Méthode de comparaison pénalisée pour la sélection d'estimateurs

Claire Lacour[1], Pascal Massart[2], Vincent Rivoirard[3], Suzanne Varet[2]

1-Université Paris-Est Marne-la-Vallée
2-Université Paris-Sud
3- Université Paris Dauphine

Outline

Context

Let $X_1, \ldots, X_n$ be data or observations.

We modelize it as random variables in $\mathbb{R}^d$
Examples : images, physical data, biological parameters, etc..

We try to "explain" these data, finding the true distribution,...

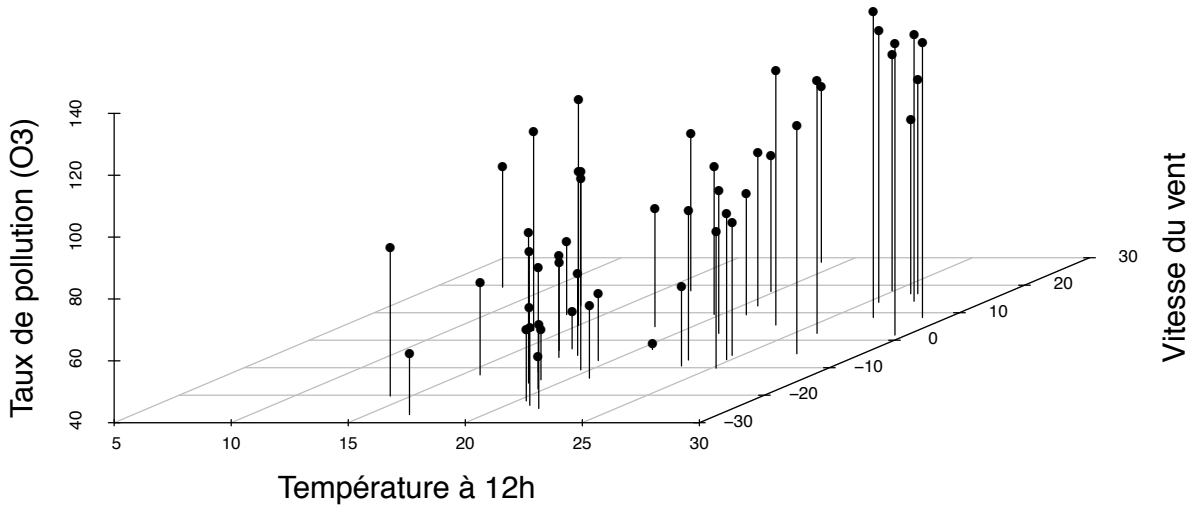
Regression

We observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent, $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$

$$Y_i = f(X_i) + \varepsilon_i \qquad 1 \leq i \leq n$$

where ε_i centered random variables, independent

Example : Y_i ozone pollution rate, X_i weather data (temperature, wind speed, etc.)



Regression

Least squares estimator : f which minimize

$$\min_{f \in S} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

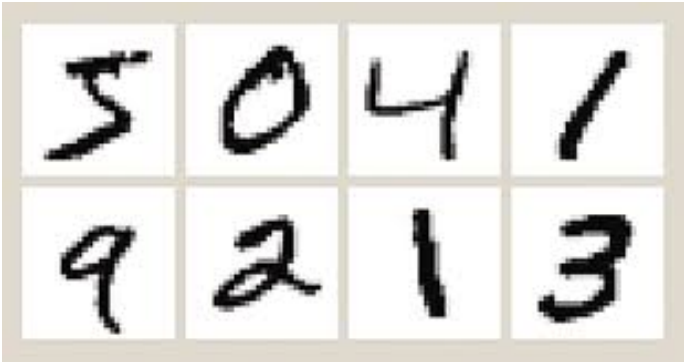
for S a subset of L^2

- $S = \{f(x) = \sum_{j=1}^{p} \theta_j x_j, \theta_j \in \mathbb{R}\}$ classical linear regression
- $S = \{f(x) = \sum_{j \in J} \theta_j x_j, \theta_j \in \mathbb{R}\}$ variable selection, only $k = |J|$ features
- $S = \{\sum_{j=1}^{k} \theta_j \phi_j, \theta_j \in \mathbb{R}\}$ for (ϕ_j) orthonormal basis (Fourier, wavelets, etc.)

Supervised learning

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent with same distribution,
$X \in \mathcal{X}$, $Y \in J$ finite
Aim : to predict which label is associated to a new observation X_{n+1}

Example: spam detection, or character recognition



X image of the handwritten digit
$Y \in \{0, \ldots, 9\}$ the true number

k-nearest neighbors algorithm:
Y_{n+1} is most frequent label among the k nearest neighbors to X_{n+1}.

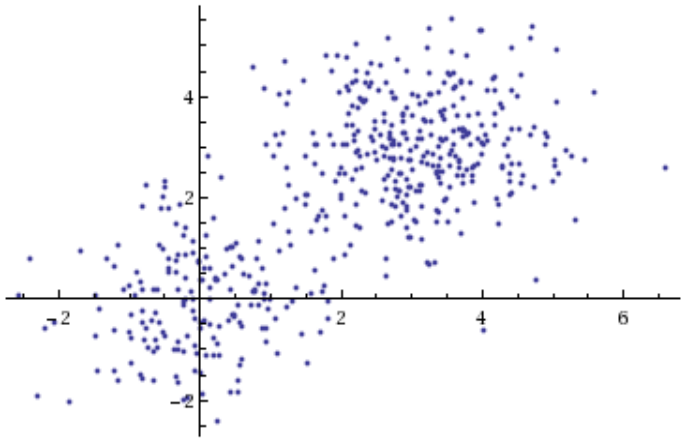
Unsupervised learning

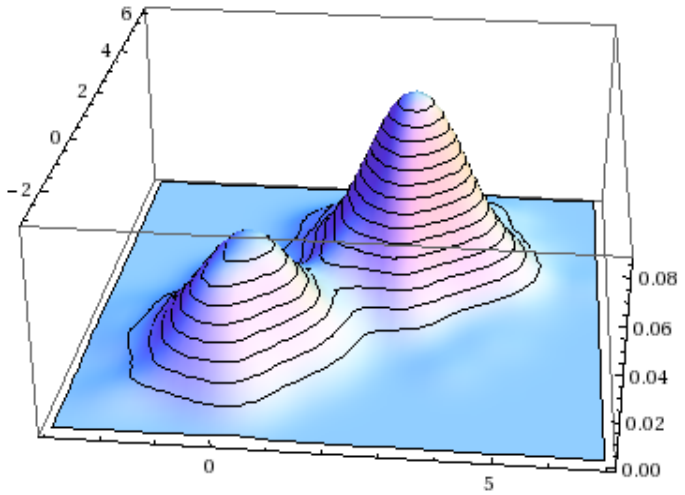
We observe $X_1, \ldots, X_n$ independent and identically distributed (i.i.d.)
Aim: group the data in classes

Methods:
- k-means,

- estimation of the data density (kernel density estimator depends of a bandwidth parameter h)

- ...





Selection methods

There is always one (or more) parameter to choose!

Some methods:

- Minimisation of a penalized contrast criterion
- Cross-validation
- Thresholding of the coefficients for a projection estimator
- Lepski's method
- ...

In all cases, the choice depends on the objective: prevision $\neq$ best model

Statistical framework

$X_1, \ldots, X_n$ i.i.d. real random variables with unknown density f
Given $h > 0$, we can estimate f by

$$\hat{f}(x) = \frac{\text{Card}\left\{i,\, x - h \leq X_i \leq x + h\right\}}{2hn} = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{\{x-h \leq X_i \leq x+h\}}$$

Statistical framework

$X_1, \ldots, X_n$ i.i.d. real random variables with unknown density f
Given $h > 0$, we can estimate f by

$$\hat{f}(x) = \frac{\text{Card}\left\{i,\, x - h \leq X_i \leq x + h\right\}}{2hn} = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{\{x-h \leq X_i \leq x+h\}}$$

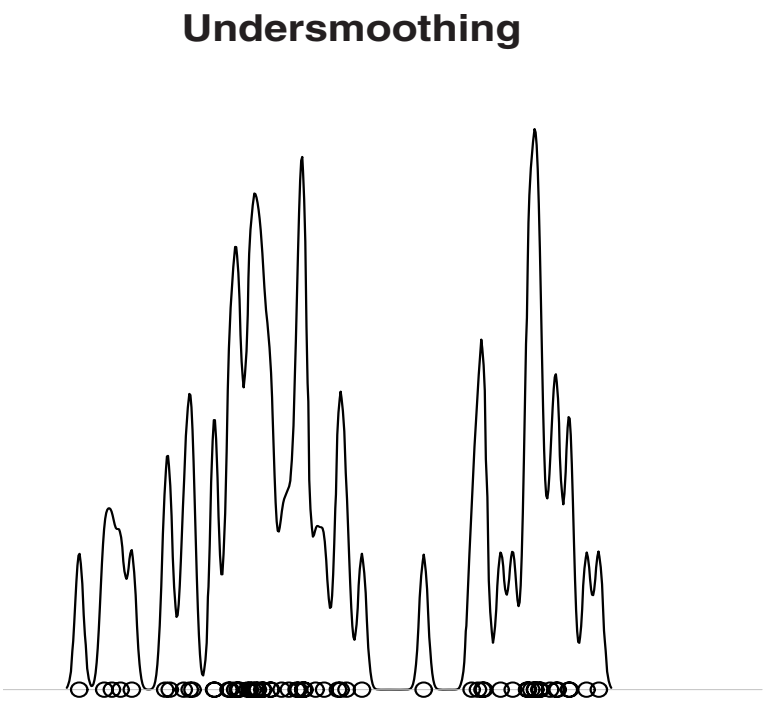
More generally $\hat{f}_h$ classical kernel estimator

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

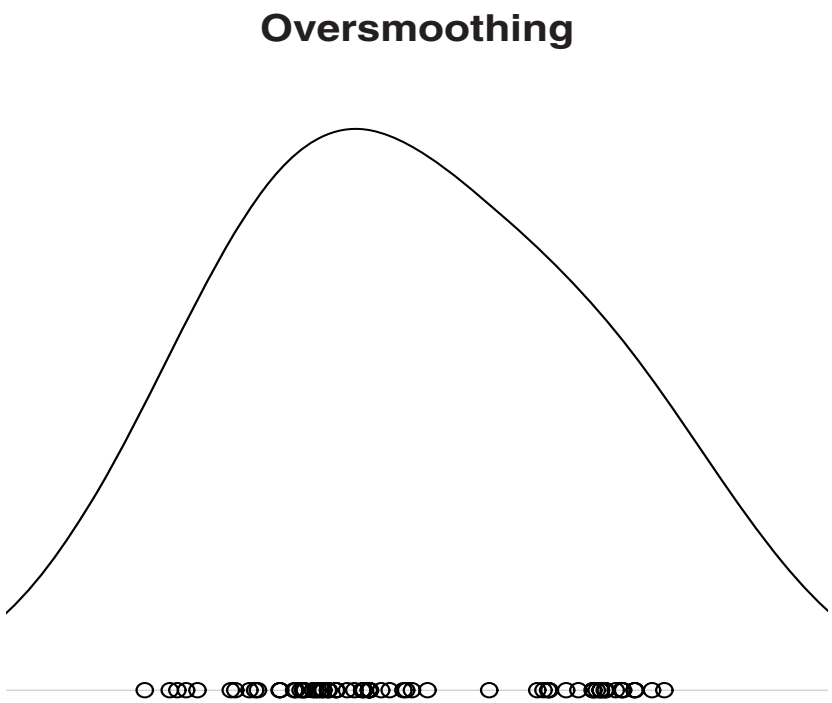
with $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$ and K a given kernel (function with integral 1)

Performance measure: quadratic loss $\|\hat{f}_h - f\|^2$ where $\|.\|$ is the L^2 norm

Bandwidth selection



h too small
(overfitting)



h too large

Bias-variance tradeoff

Let $f_h(x) := \mathbb{E}[\hat{f}_h(x)] = \int K_h(x-u)f(u)du = K_h * f(x)$

Bias-variance tradeoff:

$$\mathbb{E}\|f - \hat{f}_h\|^2 = \underbrace{\|f - f_h\|^2}_{\text{bias}} + \underbrace{\mathbb{E}\|f_h - \hat{f}_h\|^2}_{\text{variance}}$$

Bias-variance tradeoff

Let $f_h(x) := \mathbb{E}[\hat{f}_h(x)] = \int K_h(x-u)f(u)du = K_h * f(x)$

Bias-variance tradeoff:

$$\mathbb{E}\|f - \hat{f}_h\|^2 = \|f - f_h\|^2 + \mathbb{E}\|f_h - \hat{f}_h\|^2$$

bias $\searrow$ when $h \searrow$ — variance $\nearrow$ when $h \searrow$

Bias Depends of the smoothness of f
If f C^β, $\|f - f_h\| \leq Ch^\beta$

Variance $\mathbb{E}\|f_h - \hat{f}_h\|^2 \approx \frac{\|K_h\|^2}{n} = \frac{\|K\|^2}{nh}$

Bias-variance tradeoff

Let $f_h(x) := \mathbb{E}[\hat{f}_h(x)] = \int K_h(x-u)f(u)du = K_h * f(x)$

Bias-variance tradeoff:

$$\mathbb{E}\|f - \hat{f}_h\|^2 = \|f - f_h\|^2 + \mathbb{E}\|f_h - \hat{f}_h\|^2$$

bias $\searrow$ when $h \searrow$ — variance $\nearrow$ when $h \searrow$

Bias Depends of the smoothness of f
If f C^β, $\|f - f_h\| \leq Ch^\beta$

Variance $\mathbb{E}\|f_h - \hat{f}_h\|^2 \approx \frac{\|K_h\|^2}{n} = \frac{\|K\|^2}{nh}$

Best (theoretical choice) $h \approx n^{-1/(2\beta+1)}$, with β unknown
$\hookrightarrow$ we need a data-driven choice

Bias estimation

Bias-variance tradeoff:

$$\mathbb{E}\|\hat{f}_h - f\|^2 = \|f - f_h\|^2 + \mathbb{E}\|f_h - \hat{f}_h\|^2 \approx \underbrace{\|f_h - f\|^2}_{B^2(h)} + \underbrace{\frac{\|K_h\|^2}{n}}_{V(h)}$$

Idea to minimise $\mathbb{E}\|\hat{f}_h - f\|^2$: estimator $\hat{B}^2(h)$ of $B^2(h)$ and then

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \{\hat{B}^2(h) + V(h)\}$$

where $\mathcal{H}$ finite set of bandwidths

Some heuristics (1/2)

Denote $\mathcal{H} = \{h_{\min}, \ldots, h_{\max}\}$

$B^2(h) = \|f_h - f\|^2 \approx \|f_h - f_{h_{\min}}\|^2$
(if $h_{\min} \to 0$)

and

$$\mathbb{E}\|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 \quad \approx \quad \|f_h - f_{h_{\min}}\|^2 + \frac{\|K_h - K_{h_{\min}}\|^2}{n}$$

Some heuristics(2/2)

$$B^2(h) \approx \|f_h - f_{h_{\min}}\|^2 \approx \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 - \frac{\|K_h - K_{h_{\min}}\|^2}{n}$$

To minimize $\left\{B^2(h) + V(h)\right\}$ comes down to minimize

$$\|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \underbrace{\hat{V}(h) - \frac{\|K_{h_{\min}} - K_h\|^2}{n}}_{\text{pen}(h)}$$

Method: Penalized Comparison to Overfitting

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \left\{ \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \text{pen}(h) \right\}$$

where $\mathcal{H} = \{h_{\min}, \ldots, h_{\max}\}$

$\hat{f}_{h_{\min}}$: overfitting estimator

and $\text{pen}_\lambda(h) = \dfrac{\lambda \|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n}$

Link with other methods

PCO method

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \; \{\|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \text{pen}(h)\}$$

- $\hat{f}_{h_{\min}}(x) \longrightarrow \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(x)$ when $h_{\min} \to 0$
  $\hat{h} \approx \underset{h \in \mathcal{H}}{\text{argmin}} \; \{\|\hat{f}_h\|^2 - \frac{2}{n}\sum_{i=1}^n \hat{f}_h(X_i) + \|\hat{f}_{h_{\min}}\|^2 + \text{pen}(h)\}$

  penalized least-squares contrast for density estimation
- Goldenshluger-Lepski method

$$\begin{cases} \hat{B}^2(h) = \sup_{h' \in \mathcal{H}} \left[\|\hat{f}_{h \vee h'} - \hat{f}_{h'}\|^2 - \text{pen}_1(h')\right]_+ \\ \hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \; \{\hat{B}^2(h) + \text{pen}_2(h)\} \end{cases}$$

  degenerate case when $\text{pen}_1 \equiv 0$
- Link with regression: $\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \; \{\|\hat{f}_h - Y\|_n^2 + \text{pen}(h)\}$

Oracle inequality

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{argmin}} \left\{ \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \frac{\|K_h\|^2 - \|K_{h_{\min}} - K_h\|^2}{n} \right\}$$

Theorem (Lacour, Massart, Rivoirard)

Assume $\|f\|_\infty < \infty$ *and* $h_{\min} \geq \|K\|_\infty \|K\|_1 / n$. *Let* $\epsilon \in (0,1)$.
$\forall x > 0$, *with probability* $1 - C_1 |\mathcal{H}| e^{-x}$

$$\|\hat{f}_{\hat{h}} - f\|^2 \leq (1+\epsilon) \min_{h \in \mathcal{H}} \|\hat{f}_h - f\|^2 + C_2(\epsilon) \|f_{h_{\min}} - f\|^2 + C_3(K, \epsilon) \frac{\|f\|_\infty x^3}{n}$$

Proof based on concentration inequalities

Consequences

- If f has smoothness β (Nikolskii or Besov): optimal rate of convergence $n^{-\beta/(2\beta+1)}$
- Simple to implement, less comparisons than for Lepski method: numerically faster, in particular for multivariate case
- Minimal penalty different from the optimal: good news for calibration

Multivariate case

$X_1, \dots, X_n$ i.i.d. random variables in $\mathbb{R}^d$ with unknown density $f : \mathbb{R}^d \to \mathbb{R}$

$$\hat{f}_H(x) = \frac{1}{n \det H} \sum_{i=1}^{n} K(H^{-1}(x - X_i))$$

with H a symmetric positive-definite $d \times d$ matrix

$$\hat{H} = \underset{H \in \mathcal{H}}{\text{argmin}} \left\{ \|\hat{f}_H - \hat{f}_{H_{\min}}\|^2 + \frac{\|K_H\|^2 - \|K_{H_{\min}} - K_H\|^2}{n} \right\}$$

with $H_{\min}$ diagonal with minimal eigenvalues

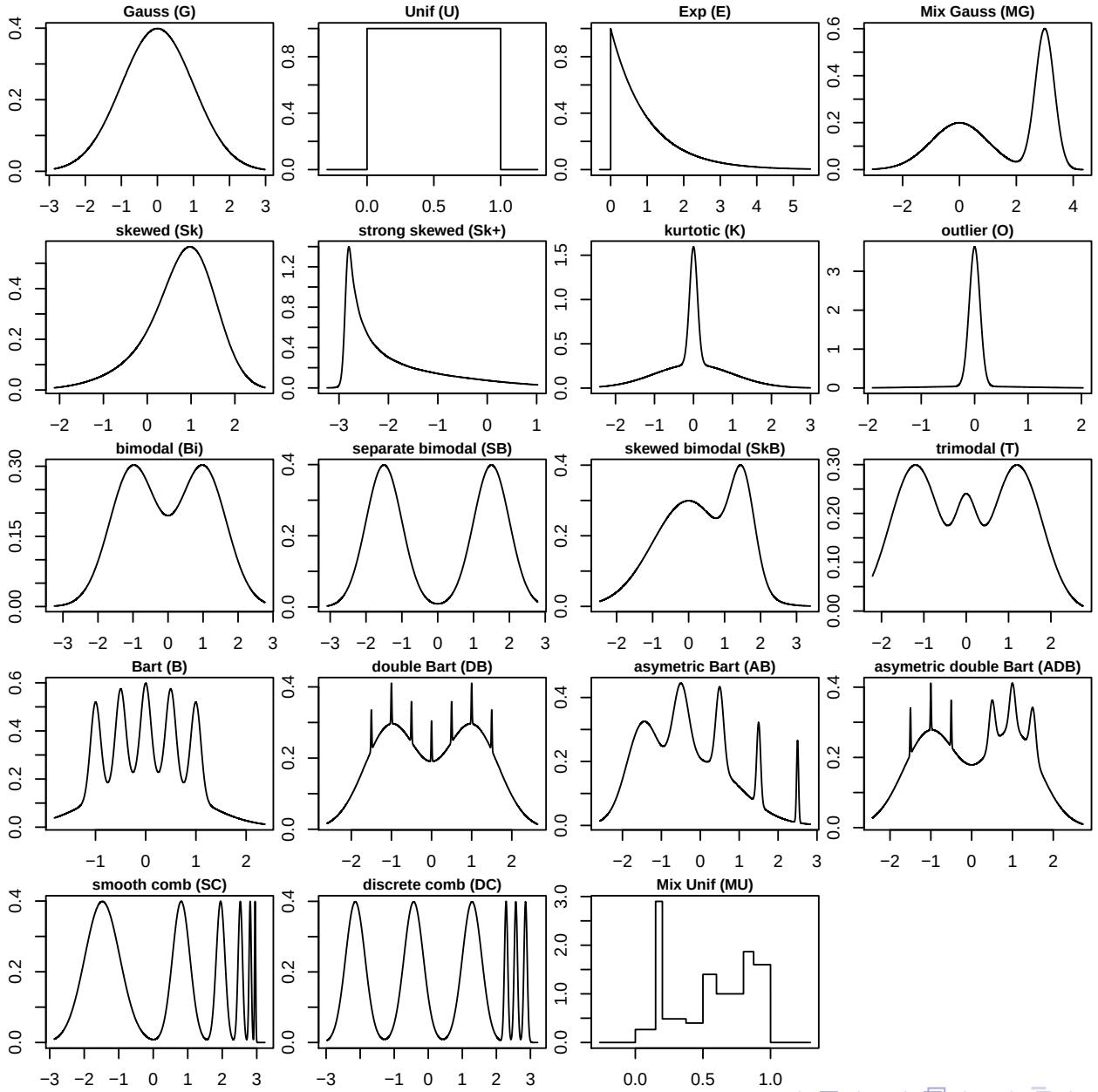
Theorem (Lacour, Massart, Rivoirard)

Assume $\|f\|_\infty < \infty$ *and* $\det(H_{\min}) \geq \|K\|_\infty \|K\|_1 / n$. *Let* $\epsilon \in (0, 1)$.
$\forall x > 0$, *with probability* $1 - C_1 |\mathcal{H}| e^{-x}$

$$\|\hat{f}_{\hat{H}} - f\|^2 \leq (1 + \epsilon) \min_{H \in \mathcal{H}} \|\hat{f}_H - f\|^2 + C_2 \|f_{H_{\min}} - f\|^2 + C_3 \frac{\|f\|_\infty x^3}{n}$$

$\hookrightarrow$ *allows anisotropic rates of convergence*

Testing densities



Comparison with other methods

$\mathcal{H}$: 400 bandwiths between $1/n$ and 1

For various other estimation methods, we compute, for each sample,

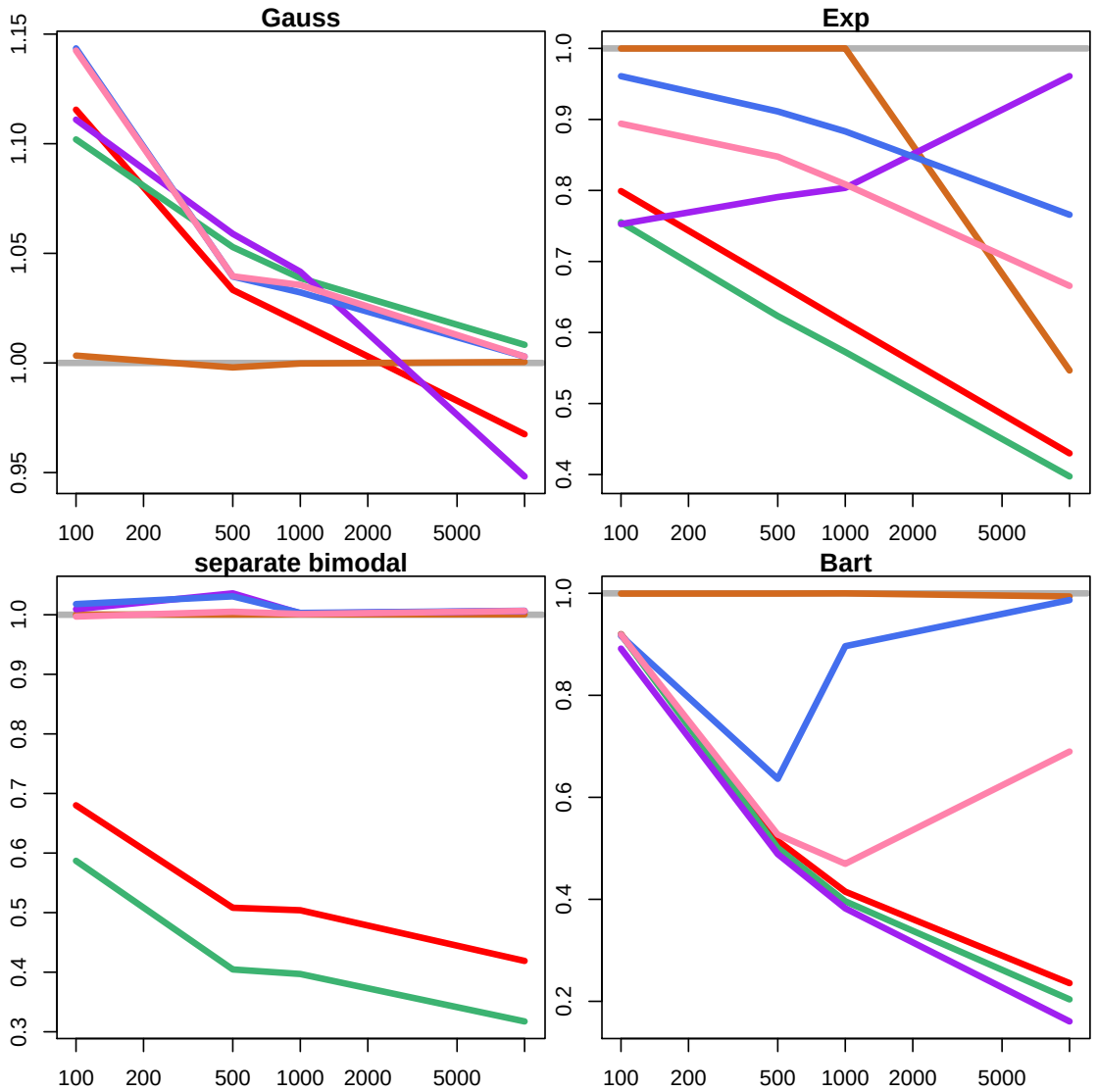
$$\frac{\|\hat{f}^{PCO} - f\|}{\|\hat{f}^{Method} - f\|}$$

and we plot the median over 20 samples

Other methods:

- **RoT**: Silverman's rule of thumb
- **BCV**: biased cross-validation
- **UCV**: unbiased cross-validation
- **PI**: Plug-in (Sheather and Jones, 1991), direct plug-in (PIdpi) or solve-the-equation (PIste)

Comparison with other methods

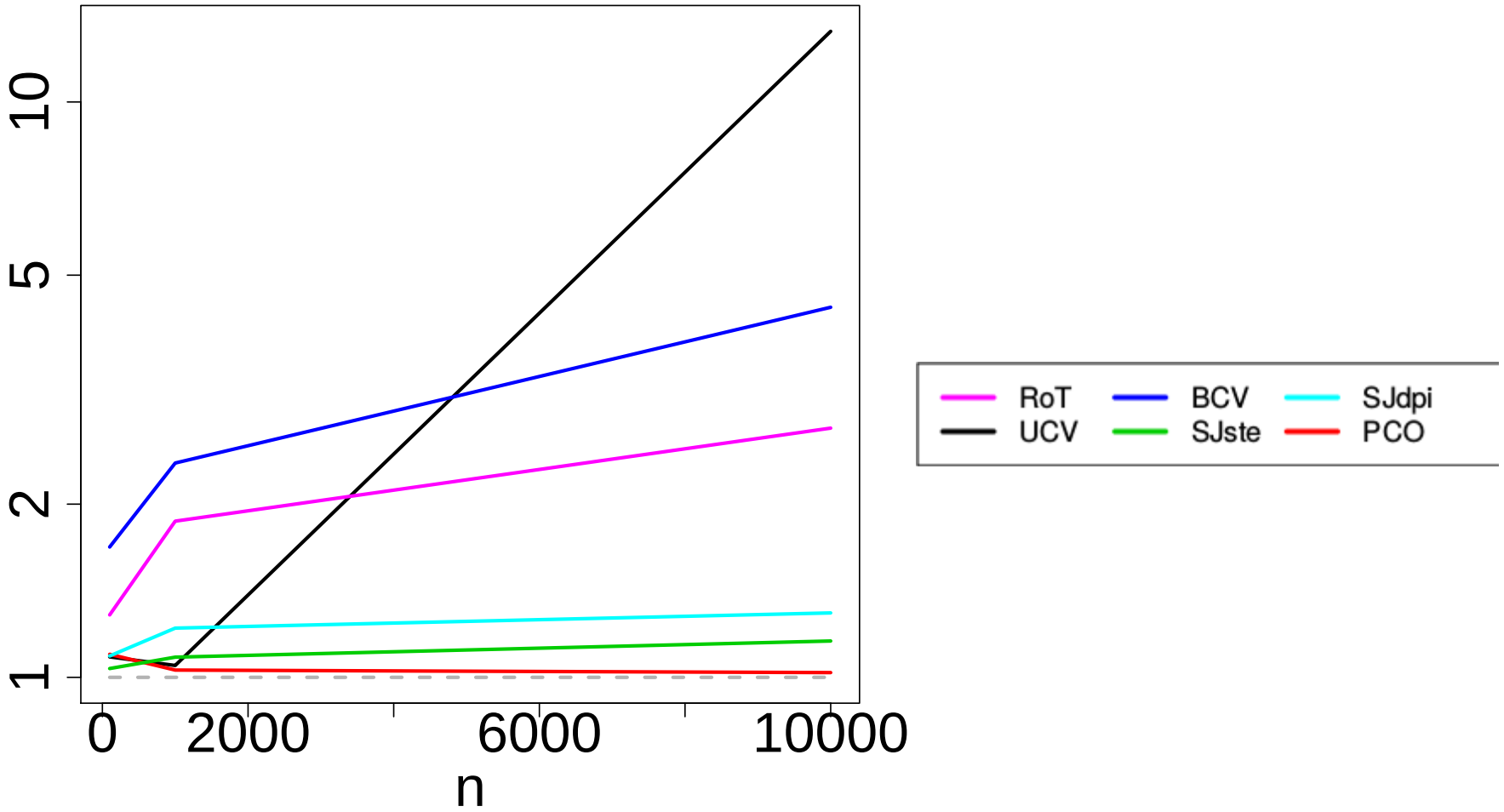


median $\left[\frac{ISE^{1/2}_{\text{PCO}}}{ISE^{1/2}_{\text{Meth}}}\right]$ versus n

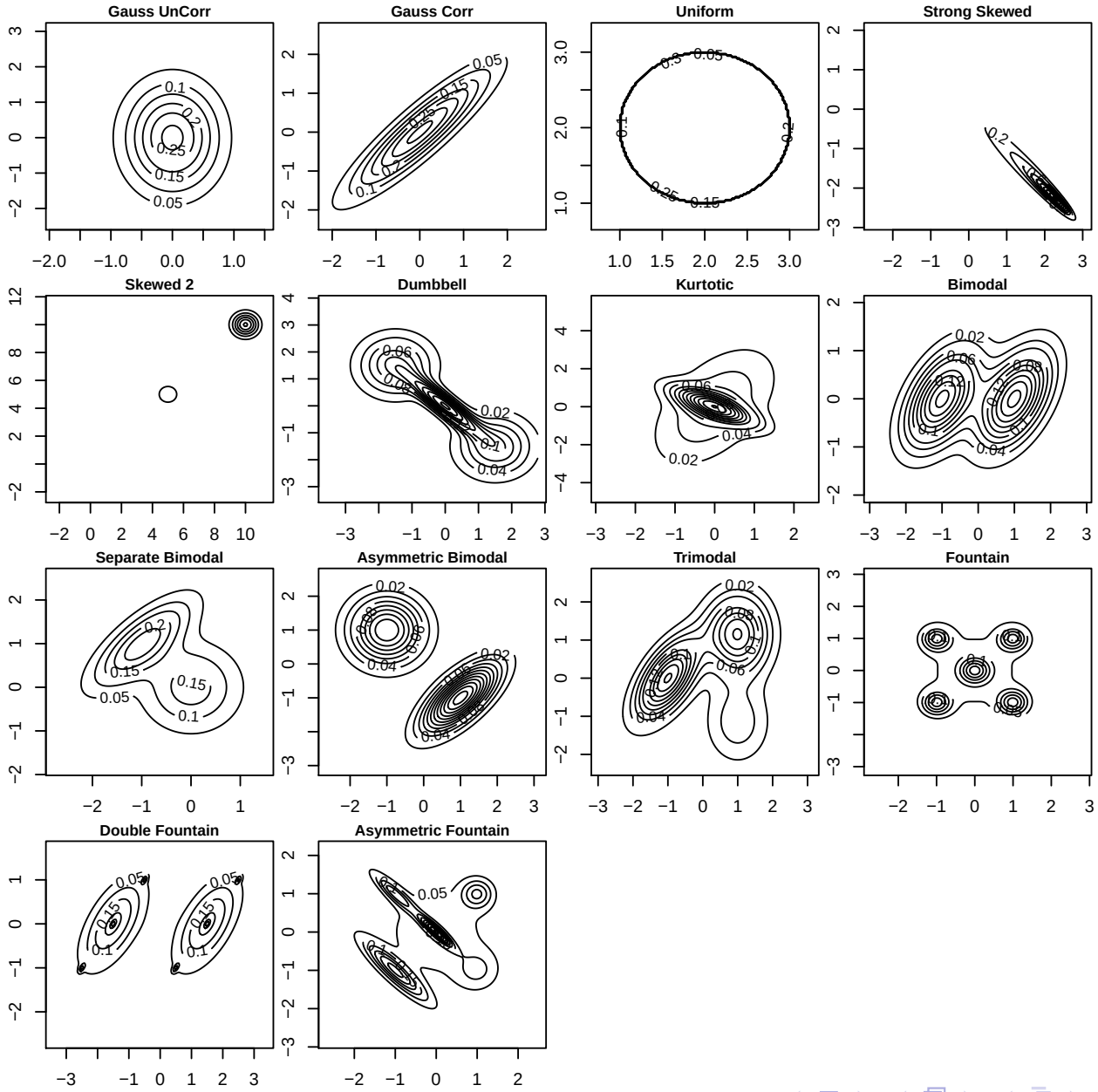
Rule of thumb 0
Rule of thumb 1
UCV
BCV
Plste
Pldpi

Comparison to the best method

Mean over 19 densities f of the ratio $\dfrac{\tilde{\mathbb{E}}\|\hat{f}^{\text{meth}} - f\|}{\min_{\text{meth}} \tilde{\mathbb{E}}\|\hat{f}^{\text{meth}} - f\|}$



Multivariate case: testing densities in dimension $d = 2$



Multivariate case: dimension $d = 2, 3, 4$

$\mathcal{H}$: 16^d bandwiths in $[(\|K\|_\infty / n^{1/d}), 1]^d$

Other methods:

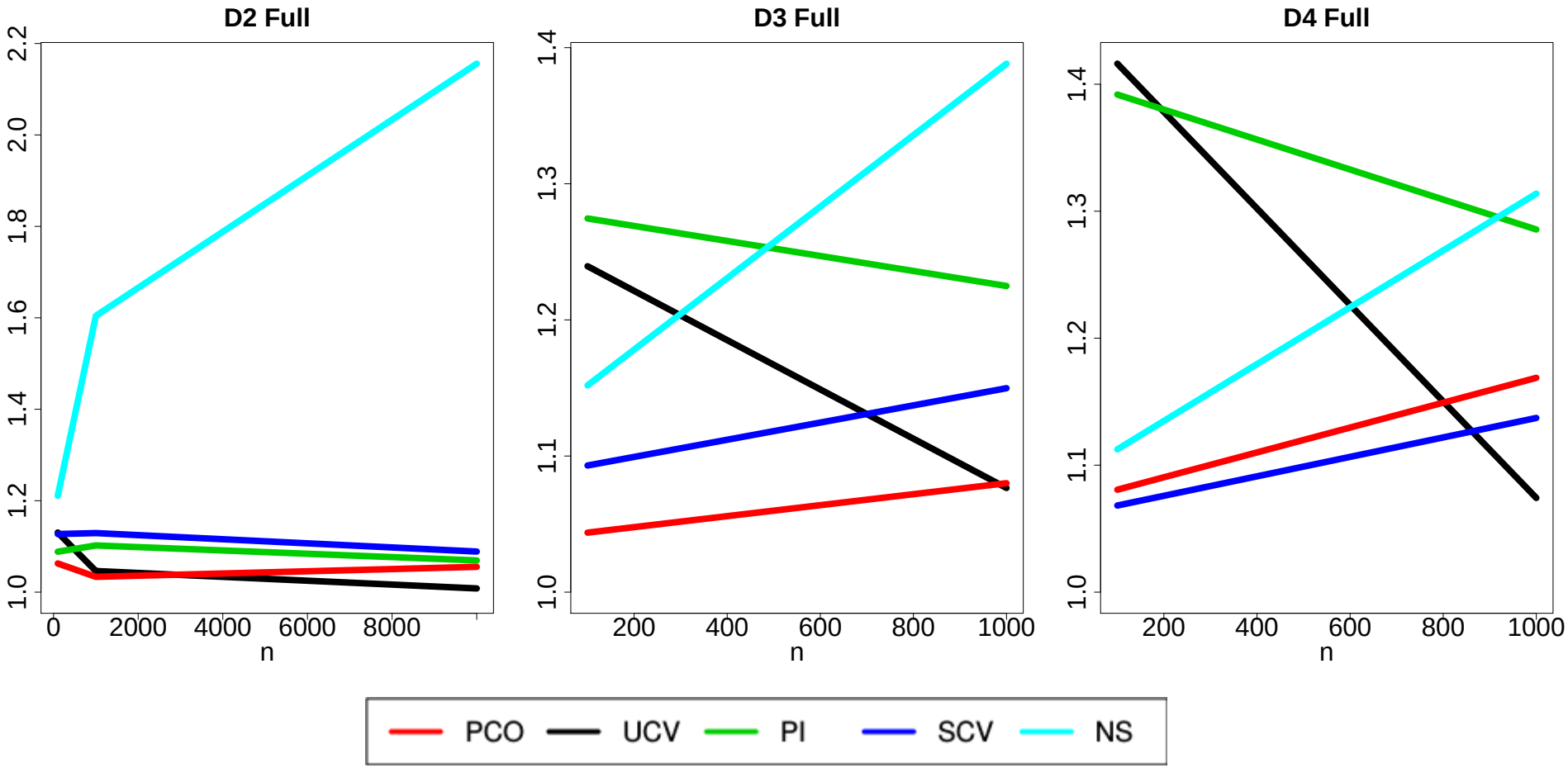
- **NS**: Normal scale = Silverman's rule of thumb
- **SCV**: smooth cross-validation
- **UCV**: unbiased cross-validation
- **PI**: Plug-in

For these methods, we compute again

$$\frac{\tilde{\mathbb{E}} \| \hat{f}^{\text{meth}} - f \|}{\min_{\text{meth}} \tilde{\mathbb{E}} \| \hat{f}^{\text{meth}} - f \|}$$

and we plot the mean over 14 testing densities

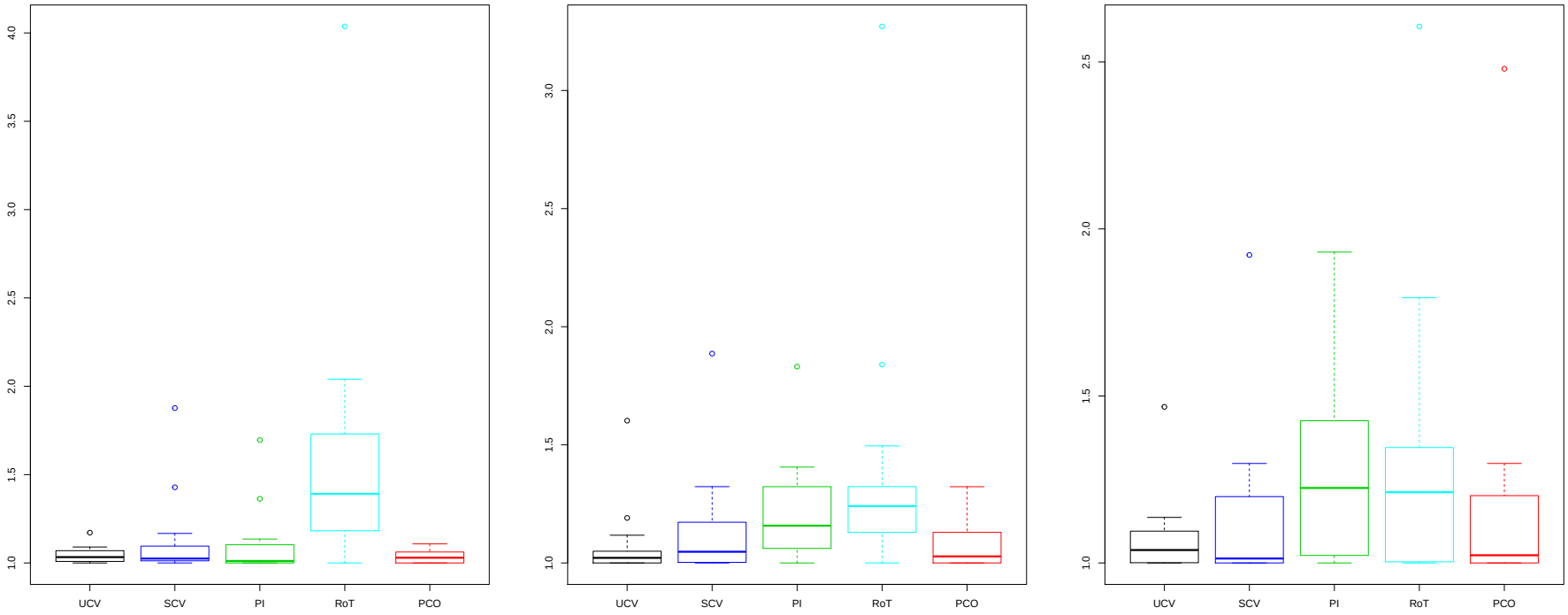
Multivariate comparison



Multivariate comparison, $n = 1000$

Boxplots of the ratio $\dfrac{\tilde{\mathbb{E}}\|\hat{f}^{\text{meth}} - f\|}{\min_{\text{meth}} \tilde{\mathbb{E}}\|\hat{f}^{\text{meth}} - f\|}$ over the 14 densities

$d = 2$ $\qquad$ $d = 3$ $\qquad$ $d = 4$



Can we generalize?

Let f be the target in a space endowed with distance d. Let $(\hat{f}_h)_{h\in\mathcal{H}}$ be a family of estimators, with $\hat{f}_{h_{\min}}$ "the closest of the data"

Selection rule: Penalized Comparison to Overfitting

$$\hat{h} = \underset{h\in\mathcal{H}}{\text{argmin}} \{d(\hat{f}_h, \hat{f}_{h_{\min}}) + \text{pen}(h)\}$$

Other models, other losses?
No result yet for other losses, but PCO for other models exists

Other models

- F. Comte, C. Prieur, A. Samson

Stochastic Processes and their Applications
Volume 127, Issue 11, November 2017, Pages 3689-3718

Adaptive estimation for stochastic damping Hamiltonian systems under partial observation

Fabienne Comte [a] ✉, Clémentine Prieur [b] ✉, Adeline Samson [c] ✉

Show more

https://doi.org/10.1016/j.spa.2017.03.011

Get rights and content

Abstract

The paper considers a process $Z_t = (X_t, Y_t)$ where X_t is the position of a particle and Y_t its velocity, driven by a hypoelliptic bi-dimensional stochastic differential equation. Under adequate conditions, the process is stationary and geometrically β-mixing. In this context, we propose an adaptive non-parametric kernel estimator of the stationary density p of Z, based on n discrete time observations with time step δ. Two observation schemes are considered: in the first one, Z is the observed process, in the second one, only X is measured. Estimators are proposed in both settings and upper risk bounds of the mean integrated squared error (MISE) are proved and discussed in each case, the second one being more difficult than the first one. We propose a data driven bandwidth selection procedure based on the Goldenshluger and Lespki (2011) method. In both cases of complete and partial observations, we can prove a bound on the MISE asserting the adaptivity of the estimator. In practice, we take advantage of a very recent improvement of the Goldenshluger and Lespki (2011) method provided by Lacour et al. (2016), which is computationally efficient and easy to calibrate. We obtain convincing simulation results in both observation contexts.

numerical use

Other models

- Luc Lehéricy

State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models

Luc Lehéricy LUC.LEHERICY@MATH.U-PSUD.FR
Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France

(...)

An advantage of this algorithm is its lower complexity, since it requires $O(M_{\max})$ computations of $\mathbf{L}^2$ norms instead of $O(M_{\max}^2)$. We do not study this method theoretically in our setting. However, the simulations (and in particular Figure 4) show that it behaves similarly to the standard state-by-state selection method in the asymptotic regime and even has a smaller error for small number of observations. In addition, the dimension jumps are much sharper for this method than for the usual state-by-state selection method (see Figure 2), which makes the calibration heuristics easier to use.

numerical use

Other models

- Thanh Mai Pham Ngoc

Adaptive kernel density estimation for directional data

Thanh Mai Pham Ngoc *

June 18, 2018

Abstract

We focus on the nonparametric density estimation problem with directional data. We propose a new rule for bandwidth selection for kernel density estimation. Our procedure is automatic and fully data driven. We obtain an oracle inequality and optimal rates of convergence for the $\mathbb{L}_2$ error. Our theoretical results are illustrated with simulations.

AMS 2000 subject classification Primary G2G07
Keywords and phrases: Bandwidth selection, directional data, kernel density estimation, concentration inequalities, penalization methods, oracle inequality

numerical and theoretical results

Other models

- Thomas Deschatre

Chapter 4

Local polynomial estimation of a doubly stochastic Poisson process

numerical and theoretical results

Abstract

We consider a doubly stochastic Poisson process with stochastic intensity $\lambda_t = nq(X_t)$ where X is a continuous Itô semimartingale and n is an integer. Both processes are observed continuously over a fixed period $[0, T]$. An estimation procedure is proposed in a non parametrical setting for the function q on an interval I where X is sufficiently observed using a local polynomial estimator. A method to select the bandwidth in a non asymptotic framework is proposed, leading to an oracle inequality. If m is the degree of the chosen polynomial, the accuracy of our estimator over the Hölder class of order β is $n^{\frac{-\beta}{2\beta+1}}$ if $m \geq \lfloor\beta\rfloor$ and $n^{\frac{-m}{2m+1}}$ if $m < \lfloor\beta\rfloor$ and is optimal in the minimax sense if $m \geq \lfloor\beta\rfloor$. A parametrical test is also proposed to test if q belongs to some parametrical family. Those results are applied to French temperature and electricity spot prices data where we infer the intensity of electricity spot spikes as a function of the temperature.

Thank you for your attention!