Electronic Journal of Statistics
Vol. 11 (2017) 4323–4346
ISSN: 1935-7524
DOI: 10.1214/17-EJS1318

Nonparametric estimating equations for circular probability density functions and their derivatives

Marco Di Marzio and Stefania Fensore

DMQTE,
Università di Chieti-Pescara,
Viale Pindaro 42, 65127 Pescara, Italy
e-mail: mdimarzio@unich.it; stefania.fensore@unich.it

Agnese Panzera

DiSIA,
Università di Firenze,
Viale Morgagni 59, 50134 Firenze, Italy
e-mail: a.panzera@disia.unifi.it

and

Charles C. Taylor

Departement of Statistics,
University of Leeds,
Leeds LS2 9JT, UK
e-mail: c.c.taylor@leeds.ac.uk

Abstract: We propose estimating equations whose unknown parameters are the values taken by a circular density and its derivatives at a point. Specifically, we solve equations which relate local versions of population trigonometric moments with their sample counterparts. Major advantages of our approach are: higher order bias without asymptotic variance inflation, closed form for the estimators, and absence of numerical tasks. We also investigate situations where the observed data are dependent. Theoretical results along with simulation experiments are provided.

Keywords and phrases: Circular kernels, density estimation, Fourier coefficients, jackknife, sin-polynomials, trigonometric moments, von Mises density.

Received July 2016.

1. Introduction

Circular data occur when the sample space is described by a circle, as opposed to the real line in standard statistics. They often arise in biology (migration paths of animals), meteorology (wind and marine current directions), and geology (orientations of joints and faults, landforms, oriented stones). Other examples include phenomena that are periodic in time, like daily and seasonal socioeconomic phenomena. If compared to a linear scale, a circular one has special features. Firstly, the beginning and the end of the measurement scale coincide, and their common location is called the *origin* (or *zero direction*). This latter is arbitrary, and likewise any designation of median, high or low values. Also note that for circular data there is no concept of standardisation. The only data transformations which do not change the relative positions of the observations are rotation and reflection.

Nonparametric *circular* density estimation could be seen as a non-mature research field. Although basic kernel density estimation is well known (see, for example, [6], [2], [8] and [14]), not much has been written on more sophisticated methods, aimed at bias reduction. This would be useful for efficient point estimation in the cases of heavy density tails, multi-modality or asymmetry. Additionally, it would improve the precision in estimating a confidence interval for the value of the density at a point, or confidence bands for the whole density. Indeed, it is well known that the bias of nonparametric methods leads to incorrect centring of the confidence interval with severe consequences on coverage probability. In the Euclidean setting small bias methods have been suggested as a remedy, for example, by [1] and [5].

In this paper we propose estimating equations for circular density estimation (and its derivatives) where local versions of population trigonometric moments are equated with their empirical counterparts. We model the unknown density by a periodic series expansion whose coefficients are the system variables. Therefore, we estimate the value of functions at a point, as opposed to the classical method of moments which is used for global parameters. It will be seen that modelling via longer series will give smaller asymptotic bias without asymptotic variance inflation. The system is linear and has a closed-form solution, so our estimators have a general formulation depending on both expansion degree and the order of the target derivative.

Concerning the novelty of our contribution, we note that polynomial-like estimators have never been obtained by explicitly starting from population moments, as we do in the present paper, but rather from least squares or likelihood scores. A reason for this could be that in the Euclidean setting the existence of whatever population moment constitutes a strict assumption, whereas for differentiable circular densities this is not the case, due to the compactness of the sample space. As a link to previous work, we note that *classic* method of moments has been applied for parameter estimation by [12] when data come from a mixture of von Mises populations and the end is to separate the two components of the mixture. Additional novelty lies in the introduction of simple estimators of circular density derivatives. A nice collection of fundamental problems where derivative estimation is central can be found in [11]. We simply remark that higher order derivatives play a key rôle in the inference on modes and density functionals. Finally, this article also contains some theory for the case of dependent data. Although most circular real datasets comprise dependent data, we note that nonparametric density estimation based on circular, dependent data seems, at the moment of writing, a totally unexplored field in circular statistics.

In Section 2 we obtain the estimators. Some interpretative reasoning connect them to previous Euclidean literature based on quite different ideas. In Section 3 we first derive general asymptotic properties, then we focus on von Mises kernel theory. Finally, in Section 4 we show how the method performs, using simulation experiments and real data case studies, also through a comparison with other methods existing in the literature.

2. The estimators

The ℓth *trigonometric moment (about the zero direction)* of the absolutely continuous distribution function F of a circular random variable Θ is defined as $\mathsf{a}_\ell + \mathrm{i}\mathsf{b}_\ell$, $\ell \in \mathbb{N}$, where $\mathrm{i}^2 = -1$, and a_ℓ and b_ℓ stand for the ℓth cosine and sine moments respectively, i.e.

$$\mathsf{a}_\ell := E[\cos(\ell\Theta)] = \int_{-\pi}^{\pi} \cos(\ell\alpha) dF(\alpha), \qquad \mathsf{b}_\ell := E[\sin(\ell\Theta)] = \int_{-\pi}^{\pi} \sin(\ell\alpha) dF(\alpha).$$

Any distribution on the circle is determined by its characteristic function, and this uniqueness property, differently from the Euclidean setting, assures that they are determined by their moments. Assuming that F admits continuous derivatives up to order $p+1$ at $\theta \in [-\pi, \pi)$, $p \in \mathbb{N}$, then $f(\alpha) := F^{(1)}(\alpha)$ can be approximated by the sin-polynomial

$$\tilde{f}_p(\alpha; \theta) := \sum_{j=0}^{p} \frac{f^{(j)}(\theta) \sin^j(\alpha - \theta)}{j!}$$

with negligible error if $|\alpha - \theta|$ is small. We can interpret $\tilde{f}_p(\alpha; \theta)$ as a (truncated) Taylor-like expansion suited for functions defined on a circle because its increment has periodic nature. In fact, the sine function preserves the sign of $\alpha - \theta$, and also takes small values when α and θ are separated by a very small arc. Notice that, if the arc contains the origin, this latter property does not hold for the simple difference.

Nonparametric estimating equations could be obtained by matching sample moments to their corresponding *approximated* theoretical ones given by

$$\int_{-\pi}^{\pi} \cos(\ell\alpha) \tilde{f}_p(\alpha; \theta) d\alpha, \qquad \int_{-\pi}^{\pi} \sin(\ell\alpha) \tilde{f}_p(\alpha; \theta) d\alpha,$$

with density derivatives at $\theta \in [-\pi, \pi)$, from order 0 up to order p, which constitute the system variables. However, recalling that $f(\alpha)$ is *satisfyingly* approximated by $\tilde{f}_p(\alpha; \theta)$ only if α belongs to a narrow neighbourhood of θ, and that, on the other hand, empirical moments estimate the *exact* ones, we see that in this strategy the sides of the equations are doomed to be rather different. A way to alleviate this drawback is to use new moment expressions, where the integrands are weighted by uni-modal, symmetric density functions centred at θ, as in the formal

Definition 2.1. Let K_κ denote a generic circular kernel with concentration parameter $\kappa \in (0, \infty)$ (see [2]). For $\ell \in \mathbb{N}$, we define the ℓth *local* trigonometric (resp. cosine and sine) moments, at $\theta \in [-\pi, \pi)$, of a circular distribution function F the quantities

$$\mathrm{a}_\ell(\theta) := \int_{-\pi}^{\pi} K_\kappa(\alpha-\theta)\cos(\ell\alpha)f(\alpha)d\alpha, \quad \mathrm{b}_\ell(\theta) := \int_{-\pi}^{\pi} K_\kappa(\alpha-\theta)\sin(\ell\alpha)f(\alpha)d\alpha.$$

The idea is that above local moments are essentially determined by the values taken by integrands over a *tight* interval centred at θ, where approximation $\tilde{f}_p$ is reliable. Tightness inversely depends on the magnitude of concentration parameter κ. As a simple parallel with Euclidean kernel density estimation, we would say that κ moves toward the same direction as the inverse of the bandwidth.

Then, our simultaneous equations would be based on quantities reported in

Definition 2.2. Let K_κ denote a generic circular kernel with concentration parameter $\kappa \in (0, \infty)$. We define *approximated* local trigonometric (resp. cosine and sine) moments of order $\ell \in \mathbb{N}$ at $\theta \in [-\pi, \pi)$ of a circular distribution with density f the quantities

$$\sum_{j=0}^{p} f^{(j)}(\theta)c_\ell^j(\theta) \quad \text{and} \quad \sum_{j=0}^{p} f^{(j)}(\theta)s_\ell^j(\theta), \tag{2.1}$$

where

$$c_\ell^j(\theta) := \int_{-\pi}^{\pi} K_\kappa(\alpha-\theta)\cos(\ell\alpha)\frac{1}{j!}\sin^j(\alpha-\theta)d\alpha,$$

and

$$s_\ell^j(\theta) := \int_{-\pi}^{\pi} K_\kappa(\alpha-\theta)\sin(\ell\alpha)\frac{1}{j!}\sin^j(\alpha-\theta)d\alpha.$$

Letting $\Theta_1, \dots, \Theta_n$ be a random sample of angles from the unknown density f, sample local trigonometric (resp. cosine and sine) moments at θ are defined as

$$\mathcal{C}_\ell(\theta) := \frac{\sum_{i=1}^n K_\kappa(\Theta_i-\theta)\cos(\ell\Theta_i)}{n} \quad \text{and} \quad \mathcal{S}_\ell(\theta) := \frac{\sum_{i=1}^n K_\kappa(\Theta_i-\theta)\sin(\ell\Theta_i)}{n}.$$

Notice that, for fixed n and θ, a trade-off in selecting the value of parameter κ arises. In fact, on the population side of above definition, we have observed that a big κ ensures accuracy of expansion. On the other hand, sample quantities require maximize the amount of data effectively participating to the estimation process, and this is produced through a small κ. This suggests to formulate κ as an increasing function of n.

Regarding asymptotic accuracy, other than the obvious convergence of the sums to integrals, we would require that quantities in Equation (2.1) satisfyingly approximate $\mathrm{a}_\ell(\theta)$ and $\mathrm{b}_\ell(\theta)$, respectively. This is easily seen to be the case if the concentration parameter κ of the kernel increases with n. As finite sample properties, obviously the use of sin-polynomials along with circular kernels will guarantee estimators that are both periodic and rotationally invariant.

Importantly, the facts that most of circular kernels can be approximated by Euclidean ones if their concentration parameter is big, and that in a neighbourhood of 0 we have $\sin(x) \approx x$ could misleadingly suggest that, under certain asymptotic conditions, they could be successfully employable also when data have periodic nature. But this does not hold because linear kernels are guaranteed to give estimates that are *a)* severely wrong in a region around the origin (potentially the whole circle); *b)* not rotationally invariant. This has a strong practical relevance when considering that the choice of origin is arbitrary.

From now on, we assume K_κ to be a second sin-order circular kernel with concentration parameter $\kappa \in (0, \infty)$ (see [2]), i.e. a 2π-periodic density function admitting the absolutely convergent Fourier series representation

$$K_\kappa(\alpha) = \frac{1 + 2\sum_{j=1}^{\infty} \gamma_j(\kappa)\cos(j\alpha)}{2\pi}.$$

Remark 2.1. The particular case of no concentration gives constant weights, formally as $\kappa \to 0$, $K_\kappa(\alpha) \to (2\pi)^{-1}$. This makes the estimate independent on the data. For example, setting both κ and p to zero *always* yields an uniform density (see also Section 2.1). Curiously, for real-line kernel methods the shape of the estimate is never perfectly independent on the specific sample at hand. In fact, a positive weight fixed over the whole support still gives estimates that are not constant for finite samples due to the unavoidable presence of boundaries.

Note that $\gamma_0(\kappa) = 1$. Moreover, the second sin-order implies that $c_0^0(\theta) = 1$ and $c_0^2(\theta) \neq 0$, while due to the symmetry of K_κ, $c_0^j(\theta) = 0$ for odd j. For $\ell \geq 0$ and $j \geq 0$, direct calculation, involving Fourier coefficients, gives general expressions as follows

$$c_\ell^j(\theta) = \begin{cases} \frac{1}{j!2^j}\sum_{s=0}^{j}\binom{j}{s}(-1)^{s+(j+1)/2}\gamma_{|\ell-j+2s|}(\kappa)\sin(\ell\theta) & \text{if } j \text{ is odd,} \\ \frac{1}{j!2^j}\sum_{s=0}^{j}\binom{j}{s}(-1)^{s+j/2}\gamma_{|\ell-j+2s|}(\kappa)\cos(\ell\theta) & \text{if } j \text{ is even,} \end{cases} \tag{2.2}$$

and

$$s_\ell^j(\theta) = \begin{cases} \frac{1}{j!2^j}\sum_{s=0}^{j}\binom{j}{s}(-1)^{s+1+(j+1)/2}\gamma_{|\ell-j+2s|}(\kappa)\cos(\ell\theta) & \text{if } j \text{ is odd,} \\ \frac{1}{j!2^j}\sum_{s=0}^{j}\binom{j}{s}(-1)^{s+j/2}\gamma_{|\ell-j+2s|}(\kappa)\sin(\ell\theta) & \text{if } j \text{ is even.} \end{cases} \tag{2.3}$$

For a pth order approximation the estimating equations have $p+1$ unknowns. With the aim of including *complete* trigonometric moments in the system, we need to start from order zero (which produces a single equation because $s_0^j(\theta) = \mathcal{S}_0(\theta) = 0$) only if the number of such unknowns is odd. In the next sections we distinguish the case of even p from odd p, but some unification in notation can be obtained by defining q as the integer part (floor) of $(p+1)/2$.

2.1. Even p

When the sin-polynomial degree p is even, simultaneous equations are obtained by including local trigonometric moments from order 0 up to order $p/2$. In matrix form these can be expressed as

$$\mathsf{B}_p\beta = \mathcal{M}\,, \tag{2.4}$$

where

$$\begin{aligned}\beta' &:= (\beta_0\;\beta_1\;\ldots\;\beta_p)\,,\\ \mathcal{M}' &:= (\mathcal{C}_0(\theta)\;\mathcal{C}_1(\theta)\;\mathcal{S}_1(\theta)\;\ldots\;\mathcal{C}_q(\theta)\;\mathcal{S}_q(\theta))\;,\end{aligned}$$

and

$$\mathsf{B}_p := \begin{pmatrix} 1 & c_0^1(\theta) & \ldots & c_0^p(\theta)\\ c_1^0(\theta) & c_1^1(\theta) & \ldots & c_1^p(\theta)\\ s_1^0(\theta) & s_1^1(\theta) & \ldots & s_1^p(\theta)\\ \vdots & \vdots & \ddots & \vdots\\ c_q^0(\theta) & c_q^1(\theta) & \ldots & c_q^p(\theta)\\ s_q^0(\theta) & s_q^1(\theta) & \ldots & s_q^p(\theta)\end{pmatrix}.$$

The density estimator is defined to be the solution for β_0, i.e.

$$\hat{f}(\theta;p) := \frac{|\mathsf{A}_p|}{|\mathsf{B}_p|}, \tag{2.5}$$

where A_p is the same as B_p but with the first column replaced by $\mathcal{M}$. Elementary orthogonality arguments prove that both A_p and B_p have full rank.

Remark 2.2. Note that using expressions (2.2) and (2.3) in the general solution (2.5) leads to explicit closed-form solutions for the estimators whatever the kernel is. In particular, Cramer's rule is not used as a numerical algorithm, since it is known to be expensive and unstable, but only as a way to represent exact solutions of the system, which will are seen to be convenient for various asymptotic considerations.

Clearly, recalling that $\int K_\kappa = 1$, we see that the case $p = 0$ yields the standard kernel density estimator

$$\hat{f}(\theta;0) = \frac{1}{n}\sum_{i=1}^{n} K_\kappa(\Theta_i - \theta).$$

When K_κ is a von Mises kernel, i.e. $\gamma_j(\kappa) = \mathcal{I}_j(\kappa)/\mathcal{I}_0(\kappa)$, with $\mathcal{I}_j$, $j \in \mathbb{N}$, denoting the modified Bessel function of first kind and order j, the estimator for $p = 2$ has the following simple form

$$\hat{f}(\theta;2) = \frac{\mathcal{I}_0(\kappa)\mathcal{I}_2(\kappa)\hat{f}(\theta;0) - \mathcal{I}_1^2(\kappa)\hat{f}(\theta;1)}{\mathcal{I}_0(\kappa)\mathcal{I}_2(\kappa) - \mathcal{I}_1^2(\kappa)}, \tag{2.6}$$

that is a linear combination of density estimators with $p = 0$ and $p = 1$ (for a definition of the latter see formula (2.8) in the next section). Such a structure (which will generally result in an estimator which is not non-negative for $p \neq 0$) is reminiscent of the *jackknife* technique, as originally formulated by [10], where a density estimator is defined as a linear combination of two distinct ones in order to get bias reduction.

The solution for β_j, $j \in (1, \ldots, p)$, of the above system gives estimator for $f^{(j)}(\theta)$, which is

$$\hat{f}^{(j)}(\theta; p) := \frac{|\mathsf{A}_p^j|}{|\mathsf{B}_p|},$$

where A_p^j is the same of B_p, but with the $(j+1)$th column replaced by $\mathcal{M}$. Notice that in the first equation of system (2.4), recalling that for odd j $c_0^j = 0$, the coefficients β_j disappear. This in turn implies the same odd derivative estimators as for the $p - 1$ degree.

2.2. *Odd p*

When the sin-polynomial degree p is odd, we consider the matching between local trigonometric moments from order 1 up to order $(p + 1)/2$, obtaining the following system expressed in matrix form

$$\mathsf{D}_p \beta = \mathcal{N}\,,$$

where D_p and $\mathcal{N}$ are defined as B_p and $\mathcal{M}$ with, respectively, the first row and first element omitted. The solution for β_0 gives the estimator

$$\hat{f}(\theta; p) := \frac{|\mathsf{C}_p|}{|\mathsf{D}_p|}, \tag{2.7}$$

where C_p has the same columns of D_p, except the first one which is replaced by $\mathcal{N}$. The same arguments as in the previous section prove that both C_p and D_p have full rank, also Remark (2.2) applies.

The use of a von Mises kernel yields simple estimators for $p = 1$ and $p = 3$, as follows

$$\hat{f}(\theta; 1) = \frac{1}{2\pi n \mathcal{I}_1(\kappa)} \sum_{i=1}^{n} \cos(\Theta_i - \theta) \exp(\kappa \cos(\Theta_i - \theta)), \tag{2.8}$$

$$\hat{f}(\theta; 3) = \frac{w(\kappa)\hat{f}(\theta; 1) + \kappa^2 \mathcal{I}_2(\kappa) \frac{1}{2\pi n} \sum_{i=1}^{n} \cos(2\Theta_i - 2\theta) \exp(\kappa \cos(\Theta_i - \theta))}{w(\kappa) + \kappa^2 \mathcal{I}_2(\kappa)^2},$$

with $w(\kappa) := 6\kappa \mathcal{I}_0(\kappa)\mathcal{I}_1(\kappa) - (12 + \kappa^2)\mathcal{I}_1(\kappa)^2$. We could interpret the solution for $p = 1$ as a standard kernel density estimate using the non-positive kernel $K_\kappa(\theta) = \cos(\theta) \exp(\kappa \cos(\theta))/(2\pi \mathcal{I}_1(\kappa))$, where the cosine function further decreases the weight assigned to observations as they get far from the estimation point. Consequently, the peaks (valleys) of the estimate become basically higher (thinner). The cubic fit is reminiscent of the *jackkinife* technique observed in expression (2.6). Concerning derivative estimators, the solution of the above system for β_j, $j \in (1, \ldots, p)$, yields

$$\hat{f}^{(j)}(\theta; p) := \frac{|\mathsf{C}_p^j|}{|\mathsf{D}_p|},$$

with C_p^j being the same as D_p, but with the $(j+1)$th column replaced by $\mathcal{N}$.

3. Large samples results

A fundamental assumption made in this section is that the concentration parameter of the kernel $\kappa = \kappa_n$ is an increasing function of sample size n. To keep notation less cumbersome, such dependence will be suppressed.

The bias of $\hat{f}(\theta; p)$ is obtained in the following

Result 3.1. *Given the random sample of angles $\Theta_1, \ldots, \Theta_n$ from the unknown density f, consider estimators (2.7) and (2.5) for odd and even p, respectively. Assume that*

i) *K_κ is a second sin-order circular kernel whose Fourier coefficients strictly increase with κ;*
ii) $\lim_{n\to\infty} \kappa = \infty$;
iii) *f is $(p+1)$-times and $(p+2)$ differentiable for odd and even p, respectively.*

Then, for odd p

$$E[\hat{f}(\theta; p)] - f(\theta) = \frac{|\tilde{\mathsf{C}}_p|}{|\mathsf{D}_p|} f^{(p+1)}(\theta) + o\left(\frac{|\tilde{\mathsf{C}}_p|}{|\mathsf{D}_p|}\right),$$

where $\tilde{\mathsf{C}}_p$ is the same as C_p, but with the first column replaced by

$$\begin{pmatrix} c_1^{p+1}(\theta) & s_1^{p+1}(\theta) & \ldots & c_{(p+1)/2}^{p+1}(\theta) & s_{(p+1)/2}^{p+1}(\theta) \end{pmatrix}',$$

while, for even p,

$$E[\hat{f}(\theta; p)] - f(\theta) = \frac{|\tilde{\mathsf{A}}_p|}{|\mathsf{B}_p|} f^{(p+2)}(\theta) + o\left(\frac{|\tilde{\mathsf{A}}_p|}{|\mathsf{B}_p|}\right),$$

with $\tilde{\mathsf{A}}_p$ being the same as A_p, but with the first column replaced by

$$\begin{pmatrix} c_0^{p+2}(\theta) & c_1^{p+2}(\theta) & s_1^{p+2}(\theta) & \ldots & c_{p/2}^{p+2}(\theta) & s_{p/2}^{p+2}(\theta) \end{pmatrix}'.$$

Proof. See Appendix. □

Circularity makes the normalization of the estimates easily tractable. Indeed, we have just seen that $E[\hat{f}(\theta;p)]$ is asymptotically formulated as a sum between $f(\theta)$ and a linear combination of the derivatives of f at θ. Now, from the fact that derivatives of a circular density function are obviously periodic functions, it follows from Fubini's theorem that the expectation of the area of the estimate is equal to one, without regard to the values of p and κ. This comes in stark contrast with Euclidean higher order bias theory.

Concerning the variance, we get the following

Result 3.2. *Given the random sample of angles $\Theta_1,\dots,\Theta_n$ from the unknown density f, consider estimator (2.7) for odd p ((2.5) for even p, respectively). Let M_{ij} denote the (i,j) minor of C_p (A_p, resp.). Then, we have*

$$Var[\hat{f}(\theta;p)] = \frac{1}{V_p^2}\left\{\sum_{i=1}^{p+1} M_{i1}^2\, Var[a_{i1}] + 2\sum_{i\neq j}(-1)^{(i+j)} M_{i1}M_{j1}\, Cov[a_{i1},a_{j1}]\right\}, \tag{3.1}$$

where a_{ij} stands for the (i,j)th entry of C_p (A_p, resp.), $V_p := |\mathsf{D}_p|$ ($V_p := |\mathsf{B}_p|$, resp.). Specifically, the possible covariance expressions, under assumptions of Result 3.1, for $(\ell,m)\in\mathbb{N}\times\mathbb{N}$ and $\ell\geq m$, omitting $O(n^{-1})$ terms, are

$$Cov[\mathcal{C}_\ell(\theta),\mathcal{C}_m(\theta)] = \frac{f(\theta)}{4n\pi}\left\{\cos((\ell-m)\theta)P(\ell,m) + \cos((\ell+m)\theta)Q(\ell,m)\right\},$$

$$Cov[\mathcal{S}_\ell(\theta),\mathcal{S}_m(\theta)] = \frac{f(\theta)}{4n\pi}\left\{\cos((\ell-m)\theta)P(\ell,m) - \cos((\ell+m)\theta)Q(\ell,m)\right\},$$

$$Cov[\mathcal{S}_\ell(\theta),\mathcal{C}_m(\theta)] = \frac{f(\theta)}{4n\pi}\left\{\sin((m-\ell)\theta)P(\ell,m) + \sin((\ell+m)\theta)Q(\ell,m)\right\},$$

$$Cov[\mathcal{S}_m(\theta),\mathcal{C}_\ell(\theta)] = \frac{f(\theta)}{4n\pi}\left\{\sin((\ell-m)\theta)P(\ell,m) + \sin((\ell+m)\theta)Q(\ell,m)\right\},$$

with

$$P(\ell,m) := \sum_{i=0}^{\ell-m}\gamma_i(\kappa)\gamma_{\ell-m-i}(\kappa) + 2\sum_{i=\ell-m+1}^{\infty}\gamma_i(\kappa)\gamma_{i-(\ell-m)}(\kappa),$$

$$Q(\ell,m) := \sum_{i=1}^{\ell+m-1}\gamma_i(\kappa)\gamma_{\ell+m-i}(\kappa) + 2\sum_{i=0}^{\infty}\gamma_i(\kappa)\gamma_{\ell+m+i}(\kappa).$$

Proof. See Appendix. □

Asymptotic normality of $\hat{f}(\theta;p)$, for both even and odd p, can be easily demonstrated starting from the fact that both $|\mathsf{A}_p|$ and $|\mathsf{B}_p|$ are linear combinations of sample averages of independent and identically distributed random variables.

3.1. Von Mises kernel theory

Although very general, Results 3.1 and 3.2 do not allow further evaluations like, for example, convergence rates or optimal smoothing. In fact, they are general with respect to the kernel, and therefore Fourier coefficients remain unspecified. However, it is well known that in local density estimation the choice of the kernel is generally not considered an important task. Specifically, we could select whatever kernel showing uni-modality, symmetry, smoothness, *rapidly* decaying tails to zero and that is able to concentrate its whole mass around zero. Now, well known circular densities that are uni-modal and symmetric are: triangular, cardioid, wrapped Cauchy, von Mises and wrapped normal.

Once said that the first one is not smooth enough, the second one is not able to indefinitely concentrate, and that the third one has too heavy tails, von Mises and wrapped normal densities remain to us. However, since from moderate concentration and upwards they have nearly identical shape, their use gives generally same results. These considerations suggest that restricting the theory to von Mises kernel is not a severe choice.

In what follows we show how using the von Mises density as the kernel leads to a number of interesting considerations. The asymptotic bias of $\hat{f}(\theta;p)$ for some values of p is obtained in the following

Result 3.3. *Given a random sample of angles $\Theta_1,\ldots,\Theta_n$, from the unknown density f, consider estimator $\hat{f}(\theta;p)$, $\theta\in[-\pi,\pi)$, with K_κ being a von Mises kernel. Assuming that $\kappa\to\infty$ as $n\to\infty$,*

$$E[\hat{f}(\theta;0)]-f(\theta)=\frac{1}{\kappa}\frac{\mathcal{I}_1(\kappa)}{2\mathcal{I}_0(\kappa)}f^{(2)}(\theta)+o\left(\frac{1}{\kappa}\right)$$

$$E[\hat{f}(\theta;1)]-f(\theta)=\frac{1}{\kappa}\frac{\mathcal{I}_2(\kappa)}{2\mathcal{I}_1(\kappa)}f^{(2)}(\theta)+o\left(\frac{1}{\kappa}\right),$$

and

$$E[\hat{f}(\theta;2)]-f(\theta)=\frac{1}{\kappa^2}\frac{\{\mathcal{I}_2^2(\kappa)-\mathcal{I}_1(\kappa)\mathcal{I}_3(\kappa)\}}{8\{\mathcal{I}_0(\kappa)\mathcal{I}_2(\kappa)-\mathcal{I}_1^2(\kappa)\}}f^{(4)}(\theta)+o\left(\frac{1}{\kappa^2}\right).$$

Proof. See Appendix. □

Comparing the asymptotic bias of the local constant and the local linear fit, we see that, despite both have the same magnitude, the local linear one is smaller than the constant one for each value of κ. This is in clear contrast with ordinary polynomial fitting, where the relative merits depend on the estimation point. For $p=2$, the asymptotic bias is $O(1/\kappa^2)$, and so the jackknife idea works by cancelling the larger bias term.

Concerning the variance, asymptotic results for some values of p are collected in the following

Result 3.4. *Given a random sample of angles $\Theta_1,\ldots,\Theta_n$, from the unknown density f, consider estimator $\hat{f}(\theta;p)$, $\theta\in[-\pi,\pi)$, equipped with a von Mises kernel. Assuming that $\kappa/n\to 0$ as $n\to\infty$*

$$Var[\hat{f}(\theta;0)]=\frac{f(\theta)}{n}\frac{\mathcal{I}_0(2\kappa)}{2\pi\mathcal{I}_0^2(\kappa)}+O\left(\frac{1}{n}\right), \tag{3.2}$$

$$Var[\hat{f}(\theta;1)] = \frac{f(\theta)}{n}\left\{\frac{\mathcal{I}_0(2\kappa)+\mathcal{I}_2(2\kappa)}{4\pi\mathcal{I}_1^2(\kappa)}\right\} + O\left(\frac{1}{n}\right), \tag{3.3}$$

and

$$Var[\hat{f}(\theta;2)] = \frac{f(\theta)}{n}\times$$
$$\left\{\frac{\mathcal{I}_0(2\kappa)[\mathcal{I}_1^2(\kappa)+2\mathcal{I}_2^2(\kappa)]+\mathcal{I}_1(\kappa)[\mathcal{I}_1(\kappa)\mathcal{I}_2(2\kappa)-4\mathcal{I}_2(\kappa)\mathcal{I}_1(2\kappa)]}{4\pi\{\mathcal{I}_1^2(\kappa)-\mathcal{I}_0(\kappa)\mathcal{I}_2(\kappa)\}^2}\right\} + O\left(\frac{1}{n}\right).$$

Proof. See Appendix. □

Using the properties of Bessel functions, we can see that, for big enough κ, the asymptotic variances in the above result have magnitude $O(\sqrt{\kappa}/n)$. However, when comparing $p = 0$ and $p = 1$, we observe a phenomenon similar to that seen for biases, but in this case the superiority is on the local constant side. Moreover, concerning asymptotic behaviour of the quadratic fit, we have

$$\lim_{\kappa\to\infty}\frac{Var[\hat{f}(\theta;2)]}{Var[\hat{f}(\theta;0)]} = \frac{27}{16}.$$

These results clearly indicate that fitting a quadratic polynomial would be convenient when the bias is severe: in fact, we observe a variance inflation, inevitably produced by the need of estimating more parameters, that with a big enough sample size will be dominated by bias reduction.

Concerning the smoothing degree, setting $R(g) := \int_{-\pi}^{\pi} g^2(\alpha)d\alpha$, for a square integrable function g, we get

Result 3.5. *Given a random sample of angles $\Theta_1,\dots,\Theta_n$ from the unknown density f, consider estimator $\hat{f}(\theta;p)$, $\theta\in[-\pi,\pi)$, having the von Mises kernel as the weight function. The value of κ which minimizes the asymptotic mean integrated squared error of $\hat{f}(\cdot;p)$ with $p\in(0,1)$ is*

$$\kappa_{AMISE} = \left(2n\pi^{1/2}R\left(f^{(2)}\right)\right)^{2/5},$$

while, for $p = 2$,

$$\kappa_{AMISE} = \left(\frac{4}{27}n\pi^{1/2}R\left(f^{(4)}\right)\right)^{2/9}.$$

Proof. See Appendix. □

It follows that the rate of convergence of $\hat{f}(\theta;p)$ is $n^{-4/5}$ for both $p = 0$ and $p = 1$, while it is $n^{-8/9}$ for $p = 2$. The rate of convergence for whatever order p is provided by the following

Result 3.6. *Given a random sample of angles $\Theta_1,\dots,\Theta_n$ from the unknown density f, consider estimator $\hat{f}(\theta;p)$, $\theta\in[-\pi,\pi)$, with K_κ being a von Mises kernel. The convergence rate of $\hat{f}(\theta;p)$ is $n^{-2(p+1)/(1+2(p+1))}$ for odd p, and $n^{-2(p+2)/(1+2(p+2))}$ for even p.*

Proof. See Appendix. □

Finally, as circular samples often come in form of dependent data – think, for example, of direction of winds or marine currents taken at a given location within a period of time – it could be of interest to ensure the effect of some general dependence structures on our methods. In what follows we assume that data come from an α-mixing stochastic process.

Result 3.7. *Let $\Theta_1, \ldots, \Theta_n$ be a realization of an α-mixing process. If*

i) the mixing coefficients satisfy $|\alpha(\omega)| \leq Q_1 \omega^{-\lambda}$, $\omega \in \mathbb{Z}$, for positive constants Q_1 and $\lambda > 2$;
ii) the joint density g of (Θ_l, Θ_m) satisfies $\|g\|_\infty < \infty$, for $l \in (1, \ldots, n)$, $m \in (1, \ldots, n)$;

then, for $p \in \mathbb{N}$, the convergence rate of estimator $\hat{f}(\theta; p)$, $\theta \in [-\pi, \pi)$, with the von Mises kernel is the same as the i.i.d. case.

Proof. See Appendix. □

A similar result for kernel density estimator in the Euclidean multivariate setting can be found, for example, in [9].

4. Practical performance

4.1. Density estimation

This simulation study compares both best possible and practical performances of the estimators. As accuracy indicator we refer to integrated mean squared error (IMSE). Kernels used are von Mises probability density functions, denoted by $vM(\mu, \kappa)$, where parameters μ and κ indicate respectively the mean direction and the concentration parameter.

When $p > 0$ our estimates are not guaranteed to be *bona-fide*. However, we have not correct them because elementary considerations say that the quantity $\int \hat{f}$ goes to one at the same rate as the asymptotic bias, and therefore the pointwise convergence rates of $\hat{f}$ to f remain unaffected up to a coefficient equal to the value of f at θ. As a final remark, we recall that the case $p = 0$ is the classical circular kernel density estimator, firstly proposed by [6]. Therefore, $p = 0$ is the best possible benchmark, useful to check practical utility of our proposals.

In the first study empirical IMSE curves are obtained as functions of κ. We see that, even if the sample sizes are moderate ($n = 100$), the standard circular kernel density estimate is not optimal. We observe that when the population is not very peaked (left panel of Figure 1) the local linear estimator wins, while for a more concentrated population (left panel of Figure 2) higher order sin-polynomial estimators are preferable. This is easily seen after noting that higher order derivatives become more important for reducing bias when the concentration parameter of the population increases. In the right panels

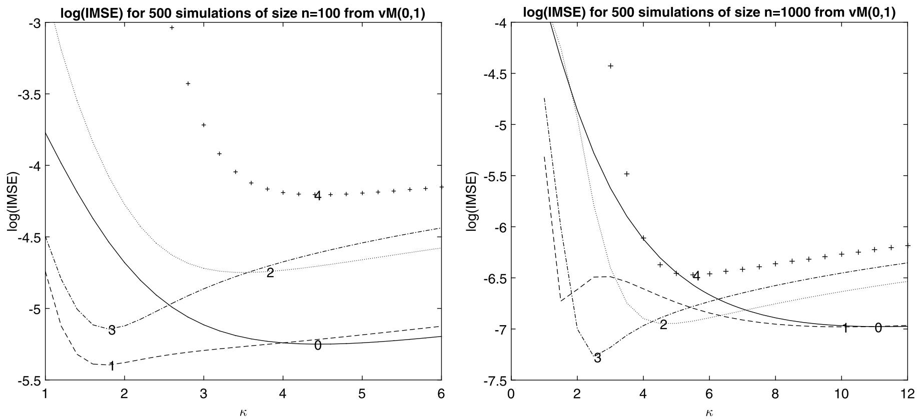

FIG 1. *log(IMSE) for a range of values of κ for $p = 0$ (solid), $p = 1$ (dashed), $p = 2$ (dotted), $p = 3$ (dotdash) and $p = 4$ (+) for 500 samples of size $n = 100$ (left) and $n = 1000$ (right) from a $vM(0, 1)$.*

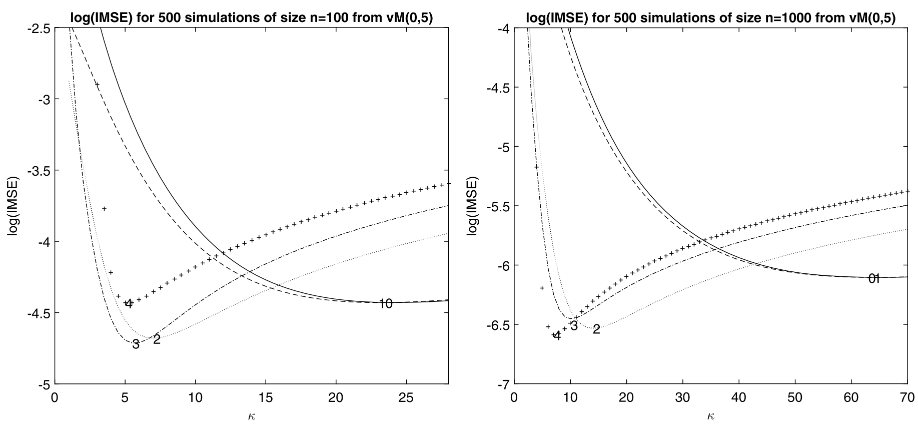

FIG 2. *log(IMSE) for a range of values of κ for $p = 0$ (solid), $p = 1$ (dashed), $p = 2$ (dotted), $p = 3$ (dotdash) and $p = 4$ (+) for 500 samples of size $n = 100$ (left) and $n = 1000$ (right) from a $vM(0, 5)$.*

of these figures we use $n = 1000$. Since a bigger sample size leads to a more accurate estimate of the derivatives, we can see that the best performances are given by estimators of higher orders, because, in contrast with the case $n = 100$, bias reduction dominates variance inflation.

Specifically for the $vM(0, 1)$ population model the sin-polynomial order $p = 3$ is the best, while in the second model performance improves with the value of p. Moreover, it is worth noting that the cosine correction of the von Mises kernel ($p = 1$) gives a slight improvement on $p = 0$ for $n = 100$, whereas same convergence rates yields equivalent behaviours for $n = 1000$.

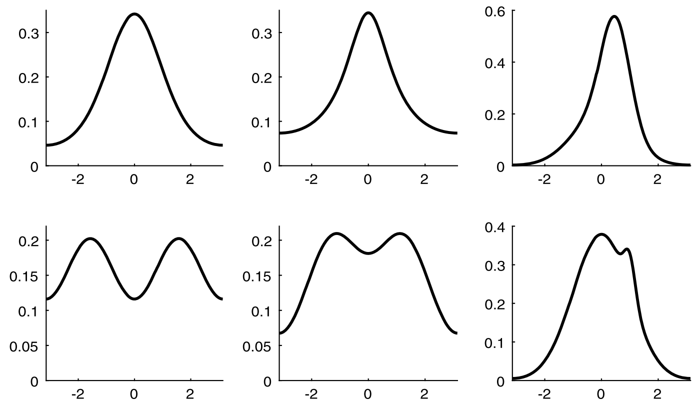

Fig 3. *Population density models with parameters (mean direction, concentration) used for the simulation on density estimation. Top line, from left: von Mises* $(0,1)$*; Wrapped Cauchy* $(0,1)$*, equal mixture between two Wrapped Normals with parameters* $(0,1)$ *and* $(0.5,0.5)$*. Bottom line, from left: equal mixture between two Wrapped Normals with parameters* $(-\pi/2,1)$ *and* $(\pi/2,1)$*, equal mixture between two Wrapped Normals with parameters* $(-2\pi/5,1)$ *and* $(2\pi/5,1)$*, mixture between two Wrapped Normals with weights* $19/20$ *and* $1/20$*, and parameters* $(0,1)$ *and* $(1,0.2)$.

In a second experiment, we estimate various densities by selecting smoothing degree by simple least squares cross-validation. Our population models are represented in Figure 3. Other than the standard von Mises, we have considered classical more difficult cases like multimodality, asymmetry or heavy tails. A promising estimator has been included as a competitor, other than standard kernel density method ($p = 0$). It is the circular local likelihood method described in [3]. Such an estimator can be defined as a polynomial estimator derived starting from a likelihood score. The Authors present closed formulas for polynomial degrees 1 and 2. We have included here both of them. We use sample sizes equal to 100 or 500. They have been taken moderate in order to check usefulness of the methods in practical situations. Our theoretical results have already described the behaviour of our proposals for large samples.

The results are collected in Tables 1 to 6. We can see that the selector works in a reasonable way, although the known tendency to undersmoothing arises when comparing median with average values of the obtained bandwidths. This tendency is more evident for small p and n. Coming to IMSE performances, the message arising from the results is as follows. The worst estimator is, by far, the classical kernel method. Our local moment approach has a clear edge on the local likelihood one. As indicated previously, bigger sample sizes and larger curvature favour the choice of higher polynomial degrees. Interestingly, the best estimator, for $n = 500$, is the third order polynomial model. This is welcome surprise, provided that sample size is not very big and the estimator has a simple closed form. Additionally, we note that third order remains nearly un-exploited in practical case studies, for both Euclidean and non-Euclidean polynomial methods.

Table 1
Comparison between proposed estimators and circular local likelihood ones (bracketed) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $vM(0,1)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	7.32	4.10	7.55
	1	5.51 (7.57)	1.66 (3.97)	3.27 (7.50)
	2	9.49 (10.41)	3.34 (3.14)	4.25 (4.21)
	3	7.09	1.77	2.23
500	0	2.05	8.10	11.45
	1	1.60 (2.08)	1.60 (8.07)	2.02 (11.30)
	2	2.06 (2.77)	4.13 (5.52)	5.19 (7.08)
	3	1.61	2.30	2.69

Table 2
Comparison between proposed estimators and (circular local likelihood ones) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $wC(0,1)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	8.04	4.24	6.56
	1	7.10 (8.14)	1.85 (4.12)	4.65 (6.35)
	2	10.08 (11.09)	3.27 (3.25)	4.19 (4.20)
	3	8.02	1.69	2.02
500	0	2.39	9.12	12.85
	1	2.67 (2.40)	1.74 (8.99)	6.02 (12.67)
	2	2.28 (3.03)	4.10 (5.77)	5.27 (7.51)
	3	2.04	2.36	2.93

TABLE 3

Comparison between proposed estimators and (circular local likelihood ones) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $0.5wN(0,1) + 0.5wN(.5,.5)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	12.52	10.90	17.13
	1	14.35 (12.98)	8.89 (10.54)	13.67 (17.16)
	2	11.44 (11.83)	4.06 (2.78)	5.89 (4.18)
	3	11.08	2.56	3.68
500	0	3.57	21.96	32.76
	1	4.30 (3.68)	20.57 (22.52)	29.77 (32.46)
	2	2.86 (3.35)	5.97 (6.43)	8.23 (9.40)
	3	3.14	4.38	5.64

TABLE 4

Comparison between proposed estimators and (circular local likelihood ones) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $0.5wN(-\pi/2,1) + 0.5wN(\pi/2,1)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	6.52	1.69	3.97
	1	6.77 (6.48)	2.36 (1.65)	5.21 (3.84)
	2	10.23 (11.24)	2.79 (3.56)	3.57 (4.08)
	3	3.90	9.56e-04	0.48
500	0	1.96	6.91	9.81
	1	1.91 (1.97)	5.47 (6.87)	8.52 (9.59)
	2	2.01 (2.87)	2.80 (5.57)	3.97 (7.03)
	3	1.74	1.42	1.62

TABLE 5

Comparison between proposed estimators and (circular local likelihood ones) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $0.5wN(-2\pi/5, 1) + 0.5wN(2\pi/5, 1)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	6.67	2.74	4.60
	1	6.33 (6.89)	2.11 (2.70)	4.38 (5.32)
	2	10.08 (10.96)	2.96 (3.28)	3.78 (3.93)
	3	6.88	1.10	1.33
500	0	1.76	6.22	8.63
	1	1.83 (1.80)	3.98 (6.26)	6.31 (8.57)
	2	2.07 (2.80)	3.44 (5.65)	4.34 (6.78)
	3	1.57	1.87	2.29

TABLE 6

Comparison between proposed estimators and (circular local likelihood ones) in terms of average integrated squared errors (×1000) over 500 samples of sizes 100 or 500 drawn from a $19/20wN(0, 1) + 1/20wN(1, 0.2)$ population. Median and mean of the smoothing degrees selected by least-squares cross-validation.

Sample size	Polynomial degree	IMSE×1000	median of bandwidths	mean of bandwidths
100	0	8.97	5.30	9.22
	1	6.61 (9.31)	1.61 (5.20)	3.36 (9.09)
	2	11.08 (11.23)	3.70 (2.76)	5.21 (3.93)
	3	8.66	2.01	2.70
500	0	2.92	11.90	19.99
	1	2.79 (3.11)	1.58 (12.40)	1.59 (21.21)
	2	3.11 (3.59)	4.91 (6.04)	7.80 (8.94)
	3	2.82	2.54	4.47

4.2. Slope and curvature estimation

Using the same samples as the first case study, we have additionally estimated first and second derivatives. The results are reported in Figures 4 and 5. We see that the interpretation given in Section 4.1 largely applies. Recalling that when p is even we have same odd derivative estimators as in the $p-1$ degree case, we show in Figure 4 only two curves.

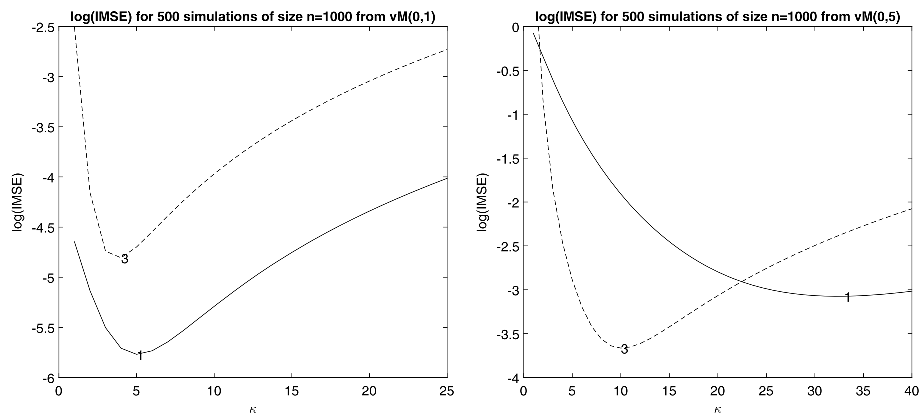

Fig 4. *log(IMSE) for a range of values of* κ *for* $p = 1$ *and* $p = 2$ *(solid),* $p = 3$ *and* $p = 4$ *(dashed) for 500 samples of size* $n = 1000$ *from a* $vM(0, 1)$ *(left) and* $vM(0, 5)$ *(right) comparing the first derivative estimation.*

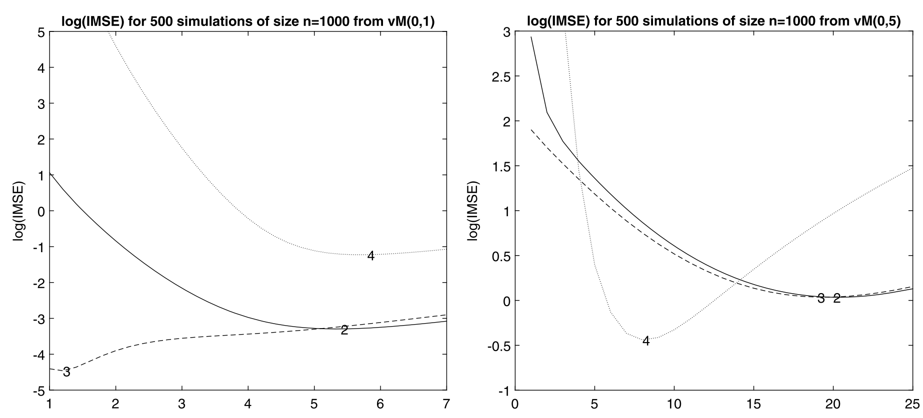

Fig 5. *log(IMSE) for a range of values of* κ *for* $p = 2$ *(solid),* $p = 3$ *(dashed) and* $p = 4$ *(dotted) for 500 samples of size* $n = 1000$ *from a* $vM(0, 1)$ *(left) and* $vM(0, 5)$ *(right) comparing the second derivative estimation.*

4.3. Application

To illustrate the methods on some real data we consider the arrival times of delayed planes in 2008. The full dataset is available from [7], but here we focus on two airlines (*American Airlines* and *Continental*) and one airport (Atlanta). The arrival times are converted to the 24-hour clock (which is periodic) and all days of the year are included in the analysis. Overall, there were 2676 (AA) and 1278 (CO) delayed flights. A brief examination of the data shows that arrival times are recorded to an integer number of minutes, and there is a tendency to round to the nearest 5 minutes.

We considered density estimation with polynomial degrees $p = 0, 1$ and 2, with smoothing parameters chosen by least squares cross-validation. The many coincident values in the data led to cross-validation selection of κ which diverged, so we "jittered" both datasets by adding random von Mises quantities with concentration 15.

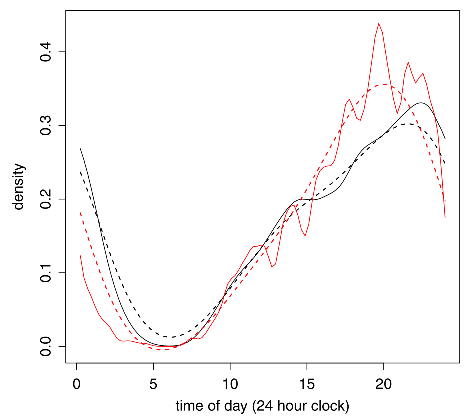

FIG 6. *Density estimation of arrival times of delayed flights for American Airlines (black) and Continental Airlines (red) using sin-polynomial degree $p = 0$ (continuous) and $p = 1$ (dashed). The smoothing parameters were chosen by least-squares cross-validation.*

The resulting density estimates of the arrival times are shown in Figure 6. The results for $p = 2$ are not shown as they were very similar to the case $p = 0$ for both datasets. Of course, the actual arrival times (even for delayed flights) will be closely related to scheduled arrivals and for the passengers the number of minutes delay is the most important information, but we note that the $p = 1$ probability density estimates are, in general, smoother than those for $p = 0$. We can also note a tendency for later arrival times — a sort of shift in the day — for American Airlines, relative to Continental Airlines. The smoothing parameters were 25.4 and 139.1 for $p = 0$ (AA and CO) and 3.13 and 1.84 for $p = 1$ (AA and CO) — it is interesting to observe that the magnitude is greater for CO for $p = 0$, but greater for AA for $p = 1$.

5. Discussion

In this paper we propose closed form estimators derived as a solution of a system of estimating equations. Other than having nice theoretical properties, our estimators seem to work satisfyingly in practical situations. As the consequence, methods for non-parametrically estimating a circular density, more sophisticated than traditional estimator presented in [6], seem a promising research field to pursue. We also introduce in circular statistics both simple formulas for local estimation of density derivatives and theory for the case of dependent observations.

Our method makes possible to efficiently employ *a priori information* on the smoothness of a circular density, especially when the curvature of the target population is pronounced. Surely, such a higher order differentiability can reveal a disadvantage for making local moments less generally applicable than traditional circular kernel method.

A promising development for our approach could lie in replacing our series expansion by a circular parametric family. Here we would estimate optimal smoothing along with the density parameters. This would amount to a totally parametric method in correspondence of a null concentration of the kernel, as opposite to the situation described in Remark 2.1.

Appendix

Proof of Result 3.1. For odd (even, respectively) p, the result follows by observing that the approximation of $E[\mathcal{C}_\ell(\theta)]$ and $E[\mathcal{S}_\ell(\theta)]$, for $\ell \in (1,\dots,(p+1)/2)$ ($\ell \in (0,\dots,p/2)$, resp.), by expansion (2) for f around θ up to order $(p+1)$ ($(p+2)$, resp), yields

$$E[\mathcal{C}_\ell(\theta)] = \sum_{j=0}^{r_p} c_\ell^j(\theta) f^{(j)}(\theta) + o\left(c_\ell^{r_p}(\theta)\right),$$

and

$$E[\mathcal{S}_\ell(\theta)] = \sum_{j=0}^{r_p} s_\ell^j(\theta) f^{(j)}(\theta) + o\left(s_\ell^{r_p}(\theta)\right),$$

with $r_p := p+1$ ($r_p := p+2$, resp.). □

Proof of Result 3.2. Equation (3.1) easily follows by the formulation of $\hat{f}(\theta;p)$, while results for covariance components are obtained by using the first term of expansion (2) to approximate f around θ. □

Proof of Result 3.3. For $p=0$ and $p=1$, starting from the closed form of $\hat{f}(\theta;0)$, and $\hat{f}(\theta;1)$ respectively, expansion (2) up to the second order leads to the result. Starting from formula (2.6), and using the bias results for $\hat{f}(\theta;0)$, and $\hat{f}(\theta;1)$, we find that the second-order bias of $\hat{f}(\theta;2)$ is zero. Then, using expansion (2) up to the fourth order yields the asymptotic bias. □

Proof of Result 3.4. When $p = 0$, the asymptotic variance directly follows by using the first term of expansion (2) for f around θ in

$$\mathrm{Var}[\hat{f}(\theta;0)] = \frac{1}{n}\int_{-\pi}^{\pi} K_\kappa^2(\alpha-\theta)f(\alpha)d\alpha - \frac{1}{n}\left\{\int_{-\pi}^{\pi} K_\kappa(\alpha-\theta)f(\alpha)d\alpha\right\}^2,$$

along with the fact that for a von Mises kernel one has

$$\int_{-\pi}^{\pi} K_\kappa^2(\alpha)d\alpha = \frac{\mathcal{I}_0(2\kappa)}{2\pi\mathcal{I}_0^2(\kappa)}.$$

When $p = 1$, by applying (3.1) with $M_{11} = s_1^1(\theta)$, $M_{21} = c_1^1(\theta)$, $V_1 = c_1^0(\theta)s_1^1(\theta) - s_1^0(\theta)c_1^1(\theta)$, where

$$c_1^0(\theta) = \frac{\mathcal{I}_1(\kappa)\cos(\theta)}{\mathcal{I}_0(\kappa)}, \qquad s_1^0(\theta) = \frac{\mathcal{I}_1(\kappa)\sin(\theta)}{\mathcal{I}_0(\kappa)},$$

$$c_1^1(\theta) = -\frac{\mathcal{I}_1(\kappa)\sin(\theta)}{\kappa\mathcal{I}_0(\kappa)}, \qquad s_1^1(\theta) = \frac{\mathcal{I}_1(\kappa)\cos(\theta)}{\kappa\mathcal{I}_0(\kappa)},$$

and, using the first term of expansion (2) for f around θ,

$$\mathrm{Var}[a_{11}] = \mathrm{Var}[\mathcal{C}_1(\theta)] = \frac{f(\theta)}{n}\left\{\frac{\mathcal{I}_0(2\kappa)+\mathcal{I}_2(2\kappa)\cos(2\theta)}{4\pi\mathcal{I}_0^2(\kappa)}\right\} + O\left(\frac{1}{n}\right),$$

$$\mathrm{Var}[a_{21}] = \mathrm{Var}[\mathcal{S}_1(\theta)] = \frac{f(\theta)}{n}\left\{\frac{\mathcal{I}_0(2\kappa)-\mathcal{I}_2(2\kappa)\cos(2\theta)}{4\pi\mathcal{I}_0^2(\kappa)}\right\} + O\left(\frac{1}{n}\right),$$

$$\mathrm{Cov}[a_{11},a_{21}] = \mathrm{Cov}[\mathcal{C}_1(\theta),\mathcal{S}_1(\theta)] = \frac{f(\theta)}{n}\frac{\mathcal{I}_2(2\kappa)\sin(2\theta)}{4\pi\mathcal{I}_0^2(\kappa)} + O\left(\frac{1}{n}\right).$$

The same result can be obviously obtained also starting from equation (2.8). When $p = 2$, the asymptotic variance can be easily obtained starting from formulation (2.6), and using (3.2) and (3.3) along with

$$\mathrm{Cov}[\hat{f}(\theta;0),\hat{f}(\theta;1)] = \frac{f(\theta)}{n}\frac{\mathcal{I}_1(2\kappa)}{2\pi\mathcal{I}_0(\kappa)\mathcal{I}_1(\kappa)} + O\left(\frac{1}{n}\right). \qquad \square$$

Proof of Result 3.5. For κ big enough we have that, for $p \in (0,1)$,

$$\mathrm{Bias}[\hat{f}(\theta;p)] = \frac{1}{2\kappa}f^{(2)}(\theta) + o\left(\frac{1}{\kappa}\right) \quad \text{and} \quad \mathrm{Var}[\hat{f}(\theta;p)] = \frac{f(\theta)}{n}\sqrt{\frac{\kappa}{4\pi}} + o\left(\frac{\sqrt{\kappa}}{n}\right),$$

while

$$\mathrm{Bias}[\hat{f}(\theta;2)] = -\frac{1}{8\kappa^2}f^{(4)}(\theta) + o\left(\frac{1}{\kappa^2}\right)$$

and

$$\mathrm{Var}[\hat{f}(\theta;2)] = \frac{f(\theta)}{n}\frac{27}{16}\sqrt{\frac{\kappa}{4\pi}} + o\left(\frac{\sqrt{\kappa}}{n}\right).$$

By minimising the resulting asymptotic mean integrated squared errors over κ we get the results. $\square$

Proof of Result 3.6. First of all, when κ is big enough, the following simple expressions hold respectively for odd j

$$c_\ell^j(\theta) \approx \frac{-\ell \mathsf{OF}(j)\sin(\ell\theta)}{j!\kappa^{(j+1)/2}}, \quad s_\ell^j(\theta) \approx \frac{\ell \mathsf{OF}(j)\cos(\ell\theta)}{j!\kappa^{(j+1)/2}}, \tag{5.1}$$

and even j

$$c_\ell^j(\theta) \approx \frac{\mathsf{OF}(j-1)\cos(\ell\theta)}{j!\kappa^{j/2}}, \quad s_\ell^j(\theta) \approx \frac{\mathsf{OF}(j-1)\sin(\ell\theta)}{j!\kappa^{j/2}}, \tag{5.2}$$

where OF stands for the *Odd Factorial*, defined by $\mathsf{OF}(2r) := (2r-1)(2r-3)\dots1$, $r \in \mathbb{N}$. As a consequence, in virtue of Result 3.1 the bias of $\hat{f}(\theta;p)$ is $O(\kappa^{-(p+1)/2})$ for odd p, while it is $O(\kappa^{-(p+2)/2})$ for even p. For the variance components, it results $P(\ell,m) = \mathcal{I}_{\ell-m}(2\kappa)/\mathcal{I}_0^2(\kappa)$, and $Q(\ell,m) = \mathcal{I}_{\ell+m}(2\kappa)/\mathcal{I}_0^2(\kappa)$. Then, starting from equation (3.1), using approximations (5.1) and (5.2) along with the fact that each of the above quantities have magnitude $O(\sqrt{\kappa}/n)$, it can be shown that the asymptotic variance of $\hat{f}(\theta;p)$ has magnitude $O(\sqrt{\kappa}/n)$ for whatever polynomial order p. Combining this result with the asymptotic bias results, we find that the value of κ minimising the asymptotic mean squared error of $\hat{f}(\theta;p)$ has order respectively $O(n^{2/(1+2(p+1))})$ for odd p, and $O(n^{2/(1+2(p+2))})$ for even p. □

Proof of Result 3.7. For $s \in (1,\dots,n)$, and integer ℓ let

$$C_\ell(\theta,\Theta_s) := K_\kappa(\Theta_s-\theta)\cos(\ell\Theta_s), \text{ and } S_\ell(\theta,\Theta_s) := K_\kappa(\Theta_s-\theta)\sin(\ell\Theta_s),$$

and using again q as the integer part of $(p+1)/2$, for even p define

$$\mathcal{L}_s' := (C_0(\theta,\Theta_s)\; C_1(\theta,\Theta_s)\; S_1(\theta,\Theta_s)\; \dots\; C_q(\theta,\Theta_s)\; S_q(\theta,\Theta_s)),$$

and let $\mathbb{A}_{s,p}$ be defined as A_p but with the first column replaced by $\mathcal{L}_s$. Moreover, for odd p, let $\mathcal{O}_s$ be defined as $\mathcal{L}_s$ but with the first element omitted, and let $\mathbb{C}_{s,p}$ be the same as C_p but with the first column replaced by $\mathcal{O}_s$.

Then, by stationarity, for even p

$$\mathrm{Var}[\hat{f}(\theta;p)] = \frac{1}{|\mathsf{B}_p|^2}\left\{\frac{1}{n}\mathrm{Var}[|\mathbb{A}_{1,p}|] + \frac{2}{n}\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\mathrm{Cov}[|\mathbb{A}_{1,p}|,|\mathbb{A}_{s+1,p}|]\right\}, \tag{5.3}$$

while, for odd p, the above identity holds with D and $\mathbb{C}$ in place of B and $\mathbb{A}$, respectively. Here we consider the case of even p, while the case of odd p can be easily obtained by following the same reasoning with due modifications. Notice that the first summand in the RHS of (5.3) corresponds to the variance of estimator for the i.i.d. case, which, when the von Mises kernel is employed, has magnitude $O(\sqrt{\kappa}/n)$. Then, we have to show that the covariance term reflecting the extra-variability due to the dependence is $o(\sqrt{\kappa}/n)$.

To this end, notice that, for even p

$$\mathrm{Cov}[|\mathbb{A}_{1,p}|,|\mathbb{A}_{s+1,p}|] = \sum_{i=1}^{p+1}\sum_{j=1}^{p+1}(-1)^{i+j}M_{i1}M_{j1}\mathrm{Cov}[\mathbb{a}_{i1,1},\mathbb{a}_{j1,s+1}],$$

where $\mathbb{a}_{ij,s}$ stands for the (i,j)th entry of $\mathbb{A}_{s,p}$, and M_{ij} is defined as in the statement of Result 3.2. We now reason in a similar way as in the proof of Theorem 5.1 in [4]. In particular, we firstly note that, due to the boundedness of sine and cosine, it holds that

$$|\mathrm{Cov}[\mathbb{a}_{i1,1}, \mathbb{a}_{j1,s+1}]| \leq \mathrm{E}[\mathbb{a}_{i1,1}\mathbb{a}_{j1,s+1}] + \mathrm{E}[\mathbb{a}_{i1,1}]^2 \leq \|g\|_\infty + \mathrm{E}[\mathbb{a}_{i1,1}]^2$$

then, in virtue of assumption ii), for a sequence of integers $u_n \to \infty$ and a constant Q_2, we have

$$\sum_{s=1}^{u_n} |\mathrm{Cov}[|\mathbb{A}_{1,p}|, |\mathbb{A}_{s+1,p}|]| \leq u_n Q_2. \tag{5.4}$$

Moreover, the use of Billingsley's inequality leads to

$$|\mathrm{Cov}[\mathbb{a}_{i1,1}, \mathbb{a}_{i1,s+1}]| \leq 4\alpha(s)\|\mathbb{a}_{i1,1}\|_\infty \|\mathbb{a}_{i1,s+1}\|_\infty \leq 4\alpha(s)\|K_\kappa\|_\infty^2,$$

and observing that for κ big enough $\mathcal{I}_0(\kappa) \simeq e^\kappa/\sqrt{2\pi\kappa}$, and recalling assumption i), we find that, for some constant Q_4,

$$\sum_{s=u_n}^{n-1} |\mathrm{Cov}[|\mathbb{A}_{1,p}|, |\mathbb{A}_{s+1,p}|]| \leq Q_4 \sum_{s=u_n}^{\infty} s^{-\lambda}\kappa = O(u_n^{-\lambda+1}\kappa). \tag{5.5}$$

Now, letting $u_n = \kappa^{\lambda/4}$, recalling assumption i), equations (5.4) and (5.5) yield the result. □

References

[1] Di Marzio, M. and Taylor, C. C. (2009). Using small bias nonparametric density estimators for confidence interval estimation. *Journal of Nonparametric Statistics*, **21:2**, 229–240. MR2488156

[2] Di Marzio, M., Panzera, A. and Taylor, C. C. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, **141**, 2156–2173. MR2772221

[3] Di Marzio, M., Fensore, S., Panzera, A. and Taylor, C. C. (2016). Practical performance of local likelihood for circular density estimation. *Journal of Statistical Computation and Simulation*, **86**, 2560–2572. MR3511013

[4] Fan, J. and Yao, Q. (2003). Nonlinear Time Series: Nonparametric and Parametric Methods. New York: Springer-Verlag. MR1964455

[5] Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics*, **20**, 675–694. MR1165587

[6] Hall, P., Watson, G. S. and Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, **74**, 751–762. MR0919843

[7] Kaggle. Airlines Delay: Airline on-time statistics and delay causes. `www.kaggle.com/giovamata/airlinedelaycauses`

[8] Oliveira, M., Crujeiras, R. M. and Rodríguez-Casal A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis*, **56**, 3898–3908. MR2957840

[9] Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185–197. MR0732897

[10] Shucany, W. R. and Sommers, J. P. (1977). Improvement of Kernel Type Density Estimators. *Journal of the American Statistical Association*, **72**, 420–423. MR0448691

[11] Singh, R. S. (1977). Applications of Estimators of a Density and its Derivatives to Certain Statistical Problems. *Journal of the Royal Statistical Society. Series B*, **39**, 357–363. MR0652725

[12] Spurr, B. D. and Koutbeiy, M. A. (1991). A comparison of various methods for estimating the parameters in mixtures of von Mises distributions. *Communications in Statistics - Simulation and Computation*, **20**, 725–741. MR1129093

[13] Sra, S. (2012). A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$. *Computational Statistics*, **27**, 177–190. MR2877816

[14] Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, **52**, 3493–3500. MR2427365