ESAIM: PS 27 (2023) 621–667
https://doi.org/10.1051/ps/2022018

ESAIM: Probability and Statistics
www.esaim-ps.org

NUMERICAL PERFORMANCE OF PENALIZED COMPARISON TO OVERFITTING FOR MULTIVARIATE KERNEL DENSITY ESTIMATION

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

Abstract. Kernel density estimation is a well known method involving a smoothing parameter (the bandwidth) that needs to be tuned by the user. Although this method has been widely used, the bandwidth selection remains a challenging issue in terms of balancing algorithmic performance and statistical relevance. The purpose of this paper is to study a recently developed bandwidth selection method, called Penalized Comparison to Overfitting (PCO). We first provide new theoretical guarantees by proving that PCO performed with non-diagonal bandwidth matrices is optimal in the oracle and minimax approaches. PCO is then compared to other usual bandwidth selection methods (at least those which are implemented in the R-package) for univariate and also multivariate kernel density estimation on the basis of intensive simulation studies. In particular, cross-validation and plug-in criteria are numerically investigated and compared to PCO. The take home message is that PCO can outperform the classical methods without algorithmic additional cost.

Mathematics Subject Classification. 62G07, 62-08, 62H12.

Received September 17, 2021. Accepted November 18, 2022.

1. Introduction

Density estimation is widely used in a variety of fields in order to study the data and extract informations on variables whose distribution is unknown. Due to its simplicity of use and interpretation, kernel density estimation is one of the most commonly used density estimation procedure. Of course we do not pretend that it is "the" method to be used in any case but that being said, if one wants to use it in a proper way, one has to take into account that its performance is conditioned by the choice of an adapted bandwidth. From a theoretical perspective, once some loss function is given, an ideal bandwidth should minimize the loss (or the expectation of the loss) between the kernel density estimator and the unknown density function. Since these "oracle" choices do not make sense in practice, statistical bandwidth selection methods consist of mimicking the oracle through the minimization of some criteria that depend only on the data. Because it is easy to compute and to analyze, the $\mathbb{L}_2$ loss has been extensively studied in the literature although it would also make sense to consider the

Keywords and phrases: Multivariate density estimation, Kernel-based density estimation, Bandwidth selection.

[1] Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405 Orsay, France.
[2] LAMA, Univ Gustave Eiffel, Univ Paris Est Creteil, CNRS, 77447 Marne-la-Vallée, France.
[3] CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France.

[*] Corresponding author: claire.lacour@u-pem.fr

Kulback-Leibler loss, the Hellinger loss or the $\mathbb{L}_1$ loss (which are somehow more intrinsic losses with respect to the distribution and do not depend on the dominating measure). For the same reasons as those mentioned above, we shall deal with the $\mathbb{L}_2$ loss in this paper and all the comparisons that we shall make between the method that we propose and other methods will be performed relatively to this loss. Focusing on the $\mathbb{L}_2$ loss, two classes of bandwidth selection have been well studied and are commonly used: cross validation and plug-in. They correspond to different ways of estimating the ISE (integrated squared error) which is just the square of the $\mathbb{L}_2$ loss between the estimator and the true density or the MISE (mean integrated squared error), which is just the expectation of the preceding quantity. The least-square cross-validation (LSCV) [2, 24] tends to minimize the ISE by replacing the occurence of the underlying density by the leave-one-out estimator. However LSCV suffers from the dispersion of the ISE even for large samples and tends to overfit the underlying density as the sample size increases. The plug-in approaches are based on the asymptotic expansion of the MISE. Since the asymptotic expansion of the MISE involves a bias term that depends on the underlying density itself one can estimate this term by plugging a *pilot* kernel estimator of the true density. Thus this plug-in approach depends on the choice of a *pilot* kernel and also on the choice of the *pilot* bandwidth. The so-called "rule of thumb" method [27] is a popular ready to be used variant of the plug-in approach in which the unknown term in the MISE is estimated as if the underlying density were Gaussian. Note that variations of previous methods can be found in the literature. See for instance [16, 17] for nice theoretical and practical reviews.

These methods have been first proposed and studied for univariate data and then extended to the multivariate case. The LSCV estimator for instance has been adapted in [29] to the multivariate case. A multivariate version of the smooth cross-validation is presented in [5, 10]. The rule of thumb is studied in [30] and the multivariate extension of the plug-in is developed in [4, 33]. Generally speaking, these methods have some well-known drawbacks: leave-one-out cross-validation tends to overfit the density for large sample while plug-in approaches depend on prior informations on the underlying density that are requested to estimate asymptotically the bias part of the MISE and which can turn to be inaccurate especially when the sample size is small.

The Penalized Comparison to Overfitting (PCO) is a selection method that has been recently developed in [19]. This approach is based on the penalization of the $\mathbb{L}_2$ loss between an estimator and the overfitting one. It does not belong to the family of cross-validation methods nor to the class of plug-in methods, but lies somewhere between these two classes. Indeed the main idea is still to mimic the oracle choice of the bandwidth but the required bias versus variance trade-off is achieved by estimating the variance with the help of a penalty term (as in a plug-in method) while the bias is estimated implicitly from the penalization procedure itself as in a cross-validation method. In [19], a theoretical study of the tuning of parameters of PCO is led. Easy to implement values are obtained, for which PCO is optimal in the oracle and minimax approaches.

What's new in this paper? As mentioned above the PCO method for quadratic risk and kernel density estimation was introduced and mathematically studied in Lacour *et al.* [19]. In this paper we further investigate this method in two distinct directions. On the one hand, we propose a modification of the method introduced and studied in [19] to deal with the case of anisotropic multivariate density estimation. In particular, to deal with distributions whose behavior can be very different from one direction to another, we consider symmetric definite positive bandwidth matrices parametrization. This consideration is important for applications since it allows adaptation to an unknown correlation structure for the data. In this context we prove a new oracle inequality which proves the optimality of our method on Nikol'skii regularity classes. We also propose a comparative numerical study to verify that the optimal choice of bandwidth which we have proved to be theoretically optimal behaves well on simulated data and this even for moderate sample sizes. For this study, we compared our method to those which, to our knowledge, are the most commonly used, knowing that many variants exist and that we cannot claim to be exhaustive.

As a conclusion of this numerical study we see that our method behaves well when compared to its competitors and offers several advantages which should be welcome for practitioners:

1. It can be used for moderately high dimensional data.
2. To a large extent, it is free-tuning.
3. Its computational cost is quite reasonable.

Moreover while the theoretical properties of the most popular methods such as the Rule of thumb, the Least-Square Cross-Validation, the Biased Cross-Validation, the Smoothed Cross-Validation and the Sheather and Jones Plug-in approach have been well studied in dimension 1, this is not the case in dimension larger than 2. Of course R-packages do exist for these methods in the multivariate case but to our knowledge similar theoretical garantees as those that we prove here for our method do not exist yet.

Concretely, the performance of each method that we analyze in our numerical study is measured in terms of the MISE of the corresponding selected kernel estimator (more precisely we use the Monte-Carlo evaluation of the MISE rather than the MISE per se). We present the results obtained for several "test" laws. We borrowed most of these laws from [21] for univariate data and [3] for bivariate data (and we use some natural extensions of them for multivariate data, up to dimension 4).

To avoid a too long study, we have only focused on comparisons between kernel estimates and simulations are performed on quite smooth density functions for which kernel rules are well suited. Of course other methodologies could be considered, as for instance density estimates based on histograms (see [24]), wavelets (see [8]) or dyadic piecewise polynomial (see [1]). Observe that histogram estimators are best ones to recover step functions while very oscillating ones should be estimated by using wavelets.

Section 2 is devoted to the presentation of all the methods used for the numerical comparison. In particular, PCO is described in Section 2.1.1 for the univariate case and in Section 2.2.1 for the multivariate case. Its implementation is discussed in Section 3.1 and the numerical study of the tuning of PCO-paremeters is led in Section 3.2. The numerical results for univariate data are detailed in Section 3.3 and in Section 3.4 for multivariate data. The extension of the oracle inequality of PCO for non-diagonal bandwidth is given in Theorem 2.1 and Corollary 2.4 of Section 2.2.1 and the associated proofs are in Appendix A. The code used for the simulations presented in this paper can be found at https://github.com/SuVaret/PCO.

Notations. The bold font denotes vectors. For any matrix A, we denote A^T the transpose of A. $Tr(A)$ denotes the trace of the matrix A.

2. Bandwidth selection

Due to their simplicity and their smoothing properties, kernel rules are among the most extensively used methodologies to estimate an unknown density, denoted f along this paper, where $f : \mathbb{R}^d \mapsto \mathbb{R}_+$. For this purpose, we consider an n sample $\mathcal{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ with $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id}) \in \mathbb{R}^d$. The kernel density estimator, $\hat{f}_H$, is given, for all $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$, by

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

where K is the kernel function, the matrix H is the kernel bandwidth belonging to a fixed grid $\mathcal{H}$ and $K_H(\boldsymbol{x}) = \frac{1}{\det(H)} K\left(H^{-1}\boldsymbol{x}\right)$. Of course, the choice of the bandwidth is essential from both theoretical and practical points of view. In the sequel, we assume that K verifies following conditions, satisfied by usual kernels:

$$\int K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1, \quad \|K\|^2 < \infty, \quad \int \|\boldsymbol{x}\|^2 |K(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} < \infty,$$

where $\|\cdot\|$ denotes the $\mathbb{L}_2$-norm on $\mathbb{R}^d$. Most of selection rules described subsequently are based on a criterion to be minimized. We restrict our attention to $\mathbb{L}_2$-criteria even if other approaches could be investigated. For this purpose, we introduce the *Integrated Square Error* (ISE) of the estimator $\hat{f}_H$ defined by

$$ISE(H) := \|\hat{f}_H - f\|^2 \tag{2.1}$$

and the mean of $ISE(H)$:

$$MISE(H) := \mathbb{E}[ISE(H)] = \mathbb{E}\|\hat{f}_H - f\|^2.$$

2.1. Univariate case

We first deal with the case $d = 1$ and we denote $\mathcal{X} = (X_1, \ldots, X_n)$ the n-sample of density f. The general case is investigated in Section 2.2. The bandwidth parameter lies in $\mathbb{R}_+^*$ and is denoted h, instead of H. In our $\mathbb{L}_2$-perspective, it is natural to use a bandwidth which minimizes $h \mapsto MISE(h)$ or $h \mapsto ISE(h)$. However, such functions strongly depend on f. We can relax this dependence by using an asymptotic expansion of the MISE:

$$AMISE(h) = \frac{\|K\|^2}{nh} + \frac{1}{4}h^4\mu_2^2(K)\|f''\|^2,$$

when f'' exists and is square-integrable, with $\mu_2(K) = \int x^2 K(x)dx$. We refer the reader to [32] who specified mild conditions under which $AMISE(h)$ is close to $MISE(h)$ when $n \to +\infty$. The main advantage of the AMISE criterion lies in the closed form of the bandwidth that minimizes it:

$$\hat{h}_{\text{AMISE}} = \left(\frac{\|K\|^2}{\mu_2^2(K)\|f''\|^2}\right)^{1/5} n^{-1/5}. \tag{2.2}$$

Note however, that $\hat{h}_{\text{AMISE}}$ still depends on f through $\|f''\|^2$. The Rule of Thumb developed in [22] and popularized by [27] (and presented subsequently) circumvents this problem. Cross-validation approaches based on a direct estimation of $ISE(h)$ constitute an alternative to bandwidth selection derived from the AMISE criterion. Both approaches can of course be combined. Before describing them precisely, we first present the PCO methodology for the univariate case.

2.1.1. Penalized Comparison to Overfitting (PCO)

Penalized Comparison to Overfitting (PCO) has been proposed by [19]. We recall main heuristic arguments of this method for the sake of completeness. We start from the classical bias-variance decomposition

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

where for any h, $f_h := K_h \star f = \mathbb{E}[\hat{f}_h]$, with $\star$ the convolution product. It is natural to consider a criterion to be minimized (on a grid) of the form

$$\text{Crit}(h) = \hat{b}_h + \hat{v}_h,$$

where $\hat{b}_h$ is an estimator of the bias b_h and $\hat{v}_h$ an estimator of the variance v_h. Minimizing such a criterion is hopefully equivalent to minimizing the MISE. Using that v_h is (tightly) bounded by $\|K\|^2/(nh)$, we naturally set $\hat{v}_h = \lambda\|K\|^2/(nh)$, with λ some tuning parameter. The difficulty lies in estimating the bias. Here we assume that h_{min}, the minimum of the bandwidths grid, is very small. In this case, $f_{h_{min}} = K_{h_{min}} \star f$ is a good approximation of f, so that $\|f_{h_{min}} - f_h\|^2$ is close to b_h. This is tempting to estimate this term by $\|\hat{f}_{h_{min}} - \hat{f}_h\|^2$ but doing so, we introduce a bias. Indeed, since

$$\hat{f}_{h_{min}} - \hat{f}_h = (\hat{f}_{h_{min}} - f_{h_{min}} - \hat{f}_h + f_h) + (f_{h_{min}} - f_h)$$

we have the decomposition

$$\mathbb{E}\|\hat{f}_{h_{min}} - \hat{f}_h\|^2 = \|f_{h_{min}} - f_h\|^2 + \mathbb{E}\|\hat{f}_{h_{min}} - \hat{f}_h - f_{h_{min}} + f_h\|^2. \tag{2.3}$$

But the centered variable $\hat{f}_{h_{min}} - \hat{f}_h - f_{h_{min}} + f_h$ can be written

$$\hat{f}_{h_{min}} - \hat{f}_h - f_{h_{min}} + f_h = \frac{1}{n}\sum_{i=1}^{n}(K_{h_{min}} - K_h)(. - X_i) - \mathbb{E}((K_{h_{min}} - K_h)(. - X_i)).$$

So, the second term in the right hand side of (2.3) is of order $n^{-1}\int(K_{h_{min}}(x) - K_h(x))^2dx$. Hence,

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

and then

$$b_h \approx \|f_{h_{min}} - f_h\|^2 \approx \|\hat{f}_{h_{min}} - \hat{f}_h\|^2 - \frac{\|K_{h_{min}} - K_h\|^2}{n}.$$

These heuristic arguments lead to the following criterion to be minimized:

$$l_{\mathrm{PCO}}(h) = \|\hat{f}_{h_{min}} - \hat{f}_h\|^2 - \frac{\|K_{h_{min}} - K_h\|^2}{n} + \lambda\frac{\|K_h\|^2}{n}. \tag{2.4}$$

Thus, our method consists in comparing every estimator of our collection to the overfitting one, namely $\hat{f}_{h_{min}}$, before adding the penalty term

$$\mathrm{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{min}} - K_h\|^2}{n}. \tag{2.5}$$

The selected bandwidth is then

$$\hat{h}_{\mathrm{PCO}} = \arg\min_{h\in\mathcal{H}} l_{\mathrm{PCO}}(h).$$

In [19], it is proved that this bandwidth choice allows to achieve the optimal minimax integrated risk, for $h_{min} = \|K\|_\infty\|K\|_1/n$ and λ well chosen. More precisely, in [19], it is shown that the risk blows up when $\lambda < 0$. So, the optimal value for λ lies in $\mathbb{R}_+$. Theorem 2 of [19] (generalized in Thm. 2.1 of Sect. 2.2.1) suggests that the optimal tuning parameter is $\lambda = 1$. It is in line with previous heuristic arguments (see the upper bound of v_h). In Section 3.3, we first conduct some numerical experiments and establish that PCO is indeed optimal for $\lambda = 1$. We then fix $\lambda = 1$ for all comparisons. We also study the choice of h_{min} and show that this parameter is not sensitive if chosen in a suitable range, so PCO becomes a free-tuning methodology.

Connections between PCO and the approach proposed by Goldenshluger and Lepski are quite strong. Introduced in [11], the Goldenshluger Lepski's methodology is a variation of the Lepski's procedure still based on pair-by-pair comparisons between estimators. More precisely, Goldenshluger and Lepski suggest to use the selection rule

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \{A(h) + V_2(h)\},$$

with

$$A(h) = \sup_{h'\in\mathcal{H}} \left\{\|\hat{f}_{h'} - \hat{f}_{h\vee h'}\|^2 - V_1(h')\right\}_+,$$

where x_+ denotes the positive part $\max(x, 0)$, $h \vee h' = \max(h, h')$ and $V_1(\cdot)$ and $V_2(\cdot)$ are penalties to be suitably chosen (Goldenshluger and Lepski essentially consider $V_2 = V_1$ or $V_2 = 2V_1$ in [11–13, 15]). The authors establish the minimax optimality of their method when V_1 and V_2 are large enough. However, observe that if $V_1 = 0$, then, under mild assumptions,

$$A(h) = \sup_{h' \in \mathcal{H}} \|\hat{f}_{h'} - \hat{f}_{h \vee h'}\|^2 \approx \|\hat{f}_{h_{min}} - \hat{f}_h\|^2$$

so that our method turns out to be exactly some degenerate case of the Goldenshluger Lespki's method. Two difficulties arise for the use of the Goldenshluger Lespki's method: Functions V_1 and V_2 depend on some parameters which are very hard to tune. Based on 2 optimization steps, its computational cost is very large. Furthermore, the larger the dimension, the more accurate these problems are. Note that the classical Lepski's method shares same issues. We lead a brief comparative numerical study in Section 3.3 that confirms that PCO gives similar results to Goldenshluger Lespki's method, but with a considerably reduced computation cost (and without need of calibration).

Other kernel rules using two bandwidths deserve to be mentioned: see [17, 18] and references therein. But their philosophy, in terms of bias estimation, penalization and choice of second bandwidth, is very different.

2.1.2. Silverman's Rule of thumb (RoT and RoT0)

The Rule of Thumb has been developed in [22] and popularized by [27]. We assume that f'' exists and is such that $\|f''\| < \infty$. The simplest way to choose h is to use a standard family of distributions to minimize $h \mapsto AMISE(h)$.

For a Gaussian kernel and f the probability density function of the normal distribution, an approximation of $\|f''\|^2$ can be plugged in (2.2) leading to a bandwidth of the form $\hat{h} = 1.06\hat{\sigma}n^{-1/5}$ where $\hat{\sigma}$ is an estimation of the standard deviation of the data. However this bandwidth leads to an oversmoothed estimator of the density for multimodal distributions. Thus it is better to use the following estimator, which works well with unimodal densities and not too badly for moderate bimodal ones:

$$\hat{h}_{\mathrm{RoT}} = 1.06 \times \min\left(\hat{\sigma}, \frac{\hat{q}_{75} - \hat{q}_{25}}{1.34}\right) \times n^{-1/5} \tag{2.6}$$

where $\hat{q}_{75} - \hat{q}_{25}$ is an estimation of the interquartile range of the data. Another variant of this approximation ([27] pp. 45–47) is:

$$\hat{h}_{\mathrm{RoT0}} = 0.9 \times \min\left(\hat{\sigma}, \frac{\hat{q}_{75} - \hat{q}_{25}}{1.34}\right) \times n^{-1/5}.$$

These two variants of the Rule of Thumb methodology are respectively denoted RoT and RoT0.

2.1.3. Least-Square Cross-Validation (UCV)

Least-square cross-validation has been developed independently in [2, 24]. In the univariate case, equation (2.1) can be expressed as

$$ISE(h) = \int \hat{f}_h^2 - 2\int \hat{f}_h f + \int f^2. \tag{2.7}$$

Since the last term of (2.7) does not depend on h, minimizing (2.7) is equivalent to minimizing

$$Q(h) = \int \hat{f}_h^2 - 2\int \hat{f}_h f.$$

The least-square cross-validation constructs an estimator of $Q(h)$ from the leave-one out estimator $\hat{f}_{-i}$:

$$l_{\mathrm{ucv0}}(h) = \int \hat{f}_h^2 - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) \tag{2.8}$$

where the leave-one out estimator is given by

$$\hat{f}_{-i}(x) = \frac{1}{n-1}\sum_{j\neq i} K_h\left(x - X_j\right).$$

Then, $\mathbb{E}[Q(h)]$ is unbiasedly estimated by $l_{\mathrm{ucv0}}(h)$, which justifies the use of l_{ucv0} for the bandwidth selection and this is the reason why this estimator is also called unbiased cross-validation (UCV) estimator. Using the expression of $\hat{f}_{-i}$ in (2.8) and replacing the factor $\frac{1}{n-1}$ with $\frac{1}{n}$ for computation ease, the following estimator $l_{\mathrm{ucv}}(h)$ is used in practice:

$$l_{\mathrm{ucv}}(h) = \frac{1}{hn^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0)$$

where $K^*(u) = (K \star \tilde{K})(u) - 2K(u)$and $\tilde{K}(u) = K(-u)$. Finally, the bandwidth selected by the least-square cross-validation is given by:

$$\hat{h}_{\mathrm{ucv}} = \arg\min_{h\in\mathcal{H}} l_{\mathrm{ucv}}(h).$$

2.1.4. Biased Cross-Validation (BCV)

The biased cross-validation was developed in [25]. It consists in minimizing the AMISE. So, we assume that f'' exists and $\|f''\| < \infty$. Since the AMISE depends on the unknown density f through $\|f''\|$, the biased cross-validation estimates $\|f''\|^2$ by

$$\widehat{\|f''\|^2} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} (\tilde{K}_h'' \star K_h'')(X_i - X_j),$$

using a twice differentiable kernel. Straightforward computations show that

$$\mathbb{E}\left[\widehat{\|f''\|^2}\right] = \|K_h \star f''\|^2,$$

which is close to $\|f''\|^2$ when h is small under mild conditions on K. This justifies the use of the objective function of BCV defined by:

$$l_{\mathrm{bcv}}(h) = \frac{\|K\|^2}{nh} + \frac{1}{4}h^4\mu_2^2(K)\widehat{\|f''\|^2}.$$

Finally, the bandwidth selected by the BCV is given by:

$$\hat{h}_{\mathrm{bcv}} = \arg\min_{h} l_{\mathrm{bcv}}(h) = \left(\frac{\|K\|^2}{\mu_2^2(K)\widehat{\|f''\|^2}}\right)^{1/5} n^{-1/5}.$$

2.1.5. Sheather and Jones Plug-in (SJ)

This estimator is, as BCV, based on the minimization of the AMISE. The difference with the BCV approach is in the estimation of $\|f''\|^2$. In this plug-in approach, $\|f''\|^2$ is estimated by the empirical mean of the fourth derivative of f, where f is replaced by a *pilot* kernel density estimate of f. Indeed, using two integrations by parts, under mild assumptions on f, we have $\mathbb{E}[f^{(4)}(X)] = \|f''\|^2$. The *pilot* kernel density estimate is defined by:

$$\hat{f}_{\text{pilot},b}(x) = \frac{1}{n}\sum_{j=1}^{n} L_b(x - X_j)$$

where L is the pilot kernel function and b the pilot bandwidth. Then, $\|f''\|^2$ is estimated by $\hat{S}(b)$ with

$$\hat{S}(\alpha) = \frac{1}{n(n-1)\alpha^5}\sum_{i=1}^{n}\sum_{j=1}^{n} L^{(4)}\left(\frac{X_i - X_j}{\alpha}\right). \tag{2.9}$$

The pilot bandwidth b is chosen in order to compensate the bias introduced by the diagonal term $i = j$ in (2.9) as explained in Section 3 of [26]. Thus, for choosing b, Sheather and Jones propose two algorithms based on the remark that, in this context, the pilot bandwidth b can be written as a quantity proportional to $h^{5/7}$ or proportional to $n^{-1/7}$. The first algorithm, called '*solve the equation*' ('ste'), consists in taking the expression $b = b(h) \propto h^{5/7}$, pluging $\hat{S}(b(h))$ in (2.2) and solving the equation. The second algorithm, 'direct plug-in', consists in taking $b \propto n^{-1/7}$, and pluging $\hat{S}(b)$ in (2.2). Thus the SJ estimators of h are given by:

$$\hat{h}_{SJste} = \left(\frac{\|K\|^2}{\mu_2^2(K)\hat{S}(c_1\hat{h}_{SJste}^{5/7})}\right)^{1/5} n^{-1/5}$$

for the 'ste' algorithm and

$$\hat{h}_{SJdpi} = \left(\frac{\|K\|^2}{\mu_2^2(K)\hat{S}(c_2 n^{-1/7})}\right)^{1/5} n^{-1/5}$$

for the 'dpi' algorithm. The constant c_1 is $c_1 = \left(\frac{2L^{(4)}(0)\mu_2^2(K)}{\mu_2(L)\|K\|^2}\right)^{1/7}\left(\frac{\|f''\|^2}{\|f'''\|^2}\right)^{1/7}$ where $\|f''\|^2$ and $\|f'''\|^2$ are estimated by $\|\widehat{f''_a}\|^2$ and $\|\widehat{f'''_b}\|^2$ with a and b the Silverman's rule of thumb bandwidths respectively. The constant c_2 is equal to $\left(\frac{2L^{(4)}(0)}{\mu_2(L)}\right)^{1/7}\left(\frac{1}{\|f'''\|^2}\right)^{1/7}$ (see Eq. (9) of [26]), where $\|f'''\|^2$ is estimated by $\|\widehat{f'''_a}\|^2$ with a the Silverman's rule of thumb bandwidth.

2.2. Multivariate case

The difficulty of the multivariate case lies in the selection of a matrix rather than a scalar bandwidth. Different classes of matrices can be used. The simplest class corresponds to matrices of the form hI_d for $h \in \mathbb{R}_+^*$. In this case, selecting the bandwidth matrix is equivalent to deriving a single smoothing parameter. However, the unknown distribution may have different behaviors according to the coordinate direction. The latter parametrization does not allow us to take this specificity into account. An extension of this class corresponds to diagonal matrices of the form $\text{diag}(h_1, \ldots, h_d)$. But this parametrization is not convenient when the directions of the density are not those of the coordinates. The most general case corresponds to the class of all symmetric definite positive matrices, which allows smoothing in arbitrary directions. A comparison of these parametrizations can be found in [31]. In this paper, we focus on diagonal and on symmetric definite positive matrices parametrization.

We now assume that the kernel $K : \mathbb{R}^d \mapsto \mathbb{R}$ satisfies

$$\int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mu_2(K) I_d$$

where $\mu_2(K)$ is a finite positive constant. In the general setting of symmetric definite positive matrices, and using the asymptotic expansion of the bias and the variance terms, the MISE is usually approximated by the AMISE function defined by

$$AMISE(H) = \frac{1}{4}\mu_2^2(K) \int [Tr(H^2 D^2 f(\boldsymbol{x}))]^2 \mathrm{d}\boldsymbol{x} + \frac{\|K\|^2}{n\mathrm{det}(H)}$$

with $D^2 f(\boldsymbol{x})$ the Hessian matrix of f. See [30] for instance. Note that $AMISE(H)$ can also be expressed as

$$AMISE(H) = \frac{1}{4}\mu_2^2(K)(\mathrm{vech}(H^2))^T \Psi_f(\mathrm{vech}(H^2)) + \frac{\|K\|^2}{n\mathrm{det}(H)}, \tag{2.10}$$

where vech is the vector half operator which transforms the lower triangular half of a matrix into a vector scanning column-wise and the matrix Ψ_f is defined by

$$\Psi_f = \int \mathrm{vech}(2D^2 f(\boldsymbol{x}) - \mathrm{diag}(D^2 f(\boldsymbol{x})))(\mathrm{vech}(2D^2 f(\boldsymbol{x}) - \mathrm{diag}(D^2 f(\boldsymbol{x}))))^T \tag{2.11}$$

with $\mathrm{diag}(A)$ the diagonal matrix formed with the diagonal elements of A.

2.2.1. Penalized Comparison to Overfitting (PCO)

The PCO methodology developed in [19] only deals with diagonal bandwidths H. We now generalize it to the more general case of symmetric positive-definite $d \times d$ matrices to compare its numerical performances to all popular methods dealing with multivariate densities. We then establish theoretical properties of PCO in oracle and minimax approaches. To the best of our knowledge, similar results have not been established for competitors of PCO.

In the sequel, we consider $\mathcal{H}$, a finite set of symmetric positive-definite $d \times d$ matrices. Let $\bar{h} \in \mathbb{R}_+^*$. Then, set $H_{min} = \bar{h} I_d$. We still consider

$$\hat{H}_{PCO} = \arg\min_{H \in \mathcal{H}} l_{PCO}(H)$$

with

$$l_{PCO}(H) = \|\hat{f}_{H_{min}} - \hat{f}_H\|^2 - \frac{\|K_{H_{min}} - K_H\|^2}{n} + \lambda \frac{\|K_H\|^2}{n}$$

and $\lambda > 0$. Define

$$f_H = \mathbb{E}[\hat{f}_H] = K_H \star f,$$

which goes to f when H goes to $\mathbf{0}_d$, under mild assumptions. The estimator $\hat{f}_{\hat{H}_{PCO}}$ verifies the following oracle inequality.

Theorem 2.1. *Assume that* $\|f\|_\infty < \infty$ *and* K *is symmetric. Assume that* $\det(H_{min}) \geq \|K\|_\infty \|K\|_1/n$. *Let* $x \geq 1$ *and* $\varepsilon \in (0,1)$. *If* $\lambda > 0$, *then, with probability larger than* $1 - C_1|\mathcal{H}|e^{-x}$,

$$\|\hat{f}_{\hat{H}_{PCO}} - f\|^2 \leq C_0(\varepsilon,\lambda) \min_{H\in\mathcal{H}} \|\hat{f}_H - f\|^2 \\ +C_2(\varepsilon,\lambda)\|f_{H_{min}} - f\|^2 + C_3(\varepsilon,K,\lambda)\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2\det(H_{min})}\right),$$

where C_1 *is an absolute constant and* $C_0(\varepsilon,\lambda) = \lambda + \varepsilon$ *if* $\lambda \geq 1$, $C_0(\varepsilon,\lambda) = 1/\lambda + \varepsilon$ *if* $0 < \lambda < 1$. *The constant* $C_2(\varepsilon,\lambda)$ *only depends on* ε *and* λ *and* $C_3(\varepsilon,K,\lambda)$ *only depends on* ε, K *and* λ.

The proof of Theorem 2.1 is given in Appendix A. Up to the constant $C_0(\varepsilon,\lambda)$, the first term of the oracle inequality corresponds to the ISE of the best estimate $\hat{f}_H$ when H describes $\mathcal{H}$. The main assumption of the theorem means that $\bar{h}$ cannot be smaller than $n^{-1/d}$ up to a constant. When $\bar{h}$ is taken proportional to $n^{-1/d}$, then the third term is of order x^3/n and is negligible with x proportional to $\log n$. The second term is also negligible when f is smooth enough and $\bar{h}$ small (see Cor. 2.4).

Remark 2.2. Note that $\arg\min_{\lambda\in\mathbb{R}_+^*} C_0(\varepsilon,\lambda) = 1$ and $C_0(\varepsilon,1) = 1+\varepsilon$. So, taking $\lambda = 1$ ensures that the leading constant of the main term of the right hand side is close to 1 when ε is small. Neglecting the other terms, this oracle inequality shows that the risk of PCO tuned with $\lambda = 1$ is not worse than the risk of the best estimate $\hat{f}_H$ up to the constant $1+\varepsilon$, for any $\varepsilon > 0$.

From Theorem 2.1, we deduce that if $\mathcal{H}$ is not too big and contains a quasi-optimal bandwidth, we can control the MISE of PCO on the Nikol'skii class of functions by assuming that K has enough vanishing moments. The anisotropic Nikol'skii class is a specific smoothness space defined as follows. Let $(e_1,\ldots,e_d)$ denote the canonical basis of $\mathbb{R}^d$. For any function $g : \mathbb{R}^d \mapsto \mathbb{R}$ and any $u \in \mathbb{R}$, we define the first order difference operator with step size u in the j-th direction by

$$\Delta_{u,j} g(x) = g(x + ue_j) - g(x), \quad j = 1,\ldots,d.$$

By induction, the k-th order difference operator with step size u in the j-th direction is defined as

$$\Delta^k_{u,j} g(x) = \Delta_{u,j}\Delta^{k-1}_{u,j} g(x) = \sum_{\ell=1}^{k} (-1)^{\ell+k}\binom{k}{\ell}\Delta_{u\ell,j} g(x).$$

We then set

Definition 2.3. For any given vectors $\mathbf{r} = (r_1,\ldots,r_d)$, $r_j \in [1,+\infty]$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$, $\beta_j > 0$, and $\mathbf{L} = (L_1,\ldots,L_d)$, $L_j > 0$, $j = 1,\ldots,d$, we say that the function $g : \mathbb{R}^d \mapsto \mathbb{R}$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{\mathbf{r},d}(\boldsymbol{\beta},\mathbf{L})$ if

(i) $\|g\|_{r_j} \leq L_j$ for all $j = 1,\ldots,d$
(ii) for every $j = 1,\ldots,d$, there exists a natural number $k_j > \beta_j$ such that

$$\|\Delta^{k_j}_{u,j} g\|_{r_j} \leq L_j|u_j|^{\beta_j}, \quad \forall u \in \mathbb{R}^d, \quad \forall j = 1,\ldots,d.$$

Note that the anisotropic Nikol'skii class is a specific class of the anisotropic Besov class (see page 488 of [14]):

$$\mathcal{N}_{\mathbf{r},d}(\boldsymbol{\beta},.) = \mathcal{B}^{\boldsymbol{\beta}}_{\mathbf{r},\infty}(\cdot).$$

From Theorem 2.1, classical adaptive minimax anisotropic rates of convergence can be obtained. To deduce a rate of convergence, we focus on a parametrization of the form $H = P^{-1}\mathrm{diag}(h_1, \dots, h_d)P$ with P some given matrix. The practical choice of P is discussed in Section 3.4. The following result is a generalization of Corollary 7 of [19]. We set $\mathbb{N}^*$ the set of positive integers. We say that a kernel K is of order ℓ if for any non-constant polynomial Q of degree smaller than ℓ,

$$\int K(\mathbf{u})Q(\mathbf{u})d\mathbf{u} = 0.$$

Corollary 2.4. *Let P be an orthogonal matrix. 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 $\ell > \max_{j=1,\dots,d} \beta_j$. Consider $H_{min} = \bar{h} I_d$ with $\bar{h}^d = \|K\|_\infty \|K\|_1/n$ and choose for $\mathcal{H}$ the following set of bandwidths:*

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

Then, if f is bounded by a constant $B > 0$,

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

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

Corollary 2.4 is proved in Appendix A. Note that if we take for P the identity matrix we obtain the results of [19]. In this case, Theorem 3 of [14] states that up to the constant M, we cannot improve the rate achieved by our procedure. So, the latter achieves the adaptive minimax rate over the class $\mathcal{N}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L})$.

Moreover there may be P such that $f \circ P^{-1}$ is smoother than f. Then it is worth estimating $f \circ P^{-1}$ than f. This is equivalent to considering the data $P\boldsymbol{X}_i$ (which is a common preprocessing), and to estimate it with kernel $K \circ P^{-1}$ and bandwidth PHP^{-1}, hence our bandwidth parametrization of the form $H = P^{-1}\mathrm{diag}(h_1, \dots, h_d)P$. The numerical interest of such full matrices is shown in Section 3.4.2. Corollary 2.4 highlights the theoretical interest of finding P such that $f \circ P^{-1}$ is as smooth as possible. Investigating the best choice for the matrix P is beyond the scope of this paper but in practice we suggest to use the covariance matrix, see Section 3.4.2.

Finally, note that the oracle inequality of Theorem 2.1 allows us to obtain a result for other function spaces as soon as the bias term can be bounded conveniently. For instance, we achieve the optimal rate for interesting classes of functions with dominating mixed-smoothness, introduced by [6].

2.2.2. Rule of thumb (RoT)

For a general parametrization, in [30], the authors derive the formula for the AMISE expansion of the MISE and also look at the particular case of the multivariate normal density with a Gaussian kernel. More precisely, the AMISE expansion given by equation (2.10) depends on f through Ψ_f. The easiest way to minimize the AMISE is to take, for f, the multivariate Gaussian density $\mathcal{N}(\boldsymbol{m}, \Sigma)$ with mean $\boldsymbol{m}$ and covariance matrix Σ in the expression of Ψ_f (see (2.11)), combined with K, the standard Gaussian kernel, in the AMISE expression (see (2.10)). Then, the AMISE-optimal bandwidth matrix is

$$\hat{H}_{RoT} = \left(\frac{4}{n(d+2)}\right)^{\frac{1}{d+4}} \hat{\Sigma}^{\frac{1}{2}},$$

where $\hat{\Sigma}$ is the empirical covariance matrix of the data [30].

2.2.3. Least-Square Cross-Validation (UCV)

The multivariate generalization of the least-square cross-validation was developed in [29]. It can easily be observed that computations leading to the Cross-Validation criterion for univariate densities can be extended without any difficulty to the case of multivariate densities and we set

$$\hat{H}_{\text{ucv}} = \arg\min_{H} l_{\text{ucv}}(H),$$

with

$$l_{\text{ucv}}(H) = \frac{1}{n^2}\sum_i \sum_j K_H^*(\boldsymbol{X}_i - \boldsymbol{X}_j) + \frac{2}{n}K_H(\mathbf{0}),$$

where $K_H^*(\mathbf{u}) = (K_H \star \tilde{K}_H)(\mathbf{u}) - 2K_H(\mathbf{u})$, still by denoting '$\star$' the convolution product and $\tilde{K}_H(\mathbf{u}) = K_H(-\mathbf{u})$.

2.2.4. Smoothed Cross-Validation (SCV)

The Smoothed Cross-Validation (SCV) approach proposed by [10] is based on the improvement of the AMISE approximation of the MISE by replacing the first term of (2.10) with the exact integrated squared bias. Then, cross-validation is used to estimate the bias term. Therefore, the objective function for the multivariate SCV methodology is

$$l_{\text{SCV}}(H) = \frac{\|K\|^2}{n\text{det}(H)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_H \star K_H \star L_G \star L_G - 2K_H \star L_G \star L_G + L_G \star L_G)(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

where L is the pilot kernel and G the pilot bandwidth matrix and the selected bandwidth is then

$$\hat{H}_{\text{SCV}} = \arg\min_{H} l_{\text{SCV}}(H).$$

See Section 3 of [10] or Sections 2 and 3 of [5] for more details. To design the pilot bandwidth matrix, [10] restrict to the case $G = g \times I_d$ for $g \in \mathbb{R}_+^*$, whereas [5] consider full matrices.

2.2.5. Plug-in (PI)

In the same spirit as the one-dimensional SJ estimator described in Section 2.1.5, the goal of the multivariate plug-in estimator is to minimize $H \mapsto AMISE(H)$ which depends on the unknown matrix Ψ_f whose elements are given by the $\psi_{\boldsymbol{r}}$'s for all $\boldsymbol{r} = (r_1, \ldots, r_d) \in \mathbb{N}^d$ such that $|\boldsymbol{r}| = \sum_{i=1}^{d} r_i = 4$ and defined by

$$\psi_{\boldsymbol{r}} = \int f^{(\boldsymbol{r})}(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad \text{where} \quad f^{(\boldsymbol{r})} = \frac{\partial^{|\boldsymbol{r}|}f}{\partial x_1^{r_1}\ldots\partial x_d^{r_d}}.$$

In [33], the elements $\psi_{\boldsymbol{r}}$ are estimated by

$$\hat{\psi}_{\boldsymbol{r}}(G) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} L_G^{(\boldsymbol{r})}(\boldsymbol{X}_i - \boldsymbol{X}_j),$$

where, as usual, L is a pilot kernel and G a pilot bandwidth matrix. Some limitations of this approach are emphasized in [33]. This is the reason why [4] alternatively suggest to estimate Ψ_f by using

$$\hat{\Psi}_4(G) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D^{\otimes 4} L_G(\boldsymbol{X}_i - \boldsymbol{X}_j),$$

with $\otimes r$ the rth Kronecker product. Section 4.1 of [4] describes the choice of G and finally the selected bandwidth is given by

$$\hat{H}_{\mathrm{PI}} = \arg\min_H \widehat{AMISE}(H)$$

where

$$\widehat{AMISE}(H) = \frac{1}{4}\mu_2^2(K)(\mathrm{vech}H^2)^T\hat{\Psi}_4(G)(\mathrm{vech}H^2) + \frac{\|K\|^2}{n\mathrm{det}(H)}.$$

3. Numerical study

To study the numerical performances of the PCO approach, a large set of testing distributions has been considered. For the univariate case, we use the benchmark densities proposed by [21] whose list is slightly extended. See Figure B.1 in Appendix and Table B.1 for the specific definition of 19 univariate densities considered in this paper. For multivariate data, we start from the 12 benchmark densities proposed by [3] and PCO is tested on an extended list of 14 densities (see Tab. B.2 and Fig. B.2). Their definition is generalized to the case of dimensions 3 and 4 (see Tabs. B.3 and B.4 respectively). We provide 3-dimensional representations of the testing densities in Figure B.3.

3.1. PCO implementation and complexity

This section is devoted to implementation aspects of PCO and we observe that its computational cost is very competitive with respect to competitors considered in this paper. We first deal with the univariate case for which three kernels have been tested, namely the Gaussian, the Epanechnikov and the biweight kernels, respectively defined by:

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right),\ \ K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{\{|u|\leq 1\}},\ \ K(u) = \frac{15}{16}(1-u^2)^2\mathbb{1}_{\{|u|\leq 1\}}.$$

For any kernel K, $\|K_h\|^2 = h^{-1}\|K\|^2$. If K is the Gaussian kernel, $\|K\|^2 = (2\sqrt{\pi})^{-1}$, and the penalty term defined in (2.5) can be easily expressed:

$$\mathrm{pen}_\lambda(h) = \frac{\lambda\|K_h\|^2 - \|K_{h_{min}} - K_h\|^2}{n} = \frac{1}{2\sqrt{\pi}n}\left(\frac{\lambda-1}{h} - \frac{1}{h_{min}} + 2\sqrt{\frac{2}{h^2+h_{min}^2}}\right).$$

For the Epanechnikov kernel, we have $\|K\|^2 = 3/5$ and

$$\mathrm{pen}_\lambda(h) = \frac{1}{n}\left(\frac{3(\lambda-1)}{5h} - \frac{3}{5h_{min}} + \frac{3}{2}\frac{h^2 - h_{min}^2/5}{h^3}\right).$$

With a biweight kernel, since $\|K\|^2 = 5/7$, the penalty term becomes

$$\text{pen}_\lambda(h) = \frac{1}{n}\left(\frac{5(\lambda-1)}{7h} - \frac{5}{7h_{min}} + \frac{15}{8}\left(\frac{1}{h} + \frac{h_{min}^4}{21h^5} - \frac{6h_{min}^2}{21h^3}\right)\right).$$

Moreover, the loss $\|\hat{f}_{h_{min}} - \hat{f}_h\|^2$ can be expressed as

$$\begin{aligned}\|\hat{f}_{h_{min}} - \hat{f}_h\|^2 =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (K_h \star K_h)(X_i - X_j) - 2(K_h \star K_{h_{min}})(X_i - X_j)\\ &+ \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (K_{h_{min}} \star K_{h_{min}})(X_i - X_j).\end{aligned}$$

With a Gaussian kernel, this formula has a simpler expression:

$$\begin{aligned}\|\hat{f}_{h_{min}} - \hat{f}_h\|^2 =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_{\sqrt{2}h}(X_i - X_j) - 2K_{\sqrt{h^2+h_{min}^2}}(X_i - X_j)\\ &+ \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_{\sqrt{2}h_{min}}(X_i - X_j).\end{aligned}$$

Omitting terms of l_{PCO} not depending on h (see (2.4)), for the Gaussian kernel, the PCO bandwidth is obtained as follows:

$$\begin{aligned}\hat{h}_{\text{PCO}} = \arg\min_{h\in\mathcal{H}} \Bigg\{& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(K_{\sqrt{2}h}(X_i - X_j) - 2K_{\sqrt{h^2+h_{min}^2}}(X_i - X_j)\right)\\ &+ \frac{1}{n\sqrt{\pi}}\sqrt{\frac{2}{h^2+h_{min}^2}} + \frac{\lambda-1}{2nh\sqrt{\pi}}\Bigg\}.\end{aligned}$$

Similarly to $\hat{h}_{\text{UCV}}$, the expression to minimize can be computed through a $O(n^2)$ algorithm. Note that when the tuning parameter is fixed to $\lambda = 1$, the PCO bandwidth is just

$$\begin{aligned}\hat{h}_{\text{PCO}} = \arg\min_{h\in\mathcal{H}} \Bigg\{& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(K_{\sqrt{2}h}(X_i - X_j) - 2K_{\sqrt{h^2+h_{min}^2}}(X_i - X_j)\right)\\ &+ \frac{1}{n\sqrt{\pi}}\sqrt{\frac{2}{h^2+h_{min}^2}}\Bigg\}.\end{aligned}$$

One can obtain similar expressions for other kernels. Regarding the set $\mathcal{H}$, the bandwidth h is chosen in a set of real numbers built from a low-discrepancy sequence and more precisely is a rescaled Sobol sequence [28] such that we obtain a uniform sampling of the interval $[\frac{1}{n}, 1]$. We add to the sequence $h_{min} = \|K\|_\infty/n$ and finally, $\text{card}(\mathcal{H}) = 400$. This choice of h_{min} stems from the theory (see [19]) since $\|K\|_1 = 1$, but a deep study of this parameter h_{min} is led in Section 3.2.2.

For the multivariate case, similar simplifications can be used. In particular, we have

$$\|K_H\|^2 = \frac{\|K\|^2}{|\det H|}.$$

Considering the Gaussian kernel for which we have $\|K\|^2 = (2\sqrt{\pi})^{-d}$, we obtain

$$\|K_{H_{min}} - K_H\|^2 = \frac{1}{|\det(H)|(2\sqrt{\pi})^d} + \frac{1}{|\det(H_{min})|(2\sqrt{\pi})^d} - \frac{2}{\sqrt{\det(H^2 + H_{min}^2)}(2\pi)^{d/2}}$$

and using easy extensions of simplifications detailed for the univariate case, we obtain

$$\begin{aligned}\|\hat{f}_{H_{min}} - \hat{f}_H\|^2 =& \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_{\sqrt{2}H}(\boldsymbol{X}_i - \boldsymbol{X}_j) - 2K_{\sqrt{H^2+H_{min}^2}}(\boldsymbol{X}_i - \boldsymbol{X}_j)\\ &+ \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_{\sqrt{2}H_{min}}(\boldsymbol{X}_i - \boldsymbol{X}_j).\end{aligned}$$

We easily obtain $\hat{H}_{\text{PCO}}$ as

$$\begin{aligned}\hat{H}_{\text{PCO}} = \arg\min_{H\in\mathcal{H}} \Bigg\{ & \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left(K_{\sqrt{2}H}(\boldsymbol{X}_i - \boldsymbol{X}_j) - 2K_{\sqrt{H^2+H_{min}^2}}(\boldsymbol{X}_i - \boldsymbol{X}_j)\right)\\ &+ \frac{2}{n\sqrt{\det(H^2 + H_{min}^2)}(2\pi)^{d/2}} + \frac{\lambda - 1}{n|\det(H)|(2\sqrt{\pi})^d}\Bigg\}.\end{aligned}$$

The construction of $\mathcal{H}$ is similar to the case of univariate data by taking $\mathcal{H}$ such that $\text{card}(\mathcal{H}) = 16^d$. According to Corollary 2.4, since $\|K\|_1 = 1$, H_{min} is chosen equal to $\bar{h}I_d$ with

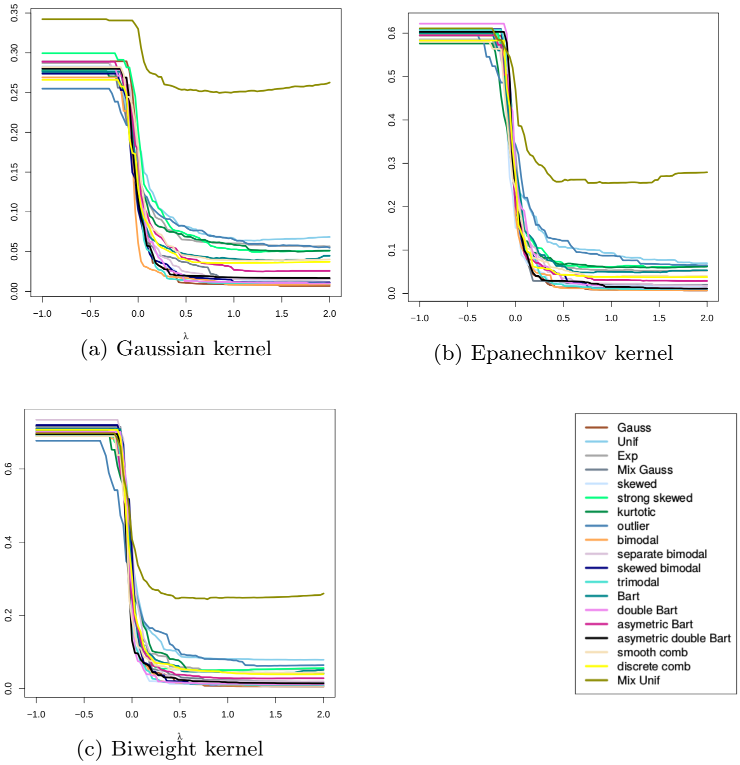
$$\bar{h}^d = \frac{\|K\|_\infty}{n}.$$

We see that the time complexity of PCO is the same as UCV, that is $O(d^3n^2|\mathcal{H}|)$. BCV and plug-in methods have the same complexity $O(d^3n^2)$, so that there is no difference between methods in terms of asymptotic complexity, except for RoT which is lighter since a single bandwidth is computed. Space complexity of PCO is also the same as UCV.

3.2. Tuning of PCO and brief numerical illustrations

3.2.1. Tuning the parameter λ

As suggested by Theorem 2 of [19] for the univariate case and Theorem 2.1 for the multivariate case, the optimal theoretical value for the tuning parameter λ is $\lambda = 1$. See arguments given in Remark 2.2 which are now confronted with a short numerical study. For this purpose, we consider the univariate case and study the risk of the PCO estimate with respect to λ. More precisely, for each benchmark density, with $n = 100$, for previous kernels, and for 20 samples, we determine successively the risk $\|f - \hat{f}\|^2$; Figure 1 provides the Monte Carlo mean of the risk over these samples in function of the PCO tuning parameter λ. We observe very similar behaviors for any density and any kernel, namely:
– very large values of the risk when $\lambda < 0$,
– an abrupt change point at $\lambda = 0$,
– and a plateau where the risk achieves its minimal value, around the value $\lambda = 1$.
Notice that the maximal range for λ considered in Figure 1 is 2 since the risk increases for larger values. We observe very similar behaviors for larger datasets (not shown). The plateau phenomenon means that in practice, considering $\lambda = 1$ instead of the true minimizer of the risk does not impact the numerical performances of PCO significantly. Thus, in subsequent numerical experiments, the tuning parameter is fixed at 1. It means that the penalty is tuning-free. This represents a great advantage compared to other kernel methodologies based for instance on Lepski-type procedures very hard to tune in practice.



(a) Gaussian kernel

(b) Epanechnikov kernel

(c) Biweight kernel

FIGURE 1. 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 three kernels with $n = 100$ observations in the univariate case.

Remark 3.1. A natural alternative would consist in detecting the abrupt jump $\hat{j}$ of one of the functions $\lambda \mapsto \hat{h}_{\text{PCO}}$ or $\lambda \mapsto \|\hat{f}_{\hat{h}} - \hat{f}_{h_{\min}}\|^2$ (see Thm. 3 of [19]) and then tuning PCO with $\lambda = \hat{j} + 1$. This alternative provides very similar results and is not considered in the sequel.

3.2.2. Tuning the parameter H_{min}

The PCO criterion depends on another crucial parameter, namely h_{min} for the univariate case and H_{min} for the multivariate case. Indeed PCO relies on comparisons with the minimal bandwidth estimator. According to Theorem 2.1 [19], h_{min} and $\det(H_{min})$ have to be larger than or equal to $\|K\|_1\|K\|_\infty/n$ and this is quite natural to consider to tune them with this lower bound. In order to test the sensitivity of PCO to this choice, some simulations have been run. More precisely, for the univariate case and 20 samples associated with the benchmark laws, the risk of PCO with different values for h_{min} has been computed and averaged over the 20 samples. For the sake of simplicity, we have only considered the Gaussian kernel, for which $\|K\|_1 = 1$. Then, each averaged risk has been normalized according to the minimal averaged risk among all possible h_{min}'s and finally averaged over all distributions. The resulting graphs for $n = 1000$ and

$$h_{min} \in \left\{ \frac{\|K\|_\infty}{n^3}, \frac{\|K\|_\infty}{n^2}, \frac{\|K\|_\infty}{2n}, \frac{\|K\|_\infty}{n}, \frac{2\|K\|_\infty}{n}, \frac{1}{n}, \frac{\|K\|_\infty}{n^{1/2}}, \frac{\|K\|_\infty}{n^{1/4}}, \frac{\|K\|_\infty}{n^{1/8}} \right\}$$

are displayed in Figure 2. Figure 2 also displays a similar study for $d \in \{2, 3, 4\}$ by replacing h_{min} with $\det(H_{min})$. For the sake of completeness, the boxplots of the risks are presented in Appendix C Figures C.1–C.4. We observe that in some situations, the value $\|K\|_\infty/\sqrt{n}$ may give slightly better results in particular for small dimensions and for some densities. However, when the density to be estimated is irregular, considering $\|K\|_\infty/\sqrt{n}$ may be a bad choice since the optimal value for the bandwidth is too small and is smaller than the minimal value of the grid. Finally, taking $h_{min} = \|K\|_\infty/n$ for $d = 1$ and $\det(H_{min}) = \|K\|_\infty/n$ provides very satisfying results for all situations, meaning that this choice is robust. One would expect the method to be very sensitive to parameters h_{min} and $\det(H_{min})$. Actually, this is not the case as shown by Figure 2 as soon as the latter lie in a convenient zone, which is a good news for practical purposes.

3.2.3. Brief numerical illustration

Before comparing PCO to classical approaches for kernel density estimation, we briefly illustrate its numerical performances in the multidimensional setting. For this purpose, we implement in Figure 3 the square root of the ISE for one realization with respect to all possible diagonal bandwidths matrices on two benchmark densities, namely Asymmetric Bimodal and Asymmetric Fountain when $n = 100$. Figure 3 shows that the bandwidth selected by PCO is – or is close to – the optimal bandwidth. This short illustration only deals with $d = 2$, one given sample size and a small set of benchmark densities. But similar behaviors are observed in more general settings. Next sections are devoted to deep numerical comparisons with the classical kernel approaches described in Section 2.

3.3. Numerical comparisons for univariate density estimation

In this section, for all methods except PCO, the bandwidth selection is performed through the `R stats` package [23].

For each testing density f and each methodology described in Section 2.1 and denoted *meth*, we compute the square root of the Integrated Square Error defined in (2.1) associated with the bandwidth selected by each methodology and viewed as a function of f. With a slight abuse of notation, we denote it $ISE^{1/2}_{\text{meth}}(f)$. With $n = 100$, Table 1 provides $\overline{ISE}^{1/2}_{\text{meth}}(f)$, the Monte Carlo mean over 20 samples, for each kernel. Since results for both variants of the Rule of Thumb approach are very close (see Fig. 4), we only give results associated with Expression (2.6). We also provide in Table D.1 in Appendix D a comparison with the Goldenshluger-Lepski approach which has inspired the PCO criterion. For this comparison we have used the following Goldenshluger-Lepski criterion

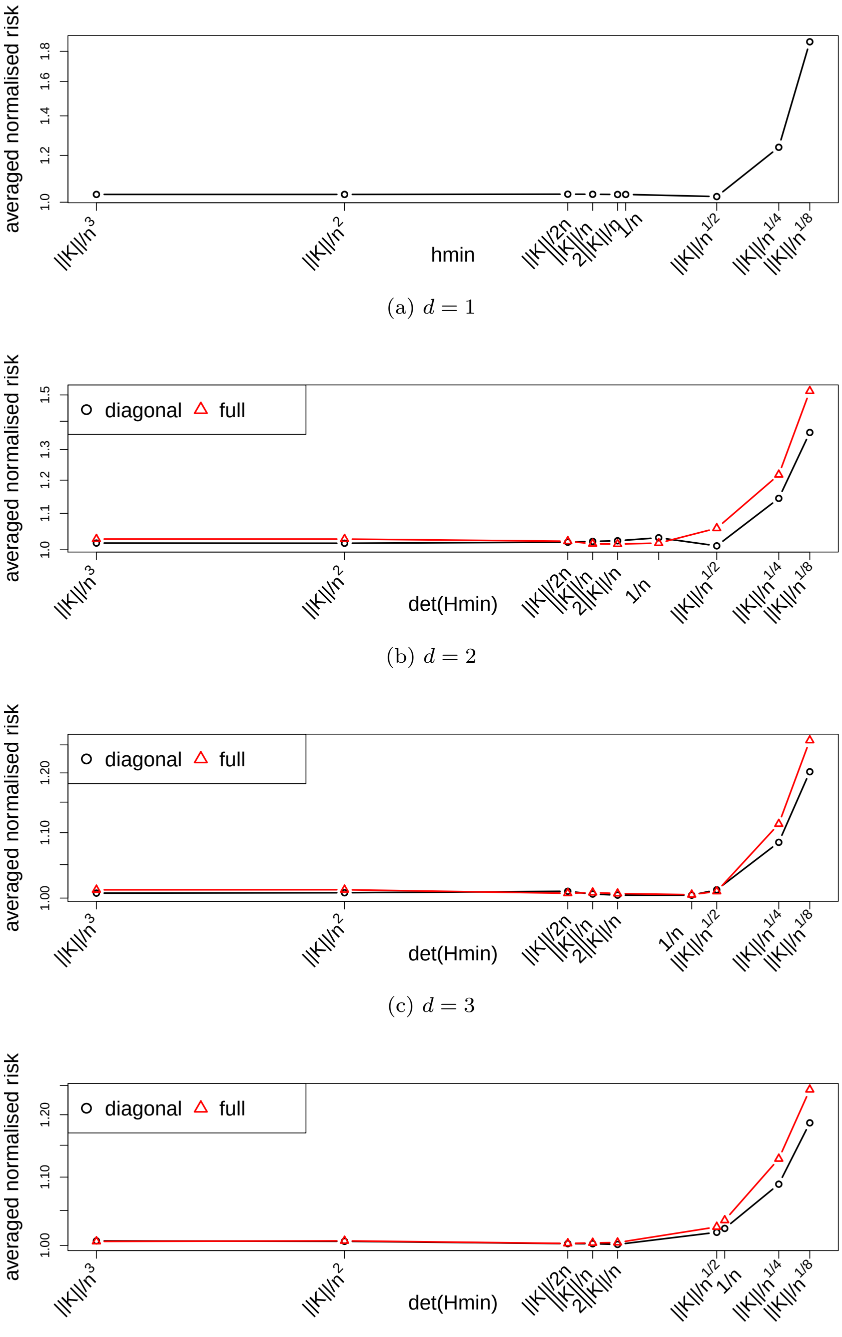
$$\hat{h}_{\text{GL}} = \arg\min_{h\in\mathcal{H}} \{A(h) + V(h)\}$$

where

$$A(h) = \sup_{h'\in\mathcal{H}} \{||\hat{f}_{h'} - \hat{f}_{h,h'}||^2 - V(h')\}_+, \qquad V(h) = \kappa\frac{||K||_1^2.||K||_2^2}{nh}$$

and $\hat{f}_{h,h'} = K_{h'} \star \hat{f}_h$. The hyperparameter κ is notoriously hard to tune. In our implementation, we choose to tune κ on fifty samples of a Gaussian distribution with a Gaussian kernel. That is, κ is taken as the mean of the fifty values that minimizes the $\mathbb{L}_2$ loss. As illustrated in Table D.1, the results of PCO and Goldenshluger-Lepski approach are very similar but with a very different computational cost (the Goldenshluger-Lepski is 750 times slower in our implementation, which approximately corresponds to the theoretical complexity: $O(n^2|\mathcal{H}|)$ for PCO against $O(n^2|\mathcal{H}|^2)$ for Goldenshluger-Lepski) thus the Goldenshluger-Lepski approach is no more used in the following.

Considering Monte Carlo mean values in bold of Table 1 (such values are not larger than 1.05 times the minimal one), we observe that overall, PCO achieves very satisfying results that are very close to those of UCV and SJste. For most of densities, BCV and RoT are outperformed by other approaches. When comparing with other methodologies, PCO is outperformed for 4 densities, namely G, U (for the Epanechnikov kernel), Sk and



(a) $d = 1$

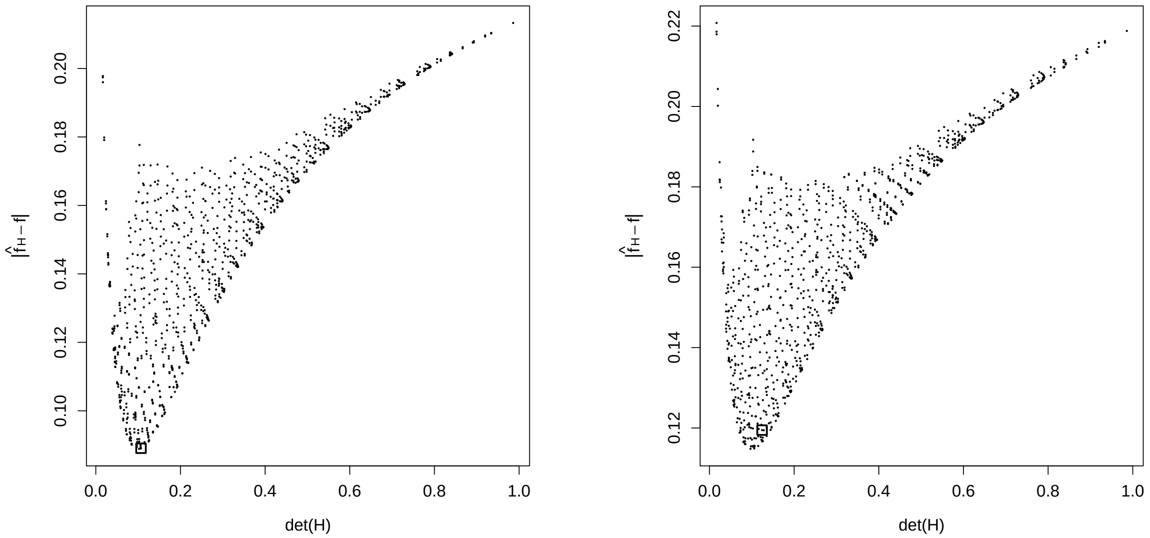
(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

FIGURE 2. Normalised risk averaged over all laws *versus* h_{min} and $\det(H_{min})$ with $n = 1000$ for $d \in \{1, 2, 3, 4\}$.

O. Observe that the smooth unimodal densities Sk and O have a shape close to G, the Gaussian one. Actually, as expected, competitors (associated with the Gaussian kernel) based on a pilot kernel tuned on Gaussian densities (RoT and SJ) outperform other methodologies for such densities. Furthermore, even when PCO, UCV



(a) ABi

(b) DF

FIGURE 3. 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 densities ABi and DF, with $n = 100$. The square corresponds to the bandwidth selected by PCO.

and SJste are not the best methodologies for a given density f, their performances are quite satisfying. It is not the case for RoT, BCV and SJdpi that achieve bad results for the densities Sk+, DC and MU. Finally, the preliminary results of Table 1 show that the kernel choice has weak influence on the results, which is confirmed for an extended simulation study lead with many values for n (not shown). So, for the subsequent results, we only focus on the Gaussian kernel.

In view of satisfying performances of PCO for $n = 100$, such preliminary results are extended to larger values of n with in mind a clear and simple though complete comparison between PCO and other methodologies. For this purpose, we still consider, for each density f, the square root of the Integrated Square Error for the bandwidth selected by each methodology, denoted $ISE^{1/2}_{\text{meth}}(f)$, and we display in Figure 4 the median over 20 samples of the ratio $ISE^{1/2}_{\text{meth}}(f)/ISE^{1/2}_{\text{PCO}}(f)$, namely

$$r^{\text{med}}_{\text{meth/PCO}}(f) := \text{median}\left(\frac{ISE^{1/2}_{\text{meth}}(f)}{ISE^{1/2}_{\text{PCO}}(f)}\right).$$

for meth $\in \{\text{RoT0}, \text{RoT}, \text{UCV}, \text{BCV}, \text{SJste}, \text{SJdpi}\}$ and $n \in \{100, 500, 1000, 10000\}$.

A brief look at the results of Figure 4 confirms that PCO, even not dramatically bad, is not the best methodology for densities 'Gauss' (G) and 'Skewed' (Sk). Similar conclusions are true for 'Outlier' (O), except when n is large. Actually, the larger n, the better the behavior of PCO with respect to all other competitors. In particular, except for Sk, PCO outperforms all competitors when $n = 10000$. For small values of n, when considering the densities 'Bimodal' (Bi), 'Skewed Bimodal' (SkB) and 'Double Bart' (DB), RoT0, SJste and SJdpi achieve better results than PCO. Otherwise, PCO is preferable. Actually, as already observed for $n = 100$, PCO and UCV behave quite similarly except for some densities for which performances of UCV deteriorate dramatically when n increases (see 'Exp' (E), 'Kurtotic' (K) and 'Outlier' (O)). Even if not reported, our simulation study shows that the variance of the $ISE^{1/2}_{\text{UCV}}(f)$ is much larger than for other methodologies. In

TABLE 1. Monte Carlo mean of $ISE^{1/2}_{\text{meth}}(f)$ over 20 trials with $n = 100$ for 6 methodologies described in Section 2.1 tested on the 19 one-dimensional densities and for different kernels ($K \in \{\text{Gaussian (G)}, \text{Epanechnikov (E)}, \text{biweight (B)}\}$). The Monte Carlo mean $\overline{ISE}^{1/2}_{\text{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\text{meth}} \overline{ISE}^{1/2}_{\text{meth}}(f)$.

	RoT			UCV			BCV		
	G	E	B	G	E	B	G	E	B
G	**0.06**	**0.07**	**0.07**	0.08	0.08	0.08	0.07	**0.07**	**0.07**
U	**0.26**	0.27	**0.28**	**0.25**	0.29	**0.28**	0.30	0.33	0.32
E	0.29	0.30	0.28	**0.24**	**0.23**	**0.22**	0.31	0.34	0.32
MG	0.26	0.29	0.28	0.13	**0.13**	0.14	0.17	0.26	0.23
Sk	**0.08**	**0.09**	**0.08**	0.10	0.11	0.08	**0.09**	**0.09**	0.08
Sk+	0.36	0.39	0.39	**0.22**	**0.24**	**0.22**	0.45	0.48	0.46
K	0.32	0.37	0.34	0.25	**0.24**	**0.20**	0.48	0.50	0.49
O	**0.20**	**0.22**	**0.23**	0.25	0.25	**0.24**	1.37	1.40	1.39
Bi	**0.09**	0.09	0.10	**0.09**	0.09	0.09	0.12	0.13	0.13
SB	0.21	0.23	0.23	**0.12**	0.12	0.13	**0.12**	**0.11**	**0.12**
SkB	0.10	0.11	0.11	0.10	**0.10**	**0.10**	0.12	0.14	0.13
T	0.10	0.11	0.11	0.10	0.10	0.10	0.13	0.15	0.14
B	0.23	**0.23**	0.23	0.21	0.23	0.22	0.23	0.24	0.24
DB	**0.09**	**0.10**	**0.11**	**0.10**	0.11	0.11	0.12	0.14	0.14
AB	**0.16**	**0.17**	**0.17**	**0.16**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**
ADB	**0.12**	0.12	**0.12**	0.13	0.12	0.13	0.15	0.16	0.15
SC	0.29	0.31	0.31	**0.20**	**0.20**	**0.21**	0.32	0.34	0.31
DC	0.34	0.36	0.35	**0.19**	**0.19**	**0.20**	0.35	0.35	0.35
MU	0.65	0.67	0.66	**0.51**	**0.51**	**0.49**	0.77	0.76	0.76

	SJste			SJdpi			PCO		
	G	E	B	G	E	B	G	E	B
G	0.07	0.08	**0.07**	**0.07**	0.07	**0.07**	0.08	0.08	0.08
U	**0.25**	**0.25**	**0.27**	**0.25**	**0.26**	**0.27**	**0.26**	0.30	**0.28**
E	**0.23**	**0.23**	**0.23**	0.25	0.25	0.24	**0.24**	**0.23**	**0.22**
MG	**0.12**	**0.14**	**0.12**	0.13	0.17	0.15	0.13	**0.13**	0.14
Sk	0.09	0.10	**0.08**	0.09	0.09	**0.08**	0.11	0.11	0.09
Sk+	**0.23**	0.27	0.26	0.26	0.31	0.30	**0.22**	**0.24**	**0.22**
K	**0.22**	**0.24**	**0.20**	0.24	0.28	0.24	0.23	**0.24**	**0.21**
O	0.22	0.23	**0.24**	**0.21**	**0.23**	**0.24**	0.24	0.26	0.26
Bi	**0.09**	**0.08**	**0.09**	**0.08**	**0.08**	**0.09**	**0.09**	0.09	**0.09**
SB	**0.12**	**0.11**	**0.11**	**0.12**	**0.11**	**0.12**	**0.12**	0.14	0.13
SkB	**0.09**	**0.10**	**0.10**	**0.09**	**0.10**	**0.10**	0.10	**0.10**	**0.10**
T	**0.09**	**0.10**	**0.10**	**0.09**	**0.10**	**0.10**	0.10	0.10	**0.10**
B	0.22	0.24	0.23	0.23	0.23	0.23	**0.20**	**0.22**	**0.21**
DB	**0.09**	**0.10**	**0.10**	**0.09**	**0.10**	**0.10**	**0.10**	0.12	0.11
AB	**0.16**	**0.17**	**0.17**	**0.16**	**0.17**	**0.17**	**0.16**	**0.17**	**0.17**
ADB	**0.12**	**0.10**	**0.12**	**0.12**	**0.10**	**0.12**	0.13	0.12	0.13
SC	0.21	0.21	0.22	0.23	0.24	0.25	**0.20**	**0.20**	**0.21**
DC	**0.20**	**0.20**	**0.20**	0.29	0.31	0.30	**0.19**	**0.19**	**0.21**
MU	0.54	0.56	0.56	0.57	0.59	0.59	**0.50**	**0.50**	**0.49**

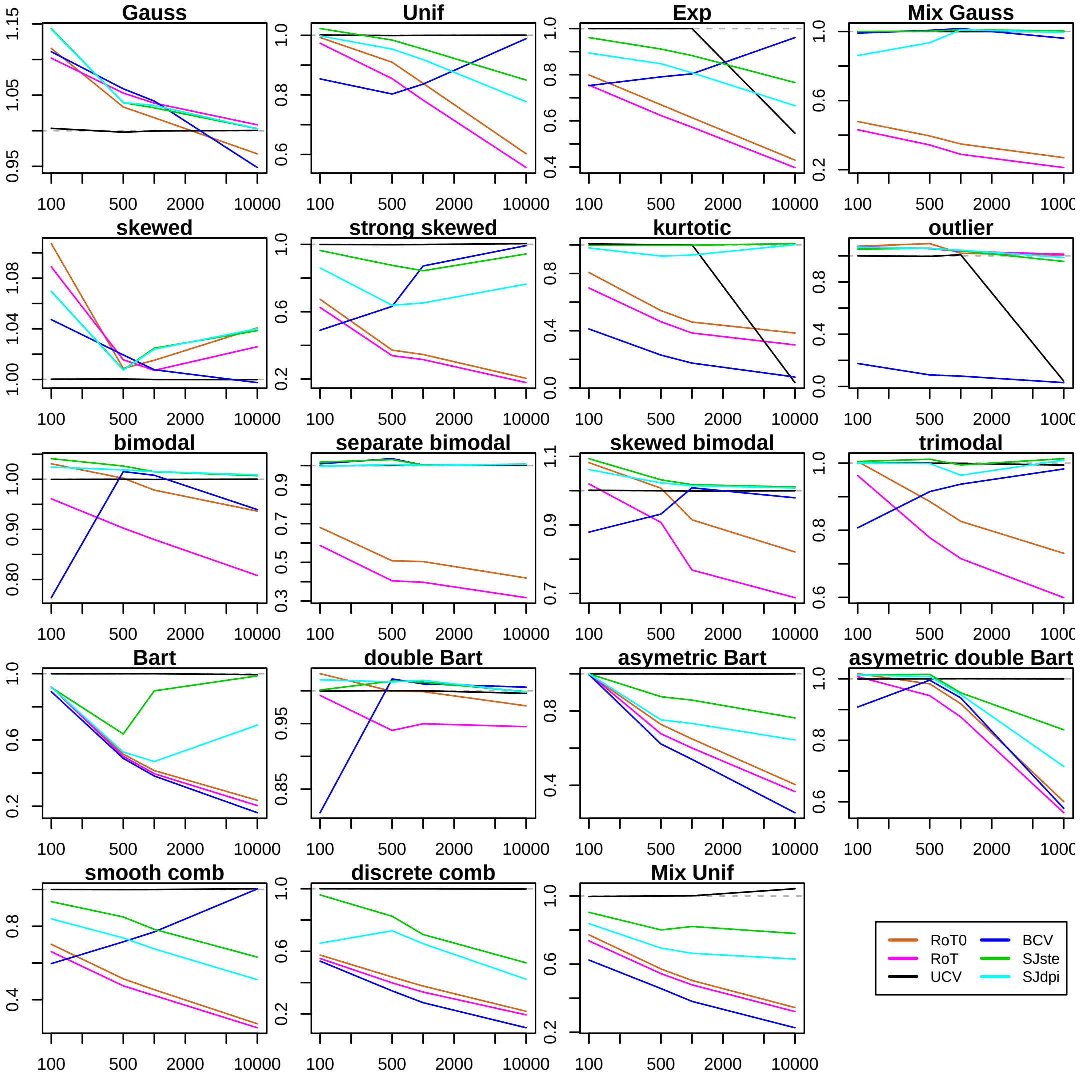


FIGURE 4. Median over 20 samples of the ratio $ISE_{\mathrm{PCO}}^{1/2}(f)/ISE_{\mathrm{meth}}^{1/2}(f)$ for meth $\in$ $\{\mathrm{RoT0}, \mathrm{RoT}, \mathrm{UCV}, \mathrm{BCV}, \mathrm{SJste}, \mathrm{SJdpi}\}$ with the Gaussian kernel versus the sample size.

particular, PCO has not to face with this issue. Stability with respect to the trial is a non-negligible advantage of PCO.

Finally, for sake of completeness, each approach is compared to the best one through the graph of the mean over all densities f of the ratio of

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

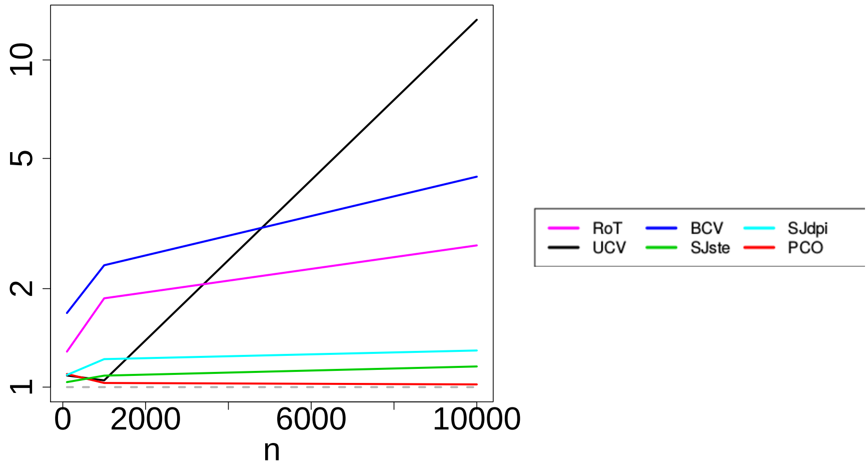


FIGURE 5. Graph of the mean over all densities f of the ratio of $r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$ for meth $\in \{\mathrm{RoT}, \mathrm{UCV}, \mathrm{BCV}, \mathrm{SJste}, \mathrm{SJdpi}, \mathrm{PCO}\}$ with the Gaussian kernel versus the sample size.

Namely, for meth $\in \{\mathrm{RoT}, \mathrm{UCV}, \mathrm{BCV}, \mathrm{SJste}, \mathrm{SJdpi}, \mathrm{PCO}\}$ and $n \in \{100, 1000, 10000\}$, we display in Figure 5:

$$\bar{r}_{\mathrm{meth}/\min} := \frac{1}{19} \sum_{f} r_{\mathrm{meth}/\min}(f).$$

At first glance, we note that instability of UCV has strong bad consequences when compared to the best method for large values of n. As explained for instance in [16], it is well-known that UCV "leads to a small bias but large variance" and "often breaks down for large samples". This synthetic figure shows that for small values of n, PCO achieves nice performances and is very competitive. It is also the case for some other methods (UCV, SJste and SJdpi), but PCO clearly outperforms any other methodology for large datasets. This result is quite surprising since PCO is not based on asymptotic approximations.

3.4. Numerical comparison for multivariate density estimation

For multivariate data, we perform selection by usual kernel methods by using the ks package of R [9]. Different sample sizes $n \in \{100, 1000, 10000\}$ with the Gaussian kernel have been tested.

3.4.1. Diagonal bandwidth matrices

In this section, we compare PCO with UCV, SCV and PI. For each methodology, the bandwidth matrix is chosen among a set of diagonal matrices. More precisely, the diagonal terms are built from a rescaled Sobol sequence such that each of them is larger than $(||K||_\infty/n)^{1/d}$ and smaller than 1. The mean over 20 samples of the square root of the ISE is given in Table D.2 for bivariate data, in Table D.3 for trivariate data and in Table D.4 for 4-dimensional data (see Appendix D). We also provide synthetic graphs to outline comparisons between each estimator. More precisely, Figure 6 displays boxplots of the ratio $\frac{\overline{ISE}^{1/2}_{\mathrm{meth}}}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}}$ for each methodology over our benchmark densities for $d = 2, 3, 4$ and for different sample sizes. We also provide simple summary graphs in Figure 7, in the same spirit as Figure 5.

Analyzing results of Table D.2 devoted to the dimension $d = 2$, we note that, as expected, performances of all methodologies improve significantly when n increases for all benchmark densities, except for UCV whose

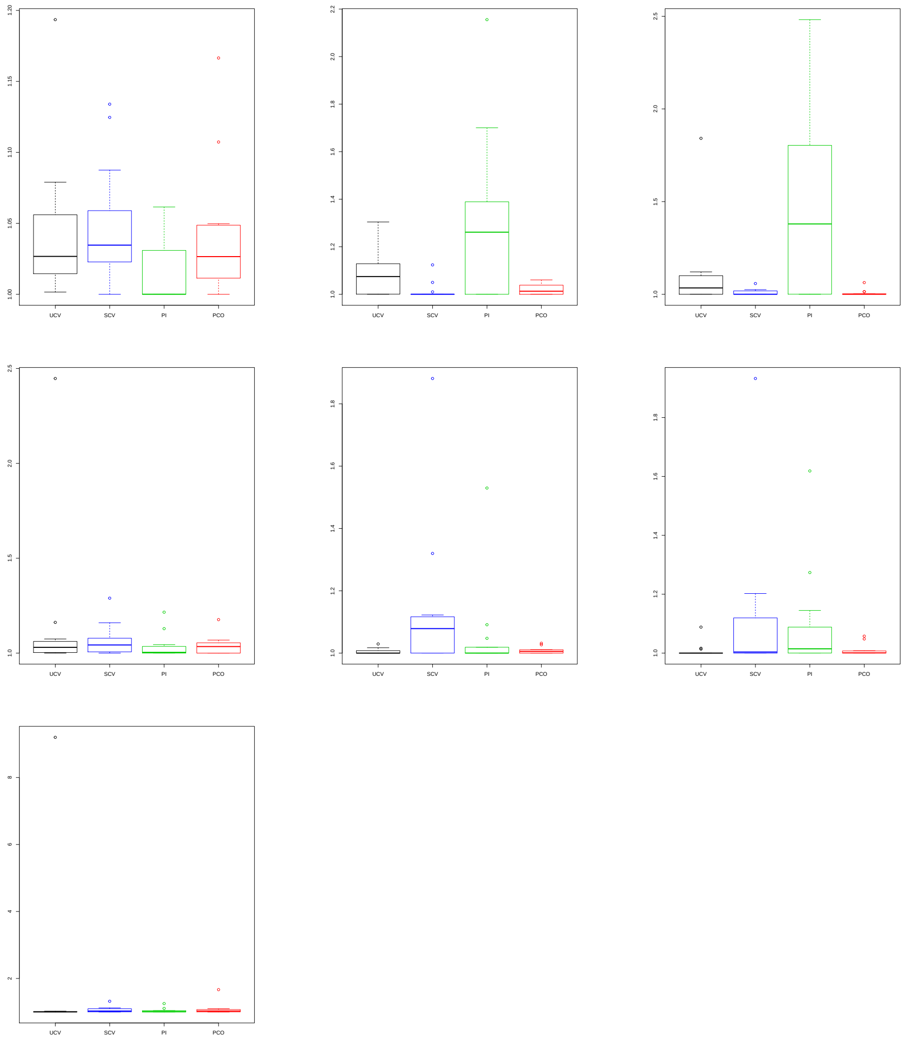
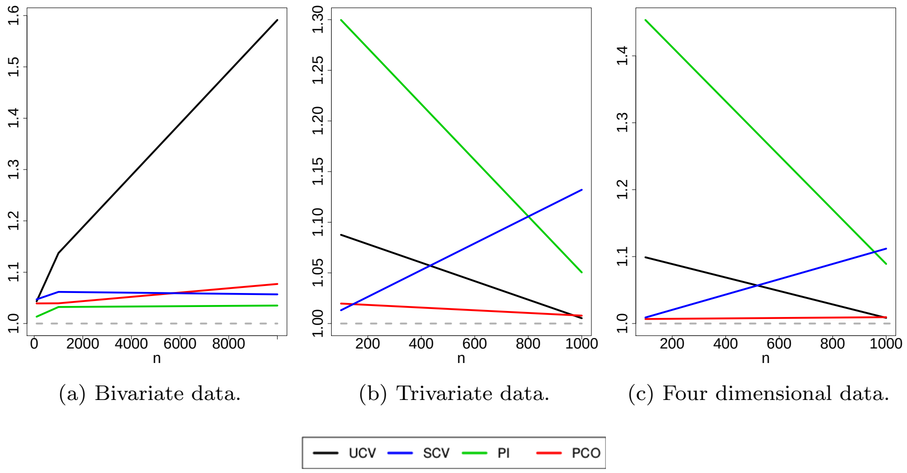


FIGURE 6. Boxplots of the ratio $\dfrac{\overline{ISE}^{1/2}_{\text{meth}}}{\min_{\text{meth}} \overline{ISE}^{1/2}_{\text{meth}}}$ over the 14 test densities described in Tables B.2, B.3 and B.4 for the diagonal case. First row: $n = 100$; second row: $n = 1000$; third row: $n = 10000$. First column: $d = 2$; second column: $d = 3$; third column: $d = 4$.

performances deteriorate for D and SK+. Actually, as explained in Section 3.3, UCV suffers from instability leading to break down issues for large datasets. PCO achieves very satisfying performances except for Sk+ and for UG when $n = 100$. It is also the case to a less extent for CG and U when $n = 10000$. These conclusions are in line with those of Section 3.3. Even if PI achieves bad results for AF (for which PCO or UCV are preferable), it remains the best methodology for bivariate data when diagonal bandwidths are considered. See the left columns of Figures 6 and 7.

Now, let us consider 3 and 4-dimensional data for which the studied sample size is not larger than 1000 to avoid too expensive computational time for all methodologies. Tables D.3 and D.4 show that all kernel strategies suffer from the curse of dimensionality for a non-negligible set of benchmark densities. See the results for irregular spiky densities Sk+, D, K and AF. Note that these densities have also strong correlations between components of the X_i's. Whereas PI achieves good results for $d = 2$, it is no longer the case for $d \geq 3$ and



(a) Bivariate data. (b) Trivariate data. (c) Four dimensional data.

FIGURE 7. Graph of the mean over all densities f of the ratio of $r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$ for meth $\in \{\mathrm{UCV}, \mathrm{SCV}, \mathrm{PI}, \mathrm{PCO}\}$ versus the sample size.

$n = 100$ due to many stability issues. This can be explained by the fact that the pilot bandwidth is proportional to identity, which is not convenient for many densities. Furthermore, whereas for $d = 2$, a closed form for $\hat{H}_{PI}$ exists, it not the case for $d \geq 3$ and optimization algorithms are necessary. When comparing methodologies between them, by analyzing Figures 6 and 7, we observe that relative performances of PI and UCV improve when n increases. When $n = 1000$, both PCO and UCV are very competitive, whereas SCV has to be avoided. However, PCO remains the best methodology for any density and for $n \in \{100, 1000\}$, except for two cases: Sk and $n = 100$ when $d = 3$ and U and $n = 100$, when $d = 4$ (see Tabs. D.3 and D.4).

To summarize, this simulation study shows that when the ratio n/d is large enough, PI is a very good strategy when considering diagonal bandwidth matrices. These graphs emphasize the remarkable property of stability of PCO. In particular, for this reason, PCO seems to be the best kernel strategy and is preferable to UCV, SCV and PI as soon as $d \geq 3$ as soon as very few is known about the density to estimate (see for instance the synthetic Fig. 7).

3.4.2. Symmetric definite positive bandwidth matrices

In this section, we investigate possible improvements of PCO for the general case, namely by considering a suitable subset of symmetric definite positive bandwidth matrices. The goal is then to detect hidden correlation structures of components of the X_i's and our strategy is based on the eigendecomposition of the covariance matrix of the data. For this purpose, let us denote $\hat{\Sigma}$ the empirical covariance matrix of the data and $\hat{\Sigma} = P^{-1} D_{\hat{\Sigma}} P$ its eigendecomposition, where $D_{\hat{\Sigma}}$ is diagonal. Then, we consider $\mathcal{H}$, the set of matrices of the form $P^{-1} D P$, where D is diagonal and the diagonal terms are built from a rescaled Sobol sequence such that each of them is larger than $(||K||_\infty/n)^{1/d}$ and smaller than 1.

Table D.8 in Appendix D emphasizes the benefits of using a full matrix rather than a diagonal one. The results in green reflect improvements of the use of a full matrix while the red ones correspond to the cases where the full matrix deteriorates numerical results. It is clear that using a full matrix can give worse results only when the sample size is small. In all other cases, the full matrix gives, at worse, similar results. It is also clear that taking a full matrix is significantly advantageous when the distribution has a strong privileged direction, as for instance CG, Sk+, D and AF.

We compare PCO with all other methodologies based on symmetric definite positive bandwidth matrices. The mean over 20 samples of the square root of the ISE is given in Table D.5 for bivariate data, in Table D.6

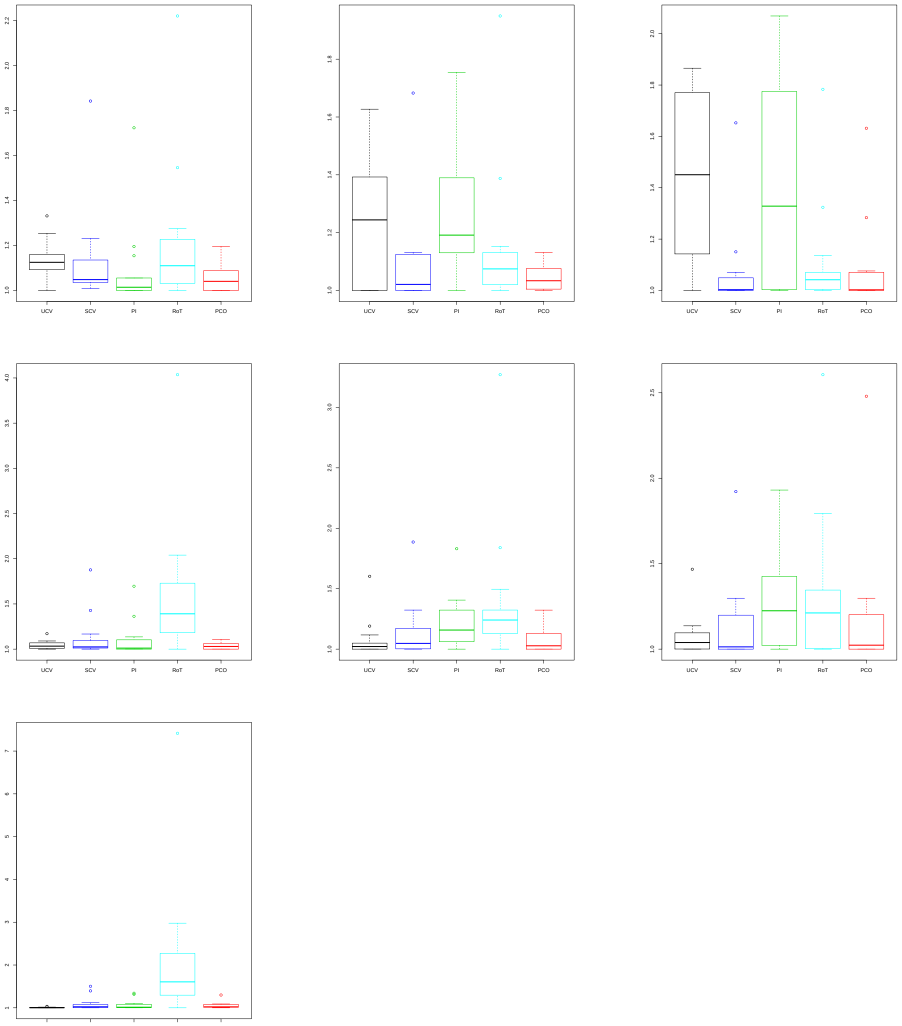
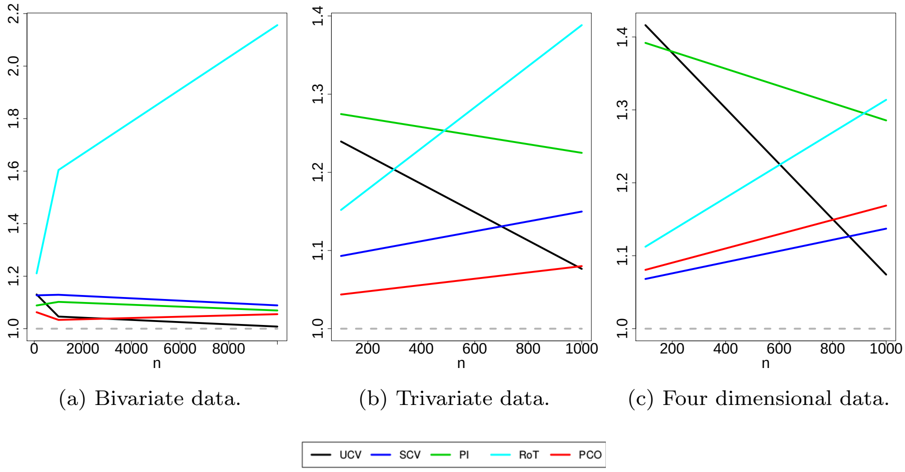


FIGURE 8. Boxplots of the ratio $\frac{\overline{ISE}^{1/2}_{\mathrm{meth}}}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}}$ over the 14 test densities described in Tables B.2, B.3 and B.4 for the general case. First row: $n = 100$, second row: $n = 1000$, third row: $n = 10000$. First column: $d = 2$, second column: $d = 3$, third column: $d = 4$.

for trivariate data and in Table D.7 for 4-dimensional data (see Appendix D), whereas Figure 8 (resp. Fig. 9) is the analog of Figure 6 (resp. Fig. 7).

Let us analyse values of the risk provided by Tables D.5, D.6 and D.7 to evaluate the gain or the loss of this new setting. First of all, we observe that for $d \geq 3$, bad results obtained by diagonal bandwidths for estimating Sk+, D, K and AF are not really improved. Secondly, as an illustrative example, let us consider purely Gaussian distributions. As expected, on the one hand, performances of PCO improve for CG as desired by using the matrix bandwidth parametrization for any value of n and any value of d. It is also the case for most of other methodologies; note however two exceptions for PI (with $n = 1000$) and UCV (with $n = 100$) for the 4-dimensional cases. On the other hand, as also expected, there is no benefit from using kernel rules with non-diagonal bandwidth for estimating the density UG. For PI and UCV, in some situations, results are even worse. More generally, we observe that results of PCO never deteriorate in this new setting with some clear



(a) Bivariate data. (b) Trivariate data. (c) Four dimensional data.

FIGURE 9. Graph of the mean over all densities f of the ratio of $r_{\mathrm{meth}/\min}(f) := \frac{\overline{ISE}^{1/2}_{\mathrm{meth}}(f)}{\min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)}$ for meth $\in \{\mathrm{UCV}, \mathrm{SCV}, \mathrm{PI}, \mathrm{RoT}, \mathrm{PCO}\}$ versus the sample size.

improvements but only in few situations (for instance for $d = 2$ and with benckmark densities Sk+, D and AF). Other procedures have mixed results with some deteriorations or some improvements. For instance, when $d = 2$, the new setting improves results of SCV and PI for Sk+ but deteriorates for Sk. We can however observe that except for CG, when $d \geq 3$, results of UCV when $n = 100$ (resp. PI when $n = 1000$) never improve and even, in many situations, deteriorate.

We now compare different methodologies by analyzing Figures 8 and 9. Hierarchy between strategies and conclusions can differ significantly from those of Section 3.4.1. We first observe that PCO is still very stable, except for the case $d = 4$ for which, in particular, estimation of Sk+ is much worse than for other strategies. RoT achieves nice results for the densities UG, CG and U (see Tabs. D.5, D.6 and D.7) but this approach, which is very unstable, is outperformed by the other ones in particular when n is large. As for the case of diagonal bandwidth matrices, PI is very satisfying for $d = 2$ but suffers from the curse of dimensionality with poor stability properties as soon as d/n is large. It is also the case for UCV but note that the latter outperforms all other strategies when d/n is small. Besides, UCV is known for working well for moderate sample size and "neither behaving well for rather small nor for rather large samples" [16] (in our context of dimension 3 or 4, $n = 1000$ is rather a moderate sample size than a large one). For many situations SCV and PCO have a similar behavior but for $d \geq 3$ and n small, SCV is preferable, maybe due to over-smoothing properties of SCV. Note however that PCO is more stable than SCV for $d \leq 3$.

To summarize, these conclusions and numerical results show that in full generality, thanks to its stability properties, PCO is probably the best strategy to adopt for most of densities and for not too large values of n.

4. Conclusion

As a general remark, we see from the boxplots resulting from our simulation studies that PCO has a stable behavior. In the univariate case, its performance is never far from optimal. In this sense, simulations corroborate what was expected from theory and validate the choice of the tuning constant in the penalty term as its optimal asymptotic value which is equal to 1. Furthermore, we have shown that the choice of the parameter h_{min} is not very sensitive and taking $h_{min} = \|K\|_\infty/n$ 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. To summarize the results for multivariate data, the performance of PCO is always close to the optimal for a diagonal bandwidth. For full bandwidth this remains true in dimension 2 and also in dimension 3 and 4 as long as the correlations are not too strong. We did not check what happens in dimension larger than 5, partly because things are becoming harder from a computational point of view and partly because kernel density estimation is unlikely to be a relevant method to be used when the dimension space increases (this is the curse of dimensionality). As compared to other methods it is not always the best competitor (but you will never beat the rule of thumb for instance when the true density happens to be Gaussian) but it has the advantage of staying competitive in any situation. In the univariate case, it is never far from cross-validation methods for small sample sizes and is better for large sample sizes while it tends to be always better than smoothed plug-in methods. Talking about future directions of research, it would be interesting to develop PCO, both from a theoretical and a practical perspective for other losses than the $\mathbb{L}_2$ loss. The cases of the $\mathbb{L}_1$ loss, which has been extensively studied by L. Devroye (see for instance [7]), or the Hellinger loss are of special interest because they correspond to some intrinsic quantities which stay invariant under some change of the dominating measure. We also believe that the PCO approach is relevant for other estimator selection problems than bandwidth selection for kernel estimation but this is another story...

Acknowledgements. The authors would like to thank the anonymous referees for valuable comments and suggestions.

Appendix A. Proofs

The notation $\square$ denotes an absolute constant that may change from line to line. We denote $\hat{H} = \hat{H}_{PCO}$ and $\langle\cdot,\cdot\rangle$ the scalar product associated with $\|\cdot\|$.

A.1 Proof of Theorem 2.1

The proof uses the lower bound (A.2) stated in the next proposition.

Proposition A.1. *Assume that K is symmetric and $\int K(\mathbf{u})d\mathbf{u} = 1$. Assume also that $\det(H_{min}) \geq \|K\|_\infty \|K\|_1/n$. Let $\Upsilon \geq (1 + 2\|f\|_\infty \|K\|_1^2)\|K\|_\infty/\|K\|^2$. For all $x \geq 1$ and for all $\eta \in (0, 1)$, with probability larger than $1 - \square|\mathcal{H}|e^{-x}$, for all $H \in \mathcal{H}$, each of the following inequalities holds:*

$$\|f - \hat{f}_H\|^2 \leq (1+\eta)\left(\|f - f_H\|^2 + \frac{\|K_H\|^2}{n}\right) + \square\frac{\Upsilon x^2}{\eta^3 n},$$
$$\|f - f_H\|^2 + \frac{\|K_H\|^2}{n} \leq (1+\eta)\|f - \hat{f}_H\|^2 + \square\frac{\Upsilon x^2}{\eta^3 n}.$$

The proof of this proposition is an easy generalization of the proof of Proposition 4.1 of [20] (combined with their Prop. 3.3) to the case of bandwidth matrices. We now give a general result for the study of $\hat{f} := \hat{f}_{\hat{H}}$, which is the analog of Theorem 9 of [19]. We set for any $H \in \mathcal{H}$,

$$\text{pen}_\lambda(H) := -\frac{\|K_{H_{min}} - K_H\|^2}{n} + \lambda\frac{\|K_H\|^2}{n}.$$

Theorem A.2. *Assume that K is symmetric and $\int K(\mathbf{u})d\mathbf{u} = 1$. Assume also that $\det(H_{min}) \geq \|K\|_\infty\|K\|_1/n$ and $\|f\|_\infty < \infty$. Let $x \geq 1$ and $\theta \in (0, 1)$. With probability larger than $1 - C_1|\mathcal{H}|\exp(-x)$, for any $H \in \mathcal{H}$,*

$$\begin{aligned}(1-\theta)\|\hat{f}_{\hat{H}} - f\|^2 \leq (1+\theta)\|\hat{f}_H - f\|^2 &+ \left(\text{pen}_\lambda(H) - 2\frac{\langle K_H, K_{H_{min}}\rangle}{n}\right)\\ &- \left(\text{pen}_\lambda(\hat{H}) - 2\frac{\langle K_{\hat{H}}, K_{H_{min}}\rangle}{n}\right) + \frac{C_2}{\theta}\|f_{H_{min}} - f\|^2\\ &+ \frac{C(K)}{\theta}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2\det(H_{min})}\right),\end{aligned}$$

where C_1 and C_2 are absolute constants and $C(K)$ only depends on K.

The oracle inequality directly follows from this theorem, see Section 5.2 in [19].

Proof of Theorem A.2

Let $\theta' \in (0,1)$ be fixed and chosen later. Following [19], we can write, for any $H \in \mathcal{H}$,

$$\|\hat{f}_{\hat{H}} - f\|^2 \leq \|\hat{f}_H - f\|^2 + \Big(\mathrm{pen}_\lambda(H) - 2\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle\Big) - \Big(\mathrm{pen}_\lambda(\hat{H}) - 2\langle \hat{f}_{\hat{H}} - f, \hat{f}_{H_{min}} - f\rangle\Big). \quad \text{(A.1)}$$

Then, for a given H, we study the term $2\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle$ that can be viewed as an ideal penalty. Let us introduce the degenerate U-statistic

$$U(H, H_{min}) = \sum_{i \neq j} \langle K_H(. - \boldsymbol{X}_i) - f_H, K_{H_{min}}(. - \boldsymbol{X}_j) - f_{H_{min}}\rangle$$

and the following centered variable

$$V(H, H') =< \hat{f}_H - f_H, f_{H'} - f > .$$

We have the following decomposition of $\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle$:

$$\begin{aligned}\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle = {} & \frac{\langle K_H, K_{H_{min}}\rangle}{n} + \frac{U(H, H_{min})}{n^2} \\ & - \frac{1}{n}\langle \hat{f}_H, f_{H_{min}}\rangle - \frac{1}{n}\langle f_H, \hat{f}_{H_{min}}\rangle + \frac{1}{n}\langle f_H, f_{H_{min}}\rangle \\ & + V(H, H_{min}) + V(H_{min}, H) + \langle f_H - f, f_{H_{min}} - f\rangle.\end{aligned}$$

We first control the last term of the first line and we obtain the following lemma.

Lemma A.3. *With probability larger than $1 - 5.54|\mathcal{H}|\exp(-x)$, for any $H \in \mathcal{H}$,*

$$\frac{|U(H, H_{min})|}{n^2} \leq \theta' \frac{\|K\|^2}{n \det(H)} + \frac{\square \|K\|_1^2 \|f\|_\infty x^2}{\theta' n} + \frac{\square \|K\|_\infty \|K\|_1 x^3}{\theta' n^2 \det(H_{min})}$$

Proof. The proof uses a concentration inequality for U-statistics. It is similar to the proof of Lemma 10 in [19], using that

$$\|K_H\|_\infty \leq \frac{\|K\|_\infty}{\det(H)} \quad \text{and} \quad \|K_H\|^2 = \frac{\|K\|^2}{\det(H)}.$$

□

We control (A.3) and (A.4) similarly to [19]. Then, from Lemma A.3, we obtain the following result. With probability larger than $1 - 9.54|\mathcal{H}|\exp(-x)$, for any $H \in \mathcal{H}$,

$$\begin{aligned}|\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle - \frac{\langle K_H, K_{H_{min}}\rangle}{n}| \leq {} & \theta' \|f_H - f\|^2 + \theta' \frac{\|K\|^2}{n \det(H)} + \left(\frac{\theta'}{2} + \frac{1}{2\theta'}\right) \|f_{H_{min}} - f\|^2 \\ & + \frac{C_1(K)}{\theta'} \left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2 \det(H_{min})}\right),\end{aligned}$$

where $C_1(K)$ is a constant only depending on K. Now, Proposition A.1 gives, with probability larger than $1 - \square|\mathcal{H}|\exp(-x)$, for any $H \in \mathcal{H}$,

$$\|f_H - f\|^2 + \frac{\|K\|^2}{n\det(H)} \leq 2\|\hat{f}_H - f\|^2 + C_2(K)\|f\|_\infty \frac{x^2}{n},$$

where $C_2(K)$ only depends on K. Hence, by applying (A.5), with probability larger than $1 - \square|\mathcal{H}|\exp(-x)$, for any $H \in \mathcal{H}$,

$$\left|\langle \hat{f}_H - f, \hat{f}_{H_{min}} - f\rangle - \frac{\langle K_H, K_{H_{min}}\rangle}{n} - \langle \hat{f}_{\hat{H}} - f, \hat{f}_{H_{min}} - f\rangle + \frac{\langle K_{\hat{H}}, K_{H_{min}}\rangle}{n}\right| \leq 2\theta'\|\hat{f}_H - f\|^2$$
$$+2\theta'\|\hat{f}_{\hat{H}} - f\|^2 + \left(\theta' + \frac{1}{\theta'}\right)\|f_{H_{min}} - f\|^2 + \frac{\tilde{C}(K)}{\theta'}\left(\frac{\|f\|_\infty x^2}{n} + \frac{x^3}{n^2\det(H_{min})}\right),$$

where $\tilde{C}(K)$ is a constant only depending on K. It remains to use (A.1) and to choose $\theta' = \frac{\theta}{4}$ to conclude.

A.2 Proof of Corollary 2.4

We shall use the following Lemma to control the bias terms.

Lemma A.4. *Let H be a symmetric positive matrix with diagonalization $H = P^{-1}DP$ with P orthogonal and D diagonal. Then*

$$\|f_H - f\| = \|\tilde{f}_D - \tilde{f}\|$$

with $\tilde{f} = f \circ P^{-1}$, $\tilde{f}_D = \tilde{K}_D \star \tilde{f}$ and $\tilde{K} = K \circ P^{-1}$. Moreover, if $f \circ P^{-1}$ belongs to the anisotropic Nikol'skii class $\mathcal{N}_{\mathbf{2},d}(\boldsymbol{\beta}, \mathbf{L})$ and K is order $\ell > \max_{j=1,\dots,d}\beta_j$ then there exists $C > 0$ such that

$$\|f_H - f\| \leq C\sum_{j=1}^{d} L_j h_j^{\beta_j}$$

where $(h_j)_{j=1}^d$ are the eigenvalues of H.

Proof of Lemma A.4

Compute for any $\mathbf{t} \in \mathbb{R}^d$,

$$f_H(\mathbf{t}) = \frac{1}{\det(H)}\int K(P^{-1}D^{-1}P(\mathbf{t} - \mathbf{u}))f(\mathbf{u})d\mathbf{u} = \frac{1}{\det(D)}\int K(P^{-1}D^{-1}(P\mathbf{t} - \mathbf{v}))f(P^{-1}\mathbf{v})d\mathbf{v}$$

$$= \frac{1}{\det(D)}\int \tilde{K}(D^{-1}(P\mathbf{t} - \mathbf{v}))\tilde{f}(\mathbf{v})d\mathbf{v} = \tilde{K}_D \star \tilde{f}(P\mathbf{t}) = \tilde{f}_D(P\mathbf{t}).$$

Thus

$$\|f_H - f\|^2 = \int |f_H(\mathbf{t}) - f(\mathbf{t})|^2 d\mathbf{t} = \int |\tilde{f}_D(P\mathbf{t}) - f(\mathbf{t})|^2 d\mathbf{t} = \int |\tilde{f}_D(\mathbf{y}) - f(P^{-1}\mathbf{y})|^2 d\mathbf{y} = \|\tilde{f}_D - \tilde{f}\|^2.$$

Note that if K is order ℓ, then $\tilde{K}$ is order ℓ. Then we apply Lemma 3 of [14] to $\tilde{f}$. □

Now, let $\mathcal{E}$ be the event corresponding to the intersection of events considered in Theorem 2.1 and Proposition A.1. For any $A > 0$, by taking x proportional to $\log n$, $\mathbb{P}(\mathcal{E}) \geq 1 - n^A$. On $\mathcal{E}$

$$\|\hat{f}_{\hat{H}} - f\|^2 \leq C_0(\varepsilon, \lambda)(1+\eta) \min_{H \in \mathcal{H}} \left(C \sum_{j=1}^{d} L_j^2 h_j^{2\beta_j} + \frac{\|K\|^2}{n \prod_{j=1}^{d} h_j} \right)$$

$$+ C_2(\varepsilon, \lambda) C \sum_{j=1}^{d} L_j^2 \bar{h}^{2\beta_j} + C' \frac{(\log n)^3}{n}.$$

But, on $\mathcal{E}^c$, for any $H \in \mathcal{H}$, $\|\hat{f}_H - f\|^2 \leq 2\|f\|^2 + 2\|K\|^2 (\|K\|_\infty \|K\|_1)^{-1} n$. Thus

$$\mathbb{E}\left[\|\hat{f}_{\hat{H}} - f\|^2\right] \leq \mathbb{E}\left[\|\hat{f}_{\hat{H}} - f\|^2 1_{\mathcal{E}}\right] + \mathbb{E}\left[\|\hat{f}_{\hat{H}} - f\|^2 1_{\mathcal{E}^c}\right]$$

$$\leq M \left(\prod_{j=1}^{d} L_j^{\frac{1}{\beta_j}} \right)^{\frac{2\bar{\beta}}{2\bar{\beta}+1}} n^{-\frac{2\bar{\beta}}{2\bar{\beta}+1}},$$

where M is a constant depending on an upper bound of f, $\boldsymbol{\beta}$, K, d and λ.

Appendix B. Testing densities

In this section, we present the testing distributions. We respectively denote $\mathcal{N}$, $\mathcal{E}$ and $\mathcal{U}$ the Gaussian, exponential and uniform distributions.

Table B.1. Definition of one-dimensional testing densities.

Dist. name	Abb.	Distribution
Gauss	G	$\mathcal{N}(0,1)$
Uniform	U	$\mathcal{U}([0,1])$
Exponential	E	$\mathcal{E}(1)$
Mix Gauss	MG	$\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{2}\mathcal{N}(3,(\frac{1}{3})^2)$
Skewed	Sk	$\frac{1}{5}\mathcal{N}(0,1)+\frac{1}{5}\mathcal{N}(\frac{1}{2},(\frac{2}{3})^2)+\frac{3}{5}\mathcal{N}(\frac{13}{12},(\frac{5}{9})^2)$
Strong skewed	Sk+	$\sum_{l=0}^{7}\frac{1}{8}\mathcal{N}(3((\frac{2}{3})^l-1),(\frac{2}{3})^{2l})$
Kurtotic	K	$\frac{2}{3}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}(0,(\frac{1}{10})^2)$
Outlier	O	$\frac{1}{10}\mathcal{N}(0,1)+\frac{9}{10}\mathcal{N}(0,(\frac{1}{10})^2)$
Bimodal	Bi	$\frac{1}{2}\mathcal{N}(-1,(\frac{2}{3})^2)+\frac{1}{2}\mathcal{N}(1,(\frac{2}{3})^2)$
Separated bimodal	SB	$\frac{1}{2}\mathcal{N}(-\frac{3}{2},(\frac{1}{2})^2)+\frac{1}{2}\mathcal{N}(\frac{3}{2},(\frac{1}{2})^2)$
Skewed bimodal	SkB	$\frac{3}{4}\mathcal{N}(0,1)+\frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{3})^2)$
Trimodal	T	$\frac{9}{20}\mathcal{N}(-\frac{6}{5},(\frac{3}{5})^2)+\frac{9}{20}\mathcal{N}(\frac{6}{5},(\frac{3}{5})^2)+\frac{1}{10}\mathcal{N}(0,(\frac{1}{4})^2)$
Bart	B	$\frac{1}{2}\mathcal{N}(0,1)+\sum_{l=0}^{4}\frac{1}{10}\mathcal{N}(\frac{l}{2}-1,(\frac{1}{10})^2)$
Double bart	DB	$\frac{49}{100}\mathcal{N}(-1,(\frac{2}{3})^2)+\frac{49}{100}\mathcal{N}(1,(\frac{2}{3})^2)+\sum_{l=0}^{6}\frac{1}{350}\mathcal{N}(\frac{l-3}{2},(\frac{1}{100})^2)$
Asymetric bart	AB	$\frac{1}{2}\mathcal{N}(0,1)+\sum_{l=-2}^{2}\frac{2^{1-l}}{31}\mathcal{N}(l+\frac{1}{2},(\frac{2^{-l}}{10})^2)$
Asymetric double bart	ADB	$\sum_{l=0}^{1}\frac{46}{100}\mathcal{N}(2l-1,(\frac{2}{3})^2)+\sum_{l=1}^{3}\frac{1}{300}\mathcal{N}(-\frac{l}{2},(\frac{1}{100})^2)+\sum_{l=1}^{3}\frac{7}{300}\mathcal{N}(\frac{l}{2},(\frac{7}{100})^2)$
Smooth comb	SC	$\sum_{l=0}^{5}\frac{2^{5-l}}{63}\mathcal{N}(\frac{65-96(\frac{1}{2})^l}{21},(\frac{32}{63}(\frac{1}{2})^l)^2)$
Discrete comb	DC	$\sum_{l=0}^{2}\frac{2}{7}\mathcal{N}(\frac{12l-15}{7},(\frac{2}{7})^2)+\sum_{l=8}^{10}\frac{1}{21}\mathcal{N}(\frac{2l}{7},(\frac{1}{21})^2)$
Mix Uniform	MU	$\frac{1}{25}\mathcal{U}([0,\frac{3}{20}])+\frac{29}{200}\mathcal{U}([\frac{3}{20},\frac{1}{5}])+\frac{17}{200}\mathcal{U}([\frac{1}{5},\frac{3}{8}])+\frac{1}{20}\mathcal{U}([\frac{3}{8},\frac{4}{8}])+\frac{7}{50}\mathcal{U}([\frac{4}{8},\frac{3}{5}])+$ $\frac{1}{5}\mathcal{U}([\frac{3}{5},\frac{4}{5}])+\frac{7}{50}\mathcal{U}([\frac{4}{5},\frac{7}{8}])+\frac{1}{5}\mathcal{U}([\frac{7}{8},1])$

TABLE B.2. Definition of bi-dimensional testing densities.

Dist. name	Abb.	Distribution
Uncorrelated Gauss	UG	$\mathcal{N}\left(\mathbf{0};(0.25,0,1)\right)$
Correlated Gauss	CG	$\mathcal{N}\left(\mathbf{0};(1,0.9,1)\right)$
Uniform	U	$\mathcal{U}(\{\boldsymbol{x} \quad \mid \quad \lVert\boldsymbol{x}-\boldsymbol{a}\rVert^2 \leq r^2, \boldsymbol{a}=(2,2), r=1\})$
Strong Skewed	Sk+	$\sum_{l=0}^{7} \frac{1}{8}\mathcal{N}\left((3\left(1-(\frac{4}{5})^l\right), -3\left(1-(\frac{4}{5})^l\right); (\frac{4}{5})^{2l}(1,-\frac{9}{10},1)\right)$
Skewed	Sk	$\frac{1}{5}\mathcal{N}\left((0,0);(1,0,1)\right)+\frac{1}{5}\mathcal{N}\left((5,5);(\frac{4}{9},0,\frac{4}{9})\right)+\frac{3}{5}\mathcal{N}\left((10,10);(\frac{25}{81},0,\frac{25}{81})\right)$
Dumbbell	D	$\frac{4}{11}\mathcal{N}\left((-\frac{3}{2},\frac{3}{2});\frac{9}{16}\boldsymbol{I}\right)+\frac{4}{11}\mathcal{N}\left((\frac{3}{2},-\frac{3}{2});\frac{9}{16}\boldsymbol{I}\right)+\frac{3}{11}\mathcal{N}\left(\mathbf{0};\frac{9}{16}(\frac{4}{5},-\frac{18}{25},\frac{4}{5})\right)$
Kurtotic	K	$\frac{2}{3}\mathcal{N}\left(\mathbf{0};\frac{9}{16}(1,1,4)\right)+\frac{1}{3}\mathcal{N}\left(\mathbf{0};\frac{9}{16}(\frac{4}{9},-\frac{1}{3},\frac{4}{9})\right)$
Bimodal	Bi	$\frac{1}{2}\mathcal{N}\left((-1,0);(\frac{4}{9},\frac{2}{9},\frac{4}{9})\right)+\frac{1}{2}\mathcal{N}\left((1,0);(\frac{4}{9},\frac{2}{9},\frac{4}{9})\right)$
Bimodal 2	Bi2	$\frac{1}{2}\mathcal{N}\left((-1,1);(\frac{4}{9},\frac{1}{3},\frac{4}{9})\right)+\frac{1}{2}\mathcal{N}\left(\mathbf{0};\frac{4}{9}\boldsymbol{I}\right)$
Asymmetric Bimodal	ABi	$\frac{1}{2}\mathcal{N}\left((1,-1);(\frac{4}{9},\frac{14}{45},\frac{4}{9})\right)+\frac{1}{2}\mathcal{N}\left((-1,1);\frac{4}{9}\boldsymbol{I}\right)$
Trimodal	T	$\frac{3}{7}\mathcal{N}\left((-1,0);\frac{1}{25}(9,\frac{63}{10},\frac{49}{4})\right) + \frac{3}{7}\mathcal{N}\left((1,\frac{2}{\sqrt{3}});\frac{1}{25}(9,0,\frac{49}{4})\right) + \frac{1}{7}\mathcal{N}\left((1,-\frac{2}{\sqrt{3}});\frac{1}{25}(9,0,\frac{49}{4})\right)$
Fountain	F	$\frac{1}{2}\mathcal{N}\left(\mathbf{0};\boldsymbol{I}\right)+\frac{1}{10}\mathcal{N}\left(\mathbf{0};\frac{1}{16}\boldsymbol{I})\right)+\sum_{i,j=1}^{2}\frac{1}{10}\mathcal{N}\left(((-1)^i,(-1)^j);\frac{1}{16}\boldsymbol{I}\right)$
Double Fountain	DF	$\frac{12}{25}\mathcal{N}\left((-\frac{3}{2},0);(\frac{4}{9},\frac{4}{15},\frac{4}{9})\right) + \frac{12}{25}\mathcal{N}\left((\frac{3}{2},0);(\frac{4}{9},\frac{4}{15},\frac{4}{9})\right) + \frac{8}{350}\mathcal{N}\left(\mathbf{0};\frac{1}{9}(1,\frac{3}{5},1)\right) + \sum_{i=-1}^{1}\frac{1}{350}\mathcal{N}\left((i-\frac{3}{2},i);\frac{1}{15}(\frac{1}{15},\frac{1}{25},\frac{1}{15})\right) + \sum_{j=-1}^{1}\frac{1}{350}\mathcal{N}\left(j+\frac{3}{2},j);\frac{1}{15}(\frac{1}{15},\frac{1}{25},\frac{1}{15})\right)$
Asymmetric Fountain	AF	$\frac{1}{2}\mathcal{N}\left(\mathbf{0};\boldsymbol{I}\right) + \frac{3}{40}\mathcal{N}\left(\mathbf{0};\frac{1}{16}(1,-\frac{9}{10},1)\right) + \frac{1}{5}\mathcal{N}\left((1,1);\frac{1}{4}(1,-\frac{9}{10},1)\right) + \frac{3}{40}\mathcal{N}\left((-1,1);\frac{1}{8}\boldsymbol{I}\right)+\frac{3}{40}\mathcal{N}\left((-1,-1);\frac{1}{8}(1,-\frac{9}{10},1)\right)+\frac{3}{40}\mathcal{N}\left((1,-1);\frac{1}{16}\boldsymbol{I}\right)$

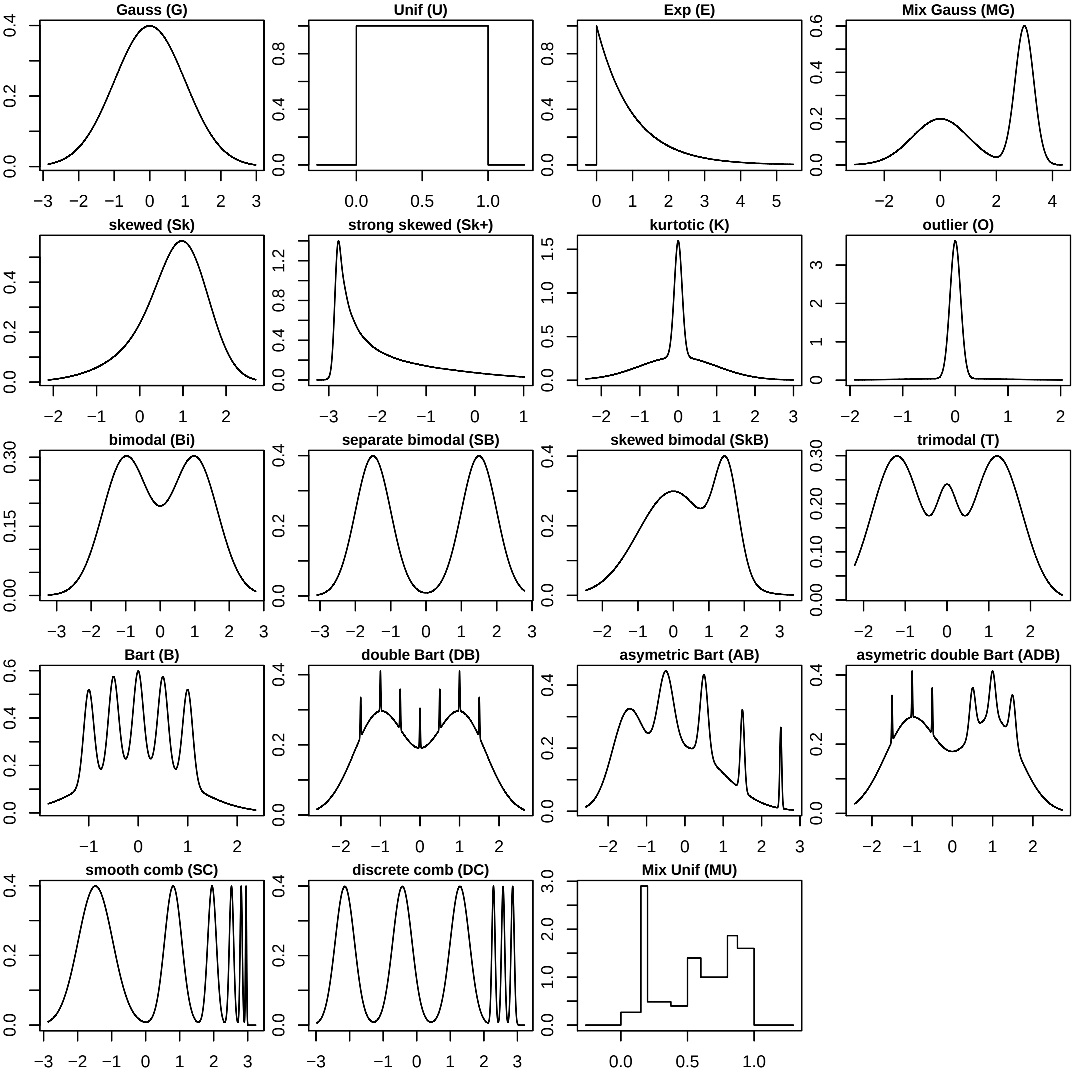


FIGURE B.1. Representation of one-dimensional testing densities.

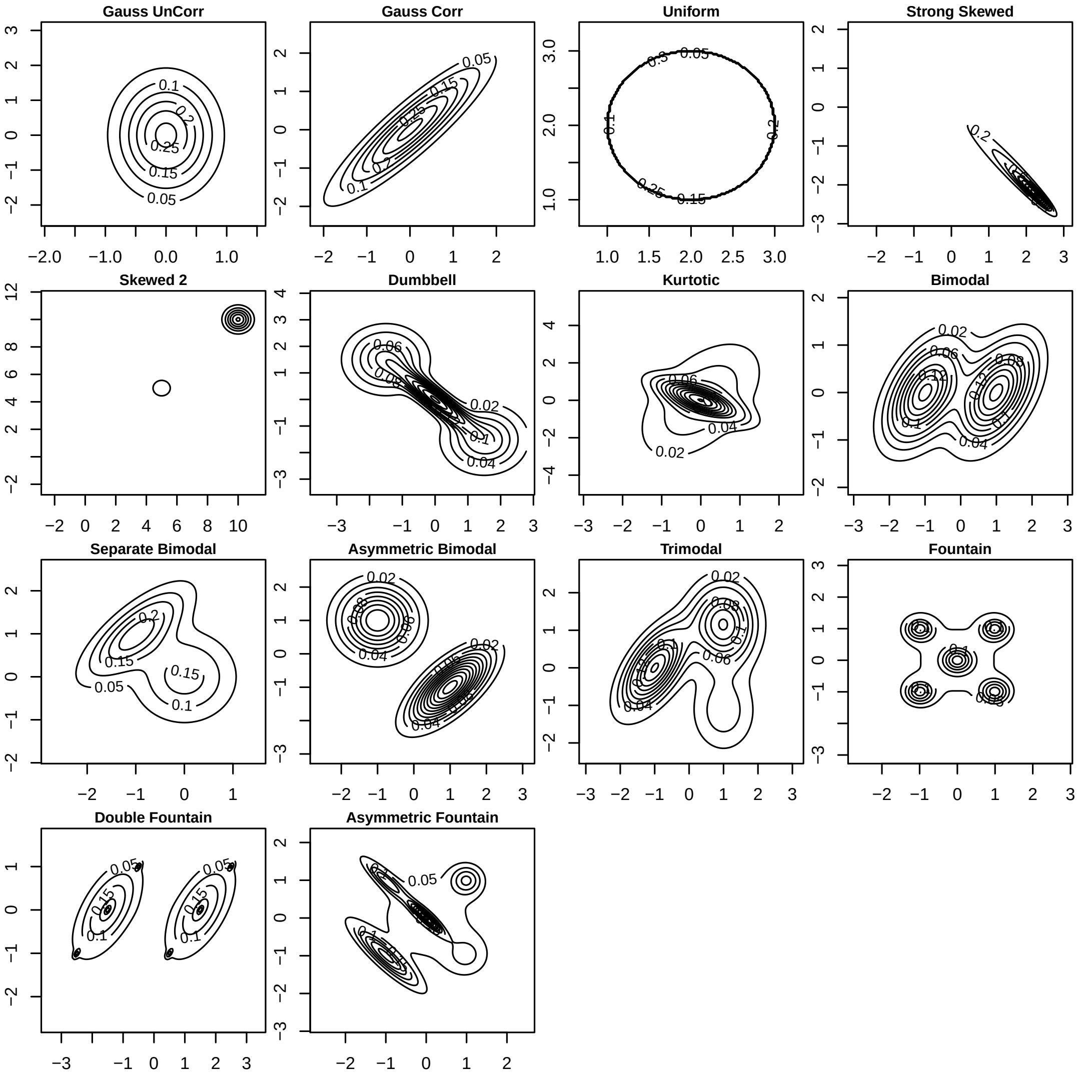
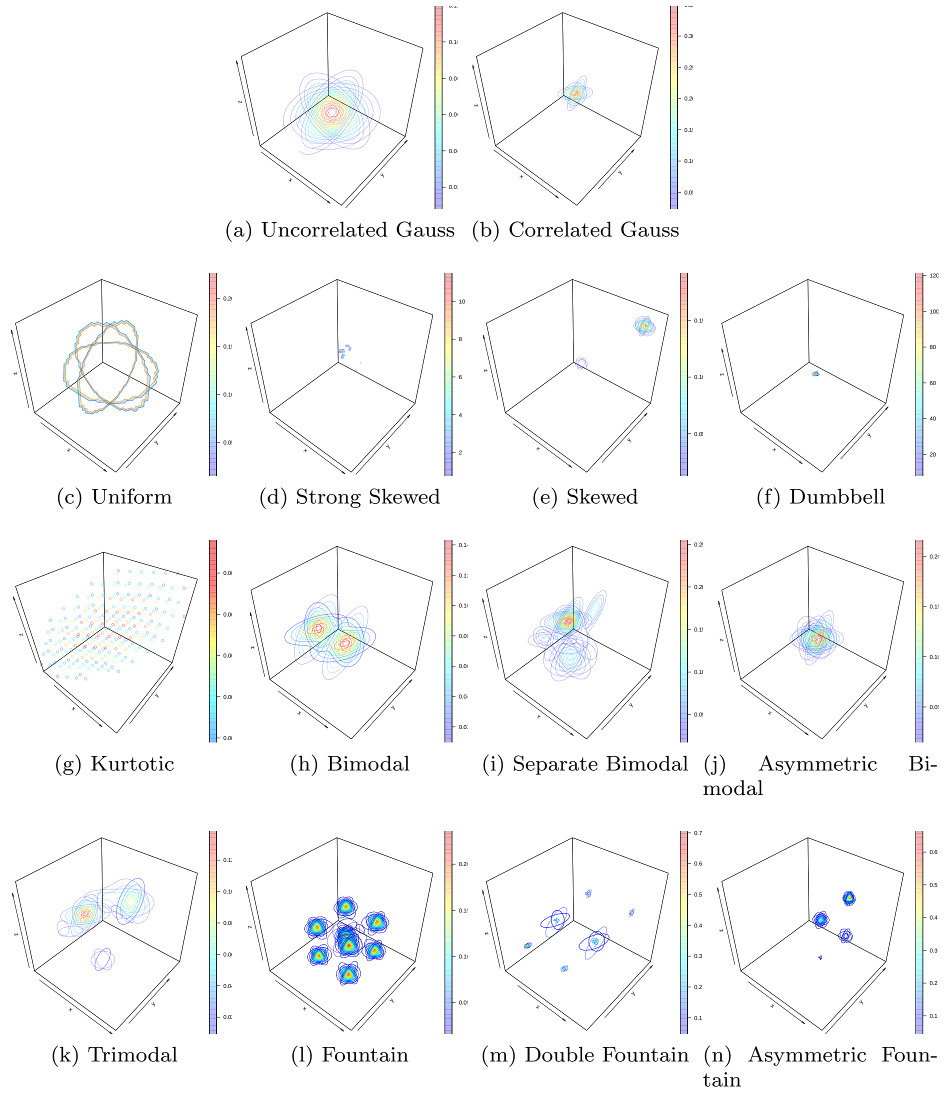


Figure B.2. Representation of bi-dimensional testing densities.

Table B.3. Definition of tri-dimensional testing densities.

Dist. name	Abb.	Distribution
Uncorrelated Gauss	UG	$\mathcal{N}\left(\mathbf{0}; (0.25, 0, 0, 1, 0, 1)\right)$
Correlated Gauss	CG	$\mathcal{N}\left(\mathbf{0}; (1, 0.9, 0.9, 1, 0.9, 1)\right)$
Uniform	U	$\mathcal{U}(\{\boldsymbol{x} \quad \mid \quad \lVert\boldsymbol{x}-\boldsymbol{a}\rVert^2 \leq r^2, \boldsymbol{a}=(2,2,2), r=1\})$
Strong Skewed	Sk+	$\sum_{l=0}^{7} \frac{1}{8}\mathcal{N}\left((m_1, m_2, m_3); (\sigma_{11}, \sigma_{21}, \sigma_{31}, \sigma_{22}, \sigma_{32}, \sigma_{33})\right)$ with $m_j = 3(-1)^{j+1}\left(1-(\frac{4}{5})^l\right)$, $\sigma_{jj} = (\frac{4}{5})^{2l}$ and $\sigma_{jk} = -\frac{9}{10}(\frac{4}{5})^{2(l-1)}$ for $j \neq k$
Skewed	Sk	$\frac{1}{5}\mathcal{N}\left(\mathbf{0}; \boldsymbol{I}\right) + \frac{1}{5}\mathcal{N}\left(\mathbf{5}; \frac{4}{9}\boldsymbol{I}\right) + \frac{3}{5}\mathcal{N}\left(\mathbf{10}; \frac{25}{81}\boldsymbol{I}\right)$
Dumbbell	D	$\frac{4}{11}\mathcal{N}\left((-\frac{3}{2}, \frac{3}{2}, -\frac{3}{2}); \frac{9}{16}\boldsymbol{I}\right) + \frac{4}{11}\mathcal{N}\left((\frac{3}{2}, -\frac{3}{2}, \frac{3}{2}); \frac{9}{16}\boldsymbol{I}\right) + \frac{3}{11}\mathcal{N}\left(\mathbf{0}; \frac{9}{16}(\frac{4}{5}, -\frac{18}{25}, -\frac{18}{25}, \frac{4}{5}, -\frac{18}{25}, \frac{4}{5})\right)$
Kurtotic	K	$\frac{2}{3}\mathcal{N}\left(\mathbf{0}; (1, 1, 1, 4, 1, 4)\right) + \frac{1}{3}\mathcal{N}\left(\mathbf{0}; (\frac{4}{9}, -\frac{1}{3}, -\frac{1}{3}, \frac{4}{9}, -\frac{1}{3}, \frac{4}{9})\right)$
Bimodal	Bi	$\frac{1}{2}\mathcal{N}\left((-1, 0, 0); (\frac{4}{9}, \frac{2}{9}, \frac{2}{9}, \frac{4}{9}, \frac{2}{9}, \frac{4}{9})\right) + \frac{1}{2}\mathcal{N}\left((1, 0, 0); (\frac{4}{9}, \frac{2}{9}, \frac{2}{9}, \frac{4}{9}, \frac{2}{9}, \frac{4}{9})\right)$
Bimodal 2	Bi2	$\frac{1}{2}\mathcal{N}\left((-1, 1, 1); (\frac{4}{9}, \frac{1}{3}, \frac{1}{3}, \frac{4}{9}, \frac{1}{3}, \frac{4}{9})\right) + \frac{1}{2}\mathcal{N}\left(\mathbf{0}; \frac{4}{9}\boldsymbol{I})\right)$
Asymmetric Bimodal	ABi	$\frac{1}{2}\mathcal{N}\left((1, -1, 1); (\frac{4}{9}, \frac{14}{45}, \frac{14}{45}, \frac{4}{9}, \frac{14}{45}, \frac{4}{9})\right) + \frac{1}{2}\mathcal{N}\left((-1, 1, -1); \frac{4}{9}\boldsymbol{I}\right)$
Trimodal	T	$\frac{3}{7}\mathcal{N}\left((-1, 0, 0); \frac{1}{25}(9, \frac{63}{10}, \frac{63}{10}, \frac{49}{4}, \frac{63}{10}, \frac{49}{4})\right) + \frac{3}{7}\mathcal{N}\left((1, \frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}); \frac{1}{25}(9, 0, 0, \frac{49}{4}, 0, \frac{49}{4})\right) + \frac{1}{7}\mathcal{N}\left((1, -\frac{2}{\sqrt{3}}, -\frac{2}{\sqrt{3}}); \frac{1}{25}(9, 0, 0, \frac{49}{4}, 0, \frac{49}{4})\right)$
Fountain	F	$\frac{1}{2}\mathcal{N}\left(\mathbf{0}; \boldsymbol{I}\right) + \frac{1}{18}\mathcal{N}\left(\mathbf{0}; \frac{1}{16}\boldsymbol{I})\right) + \sum_{i,j,k=1}^{2} \frac{1}{18}\mathcal{N}\left(((-1)^i, (-1)^j, (-1)^k); \frac{1}{16}\boldsymbol{I}\right)$
Double Fountain	DF	$\frac{12}{25}\mathcal{N}\left((-\frac{3}{2}, 0, 0); (\frac{4}{9}, \frac{4}{15}, \frac{4}{15}, \frac{4}{9}, \frac{4}{15}, \frac{4}{9})\right) + \frac{12}{25}\mathcal{N}\left((\frac{3}{2}, 0, 0); (\frac{4}{9}, \frac{4}{15}, \frac{4}{15}, \frac{4}{9}, \frac{4}{15}, \frac{4}{9})\right) + \frac{8}{350}\mathcal{N}\left(\mathbf{0}; \frac{1}{9}(1, \frac{3}{5}, \frac{3}{5}, 1, \frac{3}{5}, 1)\right) + \sum_{i=-1}^{1} \frac{1}{350}\mathcal{N}\left((i-\frac{3}{2}, i, i); \frac{1}{15}(\frac{1}{15}, \frac{1}{25}, \frac{1}{25}, \frac{1}{15}, \frac{1}{25}, \frac{1}{15})\right) + \sum_{j=-1}^{1} \frac{1}{350}\mathcal{N}\left(j+\frac{3}{2}, j, j); \frac{1}{15}(\frac{1}{15}, \frac{1}{25}, \frac{1}{25}, \frac{1}{15}, \frac{1}{25}, \frac{1}{15})\right)$
Asymmetric Fountain	AF	$\frac{1}{2}\mathcal{N}\left(\mathbf{0}; \boldsymbol{I}\right) + \frac{3}{40}\mathcal{N}\left(\mathbf{0}; \frac{1}{16}(1, -\frac{9}{10}, -\frac{9}{10}, 1, -\frac{9}{10}, 1)\right) + \frac{1}{5}\mathcal{N}\left((-1, -1, -1); \frac{1}{4}(1, -\frac{9}{10}, -\frac{9}{10}, 1, -\frac{9}{10}, 1)\right) + \sum_{k=1}^{4} \frac{9}{280}\mathcal{N}\left(((-1)^{2k}, (-1)^{(2k+1)\mathrm{div}2}, (-1)^{(2k+3)\mathrm{div}4}); \frac{1}{2^{k+2}}(1, -\frac{9}{10}, -\frac{9}{10}, 1, -\frac{9}{10}, 1)\right) + \sum_{k=1}^{3} \frac{9}{280}\mathcal{N}\left(((-1)^{2k+1}, (-1)^{(2k+2)\mathrm{div}2}, (-1)^{(2k+4)\mathrm{div}4}); \frac{1}{2^{k+2}}\boldsymbol{I}\right)$ with div the integer division



(a) Uncorrelated Gauss (b) Correlated Gauss

(c) Uniform (d) Strong Skewed (e) Skewed (f) Dumbbell

(g) Kurtotic (h) Bimodal (i) Separate Bimodal (j) Asymmetric Bimodal

(k) Trimodal (l) Fountain (m) Double Fountain (n) Asymmetric Fountain

FIGURE B.3. Representation of tri-dimensional testing densities.

TABLE B.4. Definition of quadri-dimensional testing densities.

Dist. name	Abb.	Distribution
Uncorrelated Gauss	UG	$\mathcal{N}\left(\mathbf{0};(0.25,0,0,0,0.25,0,0,1,0,1)\right)$
Correlated Gauss	CG	$\mathcal{N}\left(\mathbf{0};(1,0.9,0.9,0.9,1,0.9,0.9,1,0.9,1)\right)$
Uniform	U	$\mathcal{U}(\{\boldsymbol{x} \mid \|\boldsymbol{x}-\boldsymbol{a}\|^2 \le r^2, \boldsymbol{a}=(2,2,2,2), r=1\})$
Strong Skewed	Sk+	$\sum_{l=0}^{7}\frac{1}{8}\mathcal{N}\left((m_1,m_2,m_3,m_4);(\sigma_{11},\sigma_{21},\sigma_{31},\sigma_{41},\sigma_{22},\sigma_{32},\sigma_{42},\sigma_{33},\sigma_{43},\sigma_{44})\right)$ with $m_j = 3(-1)^{j+1}\bigl(1-(\frac{4}{5})^l\bigr)$, $\sigma_{jj} = (\frac{4}{5})^{2l}$ and $\sigma_{jk} =$ $-\frac{9}{10}(\frac{4}{5})^{2(l-1)}$ for $j \neq k$
Skewed	Sk	$\frac{1}{5}\mathcal{N}\left(\mathbf{0};\boldsymbol{I}\right)+\frac{1}{5}\mathcal{N}\bigl(\mathbf{5};\frac{4}{9}\boldsymbol{I}\bigr)+\frac{3}{5}\mathcal{N}\bigl(\mathbf{10};\frac{25}{81}\boldsymbol{I}\bigr)$
Dumbbell	D	$\frac{4}{11}\mathcal{N}\bigl((-\frac{3}{2},\frac{3}{2},-\frac{3}{2},\frac{3}{2});\frac{9}{16}\boldsymbol{I}\bigr) + \frac{4}{11}\mathcal{N}\bigl((\frac{3}{2},-\frac{3}{2},\frac{3}{2},-\frac{3}{2});\frac{9}{16}\boldsymbol{I}\bigr) +$ $\frac{3}{11}\mathcal{N}\bigl(\mathbf{0};\frac{9}{16}(\frac{4}{5},-\frac{18}{25},-\frac{18}{25},-\frac{18}{25},\frac{4}{5},-\frac{18}{25},-\frac{18}{25},\frac{4}{5},-\frac{18}{25},\frac{4}{5})\bigr)$
Kurtotic	K	$\frac{2}{3}\mathcal{N}\left(\mathbf{0};(1,1,1,1,4,1,1,4,1,4)\right)+\frac{1}{3}\mathcal{N}\bigl(\mathbf{0};(\frac{4}{9},-\frac{1}{3},-\frac{1}{3},-\frac{1}{3},\frac{4}{9},-\frac{1}{3},-\frac{1}{3},\frac{4}{9},-\frac{1}{3},\frac{4}{9})\bigr)$
Bimodal	Bi	$\frac{1}{2}\mathcal{N}\bigl((-1,0,0,0);(\frac{4}{9},\frac{2}{9},\frac{2}{9},\frac{2}{9},\frac{4}{9},\frac{2}{9},\frac{2}{9},\frac{4}{9},\frac{2}{9},\frac{4}{9})\bigr) +$ $\frac{1}{2}\mathcal{N}\bigl((1,0,0,0);(\frac{4}{9},\frac{2}{9},\frac{2}{9},\frac{2}{9},\frac{4}{9},\frac{2}{9},\frac{2}{9},\frac{4}{9},\frac{2}{9},\frac{4}{9})\bigr)$
Bimodal 2	Bi2	$\frac{1}{2}\mathcal{N}\bigl((-1,1,1,1);(\frac{4}{9},\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{4}{9},\frac{1}{3},\frac{1}{3},\frac{4}{9},\frac{1}{3},\frac{4}{9})\bigr)+\frac{1}{2}\mathcal{N}\bigl(\mathbf{0};\frac{4}{9}\boldsymbol{I})\bigr)$
Asymmetric Bimodal	ABi	$\frac{1}{2}\mathcal{N}\bigl((1,-1,1,-1);(\frac{4}{9},\frac{14}{45},\frac{14}{45},\frac{14}{45},\frac{4}{9},\frac{14}{45},\frac{14}{45},\frac{4}{9},\frac{14}{45},\frac{4}{9})\bigr) +$ $\frac{1}{2}\mathcal{N}\bigl((-1,1,-1,1);\frac{4}{9}\boldsymbol{I}\bigr)$
Trimodal	T	$\frac{3}{7}\mathcal{N}\bigl((-1,0,0,0);\frac{1}{25}(9,\frac{63}{10},\frac{63}{10},\frac{63}{10},\frac{49}{4},\frac{63}{10},\frac{63}{10},\frac{49}{4},\frac{63}{10},\frac{49}{4})\bigr) +$ $\frac{3}{7}\mathcal{N}\Bigl((1,\frac{2}{\sqrt{3}},\frac{2}{\sqrt{3}},\frac{2}{\sqrt{3}});\frac{1}{25}(9,0,0,0,\frac{49}{4},0,0,\frac{49}{4},0,\frac{49}{4})\Bigr) +$ $\frac{1}{7}\mathcal{N}\Bigl((1,-\frac{2}{\sqrt{3}},-\frac{2}{\sqrt{3}},-\frac{2}{\sqrt{3}});\frac{1}{25}(9,0,0,0,\frac{49}{4},0,0,\frac{49}{4},0,\frac{49}{4})\Bigr)$
Fountain	F	$\frac{1}{2}\mathcal{N}\left(\mathbf{0};\boldsymbol{I}\right)+\frac{1}{34}\mathcal{N}\bigl(\mathbf{0};\frac{1}{16}\boldsymbol{I})\bigr)+\sum_{i,j,k,l=1}^{2}\frac{1}{34}\mathcal{N}\bigl(((-1)^i,(-1)^j,(-1)^k,(-1)^l);\frac{1}{16}\boldsymbol{I}\bigr)$
Double Fountain	DF	$\frac{12}{25}\mathcal{N}\bigl((-\frac{3}{2},0,0,0);(\frac{4}{9},\frac{4}{15},\frac{4}{15},\frac{4}{15},\frac{4}{9},\frac{4}{15},\frac{4}{15},\frac{4}{9},\frac{4}{15},\frac{4}{9})\bigr) +$ $\frac{12}{25}\mathcal{N}\bigl((\frac{3}{2},0,0,0);(\frac{4}{9},\frac{4}{15},\frac{4}{15},\frac{4}{15},\frac{4}{9},\frac{4}{15},\frac{4}{15},\frac{4}{9},\frac{4}{15},\frac{4}{9})\bigr) +$ $\frac{8}{350}\mathcal{N}\bigl(\mathbf{0};\frac{1}{9}(1,\frac{3}{5},\frac{3}{5},\frac{3}{5},1,\frac{3}{5},\frac{3}{5},1,\frac{3}{5},1)\bigr) +$ $\sum_{i=-1}^{1}\frac{1}{350}\mathcal{N}\bigl((i-\frac{3}{2},i,i,i);\frac{1}{75}(\frac{1}{3},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{3},\frac{1}{5},\frac{1}{5},\frac{1}{3},\frac{1}{5},\frac{1}{3})\bigr) +$ $\sum_{j=-1}^{1}\frac{1}{350}\mathcal{N}\bigl(j+\frac{3}{2},j,j,j);\frac{1}{75}(\frac{1}{3},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{3},\frac{1}{5},\frac{1}{5},\frac{1}{3},\frac{1}{5},\frac{1}{3})\bigr)$
Asymmetric Fountain	AF	$\frac{1}{2}\mathcal{N}\left(\mathbf{0};\boldsymbol{I}\right)+\frac{3}{40}\mathcal{N}\bigl(\mathbf{0};\frac{1}{16}(1,-\frac{9}{10},-\frac{9}{10},-\frac{9}{10},1,-\frac{9}{10},-\frac{9}{10},1,-\frac{9}{10},1)\bigr)+$ $\frac{1}{5}\mathcal{N}\bigl((-1,-1,-1,-1);\frac{1}{4}(1,-\frac{9}{10},-\frac{9}{10},-\frac{9}{10},1,-\frac{9}{10},-\frac{9}{10},1,-\frac{9}{10},1)\bigr)+$ $\sum_{k=1}^{8}\frac{9}{600}\mathcal{N}\bigl(((-1)^{2k},(-1)^{i_k},(-1)^{j_k},(-1)^{l_k});\frac{1}{2^{k+2}}(1,-\frac{9}{10},-\frac{9}{10},-\frac{9}{10},1,-\frac{9}{10},-\frac{9}{10},$ $1,-\frac{9}{10},1)\bigr)+\sum_{k=1}^{7}\frac{9}{600}\mathcal{N}\bigl(((-1)^{2k+1},(-1)^{(2k+2)\mathrm{div}2},(-1)^{(2k+4)\mathrm{div}4},(-1)^{(2k+8)\mathrm{div}8});$ $\frac{1}{2^{k+2}}\boldsymbol{I}\bigr)$ with div the integer division, $i_k = (2k+1)\mathrm{div}2$, $j_k = (2k+3)\mathrm{div}4$ and $l_k = (2k+7)\mathrm{div}8$

Appendix C. Boxplots of Monte Carlo mean of ISE values for different h_{min} and H_{min}

The following four figures are the boxplots from which the averaged risk of Figure 2 has been obtained. They illustrate that the choice $h_{min} = ||K||/\sqrt{n}$ or $\det(H_{min}) = ||K||/\sqrt{n}$ can be a bad choice for irregular densities. This is noticeable with the mixed uniform density (MU) in dimension 1 (Fig. C.1) or with uniform (U) and fountain (F) distributions in dimension 4 (Fig. C.4).

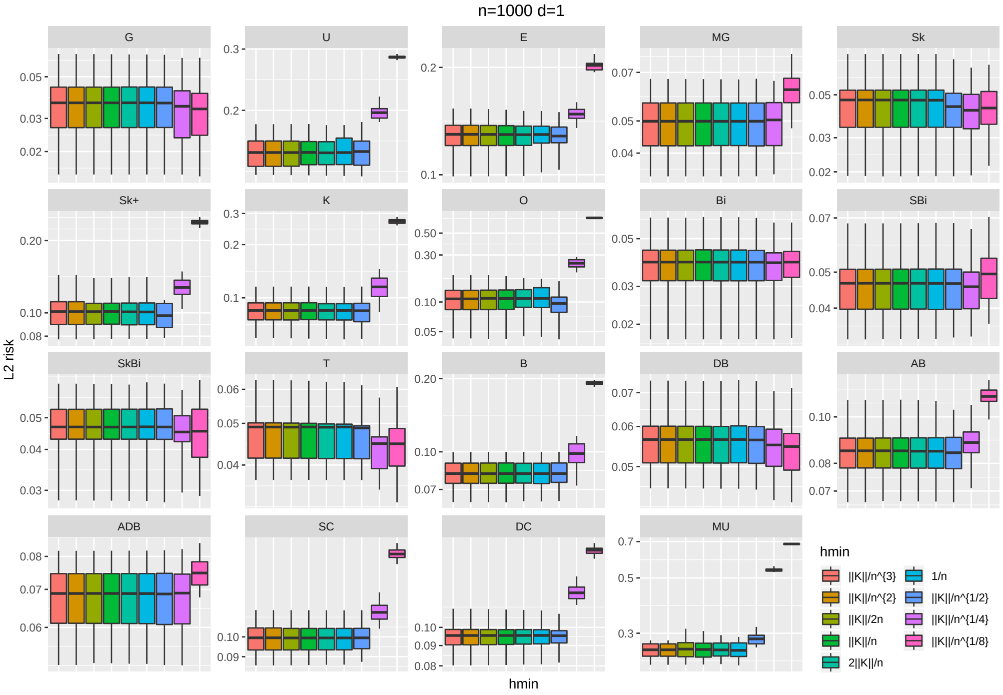


Figure C.1. Boxplots of $ISE^{1/2}_{PCO}(f)$ over 20 trials for 9 values of h_{min} tested on the 19 benchmark 1-dimensional densities for $n = 1000$.

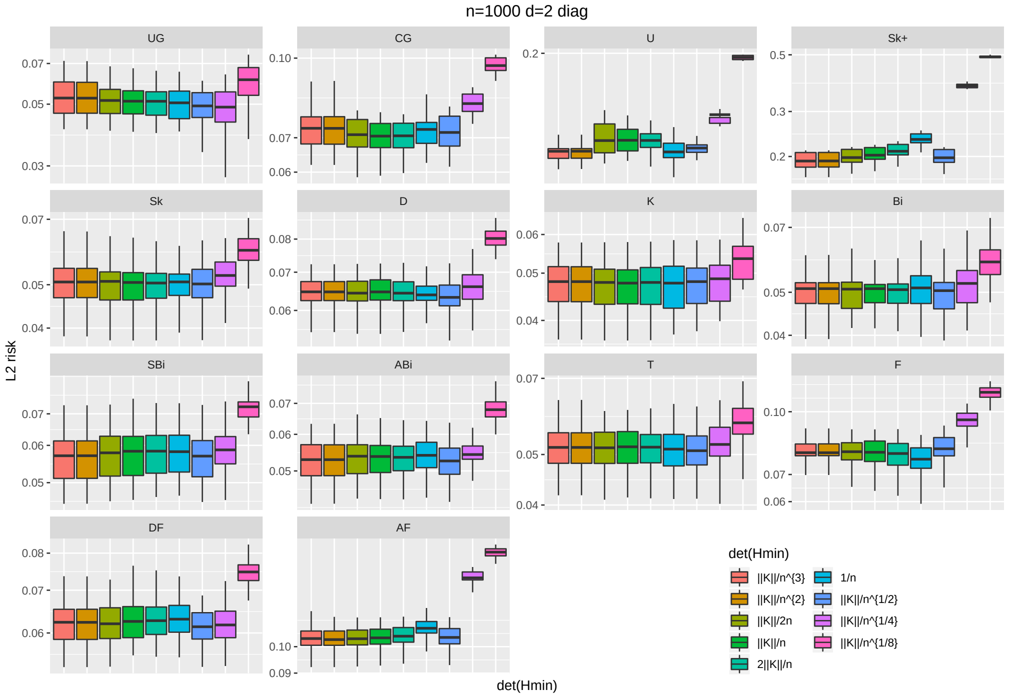


FIGURE C.2. Boxplots of $ISE_{PCO}^{1/2}(f)$ over 20 trials for 9 values of $\det(H_{min})$ with diagonal bandwidth tested on the 14 benchmark 2-dimensional densities for $n = 1000$.

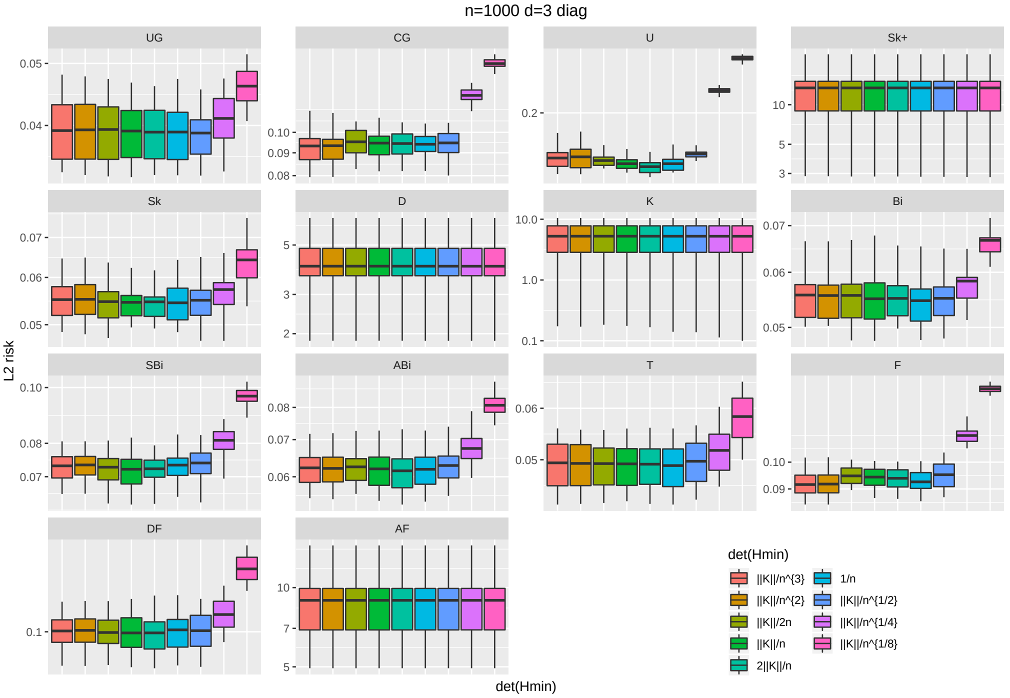


FIGURE C.3. Boxplots of $ISE_{PCO}^{1/2}(f)$ over 20 trials for 9 values of $\det(H_{min})$ with diagonal bandwidth tested on the 14 benchmark 3-dimensional densities for $n = 1000$.

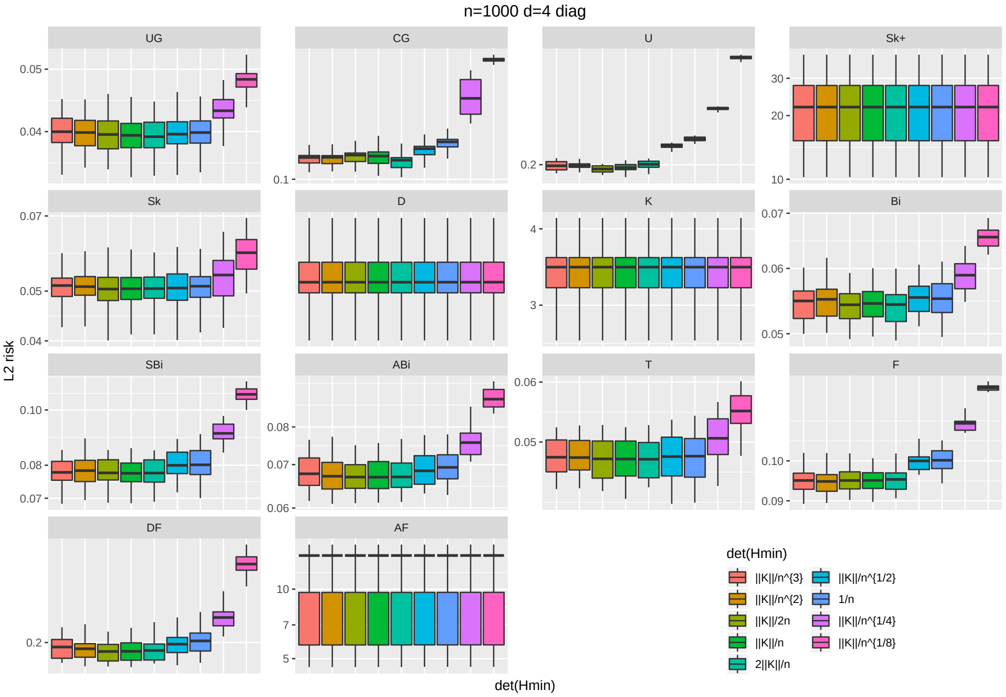


FIGURE C.4. Boxplots of $ISE_{PCO}^{1/2}(f)$ over 20 trials for 9 values of $\det(H_{min})$ with diagonal bandwidth tested on the 14 benchmark 4-dimensional densities for $n = 1000$.

Appendix D. Tables of Monte Carlo mean of ISE values

Table D.1. Monte Carlo mean of $ISE^{1/2}_{\mathrm{meth}}(f)$ over 20 trials with $n = 100$, $n = 1000$ and $n = 10000$ for PCO and Goldenshluger-Lepski methodologies tested on the 19 one-dimensional densities with Gaussian kernel.

	PCO			GL		
	$n = 100$	$n = 1000$	$n = 10000$	$n = 100$	$n = 1000$	$n = 10000$
G	0.08	0.04	0.012	0.08	0.03	0.012
U	0.26	0.15	0.085	0.26	0.16	0.088
E	0.24	0.13	0.072	0.24	0.13	0.074
MG	0.13	0.05	0.021	0.14	0.05	0.021
Sk	0.11	0.04	0.017	0.10	0.04	0.016
Sk+	0.22	0.09	0.039	0.22	0.10	0.040
K	0.23	0.09	0.036	0.24	0.09	0.036
O	0.24	0.11	0.044	0.22	0.11	0.043
Bi	0.09	0.04	0.015	0.09	0.04	0.015
SB	0.12	0.05	0.019	0.12	0.05	0.018
SkB	0.10	0.04	0.019	0.11	0.04	0.019
T	0.10	0.04	0.018	0.10	0.04	0.018
B	0.20	0.08	0.034	0.19	0.08	0.033
DB	0.10	0.06	0.041	0.10	0.05	0.040
AB	0.16	0.08	0.037	0.16	0.08	0.040
ADB	0.13	0.07	0.037	0.13	0.07	0.038
SC	0.20	0.10	0.047	0.20	0.10	0.049
DC	0.19	0.10	0.039	0.19	0.10	0.040
MU	0.50	0.26	0.150	0.51	0.27	0.149

TABLE D.2. Monte Carlo mean of $ISE^{1/2}_{\mathrm{meth}}(f)$ over 20 trials for 4 methodologies described in Section 2.2 with diagonal bandwidth tested on the 14 benchmark 2-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\mathrm{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)$.

	UCV			PI			SCV			PCO		
n	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4
UG	0.111	**0.049**	**0.023**	**0.097**	**0.047**	**0.023**	**0.093**	**0.047**	**0.023**	0.109	**0.049**	**0.023**
CG	0.142	**0.070**	**0.033**	**0.138**	**0.068**	**0.033**	**0.134**	**0.068**	**0.033**	**0.141**	**0.071**	0.035
U	0.234	**0.150**	**0.102**	**0.219**	**0.155**	0.113	**0.220**	0.159	0.115	**0.225**	**0.150**	0.110
Sk+	**0.355**	0.487	**0.095**	**0.347**	**0.199**	**0.096**	0.367	0.216	0.104	0.384	0.234	0.158
Sk	**0.108**	0.056	**0.024**	**0.105**	**0.053**	**0.024**	**0.108**	**0.053**	**0.024**	**0.109**	0.056	**0.025**
D	**0.118**	0.077	0.281	**0.115**	**0.066**	**0.031**	**0.119**	0.072	0.033	**0.118**	**0.067**	**0.031**
K	**0.101**	**0.049**	**0.024**	**0.097**	**0.051**	**0.025**	**0.099**	**0.051**	**0.025**	**0.099**	**0.049**	**0.025**
Bi	0.110	**0.050**	**0.023**	**0.102**	**0.050**	**0.023**	**0.107**	**0.052**	**0.023**	**0.108**	**0.050**	**0.024**
SBi	**0.120**	0.061	**0.027**	**0.116**	**0.057**	**0.027**	**0.120**	**0.058**	**0.027**	**0.119**	0.061	**0.027**
ABi	**0.109**	**0.058**	**0.025**	**0.108**	**0.057**	**0.025**	**0.110**	**0.057**	**0.025**	**0.109**	**0.058**	**0.025**
T	**0.099**	**0.050**	**0.024**	**0.097**	**0.048**	**0.024**	**0.102**	**0.049**	**0.024**	**0.099**	**0.050**	**0.025**
F	**0.166**	**0.078**	**0.038**	**0.173**	0.087	**0.040**	0.187	0.089	**0.040**	**0.165**	**0.077**	0.042
DF	**0.121**	**0.063**	**0.037**	**0.120**	**0.063**	**0.037**	0.130	0.066	**0.038**	**0.120**	**0.063**	**0.036**
AF	**0.179**	**0.103**	**0.053**	0.190	0.125	0.067	0.202	0.132	0.070	**0.179**	**0.108**	**0.054**

TABLE D.3. Monte Carlo mean of $ISE^{1/2}_{\mathrm{meth}}(f)$ over 20 trials for 4 methodologies described in Section 2.2 with diagonal bandwidth tested on the 14 benchmark 3-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\mathrm{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)$.

	UCV		PI		SCV		PCO	
n	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3
UG	0.073	**0.039**	0.115	**0.040**	**0.068**	**0.038**	**0.070**	**0.039**
CG	0.202	**0.092**	0.335	**0.092**	**0.163**	0.103	**0.155**	**0.092**
U	0.250	**0.180**	0.316	**0.177**	**0.234**	**0.185**	**0.235**	**0.179**
Sk+	**14.973**	**15.928**	**14.976**	**15.928**	**14.973**	**15.928**	**14.973**	**15.927**
Sk	0.107	**0.053**	0.114	0.081	**0.093**	0.100	0.098	**0.053**
D	**4.728**	**4.686**	**4.728**	**4.687**	**4.728**	**4.687**	**4.728**	**4.686**
K	**5.498**	**5.392**	**5.496**	**5.391**	**5.495**	**5.391**	**5.496**	**5.391**
Bi	0.109	**0.054**	0.145	**0.055**	**0.097**	0.060	**0.100**	**0.055**
SBi	0.132	**0.071**	0.162	**0.070**	**0.120**	0.077	**0.124**	**0.072**
ABi	0.122	**0.065**	0.140	**0.065**	**0.108**	0.073	**0.113**	**0.066**
T	0.101	**0.052**	0.120	**0.051**	**0.086**	0.056	**0.090**	**0.052**
F	**0.153**	**0.094**	**0.148**	0.103	0.167	0.124	**0.151**	**0.095**
DF	0.147	**0.101**	0.171	**0.103**	**0.140**	0.107	**0.138**	**0.101**
AF	**8.205**	**7.695**	**8.205**	**7.695**	**8.204**	**7.696**	**8.205**	**7.695**

TABLE D.4. Monte Carlo mean of $ISE^{1/2}_{\mathrm{meth}}(f)$ over 20 trials for 4 methodologies described in Section 2.2 with diagonal bandwidth tested on the 14 benchmark 4-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\mathrm{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)$.

	UCV		PI		SCV		PCO	
n	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3
UG	0.075	**0.041**	0.140	0.044	**0.069**	**0.041**	**0.069**	**0.041**
CG	**0.193**	**0.110**	0.474	**0.114**	**0.191**	0.132	**0.194**	0.116
U	**0.261**	**0.199**	0.453	**0.201**	**0.251**	**0.202**	0.267	**0.209**
Sk+	**20.373**	**25.881**	**20.375**	**25.883**	**20.374**	**25.884**	**20.373**	**25.883**
Sk	0.142	0.055	0.102	0.082	**0.077**	0.098	**0.078**	**0.051**
D	**2.503**	**2.473**	**2.504**	**2.472**	**2.503**	**2.472**	**2.503**	**2.472**
K	**3.345**	**3.341**	**3.345**	**3.341**	**3.345**	**3.341**	**3.345**	**3.341**
Bi	0.096	**0.055**	0.179	**0.056**	**0.088**	0.061	**0.086**	**0.055**
SBi	0.133	**0.080**	0.188	**0.081**	**0.121**	0.090	**0.121**	**0.080**
ABi	**0.113**	**0.069**	0.157	0.079	**0.109**	**0.069**	**0.109**	**0.069**
T	**0.077**	**0.047**	0.107	0.060	**0.077**	**0.047**	**0.075**	**0.047**
F	**0.138**	**0.094**	**0.135**	**0.095**	0.142	0.111	**0.134**	**0.095**
DF	0.234	**0.199**	0.259	**0.207**	**0.221**	**0.200**	**0.218**	**0.199**
AF	**17.498**	**15.178**	**17.497**	**15.182**	**17.497**	**15.183**	**17.497**	**15.182**

TABLE D.5. Monte Carlo mean of $ISE^{1/2}_{\mathrm{meth}}(f)$ over 20 trials for 5 methodologies described in Section 2.2 with non-diagonal bandwidth tested on the 14 benchmark 2-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\mathrm{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\mathrm{meth}} \overline{ISE}^{1/2}_{\mathrm{meth}}(f)$.

	UCV			RoT			PI		
n	10^2	10^3	10^4	10^2	10^3	10^4	10^2	10^3	10^4
UG	0.124	0.055	**0.023**	**0.093**	**0.047**	**0.022**	**0.097**	**0.047**	**0.023**
CG	0.123	0.052	**0.025**	**0.098**	**0.048**	**0.024**	**0.102**	**0.049**	**0.024**
U	0.249	**0.152**	**0.102**	**0.219**	0.161	0.128	**0.220**	**0.155**	0.113
Sk+	0.297	**0.153**	**0.068**	0.325	0.226	0.133	**0.272**	**0.151**	**0.069**
Sk	0.125	**0.057**	**0.025**	0.253	0.223	0.182	0.196	0.094	0.032
D	0.115	0.056	**0.024**	0.104	0.067	0.037	**0.098**	**0.053**	**0.024**
K	0.110	**0.048**	**0.022**	0.112	0.078	0.050	**0.100**	0.051	0.024
Bi	0.116	**0.051**	**0.022**	**0.104**	0.057	0.028	**0.100**	**0.048**	**0.022**
SBi	0.129	0.059	**0.025**	**0.119**	0.065	0.035	**0.115**	**0.054**	**0.025**
ABi	0.116	**0.055**	**0.023**	0.160	0.109	0.064	0.123	0.059	**0.024**
T	0.108	0.049	**0.023**	0.102	0.059	0.032	**0.094**	**0.045**	**0.023**
F	**0.172**	**0.078**	**0.039**	0.187	0.132	0.084	0.174	0.087	**0.040**
DF	**0.117**	**0.060**	**0.035**	0.146	0.096	0.059	**0.115**	**0.060**	**0.036**
AF	**0.164**	**0.086**	**0.042**	0.202	0.164	0.126	0.190	0.117	0.057

	SCV			PCO		
n	10^2	10^3	10^4	10^2	10^3	10^4
UG	**0.094**	**0.047**	**0.023**	0.110	**0.049**	**0.023**
CG	**0.100**	**0.048**	**0.024**	0.114	0.051	0.026
U	**0.221**	0.159	0.115	**0.226**	**0.149**	0.110
Sk+	**0.285**	**0.154**	**0.069**	0.325	**0.157**	0.088
Sk	0.210	0.104	0.037	**0.114**	**0.055**	**0.025**
D	**0.103**	**0.054**	**0.024**	**0.103**	**0.052**	**0.024**
K	**0.103**	0.052	0.024	**0.098**	**0.048**	0.024
Bi	**0.105**	**0.049**	**0.022**	0.107	0.051	0.024
SBi	**0.119**	**0.054**	**0.025**	0.123	**0.055**	**0.025**
ABi	0.125	0.058	**0.024**	**0.103**	**0.053**	**0.023**
T	**0.099**	**0.046**	**0.023**	0.103	0.049	**0.023**
F	0.187	0.089	**0.040**	**0.165**	**0.076**	0.042
DF	0.121	**0.061**	**0.036**	**0.118**	**0.062**	**0.036**
AF	0.202	0.122	0.059	**0.167**	0.095	**0.043**

TABLE D.6. Monte Carlo mean of $ISE^{1/2}_{\text{meth}}(f)$ over 20 trials for 5 methodologies described in Section 2.2 with non-diagonal bandwidth tested on the 14 benchmark 3-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\text{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\text{meth}} \overline{ISE}^{1/2}_{\text{meth}}(f)$.

	UCV		RoT		PI		SCV		PCO	
n	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3
UG	0.112	0.042	**0.069**	**0.038**	0.115	0.053	**0.069**	**0.038**	**0.071**	**0.039**
CG	0.181	0.078	**0.114**	**0.066**	0.199	0.092	**0.114**	**0.066**	0.122	**0.069**
U	0.286	**0.184**	**0.235**	**0.180**	0.315	**0.187**	**0.235**	**0.180**	**0.235**	**0.178**
Sk+	**13.882**	**14.098**	14.972	15.927	14.971	15.926	14.971	15.926	14.973	15.927
Sk	0.137	0.087	0.190	0.177	0.154	0.099	0.164	0.102	**0.097**	**0.054**
D	**4.179**	**3.541**	4.728	4.687	4.727	4.685	4.728	4.686	4.727	4.685
K	**4.862**	**4.109**	5.495	5.391	5.496	5.391	5.495	5.391	5.496	5.390
Bi	0.131	**0.054**	**0.096**	0.059	0.131	0.062	**0.094**	**0.052**	**0.098**	**0.053**
SBi	0.162	**0.067**	**0.124**	0.083	0.142	0.071	**0.118**	**0.066**	**0.122**	0.070
ABi	0.136	**0.064**	0.149	0.114	0.128	0.071	0.119	0.067	**0.107**	**0.062**
T	0.111	**0.050**	0.092	0.062	0.102	0.052	**0.086**	**0.050**	**0.087**	**0.049**
F	0.163	**0.095**	0.171	0.142	**0.149**	**0.095**	0.167	0.110	**0.152**	**0.096**
DF	0.146	**0.099**	0.153	0.121	0.158	**0.103**	**0.137**	**0.099**	**0.138**	**0.100**
AF	**7.912**	**6.563**	**8.204**	7.696	**8.205**	7.695	**8.204**	7.695	**8.204**	7.694

TABLE D.7. Monte Carlo mean of $ISE^{1/2}_{\text{meth}}(f)$ over 20 trials for 5 methodologies described in Section 2.2 with non-diagonal bandwidth tested on the 14 benchmark 4-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\text{meth}}(f)$ is in bold when it is not larger than $1.05 \times \min_{\text{meth}} \overline{ISE}^{1/2}_{\text{meth}}(f)$.

	UCV		RoT		PI		SCV		PCO	
n	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3	10^2	10^3
UG	0.128	0.046	**0.070**	**0.041**	0.144	0.068	**0.070**	**0.041**	**0.070**	**0.041**
CG	0.271	0.089	**0.146**	**0.082**	0.284	0.136	**0.145**	**0.082**	0.186	0.098
U	0.383	**0.205**	**0.257**	**0.197**	0.449	0.246	**0.253**	**0.198**	0.272	0.212
Sk+	14.264	15.316	14.187	16.195	**12.486**	**10.438**	**12.600**	11.517	20.373	25.883
Sk	0.141	0.056	0.139	0.132	0.126	0.098	0.129	0.098	**0.078**	**0.051**
D	**2.338**	**1.904**	2.503	2.472	2.503	2.471	2.503	2.472	2.503	2.472
K	**3.332**	**2.780**	**3.345**	3.341	**3.345**	3.341	**3.345**	3.341	**3.344**	3.340
Bi	0.148	0.054	**0.085**	0.056	0.154	0.074	**0.084**	**0.052**	**0.084**	**0.053**
SBi	0.184	**0.075**	0.122	0.092	0.158	0.083	**0.115**	**0.075**	**0.121**	**0.079**
ABi	0.151	**0.065**	0.137	0.113	0.136	0.081	0.119	0.075	**0.104**	**0.063**
T	0.104	0.049	0.077	0.055	0.097	0.052	**0.073**	**0.046**	**0.072**	**0.045**
F	0.162	**0.095**	0.142	0.128	**0.135**	**0.095**	**0.141**	0.111	**0.134**	**0.096**
DF	0.253	**0.198**	**0.223**	**0.207**	0.243	**0.203**	**0.218**	**0.199**	**0.217**	**0.199**
AF	**17.674**	**15.386**	**17.497**	**15.184**	**17.497**	**15.182**	**17.497**	**15.183**	**17.497**	**15.182**

Table D.8. Monte Carlo mean of $ISE^{1/2}_{\text{meth}}(f)$ over 20 trials for PCO with diagonal and full matrices bandwidths tested on the 14 benchmark 2-dimensional densities for different values of n. The Monte Carlo mean $\overline{ISE}^{1/2}_{\text{meth}}(f)$ is in green (resp. red) when using full (resp. diagonal) matrices gives better results. Black results corresponds to cases where there is no significant difference.

	$n = 10^2$		$n = 10^3$		$n = 10^4$	
	H diag	H full	H diag	H full	H diag	H full
UG	0.109	0.110	0.049	0.049	0.023	0.023
CG	0.141	0.114	0.071	0.051	0.035	0.026
U	0.225	0.226	0.150	0.149	0.110	0.110
Sk+	0.384	0.325	0.234	0.157	0.158	0.088
Sk	0.109	0.114	0.056	0.055	0.025	0.025
D	0.118	0.103	0.067	0.052	0.031	0.024
K	0.099	0.098	0.049	0.048	0.025	0.024
Bi	0.108	0.107	0.050	0.051	0.024	0.024
SBi	0.119	0.123	0.061	0.055	0.027	0.025
ABi	0.109	0.103	0.058	0.053	0.025	0.023
T	0.099	0.103	0.050	0.049	0.025	0.023
F	0.165	0.165	0.077	0.076	0.042	0.042
DF	0.120	0.118	0.063	0.062	0.036	0.036
AF	0.179	0.167	0.108	0.095	0.054	0.043

References

[1] N. Akakpo, Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Stat.* **21** (2012) 1–28.

[2] A.W. Bowman, An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** (1984) 353–360.

[3] J.E. Chacón, Data-driven choice of the smoothing parametrization for kernel density estimators. *Can. J. Stat.* **37** (2009) 249–265.

[4] J.E. Chacón and T. Duong, Multivariate plug-in bandwidth selection with unconstrained pilot matrices. *Test* **19** (2010) 375–398.

[5] J.E. Chacón and T. Duong, Unconstrained pilot selectors for smoothed cross validation. *Aust. N. Zeal. J. Stat.* **53** (2011) 331–351.

[6] G. Cleanthous, A.G. Georgiadis and E. Porcu, Minimax Density Estimation on Sobolev Spaces With Dominating Mixed Smoothness (2019). 10.48550/arXiv.1906.06835.

[7] L. Devroye, The double kernel method in density estimation, in Annales de l'IHP Probabilités et statistiques, vol. 25 (1989) 533–580.

[8] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, Density estimation by wavelet thresholding. *Ann. Stat.* **24** (1996) 508–539.

[9] T. Duong, ks: Kernel Smoothing (2017), R package version 1.10.6.

[10] T. Duong and M. Hazelton, Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Stat.* **32** (2005) 485–506.

[11] A. Goldenshluger and O. Lepski, Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** (2008) 1150–1190.

[12] A. Goldenshluger and O. Lepski, Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields* **143** (2009) 41–71.

[13] A. Goldenshluger and O. Lepski, Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** (2011) 1608–1632.

[14] A. Goldenshluger and O. Lepski, On adaptive minimax density estimation on $\mathbb{R}^d$. *Theory Probab. Appl.* **159** (2014) 479–543.

[15] A.V. Goldenshluger and O.V. Lepski, General selection rule from a family of linear estimators. *Theory Probab. Appl.* **57** (2013) 209–226.

[16] N.-B. Heidenreich, A. Schindler and S. Sperlich, Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *Adv. Stat. Anal.* **97** (2013) 403–433.

[17] M. Jones, On some kernel density estimation bandwidth selectors related to the double kernel method. *Sankhyā* **60** (1998) 249–264.

[18] M. Jones, J.S. Marron and B.U. Park, A simple root n bandwidth selector. *Ann. Stat.* **19** (1991) 1919–1932.

[19] C. Lacour, P. Massart and V. Rivoirard, Estimator selection: a new method with applications to kernel density estimation. *Sankhya* **79** (2017) 298–335.

[20] M. Lerasle, N. Malter-Magalahes and P. Reynaud-Bouret, Optimal kernel selection for density estimation. *High Dimens. Probab.* **VII: The Cargese Volume** (2016) 425–460.

[21] J. Marron and M. Wand, Exact mean integrated squared error. *Ann. Stat.* **20** (1992) 712–736.

[22] E. Parzen, On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** (1962) 1065–1076.

[23] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2015).

[24] M. Rudemo, Empirical choice of histograms and kernel density estimators. *Scand. J. Stat. Theory Appl.* **9** (1982) 65–78.

[25] D.W. Scott and G.R. Terrell, Selectors unbiased cross-validation in density estimation. *J. Am. Stat. Assoc.* **82** (1987) 1131–1146.

[26] S.J. Sheather and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Stat. Soc. B* **53** (1991) 683–690.

[27] B.W. Silverman, Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986).

[28] I.M. Sobol, On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **7** (1967) 86–112.

[29] C. Stone, An asymptotically optimal window selection rule for kernel density estimates. *Ann. Stat.* **12** (1984) 1285–1297.

[30] M. Wand, Error analysis for general multivariate kernel estimators. *J. Nonparam. Stat.* **2** (1992) 1–15.

[31] M. Wand and M. Jones, Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Am. Stat. Assoc.* **88** (1993) 520–528.

[32] M. Wand and M. Jones, Kernel smoothing. Monographs on Statistics and Applied Probability (1994).

[33] M. Wand and M. Jones, Multivariate plugin bandwidth selection. *Comput. Stat.* **9** (1994) 97–116.