_k)] \xi(k) \theta(k),$$
$$V(t) = \sum_{k=1}^{\infty} \lambda_k H_{1/t}^2(\lambda_k) [\xi^2(k) - 1]$$

are ordered processes when $H_t(\lambda)$ is a family of ordered functions.

Dichotomy Inequality

Denote for brevity $\sigma^2(t) = \mathbf{E}\xi^2(t)$, $\Delta_\xi(t_1, t_2) = \xi(t_1) - \xi(t_2)$.

Theorem

Let $\xi(u)$, $u \in [0, t]$ be an ordered process. Then for any $\lambda > 0$

$$\log \mathbf{E} \exp \left\{ \lambda \sup_{0 \leq u \leq t} \frac{\Delta_\xi(u, t)}{\sigma(t)} \right\} \leq \frac{\log(2)\sqrt{2}}{\sqrt{2}-1} +$$

$$+ \sup_{0 \leq u \leq v \leq t} \sup_{0 \leq z \leq 1/(\sqrt{2}-1)} \log \mathbf{E} \exp \left\{ z \lambda \frac{\Delta_\xi(u, v)}{[\mathbf{E}\Delta_\xi^2(u, v)]^{1/2}} \right\}.$$

Proof is based on the standard chaining argument.

Basic Properties

Theorem

Let $\xi(t)$ be an ordered process with $\xi(0) = 0$. Assume that for some $\lambda > 0$

$$\sup_{u,v} \log \mathbf{E} \exp \left\{ \lambda \frac{\Delta_{\xi}(u, v)}{[\mathbf{E} \Delta_{\xi}^2(u, v)]^{1/2}} \right\} < \infty.$$

Then there exists a constant $C(p, q)$ depending on λ such that for all $\mu > 0$

$$\mathbf{E} \sup_{t \geq 0} [\xi(t) - \mu^{q-1} \sigma^q(t)]_+^p \leq \frac{C(p, q)}{\mu^p},$$

where $[x]_+ = \max(0, x)$.

Basic Properties

Let τ be a random variable, then $\mathbf{E}\xi(\tau) \leq C\sqrt{\mathbf{E}\sigma^2(\tau)}$.

Indeed,

$$\begin{aligned} \mathbf{E}\xi(\tau) &= \inf_{\mu} \left\{ \mathbf{E}\xi(\tau) - \mu\mathbf{E}\sigma^2(\tau) + \mu\mathbf{E}\sigma^2(\tau) \right\} \\ &\leq \inf_{\mu} \left\{ \mathbf{E} \max_{t>0} [\xi(t) - \mu\sigma^2(t)] + \mu\mathbf{E}\sigma^2(\tau) \right\} \\ &\leq \inf_{\mu} \left\{ \frac{C}{\mu} + \mu\mathbf{E}\sigma^2(\tau) \right\} = C\sqrt{\mathbf{E}\sigma^2(\tau)}. \end{aligned}$$

We can use the following approximation for $\xi(\cdot)$

$$\xi(t) \approx C\xi \cdot \sigma(t), \text{ where } \xi \sim \mathcal{N}(0, 1).$$

Estimation in a linear model

Suppose we observe $Y \in \mathbb{R}^m$

$$Y = A\theta + \epsilon$$

and our goal is to estimate $\theta \in \mathbb{R}^n$.

The standard ML estimator is defined as follows

$$\hat{\theta}_0 = \arg \min_{\theta \in \mathbb{R}^n} \|Y - A\theta\|^2, \quad \text{where } \|x\|^2 = \sum_{k=1}^m x^2(k).$$

With a simple algebra we obtain

$$\hat{\theta}_0 = (A^\top A)^{-1} A^\top Y$$

/Moore (1920), Penrose (1955)/

Risk of the MP inversion

The risk of this inversion is computed by

$$\mathbf{E}\|\hat{\theta}_0 - \theta\|^2 = \mathbf{E}\|(A^\top A)^{-1}A^\top \epsilon\|^2 = \sigma^2 \sum_{k=1}^n \lambda_k,$$

where λ_k are the eigenvalues of $(A^\top A)^{-1}$, i.e.

$$\lambda_k A^\top A \psi_k = \psi_k, \quad \lambda_1 \leq \lambda_2, \dots, \leq \lambda_n$$

and $\psi_k \in \mathbb{R}^n$ are the eigenvectors of $A^\top A$.

If A is ill-posed or if n is large the risk of $\hat{\theta}_0$ may be very large.

Spectral regularization

The basic idea in the spectral regularization is to suppress large λ_k

in the risk $\sigma^2 \sum_{k=1}^n \lambda_k$.

The simplest method is

$$\hat{\theta}_\alpha = H_\alpha \hat{\theta}_0 = H_\alpha [(A^\top A)^{-1}] (A^\top A)^{-1} A^\top Y,$$

where $H_\alpha [(A^\top A)^{-1}] (s, l) = \sum_{k=1}^n H_\alpha(\lambda_k) \psi_l(k) \psi_l(k)$.

Typically $\lim_{\alpha \rightarrow 0} H_\alpha(\lambda) = 1$, $\lim_{\lambda \rightarrow \infty} H_\alpha(\lambda) = 0$ for all $\alpha > 0$.

α is called regularization parameter.

Bias-variance decomposition

For the risk of $\hat{\theta}_\alpha$ we get a standard bias-variance decomposition

$$\mathbf{E}\|\hat{\theta}_\alpha - \theta\|^2 = \sum_{k=1}^n [1 - H_\alpha(\lambda_k)]^2 \langle \theta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^n \lambda_k H_\alpha^2(\lambda_k),$$

where $\langle \theta, \psi_k \rangle = \sum_{l=1}^n \theta(l) \psi_k(l)$.

Remarks:

The spectral regularization may improve substantially $\hat{\theta}_0$ when $\langle \theta, \psi_k \rangle^2$ are small for large k .

The best regularization depends on θ and therefore it should be data-driven.

- Spectral cut-off (requires the SVD)

$$H_\alpha(\lambda) = \mathbf{1}\{\alpha\lambda \leq 1\}$$

- Tikhonov's regularization

$$\hat{\theta}_\alpha = \arg \min_{\theta} \left\{ \|Y - A\theta\|^2 + \alpha\|\theta\|^2 \right\}$$

or

$$\hat{\theta}_\alpha = [\alpha I + A^\top A]^{-1} A^\top Y, \quad H_\alpha(\lambda) = \frac{1}{1 + \alpha\lambda}$$

- Landweber's iterations (solve $A^\top Y = A^\top A\theta$)

$$\hat{\theta}_i = [I - a^{-1}A^\top A]\hat{\theta}_{i-1} + a^{-1}A^\top Y$$

The method converges if $a\lambda_1 < 1$. It is easy to check that

$$H_i(\lambda) = 1 - [1 - (a\lambda)^{-1}]^{i+1}, \quad \alpha = 1/(i + 1)$$

The main goal is to find the best method within the family spectral regularization methods

$$\hat{\theta}_\alpha = H_\alpha[(A^\top A)^{-1}](A^\top A)^{-1}A^\top Y, \quad \alpha \in \mathbb{R}^+, \quad H_0[(AA^\top)^{-1}] = I.$$

We want to find $\hat{\alpha}(Y)$ that minimizes $\mathbf{E}\|\theta - \hat{\theta}_{\hat{\alpha}(Y)}(Y)\|^2$ uniformly in $\theta \in \mathbb{R}^n$.

The empirical risk minimization principle

$$\hat{\alpha} = \arg \min_{\alpha} \left\{ \|\hat{\theta}_0 - \hat{\theta}_\alpha\|^2 + \sigma^2 \text{Pen}(\alpha) \right\},$$

where $\text{Pen}(\alpha)$ is a given function $\mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\lim_{\alpha \rightarrow 0} \text{Pen}(\alpha) = \infty$.

Main goals:

- for a given penalty, compute the risk

$$R_{Pen}(\theta) = \mathbf{E} \|\theta - \hat{\theta}_{\hat{\alpha}(Y)}(Y)\|^2$$

- find a penalty that minimizes $R_{Pen}(\theta)$ uniformly in $\theta \in \mathbb{R}^n$

Centered empirical risk. Notice that

$$\hat{\alpha} = \arg \min_{\alpha} R_{Pen}[Y, \alpha],$$

where $R_{Pen}[Y, \alpha]$ is the centered empirical risk defined by

$$R_{Pen}[Y, \alpha] = \|\hat{\theta}_0 - \hat{\theta}_{\alpha}\|^2 + Pen(\alpha)\sigma^2 - \|\theta - \hat{\theta}_0\|^2.$$

Definition

For a given penalty $Pen(\cdot)$ the excess risk function is defined by

$$\Delta_{Pen}(\mu) = \sup_{\theta \in \mathbb{R}^n} \left\{ \mathbf{E}_{\theta} \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 - (1 + \mu) \mathbf{E}_{\theta} R_{Pen}[Y, \hat{\alpha}] \right\}, \quad \mu \in \mathbb{R}^+.$$

Notice that

$$\begin{aligned} \mathbf{E}_{\theta} R_{Pen}[Y, \hat{\alpha}] &\leq \inf_{\alpha} \mathbf{E}_{\theta} R_{Pen}[Y, \alpha] \stackrel{\text{def}}{=} r_{Pen}(\theta) \\ &= \inf_{\alpha} \left\{ \mathbf{E} \|\theta - \hat{\theta}_{\alpha}\|^2 + \sigma^2 Pen(\alpha) - 2\sigma^2 \sum_{k=1}^n \lambda_k H_{\alpha}(\lambda_k) \right\}, \end{aligned}$$

and we get the following oracle inequality

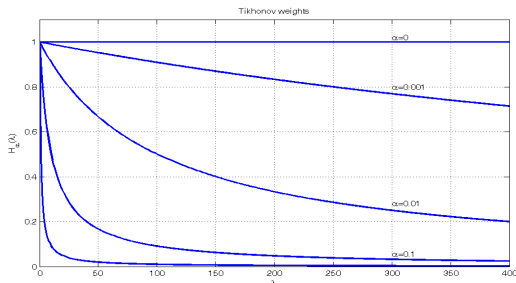
$$\mathbf{E}_{\theta} \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq r_{Pen}(\theta) + \inf_{\mu} \left\{ \Delta_{Pen}(\mu) + \mu r_{Pen}(\theta) \right\}.$$

Definition

The family $\{H_\alpha([A^\top A]^{-1}), \alpha \geq 0\}$ is called *ordered* if:

- 1 for all $\alpha \geq 0$ and $\lambda \geq 0$, $0 \leq H_\alpha(\lambda) \leq 1$
- 2 $H_{\alpha_1}(\lambda) \geq H_{\alpha_2}(\lambda)$, for all $\alpha_1 \leq \alpha_2$ and all $\lambda > 0$.

Typical examples: Tikhonov's regularization, the spectral cut-off method, the Landweber iterations.



Theorem

Let $\{H_\alpha(\cdot), \alpha \geq 0\}$ be a family of ordered smoothers. Then for some $C > 0$ and for all $\mu > 0$

$$\Delta_{Pen}(\mu) \leq \sigma^2 \Delta_{Pen}^C(\mu),$$

where

$$\Delta_{Pen}^C(\mu) \stackrel{\text{def}}{=} \mathbf{E} \sup_{\alpha} \left\{ \sum_{k=1}^n \xi^2(k) \lambda_k H_\alpha(\lambda_k) [2 + 2\mu - \mu H_\alpha(\lambda_k)] \right. \\ \left. + \frac{C \max_k \lambda_k H_\alpha^2(\lambda_k)}{\mu} - (1 + \mu) Pen(\alpha) \right\}.$$

Let $Pen(\alpha) = 2 \sum_{k=1}^n H_{\alpha}(\lambda_k) \lambda_k$. This penalty is related to the so-called unbiased risk estimation /Akaike (1965)/, since

$$\mathbf{E}_{\theta} \|\theta - \hat{\theta}_{\alpha}\|^2 = \mathbf{E} R_{Pen}[Y, \alpha] \text{ for any given } \alpha.$$

A is not severely ill-posed if there exists $\kappa < 1$ such that

$$\begin{aligned} \max_k \lambda_k H_{\alpha}^2(\lambda_k) &\leq \lambda_1 \left[\frac{1}{\lambda_1} \sum_{k=1}^n \lambda_k H_{\alpha}^2(\lambda_k) \right]^{\kappa}, \\ \sum_{k=1}^n \lambda_k^2 H_{\alpha}^2(\lambda_k) &\leq \frac{\lambda_1^2}{1 - \kappa} \left[\frac{1}{\lambda_1} \sum_{k=1}^n \lambda_k H_{\alpha}^2(\lambda_k) \right]^{1+\kappa}. \end{aligned}$$

In other words, $\lambda_k \leq \lambda_1 k^{\kappa/(1-\kappa)}$.

Unbiased Risk Estimation

Theorem

Let $\{H_\alpha[(A^\top A)^{-1}], \alpha \geq 0\}$ be a family of ordered smoothers and A is not severely ill-posed, then for some $C > 1$ and any $\mu \in (0, 1)$

$$\Delta_{Pen}(\mu) \leq \frac{C^{1/(1-\kappa)} \lambda_1 \sigma^2}{(1-\kappa)^{1/(1-\kappa)} \mu^{(1+\kappa)/(1-\kappa)}}.$$

This upper bound explodes for severely ill-posed A , i.e. as $\kappa \rightarrow 1$.

Let $\tilde{\alpha}$ be a data-driven smoothing parameter. Its performance is measured by *oracle efficiency*

$$\mathcal{E}_{or}(\tilde{\alpha}, \theta) = \frac{\inf_{\alpha} \mathbf{E} \|\hat{\theta}_{\alpha} - \theta\|^2}{\mathbf{E} \|\hat{\theta}_{\tilde{\alpha}} - \theta\|^2}.$$

Since it is impossible to compute the oracle efficiency for all $\theta \in \mathbb{R}^n$, we choose a sufficiently representative family of vectors θ

$$\theta^M(k) = \frac{M\sigma}{1 + (k/W)^m},$$

where M is called amplitude, W bandwidth, and m smoothness.

We vary M and plot

$$\mathcal{E}(M) = \mathcal{E}_{or}(\tilde{\alpha}, \theta^M)$$

The parameters $m = 6$ and $W = 6$ are assumed to be fixed.

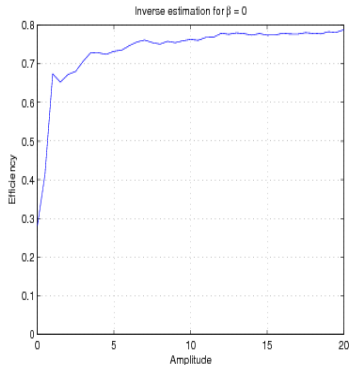
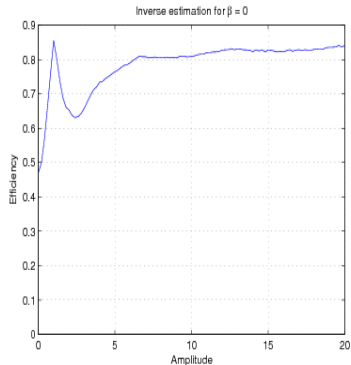
We use the spectral cut-off regularization $H_\alpha(\lambda) = \mathbf{1}\{\alpha\lambda \leq 1\}$ with the following penalties:

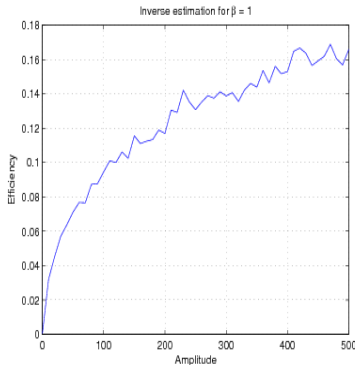
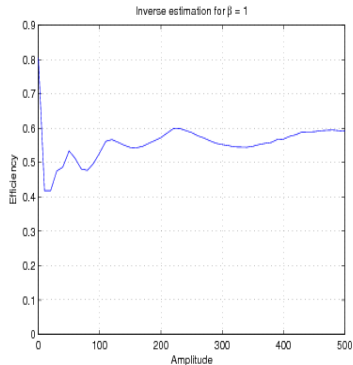


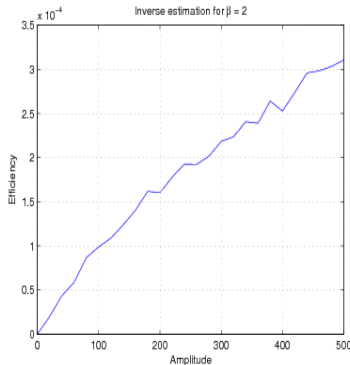
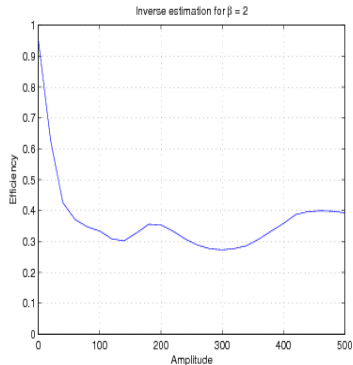
$$\underline{Pen}(\alpha) = 2 \sum_{k=1}^n \lambda_k H_\alpha(\lambda_k)$$

- an upper penalty $\overline{Pen}(\alpha)$ defined as a "minimal" penalty such that

$$\mathbf{E} \sup_{\alpha} \left\{ (2 - \mu) \sum_{k=1}^n \xi_k^2 \lambda_k H_\alpha(\lambda_k) - \overline{Pen}(\alpha) \right\} \leq \frac{1}{\mu}.$$

Direct estimation $\lambda_k = 1$ Lower penalty $\underline{Pen}(\alpha)$ Upper penalty $\overline{Pen}(\alpha)$

Inverse estimation $\lambda_k = k$ Lower penalty $\underline{Pen}(\alpha)$ Upper penalty $\overline{Pen}(\alpha)$

Inverse estimation $\lambda_k = k^2$ Lower penalty $\underline{Pen}(\alpha)$ Upper penalty $\overline{Pen}(\alpha)$

Summary: what is going on in statistics

Basic problem : estimate $\theta \in \mathbb{R}^n$ with the help of the data $Y \sim P_\theta$.

Statistics I : we are allowed to use all estimators $\hat{\theta}$.

- To define the best estimator, we need an a priori information, e.g. a probability measure $\pi(\theta)$. Then

$$\theta_\pi^*(Y) = \arg \min_{\hat{\theta}} \int \pi(\theta) \mathbf{E}_\theta \|\hat{\theta} - \theta\|^2 d\theta = \frac{\int \theta \pi(\theta) p_\theta(Y) d\theta}{\int \pi(\theta) p_\theta(Y) d\theta}.$$

- Drawback: for large n , $\theta_\pi^*(Y)$ strongly depends on $\pi(\cdot)$ which is hardly known in practice
- Basic tools: limit theorems

Statistics II : we are allowed to use only a small class of estimators

$\hat{\theta}_\alpha(Y)$, $\alpha \in \mathcal{A}$

- within this class, we look for an estimator $\hat{\theta}_{\alpha^*}(Y)$ such that

$$\mathbf{E}_\theta \|\hat{\theta}_{\alpha^*}(Y) - \theta\|^2 \rightarrow \min \text{ uniformly in } \theta \in \mathbb{R}^n.$$

- Drawbacks :
 - the optimal estimator doesn't exist
 - there is no room for asymptotic methods: consistency and convergence rates are replaced by non-asymptotic oracle inequalities
- Basic tools: non-asymptotic entropy bounds combined with extensive Monte-Carlo computations