Published paper
Di Marzio, M, Panzera, A and Taylor, CC (2011) *Kernel density estimation on the torus.* Journal of Statistical Planning & Inference, 141 (6). 2156 - 2173 .

http://dx.doi.org/10.1016/j.jspi.2011.01.002

Kernel density estimation on the torus

Marco Di Marzio$^{(1)}$, Agnese Panzera$^{(1)}$ and Charles C. Taylor$^{(2)}$

(1) Università di Chieti-Pescara; (2) University of Leeds

Abstract: Kernel density estimation for multivariate, circular data has been formulated only when the sample space is the sphere, but theory for the torus would also be useful. For data lying on a d-dimensional torus ($d \geq 1$), we discuss kernel estimation of a density, its mixed partial derivatives, and their squared functionals. We introduce a specific class of product kernels whose order is suitably defined in such a way to obtain L_2-risk formulas whose structure can be compared to their euclidean counterparts. Our kernels are based on circular densities, however we also discuss smaller bias estimation involving negative kernels which are functions of circular densities. Practical rules for selecting the smoothing degree, based on cross-validation, bootstrap and plug-in ideas are derived. Moreover, we provide specific results on the use of kernels based on the von Mises density. Finally, real-data examples and simulation studies illustrate the findings.

Key words and phrases: Circular symmetric unimodal families, conformation angles, density functionals, efficiency, minimax bounds, mixed derivatives, sin-order, toroidal kernels, *twicing*, von Mises density

1. Introduction

A *circular* observation can be seen as a point on the unit circle, and represented by an angle $\theta \in [0, 2\pi)$. Typical examples include flight direction of birds from a point of release, wind, and ocean current direction. A circular observation is periodic, *i.e.* $\theta = \theta + 2m\pi$ for $m \in \mathbb{Z}$, which sets apart circular statistical analysis from standard real-line methods. Recent accounts are given by Jammalamadaka and SenGupta (2001) and Mardia & Jupp (1999). Concerning nonparametric density estimation, there exist only a few contributions focused on data lying on the circle or on the sphere (Bai et al. (1988); Beran (1979); Hall et al. (1987); Klemelä (2000); Taylor (2008)), but nothing specific for the d-dimensional torus $\mathbb{T}^d := [-\pi, \pi]^d$. This seems strange if we note that toroidal data occur frequently and we will naturally want to know the joint distribution of two or more circular random variables. A few examples follow.

In the study of wind directions over a time period it naturally arises the need of modeling bivariate circular data. In fact, temporal variables are converted into circular variables with simple transformations such as taking the day of the year and multiplying by $2\pi/365$, or taking month of the year and multiplying by $2\pi/12$. Again in metereology, parametric families of multivariate circular densities arise in a more specific and interesting fashion in a paper by Coles (1998). In zoology countless examples arise. Fisher (1986) considers the orientations of the nests of 50 noisy scrub birds (θ) along the bank of a creek bed, toghether with the corresponding directions (ϕ) of creek flow at the nearest point to the nest. Here the joint behavior of the random variable (θ, ϕ) is of interest. In evolution biology it is of interest to study paired circular genomes. Each genome contains a population of orthologs, and a way to characterize a genome consists in observing how they are located within the genome. Such locations are usually expressed as angles, so in the study of paired genomes it arises the necessity of modeling bivariate circular populations. This has been recently accomplished in a parametric fashion by Shieh et al. (2006).

An interesting discrimination problem for data on $\mathbb{T}^2$ is presented by Sengupta and Ugwuowo (2011). They have measurements on the skull of two groups of people represented by a front angle and a side angle. Surely density estimation on the torus seems a very simple tool for their discriminant aims, but here also density estimation *per se* could be useful.

A bioinformatics example is described briefly here, and it will be taken further in Section 8. Data on the (two-dimensional) torus are commonly found in descriptions of protein structure. Here, the protein backbone is given by a set of atom co-ordinates in $\mathbb{R}^3$ which can then be converted to a sequence of *conformation angles*. The sequence of angles can be used to assign (Kabsch and Sander, 1983) the structure of that part of the backbone (for example α-helix, β-sheet) which can then give insights into the functionality of the protein. A potential higher-dimensional example is provided by NMR data which will give replicate measurements, revealing a dynamic structure of the protein. For shorter peptides the modes of variability could be studied by an analysis of the replicates, requiring density estimation on a high-dimensional torus.

With regard to methodology, orthogonal series (see, for example, Efromovich (1999)) appear to be reasonable tools for densities estimation based on toroidal data, although they do not generally give densities as the output. On the other hand, splines are not straightforward to implement in more than one dimension. The kernel density estimator, which has been widely studied for its intuitive and simple formulation, is not immediately applicable to toroidal data. This is not simply due to their periodic nature, but also because circular densities are generally not defined as scale families, thus the usual structure of the kernel estimator as the average of re-scaled densities does not directly hold in this context.

In this paper we explore the possibility of formulating a toroidal density kernel estimator whose weight functions are based on some well-known circular densities. Specifically, based on a random sample from a population with density f — supported on the multidimensional torus, and having an absolutely continuous distribution function — we address the problem of kernel estimation of any mixed partial derivative of f. We have chosen the partial derivative framework to be as general as possible, however it has also been of practical interest both in the past and more recently; see Singh (1976), Prakasa Rao (2000) and the references therein, or Duong et al. (2008). Markedly, see the very detailed discussion on the importance of multivariate kernel density derivative estimation made by Chacón et al. (2010).

In Section 2, as the starting point, we define a class of suitable kernels whose order is defined in close analogy to the linear case. In Section 3 we introduce the estimators, then derive exact mean integrated squared error for the proposed density derivatives estimator and, finally, we discuss its minimax admissibility. Here from the Fourier series expansion of the mean integrated squared error, a new smoothing concept arises, according to which the usual two-stage choice, *i.e.* the selection of kernel and bandwidth, is replaced by the single step of selecting the Fourier coefficients of the optimal kernel. Section 4 is devoted to the asymptotic properties, and some usual accuracy measures are quantified for the proposed estimators. Interestingly, even though our definition of kernel order allows for a description of L_2 risks which is reminiscent of the linear case, increased values of the order does not necessarily give smaller bias. However, Section 5 illustrates a simple and general strategy to obtain small bias estimates. In Section 6 cross-validation and plug-in ideas are employed to construct various approaches to bandwidth (degree of smoothing) selection. The von Mises density could be considered in many respects the circular counterpart of the normal, therefore it represents a natural choice for the kernel. With this motivation, in Section 7 we give specific theory for the optimal smoothing when von Mises kernels are employed. Section 8 uses some real data on conformation angles in protein backbones to illustrate the potential of kernel density estimation in a bivariate context. Section 9 contains various simulation studies such as: a comparison on the basis of *efficiency* of our estimators with trigonometric series estimators, a study on accuracy of our asymptotic approximations, a comparison among cross validation bandwidth selection rules, and finally the construction of pointwise confidence intervals.

2. Toroidal kernels

Definition 1. *A d-dimensional toroidal kernel with concentration (smoothing) parameters $\boldsymbol{C} := (\kappa_s \in \mathbb{R}_+, s = 1, \cdots, d)$, is the d-fold product $K_{\boldsymbol{C}} := \prod_{s=1}^d K_{\kappa_s}$, where $K_\kappa : \mathbb{T} \to \mathbb{R}$ is such that*

i) it admits an uniformly convergent Fourier series $\{1 + 2\sum_{j=1}^\infty \gamma_j(\kappa)\cos(j\theta)\}/(2\pi)$, $\theta \in \mathbb{T}$, where $\gamma_j(\kappa)$ is a strictly monotonic function of κ;

ii) $\int_\mathbb{T} K_\kappa = 1$, and, if K_κ takes negative values, there exists $0 < M < \infty$ such that, for all $\kappa > 0$

$$\int_\mathbb{T} |K_\kappa(\theta)|\, d\theta \le M\,;$$

iii) for all $0 < \delta < \pi$,

$$\lim_{\kappa\to\infty} \int_{\delta \le |\theta| \le \pi} |K_\kappa(\theta)|\, d\theta = 0.$$

These kernels are continuous and symmetric about the origin, so the d-fold products of von Mises, wrapped normal and wrapped Cauchy distributions are included. As more general examples, we now list families of circular densities whose d-fold products are candidates as toroidal kernels.

1. Wrapped symmetric stable family of Mardia (1972, p. 72).
2. Extensions of the von Mises distribution (Batschelet, 1981, p. 288, equation (15.7.3)).
3. Unimodal symmetric distributions in the family of Kato and Jones (2009).
4. The family of unimodal symmetric distributions of Jones and Pewsey (2005).
5. The wrapped t family of Pewsey et al. (2007).

Note that the cardioid density $(2\pi)^{-1}\{1 + 2\kappa\cos(\cdot)\}$ with $|\kappa| < 1/2, \theta \in \mathbb{T}$, is not included in our class since it does not satisfy condition *iii)*. As another relevant example, observe that orthogonal series density estimates are not included since they do not satisfy condition *i)*. Additionally, the Dirichlet kernel does not match also *ii)*.

Definition 2. ***(Sin-order)*** *Given the univariate toroidal kernel K_κ, let $\eta_j(K_\kappa) := \int_\mathbb{T} \sin^j(\theta) K_\kappa(\theta) d\theta$. We say that K_κ has sin-order q if and only if $\eta_j(K_\kappa) = 0, \quad for \quad 0 < j < q, \quad and \quad \eta_q(K_\kappa) \ne 0$.*

The following Lemma will be useful throughout the paper.

Lemma 1. *If K_κ has sin-order q, then $\eta_q(K_\kappa) = O\{(1-\gamma_q(\kappa))2^{1-q}\}$.*

Proof. See Appendix. □

Note that $K_{\boldsymbol{C}} := \prod_{s=1}^d K_{\kappa_s}$ has sin-order q if and only if K_{κ_s} has sin-order q. Higher sin-order toroidal kernels can be constructed from second sin-order ones as a direct consequence of the formulation of $\eta_j(K_\kappa)$ in (10.1) and of the result in Lemma 1, which leads to this qth sin-order kernel defined at $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{T}^d$

$$\prod_{s=1}^{d} \left\{ K_{\kappa_s}(\theta_s) - \sum_{j=1}^{q/2-1} \frac{\cos(2j\theta_s)}{\pi} [\gamma_{2j}(\kappa_s) - 1] \right\}. \tag{2.1}$$

An adaptation of the bias reduction technique of Lejeune and Sarda (1992) to the circular setting constitutes a different strategy to increase the sin-order, as in the following

Lemma 2. *Assume that K_κ has sin-order 2, let $\boldsymbol{W}_\ell$ be a matrix of order $\ell+1$ with (i,j)-th entry given by $w_{ij} := \eta_{i+j-2}(K_\kappa)$, and $\boldsymbol{U}_\ell$ be a matrix of order $\ell+1$ with (i,j)-th entry given by $u_{ij} := \sin^{i+j-2}(\theta)$ if $j=1$, and $u_{ij} = w_{ij}$ otherwise. Then, given*

$$\mathcal{K}_{\kappa,\ell}(\theta) := \frac{\mathsf{det}\,[\boldsymbol{U}_\ell]}{\mathsf{det}\,[\boldsymbol{W}_\ell]} K_\kappa(\theta),$$

we have that $\prod_{s=1}^d \mathcal{K}_{\kappa_s,\ell}(\theta_s)$ is a toroidal kernel of sin-order $\ell+1$ if ℓ is odd, and $\ell+2$ otherwise.

Proof. See Appendix. □

Notice that kernels whose sin-order is greater than 2 need to be negative in some regions of $\mathbb{T}^d$, just like euclidean higher order kernels.

3. The estimators

For a d-variate function g and a multi-index $\boldsymbol{r} = (r_1, \cdots, r_d) \in \mathbb{Z}_+^d$, we denote the mixed partial derivative of (total) order $|\boldsymbol{r}| = \sum_{s=1}^d r_s$ at $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)$ by

$$g^{(\boldsymbol{r})}(\boldsymbol{\theta}) := \frac{\partial^{|\boldsymbol{r}|}}{\partial\theta_1^{r_1} \cdots \partial\theta_d^{r_d}} g(\boldsymbol{\theta}),$$

and indicate the quadratic functional $\int_{\mathbb{T}^d} \{g^{(\boldsymbol{r})}(\boldsymbol{\theta})\}^2 d\boldsymbol{\theta}$ as $R(g^{(\boldsymbol{r})})$. Finally, as *toroidal density* we mean a probability density function whose support is $\mathbb{T}^d$. Our aim is to study the estimation of $f^{(\boldsymbol{r})}(\boldsymbol{\theta})$ and $R(f^{(\boldsymbol{r})})$.

Definition 3. ***(Kernel estimator of toroidal density mixed derivatives)*** *Let* $\{\boldsymbol{\Theta}_\ell, \ell = 1, \cdots, n\}$ *with* $\boldsymbol{\Theta}_\ell = (\Theta_{\ell 1}, \cdots, \Theta_{\ell d})$*, be a random sample from a toroidal density* f*. The kernel estimator of* $f^{(\boldsymbol{r})}$ *at* $\boldsymbol{\theta}$ *is defined as*

$$\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta}; \boldsymbol{C}) := \frac{1}{n} \sum_{\ell=1}^{n} K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\theta} - \boldsymbol{\Theta}_\ell). \tag{3.1}$$

Letting $\hat{g}$ be a nonparametric estimator of a square-integrable curve g, the mean squared error (MSE) for $\hat{g}$ at $\theta \in \mathsf{supp}[g]$ is defined by $\mathsf{MSE}[\hat{g}(\theta)] := \mathsf{E}[\{\hat{g}(\theta) - g(\theta)\}^2] = \{\mathsf{E}[\hat{g}(\theta)] - g(\theta)\}^2 + \mathsf{Var}[\hat{g}(\theta)]$, whereas the mean integrated squared error (MISE) is $\mathsf{MISE}[\hat{g}] := \int \mathsf{MSE}[\hat{g}(\theta)] d\theta$. In what follows we will derive a Fourier expansion of the exact MISE for the estimator (3.1). Before stating the main result, we need to introduce a little formalism through the following two propositions.

Proposition 1. *Given* $\boldsymbol{j} = (j_1, \cdots, j_d) \in \mathbb{Z}^d$*, for a function* f *defined on* $\mathbb{T}^d$ *we have*

$$f^{(\boldsymbol{r})}(\boldsymbol{\theta}) = \frac{i^{|\boldsymbol{r}|}}{(2\pi)^d} \sum_{\boldsymbol{j} \in \mathbb{Z}^d} \boldsymbol{j}^{\boldsymbol{r}} c_{\boldsymbol{j}} e^{i \boldsymbol{j} \cdot \boldsymbol{\theta}},$$

where $i^2 = -1$*,* $c_{\boldsymbol{j}} := \int_{\mathbb{T}^d} f(\boldsymbol{\theta}) e^{-i \boldsymbol{j} \cdot \boldsymbol{\theta}} d\boldsymbol{\theta}$*,* $\boldsymbol{j} \cdot \boldsymbol{\theta}$ *is the inner product of* $\boldsymbol{j}$ *and* $\boldsymbol{\theta}$*,* $\boldsymbol{j}^{\boldsymbol{r}} = \prod_{s=1}^d j_s^{r_s}$*, and, by convention,* $j_s^{r_s} = 1$ *for* $j_s = r_s = 0$*.*

Proposition 2. *Given the d-dimensional toroidal kernel* $K_{\boldsymbol{C}}(\boldsymbol{\theta}) = \prod_{s=1}^d K_{\kappa_s}(\theta_s)$*, let* $\gamma_{\boldsymbol{j}}(\boldsymbol{C}) := \int_{\mathbb{T}^d} K_{\boldsymbol{C}}(\boldsymbol{\theta}) e^{-i \boldsymbol{j} \cdot \boldsymbol{\theta}} d\boldsymbol{\theta} = \prod_{s=1}^d \gamma_{j_s}(\kappa_s)$*. Hence, the estimator in (*3.1*), being the convolution between the empirical version of* f *and* $K_{\boldsymbol{C}}^{(\boldsymbol{r})}$*, can be expressed as*

$$\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta}; \boldsymbol{C}) = \frac{i^{|\boldsymbol{r}|}}{(2\pi)^d} \sum_{\boldsymbol{j} \in \mathbb{Z}^d} \boldsymbol{j}^{\boldsymbol{r}} \tilde{c}_{\boldsymbol{j}} \gamma_{\boldsymbol{j}}(\boldsymbol{C}) e^{i \boldsymbol{j} \cdot \boldsymbol{\theta}}, \tag{3.2}$$

where $\tilde{c}_{\boldsymbol{j}} := n^{-1} \sum_{\ell=1}^n e^{-i \boldsymbol{j} \cdot \boldsymbol{\Theta}_\ell}$*.*

To obtain the properties of $\hat{f}^{(\boldsymbol{r})}$, we will need to assume a certain smoothness degree of f and $K_{\boldsymbol{C}}$. To this end, we require that f and $K_{\boldsymbol{C}}$ are elements of the periodic Sobolev class of order $|\boldsymbol{r}|$ on $\mathbb{T}^d$

$$\mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d) := \left\{ g \in L_2\left(\mathbb{T}^d\right) : \int_{\mathbb{T}^d} \{g^{(\boldsymbol{p})}\}^2 \leq L^2 \ , \quad \text{for} \quad 0 \leq |\boldsymbol{p}| \leq |\boldsymbol{r}| \right\}$$

where g is a toroidal density, $\boldsymbol{p} = (p_1, \cdots, p_d) \in \mathbb{Z}_+^d$, $|\boldsymbol{p}| = \sum_{s=1}^d p_s$ and $L \in (0, \infty)$.

Theorem 1. *Suppose that both f and $K_{\boldsymbol{C}}$ belong to $\mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$, then*

$$\mathsf{MISE}\left[\hat{f}^{(\boldsymbol{r})}(\cdot;\boldsymbol{C})\right] = \frac{i^{2|\boldsymbol{r}|}}{n(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left\{1-\left(\sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m}\right)^2-\left(\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}\right)^2\right\}\gamma_{\boldsymbol{j}}^2(\boldsymbol{C})\boldsymbol{j}^{2\boldsymbol{r}}$$
$$+\frac{i^{2|\boldsymbol{r}|}}{(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\{1-\gamma_{\boldsymbol{j}}(\boldsymbol{C})\}^2\left\{\left(\sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m}\right)^2+\left(\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}\right)^2\right\}\boldsymbol{j}^{2\boldsymbol{r}},$$

with $\{\alpha_{\boldsymbol{j},m},\beta_{\boldsymbol{j},m}, m=1,\cdots,2^{d-1}\}$ being the set of the coefficients in the trigonometric Fourier series expansion of f.

Proof. See Appendix. □

It is easily seen that the first summand is the integrated variance. To ensure consistency, we need to select $\boldsymbol{C}$ such that, when n increases, $n^{-1}(2\pi)^{-d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\gamma_{\boldsymbol{j}}^2(\boldsymbol{C})\boldsymbol{j}^{2\boldsymbol{r}}$ tends to zero, and $\gamma_{\boldsymbol{j}}(\boldsymbol{C})$ tends to 1 for any $\boldsymbol{j}$. On the basis of Theorem 1 we get

Result 1. *For any fixed $\boldsymbol{j}\in\mathbb{Z}^d$, the kernel Fourier coefficient minimizing MISE is*

$$\frac{\left(\sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m}\right)^2+\left(\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}\right)^2}{n^{-1}\left\{1-\left(\sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m}\right)^2-\left(\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}\right)^2\right\}+\left(\sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m}\right)^2+\left(\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}\right)^2}.$$

which closely corresponds the the MISE-optimal Fourier coefficient of trigonometric series estimators, this latter being usually expressed for zero-derivative density estimation in $\mathbb{R}$.

The above optimal Fourier coefficient suggests to estimate Fourier coefficients of the unknown density as a novel kernel estimation approach, where the tasks of smoothing selection and kernel choice are no longer separated, in contrast to standard kernel density estimation. The practical implementation amounts to an estimation made of two steps, firstly by the use of orthogonal series, after which the kernel method. Because the Fourier coefficient estimates will not necessarily lead to a non-negative density estimate, the second step could be viewed as a non-negative re-normalization of the first estimate.

A bound for the MISE of the estimator (3.1) is derived in the following

Theorem 2. *Let $K_{\boldsymbol{C}} := \prod_{j=1}^d K_\kappa \in \mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$ be a toroidal kernel of sin-order q. If*

i) $\eta_q(K_\kappa)<\infty$;

ii) $O\left(\eta_q(K_\kappa)\right) < O\left(\eta_{q+2s}(K_\kappa)\right)$ *for any* $s \geq 1$*;*

iii) $f^{(\boldsymbol{r})} \in \mathcal{S}_L^{(|q\mathbf{1}|)}(\mathbb{T}^d)$*;*

then for any $n \geq 1$

$$\mathsf{MISE}[\hat{f}^{(\boldsymbol{r})}(\cdot;\boldsymbol{C})] \leq \left\{\frac{\eta_q(K_\kappa)}{q!}Ld\right\}^2 + \frac{1}{n}R(K_{\boldsymbol{C}}^{(\boldsymbol{r})}). \tag{3.3}$$

Proof. See Appendix. □

Now, assuming that $\eta_q(K_\kappa)(q!)^{-1} = c_1 k^{-\beta}$ and $R(K_{\boldsymbol{C}}^{(\boldsymbol{r})}) = c_2 k^{\alpha}$, with $c_i \in \mathbb{R}$, $i = 1, 2$, and $(\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+$, we find that the value of κ which minimizes the RHS of (3.3) is

$$\kappa_{\min} = \left(\frac{2\beta c_1^2 L^2 d^2 n}{\alpha c_2}\right)^{1/(2\beta+\alpha)}$$

which leads to

$$\mathsf{MISE}[\hat{f}^{(\boldsymbol{r})}(\cdot;\kappa_{\min})] = O\left(n^{-2\beta/(2\beta+\alpha)}\right). \tag{3.4}$$

By formula (3.4) it follows that, when $d = 1$ and $|\boldsymbol{r}| = 0$, the estimators equipped with second sin-order kernels in the list after Definition 1 and corresponding higher sin-order kernels in Lemma 2, for which $\alpha = 1/2$ and $\beta = q/2$, attain the minimax bound for nonparametric density estimators, say $\tilde{f}_n$, formulated by Efromovich and Pinsker (1982). This is given by

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{S}_L^q} \mathsf{MISE}[\tilde{f}_n] = \mathcal{P}(q, L)(1 + o(1)), \tag{3.5}$$

where

$$\mathcal{P}(q, L) = (2q+1)\left[\frac{\pi(2q+1)(q+1)}{q}\right]^{-2q/(2q+1)} L^{2/(2q+1)}$$

is Pinsker's constant.

Now, since functionals of the form $R(f^{(\boldsymbol{p})})$, $\boldsymbol{p} = (p_1, \cdots, p_d)$, occur in many bandwidth selection strategies, we need to define an estimator also for them. However, as in the linear setting, an easy application of integration by parts shows that it will be sufficient to focus on the functionals $\psi_{\boldsymbol{r}} := \int f^{(\boldsymbol{r})}(\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$ where $|\boldsymbol{r}|$ is even.

Definition 4. ***(Kernel estimator of multivariate toroidal density functionals)*** *Given a random sample* $\{\boldsymbol{\Theta}_\ell, \ell = 1, \cdots, n\}$ *from a toroidal density* f*, the kernel estimator of the functional* $\psi_{\boldsymbol{r}}$ *is defined as*

$$\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C}) := n^{-1}\sum_{\ell=1}^{n} \hat{f}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_\ell;\boldsymbol{C}). \tag{3.6}$$

4. Asymptotic properties

Based on properties *i)–iii)* of Definition 1, we easily prove the following

Theorem 3. *If* $\lim_{n\to\infty}\kappa_s=\infty$ *for* $s=1,...,d$*, these three statements are equivalent:*

1. $f^{(\boldsymbol{r})}$ *is uniformly continuous;*

2. $\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\mathbb{T}^d}\left|\mathsf{E}\left[\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C})\right]-f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right|=0;$

3. $\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\mathbb{T}^d}\mathsf{MSE}\left[\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C})\right]=0.$

Notice, in particular, that for $|\boldsymbol{r}|=0$ we have consistency with f being merely continuous.

To derive the asymptotic distribution of $\hat{f}^{(\boldsymbol{r})}$ we need the following result, which follows from Parseval's identity.

Lemma 3. *If* $K_{\boldsymbol{C}}\in\mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$*, then* $R(K_{\boldsymbol{C}}^{(\boldsymbol{r})})=\prod_{s=1}^d Q_{\kappa_s}(r_s)$*, where, for each non-negative integer* u*,*

$$Q_\kappa(u):=\begin{cases}(2\pi)^{-1}\left\{1+2\sum_{j=1}^\infty\gamma_j^2(\kappa)\right\}, & \textit{if } u=0\ ;\\ \\ \pi^{-1}\sum_{j=1}^\infty j^{2u}\gamma_j^2(\kappa), & \textit{otherwise.}\end{cases}\tag{4.1}$$

Theorem 4. *Let* $K_{\boldsymbol{C}}:=\prod_{s=1}^d K_{\kappa_s}\in\mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$*, with* K_{κ_s} *being an univariate toroidal kernel of sin-order* q*, and* $f^{(\boldsymbol{r})}\in\mathcal{S}_L^{|q\boldsymbol{1}|}(\mathbb{T}^d)$*. Assume that* $\lim_{n\to\infty}\gamma_q(\kappa_s)=1$*, where* $\gamma_q(\kappa_s)$ *is the* q*th Fourier coefficient of* K_{κ_s}*; then*

$$\sqrt{n}\left\{\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C})-\mathsf{E}\left[\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C})\right]\right\}\xrightarrow{d}\mathcal{N}\left(0,\frac{f(\boldsymbol{\theta})}{n}\prod_{s=1}^d Q_{\kappa_s}(r_s)\right).$$

Proof. See Appendix. □

Corollary 1. *Under the assumptions of Theorem 4, we have*

$$\mathsf{MISE}\left[\hat{f}^{(\boldsymbol{r})}(\cdot;\boldsymbol{C})\right]\sim\frac{1}{(q!)^2}\int_{\mathbb{T}^d}\mathsf{tr}^2\left\{\boldsymbol{\Omega}_q\frac{\mathsf{d}^q f^{(\boldsymbol{r})}(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\}d\boldsymbol{\theta}+\frac{1}{n}\prod_{s=1}^d Q_{\kappa_s}(r_s),\tag{4.2}$$

where $\sim$ *indicates that the ratio is bounded as* $\kappa\to\infty$*,* $\boldsymbol{\Omega}_q:=\mathsf{diag}\{\eta_q(K_{\kappa_1}),\cdots,\eta_q(K_{\kappa_d})\}$*, and* $\mathsf{d}^q f^{(\boldsymbol{r})}(\boldsymbol{\theta})/\mathsf{d}\boldsymbol{\theta}^q$ *indicates the matrix derivative of order* q *of* $f^{(\boldsymbol{r})}$ *at* $\boldsymbol{\theta}$*.*

It is not straightforward to obtain the asymptotic mean integrated squared error (AMISE) of $\hat{f}^{(\boldsymbol{r})}(\cdot;\boldsymbol{C})$ using the RHS in (4.2) since the ratio of the two sides will not, in general tend to unity. This is because the rate at which $\eta_j(K_\kappa)$ decreases may not depend on j. However, if the kernel is such that $O(\eta_j(K_\kappa))$ is a strictly decreasing function of j then, together with Lemma 1, we can further establish

Theorem 5. *If the Fourier coefficients of a second sin-order kernel satisfy*

$$\lim_{\kappa\to\infty}\frac{1-\gamma_j(\kappa)}{1-\gamma_2(\kappa)}=\frac{j^2}{4} \tag{4.3}$$

then the terms on the RHS of equation (4.2) (with $q=2, |\boldsymbol{r}|=0$, and $\kappa_s=\kappa$) will define AMISE, which has a minimization given by

$$\gamma_2(\kappa)=1-2^5 3^{d+1} d\left[2^5 3^d(4d-1)+9n2\pi^d\int_{\mathbb{T}^d}\mathsf{tr}^2\left\{\frac{\mathsf{d}^2 f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\}d\boldsymbol{\theta}\right]^{-1}.$$

As expected, the optimal coefficient approaches 1 as n increases, but slower for larger d. The above condition can be extended to higher sin-order kernels, by noting that $\gamma_j(\kappa)=1$ for $1\leq j<q$ and recursively solving equation (10.1) to ensure that $\eta_j(K_\kappa)$, $j>q$ has the appropriate order.

Remark 1. *Condition (4.3) is satisfied by many symmetric, unimodal densities, though not by the wrapped Cauchy. Important cases are given by the wrapped normal and von Mises. Also the class introduced by Batschelet (1981) matches the condition, along with the unimodal symmetric densities in the family introduced by Kato and Jones (2009).*

Concerning the squared risk of the functional estimator (3.6), we have the following

Theorem 6. *Consider the estimator $\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})$ equipped with the kernel $K_{\boldsymbol{C}}:=\prod_{s=1}^d K_{\kappa_s}\in\mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$, with K_{κ_s} being a univariate toroidal kernel of sin-order q, and recall conditions i) and ii) of Theorem 4. Then, if $f^{(\boldsymbol{r})}\in\mathcal{S}_L^{|q\boldsymbol{1}|}(\mathbb{T}^d)$,*

$$\begin{aligned}\mathsf{MSE}\left[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})\right]\sim&\left[\frac{1}{n}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0})+\frac{1}{q!}\int_{\mathbb{T}^d}\mathsf{tr}\left\{\boldsymbol{\Omega}_q\frac{\mathsf{d}^q f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\}f^{(\boldsymbol{r})}(\boldsymbol{\theta})d\boldsymbol{\theta}\right]^2\\&+\frac{2}{n^2}\psi_{\boldsymbol{0}}\prod_{s=1}^d Q_{\kappa_s}(r_s)+\frac{4}{n}\left[\int_{\mathbb{T}^d}\{f^{(\boldsymbol{r})}(\boldsymbol{\theta})\}^2 f(\boldsymbol{\theta})d\boldsymbol{\theta}-\psi_{\boldsymbol{r}}^2\right].\end{aligned}\tag{4.4}$$

Moreover, letting $\mathcal{E}_1(\boldsymbol{\theta}_1):=\mathsf{E}[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\theta}_1-\boldsymbol{\Theta}_2)]$ and $\xi_1:=\mathsf{Var}[\mathcal{E}_1(\boldsymbol{\Theta}_1)]$, where $\boldsymbol{\theta}_1\in\mathbb{T}^d$ and $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are independent random variables both distributed according to the population *density f, if* $\mathsf{E}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\right] < \infty$*, we have*

$$\sqrt{n}\left\{\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C}) - \mathsf{E}\left[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})\right]\right\} \xrightarrow{d} \mathcal{N}(0, 4\xi_1).$$

Proof. See Appendix. □

5. Small bias estimates

The bias of an euclidean kernel estimator is said to have order q if the infimum of AMISE has magnitude $O\left(n^{-2q/(2q+2|\boldsymbol{r}|+d)}\right)$ and so, in this case q is also the kernel order. Now, it is interesting to note that higher sin-order kernels are not necessarily associated with smaller bias, as indeed we would expect by analogy with the linear setting. For example, the use of kernels defined as in (2.1) do not yield necessarily smaller bias.

Conversely, a simple and general bias reduction technique which does not affect the sin-order follows. If the toroidal kernel $\prod K_\kappa$ gives bias with magnitude $O\left(\kappa^{-h}\right)$, then

$$K_\kappa^t = 2K_\kappa - K_{2^{-1/h}\kappa} \tag{5.1}$$

produces bias of order $O\left(\kappa^{-h-1}\right)$. Observe that $\prod 2K_\kappa^t - K_{2^{-1/h}\kappa}^t$ yields bias order $O\left(\kappa^{-h-2}\right)$, so, iteratively, any bias order is obtainable, provided that the population density is sufficiently smooth. Notice that above kernel amounts to *twicing* of Stuetzle and Mittal (1979) if the kernel belongs to a family closed under the convolution operation, which is true for the wrapped stable family, for example. However, the von Mises density does not have this closure property, which makes successive iterations of standard twicing difficult to implement.

Concerning minimax admissibility, it is easily seen that the estimators equipped with kernels of sin-order q obtained by (5.1), whose bias order is $\beta = (q+2)/2$, attain the bound in (3.5) with q replaced by $q + 2$.

Finally, notice that Lemma 2 contains a different strategy to reduce bias, if toroidal kernels such as the d-fold products of densities listed after Definition 1 are employed.

6. Selection of the smoothing degree

To select the optimal values of the smoothing parameters κ_s, $s = 1, \cdots, d$, different strategies are available. Here we consider the case of density estimation ($|\boldsymbol{r}| = 0$), and adapt some selectors which have been widely investigated in the euclidean setting.

An intuitive selection strategy, proposed in the euclidean setting by Habbema et al. (1974) and Duin (1976), consists in choosing the values of κ_s, $s = 1, 2, \cdots, d$, which maximize the likelihood cross-validation (LCV) function $\mathsf{LCV}[\boldsymbol{C}] = n^{-1}\sum_{\ell=1}^{n}\log\hat{f}_{-\ell}(\boldsymbol{\theta}_\ell; \boldsymbol{C})$, where $\hat{f}_{-\ell}(\boldsymbol{\theta}_\ell;\boldsymbol{C})$ denotes the leave-one-out estimate of f . Concerning the efficiency of such a selector, we have the remarkable fact that, differently from the euclidean setting, in our setting both the population density and the kernel are always bounded and compactly supported for being toroidal densities, and consequently, in our scenario, the conditions for the L_1 consistency of the density estimator, as stated in Theorem 2 of Chow et al. (1983), always hold.

A different criterion for choosing the smoothing parameter is the unbiased cross-validation (UCV) introduced by Rudemo (1982) and Bowman (1984). This selector, which targets the integrated squared error (ISE) of the estimator in (3.1) with $|\boldsymbol{r}| = 0$ (given by $\mathsf{ISE}[\hat{f}(\cdot;\boldsymbol{C})] := \int_{\mathbb{T}^d}\{\hat{f}(\boldsymbol{\theta};\boldsymbol{C}) - f(\boldsymbol{\theta})\}^2 d\boldsymbol{\theta})$, leads to the minimization of the unbiased cross-validation objective function $\mathsf{UCV}[\boldsymbol{C}] = R(\hat{f}) - 2n^{-1}\sum_{\ell=1}^n \hat{f}_{-\ell}(\boldsymbol{\theta}_\ell;\boldsymbol{C})$.

Further selection strategies are those based on the minimization of an estimate of the AMISE. Here we adapt both the biased cross-validation (BCV) (Scott and Terrell, 1987) and a direct plug-in selector. However, to enable the use of AMISE estimates, we have to assume that the condition of Theorem 5 (or equivalent conditions for $q > 2$) hold. Moreover, to make our notation easier, we suppose that $\kappa_s = \kappa$ for each $s = 1, \cdots, d$, and consequently

$$\mathsf{AMISE}\left[\hat{f}(\cdot;\boldsymbol{C})\right] = \left\{\frac{\eta_q(K_\kappa)}{q!}\right\}^2 \int_{\mathbb{T}^d} \mathsf{tr}^2\left\{\frac{\mathsf{d}^q f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\} d\boldsymbol{\theta} + \frac{1}{n}\prod_{s=1}^d Q_\kappa(0)$$

where $\mathsf{tr}\{\mathsf{d}^q f(\boldsymbol{\theta})/\mathsf{d}\boldsymbol{\theta}^q\} = \sum_{s=1}^d f^{(q\boldsymbol{e}_s)}(\boldsymbol{\theta})$, whit $\boldsymbol{e}_s$ being a d-dimensional vector having 1 as s-th entry and 0 elsewhere. Now, for $s = 1, \cdots, d$ and $t > s$, the s-th squared summand and the product between the s-th and t-th summands of $\int_{\mathbb{T}^d} \mathsf{tr}\{\mathsf{d}^q f(\boldsymbol{\theta})/\mathsf{d}\boldsymbol{\theta}^q\}^2 d\boldsymbol{\theta}$, are respectively

$$\int_{\mathbb{T}^d}\left\{f^{(q\boldsymbol{e}_s)}(\boldsymbol{\theta})\right\}^2 d\boldsymbol{\theta} = \psi_{2q\boldsymbol{e}_s} \text{ and } \int_{\mathbb{T}^d} f^{(q\boldsymbol{e}_s)}(\boldsymbol{\theta}) f^{(q\boldsymbol{e}_t)}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \psi_{q\boldsymbol{e}_{st}}, \tag{6.1}$$

whose leave-one-out estimates, defined by $\hat{\psi}^*_{\boldsymbol{r}}(\kappa) := n^1 \sum_{\ell=1}^n \hat{f}^{(\boldsymbol{r})}_{-\ell}(\boldsymbol{\theta}_\ell;\kappa)$, lead to the biased cross-validation objective function

$$\mathsf{BCV}[\kappa] = \left\{\frac{\eta_q(K_\kappa)}{q!}\right\}^2 \left\{\sum_{s=1}^d \hat{\psi}^*_{2q\boldsymbol{e}_s}(\kappa) + 2\sum_{s=1}^d\sum_{t>s} \hat{\psi}^*_{q\boldsymbol{e}_{st}}(\kappa)\right\} + \frac{1}{n}\prod_{s=1}^d Q_\kappa(0).$$

Finally, by considering the estimators in Definition 4 for the quantities in (6.1) lead to the direct plug-in objective function

$$\mathsf{DPI}[\kappa] = \left\{\frac{\eta_q(K_\kappa)}{q!}\right\}^2 \left\{\sum_{s=1}^d \hat{\psi}_{2q\boldsymbol{e}_s}(\lambda_s) + 2\sum_{s=1}^d\sum_{t>s} \hat{\psi}_{q\boldsymbol{e}_{st}}(\delta_s)\right\} + \frac{1}{n}\prod_{s=1}^d Q_\kappa(0),$$

where both the λ_ss and the δ_ss are pilot bandwidths.

More selectors of smoothing degree are provided in Section 3 for the case when the von Mises kernel is employed. In particular, we discuss both the reference of a von Mises distribution, which could be considered the rule of thumb in our circular setting, and the more sophisticated bootstrap method for the choice of the optimal smoothing degree.

7. Von Mises kernel theory

Now we derive the AMISE-optimal smoothing parameter for the estimator in (3.1) when the kernel is $V_{\boldsymbol{C}}(\boldsymbol{\theta}) := \prod_{s=1}^{d} V_\kappa(\theta_s) \in \mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$, where $V_\kappa(\cdot) := \exp\{\kappa\cos(\cdot)\}/\{2\pi\mathcal{I}_0(\kappa)\}$ is the von Mises kernel with $\mathcal{I}_j(\kappa)$ being the modified Bessel function of the first kind and order j. Here, we have assumed that $\kappa_s = \kappa$, $s = 1, \cdots, d$, to simplify notation, and we have chosen a specific kernel because, in general, the smoothing parameter is not separable from the toroidal kernel function, and, therefore, rules which hold for the whole class of toroidal kernels are very hard to obtain.

Theorem 7. *Assume that $f^{(\boldsymbol{r})} \in \mathcal{S}_L^{|\boldsymbol{21}|}(\mathbb{T}^d)$, and*

i) $\lim_{n\to\infty} \kappa = \infty$;

ii) $\lim_{n\to\infty} n^{-1} R\left(V_{\boldsymbol{C}}^{(\boldsymbol{r})}\right) = 0$;

then, the AMISE optimal smoothing parameter for $\hat{f}^{(\boldsymbol{r})}(\cdot;\boldsymbol{C})$ equipped with the kernel $V_{\boldsymbol{C}}$, is

$$\kappa_{AMISE} = \left[\frac{2^{|\boldsymbol{r}|+d}\pi^{d/2} n \int_{\mathbb{T}^d} \mathsf{tr}^2\left\{\frac{\mathsf{d}^2 f^{(r)}(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\} d\boldsymbol{\theta}}{(2|\boldsymbol{r}|+d)\prod_{s=1}^{d}\mathsf{OF}(2r_s)}\right]^{2/\{4+2|\boldsymbol{r}|+d\}}, \tag{7.1}$$

where, for integer u, $\mathsf{OF}(u)$ is the product of all odd integers less or equal to u.

Proof. See Appendix. □

Result (7.1) yields $\min_{\kappa>0}\mathsf{AMISE}[\kappa] = O\left(n^{-4/(4+d+2|\boldsymbol{r}|)}\right)$. This convergence rate can be improved by employing higher sin-order kernels defined in Lemma 2. In particular, letting $\mathcal{V}_{\kappa,\ell}(\theta) := \mathsf{det}[U_\ell]/\mathsf{det}[W_\ell]V_\kappa(\theta)$, for the case $|\boldsymbol{r}| = 0$, we have

Theorem 8. *Let $\mathcal{V}_{\boldsymbol{C},\ell} := \prod \mathcal{V}_{\kappa,\ell}$. Assume that conditions i) and ii) of Theorem 7 hold, and $f \in \mathcal{S}_L^{|q\boldsymbol{1}|}(\mathbb{T}^d)$, where $q = \ell + 1$ if ℓ is odd, and $q = \ell + 2$ otherwise. Then for the estimator $\hat{f}(\cdot;\boldsymbol{C})$ equipped with $\mathcal{V}_{\boldsymbol{C},\ell}$*

$$\kappa_{AMISE} = \left[\frac{2q\{\mathsf{OF}(q)\}^2 2^d \pi^{d/2} n \int_{\mathbb{T}^d} \mathsf{tr}^2\left\{\frac{\mathsf{d}^q f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\} d\boldsymbol{\theta}}{d(q!)^2(q-1)^{d/2}}\right]^{2/\{2q+d\}} \tag{7.2}$$

Proof. See Appendix. □

It is also possible to determine the optimal smoothing degree for the small bias estimator of Section 5 building through the von Mises kernel, as in

Theorem 9. *Let* $V^t_{\boldsymbol{C}} := \prod V^t_\kappa$*. Suppose that conditions i) and ii) of Theorem 7 hold, and* $f \in \mathcal{S}^{|\mathbf{41}|}(\mathbb{T}^d)$*. Then for the estimator* $\hat{f}(\cdot;\boldsymbol{C})$ *equipped with the kernel* $V^t_{\boldsymbol{C}}$

$$\kappa_{AMISE} = \left[\frac{2^{d/2-1}\pi^{d/2} n \int_{\mathbb{T}^d}\left(\mathsf{tr}\left\{\frac{\mathsf{d}^4 f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^4}\right\} - 2\mathsf{tr}\left\{\frac{\mathsf{d}^2 f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\}\right)^2 d\boldsymbol{\theta}}{d}\right]^{2/(8+d)}. \qquad (7.3)$$

Proof. See Appendix. □

Observe that, by (7.3), the second sin-order tordoidal kernel $V^t_{\boldsymbol{C}}$ gives $\min_{\kappa>0} \mathsf{AMISE}[\kappa] = O\left(n^{-8/(8+d)}\right)$, which is equal to the minimum AMISE rate result ing from Theorem 8 for $q = 4$. Clearly, this rate can be improved by iterating the bias reduction procedure in (5.1) starting from V^t_κ.

Now letting $\mathsf{AMSE}[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})]$ denote the leading terms of RHS of (4.4), for the estimator (3.6) equipped with the kernel $V_{\boldsymbol{C}}$, we obtain

Theorem 10. *Suppose that the estimator* $\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})$ *is equipped with the kernel* $V_{\boldsymbol{C}} \in \mathcal{S}_L^{|\boldsymbol{r}|}(\mathbb{T}^d)$*. Assume that conditions i) and ii) of Theorem 7 hold, and* $f^{(\boldsymbol{r})} \in \mathcal{S}_L^{|\mathbf{21}|}(\mathbb{T}^d)$*. Then the AMSE-optimal smoothing parameter for* $\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})$ *is*

$$\kappa_{AMSE} = \begin{cases} \left[-\dfrac{i^{|\boldsymbol{r}|}2^{d/2-1}\pi^{d/2} n \sum_{s=1}^d \psi_{\boldsymbol{r}+2\boldsymbol{e}_s}}{\prod_{s=1}^d \mathsf{OF}(r_s)}\right]^{2/(2+|\boldsymbol{r}|+d)}, & \textit{if all } r_s \textit{ are even;} \\ \left[\dfrac{2^{|\boldsymbol{r}|+d}\pi^{d/2} n^2 \left(\sum_{s=1}^d \psi_{\boldsymbol{r}+2\boldsymbol{e}_s}\right)^2}{2\psi_{\mathbf{0}}(2|\boldsymbol{r}|+d)\prod_{s=1}^d \mathsf{OF}(2r_s)}\right]^{2/(4+2|\boldsymbol{r}|+d)}, & \textit{otherwise.} \end{cases} \qquad (7.4)$$

Proof. See Appendix. □

Concerning the smoothing degree selection, assuming that f is a d-fold product of von Mises densities having concentration parameters $\nu_s > 0$, $s = 1, \cdots, d$, we can get a *von Mises reference rule* to select κ. In particular, for the case $|\boldsymbol{r}| = 0$, formula (7.1) becomes a smoothing degree selector when the integrated squared trace of the Hessian matrix $\mathsf{d}^2 f(\boldsymbol{\theta})/\mathsf{d}\boldsymbol{\theta}^2$ is replaced by an estimate of it such as

$$\frac{\prod \mathcal{I}_0(2\hat{\nu}_s)\{3\sum \hat{\nu}_s^2 + \sum\sum_{s\neq t} \hat{\nu}_s\hat{\nu}_t A_1(2\hat{\nu}_s)A_1(2\hat{\nu}_t) - \sum A_1(2\hat{\nu}_s)\}}{2^{d+2}\prod \mathcal{I}_0^2(\hat{\nu}_s)},$$

where $A_j(\cdot) := \mathcal{I}_j(\cdot)/\mathcal{I}_0(\cdot)$ for each $j \in \mathbb{N}$, and $\hat{\nu}_s$ denotes an estimate of ν_s. Clearly, this selection strategy applies also for the case $|\boldsymbol{r}| \neq 0$. In particular, for the case $\boldsymbol{r} = \mathbf{1}$, with $\mathbf{1}$ denoting the d-dimensional unit vector, the above argument can be adapted by using, as an estimate of $\int_{\mathbb{T}^d} \mathsf{tr}^2 \left\{\mathsf{d}^2 f^{(\mathbf{1})}(\boldsymbol{\theta})/\mathsf{d}\boldsymbol{\theta}^2\right\} d\boldsymbol{\theta}$,

$$\frac{\prod \hat{\nu}_s\mathcal{I}_1(2\hat{\nu}_s)\{15\sum \hat{\nu}_s^2 B_3(2\hat{\nu}_s) + 9\sum\sum_{s\neq t} \hat{\nu}_s\hat{\nu}_t B_2(2\hat{\nu}_s)B_2(2\hat{\nu}_t) + 6(2d+3)\sum \hat{\nu}_s B_2(2\hat{\nu}_s) + 4d^2\}}{2^{2d+2}\prod \mathcal{I}_0^2(\hat{\nu}_s)}, \tag{7.5}$$

where $B_j(\cdot) := \mathcal{I}_j(\cdot)/\mathcal{I}_1(\cdot)$, $j \in \mathbb{Z}$.

In addiition, the bootstrap method of Taylor (1989) has closed expressions if a von Mises kernel is used. In what follows we briefly describe this. A bootstrap criteria is to select $\boldsymbol{C}$ to minimize $\int_{\mathbb{T}^d} \mathsf{E}_{\mathrm{B}}\{\hat{f}^*(\boldsymbol{\theta};\boldsymbol{C}) - \hat{f}(\boldsymbol{\theta};\boldsymbol{C})\}^2 d\boldsymbol{\theta}$ where E_{B} denotes the bootstrap expectation with respect to random samples $\{\boldsymbol{\Theta}_\ell^*, \ell = 1, \cdots, n\}$ generated from $\hat{f}(\boldsymbol{\theta};\boldsymbol{C})$. When $d = 1$ and the kernel is $V_\kappa(\theta)$ we can compute

$$\begin{aligned}\mathsf{E}_{\mathrm{B}}[\hat{f}^*(\theta;\kappa) - \hat{f}(\theta;\kappa)]^2 = \{2\pi n\mathcal{I}_0(\kappa)\}^{-2}&\left[\mathcal{I}_0(\kappa)^{-1}\sum_{\ell=1}^{n}\mathcal{I}_0\left(\kappa\{5 + 4\cos(\theta - \theta_\ell)\}^{1/2}\right)\right]\\ &+ \left\{\mathsf{E}_{\mathrm{B}}[\hat{f}^*(\theta;\kappa)] - \hat{f}(\theta;\kappa)\right\}^2 - n^{-1}\left\{\mathsf{E}_{\mathrm{B}}[\hat{f}^*(\theta;\kappa)]\right\}^2\end{aligned}$$

where

$$\mathsf{E}_{\mathrm{B}}[\hat{f}^*(\theta;\kappa)] = \{2\pi n\mathcal{I}_0(\kappa)\}^{-1}\sum_{\ell=1}^{n}\mathcal{I}_0\left(2\kappa\cos((\theta - \theta_\ell)/2)\right)/\mathcal{I}_0(\kappa).$$

Although this can be easily extended to $d > 1$ it seems hard to obtain an analytic form for its integral, so a numerical solution is required.

8. A real data case study

The *backbone* of a protein comprises a sequence of atoms, N_1–C_1^α–$\mathrm{C}_1 \cdots$–N_m–C_m^α–C_m, in which each group N_i–C_i^α–C_i is associated with a pair of *dihedral* angles and a type of amino acid. The way in which the distribution of the angles depends on the amino acid is of interest, and the kernel density estimate is both an exploratory tool to indicate differences as well as a means to identify the *nature* of differences found from a formal test. To illustrate this, we use a database of proteins which have small sequence similarity and collect together all the dihedral angles associated with each of the twenty types of amino acid. Previous attempts to model such data using a mixture of bivariate von Mises-type distributions (Mardia et al., 2007) have resulted in some success in identifying clusters which are associated with secondary structure. However, the number of components in the mixture model is problematic, and correct convergence of the EM algorithm is not assured. We have computed a kernel density estimate of these data using a von Mises kernel with the smoothing (concentration) parameter chosen by cross-validation. To illustrate some of the results, we have chosen 4 of the amino acid datasets (Alanine, Glutemate, Glycine, and Lysine). In Figure 8.1 we have plotted the contours defined so that, at level $p, (p = 0.1, 0.3, 0.5, 0.7)$ a total fraction $1 - p$ of the density is inside the contour. It can be seen that three of these distributions appear quite similar and one (Glycine) very different: that Glycine is different is well-known and well understood in terms of its chemical properties. A formal test to compare angular distributions can be obtained by using bootstrap

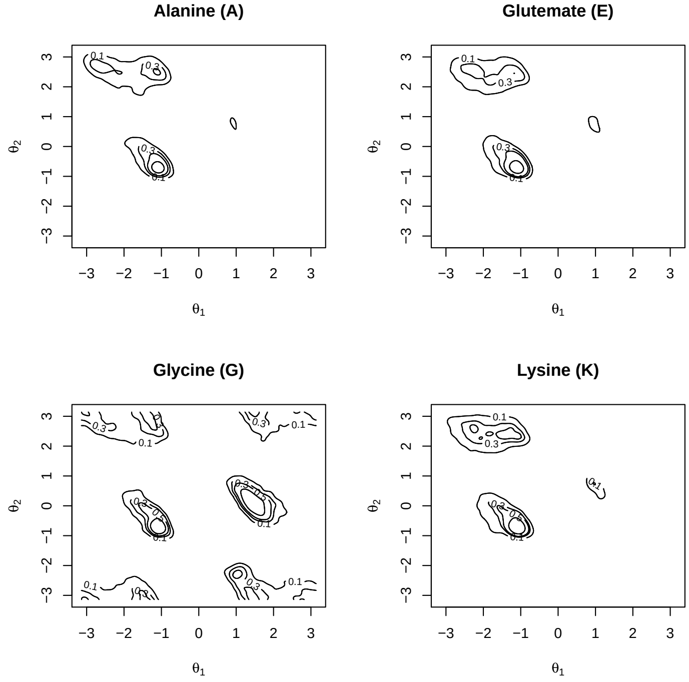

Figure 8.1: Contour plots showing the tail probabilities of kernel density estmates for four sets of dihedral angles, each corresponding to an amino acid. The sample sizes are: 8979 (A), 6183 (E), 8334 (G); 5984 (K), and corresponding smoothing parameters, chosen by cross-validation, are: $\kappa = 132, 114, 142, 121$ respectively.

resamples, for example using the energy test (Rizzo, 2002) or a similar procedure based on the difference in kernel density estimates. Such tests confirm that all four densities are indeed different.

9. Simulations

9.1 A comparison with trigonometric series estimators

The aim of this section is to compare our methods with other available ones. Trigonometric series estimators are natural competitors because we are working in periodic spaces. Since trigonometric series can be expressed as kernels, a comparison in terms of kernel efficiency is straightforward. We discuss this topic in one dimension because our kernels are products of univariate functions, and therefore not much should change in higher dimensions. The efficiency theory of euclidean kernels is based on the fact that the bandwidth and the kernel have separable contributions to the mean integrated squared error. Unfortunately, this is not the case for the MISE of estimator (3.1). In our efficiency analysis we use the MISE given by Theorem 1 and consider estimating the von Mises and the wrapped Cauchy densities with (no loss of generality) mean direction 0, specified by their concentration parameter ρ. In this context, when considering the (relative) efficiency of two circular kernels, the smoothing parameters do not "cancel" and so their equivalence needs first to be established as follows. For fixed ρ and n, we can select the bandwidth to minimize MISE for a given kernel function. The efficiency of one kernel relative to another may then be measured by taking the ratio of the minimized MISEs.

Coming to the specific summation method involved, Fejér's kernel (F_κ) — determined by $\gamma_j(\kappa) = (\kappa + 1 - j)/(\kappa + 1)\nVdash_{\{j \leq \kappa\}}$ — which is non-negative, is the obvious competitor and so the efficiency of other methods are compared to this benchmark. We also consider the Dirichlet method (D_κ) — despite some theoretical drawbacks — which is determined by $\gamma_j(\kappa) = \nVdash_{\{j \leq \kappa\}}$. Amongst the many other summation methods available, we consider the de la Vallée Poussin's sum (DV_κ) — for which $\gamma_j(\kappa) = 1$ for $j \leq \kappa$, $\gamma_j(\kappa) = 2 - j/\kappa$ for $\kappa + 1 \leq j \leq 2\kappa - 1$ and $\gamma_j(\kappa) = 0$ otherwise — because it has the best theoretical properties (see Efromovich (1999), p.43). On the other side, among our proposals, we have chosen von Mises kernel (V_κ), for which $\gamma_j(\kappa) = \mathcal{I}_j(\kappa)/\mathcal{I}_0(\kappa)$, and twiced von Mises ($V_\kappa^t$), for which $\gamma_j(\kappa) = 2\mathcal{I}_j(\kappa)/\mathcal{I}_0(\kappa) - \mathcal{I}_j(\kappa/2)/\mathcal{I}_0(\kappa/2)$, for competition. Note that D_κ, DV_κ and V_κ^t are not bona fide estimates. Concerning the usual issue whether to prefer bona fide estimators, our position is the common one, i.e.: negative estimators are of interest only if they guarantee faster convergence rates of their asymptotic risks, surely their effectiveness is doubtful with small sample sizes.

In Figure 9.2 we show the relative efficiency of the above kernels and trigonometric series for sample sizes $n = 5, 25, 125, 625$ for the von Mises and wrapped Cauchy distributions. The von Mises kernel is clearly superior to the Fejér kernel. However, the dominance of V_κ over D_κ is less than expected and we note that D_κ behaves reasonably for bigger samples, until nearly dominating F_κ for $n = 625$. Surprisingly, twicing improves on all the methods for small to medium sample sizes, and still does well for larger n. Overall, it could be the best, though DV_κ behaves better for $n = 625$ when the population is highly concentrated. Unfortunately DV_κ behaves always very poorly with low concentrated data.

Finally, we notice that to twice a twicing estimator could be a valid strategy to improve the performances still further. But we have preferred do not deepen this aspect in our simulations because simple twicing already hits reasonably the target.

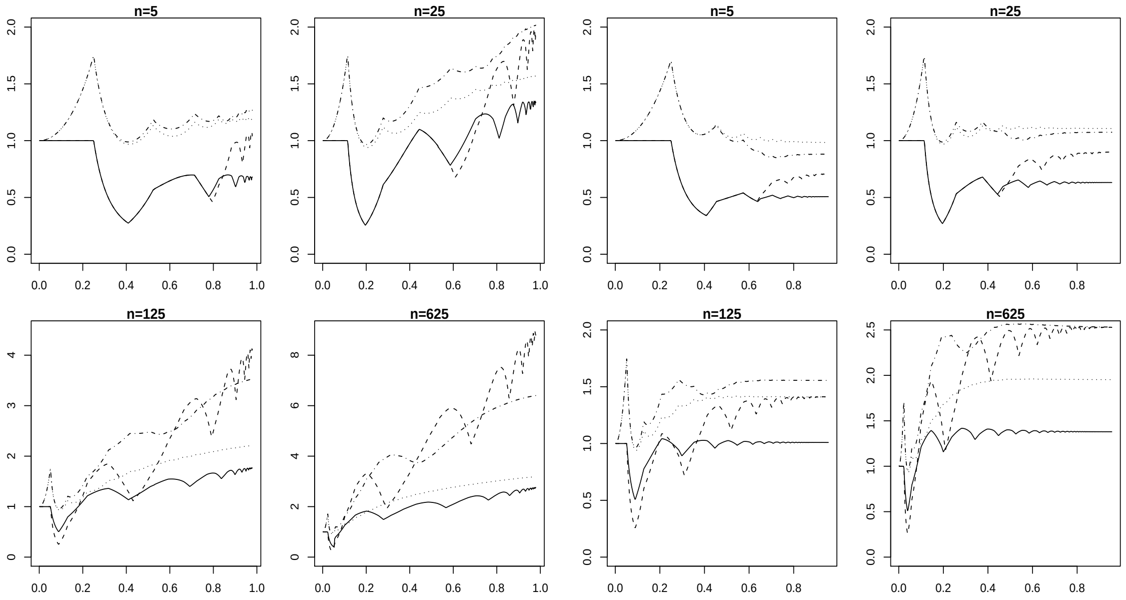

Figure 9.2: Relative efficiency of Dirichlet (——), de la Vallée Poussin (- - - - -), von Mises (· · ·), and twiced von Mises (dot-dash) kernels to the Fejér kernel, for various values of n, plotted as a function of ρ. With respect to the underlying true density, the left group corresponds to the von Mises distribution with $\rho = \mathcal{I}_1(\nu)/\mathcal{I}_0(\nu)$ and the right group to the wrapped Cauchy distribution.

9.2 Bandwidth selection

In a comparative simulation study we explore the performance of the cross-validation

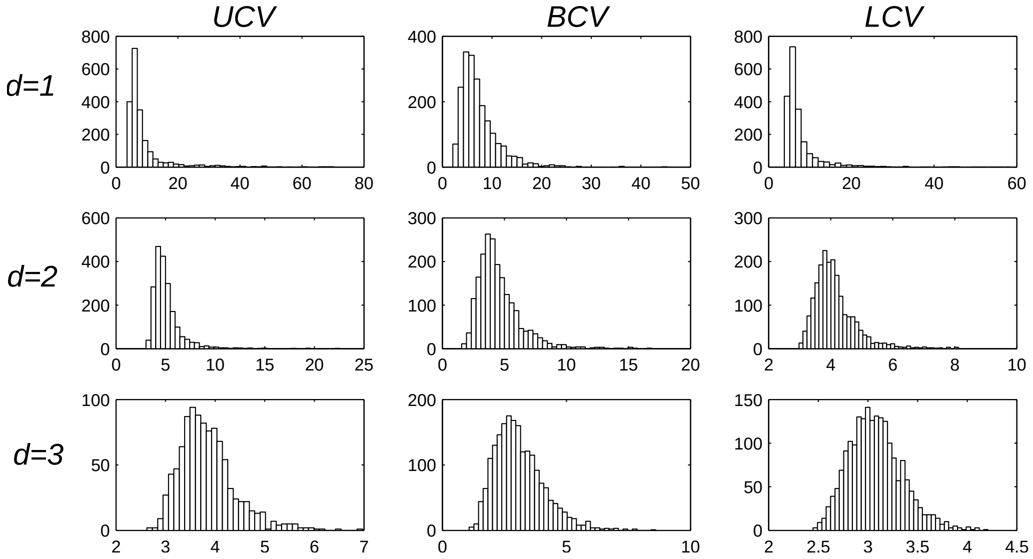

Figure 9.3: Bandwidths obtained using various selectors, biased cross-validation, unbiased cross-validation and likelihood cross-validation. Each row of histograms is based over a datased of 2000 samples with $n = 300$ drawn from a d-fold product of von Mises densities with null mean direction and and concentration parameter 1.

selectors discussed in Section 6. We have focused on them, other than for their computational simplicity — in fact they do not require any specification of pilot bandwidths — also because they could be considered reasonable for a number of theoretical respects, as convincingly argued by Loader (1999).

In particular, our target is the estimation of d-fold products of von Mises densities with null mean direction and unitary concentration parameter. Our kernel is $V_{\boldsymbol{C}}$, with $\boldsymbol{C}$ being a multiset of element κ and multiplicity d. In a first simulation study we have drawn 2000 samples with $n = 300$, and then calculated the corresponding bandwidths.

The output is represented in Figure 9.3 where each histogram corresponds to a pair (dimension, selector). A main message is that in one dimension the region where the real minimum lies could well be completely missed, and, indeed, all the selectors behave similarly. But, as dimensions increase, the estimate of the optimal bandwidth becomes more stable, markedly for LCV algorithm. In Table 9.1 we consider two more sample sizes, $n = 100$ and $n = 1000$, within the same experiment. Also, the MISE-optimal smoothing degrees are reported, not also the AMISE ones, which, however, have resulted quite similar.

On the other hand we recall that LCV does not optimizes L_2 discrepancies at all. We see that for $d = 1$ the euclidean theory is confirmed, whereas BCV has the tendency to oversmooth with respect to UCV, having also the smallest variability. Surely the average values of both of them undersmooth with respect to the MISE-optimal degree, due to the well known attitude of cross-validation algorithms to produce outliers. On the other hand, in higher dimensions UCV is seen nearly unbiased, whilst BCV slightly alleviates oversmoothing as the sample size increases. Finally, concerning LCV, we see that it appears asymptotically the most stable, producing the biggest smoothing degree for large n.

The boostrap selector was also considered for these datasets. As is the case for data in $\mathbb{R}$, the target function is zero for $\kappa = 0$ and so we would seek a local minimum for $\kappa > 0$. However, some simulated datasets had no local minimum (for example, about 15% of the datasets when $n = 100, d = 1$) and so the results are not reported here. However, for those datasets which had a local minimum, there was much less variation in the optimized smoothing parameter than for cross-validation, with generally biassed values. For example, for $n = 1000, d = 1$ the mean was 9.45, with standard deviation 0.81, which can be compared with the results in Table 9.1.

		$n = 100$	$n = 300$	$n = 1000$
$d = 1$	UCV	5.750 (5.106)	8.899 (7.606)	13.396 (9.253)
	BCV	4.614 (2.981)	7.369 (3.804)	12.256 (5.622)
	LCV	5.095 (4.066)	7.609 (5.132)	12.211 (7.121)
	MISE	4.455	6.896	11.112
$d = 2$	UCV	3.711 (1.558)	5.084 (1.535)	7.535 (1.582)
	BCV	2.770 (1.376)	4.554 (1.775)	7.095 (1.886)
	LCV	2.990 (0.653)	4.136 (0.691)	6.066 (0.827)
	MISE	3.387	4.904	7.409
$d = 3$	UCV	2.829 (0.658)	3.844 (0.570)	5.447 (1.101)
	BCV	1.912 (0.920)	3.109 (1.001)	4.814 (0.953)
	LCV	2.390 (0.325)	3.078 (0.271)	4.160 (0.281)
	MISE	2.759	3.814	5.270

Table 9.1: Performance of various smoothing selectors in toroidal density estimation. The means (standard deviations) are taken over 2000 samples of size n from d-fold products of von Mises density with mean direction 0 and concentration parameter 1.

9.3 Twicing

A small simulation to consider the impact of twicing on MISE was considered. As before we used 2000 samples for sample sizes of $n = 100, 300, 1000$ in each of $d = 1, 2, 3$ dimensions. For each dataset we compute the ISE of the estimator (3.1) with kernel $V_{\boldsymbol{C}}$ (KDE), and its twicing version (TWKDE) using formula (5.1), assuming that $\boldsymbol{C}$ is a multiset of element κ and multiplicity d. This is done for a suitable range of smoothing parameters, and then we compute the average ISE over the 2000 simulations. The results are shown in Table 9.2, and we observe that twicing can reduce the average integrated squared error by more than 20% in higher dimensions, with correspondingly more smoothing (smaller κ) being optimal. This is comparable to the case of data in $\mathbb{R}^d$.

		$n = 100$		$n = 300$		$n = 1000$	
		optimal κ	$\overline{\text{ISE}}$	optimal κ	$\overline{\text{ISE}}$	optimal κ	$\overline{\text{ISE}}$
$d = 1$	KDE	4.421	0.0050	7.053	0.0023	11.105	0.0010
	TWKDE	2.053	0.0043	2.842	0.0019	4.474	0.0008
$d = 2$	KDE	3.368	0.0034	4.947	0.0017	7.421	0.0008
	TWKDE	2.053	0.0028	2.316	0.0013	3.316	0.0006
$d = 3$	KDE	2.737	0.0015	3.895	0.0009	5.474	0.0005
	TWKDE	1.789	0.0012	2.316	0.0006	2.868	0.0003

Table 9.2: Average integrated squared error ($\overline{\text{ISE}}$) and correponding optimal smoothing parameter (κ) for the standard kernel estimator and its twicing version, for various sample sizes and dimensions d, taken over 2000 simulated datasets.

9.4 Confidence intervals

As an application of our results we can construct the following approximate, normal based, pointwise confidence interval for $f^{(\boldsymbol{r})}(\boldsymbol{\theta})$ at level $1 - \alpha$:

$$\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta}; \boldsymbol{C}) \pm z_{\alpha/2} \sqrt{\widehat{\mathsf{Var}}[\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta}; \boldsymbol{C})]},$$

where $z_{\alpha/2}$ indicates the $(1 - \alpha/2)$-quantile of the standard normal distribution.

To test practical performances we employ the same samples generated for Table 9.1. In particular we have considered confidence intervals for the case $|\boldsymbol{r}| = 0$ and $\boldsymbol{r} = \boldsymbol{1}$, and tested, for both cases, two choices of estimators: the estimator (3.1) with kernel $V_{\boldsymbol{C}}$ (KDE), and its twicing version (TWKDE) using formula (5.1), assuming that $\boldsymbol{C}$ is a multiset of element κ and multiplicity d. For these estimators we have considered an estimate of the first term in the Taylor expansion of the variance, obtaining, for KDE

$$\widehat{\mathsf{Var}}[\hat{f}(\boldsymbol{\theta};\boldsymbol{C})] = \left\{\frac{\mathcal{I}_0(2\hat{\kappa})}{2\pi\mathcal{I}_0^2(\hat{\kappa})}\right\}^d \frac{\hat{f}(\boldsymbol{\theta};\hat{\kappa})}{n},$$

whereas for TWKDE

$$\widehat{\mathsf{Var}}[\hat{f}(\boldsymbol{\theta};\boldsymbol{C})] = \left\{\frac{\mathcal{I}_0^3(\hat{\kappa}) - 4\mathcal{I}_0(\hat{\kappa}/2)\mathcal{I}_0(\hat{\kappa})\mathcal{I}_0(3\hat{\kappa}/2) + 4\mathcal{I}_0^2(\hat{\kappa}/2)\mathcal{I}_0(2\hat{\kappa})}{2\pi\mathcal{I}_0^2(\hat{\kappa}/2)\mathcal{I}_0^2(\hat{\kappa})}\right\}^d \frac{\hat{f}(\boldsymbol{\theta};\hat{\kappa})}{n}.$$

For the derivative case we have, respectively,

$$\widehat{\mathsf{Var}}[\hat{f}^{(\mathbf{1})}(\boldsymbol{\theta};\boldsymbol{C})] = \left\{\frac{\hat{\kappa}\mathcal{I}_1(2\hat{\kappa})}{4\pi\mathcal{I}_0^2(\hat{\kappa})}\right\}^d \frac{\hat{f}(\theta;\hat{\kappa})}{n},$$

and

$$\widehat{\mathsf{Var}}[\hat{f}^{(\mathbf{1})}(\boldsymbol{\theta};\boldsymbol{C})] = \left(\frac{\hat{\kappa}}{24\pi}\right)^d \left\{\frac{3\mathcal{I}_1(\hat{\kappa})}{\mathcal{I}_0^2(\hat{\kappa}/2)} - \frac{16\mathcal{I}_1(3\hat{\kappa}/2)}{\mathcal{I}_0(\hat{\kappa}/2)\mathcal{I}_0(\hat{\kappa})} + \frac{24\mathcal{I}_1(2\hat{\kappa})}{\mathcal{I}_0^2(\hat{\kappa})}\right\}^d \frac{\hat{f}(\theta;\hat{\kappa})}{n}.$$

Remark 2. *The use of above variance estimators makes confidence intervals very easy to implement, but notice that if the estimator $\hat{f}^{(\boldsymbol{r})}$ is significantly biased at $\boldsymbol{\theta}$, the bias will affect not only the location, but also the width of the interval, yielding very poor coverage rates. In addition, when a higher order kernel is employed, we could occasionally have negative variance estimates, especially with small samples. For our twicing this has happened very seldom because our smallest sample size is $n = 100$.*

Concerning the selection of the smoothing parameter κ, we have applied the LCV criterion for the case $|\boldsymbol{r}| = 0$, and, for the case $\boldsymbol{r} = \mathbf{1}$, the von Mises reference rule using (7.5) with the $\hat{\nu}_s$s being maximum likelihood estimates of the ν_ss. Our performance indicators are the average coverage $\bar{c}$ and the average width $\bar{w}$, constructed as follows. Let $\boldsymbol{\theta}_i = 2\pi(i-1)\mathbf{1}/350, i = 1, \ldots, 350$ be a set of equispaced points in $\mathbb{T}^d$. From our 2000 samples we obtain confidence intervals, with c_i and w_i indicating the observed coverage and the median width at $\boldsymbol{\theta}_i$ respectively. Now consider the weights $P_i = f(\boldsymbol{\theta}_i)/\sum_{j=1}^{350} f(\boldsymbol{\theta}_j)$, $i = 1, \ldots, 350$, then $\bar{c} = \sum_{i=1}^{350} c_i \times P_i$ and $\bar{w} = \sum_{i=1}^{350} w_i \times P_i$. The results of the simulation study are reported in Tables 9.3 and 9.4. For $d = 1$ both of the estimators give reasonable results when estimating the density — see Table 9.3 — but when coming at derivatives, the larger bias heavily affects the performance of KDE even in one dimension. However, in both cases, for higher dimensions, when the bias problem becomes more severe due to curse of dimensionality, the standard estimator gives very poor performance, whereas twicing still assures reasonable coverages provided that a big enough sample size is employed.

The good performance of twicing method with LCV criterion are due to the fact that twicing eliminates the bias due to the oversmoothing involved by LCV, whilst this latter reduces the variance inflation coming from twicing procedure.

Concerning the use of different selection criteria, we have noted that UCV and BCV algorithms give similar coverages for TWKDE, while for KDE the coverages are a little improved especially for $d = 3$. Unfortunately, these algorithms undersmooth very often, usually leading to intervals that are much wider than those of LCV.

		$n = 100$		$n = 300$		$n = 1000$	
		KDE	TWKDE	KDE	TWKDE	KDE	TWKDE
$d = 1$	$\bar{c}$	0.925	0.952	0.926	0.961	0.927	0.961
	$\bar{w}$	0.109	0.132	0.708	0.855	0.441	0.531
$d = 2$	$\bar{c}$	0.621	0.943	0.636	0.957	0.637	0.957
	$\bar{w}$	0.027	0.042	0.019	0.029	0.013	0.019
$d = 3$	$\bar{c}$	0.304	0.867	0.291	0.939	0.285	0.951
	$\bar{w}$	0.006	0.013	0.005	0.010	0.003	0.007

Table 9.3: Confidence intervals at level $1 - \alpha = 0.95$ for various sample sizes, dimensions and methods. $\bar{c}$= average coverage; $\bar{w}$= average width; KDE=standard kernel method; TWKDE=twicing kernel method; d=data dimension.

		$n = 100$		$n = 300$		$n = 1000$	
		KDE	TWKDE	KDE	TWKDE	KDE	TWKDE
$d = 1$	$\bar{c}$	0.843	0.936	0.856	0.948	0.869	0.950
	$\bar{w}$	0.164	0.252	0.123	0.187	0.0890	0.133
$d = 2$	$\bar{c}$	0.721	0.935	0.731	0.941	0.723	0.950
	$\bar{w}$	0.040	0.100	0.035	0.085	0.029	0.069
$d = 3$	$\bar{c}$	0.403	0.882	0.393	0.942	0.387	0.949
	$\bar{w}$	0.011	0.025	0.008	0.019	0.006	0.014

Table 9.4: Confidence intervals for the first derivative along each dimension at level $1 - \alpha = 0.95$ for various sample sizes, dimensions and methods. $\bar{c}$= average coverage; $\bar{w}$= average width; KDE=standard kernel method; TWKDE=twicing kernel method; d=data dimension.

9.5 Software

For the real data example we used the `optimize` function in R (R Development Core Team, 2010) to locate the minimum of the leave-one-out cross-validation function in the range $\kappa \in (1, 180)$. We made use of code from the library `CircStats` (CircStats, 2007).

For simulations we have used MATLAB R2010a. In particular, the random samples have been drawn by using the command `circ_vmrnd`, while the maximum likelihood estimates of ν_s in the von Mises reference rules for derivatives estimation have been carried out by the command `circ_kappa`. These commands are available in the freeware toolbox CIRCSTAT written by Berens (2009). The optimizations were carried out by using the function `fmincon` in the OPTIMIZATION toolbox with 0.1 as the starting value, and with the non-negativity constraint inserted. Note that `fminunc` and `fminsearch` give the same answers as `fmincon`, but, as expected, were significantly slower.

Appendix

Proof of Lemma 1 If j is odd, then $\sin^j(\theta)$ is orthogonal in $L^1(\mathbb{T})$ to each function in $\{1/2, \cos(\theta), \cos(2\theta), \cdots\}$, which implies that $\eta_j(K_\kappa) = 0$. If $j > 0$ is even, $\sin^j(\theta)$ is not orthogonal in $L^1(\mathbb{T})$ to 1/2 and to the set $\{\cos(2s),\ 0 < s \leq j/2\}$, and in particular one has

$$\int_{\mathbb{T}} \frac{\sin^j(\theta)}{2} d\theta = \binom{j-1}{j/2} \frac{\pi}{2^{j-1}} \quad \text{and} \quad \int_{\mathbb{T}} \sin^j(\theta)\cos(2s\theta) d\theta = \binom{j}{j/2+s} \frac{(-1)^{j+s}\pi}{2^{j-1}},$$

which gives

$$\eta_j(K_\kappa) = \frac{1}{2^{j-1}} \left\{ \binom{j-1}{j/2} + \sum_{s=1}^{j/2} (-1)^{j+s} \binom{j}{j/2+s} \gamma_{2s}(\kappa) \right\}. \tag{10.1}$$

Now observe that if K_κ has sin-order q, then $\gamma_j(\kappa) = 1$ for each $j < q$. Finally, recall that $\lim_{\kappa\to\infty} \gamma_q(\kappa) = 1$. □

Proof of Lemma 2 Conditions *i*), and *iii*) of Definition 1 hold for $\mathcal{K}_{\kappa,\ell}$ by construction. To prove condition *ii*) use the arguments of Lejeune and Sarda (1992). Finally, to verify condition *iv*), observe that

$$\begin{aligned}\int_{\mathbb{T}} |\mathcal{K}_{\kappa,\ell}(\theta)| d\theta &= \int_{\mathbb{T}} \left| \frac{\sum_{(j_1,\cdots,j_{\ell+1})} (-1)^{\aleph(j_1,\cdots,j_{\ell+1})} \prod_{i=1}^{\ell+1} u_{ij_i}}{\sum_{(j_1,\cdots,j_{\ell+1})} (-1)^{\aleph(j_1,\cdots,j_{\ell+1})} \prod_{i=1}^{\ell+1} w_{ij_i}} \right| K_\kappa(\theta) d\theta \\ &\leq \sum_{(j_1,\cdots,j_{\ell+1})} \frac{\left| \prod_{i=1}^{\ell+1} w_{ij_i} \right|}{\left| \sum_{(j_1,\cdots,j_{\ell+1})} (-1)^{\aleph(j_1,\cdots,j_{\ell+1})} \prod_{i=1}^{\ell+1} w_{ij_i} \right|},\end{aligned}$$

where summations are taken over all the permutations $(j_1, \cdots, j_{\ell+1})$ of the integers $(1, 2, \cdots, \ell+1)$, and $\aleph(j_1, \cdots, j_{\ell+1})$ denotes the number of inversions of the permutation. Then, since as κ increases, according to Lemma 1, w_{i,j_i} tends to 0, each summand in the last line of the above equation tends to 0 or to a positive number. □

Proof of Theorem 1 First observe that by Parseval's identity $\int_{\mathbb{T}^d}\{f(\boldsymbol{\theta})\}^2 d\boldsymbol{\theta} = (2\pi)^{-d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}||c_{\boldsymbol{j}}||^2$, where $||g||$ stands for the L_2 norm of g. Then use the results in Proposition 1 and in Proposition 2, the identities $\mathsf{E}[\tilde{c}_{\boldsymbol{j}}] = c_{\boldsymbol{j}}$, $\mathsf{E}[||\tilde{c}_j - c_{\boldsymbol{j}}||^2] = n^{-1}(1-||c_{\boldsymbol{j}}||^2)$, and some algebraic manipulations, to get

$$\begin{aligned}\mathsf{E}\left[\int_{\mathbb{T}^d}\left\{\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C}) - f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 d\boldsymbol{\theta}\right] &= \mathsf{E}\left[\frac{1}{(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left|\left|i^{|\boldsymbol{r}|}\tilde{c}_{\boldsymbol{j}}\gamma_{\boldsymbol{j}}(\boldsymbol{C})\boldsymbol{j}^{\boldsymbol{r}} - i^{|\boldsymbol{r}|}c_{\boldsymbol{j}}\boldsymbol{j}^{\boldsymbol{r}}\right|\right|^2\right]\\ &= \frac{i^{2|\boldsymbol{r}|}}{(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left(\mathsf{E}\left[\|\tilde{c}_{\boldsymbol{j}} - c_{\boldsymbol{j}}\|^2\right]\gamma_{\boldsymbol{j}}^2(\boldsymbol{C})\boldsymbol{j}^{2\boldsymbol{r}} + \{1-\gamma_{\boldsymbol{j}}(\boldsymbol{C})\}^2\|c_{\boldsymbol{j}}\|^2\boldsymbol{j}^{2\boldsymbol{r}}\right)\\ &= \frac{i^{2|\boldsymbol{r}|}}{n(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\left(1-\|c_{\boldsymbol{j}}\|^2\right)\gamma_{\boldsymbol{j}}^2(\boldsymbol{C})\boldsymbol{j}^{2\boldsymbol{r}} + \frac{i^{2|\boldsymbol{r}|}}{(2\pi)^d}\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\{1-\gamma_{\boldsymbol{j}}(\boldsymbol{C})\}^2\|c_{\boldsymbol{j}}\|^2\boldsymbol{j}^{2\boldsymbol{r}}.\end{aligned}$$

Finally, observe that $c_{\boldsymbol{j}} = \sum_{m=1}^{2^{d-1}}\alpha_{\boldsymbol{j},m} + i\sum_{m=1}^{2^{d-1}}\beta_{\boldsymbol{j},m}$, where $\alpha_{\boldsymbol{j},m}$ and $\beta_{\boldsymbol{j},m}$ denote the coefficients in the trigonometric Fourier series expansion of f whose basis contains an even and an odd number of sin functions, respectively. □

Proof of Theorem 2 For the bias term, first observe that

$$\begin{aligned}\int_{\mathbb{T}^d}\left\{\mathsf{E}[\hat{f}^{(\boldsymbol{r})}(\boldsymbol{\theta};\boldsymbol{C})] - f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 d\boldsymbol{\theta} &= \int_{\mathbb{T}^d}\left\{\int_{\mathbb{T}^d}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\alpha}-\boldsymbol{\theta})f(\boldsymbol{\alpha})d\boldsymbol{\alpha} - f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 d\boldsymbol{\theta}\\ &= \int_{\mathbb{T}^d}\left\{\int_{\mathbb{T}^d}K_{\boldsymbol{C}}(\boldsymbol{u})\left[f^{(\boldsymbol{r})}(\boldsymbol{\theta}+\boldsymbol{u}) - f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right]d\boldsymbol{u}\right\}^2 d\boldsymbol{\theta},\end{aligned}$$

then for $\boldsymbol{u} = \{u_1,\cdots,u_d\}\in\mathbb{T}^d$, $\tau\in[0,1]$, since for small α $\sin\alpha\approx\alpha$, letting $\boldsymbol{S_u} := \{\sin(u_1),\cdots,\sin(u_d)\}^T$ and recalling assumption *ii*), we use the expansion

$$f^{(\boldsymbol{r})}(\boldsymbol{\theta}+\boldsymbol{u}) = f^{(\boldsymbol{r})}(\boldsymbol{\theta}) + \sum_{p=1}^{q}\frac{(\boldsymbol{S}_{\boldsymbol{u}}^{\mathsf{T}})^{\otimes p}}{p!}\mathsf{vec}\left[\frac{\mathsf{d}^p f^{(\boldsymbol{r})}(\boldsymbol{\theta}+\tau\boldsymbol{\delta}_p)}{\mathsf{d}\boldsymbol{\theta}^p}\right]$$

where $\boldsymbol{A}^{\otimes p}$ denotes the pth Kronecker power of $\boldsymbol{A}$, $\mathsf{d}^p f^{(\boldsymbol{r})}(\boldsymbol{\alpha})/\mathsf{d}\boldsymbol{\alpha}^p$ is the matrix derivative of order p of $f^{(\boldsymbol{r})}$ at $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}_p = \boldsymbol{u}$ if $p = q$, $\boldsymbol{\delta}_p = \boldsymbol{0}$ otherwise.

Now recalling that $K_{\boldsymbol{C}}$ has sin-order q, and using the generalized Minkowski inequality and the assumption *iii*) we get

$$\begin{aligned}\int_{\mathbb{T}^d}\left\{\mathsf{E}[\hat f^{(\boldsymbol r)}(\boldsymbol\theta;\boldsymbol C)]-f^{(\boldsymbol r)}(\boldsymbol\theta)\right\}^2 d\boldsymbol\theta &= \int_{\mathbb{T}^d}\left\{\int_{\mathbb{T}^d} K_{\boldsymbol C}(\boldsymbol u)\frac{(\boldsymbol S_{\boldsymbol u}^{\mathsf{T}})^{\otimes q}}{q!}\mathsf{vec}\left[\frac{\mathsf{d}^q f^{(\boldsymbol r)}(\boldsymbol\theta+\tau\boldsymbol u)}{\mathsf{d}\boldsymbol\theta^q}\right]d\boldsymbol u\right\}^2 d\boldsymbol\theta\\ &\le \left\{\int_{\mathbb{T}^d} K_{\boldsymbol C}(\boldsymbol u)\frac{(\boldsymbol S_{\boldsymbol u}^{\mathsf{T}})^{\otimes q}}{q!}\left[\int_{\mathbb{T}^d}\left(\mathsf{vec}\left[\frac{\mathsf{d}^q f^{(\boldsymbol r)}(\boldsymbol\theta+\tau\boldsymbol u)}{\mathsf{d}\boldsymbol\theta^q}\right]\right)^2 d\boldsymbol\theta\right]^{1/2} d\boldsymbol u\right\}^2\\ &\le \left\{\frac{\eta_q(K_\kappa)}{q!}Ld\right\}^2.\end{aligned}$$

Concerning the variance term, let $\Gamma_i(\boldsymbol\theta)=K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\Theta_i-\boldsymbol\theta)-\mathsf{E}[K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\Theta_i-\boldsymbol\theta)]$, then the Γ_is, $i=1,\cdots,d$, are *i.i.d.* random variables whit mean 0 and variance $\mathsf{E}[\Gamma_i^2(\Theta_i)]\le\mathsf{E}[\{K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\Theta_i-\boldsymbol\theta)\}^2]$. Hence, for all $\boldsymbol\theta\in\mathbb{T}^d$

$$\mathsf{Var}[\hat f^{(\boldsymbol r)}(\boldsymbol\theta;\boldsymbol C)]=\frac{1}{n}\mathsf{E}[\{\Gamma_i(\boldsymbol\theta)\}^2]\le\frac{1}{n}\mathsf{E}[\{K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\Theta_i-\boldsymbol\theta)\}^2].$$

Consequently,

$$\begin{aligned}\int_{\mathbb{T}^d}\mathsf{Var}[\hat f^{(\boldsymbol r)}(\boldsymbol\theta;\boldsymbol C)]d\boldsymbol\theta &\le \frac{1}{n}\int_{\mathbb{T}^d}\left[\int_{\mathbb{T}^d}\{K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\alpha-\boldsymbol\theta)\}^2 f(\boldsymbol\alpha)d\boldsymbol\alpha\right]d\boldsymbol\theta\\ &= \frac{1}{n}\int_{\mathbb{T}^d} f(\boldsymbol\alpha)\left[\int_{\mathbb{T}^d}\{K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\alpha-\boldsymbol\theta)\}^2 d\boldsymbol\theta\right]d\boldsymbol\alpha\\ &= \frac{1}{n}\int_{\mathbb{T}^d}\{K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol u)\}^2 d\boldsymbol u. \qquad\square\end{aligned}$$

Proof of Theorem 4 Put $\boldsymbol S_{\boldsymbol\alpha}:=\{\sin(\alpha_1),\cdots,\sin(\alpha_d)\}^{\mathsf{T}}$, $\boldsymbol\Omega_p:=\mathsf{diag}\{\eta_p(\kappa_1),\cdots,\eta_p(\kappa_d)\}$ and $X_\ell=K_{\boldsymbol C}^{(\boldsymbol r)}(\boldsymbol\theta-\boldsymbol\Theta_\ell)$, $\ell=1,\cdots,n$. Now, $X_1,\cdots,X_n$ are *i.i.d.*, and recalling assumptions *i*) and Lemma 1, a change of variables leads to

$$\begin{aligned}\mathsf{E}[X_1] &= K_{\boldsymbol C}^{(\boldsymbol r)} * f = K_{\boldsymbol C} * f^{(\boldsymbol r)}\\ &\sim \int_{\mathbb{T}^d} K_{\boldsymbol C}(\boldsymbol u)\left\{f^{(\boldsymbol r)}(\boldsymbol\theta)+\sum_{p=1}^{q}\frac{(\boldsymbol S_{\boldsymbol u}^{\mathsf{T}})^{\otimes p}}{p!}\mathsf{vec}\left[\frac{\mathsf{d}^p f^{(\boldsymbol r)}(\boldsymbol\theta)}{\mathsf{d}\boldsymbol\theta^p}\right]+O\left(\frac{(\boldsymbol S_{\boldsymbol u}^{T})^{\otimes(q+1)}}{(q+1)!}\right)\right\}d\boldsymbol u\\ &= f^{(\boldsymbol r)}(\boldsymbol\theta)+\frac{1}{q!}\mathsf{tr}\left\{\boldsymbol\Omega_q\frac{\mathsf{d}^q f^{(\boldsymbol r)}(\boldsymbol\theta)}{\mathsf{d}\boldsymbol\theta^q}\right\}+o\left(2^{1-q}\{1-\gamma_q(\kappa_s)\right).\end{aligned}$$

We have used the expansion in squared brackets because, due to points *i*) and *iii*) of Definition 3.1, the integrand is non-zero over $[-\lambda,\lambda]^d$ where $\lim_{\kappa\to\infty}\lambda=0$, and therefore, for each $s = 1, \cdots, d$, u_s can be considered an element of a sequence approaching zero as κ increases.

For the variance, recalling Lemma 3, we obtain

$$\begin{aligned}\mathsf{Var}[X_1] &= \int_{\mathbb{T}^d} \left\{K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\beta}-\boldsymbol{\theta})\right\}^2 f(\boldsymbol{\beta})d\boldsymbol{\beta} - \{\mathsf{E}\left[X_1\right]\}^2 \\ &\sim \int_{\mathbb{T}^d} \left\{K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{u})\right\}^2 \left[f(\boldsymbol{\theta}) + O\left(\boldsymbol{S}_{\boldsymbol{u}}^T\right)\right] d\boldsymbol{u} - \left\{f^{(\boldsymbol{r})}(\boldsymbol{\theta}) + O\left(2^{1-q}\{1-\gamma_q(\kappa_s)\}\right)\right\}^2 \\ &= f(\boldsymbol{\theta}) \prod_{s=1}^{d} Q_{\kappa_s}(r_s) - \left\{f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 + O\left(2^{1-q}\{1-\gamma_q(\kappa_s)\}\right) .\end{aligned}$$

Now note that, under conditions $i)$ and $ii)$, for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathsf{E}\left[X_1^2 \mathbb{1}_{\left\{|X_1 - \mathsf{E}[X_1]| > \epsilon\sqrt{n\prod Q_{\kappa_s}(r_s)}\right\}}\right] = 0,$$

and by applying Lindeberg's central limit theorem the result directly follows. □

Proof of Theorem 6 Firstly observe that

$$\hat{\psi}_r(\boldsymbol{C}) = n^{-1}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) + n^{-2}\sum_{\ell\neq\mu}\sum K_{\boldsymbol{C}}^{(\boldsymbol{r})}\left(\boldsymbol{\Theta}_\ell - \boldsymbol{\Theta}_\mu\right),$$

and hence

$$\mathsf{E}\left[\hat{\psi}_r(\boldsymbol{C})\right] = n^{-1}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) + (1-n^{-1})\mathsf{E}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}\left(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\right)\right].$$

Then, using the expansion in the proof of Theorem 4 with $f^{(\boldsymbol{r})}$ replaced by f, a change of variable leads to

$$\mathsf{E}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\right] \sim \psi_{\boldsymbol{r}} + \frac{1}{q!}\int_{\mathbb{T}^d} \mathsf{tr}\left\{\boldsymbol{\Omega}_q \frac{\mathsf{d}^q f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\} f^{(\boldsymbol{r})}(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1),$$

and hence

$$\mathsf{E}\left[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})\right] - \psi_{\boldsymbol{r}} \sim n^{-1}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) + \frac{1}{q!}\int_{\mathbb{T}^d} \mathsf{tr}\left\{\boldsymbol{\Omega}_q \frac{\mathsf{d}^q f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^q}\right\} f^{(\boldsymbol{r})}(\boldsymbol{\theta})d\boldsymbol{\theta} + o(1).$$

To derive the variance, we firstly observe that

$$\mathsf{Var}\left[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})\right] = \frac{2(n-1)}{n^3}\mathsf{Var}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\right] + \frac{4(n-1)(n-2)}{n^3}\mathsf{Cov}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2), K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_3)\right]. \tag{10.2}$$

By considering each component of (10.2) in turn, by a change of variable and recalling Lemma 3, we first obtain

$$\mathsf{E}\left[\left\{K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\right\}^2\right] = \int_{\mathbb{T}^d}\int_{\mathbb{T}^d} \left\{K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\beta}-\boldsymbol{\theta})\right\}^2 f(\boldsymbol{\beta})f(\boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta} = \psi_{\boldsymbol{0}}\prod_{s=1}^{d} Q_{\kappa_s}(r_s),$$

while

$$\begin{aligned}\mathsf{E}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_2-\boldsymbol{\Theta}_3)\right] &= \int_{\mathbb{T}^d}\int_{\mathbb{T}^d}\int_{\mathbb{T}^d} K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\beta}-\boldsymbol{\theta})K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\theta}-\boldsymbol{\lambda})f(\boldsymbol{\beta})f(\boldsymbol{\theta})f(\boldsymbol{\lambda})d\boldsymbol{\beta}d\boldsymbol{\theta}d\boldsymbol{\lambda}\\ &= \int_{\mathbb{T}^d}\int_{\mathbb{T}^d}\int_{\mathbb{T}^d} K_{\boldsymbol{C}}(\boldsymbol{u})K_{\boldsymbol{C}}(\boldsymbol{v})f^{(\boldsymbol{r})}(\boldsymbol{\theta}+\boldsymbol{u})f(\boldsymbol{\theta})f^{(\boldsymbol{r})}(\boldsymbol{\theta}-\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{\theta}\\ &\sim \int_{\mathbb{T}^d}\left\{f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 f(\boldsymbol{\theta})d\boldsymbol{\theta}+o(1).\end{aligned}$$

Hence, using $\mathsf{E}[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\Theta}_1-\boldsymbol{\Theta}_2)]=\psi_{\boldsymbol{r}}+o(1)$, we finally get

$$\mathsf{Var}[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})] \sim \frac{2}{n^2}\psi_{\boldsymbol{0}}\prod_{s=1}^{d} Q_{\kappa_s}(r_s)+\frac{4}{n}\left[\int_{\mathbb{T}^d}\left\{f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 f(\boldsymbol{\theta})d\boldsymbol{\theta}-\psi_{\boldsymbol{r}}^2\right]+o(1).$$

Concerning the asymptotic distribution, first observe that the estimator in (3.6) is a V-statistic of order 2, then note that $\mathcal{E}_1(\boldsymbol{\theta}_1):=\mathsf{E}\left[K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\theta}_1-\boldsymbol{\Theta}_2)\right]=\int_{\mathbb{T}^d}K_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)f(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2 = K_{\boldsymbol{C}}^{(\boldsymbol{r})}*f$, and hence that $\mathcal{E}_1(\boldsymbol{\Theta}_1)$ is not degenerate, then apply the result in Section 5.7.3 of Serling (1980). □

Proof of Theorem 7 Observe that the von Mises kernel is a second sin-order toroidal kernel with $\eta_2(V_\kappa)=\mathcal{I}_1(\kappa)/\{\kappa\mathcal{I}_0(\kappa)\}$, and use Corollary 1 to get

$$\mathsf{AMISE}\left[\hat{f}(\cdot;\boldsymbol{C})\right]=\frac{1}{4}\left\{\frac{\mathcal{I}_1(\kappa)}{\kappa\mathcal{I}_0(\kappa)}\right\}^2\int_{\mathbb{T}^d}\mathsf{tr}^2\left\{\frac{\mathsf{d}^2 f^{(\boldsymbol{r})}(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\}d\boldsymbol{\theta}+\frac{1}{n}R\left(V_{\boldsymbol{C}}^{(\boldsymbol{r})}\right). \tag{10.3}$$

Now, replace $\mathcal{I}_1(\kappa)/\mathcal{I}_0(\kappa)$ by 1 with an error of magnitude $O\left(\kappa^{-1}\right)$, and notice that for a big enough κ

$$R\left(V_{\boldsymbol{C}}^{(\boldsymbol{r})}\right)\approx\prod_{s=1}^{d}\frac{\mathsf{OF}(2r_s)\kappa^{(2r_s+1)/2}}{2^{r_s+1}\pi^{1/2}} \tag{10.4}$$

and minimize the RHS of (10.3). □

Proof of Theorem 8 Observe that for a big enough κ $\eta_q(\mathcal{V}_{\kappa,\ell})\approx -i^q\eta_q(V_\kappa)$,and $R\left(\mathcal{V}_{\boldsymbol{C},\ell}\right)\approx(q-1)^{d/2}R(V_{\boldsymbol{C}})$, where $\eta_q(V_\kappa)\approx\frac{\mathsf{OF}(q)}{\kappa^{q/2}}$ and $R(V_{\boldsymbol{C}})\approx\kappa^{d/2}(4\pi)^{-d/2}$, then reason as in the proof of Theorem 7. □

Proof of Theorem 9 First of all notice that, since $O\left(\eta_2(V_\kappa^t)\right)=O\left(\eta_4(V_\kappa^t)\right)>O\left(\eta_{2s+2}(V_\kappa^t)\right)$, $s\geq 2$, the bias term in the AMISE formula can be derived by considering the expansion of f in the proof of Theorem 3 with matrix derivatives up to order 4, to get

$$\mathsf{E}[\hat{f}(\boldsymbol{\theta};\boldsymbol{C})]-f(\boldsymbol{\theta})\sim\frac{\eta_2(V_\kappa^t)}{2}\mathsf{tr}\left\{\frac{\mathsf{d}^2 f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\}+\frac{\eta_4(V_\kappa^t)}{4!}\mathsf{tr}\left\{\frac{\mathsf{d}^4 f(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^4}\right\}+O\left(\eta_2^2(V_\kappa^t)\right).$$

Finally observe that for a big enough κ $\eta_2\left(V_\kappa^t\right) \approx k^{-2}$, $\eta_4\left(V_\kappa^t\right) \approx -6k^{-2}$ and $R\left(V_{\boldsymbol{C}}^t\right) \approx 2^{d/2}R\left(V_{\boldsymbol{C}}\right)$, then reason as in the proof of Theorem 7. □

Proof of Theorem 10 Recall that for the von Mises kernel $\eta_2(V_\kappa) = \mathcal{I}_1(\kappa)/\{\kappa\mathcal{I}_0(\kappa)\}$, and observe that

$$\int_{\mathbb{T}^d} \mathsf{tr}\left\{\frac{\mathsf{d}^2 f^{(\boldsymbol{r})}(\boldsymbol{\theta})}{\mathsf{d}\boldsymbol{\theta}^2}\right\} f^{(\boldsymbol{r})}(\boldsymbol{\theta})d\boldsymbol{\theta} = \sum_{s=1}^{d} \psi_{\boldsymbol{r}+2\boldsymbol{e}_s}$$

then follow the proof of Theorem 6 to get

$$\mathsf{AMSE}\left[\hat{\psi}_{\boldsymbol{r}}(\boldsymbol{C})\right] = \left[\frac{1}{n}V_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) + \frac{\mathcal{I}_1(\kappa)}{2\kappa\mathcal{I}_0(\kappa)}\sum_{s=1}^{d}\psi_{\boldsymbol{r}+2\boldsymbol{e}_s}\right]^2 + \frac{2}{n^2}\psi_{\boldsymbol{0}}R\left(V_{\boldsymbol{C}}^{(\boldsymbol{r})}\right) + \frac{4}{n}\left[\int_{\mathbb{T}^d}\left\{f^{(\boldsymbol{r})}(\boldsymbol{\theta})\right\}^2 f(\boldsymbol{\theta})d\boldsymbol{\theta} - \psi_{\boldsymbol{r}}^2\right].$$

Now, to derive the AMSE-optimal smoothing parameter, first replace $\mathcal{I}_1(\kappa)/\mathcal{I}_0(\kappa)$ by 1 with an error of magnitude $O\left(\kappa^{-1}\right)$ then, if all r_s are even, use

$$V_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) \approx i^{|\boldsymbol{r}|}(2\pi)^{-d/2}\kappa^{(|\boldsymbol{r}|+d)/2}\prod_{s=1}^{d}\mathsf{OF}(r_s)$$

which holds for a big enough κ. Finally, note that $V_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0})$ and $\psi_{\boldsymbol{r}+2\boldsymbol{e}_s}$ are of opposite sign, and take as optimal the value of κ which eliminates the first two summands in squared brackets in the AMSE equation.

If at least one r_s is odd, observe that $V_{\boldsymbol{C}}^{(\boldsymbol{r})}(\boldsymbol{0}) = 0$, use the result in (10.4), then minimize the components of AMSE depending on κ. □

References

Bai, Z.D., Rao, R.C. and Zhao, L.C. (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, **27**, 24–39.

Batschelet, E. (1981). *Circular Statistics in Biology.* Academic Press, London.

Beran, R. (1979). Exponential models for directional data. *The Annals of Statistics*, **7**, 1162–1178.

Berens, P. (2009). CircStat: A Matlab Toolbox for Circular Statistics. *Journal of Statistical Software*, **31**, Issue 10.

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

Chacón, J.E., Duong, T. and Wand, M.P. (2010). Asymptotics for general multivariate kernel density derivative estimators. *submittted.*

Chow, Y.S., Geman, S. and Wu, L.D. (1983). Consistent cross-validated density estimation. *The Annals of Statistics*, **11**, 25–38.

CircStats: Circular Statistics (2007), S-plus original by Ulric Lund and R port by Claudio Agostinelli, R package version 0.2-3.

Coles, S. (1998). Inference for circular distributions and processes. *Statistics and Computing*, **8**, 105–113.

Duin, R.P.W. (1976). On the choice of smoothing parameter for Parzen estimators of probability density functions. *IEEE trans. Compt.*, **C-25**, 1175–1179.

Duong, T., Cowling, A., Koch, I. and Wand, M.P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, **52**, 4225–4242.

Efromovich, S. (1999). *Nonparametric curve estimation.* Springer, New York.

Efromovich, S. and Pinsker, M.S. (1982). Estimation of square-integrable probability density of a random variable. *Problems of Information Trasmission*, **18**, 175–189.

Fisher, N.I. (1993). *Statistical analysis of circular data.* Cambridge University Press.

Habbema, J.D.F., Hermans, J. and Van Der Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In *COMPSTAT 1974, Proceedings in Computational Statistics, Vienna*, (G. Bruckman ed.) pages 101–110. Physica, Heidelberg.

Hall, P., Watson, G.S. and Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, **74**, 751–762.

Jammalamadaka, S.R. and SenGupta, A. (2001). *Topics in Circular Statistics.* World Scientific, Singapore.

Jones, M.C. and Pewsey, A. (2005). A family of symmetric distributions on the circle. *Journal of the American Statistical Association*, **100**, 1422–1428.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kato, S. and Jones, M.C. (2009). A family of distributions on the circle with links to, and applications arising from, möbius transformation. *Journal of the American Statistical Association*, , .

Klemelä, J. (2000). Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, **73**, 18–40.

Lejeune, M. and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, **14**, 457–471.

Loader, C.R. (1999). Bandwidth selection: classical or plug-in? *The Annals of Statistics*, **27**, 415-438.

Mardia, K.V. (1972). *Statistics of Directional Data.* Academic Press, London.

Mardia, K.V. and Jupp, P.E. (1999). *Directional Statistics.* John Wiley, New York, NY.

Mardia, K.V., Taylor, C.C. and Subramaniam, G.K. (2007). Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. *Biometrics*, **63**, 505–512.

Pewsey, A., Lewis, T. and Jones, M.C. (2007). The wrapped t family of circular distributions. *Australian & New Zealand Journal of Statistics*, **49**, 79–91.

Prakasa Rao, B.L.S. (2000). Nonparametric estimation of partial derivatives of a multivariate probability density by the method of wavelets. In M.L. Puri, editor, *Asymptotics in Statistics and Probability, Festschrift for G.G.Roussas*, pages 321–330. VSP, The Netherlands.

R: *A Language and Environment for Statistical Computing* (2010), R Development Core Team, Vienna, Austria. `http://www.R-project.org`.

Rizzo, M.L. (2002). A Test of Homogeneity for Two Multivariate Populations. Proceedings of the American Statistical Association, Physical and Engineering Sciences Section [CD-ROM], Alexandria, VA: American Statistical Association.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.

Sain, S.R., Baggerly, K.A. and Scott, D.W. (1992). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, **89**, 807–817.

Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131–1146.

Sengupta, A. and Ugwuowo, F.I. (2011). A Classification Method for Directional Data with Application to the Human Skull. *Communications in Statistics-Theory and Methods*, **40**, 457–466.

Serfling, R.J. (1980). *Approximation Theorems for Mathematical Statistics.* John Wiley, New York, NY.

Shieh, G.S., Zheng, S.R. and Shimizu, K. (2006). *A Bivariate Generalized von Mises with Applications to Circular Genomes.* Technical Report 06-06, Institute of Statistical Science, Academia Sinica.

Singh, R.S. (1976). Nonparametric estimation of mixed partial derivatives of a multivariate density. *Journal of Multivariate Analysis*, **6**, 111–122.

Stuetzle, W. and Mittal, Y. (1979). Some comments on the asymptotic behavior of robust smoothers. In *Smoothing Techniques for Curve Estimation. Proceedings, Heidelberg 1979*, Lecture Notes in Mathematics 757, pages 191–195. Springer-Verlag, Berlin.

Taylor, C.C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, **76**, 705–712.

Taylor, C.C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, **52**, 3493–3500.

(1) DMQTE, Università di Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy

E-mail: mdimarzio@unich.it ; agnesepanzera@yahoo.it

(2) Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

E-mail: charles@maths.leeds.ac.uk (corresponding author)

Tel: +44 (0)113 343 5168 Fax: +44 (0)113 343 5090