Journal of Multivariate Analysis 173 (2019) 248–267

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Adaptive optimal kernel density estimation for directional data

Thanh Mai Pham Ngoc

Laboratoire de mathématique d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France

ARTICLE INFO

Article history:
Received 7 August 2018
Available online 27 February 2019

AMS 2010 subject classifications:
primary 62G07

Keywords:
Bandwidth selection
Directional data
Kernel density estimation
Oracle inequality
Penalization methods

ABSTRACT

This paper considers nonparametric density estimation with directional data. A new rule is proposed for bandwidth selection for kernel density estimation. The procedure is automatic, fully data-driven, and adaptive to the degree of smoothness of the density. An oracle inequality and optimal rates of convergence for the L_2 error are derived. These theoretical results are illustrated with simulations.

1. Introduction

Directional data arise in many fields such as wind direction for the circular case, and astrophysics, paleomagnetism, geology for the spherical case. Much effort has been made to devise statistical methods to tackle the density estimation problem. We refer to [19] and more recently to [18] for a comprehensive review. Nonparametric procedures have been well developed.

In this article we focus on kernel density estimation. Various works [3,11] have used projection methods on localized bases adapted to the sphere. Classical references for kernel density estimation with directional data include the seminal papers [2,9]. It is well known that the choice of the bandwidth is a key and intricate issue when using kernel methods. In practice, various techniques for selecting the bandwidth have been suggested since the popular cross-validation rule in [9]. Let us mention the plug-in and refined cross-validatory methods in [21,24] for the circular case, and [5] on the torus.

Recently, García-Portugués [6] devised an equivalent of the rule-of-thumb of [23] for directional data, and Amiri et al. [1] explored computational problems with recursive kernel estimators based on the cross-validation procedure of [9]. To the best of our knowledge, however, the various rules that have been proposed so far for selecting the bandwidth in practice have not been assessed from a theoretical point of view. In particular, there are no results proving that cross-validation is adaptively rate-optimal, even in the linear case. From a theoretical point of view, Klemelä [13] studied convergence rates for L_2 error over some regularity classes. Unfortunately, the asymptotically optimal bandwidth in [13] depends on the density and its degree of smoothness, which is infeasible in practice.

In the linear case, kernel bandwidth selection rules have been proposed, leading to adaptive estimators which attain optimal rates of convergence. By adaptive we mean that the estimator is adaptive to the degree of smoothness of the underlying density: the method does not require the specification of the regularity of the density. In this regard, we may

E-mail address: thanh.pham_ngoc@math.u-psud.fr.

https://doi.org/10.1016/j.jmva.2019.02.009

cite the remarkable series of papers [8,15,16] and the recent work of Lacour et al. [14]. The drawback of the methods in [8,15,16] is that they involve tuning parameters. It is well known that in nonparametric statistics, minimax and oracle theoretical results rarely give optimal choices for tuning parameters from a practical point of view with very conservative choices. The major interest of the procedure in [14] is that it is free of tuning parameters, which constitutes a great advantage in practice. The approach in [14] called PCO (Penalized Comparison of Overfitting) is based on concentration inequalities for U-statistics.

In the present paper, we aim at filling the gap between theory and practice in the directional kernel density estimation literature. Our goal is to construct a fully data-driven bandwidth selection rule providing an adaptive estimator which reaches minimax rates of convergence for L_2 risk over some regularity classes. This motivates our choice to adapt the method of Lacour et al. [14] to the directional setting. Our procedure is simple to implement and in examples based on simulations, it shows quite good performances in a reasonable computation time.

This paper is organized as follows. In Section 2, we present our estimation procedure. In Section 3 we provide an oracle inequality and rates of convergences of our estimator for the MISE (Mean Integrated Squared Error). Section 4 gives some numerical illustrations. Section 5 gives the proofs of theorems. Finally, the Appendix gathers technical propositions and lemmas.

The following notation is used throughout. For two integers a, b, we denote $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For arbitrary $y \in \mathbb{R}$, $\lfloor y \rfloor$ denotes the integer part of y. Depending on the context, $\|\cdot\|$ denotes the classical L_2 norm on $\mathbb{R}$ or $\mathbb{S}^{d-1}$. For any integer $d \geq 2$, we denote the unit sphere of $\mathbb{R}^d$ by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = 1\}$ and the associated scalar product by $\langle \cdot, \cdot \rangle$. For a vector $x \in \mathbb{R}^d$, $\|x\|$ stands for the Euclidean norm on $\mathbb{R}^d$ while $\|\cdot\|_\infty$ is the usual L_∞-norm on $\mathbb{S}^{d-1}$. Finally, the scalar product of two vectors x and y, is denoted by $x^\top y$, where $^\top$ is the transpose operator.

2. Estimation procedure

We are given n mutually independent and identically distributed observations $X_1, \ldots, X_n$ on $\mathbb{S}^{d-1}$ for some integer $d \geq 2$. The X_is are absolutely continuous with respect to the Lebesgue measure ω_d on $\mathbb{S}^{d-1}$ with common density f. Therefore, a directional density f satisfies

$$\int_{\mathbb{S}^{d-1}} f(x)\omega_d(dx) = 1.$$

We aim at constructing an adaptive kernel estimator of the density f with a fully data-driven choice of bandwidth.

2.1. Directional approximation kernel

We present some classical conditions that are required for the kernel.

Assumption 1. The kernel $K : [0, \infty) \to [0, \infty)$ is a bounded and Riemann integrable function such that

$$0 < \int_0^\infty x^{(d-3)/2} K(x) dx < \infty.$$

Assumption 1 is usual in kernel density estimation with directional data; see, e.g., Assumptions D1–D3 in [7] and Assumption A1 in [1]. An example of kernel which satisfies Assumption 1 is the popular von Mises kernel $K(x) = e^{-x}$.

2.2. Family of directional kernel estimators

We consider the following standard directional kernel density estimator $\left(K_h(x, y) = K\left\{(1 - x^\top y)/h\right\}\right)$. For all $x \in \mathbb{S}^{d-1}$,

$$\hat{f}_h(x) = \frac{c_0(h)}{n} \sum_{i=1}^n K\left(\frac{1 - x^\top X_i}{h^2}\right) = \frac{c_0(h)}{n} \sum_{i=1}^n K_{h^2}(x, X_i),$$

where K is a kernel satisfying Assumption 1 and $c_0(h)$ a normalizing constant such that $\hat{f}_h(x)$ integrates to unity:

$$c_0^{-1}(h) = \int_{\mathbb{S}^{d-1}} K_{h^2}(x, y)\omega_d(dy).$$

It remains to select a convenient value for the bandwidth h.

Remark 1. Note that $c_0(h)$ does not depend on x. The "tangent-normal" decomposition (see [19]) says that if y is a vector and x a fixed element of $\mathbb{S}^{d-1}$, then denoting $t = x^\top y$ their scalar product, we may always write $y = tx + (1 - t^2)^{1/2}\xi$, where ξ is a unit vector orthogonal to x. Further, the area element on $\mathbb{S}^{d-1}$ can be written as

$$\omega_d(dx) = (1 - t^2)^{(d-3)/2} dt\ \omega_{d-1}(d\xi).$$

Thus, using these conventions, one obtains

$$
\begin{aligned}
c_0^{-1}(h) &= \int_{\mathbb{S}^{d-1}} K\left(\frac{1-x^\top y}{h^2}\right)\omega_d(dy) \\
&= \int_{\mathbb{S}^{d-2}}\int_{-1}^{1} K\left[\frac{1-x^\top\{tx+(1-t^2)^{1/2}\xi\}}{h^2}\right](1-t^2)^{(d-3)/2}dt\ \omega_{d-1}(d\xi) \\
&= \int_{\mathbb{S}^{d-2}}\omega_{d-1}(d\xi)\int_{-1}^{1} K\left(\frac{1-tx^\top x}{h^2}\right)(1-t^2)^{(d-3)/2}dt \\
&= \sigma_{d-2}\int_{-1}^{1} K\left(\frac{1-t}{h^2}\right)(1-t^2)^{(d-3)/2}dt,
\end{aligned}
$$

where $\sigma_{d-1}=\omega_d(\mathbb{S}^{d-1})$ denotes the area of $\mathbb{S}^{d-1}$. We recall that $\sigma_{d-1}=(2\pi^{d/2})/\Gamma(d/2)$ with Γ the Gamma function.

2.3. Bandwidth selection

In kernel density estimation, a delicate step consists in selecting the proper bandwidth h for $\hat{f}_h$. We present our data-driven choice of bandwidth $\hat{h}$ inspired from [14]. We name our procedure SPCO (Spherical Penalized Comparison to Overfitting). Consider a set $\mathcal{H}$ of bandwidths defined by

$$
\mathcal{H}=\left\{h:\left\{\frac{\|K\|_\infty}{n}\frac{1}{R_0(K)}\right\}^{1/(d-1)}\le h\le 1,\ \text{and } 1/h \text{ is an integer}\right\}, \tag{1}
$$

with $R_0(K)=2^{(d-3)/2}\sigma_{d-2}\int_0^\infty x^{(d-3)/2}K(x)dx$. We obtain the selected bandwidth by setting, for $\lambda\in\mathbb{R}$,

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

where $h_{\min}=\min\mathcal{H}$ and the penalty term $\operatorname{pen}_\lambda(h)$ is defined, for $h\in\mathcal{H}$, as

$$
\operatorname{pen}_\lambda(h)=\frac{\lambda c_0^2(h)c_2(h)}{n}-\frac{1}{n}\int_{\mathbb{S}^{d-1}}\{c_0(h_{\min})K_{h_{\min}^2}(x,y)-c_0(h)K_{h^2}(x,y)\}^2\omega_d(dy), \tag{3}
$$

with $c_2(h)=\int_{\mathbb{S}^{d-1}}K_{h^2}^2(x,y)\omega_d(dy)$.

Our SPCO estimator of f is $\hat{f}_{\hat{h}}$. The procedure SPCO involves a real parameter λ. In Section 3, we study how to choose the optimal value of λ leading to a fully data-driven procedure.

Remark 2. Let us give some explanations about the terms involved in the expression of the selection rule (2). One can decompose the risk $\mathrm{E}\|f-\hat{f}_h\|^2$ with the classical bias–variance decomposition. Hence, heuristically, the idea is to find the best bandwidth h minimizing an estimate of the bias–variance decomposition of the risk. Developing the quantity

$$
\|\hat{f}_h-\hat{f}_{h_{\min}}\|^2-\frac{1}{n}\int_{\mathbb{S}^{d-1}}\{c_0(h_{\min})K_{h_{\min}^2}(x,y)-c_0(h)K_{h^2}(x,y)\}^2\omega_d(dy),
$$

one realizes that it is in fact an estimator of the bias. Since the variance is bounded by $c_0^2(h)c_2(h)/n$, the term $\lambda c_0^2(h)c_2(h)/n$ acts as an estimator of the variance term. For more details, see Section 3.1 in [14].

Note that again $c_2(h)$ and $\operatorname{pen}_\lambda(h)$ do not depend on x using Remark 1. Indeed, similar computations lead to

$$
c_2(h)=\sigma_{d-2}\int_{-1}^{1}K^2\left(\frac{1-t}{h^2}\right)(1-t^2)^{(d-3)/2}dt,
$$

and

$$
\operatorname{pen}_\lambda(h)=\frac{\lambda c_0^2(h)c_2(h)}{n}-\frac{\sigma_{d-2}}{n}\int_{-1}^{1}\left\{c_0(h_{\min})K\left(\frac{1-t}{h_{\min}^2}\right)-c_0(h)K\left(\frac{1-t}{h^2}\right)\right\}^2(1-t^2)^{(d-3)/2}dt.
$$

3. Rates of convergence

3.1. Oracle inequality

First, we state an oracle-type inequality which highlights the bias–variance decomposition of the L_2 risk when $\lambda>0$. In what follows, $|\mathcal{H}|$ denotes the cardinality of the set $\mathcal{H}$. We denote $f_h=\mathrm{E}(\hat{f}_h)$.

Theorem 1. *Assume that the kernel K satisfies Assumption 1 and $\|f\|_\infty < \infty$. Let $x \geq 1$ and $\varepsilon \in (0,1)$. Then there exists n_0 independent of f such that, for $n \geq n_0$, with probability larger than $1 - C_1|\mathcal{H}|e^{-x}$,*

$$\|\hat{f}_{\hat{h}} - f\|^2 \leq C_0(\varepsilon,\lambda)\min_{h\in\mathcal{H}}\|\hat{f}_h - f\|^2 + C_2(\varepsilon,\lambda)\|f_{h_{\min}} - f\|^2 + C_3(\varepsilon,K,\lambda)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}, \quad (4)$$

where C_1 is an absolute constant and $C_0(\varepsilon,\lambda) = \lambda + \varepsilon$ if $\lambda \geq 1$, $C_0(\varepsilon) = 1/\lambda + \varepsilon$ 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.

This oracle inequality bounds the quadratic risk of SPCO estimator by the infimum over $\mathcal{H}$ of the tradeoff between the approximation term $\|f_{h_{\min}} - f\|^2$ and the variance term $\|\hat{f}_h - f\|^2$ provided that $\lambda > 0$. In fact, we need that $\lambda > 0$ to use concentration inequalities to prove the oracle inequality. The terms $C_3(\varepsilon,K,\lambda)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}$ are remainder terms. Hence, this oracle inequality justifies our selection rule. For further details about oracle inequalities and model selection see [20].

Nonetheless one could wonder what would happen if $\lambda < 0$. The next theorem shows that we cannot choose λ too small ($\lambda < 0$) as it would lead to select a bandwidth close to $h_{\min}$ with high probability. One would obtain an overfitting estimator. To this purpose, we suppose

$$\|f - f_{h_{\min}}\|^2 \frac{n}{c_0^2(h_{\min})c_2(h_{\min})} = o(1). \quad (5)$$

Let us focus on Assumption (5). For $h \in \mathcal{H}$, the bias of $\hat{f}_h$ is equal to $\|f - f_h\|^2$. As $f_{h_{\min}}$ is the best approximation of f among the grid $\mathcal{H}$, the smallest bias for $\hat{f}_h,\ h \in \mathcal{H}$ is equal to $\|f - f_{h_{\min}}\|^2$. Since the variance of $\hat{f}_h$ is of order $c_0^2(h)c_2(h)/n$, this assumption means that the smallest bias $\|f - f_{h_{\min}}\|^2$ is negligible with respect to the corresponding integrated variance. Thus this assumption is mild.

Theorem 2. *Assume that the kernel K satisfies Assumption 1 and $\|f\|_\infty < \infty$. Assume also (5) and, for $\beta > 0$,*

$$\|K\|_\infty/\{nR_0(K)\} \leq h_{\min}^{d-1} \leq (\ln n)^\beta/n.$$

Then if we consider $\mathrm{pen}_\lambda(h)$ defined in (3) with $\lambda < 0$, we have for n large enough, with probability larger than $1 - C_1|\mathcal{H}|e^{-(n/\ln n)^{1/3}}$,

$$\hat{h} \leq C(\lambda)h_{\min} \leq C(\lambda)\{(\ln n)^\beta/n\}^{1/(d-1)},$$

where C_1 is an absolute constant and $C(\lambda) = \{1.23\,(2.1 - 1/\lambda)\}^{1/(d-1)}$.

Remark 3. Theorem 2 invites us to discard $\lambda < 0$. Indeed, setting λ to negative values leads the procedure to select with large probability a bandwidth $\hat{h}$ close to $h_{\min}$. As a result, we would obtain an overfitting estimator which behaves very poorly. Now considering oracle inequality (4), $\lambda = 1$ yields the minimal value of the leading constant $C_0(\varepsilon,\lambda) = \lambda + \varepsilon$. Thus, the theory urges us to take the optimal value $\lambda = 1$ in the SPCO procedure. Actually, we will see in the numerical section that the choice $\lambda = 1$ is quite efficient.

3.2. Free of tuning parameters estimator and rates of convergence

Results of Section 3.1 about the optimality of $\lambda = 1$ enable us to devise our estimator free of tuning parameters. We call it $\hat{f}_{\check{h}}$ with bandwidth $\check{h}$ defined as in (2) with $\lambda = 1$.

We now compute rates of convergence for the MISE of our estimator $\hat{f}_{\check{h}}$ over some smoothness classes. In [13], suitable smoothness classes are defined for the study of the MISE. In particular, these regularity classes involve a concept of an "average" of directional derivatives which was first defined in [9]. Let us recall the definition of these smoothness classes in [13].

Let $\eta \in \mathbb{S}^{d-1}$ and $T_\eta = \{\xi \in \mathbb{S}^{d-1} : \xi \perp \eta\}$. Let $\phi_\eta : \mathbb{S}^{d-1} \setminus \{\eta, -\eta\} \to T_\eta \times (0,\pi)$ be a parameterization of $\mathbb{S}^{d-1}$ defined by

$$\phi_\eta^{-1}(\xi,\theta) = \eta\cos(\theta) + \xi\sin(\theta).$$

When $g : \mathbb{R}^d \to \mathbb{R}$ and $x, \xi \in \mathbb{R}^d$, define the derivative of g at x in the direction of ξ to be $D_\xi g(x) = \lim_{h\to 0} h^{-1}\{g(x + h\xi) - g(x)\}$ and $D_\xi^s g = D_\xi D_\xi^{s-1} g$, for some integer $s \geq 2$.

We will now define the derivative of order s.

Definition 1. Let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ and define $\bar{f} : \mathbb{R}^d \to \mathbb{R}$ by $\bar{f}(x) = f(x/\|x\|)$. The derivative of order s is $D^s f : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by

$$D^s f(x) = \frac{1}{\sigma_{d-1}} \int_{T_x} D_\xi^s \bar{f}(x)\omega_d(d\xi),$$

where $T_x = \{\xi \in \mathbb{S}^{d-1} : \xi \perp x\}$.

Definition 2. When $f : \mathbb{S}^{d-1} \to \mathbb{R}$, define $\tilde{D}^s f : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ by

$$\tilde{D}^s f(x,\theta) = \frac{1}{\sigma_{d-1}} \int_{T_x} D^s_{\phi_x^{-1}(\xi,\theta+\frac{\pi}{2})} \bar{f}\{\phi_x^{-1}(\xi,\theta)\}\omega_d(d\xi).$$

We are now able to define the smoothness class $\mathbf{F}_2(s)$; see [13].

Definition 3. Let $s \geq 2$ be even and $p \in [1,\infty]$. Let $\mathbf{F}_2(s)$ be the set of functions $f : \mathbb{S}^{d-1} \to \mathbb{R}$ such that (i) $\|D^i f\| < \infty$ for all $i \in \{0,\ldots,s\}$; (ii) for all $x \in \mathbb{S}^{d-1}$ and all $\xi \in T_x$, $\partial^s f\{\phi_x^{-1}(\xi,\theta)\}/\partial\theta^s$ is continuous as a function of $\theta \in \mathbb{R}$; (iii) $\|\tilde{D}^s f(\cdot,\theta)\|$ is bounded for $\theta \in [0,\pi]$ and (iv) $\lim_{\theta\to 0} \|\tilde{D}^s f(\cdot,\theta) - D^s f\| = 0$.

To achieve optimal rates of convergence over the class $\mathbf{F}_2(s)$, we need supplementary conditions on the kernel to deal with the bias term. The idea of reducing the bias in the Euclidean case using a class s kernel dates back to [4,22]. In the directional case, this has been early pointed out in [9]. Following [13], we will define what is called a kernel of class s. For all $i \in \mathbb{N}$, let

$$\alpha_i(K) = \int_0^\infty x^{(i+d-3)/2} K(x)dx.$$

Assumption 2. Let $s \geq 0$ be even. The kernel K is of class s, i.e., it is a measurable function $K : [0,\infty) \to \mathbb{R}$ which satisfies:

(i) $\alpha_i(K) < \infty$ for $i \in \{0, s\}$;
(ii) $\alpha_0(K) \neq 0$;
(iii) $\int_0^{h^{-2}} x^{(2i+d-3)/2} K(x)dx = o(h^{s-2i})$ for $i \in \{1,\ldots,s/2-1\}$, when $h \to 0$.

In Assumption 2, s must be even because $D^s f(x) = 0$ for all $x \in \mathbb{S}^{d-1}$ when $s \geq 1$ is odd; see Chapter 2 in [12]. Furthermore, note that von Mises kernel is of order 2.

Now, a direct application of the oracle inequality in Theorem 1 allows us to derive rates of convergence for the MISE of $\hat{f}_{\check{h}}$.

Theorem 3. *Consider a kernel K satisfying Assumptions 1 and 2. For $B > 0$, denote by $\tilde{\mathbf{F}}_2(s,B)$ the set of densities bounded by B and belonging to $\mathbf{F}_2(s)$. Then we have*

$$\sup_{f\in\tilde{\mathbf{F}}(s,B)} \mathrm{E}(\|\hat{f}_{\check{h}} - f\|^2) \leq C(s,K,d,B) n^{-2s/(2s+d-1)},$$

with $C(s,K,d,B)$ a constant depending on s, K, d and B.

Theorem 3 shows that the estimator $\hat{f}_{\check{h}}$ achieves the optimal rate of convergence for estimating a density on $\mathbb{S}^{d-1}$ with an s order smoothness; matching lower bounds are proved in Chapter 6 of [12] and in [3]. Hence estimating on the d-dimensional sphere appears to be analogous to inference in $(d-1)$-dimensional space. Furthermore, our statistical procedure is adaptive to the smoothness s. It means that it does not require the specification of s.

4. Numerical results

We investigate the numerical performances of our fully data-driven estimator $\hat{f}_{\check{h}}$ defined in Section 3.2. We compare $\hat{f}_{\check{h}}$ with the widely used cross-validation estimator and with an "oracle" to be defined later on. We focus on the unit sphere $\mathbb{S}^2$, i.e., the case $d = 3$.

We consider various densities. The first one is the von Mises–Fisher density

$$f_{1,vM} = \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})} e^{\kappa x^\top \mu},$$

with $\kappa = 2$ and $\mu = (1,0,0)^\top$; see Fig. 1. We recall that κ is the concentration parameter and μ the directional mean. Note that the smaller the concentration parameter is, the closer to the uniform density the von Mises–Fisher density is. We also estimate the mixture of two von Mises–Fisher densities, viz.

$$f_{2,vM} = \frac{4}{5} \times \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})} e^{\kappa x^\top \mu} + \frac{1}{5} \times \frac{\kappa'}{2\pi(e^{\kappa'} - e^{-\kappa'})} e^{\kappa' x^\top \mu'},$$

with $\kappa' = 0.7$ and $\mu' = (-1,0,0)^\top$. Note that $f_{1,vM}$ is rotationally symmetric and $f_{2,vM}$ also since μ and μ' are antipodal. Finally, let us consider a non rotationally symmetric density

$$f_{3,vM} = \frac{4}{5} \times \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})} e^{\kappa x^\top \mu} + \frac{1}{5} \times \frac{\kappa'}{2\pi(e^{\kappa'} - e^{-\kappa'})} e^{\kappa' x^\top \mu''},$$

with $\mu'' = (0, 1/\sqrt{2}, 1/\sqrt{2})^\top$.

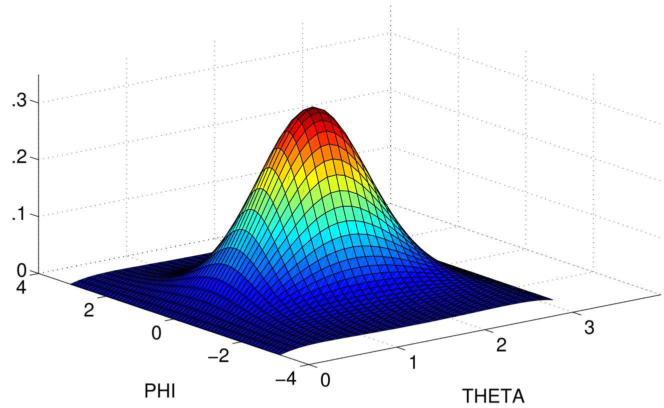


Fig. 1. The density $f_{1,vM}$ in spherical coordinates.

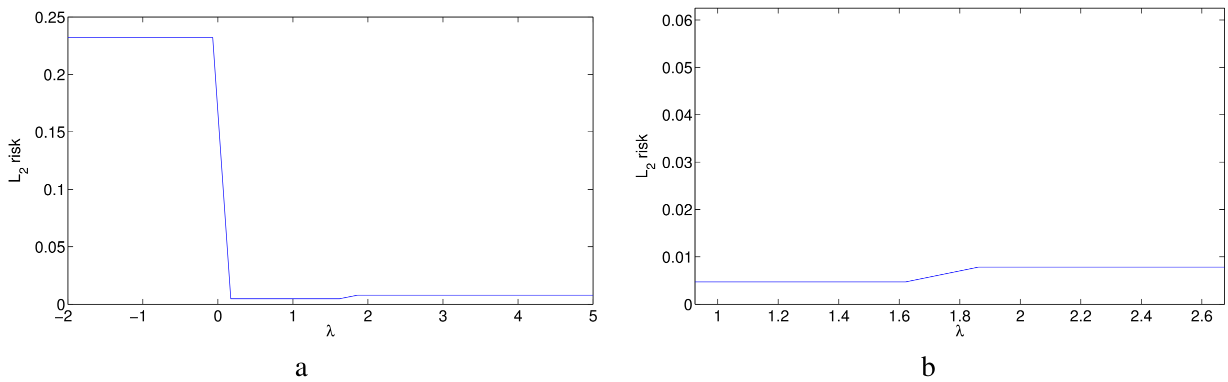


Fig. 2. (a) Empirical L_2-risk of $\hat{f}_{\check{h}}$ to estimate $f_{1,vM}$ in function of λ; (b) A zoom.

Now let us define what the "oracle" $\hat{f}_{h_{\text{oracle}}}$ is. The bandwidth h_{oracle} is defined as

$$h_{\text{oracle}} = \operatorname*{argmin}_{h \in \mathcal{H}} \|\hat{f}_h - f\|^2.$$

This bandwidth can be viewed as the "ideal" one since it uses the specification of the density of interest f which is here $f_{1,vM}, f_{2,vM}$ or $f_{3,vM}$. Hence, the performances of $\hat{f}_{h_{\text{oracle}}}$ are used as a benchmark.

In the sequel we present detailed results for $f_{1,vM}$, namely risk curves and graphic reconstructions and we compute MISE for $f_{1,vM}, f_{2,vM}$ or $f_{3,vM}$. We use the von Mises kernel $K(x) = e^{-x}$.

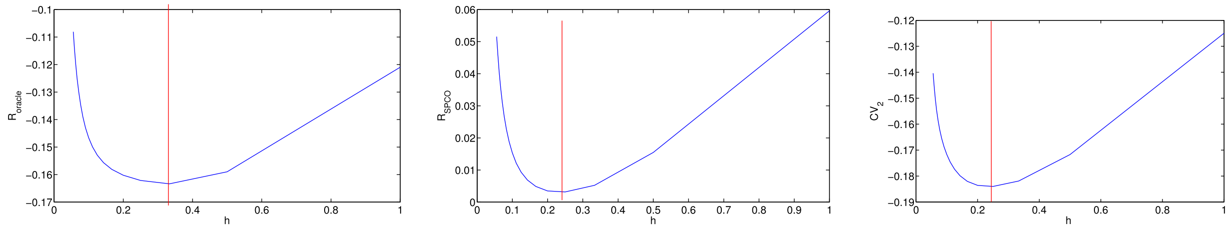
Before presenting the performances of the various procedures, we shall remind that theoretical results of Section 3.1 have shown that setting $\lambda = 1$ in the SPCO algorithm is optimal. We would like to show how simulations actually support this conclusion. Indeed, Fig. 2 displays the empirical L_2 risk of $\hat{f}_{\hat{h}}$ to estimate $f_{1,vM}$ in function of parameter λ for $n = 100$ and 100 Monte Carlo replications. Fig. 2(a) shows a "dimension jump" and that the minimal risk is reached in a stable zone around $\lambda = 1$: negative values of λ lead to an overfitting estimator ($\hat{h}$ is chosen close to $h_{\min}$ as shown in Theorem 2) with poor performances whereas large values of λ make the risk increase again; see a zoom on Fig. 2(b). Next, considering the MISE computations, we will see that $\lambda = 1$ yields quite good results.

In Lemma D of the Appendix, we develop the expression (2) to be minimized to implement our estimator $\hat{f}_{\check{h}}$. We now recall the cross-validation criterion of [9]. Let

$$\hat{f}_{h,i}(x) = \frac{c_0(h)}{n-1} \sum_{j \neq i}^{n} e^{-(1 - x^\top X_j)/h^2}.$$

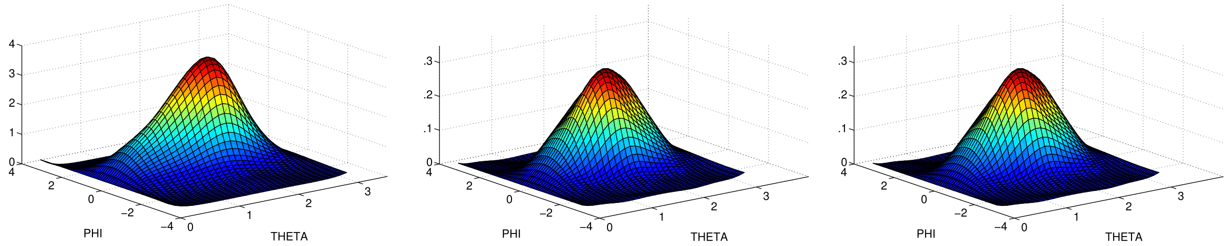
Then

$$CV_2(h) = \|\hat{f}_h\|^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,i}(x).$$



a b c

Fig. 3. Risks curves in function of h for $f_{1,vM}$, $n = 500$: (a) R_{oracle}; (b) R_{SPCO}; (c) CV_2. Vertical red lines represent the bandwidth value h minimizing each curve . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



a b c

Fig. 4. Reconstruction of $f_{1,vM}$, $n = 500$: (a) $\hat{f}_{h_{\text{oracle}}}$, $h_{\text{oracle}} = 0.33$; (b)/SPCO $\hat{f}_{\hat{h}}$, $\hat{h} = 0.25$; (c) cross-validation $\hat{f}_{h_{CV2}}$, $h_{CV2} = 0.25$.

Table 1
MISE over 100 Monte Carlo repetitions to estimate $f_{1,vM}$.

	$n = 50$	$n = 100$	$n = 500$
Oracle	0.0088	0.0064	0.0027
SPCO	0.0160	0.0091	0.0048
Cross-validation	0.0191	0.0099	0.0053

Table 2
MISE over 100 Monte Carlo repetitions to estimate $f_{2,vM}$.

	$n = 50$	$n = 100$	$n = 500$
Oracle	0.0086	0.0051	0.0027
SPCO	0.0122	0.0083	0.0043
Cross-validation	0.0139	0.0096	0.0047

Table 3
MISE over 100 Monte Carlo repetitions to estimate $f_{3,vM}$.

	$n = 50$	$n = 100$	$n = 500$
Oracle	0.0099	0.0075	0.0043
SPCO	0.0153	0.0107	0.0063
Cross-validation	0.0185	0.0122	0.0066

Note that $CV_2(h)+\|f\|^2$ is an unbiased estimate of the MISE of $\hat{f}_h$. The cross-validation procedure to select the bandwidth h consists in minimizing CV_2 with respect to h. We call this selected value h_{CV_2}.

In the rest of this section, SPCO will denote the estimation procedure related to $\hat{f}_{\hat{h}}$. In Fig. 3, for $n = 500$ we plot as a function of h: $R_{\text{oracle}} = \|\hat{f}_h - f_{1,vM}\|^2 - \|f_{1,vM}\|^2$ for the oracle, $R_{SPCO} = \|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 + \text{pen}_{\lambda=1}(h)$ for SPCO and $CV_2(h)$ for cross-validation. We point out on each plot the value of h that minimizes each quantity. In Fig. 4, we plot in spherical coordinates, for $n = 500$, density reconstructions of $f_{1,vM}$ for the oracle, SPCO and cross-validation. Eventually, in Tables 1–3, we compute MISE of the oracle, SPCO and cross-validation to estimate $f_{1,vM}$, $f_{2,vM}$ and $f_{3,vM}$ for $n \in \{50, 100, 500\}$ over 100 Monte Carlo runs.

When analyzing the results, SPCO performs nicely. In particular, inspection of Tables 1–3 shows that SPCO is close to the oracle and is always slightly better than cross-validation for all densities.

5. Proofs of theorems

In the sequel, Ξ denotes an absolute constant which may change from line to line. The following proofs of theorems rely on technical propositions and lemmas which are gathered for sake of clarity in the Appendix. More specifically, proofs of Theorems 1 and 2 use Lemma A, Propositions A and B, and proof of Theorem 3 uses Proposition C.

Proof of Theorem 1. We set $\tau = \lambda - 1$. Let $\varepsilon \in (0, 1)$ and $\theta \in (0, 1)$ depending on ε to be specified later. Developing the expression of $\mathrm{pen}_\lambda(h)$ given in (3), we have

$$\begin{aligned}\mathrm{pen}_\lambda(h) &= \lambda c_0^2(h)c_2(h)/n - c_0^2(h_{\min})c_2(h_{\min})/n - c_0^2(h)c_2(h)/n + 2\langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n\\ &= \tau c_0^2(h)c_2(h)/n - c_0^2(h_{\min})c_2(h_{\min})/n + 2\langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n.\end{aligned}$$

Using Proposition B and the expression of $\mathrm{pen}_\lambda(h)$ given above, we obtain with probability greater than $1 - \Xi\,|\mathcal{H}|\exp(-x)$, for any $h \in \mathcal{H}$,

$$\begin{aligned}&(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau c_0^2(\hat{h})c_2(\hat{h})/n \le (1+\theta)\|\hat{f}_h - f\|^2 + \tau c_0^2(h)c_2(h)/n + C_2/\theta\,\|f_{h_{\min}} - f\|^2\\ &\quad + C(K)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}/\theta,\end{aligned}\tag{6}$$

with C_2 an absolute constant and $C(K)$ a constant only depending on K. We first consider the case $\tau \ge 0$. Using (A.5) of Proposition A, with probability $1 - \Xi\,|\mathcal{H}|e^{-x}$ one has

$$\tau c_0^2(h)c_2(h)/n \le \tau(1+\theta)\|f - \hat{f}_h\|^2 + \tau C'(K)\|f\|_\infty x^2/(\theta^3 n),$$

where $C'(K)$ is a constant only depending on the kernel K. As $\tau c_0^2(\hat{h})c_2(\hat{h})/n \ge 0$, thus (6) becomes

$$\begin{aligned}&(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 \le \{1+\theta+\tau(1+\theta)\}\|\hat{f}_h - f\|^2 + C_2\|f_{h_{\min}} - f\|^2/\theta\\ &\quad + C(K)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}/\theta + \tau C'(K)\|f\|_\infty x^2/(\theta^3 n).\end{aligned}$$

With $\theta = \varepsilon/(\varepsilon + 2 + 2\tau)$, we obtain

$$\|\hat{f}_{\hat{h}} - f\|^2 \le (1+\tau+\varepsilon)\|\hat{f}_h - f\|^2 + \frac{C_2(\varepsilon+2+2\tau)^2}{(2+2\tau)\varepsilon}\|f_{h_{\min}} - f\|^2 + C''(K,\varepsilon,\tau)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\},$$

with $C''(K, \varepsilon, \tau)$ a constant depending only on K, ε, τ.

Let us now study the case $\tau \in (-1, 0]$. Using (A.5) of Proposition A with $h = \hat{h}$, we have with probability $1 - \Xi\,|\mathcal{H}|e^{-x}$,

$$\tau c_0^2(\hat{h})c_2(\hat{h})/n \ge \tau(1+\theta)\|f - \hat{f}_{\hat{h}}\|^2 + \tau C'(K)\|f\|_\infty x^2/(\theta^3 n).$$

Consequently, as $\tau c_0^2(h)c_2(h)/n \le 0$, (6) becomes

$$\begin{aligned}&\{1-\theta+\tau(1+\theta)\}\|\hat{f}_{\hat{h}} - f\|^2 \le (1+\theta)\|\hat{f}_h - f\|^2 + C_2\|f_{h_{\min}} - f\|^2/\theta\\ &\quad + C(K)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}/\theta - \tau C'(K)\|f\|_\infty x^2/(\theta^3 n).\end{aligned}$$

With $\theta = \varepsilon(\tau+1)^2/\{2+\varepsilon(1-\tau^2)\} < 1$, we obtain with probability $1 - \Xi\,|\mathcal{H}|e^{-x}$,

$$\|\hat{f}_{\hat{h}} - f\|^2 \le \left(\frac{1}{1+\tau} + \varepsilon\right)\|\hat{f}_h - f\|^2 + C''(\varepsilon,\tau)\|f_{h_{\min}} - f\|^2 + C'''(K,\varepsilon,\tau)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\},$$

with $C''(\varepsilon, \tau)$ a constant depending on ε and τ and $C'''(K, \varepsilon, \tau)$ a constant depending on K, ε and τ. This completes the proof of Theorem 1. □

Proof of Theorem 2. We still set $\tau = \lambda - 1$. We set $\theta \in (0, 1)$ such that $\theta < -(1+\tau)/5$. We consider inequality (6) written with $h = h_{\min}$. One obtains

$$\begin{aligned}&(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau c_0^2(\hat{h})c_2(\hat{h})/n \le (1+\theta)\|\hat{f}_{h_{\min}} - f\|^2 + \tau c_0^2(h_{\min})c_2(h_{\min})/n\\ &\quad + C_2\|f_{h_{\min}} - f\|^2/\theta + C(K)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}/\theta.\end{aligned}$$

Now consider Eq. (A.4) with $h = h_{\min}$, one gets

$$\|f - \hat{f}_{h_{\min}}\| \le (1+\theta)\{\|f - f_{h_{\min}}\|^2 + c_0^2(h_{\min})c_2(h_{\min})/n\} + C'(K)\|f\|_\infty x^2/(\theta^3 n).$$

Combining the two inequalities above, we have

$$\begin{aligned}&(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau c_0^2(\hat{h})c_2(\hat{h})/n \le \{(1+\theta)^2 + C_2/\theta\}\|f - f_{h_{\min}}\|^2 + \{\tau + (1+\theta)^2\}c_0^2(h_{\min})c_2(h_{\min})/n\\ &\quad + C(K)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}/\theta + (1+\theta)C'(K)\|f\|_\infty x^2/(\theta^3 n).\end{aligned}$$

Now let us define $u_n = \|f_{h_{\min}} - f\|^2/\{c_0^2(h_{\min})c_2(h_{\min})/n\}$. We have by Assumption (5) that $u_n \to 0$ when $n \to \infty$. Then we get

$$(1-\theta)\|\hat{f}_{\hat{h}} - f\|^2 + \tau c_0^2(\hat{h})c_2(\hat{h})/n \le [\{(1+\theta)^2 + C_2/\theta\}u_n + \tau + (1+\theta)^2]c_0^2(h_{\min})c_2(h_{\min})/n$$
$$+C(K,\theta)\{\|f\|_\infty x^2/n + x^3 c_0(h_{\min})/n^2\}. \quad (7)$$

Now we consider using Eq. (A.5) from Proposition A with $h = \hat{h}$ and $\eta = 1$, we get

$$c_0(\hat{h})c_2(\hat{h})/n \le 2\|f - \hat{f}_{\hat{h}}\| + C'(K)\|f\|_\infty x^2/n.$$

Then

$$\|f - \hat{f}_{\hat{h}}\| \ge c_0(\hat{h})c_2(\hat{h})/(2n) - C'(K)\|f\|_\infty x^2/n$$

and hence (7) becomes

$$\{(1-\theta)/2 + \tau\}c_0^2(\hat{h})c_2(\hat{h})/n \le [\{(1+\theta)^2 + C_2/\theta\}u_n + \tau + (1+\theta)^2]c_0^2(h_{\min})c_2(h_{\min})/n$$
$$+C'(K,\theta)\{\|f\|_\infty x^2/n + x^3 c_0(h_{\min})/n^2\}.$$

However, we assumed that $u_n = o(1)$. Thus for n large enough, $\{(1+\theta)^2 + C_2/\theta\}u_n \le \theta$. We are now going to bound the remainder terms $C'(K,\theta)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}$. We have

$$\begin{aligned}C'(K,\theta)\{\|f\|_\infty x^2/n + c_0(h_{\min})x^3/n^2\}&\frac{n}{c_0^2(h_{\min})c_2(h_{\min})}\\ &= C''(K,\theta,\|f\|_\infty)\left\{\frac{x^2}{c_0^2(h_{\min})c_2(h_{\min})} + \frac{x^3}{nc_0(h_{\min})c_2(h_{\min})}\right\}\\ &\le C''(K,\theta,\|f\|_\infty)(x^2 h_{\min}^{d-1} + x^3/n),\end{aligned}$$

for n large enough using (A.1) and (A.2) from Lemma A. But $h_{\min}^{d-1} \le (\ln n)^\beta/n$ and setting $x = (n/\ln n)^{1/3}$, we get

$$C''(K,\theta,\|f\|_\infty)(x^2 h_{\min}^{d-1} + x^3/n) \le C''(K,\theta,\|f\|_\infty)\{(\ln n)^{\beta-2/3}/n^{1/3} + 1/\ln n\} = o(1) \le \theta,$$

for n large enough. Consequently there exists N such that for $n \ge N$, with probability larger than $1 - \Xi\,|\mathcal{H}|e^{-(n/\ln n)^{1/3}}$,

$$\{(1-\theta)/2 + \tau\}c_0^2(\hat{h})c_2(\hat{h})/n \le \{\theta + \tau + (1+\theta)^2 + \theta\}c_0^2(h_{\min})c_2(h_{\min})/n \le (1 + \tau + 5\theta)c_0^2(h_{\min})c_2(h_{\min})/n.$$

Using (A.3) of Lemma A, we have, for n large enough,

$$0.9\,h^{1-d}R(K) \le c_0^2(h)c_2(h) \le 1.1\,h^{1-d}R(K).$$

Thus we finally get, for n large enough,

$$0.9\{(1-\theta)/2 + \tau\}\hat{h}^{1-d} \le 1.1(1 + \tau + 5\theta)h_{\min}^{1-d} \quad \Leftrightarrow \quad \{(1-\theta)/2 + \tau\}\hat{h}^{1-d} \le 1.23(1 + \tau + 5\theta)h_{\min}^{1-d}.$$

But $(1-\theta)/2 + \tau < 1 + \tau < 0$, and because we have chosen θ such that $1 + \tau + 5\theta < 0$ (for instance $\theta = -(\tau+1)/10$), one gets

$$\hat{h} \le \left\{\frac{1.23(1+\tau+5\theta)}{(1-\theta)/2 + \tau}\right\}^{1/(d-1)} h_{\min}.$$

With $\theta = -(\tau+1)/10$, the inequality above becomes for n large enough

$$\hat{h} \le \{1.23\,(2.1 - 1/\lambda)\}^{1/(d-1)}\,h_{\min},$$

which completes the proof of Theorem 2. □

Proof of Theorem 3. Let $f \in \tilde{\mathbf{F}}_2(s,B)$ and $\mathcal{E}$ be the event corresponding to the intersection of events in Theorem 1 and Proposition A. Let $\mathcal{E}^\complement$ denote the complementary of $\mathcal{E}$. For any $A > 0$, by taking x proportional to $\ln n$, $\Pr(\mathcal{E}) \ge 1 - n^{-A}$. We have

$$\mathrm{E}(\|\hat{f}_{\tilde{h}} - f\|^2) \le \mathrm{E}(\|\hat{f}_{\tilde{h}} - f\|^2 1_{\mathcal{E}}) + \mathrm{E}(\|\hat{f}_{\tilde{h}} - f\|^2 1_{\mathcal{E}^\complement}).$$

Let us deal with the second term of the right-hand side. We have $\|\hat{f}_h - f\|^2 \le 2(\|\hat{f}_h\|^2 + \|f\|^2)$. However,

$$\begin{aligned}\|\hat{f}_h\|^2 &= \frac{c_0^2(h)}{n^2}\sum_{i,j}\int_{\mathbb{S}^{d-1}} K_{h^2}(x,X_i)K_{h^2}(x,X_j)\omega_d(dx)\\ &\le \frac{c_0(h)}{n^2}\,\|K\|_\infty \sum_{i,j} c_0(h)\int_{\mathbb{S}^{d-1}} K_{h^2}(x,X_j)\omega_d(dx) \le c_0(h)\|K\|_\infty \le 2n,\end{aligned}$$

since $c_0(h) \le 2n/\|K\|_\infty$, using (A.1) and (1) for n large enough. Thus $\|\hat{f}_h - f\|^2 \le 2n + 2\|f\|^2$, which gives the result on $\mathcal{E}^\complement$. Now on $\mathcal{E}$, for $n \ge n_0$ (n_0 not depending on f) Propositions A and C yield that

$$\begin{aligned}\min_{h\in\mathcal{H}} \|\hat{f}_h - f\|^2 &\le (1+\eta)\min_{h\in\mathcal{H}}\{\|f - f_h\|^2 + c_0^2(h)c_2(h)/n\} + \Xi\,\Upsilon(\ln n)^2/(\eta^3 n)\\ &\le C\min_{h\in\mathcal{H}}(h^{2s} + h^{1-d}/n) + \Xi\,\Upsilon(\ln n)^2/(\eta^3 n).\end{aligned}$$

Minimizing in h the right-hand side of the last inequality gives the result on $\mathcal{E}$. $\square$

Acknowledgments

The author thanks the Editor-in-Chief, Christian Genest, an Associate Editor and two anonymous referees for their constructive comments, which helped to improve the manuscript. The author would also like to thank Claire Lacour and Vincent Rivoirard for interesting suggestions and remarks.

Appendix

This Appendix gathers technical results needed to prove the theorems. We shall start with a lemma that collects some standard properties about constants c_0 and c_2.

Lemma A. *We have, as $h \to 0$,*

$$c_0^{-1}(h) = R_0(K)h^{d-1} + o(1), \tag{A.1}$$

where $R_0(K) = 2^{(d-3)/2}\sigma_{d-2}\alpha_0(K)$. We also have, as $h \to 0$,

$$c_2(h) = R_1(K)h^{d-1} + o(1), \tag{A.2}$$

with $R_1(K) = 2^{(d-3)/2}\sigma_{d-2}\int_0^\infty x^{(d-3)/2}K^2(x)dx$. Eventually we have, when $h \to 0$,

$$c_0^2(h)c_2(h) = h^{1-d}R(K) + o(1), \tag{A.3}$$

with $R(K) = R_1(K)/R_0^2(K)$.

The proof of Lemma A can be found in the proof of Proposition 4.1 of [1].

Propositions A and B are essential to prove Theorems 1 and 2. Let us start with Proposition A, which is a counterpart of Proposition 4.1 of [17] for $\mathbb{S}^{d-1}$.

Proposition A. *Assume that the kernel K satisfies Assumption 1. Let $\Upsilon \ge (1 + 2\|f\|_\infty) \vee 8\pi\|K\|_\infty R_0(K)/R_1(K)$. There exists n_0 such that, for $n \ge n_0$ (n_0 not depending on f), all $x \ge 1$ and for all $\eta \in (0,1)$ with probability larger than $1 - \Xi\,|\mathcal{H}|e^{-x}$, for all $h \in \mathcal{H}$ each of the following inequalities holds:*

$$\|f - \hat{f}_h\| \le (1+\eta)\{\|f - f_h\|^2 + c_0^2(h)c_2(h)/n\} + \Xi\,\Upsilon x^2/(\eta^3 n) \tag{A.4}$$

$$\|f - f_h\|^2 + c_0^2(h)c_2(h)/n \le (1+\eta)\|f - \hat{f}_h\|^2 + \Xi\,\Upsilon x^2/(\eta^3 n). \tag{A.5}$$

Proof of Proposition A. To prove Proposition A, we need to verify Assumptions (11)–(16) of [17]. We recall that

$$f_h = \mathrm{E}(\hat{f}_h) = c_0(h)\int_{y\in\mathbb{S}^{d-1}} f(y)K_{h^2}(x,y)\omega_d(dy).$$

Let us check Assumption (11) of [17]. This one amounts to prove that, for some Γ and Υ,

$$\Gamma(1 + \|f\|_\infty) \vee \sup_{h\in\mathcal{H}}\|f_h\|^2 \le \Upsilon.$$

We have

$$\begin{aligned}\|f_h\|^2 &\le \|f\|_\infty \int_x \underbrace{\left\{\int_y c_0(h)K_{h^2}(x,y)\omega_d(dy)\right\}}_{=1}\left\{\int_y f(y)c_0(h)K_{h^2}(x,y)\omega_d(dy)\right\}\omega_d(dx)\\ &\le \|f\|_\infty \int_y f(y)\int_x c_0(h)K_{h^2}(x,y)\omega_d(dx)\omega_d(dy) \le \|f\|_\infty.\end{aligned}$$

Therefore, Assumption (11) in [17] holds with $\Gamma = 1$ and $\Upsilon \ge 1 + \|f\|_\infty$.

Let us check Assumption (12) of [17]. We have to prove that

$$\int c_0^2(h)K_{h^2}^2(x,x)\omega_d(dx) \le \Upsilon n \iint c_0^2(h)K_{h^2}^2(x,y)\omega_d(dx)f(y)\omega_d(dy).$$

Given that

$$\iint c_0^2(h)K_{h^2}^2(x,y)\omega_d(dx)f(y)\omega_d(dy) = c_0^2(h)c_2(h)\int f(y)\omega_d(dy) = c_0^2(h)c_2(h),$$

and

$$\int c_0^2(h)K_{h^2}^2(x,x)\omega_d(dx) = 4\pi c_0^2(h)K^2(0),$$

Assumption (12) amounts to check that

$$4\pi c_0^2(h)K^2(0) \le \Upsilon n c_0^2(h)c_2(h) \quad \Leftrightarrow \quad \Upsilon \ge 4\pi K^2(0)/\{nc_2(h)\}.$$

But using (A.2), we have

$$c_2(h) = R_1(K)h^{d-1} + o(1),$$

when $h \to 0$ uniformly in h. Thus there exist n_1, n_1 independent of f such that, for $n \ge n_1$, $c_2(h) \ge R_1(K)h^{d-1}/2$. Now using that $h^{d-1} \ge \|K\|_\infty/\{R_0(K)n\}$ and $K(0) \le \|K\|_\infty$, it is sufficient to have $\Upsilon \ge 8\pi\|K\|_\infty R_0(K)/R_1(K)$ to ensure Assumption (12) in [17].

Assumption (13) in [17] consists to prove that

$$\|f_h - f_{h'}\|_\infty \le \Upsilon \vee \sqrt{\Upsilon n}\|f_h - f_{h'}\|_2.$$

For any $h \in \mathcal{H}$ and any $x \in \mathbb{S}^{d-1}$, we have $\|f_h\|_\infty \le \|f\|_\infty$. Therefore, Assumption (13) in [17] holds for $\Upsilon \ge 2\|f\|_\infty$.

Assumptions (14) and (15) of [17] consist in proving respectively that

$$\mathrm{E}\left\{c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,Y)\omega_d(dz)\right\}^2 \le \Upsilon c_0^2(h)c_2(h)$$

and

$$\sup_{x\in\mathbb{S}^{d-1}} \mathrm{E}\left\{c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,x)\omega_d(dz)\right\}^2 \le \Upsilon n.$$

We have

$$c_0^2(h)\int K_{h^2}(x,z)K_{h^2}(z,y)\omega_d(dz) \le c_0^2(h)c_2(h) \wedge c_0(h)\|K\|_\infty.$$

Indeed if $y = z$, then $\int K_{h^2}(x,z)K_{h^2}(z,y)\omega_d(dz) = \int K_{h^2}^2(x,z)\omega_d(dz) = c_2(h)$. Otherwise,

$$c_0^2(h)\int K_{h^2}(x,z)K_{h^2}(z,y)\omega_d(dz) \le c_0(h)\|K\|_\infty \underbrace{c_0(h)\int K_{h^2}(x,z)\omega_d(dz)}_{=1} = c_0(h)\|K\|_\infty.$$

Furthermore, (A.1) entails that there exists n_2 independent of f such that, for $n \ge n_2$, $c_0^{-1}(h) \ge R_0(K)h^{d-1}/2$ and consequently $c_0(h) \le 2n/\|K\|_\infty$, using (1). Thus, for $n \ge n_2$,

$$c_0^2(h)\int K_{h^2}(x,z)K_{h^2}(z,y)\omega_d(dz) \le c_0^2(h)c_2(h) \wedge 2n. \tag{A.6}$$

We have

$$\begin{aligned}
\mathrm{E}\left\{c_0^2(h)\int_z K_{h^2}(X,z)K_{h^2}(z,x)\omega_d(dz)\right\} &= c_0^2(h)\int_z\left\{\int_y K_{h^2}(y,z)f(y)\omega_d(dy)\right\}K_{h^2}(z,x)\omega_d(dz)\\
&\le \|f\|_\infty c_0(h)\int_z c_0(h)\int_y K_{h^2}(y,z)\omega_d(dy)K_{h^2}(z,x)\omega_d(dz)\\
&\le \|f\|_\infty.
\end{aligned} \tag{A.7}$$

Therefore, for $n \ge n_2$,

$$\begin{aligned}
\sup_{x\in\mathbb{S}^{d-1}} \mathrm{E}\left[c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,x)\omega_d(dz)\right]^2 &\le \sup_{(x,y)}\left\{c_0^2(h)\int K_{h^2}(x,z)K_{h^2}(z,y)\omega_d(dz)\right\}\\
&\quad\times \sup_x \mathrm{E}\left\{c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,x)\omega_d(dz)\right\}\\
&\le \{c_0^2(h)c_2(h) \wedge 2n\}\|f\|_\infty,
\end{aligned}$$

using (A.6) and (A.7). Moreover, we have

$$\mathrm{E}\left\{c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,Y)\omega_d(dz)\right\}^2 \le \sup_x \mathrm{E}\left\{c_0^2(h)\int K_{h^2}(X,z)K_{h^2}(z,x)\omega_d(dz)\right\}^2$$
$$\le \{c_0^2(h)c_2(h)\wedge 2n\}\|f\|_\infty$$

using (A.6) and (A.7). Hence Assumptions (14) and (15) in [17] hold for $\Upsilon \ge 2\|f\|_\infty$.

Let $t\in\mathbb{B}_{c_0(h)K_{h^2}}$ be the set of functions t such that $t(x)=\int a(z)c_0(h)K_{h^2}(z,x)\omega_d(dz)$ for some $a\in L_2(\mathbb{S}^{d-1})$ with $\|a\|\le 1$. Now let $a\in L_2(\mathbb{S}^{d-1})$ be such that $\|a\|=1$ and $t(y)=\int a(x)c_0(h)K_{h^2}(x,y)\omega_d(dx)$ for all $y\in\mathbb{S}^{d-1}$. To verify Assumption (16) in [17] we have to prove that

$$\sup_{t\in\mathbb{B}_{c_0(h)K_{h^2}}}\int t(x)f(x)\omega_d(dx)\le \Upsilon\vee\sqrt{\Upsilon c_0^2(h)c_2(h)}.$$

Using the Cauchy–Schwarz inequality, one gets

$$t(y)\le\sqrt{\int_{\mathbb{S}^{d-1}}a^2(x)\omega_d(dx)}\sqrt{c_0^2(h)\int_{\mathbb{S}^{d-1}}K_{h^2}^2(x,y)\omega_d(dx)}\le\sqrt{c_0^2(h)c_2(h)}.$$

Thus for any $t\in\mathbb{B}_{c_0(h)K_{h^2}}$

$$\int t^2(x)f(x)\omega_d(dx)\le\|t\|_\infty\langle|t|,f\rangle\le\sqrt{c_0^2(h)c_2(h)}\,\|f\|\times\|t\|,$$

but using the Cauchy–Schwarz inequality and Fubini, one gets

$$\begin{aligned}\|t\| &= \int_x\left\{\int_y a(y)c_0(h)K_{h^2}(x,y)\omega_d(dy)\right\}^2\omega_d(dx)\\ &\le \int_x\left\{\int_y a^2(y)c_0(h)K_{h^2}(x,y)\omega_d(dy)\right\}\left\{\int_y c_0(h)K_{h^2}(x,y)\omega_d(dy)\right\}\omega_d(dx)\\ &= \int_x\int_y a^2(y)c_0(h)K_{h^2}(x,y)\omega_d(dy)\omega_d(dx)\\ &= \int_y a^2(y)\left\{\int_x c_0(h)K_{h^2}(x,y)\omega_d(dx)\right\}\omega_d(dy)=1.\end{aligned}$$

Furthermore,

$$\int t^2(x)f(x)\omega_d(dx)\le\|f\|\sqrt{c_0^2(h)c_2(h)}\le\sqrt{\|f\|_\infty}\sqrt{c_0^2(h)c_2(h)}\le\sqrt{\Upsilon c_0^2(h)c_2(h)},$$

and hence Assumption (16) in [17] is verified.

Finally, Assumptions (11)–(16) from [17] hold in the spherical setting, for $n\ge n_0=\max(n_1,n_2)$ and if $\varGamma=1$ and

$$\Upsilon\ge(1+2\|f\|_\infty)\vee 8\pi\|K\|_\infty R_0(K)/R_1(K).$$

This enables us to use Proposition 4.1 of [17] which gives Proposition A. □

The next proposition gives a general result on the estimator $\hat f_{\hat h}$.

Proposition B. *Assume that the kernel K satisfies Assumption 1 and $\|f\|_\infty<\infty$. Let $x\ge 1$ and $\theta\in(0,1)$. With probability larger than $1-C_1|\mathcal{H}|e^{-x}$, with C_1 an absolute constant, for any $h\in\mathcal{H}$,*

$$\begin{aligned}(1-\theta)\|\hat f_{\hat h}-f\|^2 \le (1+\theta)\|\hat f_h-f\|^2 &+ \{pen_\lambda(h)-2\langle c_0(h)K_{h^2},c_0(h_{\min})K_{h^2_{\min}}\rangle/n\}\\ &-\{pen_\lambda(\hat h)-2\langle c_0(\hat h)K_{\hat h^2},c_0(h_{\min})K_{h^2_{\min}}\rangle/n\}\\ &+C_2\|f_{h_{\min}}-f\|^2/\theta+C(K)\{\|f\|_\infty x^2/n+x^3c_0(h_{\min})/n^2\}/\theta,\end{aligned}$$

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

In order to avoid any confusion, we recall that $K_{h^2}=K_{h^2}(\cdot,\cdot)$ and

$$\langle c_0(h)K_{h^2},c_0(h_{\min})K_{h^2_{\min}}\rangle=\int_{\mathbb{S}^{d-1}}c_0(h)K_{h^2}(x,y)c_0(h_{\min})K_{h^2_{\min}}(x,y)\omega_d(dy).$$

Once again, we would like to draw the attention to the fact that the quantity

$$\int_{\mathbb{S}^{d-1}} c_0(h)K_{h^2}(x,y)c_0(h_{\min})K_{h^2_{\min}}(x,y)\omega_d(dy)$$

does not depend on x. Indeed, we have, using Remark 1

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} K_{h^2}(x,y)K_{h^2_{\min}}(x,y)\omega_d(dy) &= \int_{\mathbb{S}^{d-1}} K\left(\frac{1-x^\top y}{h^2}\right)K\left(\frac{1-x^\top y}{h^2_{\min}}\right)\omega_d(dy)\\ &= \sigma_{d-2}\int_{-1}^{1} K\left(\frac{1-t}{h^2}\right)K\left(\frac{1-t}{h^2_{\min}}\right)(1-t^2)^{(d-3)/2}dt.\end{aligned}$$

Proof of Proposition B. The proof follows the proof of Theorem 9 in [14] adapted to $\mathbb{S}^{d-1}$. Let $\theta' \in (0,1)$ be fixed and chosen later. Using the definition of $\hat{h}$, we can write, for any $h \in \mathcal{H}$

$$\begin{aligned}\|\hat{f}_{\hat{h}}-f\|^2 + pen_\lambda(\hat{h}) &= \|\hat{f}_{\hat{h}}-\hat{f}_{h_{\min}}\|^2 + \mathrm{pen}_\lambda(\hat{h}) + \|\hat{f}_{h_{\min}}-f\|^2 + 2\langle \hat{f}_{\hat{h}}-\hat{f}_{h_{\min}}, \hat{f}_{h_{\min}}-f\rangle\\ &\le \|\hat{f}_{h}-\hat{f}_{h_{\min}}\|^2 + pen_\lambda(h) + \|\hat{f}_{h_{\min}}-f\|^2 + 2\langle \hat{f}_{\hat{h}}-\hat{f}_{h_{\min}}, \hat{f}_{h_{\min}}-f\rangle\\ &\le \|\hat{f}_{h}-f\|^2 + 2\|f-\hat{f}_{h_{\min}}\|^2 + 2\langle \hat{f}_{h}-f, f-\hat{f}_{h_{\min}}\rangle\\ &+ pen_\lambda(h) + 2\langle \hat{f}_{\hat{h}}-\hat{f}_{h_{\min}}, \hat{f}_{h_{\min}}-f\rangle.\end{aligned}$$

Consequently,

$$\|\hat{f}_{\hat{h}}-f\|^2 \le \|\hat{f}_{h}-f\|^2 + \{pen_\lambda(h) - 2\langle \hat{f}_{h}-f, \hat{f}_{h_{\min}}-f\rangle\} - \{pen_\lambda(\hat{h}) - 2\langle \hat{f}_{\hat{h}}-f, \hat{f}_{h_{\min}}-f\rangle\}. \tag{A.8}$$

Then for a given h, we study the term $2\langle \hat{f}_h - f, \hat{f}_{h_{\min}}-f\rangle$. Let us introduce the degenerate U-statistic

$$U(h,h_{\min}) = \sum_{i\neq j}\left\langle c_0(h)K_{h^2}(.,X_i)-f_h,\, c_0(h_{\min})K_{h_{\min}}(.,X_j)-f_{h_{\min}}\right\rangle$$

and the centered variable $V(h,h') = \langle \hat{f}_h - f_h, f_{h'}-f\rangle$. We first center the terms

$$\langle \hat{f}_h - f, \hat{f}_{h_{\min}}-f\rangle = \langle \hat{f}_h - f_h, \hat{f}_{h_{\min}}-f_{h_{\min}}\rangle + V(h,h_{\min}) + V(h_{\min},h) + \langle f_h - f, f_{h_{\min}}-f\rangle.$$

Now

$$\begin{aligned}\langle \hat{f}_h - f_h, \hat{f}_{h_{\min}}-f_{h_{\min}}\rangle &= \frac{1}{n^2}\sum_{i,j=1}^{n}\langle c_0(h)K_{h^2}(.,X_i)-f_h,\, c_0(h_{\min})K_{h_{\min}}(.,X_j)-f_{h_{\min}}\rangle\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\langle c_0(h)K_{h^2}(.,X_i)-f_h,\, c_0(h_{\min})K_{h_{\min}}(.,X_i)-f_{h_{\min}}\rangle + U(h,h_{\min})/n^2.\end{aligned}$$

Then

$$\langle \hat{f}_h - f_h, \hat{f}_{h_{\min}}-f_{h_{\min}}\rangle = \langle c_0(h)K_h^2, c_0(h_{\min})K_{h^2_{\min}}\rangle/n - \langle \hat{f}_h, f_{h_{\min}}\rangle/n - \langle f_h, \hat{f}_{h_{\min}}\rangle/n + \langle f_h, f_{h_{\min}}\rangle/n + U(h,h_{\min})/n^2.$$

Finally, we obtain

$$\begin{aligned}\langle \hat{f}_h - f, \hat{f}_{h_{\min}}-f\rangle &= \langle c_0(h)K_h^2, c_0(h_{\min})K_{h^2_{\min}}\rangle/n + U(h,h_{\min})/n^2 &\text{(A.9)}\\ &- \langle \hat{f}_h, f_{h_{\min}}\rangle/n - \langle f_h, \hat{f}_{h_{\min}}\rangle/n + \langle f_h, f_{h_{\min}}\rangle/n &\text{(A.10)}\\ &+ V(h,h_{\min}) + V(h_{\min},h) + \langle f_h - f, f_{h_{\min}}-f\rangle. &\text{(A.11)}\end{aligned}$$

We first control the last term of (A.9) involving a U-statistic. This is done in the next lemma.

Lemma B. *With probability greater than $1-5.54|\mathcal{H}|e^{-x}$, for any h in $\mathcal{H}$,*

$$\left|U(h,h_{\min})/n^2\right| \le \theta' c_0^2(h)c_2(h)/n + \Xi\,\|f\|_\infty x^2/(\theta' n) + \Xi\, c_0(h_{\min})\|K\|_\infty x^3/(\theta' n^2).$$

Proof of Lemma B. We have

$$\begin{aligned}U(h,h_{\min}) &= \sum_{i\neq j}\langle c_0(h)K_h^2(.,X_i)-f_h,\, c_0(h_{\min})K_{h^2_{\min}}(.,X_j)-f_{h_{\min}}\rangle\\ &= \sum_{i=2}^{n}\sum_{j<i} G_{h,h_{\min}}(X_i,X_j) + G_{h_{\min},h}(X_i,X_j),\end{aligned}$$

where

$$G_{h,h'}(s,t) = \langle c_0(h)K_{h^2}(.,s) - f_h, c_0(h')K_{h'^2}(.,t) - f_{h'} \rangle.$$

We apply Theorem 3.4 of [10]

$$\Pr\{|U(h,h_{\min})| \geq \Xi\,(C\sqrt{x} + Dx + Bx^{3/2} + Ax^2)\} \leq 5.54e^{-x},$$

with A, B, C and D defined subsequently. First, we have

$$\begin{aligned}\|f_{h_{\min}}\|_\infty = \|\mathrm{E}(\hat{f}_{h_{\min}})\|_\infty &= \|c_0(h_{\min})\int_{\mathbb{S}^{d-1}} K_{h^2_{\min}}(x,y)f(y)\omega_d(dy)\|_\infty \\ &\leq c_0(h_{\min})\|K\|_\infty \int_{\mathbb{S}^{d-1}} f(y)\omega_d(dy) \leq c_0(h_{\min})\|K\|_\infty.\end{aligned}$$

We have

$$A = \|G_{h,h_{\min}} + G_{h_{\min},h}\|_\infty \leq \|G_{h,h_{\min}}\|_\infty + \|G_{h_{\min},h}\|_\infty = 2\|G_{h,h_{\min}}\|_\infty,$$

because $G_{h_{\min},h} = G_{h,h_{\min}}$. We have

$$\begin{aligned}\|G_{h,h_{\min}}\|_\infty &= \sup_{s,t}\left|\int_{\mathbb{S}^{d-1}} \left\{c_0(h)K_{h^2}(u,s) - f_h(u)\right\}\{c_0(h_{\min})K_{h^2_{\min}}(u,t) - f_{h_{\min}}(u)\}\omega_d(du)\right| \\ &\leq \sup_{u,t}|c_0(h_{\min})K_{h^2_{\min}}(u,t) - f_{h_{\min}}(u)| \sup_s \int |c_0(h)K_{h^2}(u,s) - f_h(u)|\omega_d(du) \\ &\leq \left\{c_0(h_{\min})\|K\|_\infty + \|f_{h_{\min}}\|_\infty\right\}\left\{\sup_s c_0(h)\int K_{h^2}(u,s)\omega_d(du)\right. \\ &\qquad \left. + c_0(h)\iint K_{h^2}(u,y)f(y)\omega_d(dy)\omega_d(du)\right\} \\ &\leq 2c_0(h_{\min})\|K\|_\infty\left\{1 + \int f(y)c_0(h)\int K_{h^2}(u,y)\omega_d(du)\omega_d(dy)\right\} \\ &\leq 4c_0(h_{\min})\|K\|_\infty.\end{aligned}$$

Consequently we have that $A \leq 8c_0(h_{\min})\|K\|_\infty$ and $Ax^2/n^2 \leq 8x^2c_0(h_{\min})\|K\|_\infty/n^2$. We define

$$B^2 = (n-1)\sup_t \mathrm{E}[\{G_{h,h_{\min}}(t,X_2) + G_{h_{\min},h}(t,X_2)\}^2].$$

For any t, we have

$$\begin{aligned}\mathrm{E}[G^2_{h,h_{\min}}(t,X_2)] &= \mathrm{E}\left[\left[\int \{c_0(h)K_{h^2}(u,t) - f_h(u)\}[c_0(h_{\min})K_{h^2_{\min}}(u,X_2) - \mathrm{E}\{c_0(h_{\min})K_{h^2_{\min}}(u,X_2)\}]\omega_d(du)\right]^2\right] \\ &\leq \mathrm{E}\left[\int \{c_0(h)K_{h^2}(u,t) - f_h(u)\}^2\omega_d(du)\int [c_0(h_{\min})K_{h^2_{\min}}(u,X_2) - \mathrm{E}\{c_0(h_{\min})K_{h^2_{\min}}(u,X_2)\}]^2\omega_d(du)\right] \\ &\leq 2\left\{\int c_0^2(h)K^2_{h^2}(u,t)\omega_d(du) + \int f_h^2(u)\omega_d(du)\right\}\times \\ &\qquad \int \mathrm{E}\left[c_0(h_{\min})K_{h^2_{\min}}(u,X_2) - \mathrm{E}\{c_0(h_{\min})K_{h^2_{\min}}(u,X_2)\}\right]^2\omega_d(du) \\ &\leq 2\left[\int c_0^2(h)K^2_{h^2}(u,t)\omega_d(du) + \int_u\left\{c_0(h)\int_y K_{h^2}(u,y)f(y)\omega_d(dy)\right\}^2\ \omega_d(du)\right] \\ &\qquad \times \int \mathrm{E}\{c_0^2(h_{\min})K^2_{h_{\min}}(u,X_2)\}\omega_d(du) \\ &\leq 4c_0^2(h)c_2(h)c_0^2(h_{\min})c_2(h_{\min}).\end{aligned}$$

Therefore

$$B^2 \leq 8(n-1)c_0^2(h)c_2(h)c_0^2(h_{\min})c_2(h_{\min})$$

and

$$B^2x^3/n^4 \leq 8c_0^2(h)c_2(h)c_0^2(h_{\min})c_2(h_{\min})x^3/n^3.$$

Now using $\sqrt{ab} \leq \theta a/2 + \theta^{-1}b/2$, we obtain

$$Bx^{3/2}/n^2 \leq \theta' c_0^2(h)c_2(h)/(3n) + 6c_0^2(h_{\min})c_2(h_{\min})x^3/\theta' n^2.$$

Now we have

$$\begin{aligned}
C^2 &= \sum_{i=2}^{n}\sum_{j=1}^{i-1} \mathrm{E}[\{G_{h,h_{\min}}(X_i,X_j) + G_{h_{\min},h}(X_i,X_j)\}^2] \\
&\leq \Xi\, n^2 \mathrm{E}\{G^2_{h,h_{\min}}(X_1,X_2)\} \\
&= \Xi\, n^2 \mathrm{E}\left[\left[\int \{c_0(h)K_{h^2}(u,X_1) - f_h(u)\}\{c_0(h_{\min})K_{h^2_{\min}}(u,X_2) - f_{h_{\min}}(u)\}\omega_d(du)\right]^2\right] \\
&= \Xi\, n^2 \mathrm{E}\Bigg[\Bigg[\int c_0(h)K_{h^2}(u,X_1)c_0(h_{\min})K_{h^2_{\min}}(u,X_2)\omega_d(du) \\
&\qquad - \int c_0(h)K_{h^2}(u,X_1)\left\{\int c_0(h_{\min})K_{h^2_{\min}}(u,y)f(y)\omega_d(dy)\right\}\omega_d(du) \\
&\qquad - \int c_0(h_{\min})K_{h^2_{\min}}(u,X_2)\left\{\int c_0(h)K_{h^2}(u,y)f(y)\omega_d(dy)\right\}\omega_d(du) \\
&\qquad + \int_u \left\{\int_y c_0(h_{\min})K_{h^2_{\min}}(u,y)f(y)\omega_d(dy)\right\}\left(\int_y c_0(h)K_{h^2}(u,y)f(y)\omega_d(dy)\right)\omega_d(du)\Bigg]^2\Bigg]. \\
&\leq \Xi\, n^2(A_1 + A_2 + A_3 + A_4).
\end{aligned}$$

We have, for A_2,

$$\begin{aligned}
&\mathrm{E}\left[\int c_0(h)K_{h^2}(u,X_1)\left\{\int c_0(h_{\min})K_{h^2_{\min}}(u,y)f(y)\omega_d(dy)\right\}\omega_d(du)\right]^2 \\
&\leq \|f\|_\infty^2 \mathrm{E}\left[\int c_0(h)K_{h^2}(u,X_1)\left\{\int c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(dy)\right\}\omega_d(du)\right]^2 \\
&= \|f\|_\infty^2 \mathrm{E}\left[\int c_0(h)K_{h^2}(u,X_1)\omega_d(du)\right]^2 \\
&\leq \|f\|_\infty^2 \int \left\{\int c_0(h)K_{h^2}(u,y)\omega_d(du)\right\}^2 f(y)\omega_d(dy) = \|f\|_\infty^2.
\end{aligned}$$

With similar computations, we obtain the same bound for A_3. As for A_4, we get

$$\begin{aligned}
&\mathrm{E}\left[\int_u \left\{\int c_0(h_{\min})K_{h^2_{\min}}(u,y)f(y)\omega_d(dy)\right\}\left\{\int c_0(h)K_{h^2}(u,y)f(y)\omega_d(dy)\right\}\omega_d(du)\right]^2 \\
&\leq \mathrm{E}\left[\int \|f\|_\infty \left\{\int c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(dy)\right\}\left\{\int c_0(h)K_{h^2}(u,y)f(y)\omega_d(dy)\right\}\omega_d(du)\right]^2 \\
&\leq \|f\|_\infty^2 \mathrm{E}\left[\iint c_0(h)K_{h^2}(u,y)f(y)\omega_d(dy)\omega_d(du)\right]^2 \\
&\leq \|f\|_\infty^2 \left\{\int f(y)\int c_0(h)K_{h^2}(u,y)\omega_d(du)\omega_d(dy)\right\}^2 = \|f\|_\infty^2.
\end{aligned}$$

Hence

$$C^2 \leq \Xi\, n^2 \mathrm{E}\left[\left\{\int c_0(h)K_{h^2}(u,X_1)c_0(h_{\min})K_{h_{\min}}(u,X_2)\omega_d(du)\right\}^2\right] + \Xi\, \|f\|_\infty^2 \times n^2.$$

It remains to bound A_1. We have, using the Cauchy–Schwarz inequality,

$$
\begin{aligned}
&\mathrm{E}\left[\left\{\int c_0(h)K_{h^2}(u,X_1)c_0(h_{\min})K_{h_{\min}}(u,X_2)\omega_d(du)\right\}^2\right]\\
&= \int_y\int_x\left\{\int_u c_0(h)K_{h^2}(u,x)c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(du)\right\}^2 f(x)\omega_d(dx)f(y)\omega_d(dy)\\
&\le \|f\|_\infty\int_y\int_x\left\{\int_u c_0^2(h)K^2_{h^2}(u,x)c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(du)\times\right.\\
&\quad\left.\int_u c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(du)\right\}\omega_d(dx)f(y)\omega_d(dy)\\
&\le \|f\|_\infty\int_y\int_x\int_u c_0^2(h)K^2_{h^2}(u,x)c_0(h_{\min})K_{h^2_{\min}}(u,y)\omega_d(du)\omega_d(dx)f(y)\omega_d(dy)\\
&\le \|f\|_\infty c_0^2(h)c_2(h).
\end{aligned}
$$

Finally,

$$C \le \Xi\, n\|f\|_\infty^{1/2}c_0(h)\sqrt{c_2(h)} + \Xi\,\|f\|_\infty n.$$

Hence, given that $x \ge 1$, we get

$$
\begin{aligned}
C\sqrt{x}/n^2 &\le \Xi\,\|f\|_\infty^{1/2}c_0(h)\sqrt{c_2(h)}\sqrt{x}/n + \Xi\,\|f\|_\infty\sqrt{x}/n\\
&\le \theta' c_0^2(h)c_2(h)/(3n) + \Xi\,\|f\|_\infty x/(\theta' n) + \Xi\,\|f\|_\infty\sqrt{x}/n\\
&\le \theta' c_0^2(h)c_2(h)/(3n) + \Xi\,\|f\|_\infty x/(\theta' n).
\end{aligned}
$$

Now let us consider

$$\mathcal{S} = \left\{a = (a_i)_{2\le i\le n}, b = (b_i)_{1\le i\le n-1} : \sum_{i=2}^n \mathrm{E}\{a_i^2(X_i)\} \le 1, \sum_{i=1}^{n-1}\mathrm{E}\{b_i^2(X_i)\} \le 1\right\}.$$

We have

$$D = \sup_{(a,b)\in\mathcal{S}}\left[\sum_{i=2}^n\sum_{j=1}^{i-1}\mathrm{E}\{(G_{h,h_{\min}}(X_i,X_j) + G_{h_{\min},h}(X_i,X_j))a_i(X_i)b_j(X_j)\}\right].$$

We have, for $(a,b)\in\mathcal{S}$,

$$
\begin{aligned}
&\sum_{i=2}^n\sum_{j=1}^{i-1}\mathrm{E}\{G_{h,h_{\min}}(X_i,X_j)a_i(X_i)b_j(X_j)\}\\
&\quad\le \sum_{i=2}^n\sum_{j=1}^{i-1}\mathrm{E}\left\{\int |c_0(h)K_{h^2}(u,X_i) - f_h(u)||a_i(X_i)||c_0(h_{\min})K(u,X_j) - f_{h_{\min}}(u)||b_j(X_j)|\omega_d(du)\right\}\\
&\quad\le \sum_{i=2}^n\sum_{j=1}^{n-1}\int \mathrm{E}\left\{|c_0(h)K_{h^2}(u,X_i) - f_h(u)||a_i(X_i)|\right\}\mathrm{E}\{|c_0(h_{\min})K_{h^2_{\min}}(u,X_j) - f_{h_{\min}}(u)||b_j(X_j)|\}\omega_d(du),
\end{aligned}
$$

and for any u, using the Cauchy–Schwarz inequality, we get

$$
\begin{aligned}
\sum_{i=2}^n\mathrm{E}\left\{|c_0(h)K_{h^2}(u,X_i) - f_h(u)||a_i(X_i)|\right\} &\le \sqrt{n}\sqrt{\sum_{i=2}^n\mathrm{E}\{|c_0(h)K_{h^2}(u,X_i) - f_h(u)|^2\}\mathrm{E}\{a_i^2(X_i)\}}\\
&\le \sqrt{n}\sqrt{\sum_{i=2}^n\mathrm{E}\{c_0^2(h)K^2_{h^2}(u,X_i)\}\mathrm{E}\{a_i^2(X_i)\}}\\
&\le \sqrt{n}\sqrt{\|f\|_\infty c_0^2(h)c_2(h)\sum_{i=2}^n\mathrm{E}\{a_i^2(X_i)\}}\\
&\le \sqrt{c_0^2(h)c_2(h)}\sqrt{n\|f\|_\infty}.
\end{aligned}
$$

Now since

$$\int f_{h_{\min}}(u)\omega_d(du) = 1,$$

and

$$\int \mathrm{E}\{c_0(h_{\min})K_{h^2_{\min}}(u, X_j)\}\omega_d(du) = 1,$$

we have

$$\sum_{j=1}^{n-1}\int \mathrm{E}\{|c_0(h_{\min})K_{h^2_{\min}}(u, X_j) - f_{h_{\min}}(u)|\omega_d(du)|b_j(X_j)|\} \le 2\sum_{j=1}^{n-1}\mathrm{E}\{|b_j(X_j)|\} \le 2\sqrt{n}\sqrt{\sum_{j=1}^{n-1}\mathrm{E}\{|b_j^2(X_j)|\}} \le 2\sqrt{n}.$$

Finally,

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathrm{E}\{G_{h,h_{\min}}(X_i, X_j)a_i(X_i)b_j(X_j)\} \le 2n\sqrt{\|f\|_\infty}\sqrt{c_0^2(h)c_2(h)},$$

and

$$Dx/n^2 \le 4\sqrt{\|f\|_\infty}\sqrt{c_0^2(h)c_2(h)}/nx \le \theta' c_0^2(h)c_2(h)/(3n) + 12\|f\|_\infty x^2/(\theta' n).$$

In summary, we have proved

$$Ax^2/n^2 \le 8x^2 c_0(h_{\min})\|K\|_\infty/n^2, \quad Bx^{3/2}/n^2 \le \theta' c_0^2(h)c_2(h)/(3n) + 6c_0^2(h_{\min})c_2(h_{\min})x^3/n^2\theta'$$
$$C\sqrt{x}/n^2 \le \theta' c_0^2(h)c_2(h)/(3n) + \Xi\,\|f\|_\infty x/(\theta' n), \quad Dx/n^2 \le \theta' c_0^2(h)c_2(h)/(3n) + 12\|f\|_\infty x^2/(\theta' n).$$

But

$$c_0^2(h_{\min})c_2(h_{\min}) = c_0(h_{\min})\int K_{h^2_{\min}}(x, y)c_0(h_{\min})K_{h^2_{\min}}(x, y)\omega_d(dy) \le c_0(h_{\min})\|K\|_\infty.$$

Thus finally, with probability larger than $1 - 5.54|\mathcal{H}|e^{-x}$, we have, for any $h \in \mathcal{H}$,

$$|U(h, h_{\min})/n^2| \le \theta' c_0^2(h)c_2(h)/n + \Xi\,\|f\|_\infty x^2/(\theta' n) + \Xi\, c_0(h_{\min})\|K\|_\infty x^3/(\theta' n^2).$$

This ends the proof of Lemma B. □

Back to (A.9), we have the following control.

Lemma C. *With probability greater that $1 - 9.54|\mathcal{H}|e^{-x}$, for any $h \in \mathcal{H}$,*

$$\left|\langle \hat f_h - f, \hat f_{h_{\min}} - f\rangle - \langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n\right| \le \theta'\|f_h - f\|^2 + \theta' c_0^2(h)c_2(h)/n$$
$$+\{\theta'/2 + 1/(2\theta')\}\|f_{h_{\min}} - f\|^2 + Cx^2\|f\|_\infty/(\theta' n) + C(K)c_0(h_{\min})x^3/n^2, \tag{A.12}$$

where C is an absolute constant and $C(K)$ a constant only depending on K.

Proof of Lemma C. We have first to control (A.10) and (A.11), namely

$$\langle \hat f_h, f_{h_{\min}}\rangle/n - \langle \hat f_h, f_{h_{\min}}\rangle/n + \langle f_h, f_{h_{\min}}\rangle/n$$

and

$$V(h, h_{\min}) + V(h_{\min}, h) + \langle f_h - f, f_{h_{\min}} - f\rangle.$$

Let h and h' be fixed. We have

$$\langle \hat f_h, f_{h'}\rangle = \frac{1}{n}\sum_{i=1}^{n}\int c_0(h)K_{h^2}(x, X_i)f_{h'}(x)\omega_d(dx).$$

Therefore,

$$|\langle \hat f_h, f_{h'}\rangle| \le \|f_{h'}\|_\infty = \left\|\int c_0(h)K_{h'^2}(u, y)f(y)\omega_d(dy)\right\|_\infty \le \|f\|_\infty,$$

and

$$|\langle f_h, f_{h_{\min}}\rangle| \le \|f_h\|_\infty \int f_{h_{\min}}(u)\omega_d(du) \le \|f\|_\infty,$$

which gives the control of (A.10), viz.

$$|\langle \hat{f}_h, f_{h_{\min}}\rangle/n - \langle \hat{f}_h, f_{h_{\min}}\rangle/n + \langle f_h, f_{h_{\min}}\rangle/n| \leq 3\|f\|_\infty/n.$$

It remains to bound the three terms of (A.11). We get

$$V(h,h') = \langle \hat{f}_h - f_h, f_{h'} - f\rangle = \frac{1}{n}\sum_{i=1}^{n}\left[g_{h,h'}(X_i) - \mathrm{E}\{g_{h,h'}(X_i)\}\right]$$

with $g_{h,h'}(x) = \langle c_0(h)K_{h^2}(\cdot,x), f_{h'} - f\rangle$. We have $\|g_{h,h'}\|_\infty \leq \|f_{h'} - f\|_\infty \leq 2\|f\|_\infty$.

Furthermore using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\mathrm{E}\{g^2_{h,h'}(X_1)\} &= \int_y \left[\int_x c_0(h)K_{h^2}(x,y)\{f_{h'}(x) - f_h(x)\}\omega_d(dx)\right]^2 f(y)\omega_d(dy)\\
&\leq \|f\|_\infty \int_y \left[\int_x c_0(h)K_{h^2}(x,y)\{f_{h'}(x) - f_h(x)\}\omega_d(dx)\right]^2 \omega_d(dy)\\
&\leq \|f\|_\infty c_0(h)\int_y \left[\int_x K_{h^2}(x,y)\{f_{h'}(x) - f_h(x)\}^2\omega_d(dx)\right]\\
&\quad\times \left\{c_0(h)\int_x K_{h^2}(x,y)\omega_d(dx)\right\}\omega_d(dy)\\
&\leq \|f\|_\infty c_0(h)\int_y\int_x K_{h^2}(x,y)\{f_{h'}(x) - f_h(x)\}^2\omega_d(dx)\omega_d(dy)\\
&\leq \|f\|_\infty\|f_{h'} - f_h\|^2.
\end{aligned}$$

Consequently, with probability larger than $1 - 2e^{-x}$, Bernstein's inequality [20] leads to

$$|V(h,h')| \leq \sqrt{2x\|f\|_\infty\|f_{h'} - f\|^2/n} + 2x\|f\|_\infty/(3n) \leq \theta'\|f_{h'} - f\|^2/2 + C\|f\|_\infty x/(\theta' n).$$

The bound on $V(h,h')$ obtained above is first applied with $h' = h_{\min}$; then we invert the roles of h and $h_{\min}$. Besides, we have

$$|\langle f_h - f, f_{h_{\min}} - f\rangle| \leq \theta'\|f_h - f\|^2/2 + \|f_{h_{\min}} - f\|^2/(2\theta').$$

Finally using Lemma B, we get with probability larger than $1 - 9.54|\mathcal{H}|e^{-x}$,

$$\begin{aligned}
&|\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle - \langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n|\\
&\leq \theta' c_0^2(h)c_2(h)/n + \Xi\ \|f\|_\infty x^2/(\theta' n) + \Xi\ c_0(h_{\min})\|K\|_\infty x^3/\theta' n^2\\
&\qquad + 3\|f\|_\infty/n + \{\theta'/2 + 1/(2\theta')\}\|f_{h_{\min}} - f\|^2 + 2\|f\|_\infty x/(\theta' n) + \theta'\|f_h - f\|^2\\
&\leq \theta'\|f_h - f\|^2 + \theta' c_0^2(h)c_2(h)/n + \{\theta'/2 + 1/(2\theta')\}\|f_{h_{\min}} - f\|^2 + Cx^2\|f\|_\infty/(\theta' n) + C(K)c_0(h_{\min})x^3/n^2,
\end{aligned}$$

which completes the proof of Lemma C. □

Now Proposition A gives with probability larger than $1 - \Xi\ |\mathcal{H}|e^{-x}$, for any $h \in \mathcal{H}$,

$$\|f - f_h\|^2 + c_0^2(h)c_2(h)/n \leq 2\|f - \hat{f}_h\|^2 + C_2(K)\|f\|_\infty x^2/n,$$

where $C_2(K)$ depends only on K. Hence by applying Lemma C with h first and then $\hat{h}$ we obtain with probability larger than $1 - \Xi\ |\mathcal{H}|e^{-x}$, for any $h \in \mathcal{H}$,

$$\begin{aligned}
&|\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle - \langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n - \langle \hat{f}_{\hat{h}} - f, \hat{f}_{h_{\min}} - f\rangle + \langle c_0(\hat{h})K_{\hat{h}^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n|\\
&\leq 2\theta'\|\hat{f}_h - f\|^2 + 2\theta'\|\hat{f}_{\hat{h}} - f\|^2 + (\theta' + 1/\theta')\|f_{h_{\min}} - f\|^2 + \tilde{C}(K)\{\|f\|_\infty x^2/n + x^3 c_0(h_{\min})/n^2\}/\theta',
\end{aligned} \tag{A.13}$$

where $\tilde{C}(K)$ is a constant only depending on K. Now back to (A.8) and using (A.13), we have

$$\begin{aligned}\|\hat{f}_{\hat{h}} - f\|^2 \le\ & \|\hat{f}_h - f\|^2 + pen_\lambda(h) \\ & - 2\{\langle \hat{f}_h - f, \hat{f}_{h_{\min}} - f\rangle - \langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n\} - 2\langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n \\ & - pen_\lambda(\hat{h}) + 2\{\langle \hat{f}_{\hat{h}} - f, \hat{f}_{h_{\min}} - f\rangle - \langle c_0(h)K_{\hat{h}^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n\} \\ & + 2\langle c_0(\hat{h})K_{\hat{h}^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n \\ \le\ & \|\hat{f}_h - f\|^2 + pen_\lambda(h) \\ & - 2\langle c_0(h)K_{h^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n - pen_\lambda(\hat{h}) + 2\langle c_0(\hat{h})K_{\hat{h}^2}, c_0(h_{\min})K_{h^2_{\min}}\rangle/n \\ & + 4\theta'\|\hat{f}_h - f\|^2 + 4\theta'\|\hat{f}_{\hat{h}} - f\|^2 + 2(\theta' + 1/\theta')\|f_{h_{\min}} - f\|^2 \\ & + \tilde{C}(K)\{\|f\|_\infty x^2/n + x^3 c_0(h_{\min})/n^2\}/\theta'.\end{aligned}$$

Choosing $\theta' = \theta/4$ yields the result. This completes the proof of Proposition B. □

The next proposition gives a bound for the bias term (see [13]) that is used to obtain rates of convergence. Define for $f : \mathbb{S}^{d-1} \to \mathbb{R}$ and s even,

$$\mathcal{D}^s f = \sum_{i=1}^{s/2} \frac{2i}{(2i)!} \gamma_{2i,s/2-i} D^{2i} f,$$

where $\gamma_0 = 1$ and

$$\gamma_i = \sum_{\alpha_1+\cdots+\alpha_{d-1}=i} \frac{(-1)^{\alpha_1}}{(2\alpha_1+1)!} \cdots \frac{(-1)^{\alpha_{d-1}}}{(2\alpha_{d-1}+1)!}.$$

Proposition C. *Assume that $f \in \mathbf{F}_2(s)$. Let K be a class s kernel, where $s \ge 2$ is even. Then*

$$\lim_{h\to 0} \|h^{-s}|\mathrm{E}(\hat{f}_h) - f| - |\alpha_0^{-1}(K)\alpha_s(K)\mathcal{D}^s f|\| = 0.$$

For $d = 3$ and for von Mises kernel, SPCO algorithm turns to be simple to compute. Straightforward computations yield the next lemma, which specifies the various quantities involved in the procedure.

Lemma D. *For $\mathbb{S}^2$ and $K(x) = e^{-x}$, we have*

$$\begin{aligned}\|\hat{f}_h - \hat{f}_{h_{\min}}\|^2 =\ & \frac{4\pi c_0^2(h)}{n^2} e^{-2/h^2} h^2 \sum_{i,j} \frac{\sinh(|X_i + X_j|/h^2)}{|X_i + X_j|} \\ & + \frac{4\pi c_0^2(h_{\min})}{n^2} e^{-2/h^2_{\min}} h^2_{\min} \sum_{i,j} \frac{\sinh(|X_i + X_j|/h^2_{\min})}{|X_i + X_j|} \\ & - \frac{8\pi}{n^2} c_0^2(h) c_0^2(h_{\min}) e^{-1/h^2} e^{-1/h^2_{\min}} \sum_{i,j} \frac{\sinh(|X_i/h^2 + X_j/h^2_{\min}|)}{|X_i/h^2 + X_j/h^2_{\min}|},\end{aligned}$$

with $c_0(h)^{-1} = 4\pi e^{-1/h^2} h^2 \sinh(1/h^2)$ and $c_2(h) = 2\pi e^{-2/h^2} h^2 \sinh(2/h^2)$.

References

[1] A. Amiri, B. Thiam, T. Verdebout, On the estimation of the density of a directional data stream, Scand. J. Stat. 44 (2017) 249–267.
[2] Z.D. Bai, C.R. Rao, L.C. Zhao, Kernel estimators of density function of directional data, J. Multivariate Anal. 27 (1988) 24–39.
[3] P. Baldi, G. Kerkyacharian, D. Marinucci, D. Picard, Adaptive density estimation for directional data using needlets, Ann. Statist. 37 (2009) 3362–3395.
[4] M.S. Bartlett, Statistical estimation of density functions, Sankhyā Ser. A 25 (1963) 245–254.
[5] M.. Di Marzio, A. Panzera, C.C. Taylor, Kernel density estimation on the torus, J. Statist. Plann. Inference 141 (2011) 2156–2173.
[6] E. García-Portugués, Exact risk improvement of bandwidth selectors for kernel density estimation with directional data, Electron. J. Stat. 7 (2013) 1655–1685.
[7] E. García-Portugués, R.M. Crujeiras, W. González-Manteiga, Kernel density estimation for directional–linear data, J. Multivariate Anal. 121 (2013) 152–175.
[8] A. Goldenshluger, O. Lepski, Universal pointwise selection rule in multivariate function estimation, Bernoulli 14 (2008) 1150–1190.
[9] P. Hall, G.S. Watson, J. Cabrera, Kernel density estimation with spherical data, Biometrika 74 (1987) 751–762.
[10] C. Houdré, P. Reynaud-Bouret, Exponential inequalities, with constants, for U-statistics of order two, in: E. Giné, C. Houdré, D. Nualart (Eds.), Stochastic Inequalities and Applications, Birkhäuser, Basel, 2003, pp. 55–69.
[11] G. Kerkyacharian, T.M. Pham Ngoc, D. Picard, Localized spherical deconvolution, Ann. Statist. 39 (2011) 1042–1068.

[12] J.S. Klemelä, Estimation of Densities and Functionals of Densities with Spherical Data (Ph.D. Thesis), ProQuest LLC, Ann Arbor, MI, 1997, Helsingin Yliopisto, Finland.
[13] J.S. Klemelä, Estimation of densities and derivatives of densities with directional data, J. Multivariate Anal. 73 (2000) 18–40.
[14] C. Lacour, P. Massart, V. Rivoirard, Estimator selection: A new method with applications to kernel density estimation, Sankhyā 79 (2017) 298–335.
[15] O.V. Lepskiĭ, A problem of adaptive estimation in Gaussian white noise, Teor. Veroyatn. Primen. 35 (1990) 459–470.
[16] O.V. Lepskiĭ, Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates, Teor. Veroyatn. Primen. 36 (1991) 645–659.
[17] M. Lerasle, N.M. Magalhães, P. Reynaud-Bouret, Optimal kernel selection for density estimation, in: C. Houdré, D.M. Mason, P. Reynaud-Bouret, J. Rosiński (Eds.), High Dimensional Probability VII, Springer, Cham, 2016, pp. 425–460.
[18] C. Ley, T. Verdebout, Modern Directional Statistics, CRC Press, Boca Raton, FL, 2017.
[19] K.V. Mardia, P.E. Jupp, Directional Statistics, Wiley, Chichester, 2000.
[20] P. Massart, Concentration Inequalities and Model Selection, Springer, Berlin, 2007.
[21] M. Oliveira, R.M. Crujeiras, A. Rodríguez Casal, A plug-in rule for bandwidth selection in circular density estimation, Comput. Statist. Data Anal. 56 (2012) 3898–3908.
[22] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[23] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.
[24] C.C. Taylor, Automatic bandwidth selection for circular density estimation, Comput. Statist. Data Anal. 52 (2008) 3493–3500.