

# Ordered processes and high dimensional linear models

Yuri Golubev

CNRS, Université Aix-Marseille 1

*SAPS VI, 21-23 March 2007*

This 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

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

- *$m$  and  $n$  are assumed to be large.*

# Ordered Processes: an example

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

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

Consider  $\xi(t) = \xi \times t$ ,  $t \geq 0$ , where  $\xi$  is  $\mathcal{N}(0, 1)$ . It is also very easy to check

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

*Question: what is  $\xi(t)$  satisfying*

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

## 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

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

is an ordered process if  $H \geq 1/2$ .

- ordered processes related to the so-called ordered smoothers introduced by Kneip (1995).

# Ordered Smoothers

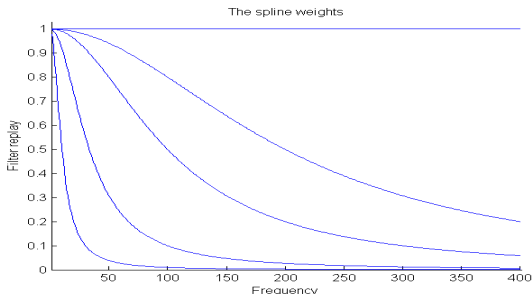
The family of functions  $H_\alpha(\lambda)$ ,  $\alpha, \lambda \in \mathbb{R}^+$  is called *ordered smoothers* if



$$0 \leq H_\alpha(\lambda) \leq 1$$

- for all  $\lambda \in \mathbb{R}^+$

$$H_{\alpha_1}(\lambda) \geq H_{\alpha_2}(\lambda), \quad \alpha_1 \leq \alpha_2, .$$



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

$$\eta_0(t) = \sum_{k=1}^{\infty} [1 - H_t(\lambda_k)] \xi(k) \theta(k),$$

$$\eta_2(t) = \sum_{k=1}^{\infty} \lambda_k H_{1/t}^2(\lambda_k) (\xi^2(k) - 1)$$

are ordered processes.

# Dichotomy Inequality

Denote for brevity

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

## Theorem

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

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

## 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  $C$  depending on  $\lambda$  such that for all  $\mu > 0$

$$\mathbf{E} \sup_{t \geq 0} [\xi(t) - \mu \sigma^q(t)]_+^p \leq \frac{C[2q(p+2) - 4]^{q(p+2)-2}}{\mu^{p/(q-1)}},$$

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

Let  $\tau$  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 the ordered process  $\xi(\cdot)$

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

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

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

/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  $\lambda_k$  are the eigenvalues of  $(A^\top A)^{-1}$

$$\lambda_k A^\top A \psi_k = \psi_k$$

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

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

The basic idea to improve  $\hat{\theta}_0$  is to make the variance  $\sigma^2 \sum_{k=1}^n \lambda_k$  smaller suppressing largest  $\lambda_k$ .

The simplest method is based on *linear filtering*

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

where

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

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

# Bias-variance decomposition

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

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

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

*The spectral regularization make sense if  $\langle \theta, \psi_k \rangle^2 \approx 0$  for large  $k$ .*

*The best regularization depends on  $\theta$ .*

# Basic spectral regularization methods

- Spectral cut-off

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

- 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 are defined by

$$\theta_i = [I - a^{-1}A^\top A]\theta_{i-1} + a^{-1}A^\top Y$$

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

$$H_i(\lambda) = 1 - \left(1 - \frac{1}{a\lambda}\right)^{i+1}$$

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 such that  $\lim_{\alpha \rightarrow 0} \text{Pen}(\alpha) = \infty$ .

## Problems:

- for a given penalty, compute the risk

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

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

Notice that

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

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

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

## Definition

For any  $\mu > 0$  and a given penalty  $Pen(\cdot)$  the excess risk is defined by

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

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 therefore uniformly in  $\theta$

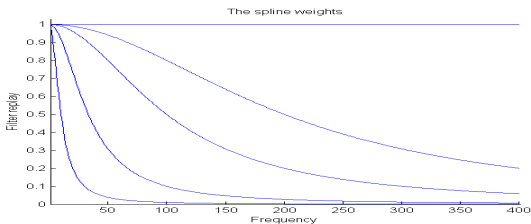
$$\mathbf{E}_\theta \|\theta - \hat{\theta}_{\hat{\alpha}}\|^2 \leq r_{Pen}(\theta) + \inf_{\mu} \left\{ \Delta_{Pen}(\mu) + \mu r_{Pen}(\theta) \right\}.$$

## Definition

The family of smoothers  $\{H_\alpha(\cdot), \alpha \geq 0\}$  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 of ordered smoothers are provided the Tikhonov 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

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

# Unbiased risk estimation

Suppose

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

This penalty is related to the unbiased risk estimation, since for any given  $\alpha$

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

We say  $A$  is not severely ill-posed if there exists  $\varkappa < 1$  such that for all  $\alpha \geq 0$

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

## Theorem

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

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

# How does it work

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  $\theta$

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

where  $A$  is called amplitude,  $W$  bandwidth, and  $m$  smoothness.

# How does it work

We vary  $A$  and plot

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

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

We use the spectral cut-off regularization

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

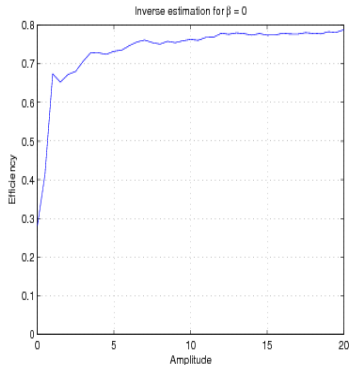
with the following penalties:

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

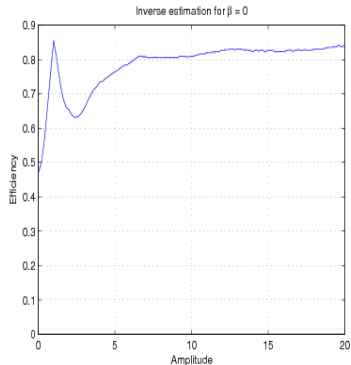
and the upper penalty which is defined as a "minimal" penalty such that

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

# Direct estimation $\lambda_k = 1$

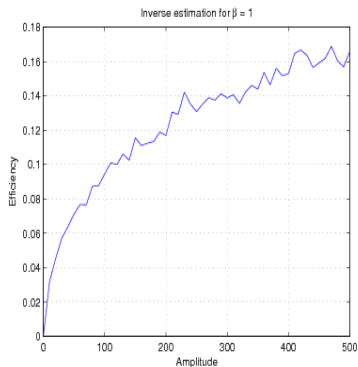


Lower penalty  $\underline{Pen}(\alpha)$

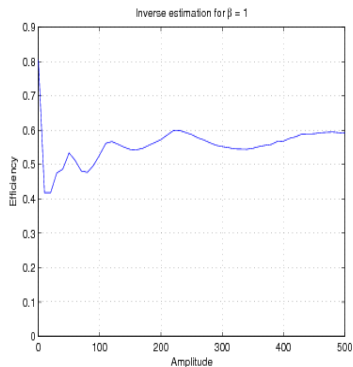


Upper penalty  $\overline{Pen}(\alpha)$

# Inverse estimation $\lambda_k = k$

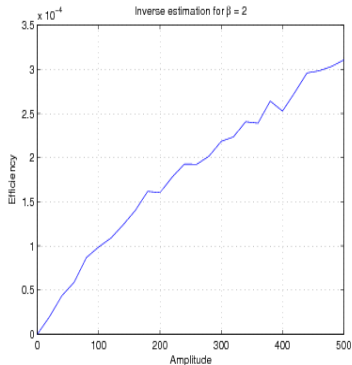


Lower penalty  $\underline{Pen}(\alpha)$

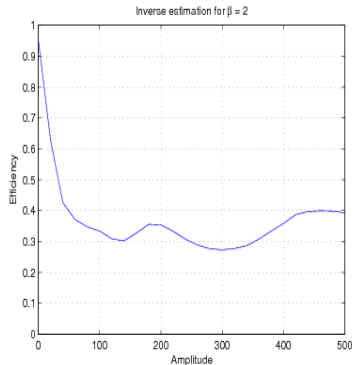


Upper penalty  $\overline{Pen}(\alpha)$

# 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  $Y \sim P_\theta$*

- Classical statistics : we are allowed to use all estimators  $\hat{\theta}$ 
  - the estimators within this class are not comparable. To define the best estimator, we need an a priori information, e.g. a probability measure  $\pi(\theta)$ . Then the best estimator is given by

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

- for large  $n$ ,  $\theta_\pi^*(Y)$  strongly depends on  $\pi(\cdot)$ .
- Modern approach : 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

$$\mathbf{E}_\theta \|\hat{\theta}_{\alpha^*}(Y) - \theta\|^2 \lesssim \mathbf{E}_\theta \|\hat{\theta}_\alpha(Y) - \theta\|^2 \text{ uniformly in } \theta.$$