Département de Mathématiques d'Orsay

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

Estimator Selection in High-Dimensional Settings

Christophe Giraud

Université Paris Sud

Journées MAS, Toulouse, 2014

Big Data?

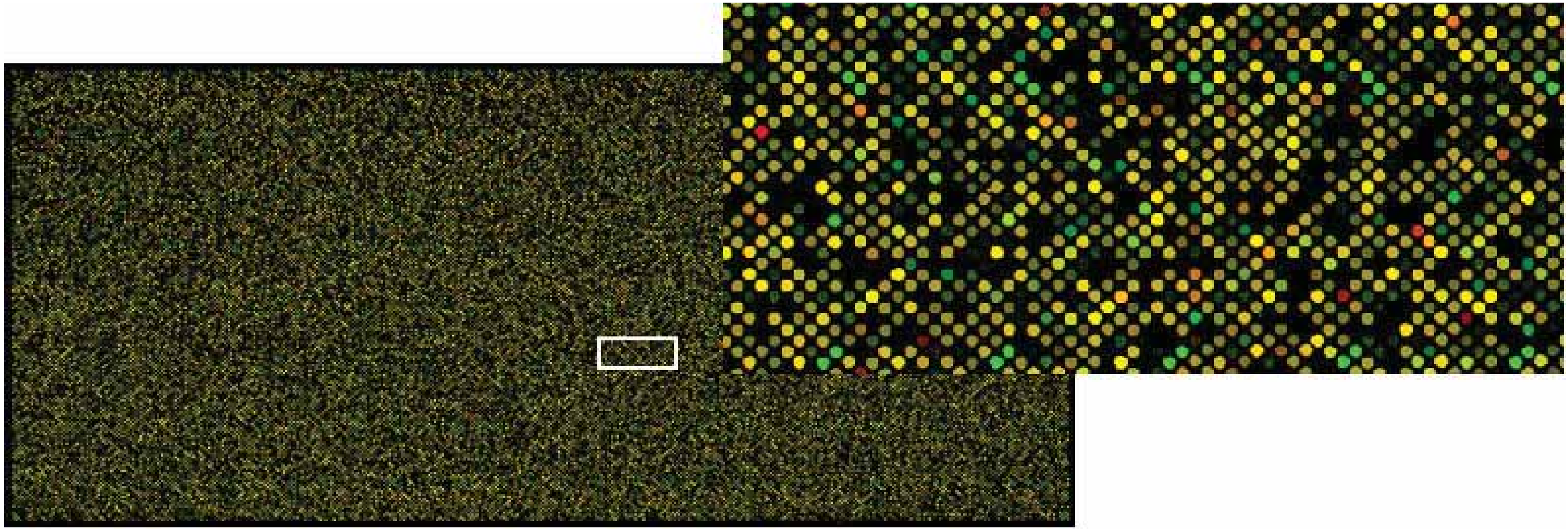
High-Dimensional Statistics

~~Big Data?~~

"Wide" Data

High-dimensional data

- **Biotech data :** Biotech devices can sense up to tens of thousands of "features" $\gg n =$ number of "individuals".
- **Images :** medical images, massive astrophysic images, video surveillance images, etc. Each image is made of thousands up to millions of pixels or voxels.
- **Consumers preferences data :** websites and loyalty programs collect huge amounts of informations on the preferences and the behaviors of customers. Ex: recommendation systems for movies, books, musics.
- **Business data :** optimal exploitation of internal and external data (logistic and transportation, insurance, finance, etc)
- **Crowdsourcing data :** massive online participative data sets (recorded by volunteers). Ex: eBirds collects online millions of bird counts across Northern America.



Whole Human Genome Microarray covering over 41,000 human genes and transcripts on a standard 1" x 3" glass slide format.

Blessing?

☺ we can sense thousands of variables on each "individual" : potentially we will be able to scan every variables that may influence the phenomenon under study.

☹ the curse of dimensionality : separating the signal from the noise is <u>in general</u> almost impossible in high-dimensional data and computations can rapidly exceed the available resources.

Curse 1 : fluctuations cumulate

Exemple : linear regression $Y = \mathbf{X}\beta^* + \varepsilon$ with $\mathbf{cov}(\varepsilon) = \sigma^2 I_n$.
The Least-Square estimator $\widehat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2$ has a risk

$$\mathbb{E}\left[\|\widehat{\beta} - \beta^*\|^2\right] = \operatorname{Tr}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\sigma^2.$$

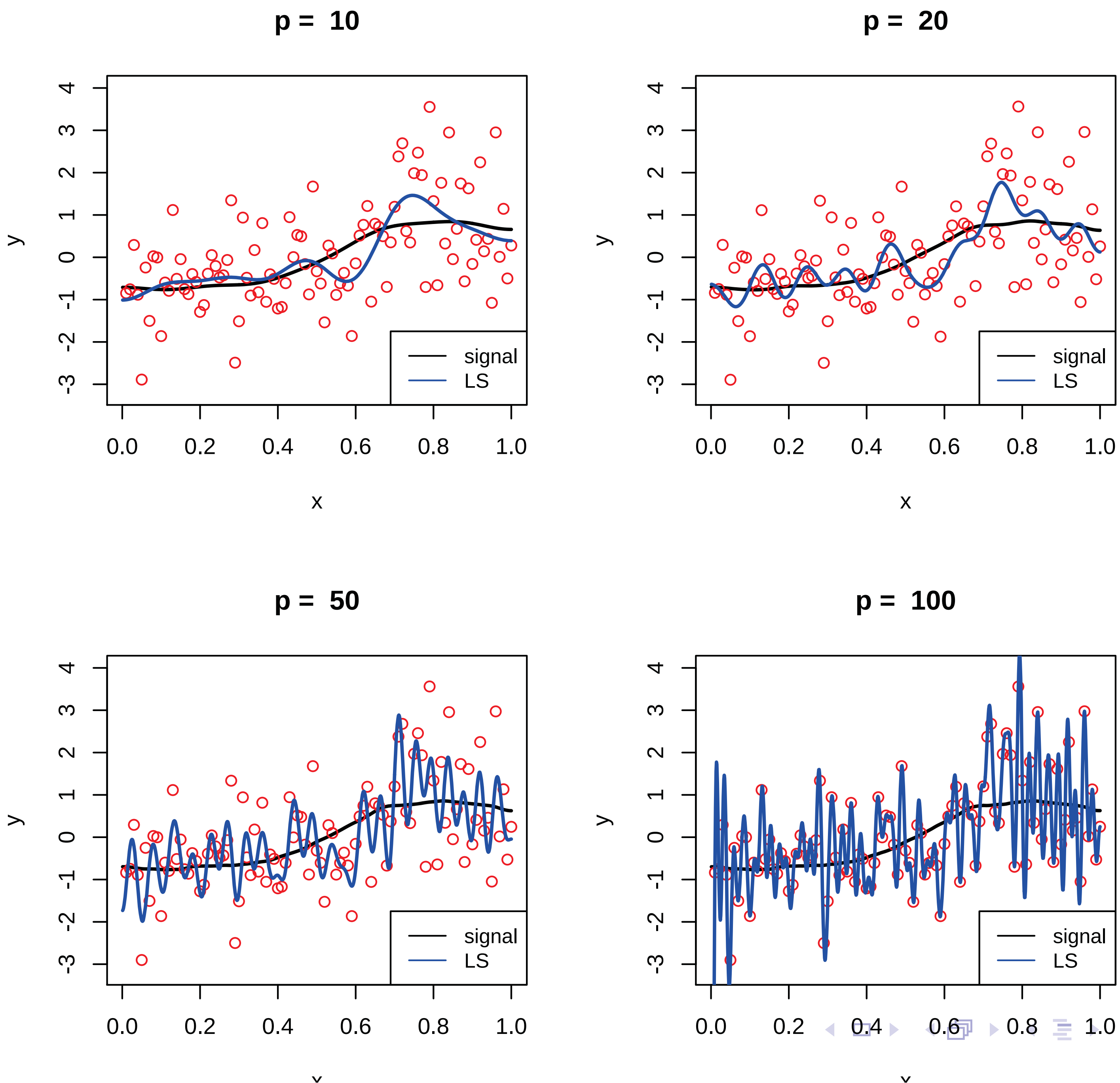
Illustration :

$$Y_i = \sum_{j=1}^{p} \beta_j^* \cos(\pi j i/n) + \varepsilon_i = f_{\beta^*}(i/n) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

with

- $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d with $\mathcal{N}(0,1)$ distribution
- β_j^* independent with $\mathcal{N}(0, j^{-4})$ distribution

Curse 1 : fluctuations cumulate



Curse 2 : locality is lost

Observations $(Y_i, X^{(i)}) \in \mathbb{R} \times [0,1]^p$ for $i = 1, \ldots, n$.

Model: $Y_i = f(X^{(i)}) + \varepsilon_i$ with f smooth.

Local averaging: $\widehat{f}(x) = \text{average of}\left\{Y_i : X^{(i)} \text{ close to } x\right\}$

Curse 2 : locality is lost

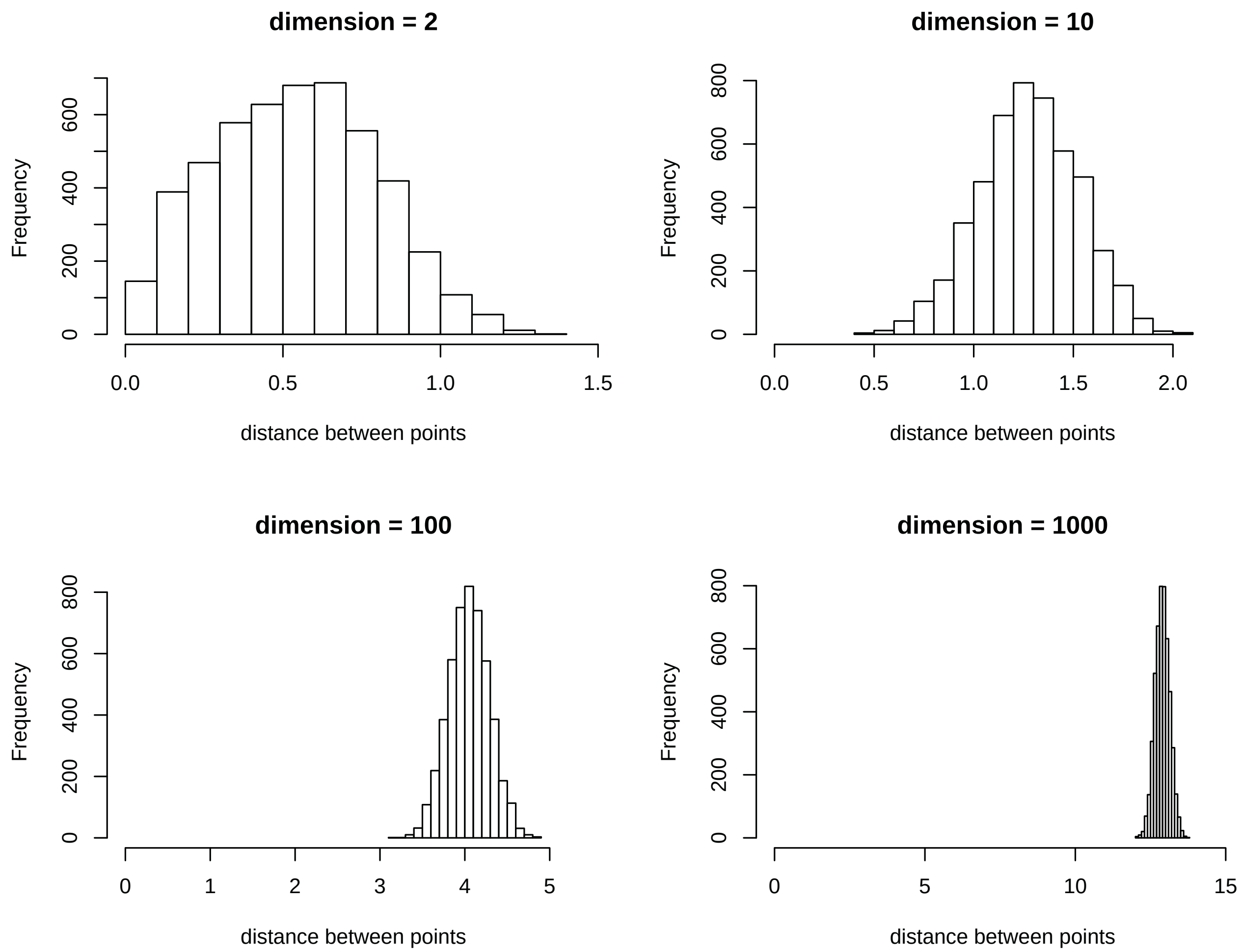


Figure: Histograms of the pairwise-distances between $n = 100$ points sampled uniformly in the hypercube $[0, 1]^p$, for $p = 2, 10, 100$ and 1000.

Curse 2 : locality is lost

Number n of points $x_1, \ldots, x_n$ required for covering $[0,1]^p$ by the balls $B(x_i, 1)$:

$$n \geq \frac{\Gamma(p/2+1)}{\pi^{p/2}} \overset{p\to\infty}{\sim} \left(\frac{p}{2\pi e}\right)^{p/2} \sqrt{p\pi}$$

p	20	30	50	100	200
n	39	45630	$5.7\,10^{12}$	$42\,10^{39}$	larger than the estimated number of particles in the observable universe

Some other curses

- Curse 3 : an accumulation of rare events may not be rare (false discoveries, etc)
- Curse 4 : algorithmic complexity must remain low

Low-dimensional structures in high-dimensional data

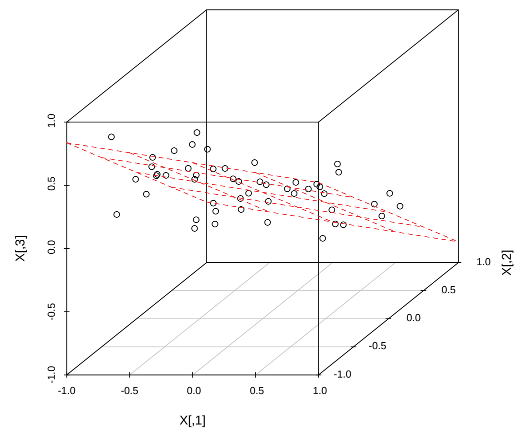
Hopeless?

Low dimensional structures : high-dimensional data are usually concentrated around low-dimensional structures reflecting the (relatively) small complexity of the systems producing the data

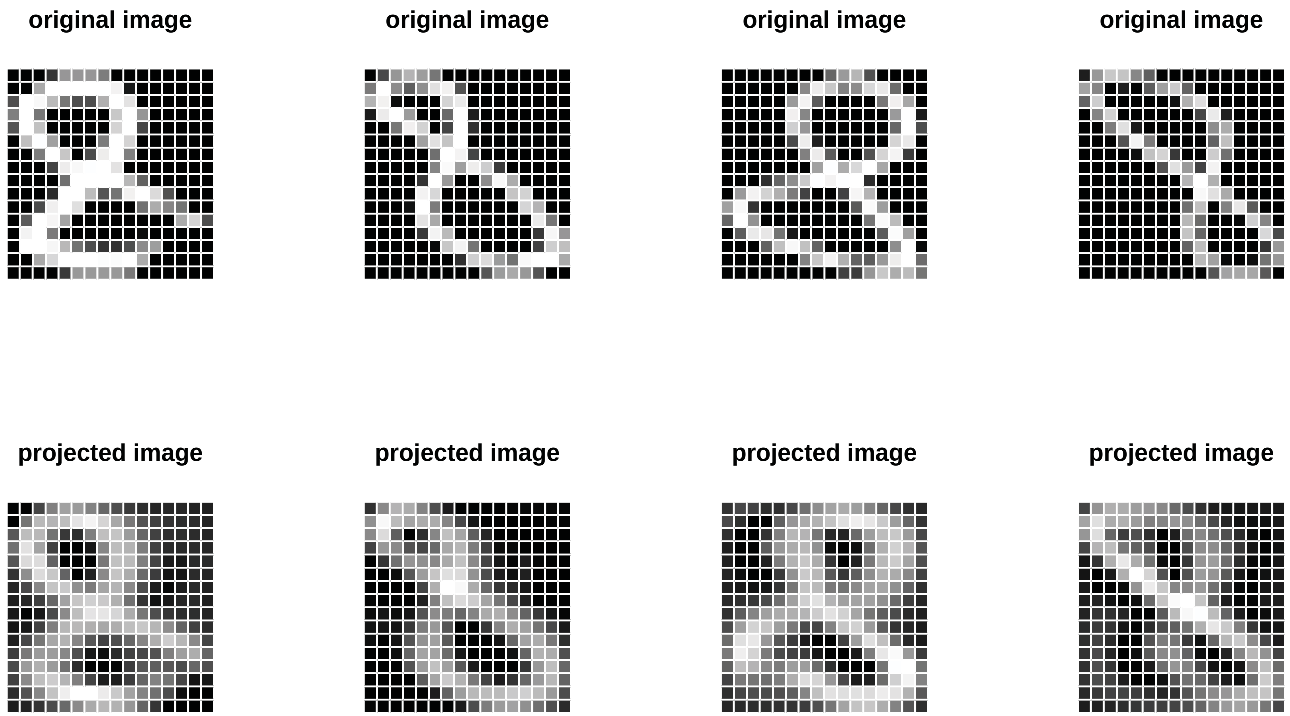
- geometrical structures in an image,
- regulation network of a "biological system",
- social structures in marketing data,
- human technologies have limited complexity, etc.

Dimension reduction :

- "unsupervised" (PCA)
- "estimation-oriented"



PCA in action



MNIST : 1100 scans of each digit. Each scan is a 16×16 image which is encoded by a vector in $\mathbb{R}^{256}$. The original images are displayed in the first row, their projection onto 10 first principal axes in the second row.

"Estimation-oriented" dimension reduction

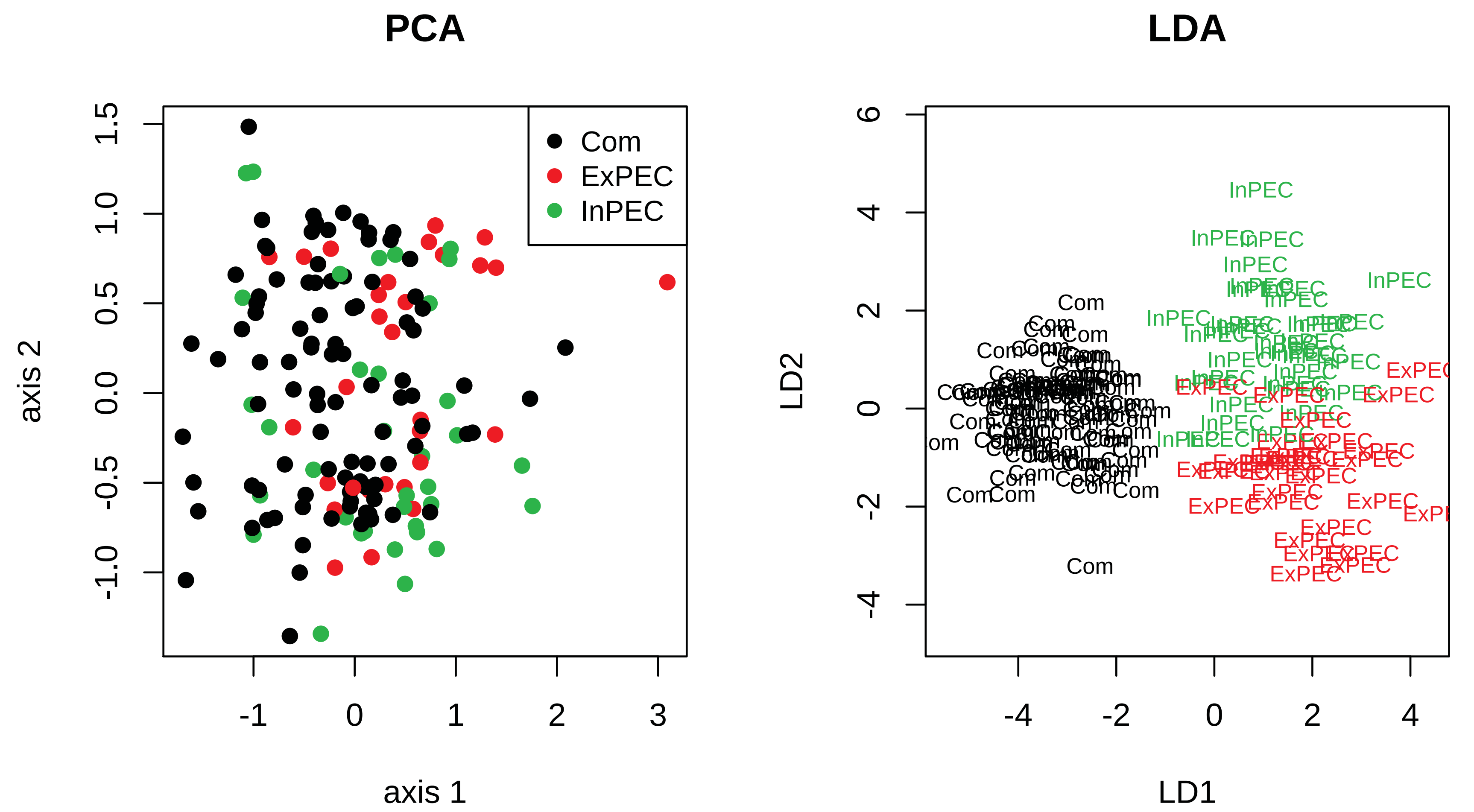


Figure: 55 chemical measurements of 162 strains of *E. coli*.
Left : the data is projected on the plane given by a PCA.
Right : the data is projected on the plane given by a LDA.

A paradigm shift

Classical statistics:

- a small number p of parameters
- a large number n of observations.

High-dimensional data:

- a huge number p of parameters
- a sample size n with $n \asymp p$ or $n \ll p$.

classical asymptotic analyses do not fit!

Statistical setting:

- either $n, p \to \infty$ with $p \sim g(n)$ (yet sensitive to g)
- or treat n and p as they are (yet analysis is more involved)

Central issue:

identify (at least approximately) the low-dimensional structures

Estimator selection

Unknown structures

- the low-dimensional structures are unknown
- the "class" of hidden structures is possibly unknown
  - various sparsity patterns
  - smoothness
  - low-rank
  - etc

Strong effort

For each "class" of structures, many computationally efficient estimators adapting to the hidden structures have been developed

Estimator selection

Difficulties

- No procedure is universally better than the others
- A sensible choice of the tuning parameters usually depends on some unknown characteristics of the data (sparsity, smoothness, variance, etc)

Estimator selection / aggregation

needed for

- adapting to the possibly unknown "classes" of structures
- choosing the best among several estimators
- tuning parameters

Ideal objective

- Select the "best" estimator among a collection $\{\hat{f}_\lambda, \lambda \in \Lambda\}$.

1- a simple setting

Regression framework

Regression setting

- $Y_i = F(x_i) + \varepsilon_i$, for $i = 1, \ldots, n$, with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d.
- $F : \mathcal{X} \to \mathbb{R}$ and $\sigma^2 = \mathrm{var}(\varepsilon_i)$ are unknown
- we want to estimate F or $f^* = (F(x_1), \ldots, F(x_n))$

Ex 1 : sparse linear regression

- $F(x) \approx \langle \beta^*, \phi(x) \rangle$ with β^* "sparse" in some sense and $\phi(x) \in \mathbb{R}^p$ with possibly $p > n$

Ex 2 : non-parametric regression

- $F : \mathcal{X} \to \mathbb{R}$ is smooth

A plethora of estimators

For each "class" of structures, there is a strong effort for developing computationally efficient estimators which adapt to the hidden structures

Sparse linear regression

- Coordinate sparsity: Lasso, Dantzig, Elastic-Net, Exponential-Weighting, Random Forest, etc.
- Structured sparsity: Group-lasso, Fused-Lasso, Hierarchical-Group Lasso, Bayesian estimators, etc

Non-parametric regression

- Spline smoothing, Nadaraya kernel smoothing, kernel ridge estimators, nearest neighbors, L^2-basis projection, Sparse Additive Models, etc

Important practical issues

Which class of structures?

- coordinate sparse linear regression?
- group-sparse linear regression?
- smoothing?

Which estimator should be used?

- Sparse regression: Lasso? Exponential-Weighting? Random-Forest?
- Non-parametric regression: Kernel regression? (which kernel?) Spline smoothing?

Which "tuning" parameter?

- which penalty level for the lasso?
- which bandwith for kernel regression?
- etc

A simple example

Model

$$Y = \mathbf{X}\beta^* + \varepsilon$$

Scale-invariance

An estimator $(Y, \mathbf{X}) \to \widehat{\beta}(Y, \mathbf{X})$ is scale invariant if

$$\widehat{\beta}(sY, \mathbf{X}) = s\widehat{\beta}(Y, \mathbf{X}).$$

Lasso estimator

$$\widehat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left\{ \|Y - \mathbf{X}\beta\|^2 + 2\lambda\|\beta\|_1 \right\}, \quad \lambda > 0$$

is not scale-invariant

$$\widehat{\beta}_\lambda(sY, \mathbf{X}) = s\widehat{\beta}_{\lambda/s}(Y, \mathbf{X}).$$

A simple example

The compatibility constant

$$\kappa[S] = \min_{u \in \mathcal{C}(S)} \left\{ |S|^{1/2} \|\mathbf{X}u\|_2 / \|u_S\|_1 \right\},$$

where $\mathcal{C}(S) = \{u : \ \|u_{S^c}\|_1 < 4\|u_S\|_1\}$.

Theorem (Koltchinskii, Lounici, Tsybakov)

Assumptions

- $\varepsilon_1, \dots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$
- columns of $\mathbf{X}$ with norm 1

Then, for $\lambda = 3\sigma\sqrt{2\log(p) + 2L}$ with probability at least $1 - e^{-L}$

$$\|\mathbf{X}(\widehat{\beta}_\lambda - \beta^*)\|^2 \leq \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|^2 + \frac{18\sigma^2(L + \log(p))}{\kappa[\mathrm{supp}(\beta)]^2} |\beta|_0 \right\}$$

Selection strategies

Resampling strategies

- V-fold Cross-Validation
- and many others

Complexity penalization

- Penalized log-likelihood (AIC, BIC, etc)
- LinSelect
- pairwise comparison : Goldenshluger-Lepski's method, Birgé-Baraud's method

Some specific strategies

- Square-root / scaled (group-)Lasso

Principle of complexity penalization

Model

$$Y = f^* + \varepsilon$$

Typical selection criterion

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda} \left\{ \| Y - \widehat{f}_\lambda \|^2 + \mathrm{pen}(\lambda) \right\}$$

Main issue

To shape $\lambda \to \mathrm{pen}(\lambda)$ in order to have

$$\mathbb{E}\left[\| f^* - \widehat{f}_{\widehat{\lambda}} \|^2 \right] \leq C \min_{\lambda} \mathbb{E}\left[\| f^* - \widehat{f}_\lambda \|^2 \right] + \text{something small}$$

Principle of complexity penalization

If $\widehat{\lambda} \in \operatorname{argmin}_\lambda \left\{ \|Y - \widehat{f}_\lambda\|^2 + \mathrm{pen}(\lambda) \right\}$ then

$$\|f^* - \widehat{f}_{\widehat{\lambda}}\|^2 \leq \|f^* - \widehat{f}_\lambda\|^2 + 2\langle \varepsilon, f^* - \widehat{f}_\lambda \rangle + \mathrm{pen}(\lambda) + 2\langle \varepsilon, \widehat{f}_{\widehat{\lambda}} - f^* \rangle - \mathrm{pen}(\widehat{\lambda}).$$

Which penalty $\mathrm{pen}(\lambda)$?

Find $\mathrm{pen}(\lambda)$ such that there exist some $\{Z_\lambda : \lambda \in \Lambda\}$ fulfilling for some $a < 1$, $c \geq 0$

- $\mathbb{E}\left[\sup_{\lambda \in \Lambda} Z_\lambda\right] \leq c\sigma^2$
- $2\langle \varepsilon, \widehat{f}_\lambda - f^* \rangle - \mathrm{pen}(\lambda) \leq a\|\widehat{f}_\lambda - f^*\|^2 + Z_\lambda, \quad \text{for all } \lambda \in \Lambda.$

Hence

$$(1-a)\,\mathbb{E}\left[\|f^* - \widehat{f}_{\widehat{\lambda}}\|^2\right] \leq \mathbb{E}\left[\|f^* - \widehat{f}_\lambda\|^2\right] + \mathrm{pen}(\lambda) + c\sigma^2$$

Principle of complexity penalization

Since

$$2\langle \varepsilon, \widehat{f}_\lambda - f^* \rangle \leq a \|\widehat{f}_\lambda - f^*\|^2 + a^{-1} \left\langle \varepsilon, \frac{\widehat{f}_\lambda - f^*}{\|\widehat{f}_\lambda - f^*\|} \right\rangle$$

the penalty must control the fluctuations of

$$\left\langle \varepsilon, \frac{\widehat{f}_\lambda - f^*}{\|\widehat{f}_\lambda - f^*\|} \right\rangle .$$

Concentration inequalities help!

Back to the simple example

Restricted eigenvalue

For $k^* = n/(3\log(p))$ we set $\phi_* = \sup\{\|Xu\|_2/\|u\|_2 : u\ k^*\text{-sparse}\}$

Theorem

(Y. Baraud, C.G, S. Huet, N. Verzelen + T. Sun, C-H. Zhang)

If we assume that

- $|\beta^*|_0 \leq C_1\, \kappa^2[\mathrm{supp}(\beta^*)] \times \frac{n}{\phi_* \log(p)}$,

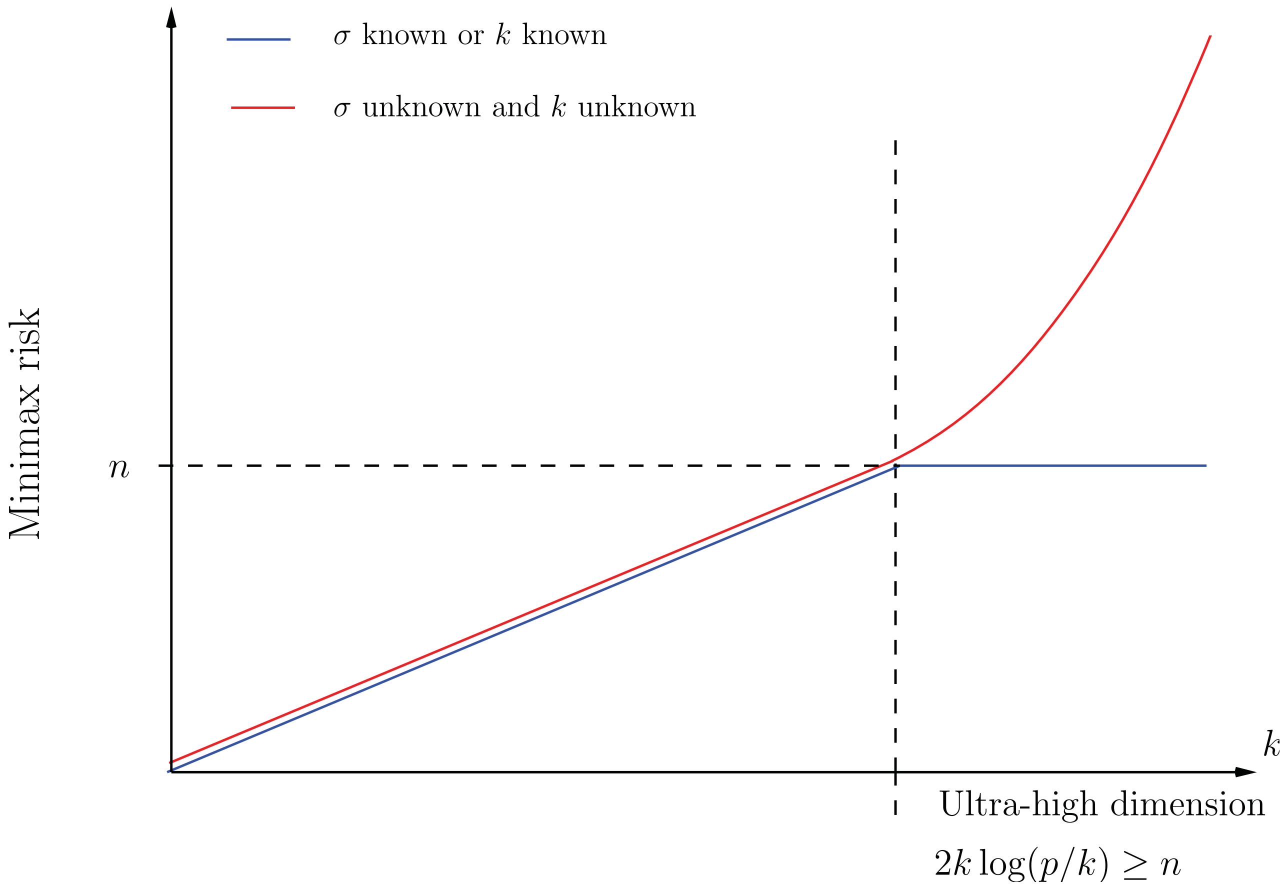
then there is a selection criterion such that with high probability,

$$\|\mathbf{X}(\widehat{\beta} - \beta^*)\|_2^2 \leq C \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta^* - \beta)\|_2^2 + C_2 \frac{\phi_* |\beta|_0 \log(p)}{\kappa^2[\mathrm{supp}(\beta)]} \sigma^2 \right\}$$

Impact of the unknown variance?

Case of coordinate-sparse linear regression (N. Verzelen)

k-sparse signal $f^* = \mathbf{X}\beta^*$



Minimax prediction risk over k-sparse signal as a function of k

2- a more complex setting

Joint work with François Roueff and Andrés Sanchez-Perez.

Another strategy

Issue

The analysis of the fluctuations of the estimators can be intractable in some settings

Example

Non-linear estimators for Time Varying AutoRegressive processes

$$X_t = \sum_{j=1}^{d} \theta_j(t) X_{t-j} + \xi_t$$

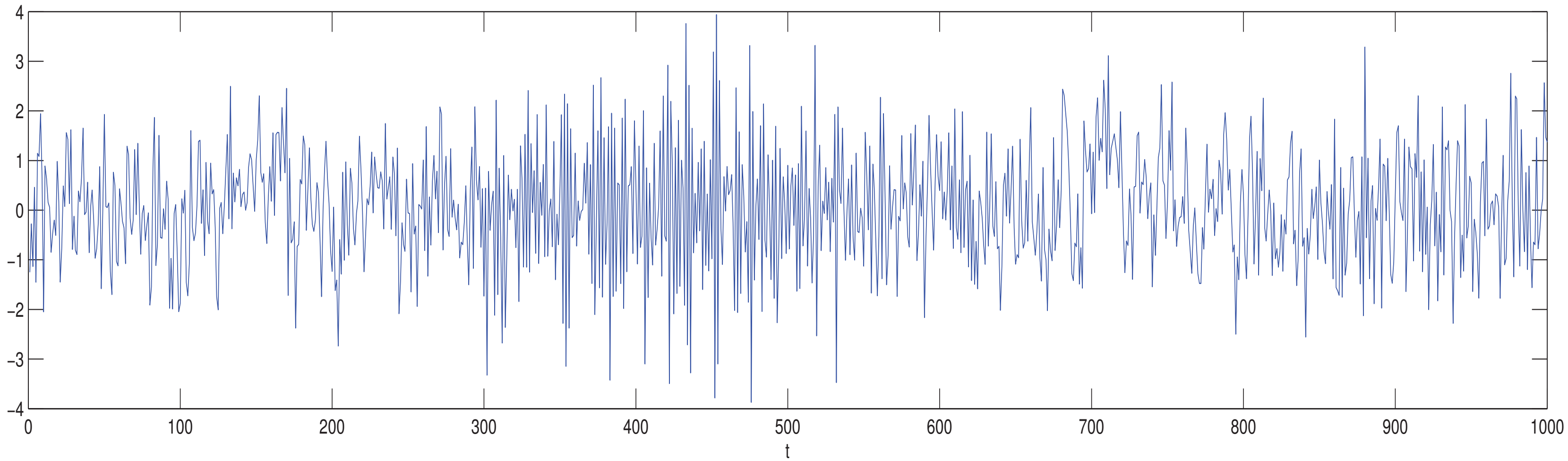


Figure: A TVAR process

Estimation for TVAR

Notations

$\mathbf{X}_{t-1} = (X_{t-1}, \ldots, X_{t-d})$ and $\theta_{t-1} = (\theta_1(t), \ldots, \theta_d(t))$ so that

$$X_t = \langle \theta_{t-1}, \mathbf{X}_{t-1} \rangle + \xi_t$$

Normalized Least Mean Squares estimators (NLMS)

$\widehat{X}_t^{(\lambda)} = \langle \widehat{\theta}_{t-1}^{(\lambda)}, \mathbf{X}_{t-1} \rangle$ where for $\lambda > 0$

$$\widehat{\theta}_t^{(\lambda)} = \widehat{\theta}_{t-1}^{(\lambda)} + \lambda \left(X_t - \langle \widehat{\theta}_{t-1}^{(\lambda)}, \mathbf{X}_{t-1} \rangle \right) \frac{\mathbf{X}_{t-1}}{1 + \lambda \left| \mathbf{X}_{t-1} \right|_2^2} .$$

Optimality

Moulines, Priouret and Roueff (2005) have shown some optimality of $\widehat{\theta}^{(\lambda)}$ for a suitable λ depending on some unknown quantities.

Estimator selection

How to choose λ?

Hard to quantify precisely the fluctuations of a TVAR...

A strategy

Use some technics from individual sequence forecasting.

Individual sequence model

The observations $x_1, x_2, \ldots$ are deterministic!

A simple result (1)

Observations

$x_1, x_2, \ldots$ with values in $[-B, B]$

Predictors

$\hat{x}_t^{(\lambda)}$ for $\lambda \in \Lambda$, with values in $[-B, B]$.

Aggregation

$$\hat{x}_t = \sum_{\lambda \in \Lambda} w_\lambda(t) \hat{x}_t^{(\lambda)} \quad \text{with } w_\lambda(t) \propto \pi_\lambda e^{-\eta \sum_{s=1}^{t-1} (x_s - \hat{x}_s^{(\lambda)})^2}$$

where π is a probability distribution on Λ

A simple result (2)

Theorem (Catoni 97)

For $\eta = 1/(8B^2)$ we have

$$\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2 \leq \min_{\lambda\in\Lambda}\left\{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t^{(\lambda)})^2 + \frac{8B^2}{T}\log(\pi_\lambda^{-1})\right\}$$

proof

$x \longrightarrow \exp(-x^2)$ is concave on $[-2^{-1/2}, 2^{-1/2}]$.

Issues

- Some processes like TVAR are not bounded
- Even if a process is bounded by some B, the choice $\eta \asymp B^{-2}$ can be suboptimal

Extension to unbounded stochastic settings

Sublinear model

$(X_t)_{t\in\mathbb{Z}}$ satisfies

$$|X_t| \leq \sum_{j\in\mathbb{Z}} A_t(j)\, Z_{t-j} \ ,$$

- $Z_t \geq 0$, independent
- $A_t(j) \geq 0$ and $$A_* := \sup_{t\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} A_t(j) < \infty \ . \tag{1}$$

Examples of sublinear processes (1)

Linear processes with time varying coefficients

$$X_t = \sum_{j \in \mathbb{Z}} a_t(j)\, \xi_{t-j} \ ,$$

- $(\xi_t)_{t \in \mathbb{Z}}$ independents, standardized,
- $(a_t(j))_{t,j}$ satisfies (1) with $A_t(j) = |a_t(j)|$.

Classical assumption :

$$\sup_{T \geq 1} \sup_{j \in \mathbb{Z}} \sum_{t=1}^{T} |a_{t,T}(j) - a(t/T, j)| < \infty \ .$$

where $u \longrightarrow a(u, j)$ is smooth.

Examples of sublinear processes (2)

TVAR model

$\theta = (\theta_1, \dots \theta_d) : (-\infty, 1] \to \mathbb{R}, \sigma : (-\infty, 1] \to \mathbb{R}_+$,
$(\xi_t)_{t\in\mathbb{Z}}$ independents, $\mathbb{E}[\xi_t] = 0, \mathbb{E}[\xi_t^2] = 1$.

1. For all $t \le T$,

$$X_{t,T} = \sum_{j=1}^{d} \theta_j \left(\frac{t-1}{T}\right) X_{t-j,T} + \sigma\left(\frac{t}{T}\right) \xi_t \ .$$

2.

$$\lim_{M\to\infty} \sup_{t\le T} \mathbb{P}(|X_{t,T}| > M) = 0 \ .$$

Examples of sublinear processes (2)

Regularity-stability condition

$\theta \in \mathcal{C}\left(\beta, R, \delta, \sigma_-, \sigma_+\right) =$
$\left\{(\boldsymbol{\theta}, \sigma) : (-\infty, 1] \to \mathbb{R}^d \times [\sigma_-, \sigma_+] \;:\; \boldsymbol{\theta} \in \Lambda_d(\beta, R) \cap s_d(\delta)\right\},$
where

- $\Lambda_d(\beta, R) =$
$$\left\{ f : (-\infty, 1] \to \mathbb{R}^d, \sup_{s \neq s'} \frac{\left|f^{(\lceil\beta\rceil - 1)}(s) - f^{(\lceil\beta\rceil - 1)}(s')\right|}{|s - s'|^{\beta + 1 - \lceil\beta\rceil}} \leq R \right\}$$
- $s_d(\delta) = \left\{\theta : (-\infty, 1] \to \mathbb{R}^d, \Theta(z; u) \neq 0, \forall |z| < \delta^{-1}, u \in [0, 1]\right\}$
where $\Theta(z; u) = 1 - \sum_{j=1}^{d} \theta_j(u) z^j$.

Examples of sublinear processes (2)

Proposition

If $(\theta, \sigma) \in \mathcal{C}(\beta, R, \delta, 0, \sigma_+)$*, for* $\delta \in (0,1)$

$$X_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j)\, \sigma\left(\frac{t-j}{T}\right) \xi_{t-j} \ ,$$

where for any $\rho \in (\delta, 1)$*,* $\exists \bar{K} = \bar{K}(\rho, \delta, \beta, R) > 0$ *such that,* $\forall t \leq T$ *and* $j \geq 0$*,*

$$|a_{t,T}(j)| \leq \bar{K} \rho^j \ .$$

Hence

$$\sup_t \sum_{j \geq 0} |a_{t,T}(j)\, \sigma(t - j/T)| \leq \frac{\bar{K}\sigma_+}{1-\rho}.$$

Sublinearity of the predictors

L-Lipschitz predictors

- $L = (L_s)_{s \geq 1}$ non-negative with
$$L_* = \sum_{j \geq 1} L_j < \infty \,.$$
- A predictor $\widehat{X}_t$ of X_t from $(X_s)_{s \leq t-1}$ is L-Lipschitz if
$$\left|\widehat{X}_t\right| \leq \sum_{s \geq 1} L_s \left|X_{t-s}\right| \,.$$

Risk bound with p-th moment

Theorem (C.G., F. Roueff, A. Sanchez-Perez)

If $m_p := \sup_{t\in\mathbb{Z}} \mathbb{E}[Z_t^p] < \infty$, $p \geq 2$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[(\widehat{X}_t - X_t)^2\right] \leq \min_{\lambda\in\Lambda}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[(\widehat{X}_t^{(\lambda)} - X_t)^2\right] + \frac{\log(\pi_\lambda^{-1})}{T\eta}\right\} + T\,(8\eta)^{p/2-1} A_*^p (1+L_*)^p m_p\,.$$

Choice of η

For $\pi_\lambda = |\Lambda|^{-1}$, the optimal choice is

$$\eta = \frac{1}{8^{1-2/p}\,(1+L_*)^2\,A_*^2 m_p^{2/p}} \left(\frac{\log|\Lambda|}{T^2}\right)^{2/p},$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \le \inf_{\lambda\in\Lambda}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(\lambda)} - X_t\right)^2\right] + C_1\,\frac{\log(|\Lambda|)^{1-2/p}}{T^{1-4/p}}$$

with $C_1 = 2\times 8^{(p-2)/p}\,(1+L_*)^2\,A_*^2 m_p^{2/p}$.

Risk bound with exponential moment

Theorem (C.G., F. Roueff, A. Sanchez-Perez)

If

1. $\phi(\zeta) := \sup_{t\in\mathbb{Z}} \mathbb{E}[\mathrm{e}^{\zeta Z_t}] < \infty$, *for some* $\zeta > 0$,
2. $\pi_\lambda = |\Lambda|^{-1}$
3. $$\eta = \frac{\zeta^2}{8(A^*)^2(L_* + 1)^2} \max\left\{2, \log\left(\frac{T^2\phi(\zeta)}{8\log N}\right)\right\}^{-2}$$

then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \leq \min_{\lambda\in\Lambda}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\widehat{X}_t^{(\lambda)} - X_t\right)^2\right] + C_2\frac{\log|\Lambda|}{T} \max\left\{2, \log\left(\frac{T^2\phi(\zeta)}{8\log|\Lambda|}\right)\right\}^2$$

with $C_2 = 16A_*^2(1 + L_*)^2\zeta^{-2}$.

Conclusion

Corollary

The estimator $\widehat{X}$ built from the aggregation of the $\widehat{X}^{(\lambda)}$ adapts to the regularity of θ

Caveat

The choice of η depends on a moment of ξ_t (exponential moment or *p*-th moment). Yet, an upper-bound is enough.

$d = 3, T = 1000, \sigma = 1, \delta = 0.5, N = 7$

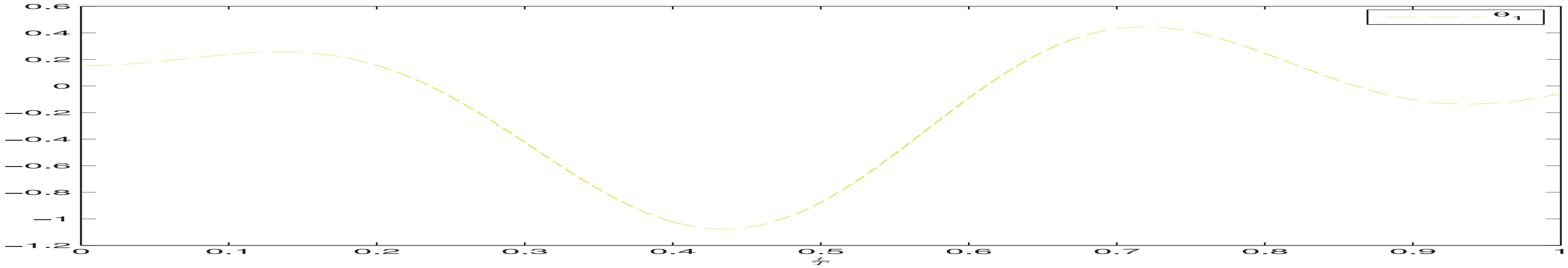


Figure: θ_1

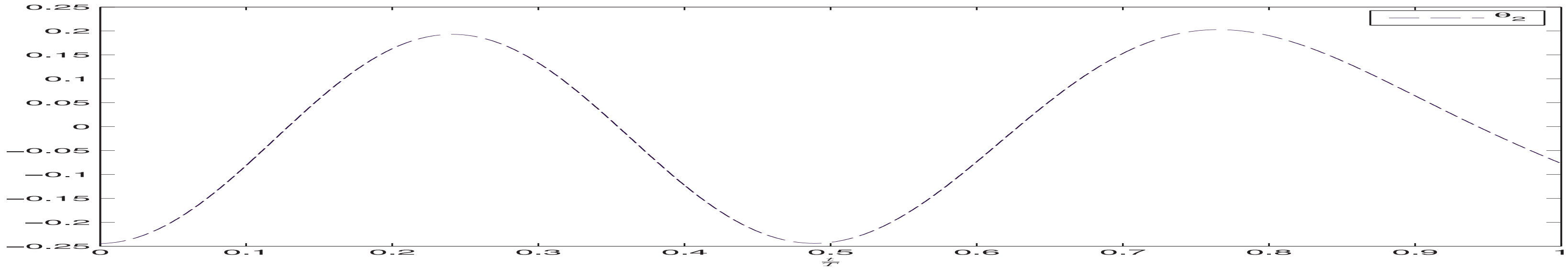


Figure: θ_2

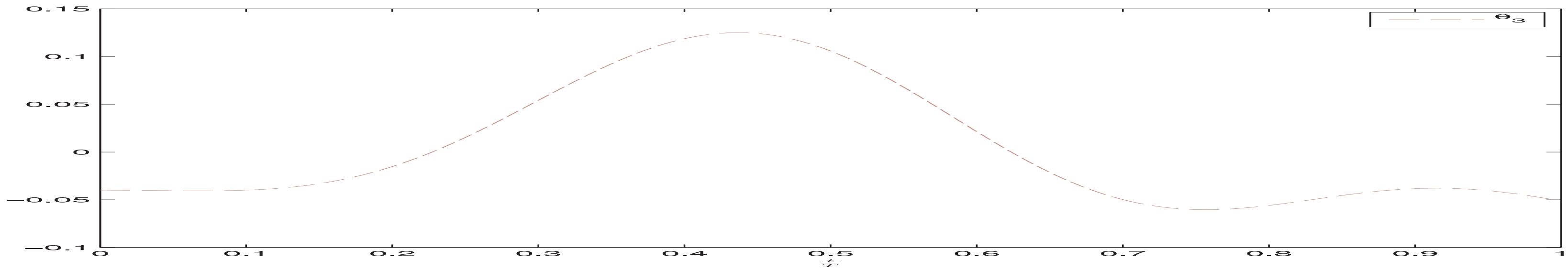


Figure: θ_3

Processes and estimation

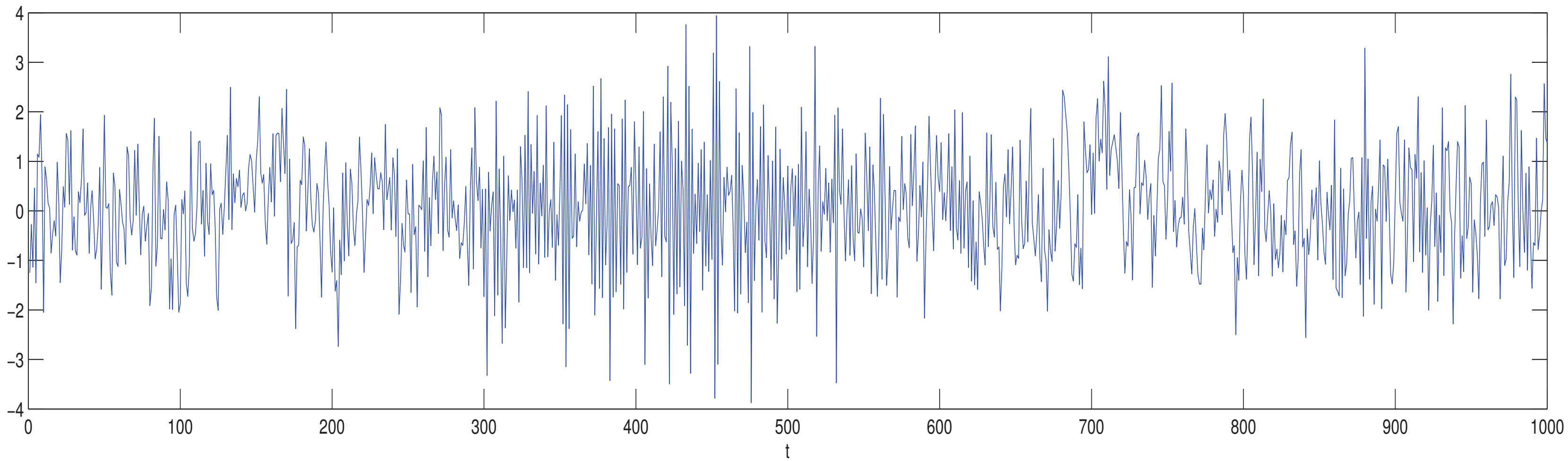


Figure: A TVAR process

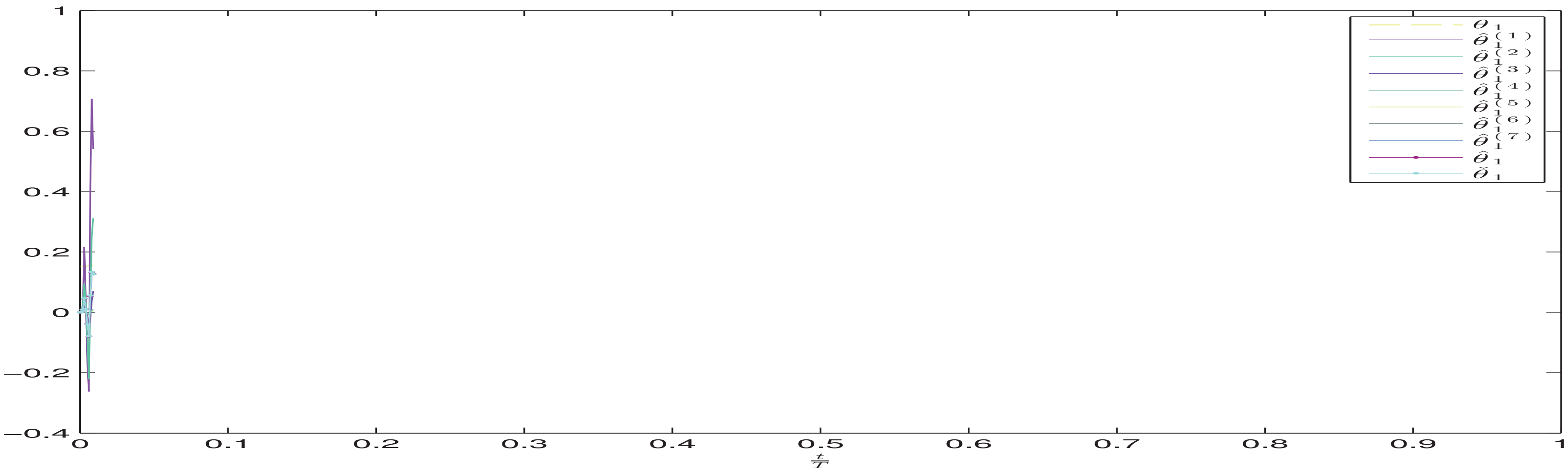


Figure: NLMS estimators for θ_1

Processes and estimation

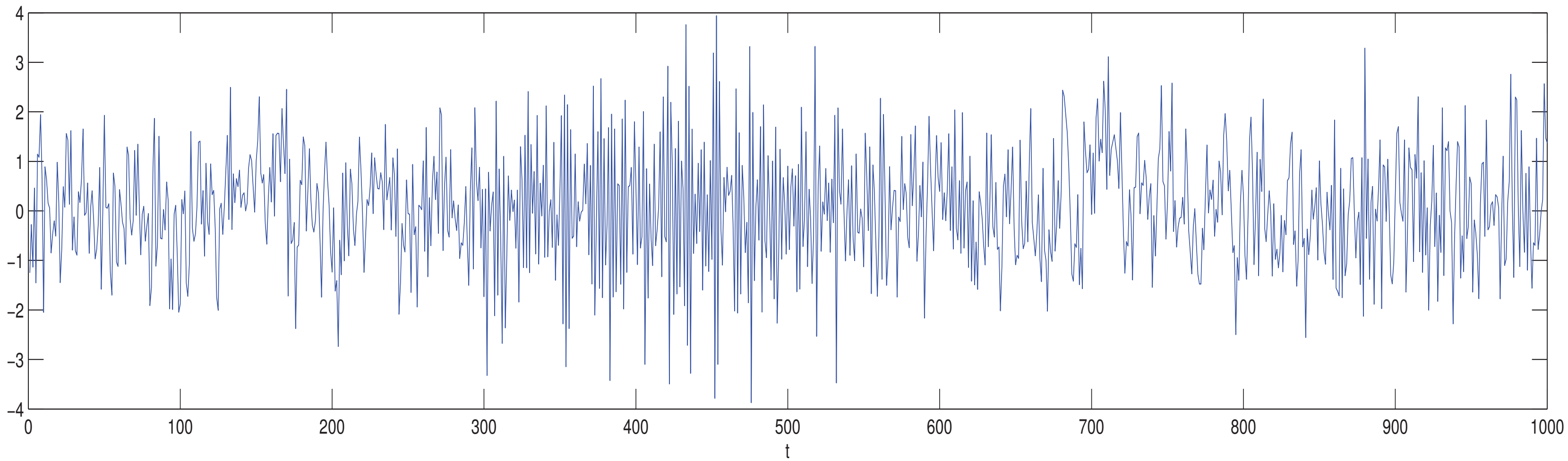


Figure: A TVAR process

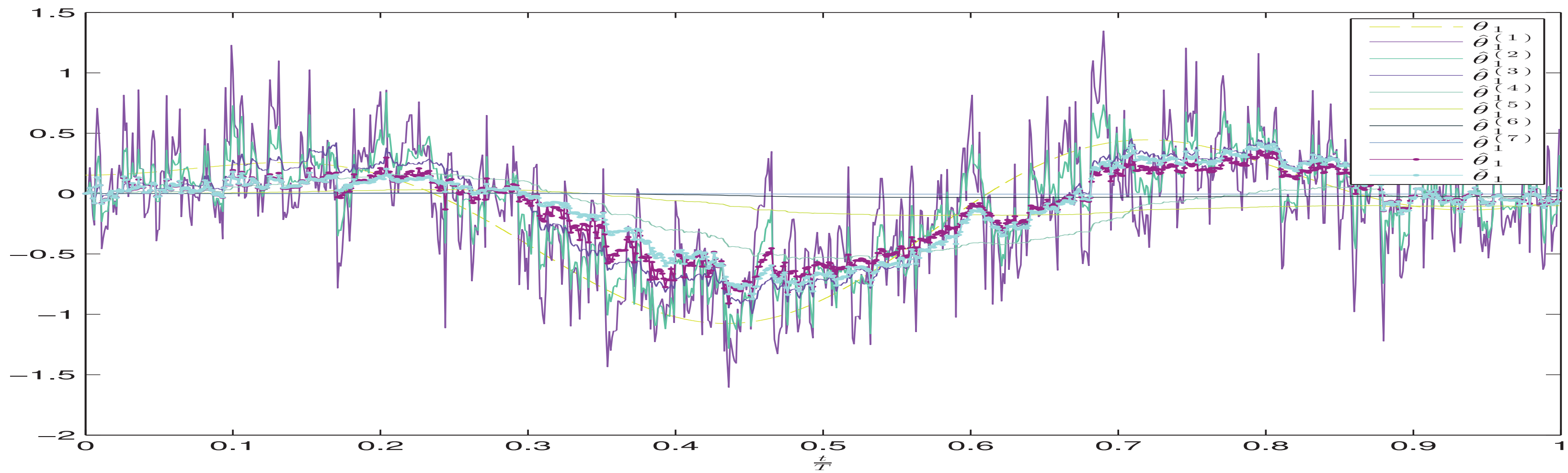


Figure: NLMS estimators for θ_1

Estimation

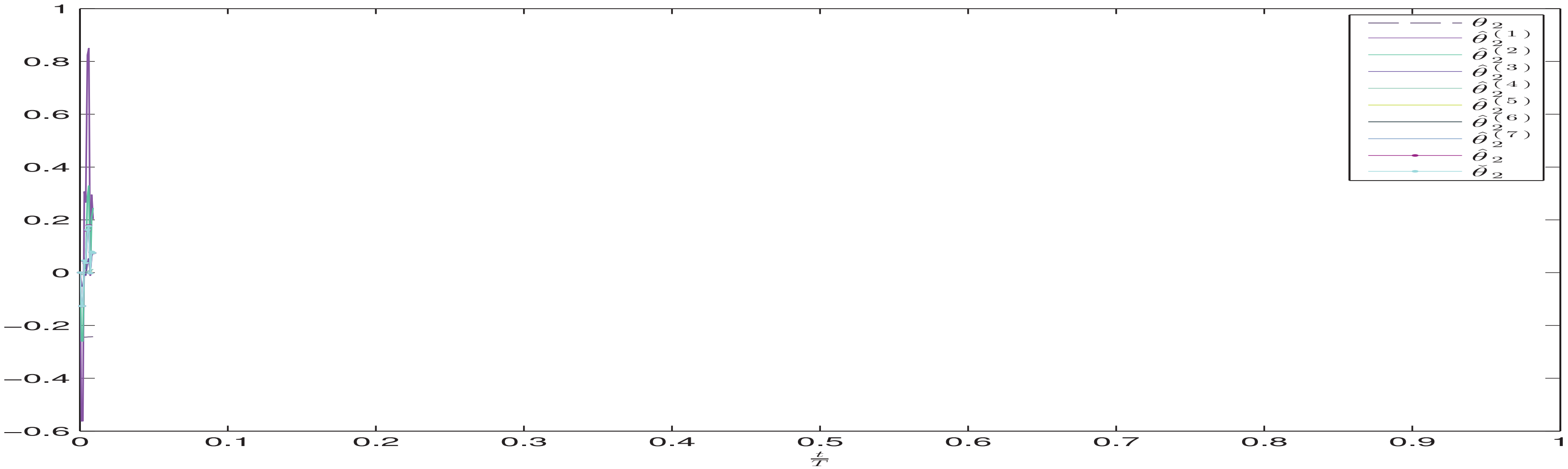


Figure: NLMS estimators for θ_2

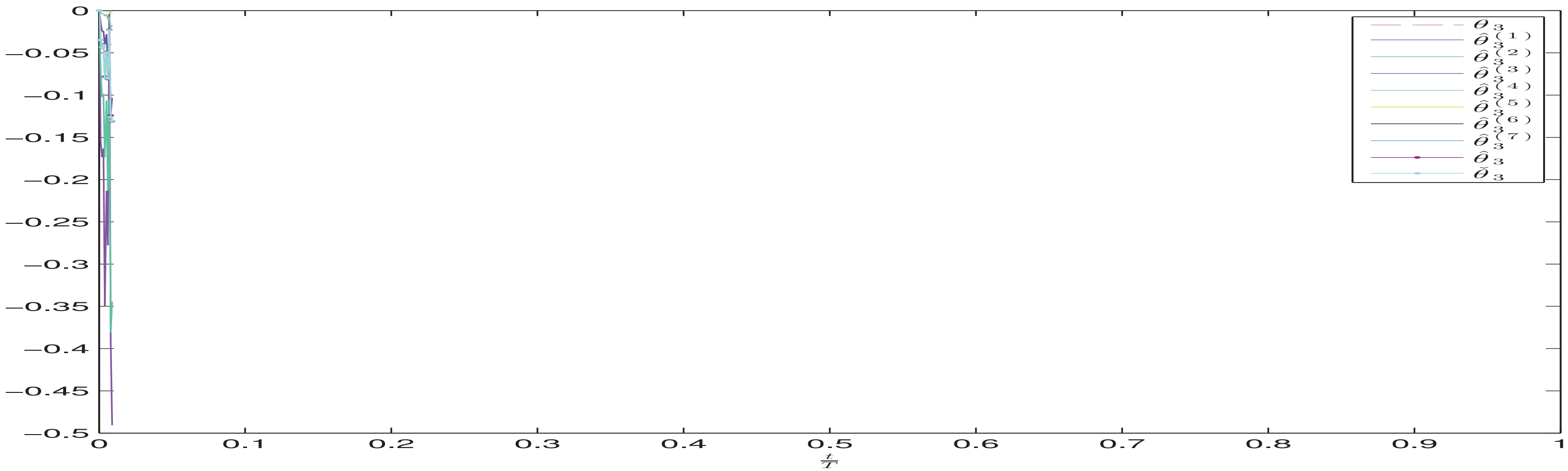


Figure: NLMS estimators for θ_3

Estimation

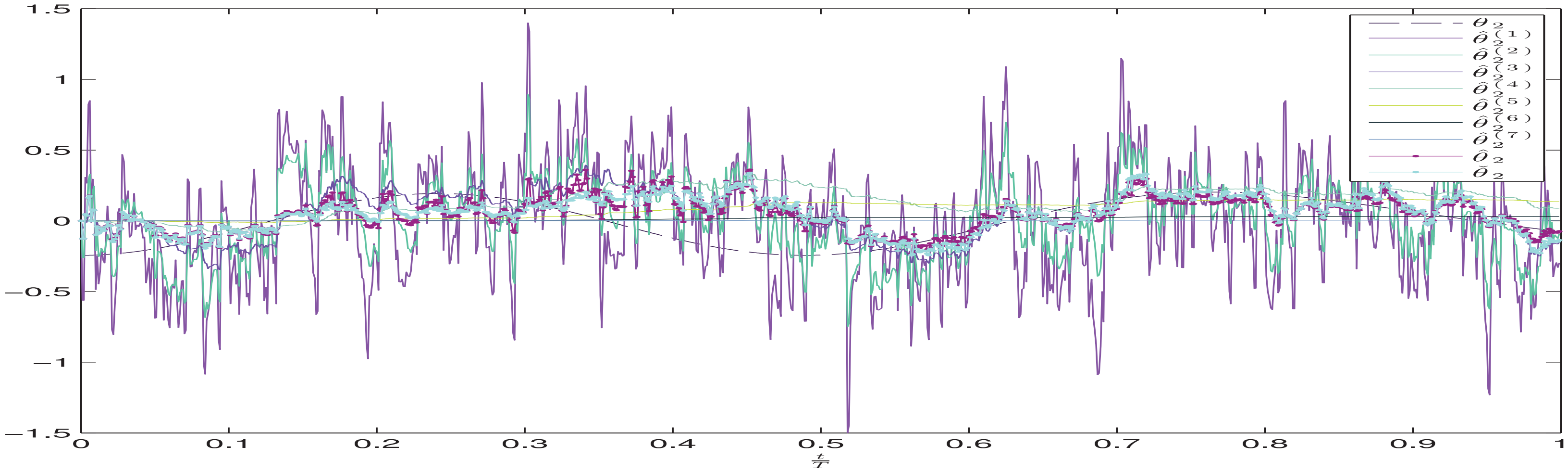


Figure: NLMS estimators for θ_2

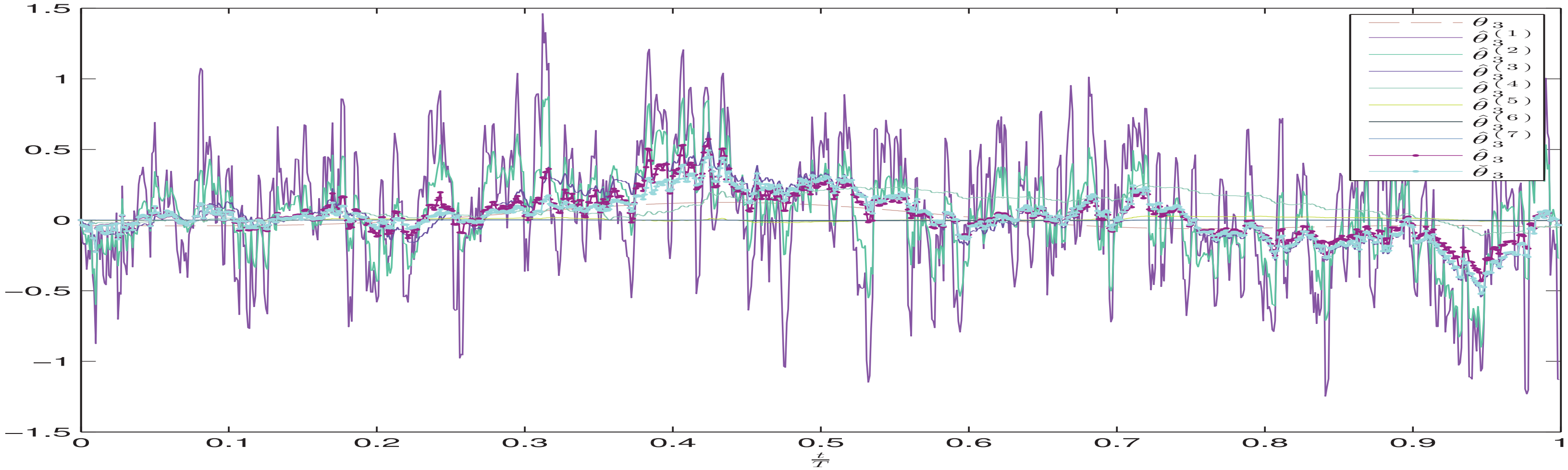


Figure: NLMS estimators for θ_3

Boxplot of the cumulative error

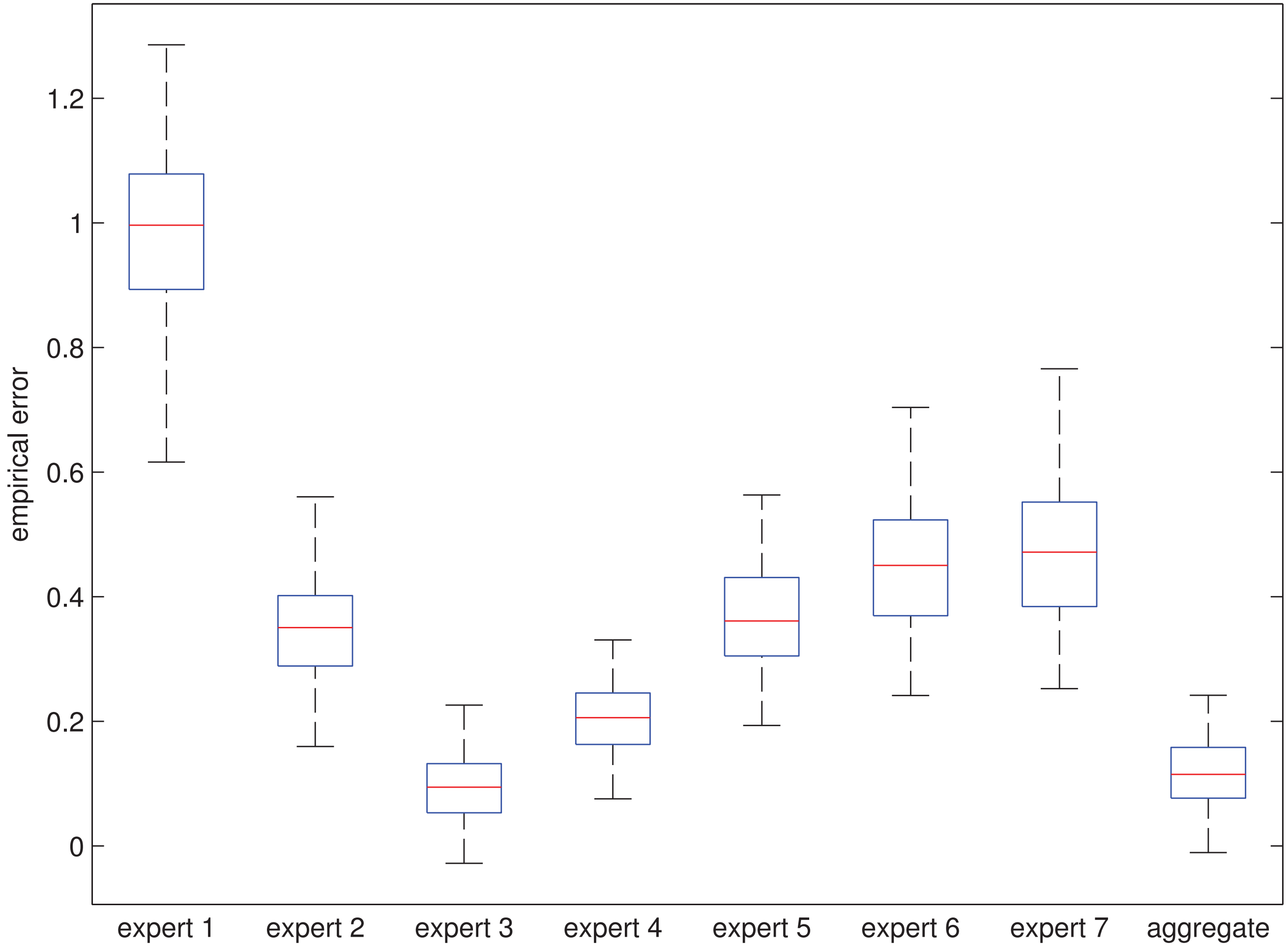


Figure: Empirical cumulative error

Introduction to high-dimensional statistics

Available online :

`http://www.cmap.polytechnique.fr/~giraud/MSV/LectureNotes.pdf`

Contents