On the use of overfitting for estimator selection in multivariate density estimation

Claire Lacour

Université Gustave Eiffel (Paris East)

Joint work with

V. Rivoirard, P. Massart, S. Varet

Multivariate density estimation

- We consider an n-sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with $\mathbf{X}_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d$. We denote by $f : \mathbb{R}^d \longmapsto \mathbb{R}_+$ the density of the $\mathbf{X}_i$'s to be estimated.
- We consider K a bounded kernel function, so that $K \in \mathbb{L}_1$ and it satisfies

$$\int_{\mathbb{R}^d} K(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$$

- The kernel density estimator $\widehat{f}_H$ is given, for all $\mathbf{x} \in \mathbb{R}^d$, by

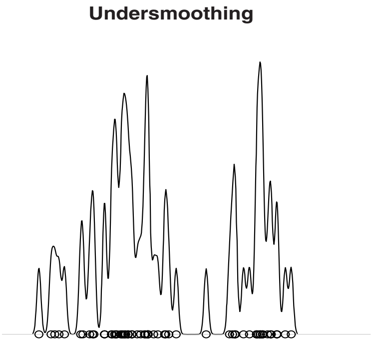
$$\widehat{f}_H(\mathbf{x}) = \frac{1}{n\det(H)} \sum_{i=1}^{n} K\left(H^{-1}(\mathbf{x} - \mathbf{X}_i)\right) = \frac{1}{n} \sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{X}_i)$$

  where the matrix H is the kernel bandwidth belonging to a fixed grid $\mathcal{H}$ of invertible matrices and

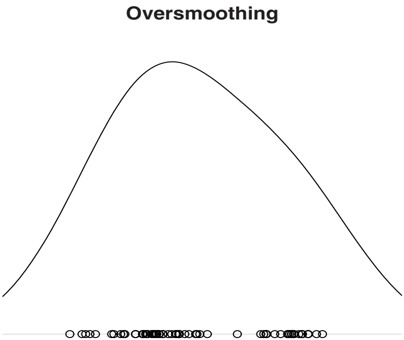
$$K_H(\mathbf{x}) = \frac{1}{\det(H)} K\left(H^{-1}\mathbf{x}\right)$$

- One of main critical points is the choice of the bandwidth.

Choice of the bandwidth (univariate illustration)



too small bandwidth
overfitting



too large bandwidth

Multivariate density estimation

- The kernel density estimator, $\widehat{f}_H$, is given, for all $\mathbf{x} \in \mathbb{R}^d$, by

$$\widehat{f}_H(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{X}_i)$$

One of main critical points is the choice of the bandwidth H
We denote by $\|.\|$ the $\mathbb{L}_2$ norm

- We wish to select $\widehat{H} \in \mathcal{H}$ so that
  1. $\widehat{f}_{\widehat{H}}$ is optimal in the oracle setting meaning that with large probability

$$\|\widehat{f}_{\widehat{H}} - f\|^2 \quad \leq \quad \min_{H \in \mathcal{H}} \|\widehat{f}_H - f\|^2 + \texttt{negligible terms}$$

  2. the selection of $\widehat{H}$ is free-tuning
  3. the computational cost is reasonable

Classical approaches for (univariate) density estimation

- V-fold Cross-validation based on the least-squares contrast: Split $\{1, \ldots, n\}$ into V subsets, $B_1, \ldots, B_V$ and compute for each B_k the kernel rule on the training set $((\mathbf{X}_i)_{i \in B_\ell})_{\ell \neq k}$

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{V} \sum_{k=1}^{V} \mathcal{L}SC_{B_k}(\widehat{f}_h^{(-B_k)})$$

- Plug-in methods based on the minimisation of the asymptotic expansion of the MISE

- The classical Lepski's method consists in selecting the bandwidth $\widehat{h}$ by using the rule

$$\widehat{h} = \max \left\{ h \in \mathcal{H} : \quad \|\widehat{f}_{h'} - \widehat{f}_h\|^2 \leq V_1(h') \ \text{ for any } h' \in \mathcal{H} \text{ s.t. } h' \leq h \right\}$$

The Goldenshluger-Lepski's methodology is a variation of the Lepski's procedure:

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \{A(h) + V_2(h)\},$$

$$A(h) = \sup_{h' \in \mathcal{H}} \left\{ \|\widehat{f}_{h'} - K_h \star \widehat{f}_{h'}\|^2 - V_2(h') \right\}_+$$

Classical approaches for density estimation

- *V*-fold Cross-validation based on the least-squares contrast
- Plug-in methods, minimisation of the asymptotic expansion of the MISE
- The classical Lepski's method
  or the Goldenshluger-Lepski's methodology

These approaches are

- hard to tune,
- or not optimal in the oracle setting,
- or time-consuming.

$\hookrightarrow$ New method PCO (Penalized Comparison to Overfitting):
an alternative based on comparisons to the overfitting estimator

Heuristic, for $d = 1$

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

$f_h := \mathbb{E}(\hat{f}_h) = K_h \star f$

Oracle inequality in the univariate case

We consider $\mathcal{H}$ a finite set of positive reals and $h_n = \min \mathcal{H}$. We set

$$\widehat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \left\{ \|\widehat{f}_{h_n} - \widehat{f}_h\|^2 - \frac{\|K_{h_n} - K_h\|^2}{n} + \lambda \frac{\|K_h\|^2}{n} \right\}$$

Theorem

Assume that $\|f\|_\infty < \infty$ *and* $h_n \geq \|K\|_\infty \|K\|_1 / n$*. Let* $\epsilon \in (0,1)$*. If* $\lambda > 0$*,* $\forall x \geq 1$*, with probability larger than* $1 - c|\mathcal{H}|e^{-x}$*,*

$$\begin{aligned} \|\widehat{f}_{\hat{h}} - f\|^2 \quad &\leq \quad C_0(\epsilon, \lambda) \min_{h \in \mathcal{H}} \|\widehat{f}_h - f\|^2 \\ &\quad + C_1(\epsilon, \lambda) \|f_{h_n} - f\|^2 + C_2(\epsilon, K, \lambda) \frac{\|f\|_\infty x^3}{n} \end{aligned}$$

with the oracle constant $C_0(\epsilon, \lambda) = \lambda + \epsilon$ *if* $\lambda \geq 1$*,* $C_0(\epsilon, \lambda) = 1/\lambda + \epsilon$ *if* $0 < \lambda \leq 1$

In particular, the choice $\lambda = 1$ leads to an optimal estimate in the oracle setting.

Elements of the proof

For any $h \in \mathcal{H}$, a fast computation leads to

$$\|\hat{f}_{\hat{h}} - f\|^2 \leq \|\hat{f}_h - f\|^2 + \Big(\mathrm{pen}_\lambda(h) - 2\langle \hat{f}_h - f, \hat{f}_{h_n} - f\rangle\Big) - \Big(\mathrm{pen}_\lambda(\hat{h}) - 2\langle \hat{f}_{\hat{h}} - f, \hat{f}_{h_n} - f\rangle\Big)$$

$\hookrightarrow$ control $\langle \hat{f}_h - f, \hat{f}_{h_n} - f\rangle = \langle \hat{f}_h - f_h, \hat{f}_{h_n} - f_{h_n}\rangle + \ldots$

- control the U-statistic

$$U(h, h_n) = \sum_{i \neq j} \langle K_h(. - X_i) - f_h, K_{h_n}(. - X_j) - f_{h_n}\rangle$$

  $\hookrightarrow$ concentration inequality from Houdré and Reynaud-Bouret (2003)

- control the empirical sum $V(h, h') =< \hat{f}_h - f_h, f_{h'} - f >$

  $\hookrightarrow$ Bernstein's inequality

- use of the following lower bound

$$\|f - f_h\|^2 + \frac{\|K_h\|^2}{n} \leq (1 + \epsilon)\|f - \hat{f}_h\|^2 + \frac{C(K, \|f\|_\infty)x^2}{\epsilon^3 n} \qquad \text{w.h.p.}$$

  from Lerasle et al. (2015)

Minimal penalty

- Oracle inequality is obtained when the penalty it tuned with $\lambda > 0$, with

$$\mathrm{pen}_\lambda(h) = \frac{\lambda \|K_h\|^2 - \|K_{h_n} - K_h\|^2}{n}$$

- Take h_n so that for some $\beta > 0$,

$$\frac{\|K\|_\infty \|K\|_1}{n} \leq h_n \leq \frac{(\log n)^\beta}{n}$$

  and assume $nh_n \|f_{h_n} - f\|^2 = o(1)$ (Bias$(h_n) \ll$ Variance(h_n))
  If $\lambda < 0$, then, with probability larger than $1 - c|\mathcal{H}| \exp(-(n/\log n)^{1/3})$,

$$\widehat{h} \leq C(\lambda) h_n \leq C(\lambda) \frac{(\log n)^\beta}{n}$$

  where c is an absolute constant and $C(\lambda) = 2.1 - 1/\lambda$. This penalty leads to an overfitting estimator and

$$\liminf_{n \to +\infty} \mathbb{E}\left[\|\widehat{f}_{\hat{h}} - f\|^2\right] > 0 \quad \text{(risk explosion)}$$

- PCO is tuned by using $\lambda = 1$ leading to the optimal penalty

$$\mathrm{pen}_{\mathrm{opt}}(h) = \frac{2\langle K_h, K_{h_n}\rangle}{n}$$

The multivariate case: oracle setting

- Previous oracle inequalities can be extended to the multivariate case where $f : \mathbb{R}^d \longmapsto \mathbb{R}_+$ is the density of the $\mathbf{X}_i$'s with $\mathbf{X}_i = (X_{i1}, \dots, X_{id}) \in \mathbb{R}^d$.
- We consider $\mathcal{H}$, a finite set of symmetric positive-definite $d \times d$ matrices. Set $H_n = \bar{h} I_d$ and

$$\widehat{H} = \arg\min_{H \in \mathcal{H}} \left\{ \|\widehat{f}_{H_n} - \widehat{f}_H\|^2 - \frac{\|K_{H_n} - K_H\|^2}{n} + \lambda \frac{\|K_H\|^2}{n} \right\}$$

Theorem

Assume that $\|f\|_\infty < \infty$ *and* $\bar{h}^d \geq \|K\|_\infty \|K\|_1 / n$. *Let* $\epsilon \in (0,1)$. *If* $\lambda > 0$, $\forall x \geq 1$, *with probability larger than* $1 - c|\mathcal{H}|e^{-x}$,

$$\begin{aligned} \|\widehat{f}_{\widehat{H}} - f\|^2 \quad &\leq \quad C_0(\epsilon, \lambda) \min_{H \in \mathcal{H}} \|\widehat{f}_H - f\|^2 \\ &+ C_1(\epsilon, \lambda) \|f_{H_n} - f\|^2 + C_2(\epsilon, K, \lambda) \left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 \det(H_n)} \right), \end{aligned}$$

with $C_0(\epsilon, \lambda) = \lambda + \epsilon$ *if* $\lambda \geq 1$, $C_0(\epsilon, \lambda) = 1/\lambda + \epsilon$ *if* $0 < \lambda \leq 1$.

- In particular, the choice $\lambda = 1$ leads to an optimal estimate in the oracle setting.

The multivariate case: minimax setting

- We consider the minimax setting and construct a set of bandwidths leading to an optimal kernel estimate based on the PCO methodology.
- Let P an orthogonal matrix. Consider $H_n = \bar{h} I_d$ with $\bar{h}^d = \|K\|_\infty \|K\|_1 / n$ and choose for $\mathcal{H}$ the following set of bandwidths:

$$\mathcal{H} = \left\{ H = P^{-1} \mathrm{diag}(h_1, \ldots, h_d) P : \prod_{j=1}^{d} h_j \geq \bar{h}^d \text{ and } h_j^{-1} \in \mathbb{N}^* \ \forall j = 1, \ldots, d \right\}$$

Consider the PCO bandwidth (tuned with $\lambda = 1$)

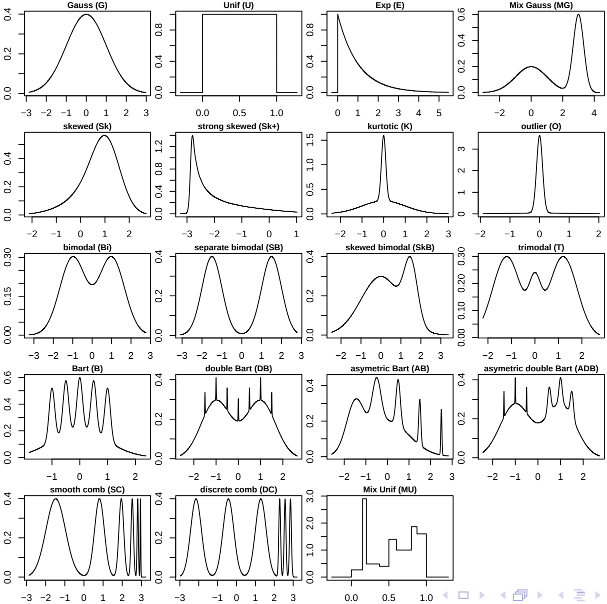
$$\widehat{H} = \operatorname*{arg\,min}_{H \in \mathcal{H}} \left\{ \|\widehat{f}_{H_n} - \widehat{f}_H\|^2 + \frac{2\langle K_H, K_{H_n} \rangle}{n} \right\}$$

- Assume that $f \circ P^{-1}$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L})$. Assume that the kernel K is order $\ell > \max_{j=1,\ldots,d} \beta_j$. Then, if for $B > 0$, $\|f\|_\infty \leq B$,

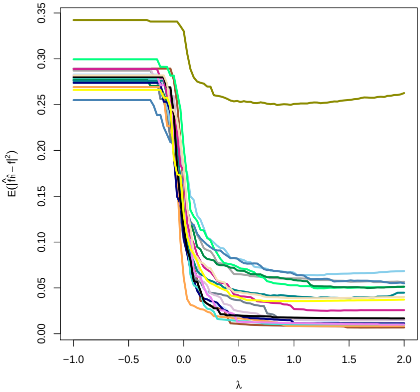
$$\mathbb{E}\left[\|\widehat{f}_{\widehat{H}} - f\|^2 \right] \leq M \left(\prod_{j=1}^{d} L_j^{\frac{1}{\beta_j}} \right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}} n^{-\frac{2\bar{\beta}}{2\bar{\beta}+1}},$$

where M is a constant only depending on $\boldsymbol{\beta}$, K, B, and d and $\bar{\beta} = (\sum_{j=1}^{d} 1/\beta_j)^{-1}$

Numerical study: benchmark univariate densities



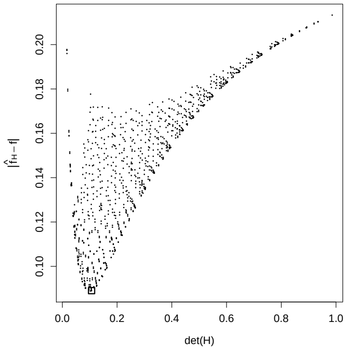
Numerical study: tuning for the univariate case

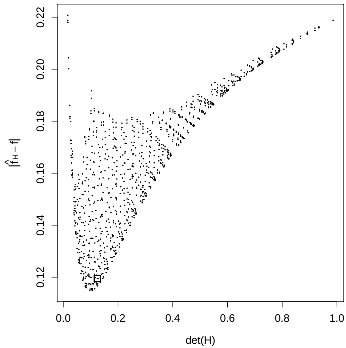


For each benchmark density f, estimated $\mathbb{L}_2$-risk of the PCO estimate by using the Monte Carlo mean over 20 samples in function of the tuning parameter λ, for the Gaussian kernel with $n = 100$ observations in the univariate case

Densities

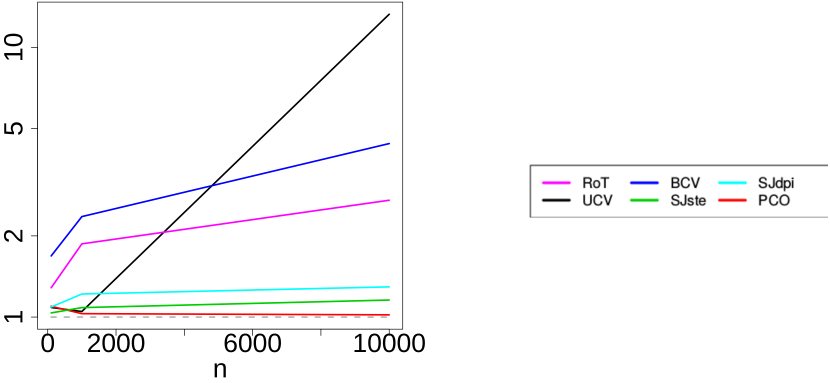
Numerical study: tuning for the bivariate case





Square root of the ISE against $\det(H)$ for all $H \in \mathcal{H}$ with $\mathcal{H}$ a set of 2×2 diagonal matrices for two different densities, with $n = 100$. The square corresponds to the bandwidth selected by PCO with $\lambda = 1$

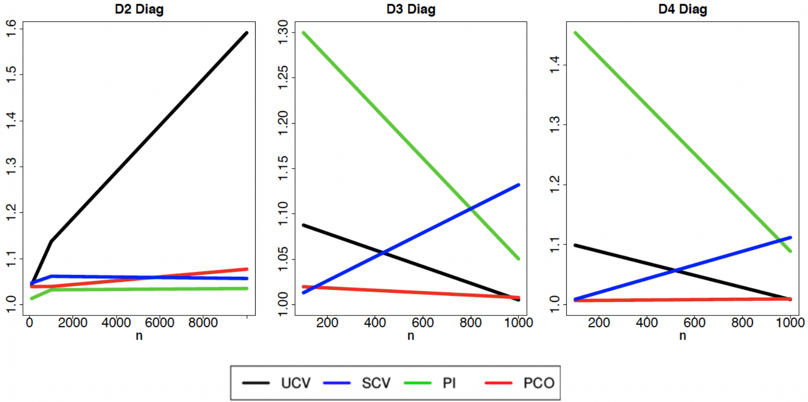
Numerical study: the univariate case



For $\mathrm{meth} \in \{\mathsf{RoT}, \mathsf{UCV}, \mathsf{BCV}, \mathsf{SJste}, \mathsf{SJdpi}, \mathsf{PCO}\}$ (implemented in the package ks) with the Gaussian kernel, graph versus the sample size of the mean over all 19 densities f of the ratio of

$$r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$$

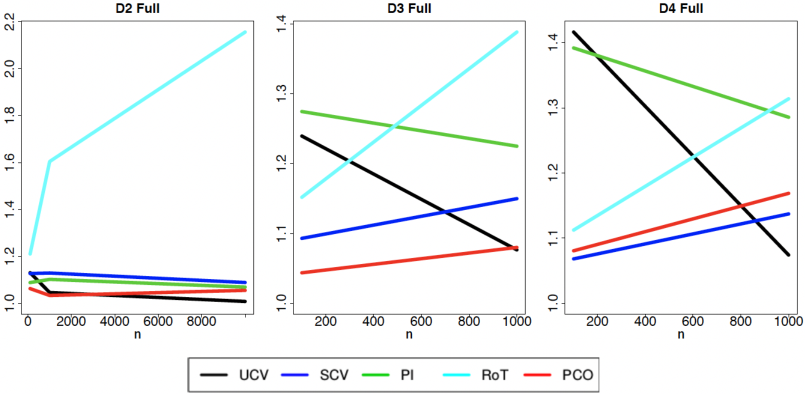
The multivariate case: $d \in \{2, 3, 4\}$ - Diagonal matrices



For $\mathrm{meth} \in \{\mathsf{UCV}, \mathsf{SCV}, \mathsf{PI}, \mathsf{PCO}\}$ with the Gaussian kernel, graph versus the sample size of the mean over all 14 densities f of the ratio of

$$r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$$

The multivariate case: $d \in \{2, 3, 4\}$ - Full matrices



For $\mathrm{meth} \in \{\mathsf{UCV}, \mathsf{SCV}, \mathsf{PI}, \mathsf{RoT}, \mathsf{PCO}\}$ with the Gaussian kernel, graph versus the sample size of the mean over all 14 densities f of the ratio of

$$r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$$

Conclusions from our numerical study

- Simulations corroborate what was expected from theory and validate the choice of the tuning constant $\lambda = 1$ in the penalty term.
- The choice of the parameter h_n is not very sensitive and taking $h_n = \|K\|_\infty \|K\|_1 / n$ is suitable and robust.
- These parameters being tuned once for all, PCO becomes a ready to be used method which is further more easy to compute.
- As compared to other methods, PCO has a stable behavior and its performance is never far from being optimal. PCO is not always the best competitor but it has the advantage of staying competitive in any situation.

Conclusions and perspectives

- PCO offers several advantages which should be welcome for practitioners:
  1. It can be used for moderately high dimensional data
  2. PCO is optimal in oracle and minimax settings and achieves nice numerical performances
  3. To a large extent, it is free-tuning
  4. Its computational cost is quite reasonable
- PCO has been used in various settings: nonparametric regression, deconvolution and other settings: Comte, Prieur and Samson (2017), Deschatre (2017), Lehéricy (2018), Pham Ngoc (2019), Halconruy and Marie (2020), Comte and Marie (2020, 2021), Divol (2021)
- Future directions of research: interesting to develop PCO, both from a theoretical and a practical perspective for other losses than the $\mathbb{L}_2$-loss (Hellinger and $\mathbb{L}_p$-losses for $p \neq 2$). Work in progress.

Thank you for your attention!

References:

- Lacour C., Massart P. and Rivoirard V. (2017) Estimator selection: a new method with applications to kernel density estimation. *Sankhya A (special issue on Application of concentration inequalities and empirical processes to modern statistics)*, 79, no 2, 298-335

- Varet, S., Lacour C., Massart P. and Rivoirard V. (2022) Numerical performance of Penalized Comparison to Overfitting for multivariate kernel density estimation. *Submitted*