CrossMark

ORIGINAL PAPER

Circular local likelihood

Marco Di Marzio[1] · **Stefania Fensore[1]** ·
Agnese Panzera[2] · **Charles C. Taylor[3]**

Abstract We introduce a class of local likelihood circular density estimators, which includes the kernel density estimator as a special case. The idea lies in optimizing a spatially weighted version of the log-likelihood function, where the logarithm of the density is locally approximated by a periodic polynomial. The use of von Mises density functions as weights reduces the computational burden. Also, we propose closed-form estimators which could form the basis of counterparts in the multidimensional Euclidean setting. Simulation results and a real data case study are used to evaluate the performance and illustrate the results.

Keywords Bessel functions · Circular data · Density estimation · Log-likelihood · Numerical integration · Product kernels · von Mises density

Mathematics Subject Classification 62G07

✉ Charles C. Taylor
  charles@maths.leeds.ac.uk

  Marco Di Marzio
  marco.dimarzio@unich.it

  Stefania Fensore
  stefania.fensore@unich.it

  Agnese Panzera
  a.panzera@disia.unifi.it

[1] DMQTE, Università di Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy

[2] DiSIA, Università di Firenze, Viale Morgagni 59, 50134 Florence, Italy

[3] Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

⚫ Springer

1 Introduction

A circular observation can be represented by a point on the unit circle and measured by an angle $\theta \in [-\pi, \pi)$, after both an origin and an orientation have been chosen. Its real-line representation is provided by the equivalence class $\{2m\pi + \theta, m \in \mathbb{Z}\}$, and therefore standard linear methods are not suitable for circular data analysis.

Classic examples include flight direction of birds, wind and ocean current direction. Time of day, or time of year are also obvious candidates for directional modelling. When, along with a direction, we report also the time of the day when it has been recorded, we are collecting two-dimensional circular data. In zoology many multi-dimensional instances arise. For example, Fisher (1993) considers the orientations of the nests of noisy scrub birds along the bank of a creek bed, together with the corresponding directions of creek flow at the nearest point to the nest: here the joint behaviour of these random variables is of interest. Multidimensional circular data are also commonly found in the analysis of protein structure (Lovell et al. 2003). In political science, Gill and Hangartner (2010) study directional party preferences in a two-dimensional ideological space for the German Bundestag elections.

Maximum likelihood estimation is a common approach in many statistical problems, although it requires an assumption that the unknown target belongs to a restricted class of functions. To obtain more general models, Tibshirani and Hastie (1987) introduced the concept of *local likelihood*. They proposed to fit a regression function using only the observations falling within a certain window around the estimation point. In the context of density estimation, local likelihood requires spatially weighting the log-densities. Depending on the smoothing degree, the methodology can be viewed, in practice, as basically parametric or nonparametric.

The log-densities can be modelled in various ways corresponding to various techniques. Hjort and Jones (1996) have established a general framework, where a parametric family is locally modelled, by allowing its parameters change along the support. Loader (1996a) focused on the use of log-polynomials. Eguchi and Copas (1998) proposed an alternative construction and focus on properties related to asymptotics when the smoothing degree is fixed. Delicado (2006) proposed a unified formulation of these local likelihood approaches based on the concept of sample truncation.

In the present paper, we discuss local polynomial likelihood in order to introduce small-biased circular density estimation. The current literature on nonparametric circular density estimation is substantially limited to the classical kernel estimator introduced by Hall et al. (1987), and contributions introducing more sophisticated nonparametric methods, able to arbitrarily reduce the asymptotic bias without asymptotic variance inflation, have, until now, never been systematically studied. Our proposed methods make it possible to employ a priori knowledge about the smoothness of the target density. It seems that in the current literature there is no specific focus on local methods which allow for this. A strong, technical motivation could be that in the circular setting there is no exact counterpart for the concept of kernel order, as in the Euclidean setting. On the other hand, the absence of boundaries in the support of directional distributions (circle, sphere or torus) could make nonparametric estimation less challenging. Therefore, the serious problem of boundary bias is unknown for non-

Euclidean data. However, small bias estimation in nonparametric circular *regression* has been introduced by Di Marzio et al. (2009, 2013).

Recently Di Marzio et al. (2016) have presented a computational study where density estimation based on local polynomial likelihood is investigated for two practical issues not treated in this paper: the impact of the density normalization step when the sample size is moderate; and the effectiveness in identifying the number of population modes.

In Sect. 2 we present the model together with some major features of the estimators, while Sect. 3 is devoted to asymptotic accuracy. In Sect. 4 we show how some numerical aspects can be greatly simplified if the d-fold product ($d \geq 1$) of von Mises densities is used as the weight function. After some asymptotic approximations and interpretations, two new estimators are proposed, which could inspire similar counterparts in the multidimensional Euclidean setting. Section 5 is devoted to numerical experiments where the main theoretical properties are confirmed in small to moderate sample sizes. Finally, Sect. 6 contains a real data example related to the three-dimensional structure of proteins.

2 The model

Let f be a circular population density, i.e. a non-negative, 2π-periodic function with $\int_{[-\pi,\pi)^d} f = 1, d \geq 1$. We want to estimate f at $\boldsymbol{\theta} \in [-\pi, \pi)^d$, using a realization $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ of a random sample $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_n$ drawn from f.

Once the domain of f has been partitioned into S equal cells, say $C_1, \ldots, C_S$, let n_s and P_s denote the cell counts and probabilities, respectively. Due to the mean value theorem we can write $P_s = f(\boldsymbol{\theta}_s)(2\pi)^d/S$ for some $\boldsymbol{\theta}_s \in C_s$. The likelihood is $c \prod_{s=1}^{S} P_s^{n_s}$ subject to $\sum_{s=1}^{S} P_s = 1$, where c is a multinomial coefficient. Using Lagrange multipliers, this leads to a penalized log-likelihood

$$\mathcal{L}(f) = \log c + \sum_{s=1}^{S} \{n_s \log P_s - n P_s\}.$$

Assuming that the number of cells is sufficiently large so that not more that one observation falls in each one, the sum can be taken over the n observations, $n_s = 1$, and P_s can be replaced by P_i, the probability for the cell containing the ith observation.

For $\boldsymbol{\beta} \in [-\pi, \pi)^d$ with jth entry denoted by $\boldsymbol{\beta}^{(j)}$, we define the *kernel* function $K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\beta}) = \prod_{j=1}^{d} K_{\kappa_j}\left(\boldsymbol{\beta}^{(j)}\right)$, where K_{κ_j} is a *circular kernel* with zero mean direction and concentration parameter $\kappa_j \geq 0$; see Definition 1 given by Di Marzio et al. (2011). The weight function K_{κ_j} is usually chosen to be a continuous density function whose support is the circle with the property that as $\kappa_j \to \infty$ the density tends to concentrate at the mode. If we spatially weight each summand of $\mathcal{L}(f)$ by $K_{\kappa_1,\ldots,\kappa_d}$, and approximate the second sum with an integral, ignoring the constant term, we obtain the following definition of a *local* log-likelihood at $\boldsymbol{\theta} \in [-\pi, \pi)^d$,

$$\mathcal{L}_{\boldsymbol{\theta}}(f) = \sum_{i=1}^{n} K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) \log f(\boldsymbol{\theta}_i) - n \int_{[-\pi,\pi)^d} K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) f(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}.$$

The number of observations contributing to the estimate in the jth direction is related to the magnitude of the concentration κ_j.

The main motivation for defining an estimator of $f(\boldsymbol{\theta})$ as the maximizer of $\mathcal{L}_{\boldsymbol{\theta}}(f)$ over f lies in the property $\mathrm{E}_f[\mathcal{L}_{\boldsymbol{\theta}}(f)] \geq \mathrm{E}_f[\mathcal{L}_{\boldsymbol{\theta}}(g)]$ for all non-negative functions g, with equality holding when $f(\boldsymbol{u}) = g(\boldsymbol{u})$, for any $\boldsymbol{u} \in [-\pi, \pi)^d$. This is shown by noting that $x \geq \log x + 1$ if $x > 0$. The condition $x > 0$ extends our methodology also to non-negative regression function estimation.

As a model for $\log f$ consider a (2π-periodic) pth degree sin-polynomial (Di Marzio et al. 2009)

$$\mathcal{P}_p(\boldsymbol{\lambda}) = a_0 + \sum_{s=1}^{p} \frac{(S'_{\boldsymbol{\lambda}})^{\otimes s} \, \boldsymbol{a}_s}{s!},$$

with $\boldsymbol{\lambda} \in \mathbb{R}^d$, $a_0 \in \mathbb{R}$, $\boldsymbol{a}_s \in \mathbb{R}^{d^s}$, $s \in (1, \ldots, p)$, $S_{\boldsymbol{\lambda}} = \left(\sin\left(\lambda^{(1)}\right), \ldots, \sin\left(\lambda^{(d)}\right)\right)'$, and $S_{\boldsymbol{\lambda}}^{\otimes s}$ denoting the sth order Kronecker power of $S_{\boldsymbol{\lambda}}$. We call $\mathcal{P}_p$ a *sin-polynomial* because the functions $\sin^s$ are reminiscent of the monomial bases for ordinary polynomials. Likewise, we associate the terms *linear* and *quadratic*, respectively, to $\mathcal{P}_1$ and $\mathcal{P}_2$. We use sin functions since the absolute value of the sine of a difference depends only on the magnitude of the smallest arc between the two respective points. The use of a simple difference, as for standard local polynomial modelling, would not be suited to angles because it depends on whether the origin belongs to above the smallest arc or not. In both cases, the sign depends on the orientation choice, that is also arbitrary, but this is not relevant due to the symmetry of our weight functions.

Since f is determined by $\boldsymbol{a} = (a_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)'$, we get

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{a}) = \sum_{i=1}^{n} K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) \mathcal{P}_p(\boldsymbol{\theta}_i - \boldsymbol{\theta})$$

$$- n \int_{[-\pi,\pi)^d} K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp(\mathcal{P}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})) \mathrm{d}\boldsymbol{\alpha}.$$

Differentiating $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{a})$ with respect to the elements of $\boldsymbol{a}$, and setting these partial derivatives equal to $\boldsymbol{0}$, leads to $\sum_{s=0}^{p} d^s$ equations:

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{A}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta})$$

$$= \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\mathcal{P}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\alpha}, \tag{1}$$

where $\mathcal{A}(\boldsymbol{\lambda}) = \mathrm{vec}\left(1, S'_{\boldsymbol{\lambda}}, \ldots, \left(S'_{\boldsymbol{\lambda}}\right)^{\otimes p} / p!\right).$
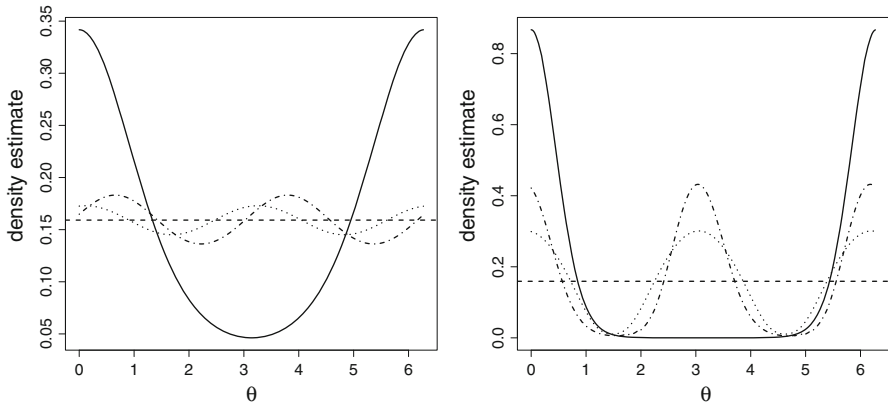
Fig. 1 (Normalized) density estimates for samples of size 100 from a von Mises distribution with concentration parameter 1 (left) and 5 (right), for $\mathcal{P}_0$ (dashed), $\mathcal{P}_1$ (dotted), and $\mathcal{P}_2$, (dash-dot) with smoothing parameter $\kappa = 0$. True density is continuous line

If $\log f$ is smooth enough at $\boldsymbol{\theta}$, the sin-polynomial $\mathcal{P}_p$ represents a series expansion of $\log f$ up to order p, and solving the system of Eq. (1) for $\boldsymbol{a}$ gives the estimates $\hat{\boldsymbol{a}} = (\hat{a}_0, \hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_p)'$ of $\tilde{\boldsymbol{a}} = (\tilde{a}_0, \tilde{\boldsymbol{a}}_1, \ldots, \tilde{\boldsymbol{a}}_p)'$, where, for $\boldsymbol{\theta} \in [-\pi, \pi)^d$, $\tilde{a}_0 = \log f(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{a}}_s$ is the vector of the mixed partial derivatives of total order s of $\log f$ at $\boldsymbol{\theta}$. For example, $\tilde{\boldsymbol{a}}_1$ is the gradient vector, and $\tilde{\boldsymbol{a}}_2 = \text{vec}(\mathbf{H})$, where $\mathbf{H}$ denotes the Hessian matrix. Arguments in Loader (1996a) assure the existence and uniqueness of $\hat{\boldsymbol{a}}$ since cartesian products of circles are compact. Setting $g = \log f$, and $\hat{g}(\boldsymbol{\theta}) = \hat{a}_0$ the density estimate at $\boldsymbol{\theta} \in [-\pi, \pi)^d$ is then given by

$$\hat{f}(\boldsymbol{\theta}) = \frac{\exp(\hat{g}(\boldsymbol{\theta}))}{\int_{[-\pi,\pi)^d} \exp(\hat{g}(\boldsymbol{\theta})) d\boldsymbol{\theta}}. \tag{2}$$

When $p = 0$, formula (2) simplifies to the standard kernel estimator (Di Marzio et al. 2011), whereas for $p > 0$ it generally becomes nonlinear and the denominator is required to make it a *bona fide* density. It is *rotationally invariant*, that is $\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}^*$, where $\hat{\boldsymbol{a}}^*$ is the estimate using translated data $\boldsymbol{\theta}_1 + \boldsymbol{\omega}, \boldsymbol{\theta}_2 + \boldsymbol{\omega}, \ldots, \boldsymbol{\theta}_n + \boldsymbol{\omega}, \boldsymbol{\omega} \in [-\pi, \pi)^d$. Thus, if we rotate the initial direction by $\boldsymbol{\omega}$ the estimate is not affected. This, in circular statistics, is an important property as the choice of the origin is arbitrary.

Maximum smoothing ($\kappa = 0$) yields non-uniform estimates when $p > 0$, whereas, if $p = 0$, it gives $1/(2\pi)$ for any data. When $\kappa = 0$ the contribution of one observation to the estimate does not depend on its location, and this makes the inferential problem fully parametric. Therefore, the use of a constant weight function in the estimate amounts to selecting an element within the parametric family of sin-polynomials we are modelling as a global estimate. When $p = 0$ this family contains only the element $1/(2\pi)$, while $p > 0$ produces a sinusoidal estimate. Figure 1 shows an example using von Mises data, in which the amplitude of the sinusoidal behaviour depends on the population.

3 Accuracy

As a starting point, we establish that asymptotic properties of estimator (2) can be conveniently expressed by referring to those of the estimators of log-densities. This can be seen by a very general argument that requires consistency of the estimator and smoothness of both the population and the estimate. To simplify notation, we initially consider the one-dimensional case.

Using again $g = \log f$, in virtue of Corollary 1, we have that, for large n, $R_n = \hat{g} - g \approx 0$ and so $\exp(g + R_n) \approx f \times (1 + R_n)$. This shows that the rate of convergence of the log-density estimator at $\theta \in [-\pi, \pi)$ does not change when we exponentiate it, whereas its magnitude varies due to the multiplicative factor $f(\theta)$. Clearly, this transformation improves the estimation at the tails, and, more generally, over the regions where $f(\theta) < 1$. Such regions are generally a large part of the support when we note that our densities are continuous functions over $[-\pi, \pi)$. Concerning the convergence rate of the normalized estimator, we see that

$$f \times (1 + R_n) \left(1 - \int f \times R_n\right) = f \times (1 + O(R_n)),$$

so the rate of convergence does not change even after normalization. Coming to the magnitude of the mean integrated squared error (MISE), it is interesting to note that above expression makes it possible to invoke Theorem 1 of Glad et al. (2003). They prove that, when the un-normalized area is bigger than one, then it does not worsen if we add to it a fixed quantity that makes its integral equal to one. This result can be considered very strong because it holds for any sample size. Obviously, if the area of our estimate is smaller than one, this theorem will not apply; however, we can still be confident that severe, often negative, bias at the peaks has been reduced. Comparisons of normalized and un-normalized estimators can be found in Di Marzio et al. (2016).

In general system (1) has only numerical roots when $p > 0$, and direct accuracy calculations are impossible, and so we expand (1) to obtain an expression for the estimation error. We consider un-normalized estimators throughout this section.

3.1 Asymptotics

The starting point for obtaining asymptotic properties is the expansion of system (1), for $\hat{a}$ around $\tilde{a}$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{A}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta})$$

$$- \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) d\boldsymbol{\alpha}$$

$$- \tilde{\mathbf{J}}(\hat{a} - \tilde{a}) + o(\hat{a} - \tilde{a}) = \mathbf{0},$$

where $\tilde{\mathcal{P}}_p(\lambda) = \tilde{a}_0 + \sum_{s=1}^{p}(S'_\lambda)^{\otimes s}\tilde{a}_s/s!$, and $\tilde{\mathbf{J}} = \int_{[-\pi,\pi)^d} \mathcal{A}(\alpha - \theta)\mathcal{A}(\alpha - \theta)' K_{\kappa_1,...,\kappa_d}(\alpha - \theta) \exp\left(\tilde{\mathcal{P}}_p(\alpha - \theta)\right) d\alpha$ is the Jacobian matrix of the local likelihood system at $a = \tilde{a}$. It follows that

$$\hat{a} - \tilde{a} \approx \tilde{\mathbf{J}}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{A}(\theta_i - \theta)K_{\kappa_1,...,\kappa_d}(\theta_i - \theta) \right.$$

$$\left. - \int_{[-\pi,\pi)^d} \mathcal{A}(\alpha - \theta)K_{\kappa_1,...,\kappa_d}(\alpha - \theta)\exp\left(\tilde{\mathcal{P}}_p(\alpha - \theta)\right) d\alpha \right). \qquad (3)$$

Starting from Eq. (3), asymptotic bias and variance, for the case of order $p \geq 1$ of the approximating sin-polynomial, are provided by

Theorem 1 *Define as e_i the (i, i)th entry of $\int_{[-\pi,\pi)^d} K^2_{\kappa_1,...,\kappa_d}(\alpha)\mathcal{A}(\alpha)\mathcal{A}(\alpha)'d\alpha$, and assume that*

(a) $\lim_{n\to\infty} \kappa_j = \infty$ *for $j \in (1, \ldots, d)$;*
(b) $\lim_{n\to\infty} n^{-1}e_i = 0$ *for $i \in (1, \ldots, \sum_{s=0}^{p} d^s)$.*

Moreover, assume that, for odd p, all the mixed derivatives of total order $p + 1$ of the log-likelihood function exist and are continuous in $[-\pi, \pi)^d$, and, for even p, this also holds for all the mixed derivatives of total order $p + 2$, then, we have

(i) *for odd p*

$$\mathrm{E}[\hat{a}] - \tilde{a} \approx \tilde{\mathbf{J}}^{-1} \int_{[-\pi,\pi)^d} \mathcal{A}(\alpha)K_{\kappa_1,...,\kappa_d}(\alpha)(S'_\alpha)^{\otimes p+1} \frac{\tilde{a}_{p+1}}{(p+1)!} d\alpha \, f(\theta)$$

(ii) *for even $p > 0$*

$$\mathrm{E}[\hat{a}] - \tilde{a} \approx \tilde{\mathbf{J}}^{-1} \int_{[-\pi,\pi)^d} \mathcal{A}(\alpha)K_{\kappa_1,...,\kappa_d}(\alpha) \left\{ f(\theta)(S'_\alpha)^{\otimes p+2} \frac{\tilde{a}_{p+2}}{(p+2)!} \right.$$

$$\left. + S'_\alpha D_f(\theta)(S'_\alpha)^{\otimes p+1} \frac{\tilde{a}_{p+1}}{(p+1)!} \right\} d\alpha,$$

where $D_f(\theta)$ denotes the gradient of f at θ, and $\mathbf{1}$ is a vector of ones of length d^{p+1} for odd p, and d^{p+2} for even $p > 0$.
Moreover, for both (i) and (ii)

$$\mathrm{Var}[\hat{a}] \approx \frac{1}{n}f(\theta)\tilde{\mathbf{J}}^{-1} \int_{[-\pi,\pi)^d} \mathcal{A}(\alpha)\mathcal{A}(\alpha)' K^2_{\kappa_1,...,\kappa_d}(\alpha)d\alpha \, \tilde{\mathbf{J}}^{-1}$$

Proof See "Appendix". $\qquad\qquad\qquad\square$

Using results in Theorem 1, component-wise consistency of $(\hat{a}_0, \hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_p)'$ comes from a direct application of Chebychev's inequality, as stated in

Corollary 1 *If assumptions (a) and (b) of Theorem 1 hold, then $\hat{a}_0 \overset{p}{\to} \tilde{a}_0$ and $\hat{\boldsymbol{a}}_s \overset{p}{\to} \tilde{\boldsymbol{a}}_s$ for any $s \in (1, \ldots, p)$.*

Remark 1 If we focus on $\hat{a}_0$, the results in Theorem 1 can be simplified as follows. For a multiindex $\boldsymbol{j} = (j_1, \ldots, j_d)$, and a kernel $K_{\kappa_1, \ldots, \kappa_d}$, set

$$
\eta_{\boldsymbol{j}}(K_{\kappa_1, \ldots, \kappa_d}) = \int\limits_{[-\pi, \pi)^d} K_{\kappa_1, \ldots, \kappa_d}(\boldsymbol{\alpha}) \prod_{i=1}^d \sin^{j_i}\left(\boldsymbol{\alpha}^{(i)}\right) d\boldsymbol{\alpha},
$$

and notice that, due to the symmetry of $K_{\kappa_1, \ldots, \kappa_d}$, $\eta_{\boldsymbol{j}}(K_{\kappa_1, \ldots, \kappa_d}) = 0$ if j_i is odd for at least one $i \in (1, \ldots, d)$. Now, denote as $v^{(\boldsymbol{j})}(\boldsymbol{\theta})$ the mixed partial derivative of total order $|\boldsymbol{j}| = \sum_{i=1}^d j_i$ of a function v at $\boldsymbol{\theta}$, and set $\boldsymbol{i} = (i_1, \ldots, i_d)$ and $g = \log f$. Then, using the results in Theorem 1, along with the approximation

$$
\tilde{\mathbf{J}} \approx f(\boldsymbol{\theta}) \int\limits_{[-\pi, \pi)^d} \mathcal{A}(\boldsymbol{\alpha}) \mathcal{A}(\boldsymbol{\alpha})' K_{\kappa_1, \ldots \kappa_d}(\boldsymbol{\alpha}) d\boldsymbol{\alpha},
$$

the leading term of the bias of $\hat{a}_0$ is

$$
\begin{cases}
\displaystyle\sum_{|\boldsymbol{j}|=(p+1)/2} \eta_{2\boldsymbol{j}}(K_{\kappa_1, \ldots, \kappa_d}) \frac{g^{(2\boldsymbol{j})}(\boldsymbol{\theta})}{(2\boldsymbol{j})!} & \text{for odd } p, \\[3ex]
\displaystyle\sum_{|\boldsymbol{j}|=(p+2)/2} \eta_{2\boldsymbol{j}}(K_{\kappa_1, \ldots, \kappa_d}) \frac{1}{f(\boldsymbol{\theta})} \left\{ \frac{g^{(2\boldsymbol{j})}(\boldsymbol{\theta}) f(\boldsymbol{\theta})}{(2\boldsymbol{j})!} \right. \\[2ex]
\quad \left. + \displaystyle\sum_{|\boldsymbol{i}|=1 : \boldsymbol{i} \leq \boldsymbol{j}} \frac{g^{(2\boldsymbol{j}-\boldsymbol{i})}(\boldsymbol{\theta}) f^{(\boldsymbol{i})}(\boldsymbol{\theta})}{(2\boldsymbol{j}-\boldsymbol{i})!} \right\} & \text{for even } p > 0,
\end{cases}
$$

where $\boldsymbol{j}! = \prod_{i=1}^d j_i!$, and $\boldsymbol{i} \leq \boldsymbol{j}$ means that $i_s \leq j_s$ for each $s \in (1, \ldots, d)$, while, in either case, the leading term of the variance is

$$
\frac{1}{n f(\boldsymbol{\theta})} \int\limits_{[-\pi, \pi)^d} K^2_{\kappa_1, \ldots, \kappa_d}(\boldsymbol{\alpha}) d\boldsymbol{\alpha}.
$$

Remark 2 The above results can be further simplified if $K_{\kappa_1, \ldots, \kappa_d}$ is the d-fold product of von Mises kernels with $\kappa_i = \kappa > 0$ for each $i \in (1, \ldots, d)$, i.e. $K_{\kappa_1, \ldots, \kappa_d}(\boldsymbol{\theta}) = [2\pi I_0(\kappa)]^{-d} \exp\left(\kappa \sum_{j=1}^d \cos\left(\boldsymbol{\theta}^{(j)}\right)\right)$, where $I_s(\cdot)$ is the modified Bessel function of the first kind and order s. Then, for $\boldsymbol{j} \geq \boldsymbol{0}$, it holds that

$$
\eta_{2\boldsymbol{j}}(K_{\kappa_1, \ldots, \kappa_d}) = \prod_{i=1}^d \frac{\mathsf{OF}(2j_i) I_{j_i}(\kappa)}{\kappa^{j_i} I_0(\kappa)}, \quad \text{and}
$$

$$\int\limits_{[-\pi,\pi)^d} K_{\kappa_1,\dots,\kappa_d}^2(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\alpha} = \left[\frac{I_0(2\kappa)}{2\pi I_0^2(\kappa)}\right]^d,$$

where $\mathsf{OF}(z)$ stands for the odd factorial of $z \in \mathbb{Z}^+$, with $\mathsf{OF}(0) = 1$.

For large enough κ, $I_0^d(2\kappa)/I_0^{2d}(\kappa) \approx (\pi\kappa)^{d/2}$ and $I_j(\kappa)/I_0(\kappa)$, $j \in (1,\dots,d)$, can be approximated by 1 with an error of magnitude $O(1/\kappa)$. These approximations, along with the assumptions in Theorem 1, give an asymptotic bias of $O\left(\kappa^{-(p+1)/2}\right)$ for odd p and $O\left(\kappa^{-(p+2)/2}\right)$ for even p, while, in both cases, the asymptotic variance is $O\left(n^{-1}\kappa^{d/2}\right)$. As a consequence, the value of κ which minimizes the asymptotic mean squared error of $\hat{a}_0$ is $O\left(n^{2/(2(p+1)+d)}\right)$ for odd p, and $O\left(n^{2/(2(p+2)+d)}\right)$ for even p, which lead to rates of convergence of orders $n^{-2(p+1)/(2(p+1)+d)}$ and $n^{-2(p+2)/(2(p+2)+d)}$, respectively.

As previously noted, when $p = 0$ system (1) has a closed form solution. This allows direct calculations, without using Theorem 1. For $d = 1$, the leading terms of bias and variance are, respectively, $1/2 f''(\theta)/f(\theta) \int \sin^2 K_\kappa$, and $1/(nf(\theta)) \int K_\kappa^2$. Because the local linear fit has the bias leading term equal to $1/2 \left(f''(\theta)/f(\theta) - f'(\theta)^2/f(\theta)^2\right) \int \sin^2 K_\kappa$ and the same variance, the respective convergence rates are the same. Therefore, apart from the stationary points, the $\mathcal{P}_0$ fit is asymptotically superior to the (un-normalized) $\mathcal{P}_1$ one where the population density is concave, as is usually the case around the modes.

The previous results can be formulated for $\hat{f}$ instead of $\hat{a}_0$. When $d = 1$, the leading terms of the biases of the un-normalized estimates, up to order two, are

$$\begin{cases} \dfrac{f''(\theta)}{2} \displaystyle\int_{-\pi}^{\pi} \sin^2(u) K_\kappa(u)\mathrm{d}u & \text{if } p = 0 \\[4mm] \dfrac{1}{2}\left(f''(\theta) - \dfrac{f'(\theta)^2}{f(\theta)}\right) \displaystyle\int_{-\pi}^{\pi} \sin^2(u) K_\kappa(u)\mathrm{d}u & \text{if } p = 1 \\[4mm] \dfrac{2f'(\theta)^4 - 3(f(\theta)f''(\theta))^2 + f(\theta)^3 f^{(4)}(\theta)}{4!f(\theta)^3} \displaystyle\int_{-\pi}^{\pi} \sin^4(u) K_\kappa(u)\mathrm{d}u & \text{if } p = 2, \end{cases}$$

whereas the asymptotic variances are all equal to $f(\theta)/n \int K_\kappa^2$.

3.2 Smoothing degree selection

In order to select the smoothing degree, we prefer likelihood cross-validation since it does not require explicit estimation of higher order derivatives, as happens for any *plug-in* approach, and explicitly takes account of the risk function we use for our sin-polynomial modelling. We start with a caveat as follows. The local likelihood estimator is nonlinear in nature when $p > 0$. Consequently, when the smoothing parameter(s) is (are) fixed, if $\hat{f}_i$ are the normalized estimates from N samples of size n_i, then the (normalized) estimate using all the data from the combined samples is *not* the same as $\sum_{i=1}^{N} n_i \hat{f}_i / \sum n_i$, as would be the case for $p = 0$. This anomaly, which leads to an increased computational burden when cross-validation is used, is also apparent in the Euclidean setting.

We could use a normalized estimate, or just penalize the un-normalized one. Under the first perspective, the likelihood cross-validation criterion for density estimation suggests maximizing the leave-one-out log-likelihood

$$\sum_{i=1}^{n} \left\{ \log \hat{f}_{-i}(\boldsymbol{\theta}_i) - \log \int \hat{f}_{-i}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} \right\}$$

over $\{\kappa_1, \ldots, \kappa_d\}$, where $\hat{f}_{-i}$ indicates an estimate obtained after removing the ith observation. The second approach leads to a penalized likelihood given by the target function $\sum_{i=1}^{n} \log \hat{f}_{-i}(\boldsymbol{\theta}_i) - \lambda \left(\int \hat{f}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} - 1 \right)$, where λ is some penalty; here, the difficulty lies in choosing an appropriate λ.

The first approach appears more direct, but turns out to be very computationally intensive; this is a consequence of the caveat explained above. However, passing to the logarithm we can approximate the second term by $n \log \int \hat{f}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}$. Noting that $\int \hat{f} \approx 1$, a Taylor series approximation of the logarithm leads to

$$\sum_{i=1}^{n} \log \hat{f}_{-i}(\boldsymbol{\theta}_i) - n \left(\int_{[-\pi,\pi)^d} \hat{f}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} - 1 \right). \tag{4}$$

This can be seen as the same as a penalized likelihood when $\lambda = n$. Formula (4) has been presented by Loader (1996b, p. 90) as a direct application of standard cross-validation to his log-likelihood model $\sum \log f(X_i) - n \left(\int f(u) \mathrm{d}u - 1 \right)$, that is slightly different from our $\mathcal{L}(f)$.

4 Computational aspects and interpretation

System (1) is nonlinear and contains a number of integrals; hence, closed-form solutions are in general unavailable. Nevertheless, when products of von Mises densities are used as kernels, it is possible to alleviate this issue.

In Sect. 4.1 we indicate a way to avoid numerical integration based on the properties of Bessel functions when $\mathcal{P}_1$ is used. This strategy does not apply for higher order sin-polynomials ($p > 1$) because cross-terms do not allow us to obtain separable integrals. Even avoiding cross-terms would not work since the resulting integrals do not have any explicit expression. In Sect. 4.2, based on asymptotic arguments, we present, for $\mathcal{P}_1$ and $\mathcal{P}_2$ fits, a simple way to obtain closed-form solutions without resorting to numerical integration.

4.1 Local linear fit

A local linear fit for f at $\boldsymbol{\theta} \in [-\pi, \pi)^d$ can be obtained starting from the solution for a_0, of $d + 1$ equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{A}(\boldsymbol{\theta}_i - \boldsymbol{\theta})K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) = \int\limits_{[-\pi,\pi)^d}\mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta})K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta})$$

$$\times \exp\left(a_0 + \sum_{j=1}^{d}a_1^{(j)}\sin\left(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)d\boldsymbol{\alpha}.$$

(5)

We will denote the quantities on the LHS by the statistics

$$M_0 = \frac{1}{n}\sum_{i=1}^{n}K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta})$$

(6)

and, for $j \in (1,\dots,d)$,

$$M_p^{(j)} = \frac{1}{n}\sum_{i=1}^{n}\sin^p\left(\boldsymbol{\theta}_i^{(j)} - \boldsymbol{\theta}^{(j)}\right)K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta}).$$

(7)

Using a von Mises kernel, and denoting $\{(2\pi)^d\prod_{j=1}^{d}I_0(\kappa_j)\}^{-1}$ by $\mathcal{B}$, the quantities in the RHS of system (5), respectively, become

$$\exp(a_0)\mathcal{B}\prod_{j=1}^{d}\int_{-\pi}^{\pi}\exp\left(\kappa_j\cos\left(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)\exp\left(a_1^{(j)}\sin\left(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)d\boldsymbol{\alpha}^{(j)},$$

and

$$\exp(a_0)\mathcal{B}\int_{-\pi}^{\pi}\exp\left(\kappa_i\cos\left(\boldsymbol{\alpha}^{(i)} - \boldsymbol{\theta}^{(i)}\right)\right)\exp\left(a_1^{(i)}\sin\left(\boldsymbol{\alpha}^{(i)} - \boldsymbol{\theta}^{(i)}\right)\right)$$

$$\times \sin\left(\boldsymbol{\alpha}^{(i)} - \boldsymbol{\theta}^{(i)}\right)d\boldsymbol{\alpha}^{(i)}\prod_{j\neq i}^{d}\int_{-\pi}^{\pi}\exp\left(\kappa_j\cos\left(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)$$

$$\times \exp\left(a_1^{(j)}\sin\left(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)d\boldsymbol{\alpha}^{(j)},$$

for $i \in (1,\dots,d)$. Hence, expressing the integrals as Bessel functions, the above quantities can be, respectively, rewritten as

$$\exp(a_0)\mathcal{B}(2\pi)^d\prod_{j=1}^{d}I_0\left(\left\|\left(\kappa_j \ a_1^{(j)}\right)\right\|\right),$$

and

$$\exp(a_0)\mathcal{B}(2\pi)^d I_1\left(\left\|\left(\kappa_i \ \boldsymbol{a}_1^{(i)}\right)\right\|\right) \sin\left(\operatorname{atan2}\left(\boldsymbol{a}_1^{(i)}, \kappa_i\right)\right) \prod_{\substack{j=1 \\ j\neq i}}^{d} I_0\left(\left\|\left(\kappa_j \ \boldsymbol{a}_1^{(j)}\right)\right\|\right),$$

where $\operatorname{atan2}(y, x)$ gives the angle (in radians) between the x-axis and the vector from the origin to (x, y). Then, taking the ratio gives

$$\frac{\boldsymbol{M}_1^{(j)}}{M_0} = \frac{I_1\left(\left\|\left(\kappa_j \ \boldsymbol{a}_1^{(j)}\right)\right\|\right) \sin\left(\operatorname{atan2}\left(\boldsymbol{a}_1^{(j)}, \kappa_j\right)\right)}{I_0\left(\left\|\left(\kappa_j \ \boldsymbol{a}_1^{(j)}\right)\right\|\right)}.$$

As for the existence conditions, due to the circular kernel definition, $M_0 > 0$. This quantity has to be solved in order to obtain $\hat{\boldsymbol{a}}_1^{(j)}$. Such an approach gives the numerical solutions for all the partial derivatives ($j \in (1, \ldots, d)$). Finally, substituting these into the first equation of system (5), we obtain

$$\exp(\hat{a}_0) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\prod_{j=1}^{d}\exp\left(\kappa_j \cos\left(\boldsymbol{\theta}_i^{(j)} - \boldsymbol{\theta}^{(j)}\right)\right)\right]}{(2\pi)^d \prod_{j=1}^{d} I_0\left(\left\|\left(\kappa_j \ \hat{\boldsymbol{a}}_1^{(j)}\right)\right\|\right)}. \tag{8}$$

This expression suggests that $\mathcal{P}_1$ modelling can be seen as a correction of the kernel density estimator which basically reduces the estimate where the density gradient has nonzero norm, and leaves it unchanged at the maxima and minima. Thus, if κ_j are the same both for $p = 1$ and $p = 0$, the un-normalized area of the case $p = 1$ is strictly smaller than one. Hence, normalization would result in bias reduction (increase) near the mode (along the valleys) and in variance inflation, drastically contrasting with the un-normalized fit that has same bias as case $p = 0$ at stationary points.

4.2 Asymptotic approach

Closed-form solutions for system (1) do not exist for usual circular kernels if $p > 0$. This is in contrast with the Euclidean setting, where the use of the Gaussian kernel makes them available if we implement $\mathcal{P}_1$ and $d \in (1, 2)$, or $p = 2$ and $d = 1$. Hjort and Jones (1996) report them; see their formulas (5.1), (5.2) and (7.3). In this section we obtain closed-form approximate solutions when the von Mises kernel is used by appealing to some asymptotic arguments.

Like any Euclidean kernel, as n increases circular kernels also concentrate, giving significant weight only to those observations which are close to the estimation point. For large sample sizes, this allows these approximations

$$\begin{cases} \cos(u) \approx 1 - \frac{1}{2}\operatorname{atan2}(\sin(u), \cos(u))^2 \\ I_0(\kappa) \approx (2\pi\kappa)^{-1/2}\exp(\kappa) \\ \sin(u) \approx \operatorname{atan2}(\sin(u), \cos(u)). \end{cases} \tag{9}$$

The resulting functions lead to closed forms for both integrals and estimators when $\mathcal{P}_1$ or $\mathcal{P}_2$ are modelled. We use periodic functions since the difference between two angles can fall outside the interval $[-\pi, \pi)$. Numerically, it could be convenient to rescale the kernel in order to avoid the weight values diverging too much which would affect stability. For this reason, we have nevertheless included a scale factor, although $\int K_{\kappa_j} \neq 1$.

Result 1 *Consider the log-likelihood system* (1) *with* $p = 1, d \geq 1$, *and a d-fold product of von Mises densities as the weight function. The use of approximations* (9) *within the integrands in the RHSs of the system leads to the following closed form solutions*:

$$\hat{a}_0 = \log M_0 - \frac{1}{2} \sum_{j=1}^{d} \kappa_j \left(\frac{M_1^{(j)}}{M_0} \right)^2, \tag{10}$$

$$\hat{a}_1^{(j)} = \kappa_j \frac{M_1^{(j)}}{M_0}, \quad for\ j \in (1, \dots, d), \tag{11}$$

with M_0 *and* $M_1^{(j)}$ *defined in Eqs.* (6) *and* (7).

Since the $\hat{a}_0$ and $\hat{a}_1^{(j)}$ formulations do not have an intuitive nature, it is of interest to further examine their structure. They are consistent for $\tilde{a}_0$ and $\tilde{a}_1^{(j)}$, respectively, after observing the limits $M_0 \xrightarrow{P} f(\boldsymbol{\theta})$ and $M_1^{(j)} \xrightarrow{P} \partial f(\boldsymbol{\theta})/\partial \theta^{(j)} \int \sin^2 \left(\alpha^{(j)} \right) K_{\kappa_j} \left(\alpha^{(j)} \right) d\alpha^{(j)}$. This latter integral, for large κ_j, is approximately equal to $1/\kappa_j$, which is consistent with the rate of convergence in the linear case.

Result 2 *Consider the log-likelihood system* (1) *with* $p = 2, d \geq 1$, *and a d-fold product of von Mises densities as the weight function. Using the approximations* (9) *leads to these expressions for the RHS of this system*:

$$\int\limits_{[-\pi,\pi)^d} K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\mathcal{P}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) d\boldsymbol{\alpha} \approx D,$$

$$\int\limits_{[-\pi,\pi)^d} \left(S'_{\boldsymbol{\alpha}-\boldsymbol{\theta}}\right)^{\otimes 1} K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\mathcal{P}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) d\boldsymbol{\alpha} \approx D\mathbf{C}^{-1}\boldsymbol{a}_1,$$

$$\int\limits_{[-\pi,\pi)^d} \left(S'_{\boldsymbol{\alpha}-\boldsymbol{\theta}}\right)^{\otimes 2} K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\mathcal{P}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) d\boldsymbol{\alpha}$$

$$\approx D\text{vec}\left(\mathbf{C}^{-1} + \mathbf{C}^{-1}\boldsymbol{a}_1\boldsymbol{a}_1'\mathbf{C}^{-1}\right),$$

where D indicates the following quantity

$$\exp\left(a_0 - \boldsymbol{a}_1'\boldsymbol{\theta} - \frac{1}{2}\left(\boldsymbol{\theta}'\mathbf{C}\boldsymbol{\theta} + (\boldsymbol{a}_1 + \mathbf{C}\boldsymbol{\theta})'\,\mathbf{C}^{-1}\,(\boldsymbol{a}_1 + \mathbf{C}\boldsymbol{\theta})\right)\right) \det(\mathbf{C})^{-1/2} \prod_{j=1}^{d} \kappa_j^{1/2},$$

and $\mathbf{C} = \mathrm{diag}(\kappa_1, \ldots, \kappa_d) - \mathbf{A}$, *with*

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_2^{(1)} & \cdots & \boldsymbol{a}_2^{(d)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{a}_2^{(d(d-1)+1)} & \cdots & \boldsymbol{a}_2^{(d \times d)} \end{pmatrix}.$$

If $d = 1$ then the above RHSs give these closed-form solutions:

$$\hat{a}_0 = \frac{1}{2} \left\{ \frac{M_1^2}{M_1^2 - M_0 M_2} + \log \left(\frac{M_0^4}{\kappa (M_0 M_2 - M_1^2)} \right) \right\}, \tag{12}$$

$$\hat{a}_1 = \frac{M_0 M_1}{M_0 M_2 - M_1^2}, \tag{13}$$

$$\hat{a}_2 = \kappa + \frac{M_0^2}{M_1^2 - M_0 M_2},$$

with $M_p = 1/n \sum_{i=1}^n \sin^p(\theta_i - \theta) K_\kappa(\theta_i - \theta)$ for $p \in (0, 1, 2)$.

The existence condition $M_1^2 - M_0 M_2 < 0$ is asymptotically satisfied since $M_1^2 = O(1/\kappa^2)$ and $M_0 M_2 = O(1/\kappa)$. A check of their consistency requires to additionally know that $M_2 \xrightarrow{p} f(\theta) \int \sin^2 K_\kappa$. A brief examination also reveals that estimator (13) has first-order approximation exactly equal to $d \log f(\theta)/d\theta$, differently from that seen for (11), where this is only asymptotically true. Substituting Eq. (11) into (10) we can write

$$\hat{a}_0 = \log M_0 - \frac{1}{2} \frac{\hat{a}_1^2}{\kappa},$$

and then model $\mathcal{P}_1$ can be described as the kernel estimator plus a correction based on slope. Similarly, model $\mathcal{P}_2$ can be written as

$$\hat{a}_0 = \log M_0 - \frac{1}{2} \frac{\widehat{a_1^2}}{\kappa} + \log \sqrt{1 - \frac{\hat{a}_2}{\kappa}}, \tag{14}$$

where $\widehat{a_1^2}$, which is a different quantity from $\hat{a}_1^2$, denotes a consistent estimator (via Slutsky's theorem) of $\tilde{a}_1^2$ given by the product of the derivative estimators (11) and (13). This formulation suggests that the quadratic estimator can be seen as a correction of a "linear" estimator where the bias introduced both at the minima and peaks is reduced by a logarithmic term involving second derivative estimation. Specifically, slope and curvature corrections have the same magnitude for big κ, and their relative impact on the estimate is described by the ratio $\widehat{a_1^2}/\hat{a}_2$. Also, $\mathcal{P}_2$ fits tend to have bigger area than linear ones when f is convex nearby the maxima, as often happens.

The above considerations lead to a couple of new estimators of $\tilde{a}_0$ which do not have Euclidean counterparts. A promising competitor of the linear fit (10), based on

the fact that estimator (13) is more efficient than its counterpart (11), could be:

$$L_0 = \log M_0 - \frac{1}{2} \sum_{j=1}^{d} \frac{\boldsymbol{M}_1^{(j)2}}{M_0 \boldsymbol{M}_2^{(j)} - \boldsymbol{M}_1^{(j)2}}.$$

A multidimensional, quadratic estimator can be conceived as a direct generalization of the one-dimensional case:

$$Q_0 = L_0 + \frac{1}{2} \log \sum_{j=1}^{d} \frac{M_0^2}{\kappa_j \left(M_0 \boldsymbol{M}_2^{(j)} - \boldsymbol{M}_1^{(j)2} \right)}.$$

5 Simulations

Firstly, we examine the efficiency of our (normalized) estimators on 200 samples of n observations drawn from a von Mises population with null mean direction and concentration parameter equal to five ($vM(0, 5)$). Figure 2 shows the estimated log(MISE) for $n = 100$ and $n = 500$. It can be seen that the approximation of $\exp(\hat{a}_0)$ using Eq. (10) is very good for both values of n. As expected, a larger κ (corresponding to less smoothing) is required for larger sample sizes. Despite their asymptotic equivalence, the use of $\mathcal{P}_1$ improves on the standard kernel density estimate, whereas the $\mathcal{P}_2$ performance is even better. This is despite the fact that the normalization step has a bigger (beneficial) impact for $p = 1$ than for $p = 2$ in terms of bias reduction. Overall, the estimator L_0 has the best performance.

In the second simulation experiment, we consider some mixture distributions. In the first case, we use an equal mixture between wrapped Cauchy with mean direction 0 and concentration 0.225 (WC(0, 0.225)) and uniform (UC($-\pi, \pi$)) distributions. This
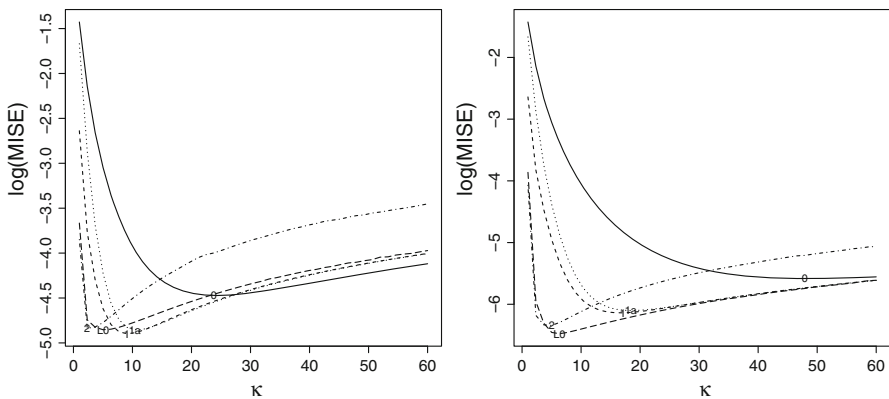

Fig. 2 log(MISE) for a range of values of κ for $p = 0$ (solid), $p = 1$ using Eq. (8) (dashed), $p = 1$ using Result 1 (dotted), L_0 (longdash) and $p = 2$ using Eq. (12) (dotdash) for 200 samples of size $n = 100$ (left) and $n = 500$ (right) from a $vM(0, 5)$
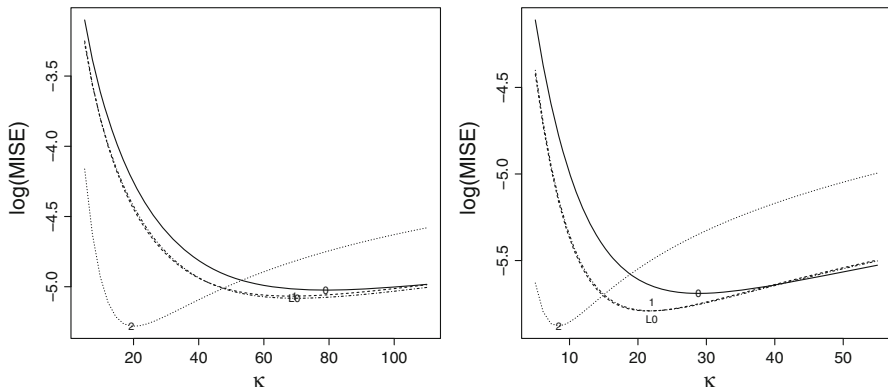
Fig. 3 log(MISE) for a range of values of κ for $p = 0$ (solid), $p = 1$ using Eq. (8) (dashed), L_0 (dotdash) and $p = 2$ using Eq. (12) (dotted) for 200 samples of size $n = 500$ from equal mixture of a WC(0, 0.225) and a uniform distribution (left) and an equal mixture of $vM(\pm \pi/3, 5)$ (right)

population model is not as well behaved as a von Mises one as it has very thick tails, and the integral of the squared second derivative is larger. In such a context, we may expect that the case $p = 2$ would be superior to other models. Simulations confirm that this is indeed the case; see Fig. 3. The second example uses a equal mixture of von Mises densities to form a bimodal distribution. In this case, also, $p = 2$ is the best, with $p = 0$ the worst.

The final experiments are designed to gain knowledge about practical performance in the case that the smoothing degree is data-driven. We consider sixteen models, eight of which are bivariate. They are unimodal or multimodal and are more or less (rotationally) symmetric around the origin. We use, for each sample, two bandwidth selectors: the maximizer of likelihood cross-validation (LCV) given by Eq. (4) (which uses an un-normalized estimator); and classical least-squares cross-validation (LSCV).

Computations were made using MATLAB, and some example code is available from http://www1.maths.leeds.ac.uk/~charles/TESTpaper/matlab.zip. The results are presented in terms of average integrated squared errors evaluated using normalized estimates; see Tables 1 (univariate populations) and 2 (bivariate populations). After noting that the relative merits of the estimators do not depend on which selector has been used, the main message is that the standard kernel is the worst density estimator, even with $n = 100$, when asymptotic performance is not relevant. First-order fits, i.e. $\mathcal{P}_1$, always have satisfactory performance because the bias reduction at the peaks, which is mainly due to normalization, is often decisive. From additional results not reported in Table 1, it appears that estimator L_0 seems better suited than $\mathcal{P}_1$ when the population shape is simpler but their performances are very similar.

Quadratic modelling behaves unexpectedly well, being the best one in the majority of cases. In general, it is more efficient when the roughness is more pronounced. Indeed, in the case of the uniform population, which can be considered as a proper counterexample in that all derivatives are zero, $\mathcal{P}_0$ and $\mathcal{P}_1$ fits have very similar behaviour (for both the smoothing degree selectors) which outperform Q_0.

Table 1 Average integrated squared errors (1000×) over 200 samples of sizes 100 or 500 drawn from various univariate population models (*WN* wrapped normal)

Population model	Shape	n	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{Q}_0$
$UC(-\pi, \pi)$		100	3.02	2.94	7.89
			(3.20)	(3.12)	(3.79)
		500	0.68	0.67	2.09
			(0.69)	(0.70)	(0.73)
$vM(0, 5)$	U, S	100	15.43	10.98	9.89
			(17.59)	(12.89)	(10.80)
		500	4.66	2.77	2.23
			(4.82)	(3.14)	(2.67)
$\frac{1}{2}WC(0, 0.225) + \frac{1}{2}UC(-\pi, \pi)$	U, S	100	40.21	38.37	32.19
			(35.75)	(34.12)	(31.74)
		500	20.08	18.87	16.15
			(17.93)	(16.94)	(15.75)
$\frac{3}{5}WN(0, 1) + \frac{2}{5}WN(1, 0.5)$	U, A	100	9.50	8.12	8.54
			(11.11)	(9.91)	(10.35)
		500	2.75	2.47	2.18
			(3.08)	(2.83)	(2.44)
$\frac{1}{2}WN(0, 0.3) + \frac{1}{2}WN(1, 0.3)$	B, S	100	22.07	22.30	24.27
			(23.73)	(23.90)	(23.73)
		500	6.22	6.46	5.86
			(6.57)	(6.34)	(5.65)
$\frac{1}{2}WN(0, 0.3) + \frac{1}{2}WN(2, 0.6)$	B, A	100	19.39	18.51	18.65
			(21.36)	(20.07)	(18.46)
		500	5.32	4.82	4.25
			(5.54)	(4.95)	(4.35)
$\frac{1}{4}WN(-2, 0.3) + \frac{1}{2}WN(0, 0.3)$ $+ \frac{1}{4}WN(2, 0.3)$	T, S	100	20.89	17.88	20.53
			(22.95)	(19.13)	(21.13)
		500	5.82	4.42	4.83
			(6.16)	(4.53)	(4.74)
$\frac{1}{5}WN\left(\frac{2\pi}{5}, 0.2\right) + \frac{1}{5}WN\left(\frac{4\pi}{5}, 0.2\right)$ $+ \frac{1}{5}WN\left(\frac{6\pi}{5}, 0.2\right) + \frac{1}{5}WN\left(\frac{8\pi}{5}, 0.2\right)$ $+ \frac{1}{5}WN(2\pi, 0.2)$	F, S	100	27.18	25.19	29.17
			(28.47)	(25.82)	(28.89)
		500	8.12	6.91	7.73
			(8.37)	(6.77)	(7.44)

Smoothing degree is selected by likelihood (least-squares) cross-validation

U unimodal, *B* bimodal, *T* trimodal, *F* five-modes, *S* rotationally symmetric, *A* rotationally asymmetric

Table 2 Average integrated squared errors ($1000\times$) over 200 samples of sizes 100 or 500 drawn from various bivariate population models (BvM stands for bivariate von Mises by Singh et al. (2002))

Population model	Shape	n	$\mathcal{P}_0$	$\mathcal{P}_1$	Q_0
WN(0, 1) × WN(0, 1)	U, S	100	4.46	3.41	3.92
			(4.98)	(3.86)	(4.32)
		500	1.75	1.07	1.05
			(1.80)	(1.10)	(1.11)
$vM(0, 1) \times vM(0, 10)$	U, A	100	20.15	14.31	9.00
			(22.41)	(16.53)	(11.56)
		500	7.41	4.37	2.33
			(7.57)	(4.57)	(2.60)
$\left(\frac{3}{5}WN(0, 1) + \frac{2}{5}WN(1, .5)\right) \times$	U, A	100	8.95	7.94	8.23
			(10.28)	(9.46)	(9.79)
$\times \left(\frac{3}{5}WN(0, 1) + \frac{2}{5}WN(1, .5)\right)$		500	6.06	5.55	5.05
			(7.50)	(7.24)	(7.16)
$\frac{1}{2}$(WN(0, 0.5) × WN(0, 0.5)) + $\frac{1}{2}$(WN(π, 0.5)	B, A	100	15.25	10.97	12.04
× WN(π, 0.5))			(15.43)	(11.28)	(11.92)
		500	5.62	3.24	2.62
			(5.77)	(3.29)	(2.58)
$\frac{1}{2}$(WN(0, 0.4) × WN(0, 0.7)) + $\frac{1}{2}$(WN(π, 0.5)	B, A	100	13.78	10.07	10.57
× WN(π, 0.6))			(14.69)	(11.04)	(10.75)
		500	5.20	2.98	2.29
			(5.34)	(3.18)	(2.35)
$\frac{1}{2}$(WN(0, 0.5) + WN($\frac{\pi}{2}$, 0.5)) × WN(0, 0.5)	B, A	100	14.02	11.15	13.43
			(11.31)	(10.40)	(15.60)
		500	5.22	3.97	5.66
			(5.33)	(4.09)	(4.28)
$BvM(0, 0, 1.9, 1.9, 2)$	U, S	100	8.96	6.60	6.81
			(9.26)	(7.00)	(7.24)
		500	3.53	2.14	2.22
			(3.32)	(2.20)	(2.15)
$BvM(0, 0, 4.9, 4.9, 5)$	U, S	100	21.43	14.82	15.63
			(23.43)	(16.16)	(17.12)
		500	8.34	4.69	5.57
			(8.26)	(4.82)	(4.72)

Smoothing degree is selected by likelihood (least-squares) cross-validation

U unimodal, B bimodal, S rotationally symmetric, A rotationally asymmetric

Table 3 Number of matches of estimator, over 200 samples, between optimal ISE estimator and optimal LCV (or LSCV) function

Population model	n	LCV	LSCV
$vM(0, 5)$	100	20	75
	500	25	96
$\frac{1}{2}\mathrm{WN}(0, 0.3) + \frac{1}{2}\mathrm{WN}(2, 0.6)$	100	48	72
	500	71	102
$vM(0, 5) \times vM(0, 5)$	100	13	64
	500	29	88
$\frac{1}{2}(\mathrm{WN}(0, 0.4) \times \mathrm{WN}(0, 0.7)) + \frac{1}{2}(\mathrm{WN}(\pi, 0.5) \times \mathrm{WN}(\pi, 0.6))$	100	4	48
	500	31	78

Top, univariate samples; bottom, bivariate samples

The saddle-shaped population (bottom three in Table 2) has large regions where the asymptotic bias is mainly due only to first-order properties of the model since second derivatives have different signs at opposite sides. In such case, variance inflation of the quadratic estimator dominates its bias reduction and the overall performance degrades. The last two models in Table 2 deserve particular attention because they concern correlated variables. Specifically, we use the bivariate von Mises model proposed by Singh et al. (2002). Similar to a bivariate Gaussian family, we have five parameters: two locations, two concentrations, and a "correlation". Our two cases— both bimodal—are, respectively, featured by small concordance (concentrations equal to 1.9 and correlation equal to 2), and moderate concordance (concentrations equal to 4.9 and correlation equal to 5). We see that in these cases our proposals are by far superior to the standard kernel method reaching an improvement of nearly 40% in the case of moderate concordance.

A comparison between the two smoothing selectors shows that LCV performs a little better in the majority of cases, although for $n = 500$ they appear almost equivalent. The fact that we measure discrepancy by squared errors suggests that we still could try to select at least p by LSCV. In order to investigate the effectiveness of this approach, we have considered, for each sample of the previous simulation, the four integrated squared error curves as functions of κ (one for each estimator), and then selected the p associated to the smallest minimum. Then we have checked if such p is the same as the one optimizing LSCV or LCV curves over κ. In Table 3 we report the number of such matches for a few populations. Since the results suggest that LSCV is the best in this regard, we might thus envisage choosing the estimator based on the optimal LSCV, and then choosing the smoothing parameter for that estimator based on LCV. This idea is taken up further in the next section.

6 A real data case study

A protein is comprised of a chain of 20 different amino acids, which join together along a "backbone". The atoms on the backbone are a sequence of three atoms (nitrogen–

carbon–carbon) to which other atoms are linked. In particular, to one of the carbon atoms a "side chain" is formed, and the structure of this side chain defines the type of amino acid. The sequence of dihedral angles (ϕ, ψ, ω) along the backbone is determined by the relative positions of the atoms, and this determines the overall shape of the protein after folding. Since ω is highly predictable, with little variation, a study of $-\pi < \phi, \psi < \pi$ is useful for many purposes, particularly in validation of newly determined protein structures.

A plot of pairs $(\phi_i, \psi_i), i = 1, \ldots$—known as a Ramachandran plot—for any protein reveals several subgroups, and mixture models of von Mises distributions have been used to summarize these from a parametric perspective. Work by Mardia et al. (2007) used an EM algorithm to fit a mixture, with the number of components being determined by AIC. Lennox et al. (2010) proposed a Dirichlet process to determine the number of components. A related Bayesian approach was developed by Boomsma et al. (2008) which allowed for many more components in a hidden Markov model, in which mixtures were trained on a set of angles from each amino acid. Kernel density estimation has also been used on the protein data (Taylor et al. 2012) where interest lay in considering a bivariate density estimate conditional on the amino acid type. This has been considered as an alternative approach to validation (Lovell et al. 2003), in which PROCHECK—based on histograms—is currently used. Finally, we note that Fernández-Durán and Gregorio-Domínguez (2016) have anaylzed similar datasets using trigonometric sums, with visual results that exhibit a periodic structure.

In this case study, we use data from 500 high-quality, "representative" proteins available from the Richardson Laboratory.[1] Each protein is represented by a sequence of bivariate angles, each of which is associated with an amino acid. These are pooled together, and we thus obtain 20 datasets corresponding to the 20 amino acids. It should be noted that, within a protein, we expect observations not to be independent. However, in each of the 20 datasets we are pooling data from 500 unrelated proteins, and so strong dependence between observations in the same dataset will only occur when an amino acid occurs consecutively along the backbone; this is relatively uncommon. For each dataset, we can obtain a bivariate density estimate using the methods described in this paper, in which the smoothing parameter is selected by cross-validation. Our objective is to examine the differences in the estimates, and to see how these results may relate to previous work.

Obviously, we are unable to compare density estimates with the true densities for these data, as will be the case in any real-life application. One of the uses of density estimation is to obtain information about subgroups, or clusters, within the data. This could be subsequently used in the fitting of (parametric) models, for example. A natural way to investigate this is to identify the location (and height) of local modes. The ability of our estimators to identify bumps has been discussed, by presenting extensive simulation evidence, in Di Marzio et al. (2016). This motivates our focus on such data, where amino acid distributions are partially characterized by modes. As it is customary in peak recognition, we applied two filters, a global and a local

---

[1] https://www.kinemage.biochem.duke.edu/databases/top500.php.

Table 4 Amino acid recognition of subgroups through bivariate density estimation

Amino	n	$\mathcal{P}_0$	$\mathcal{P}_1$	L_0	Q_0
A	8979	10 (5.53) *23.7*	8 (5.13) *25.0*	5 (3.26) *8.9*	5 (4.17) *15.6*
G	8334	18 (1.27) *1.1*	10 (1.12) 2.2	5 (0.84) *0.8*	6 (1.07) *1.4*
I	5570	6 (3.96) *8.6*	3 (3.72) *8.8*	2 (1.70) *1.4*	3 (2.71) *3.8*
N	4796	13 (1.68) *1.5*	9 (1.54) *1.6*	4 (0.91) *0.4*	6 (1.30) *0.9*

Sample sizes, number of modes (maximum), and *roughness* in the estimates using various methods. The smoothing parameters are selected by likelihood cross-validation

one, in order to reduce false positives. So, for a global threshold for the modes of estimator $E \in \{\mathcal{P}_0, \mathcal{P}_1, L_0, Q_0\}$ we used $c_0 m_E$ where $c_0 = 0.005$, and m_E indicates the maximum among the estimates made over an equispaced grid (of 50×50) locations using E. Secondly, to avoid locations which were more akin to saddle points, we required that, at a mode, say (ϕ_m, ψ_m), $\hat{f}(\phi_m, \psi_m) - \max_{\delta_m} \hat{f} > c_1 m_E$, where δ_m represents the set of neighbouring points around a mode, and $c_1 = 0.0001$. Table 4 gives the maximum of $\hat{f}$, the number of modes, and the roughness for some of the amino acids. It can be seen that $\mathcal{P}_0$ has the largest maxima and the largest number of modes, but that $\mathcal{P}_1$ has the largest roughness. Correspondingly, L_0 has the lowest maximum, the fewest number of modes and the smallest roughness, with Q_0 generally being closer to L_0. We found that the number of elements in the union of mode locations, over all 20 amino acid datasets, for each of the methods in Table 4 is: 111, 74, 43 and 38, respectively. We note that the hidden Markov model of Boomsma et al. (2008) used 50 components in a mixture of bivariate von Mises distributions which seems to be consistent with these values, except for the standard kernel.

The most striking difference, however, is in the visual appearance of the density estimates, in which the difference in roughness is very evident. We first note that the height of the highest mode is much greater than the rest of the density, so in order to visualize the whole density estimate we have taken cube roots throughout. To illustrate the differences, we focus on two amino acid datasets: Alanine (A), and Asparagine (N). Figures 4 and 5 show (transformed) contour plots of estimates $\mathcal{P}_0$ and Q_0, as well as slices through the density estimates at the indicated values of ϕ and ψ. The slices were chosen to pass through (or close to) a mode. Comparison of the contour plots confirms that the estimates for Q_0 are much smoother than those for $\mathcal{P}_0$. The profile densities, which have been chosen to pass near to a local mode, show an "adaptive" smoothing character of the Q_0 estimate, in which the tails of the density are noticeably smoother, while the height of the modes are not much less than those for $\mathcal{P}_0$. We note, also, the possibility of spurious modes far in the tails of Q_0 which may arise for similar reasons to the sinusoidal behaviour seen in Fig. 1. In all these interpretations, it should be remembered that the comparisons are made on the basis of an LCV choice of smoothing parameter.
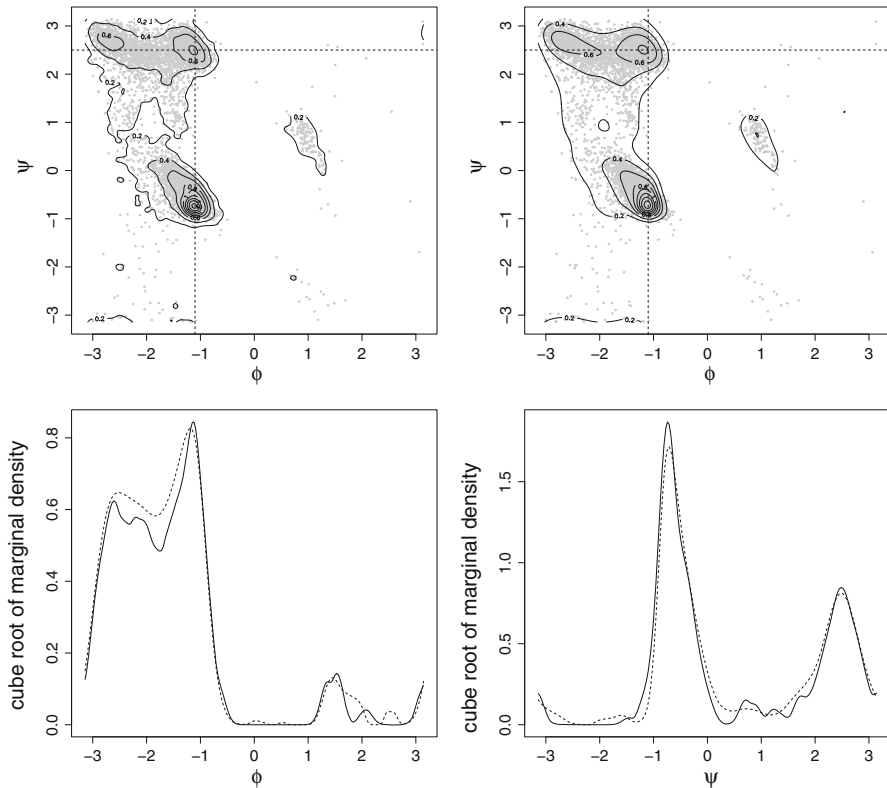
Fig. 4 Top: transformed (cube-root) contour plots of the bivariate density estimates for amino acid A (alanine) using $\mathcal{P}_0$ (left) and Q_0 (right). Bottom: profile densities for $\mathcal{P}_0$ (continuous) and Q_0 (dashed) corresponding to $\psi = 2.5$ (left) and $\phi = -1.1$ (right)

7 Discussion

Nonparametric density estimation for circular data has previously focussed on classical kernel density estimation (using a circular kernel). In this paper, we propose a family of estimators that have better theoretical properties in the sense of an arbitrarily small asymptotic bias without asymptotic variance inflation. Such methods have the potential to be more useful when the shape of the density shows unfriendly features like big curvature or large tails. A disadvantage of our proposal is that it is more restrictive than traditional circular kernel method because it requires that population densities are as smooth as required by the sin-polynomial order. Moreover, it is not guaranteed to give favourable results with small sample sizes. However, in our simulations it seems that improvements come with reasonable sample sizes like 100 and 500. A final drawback is that our estimators are more computationally expensive since they have not, in general, closed forms, although in the paper we have seen that use of the von Mises kernel gives an efficient closed form (in approximate or exact form) for $p \leq 2$. Thus, the computational effort in our simulation examples is never more than twice that of the standard ($p = 0$) case.
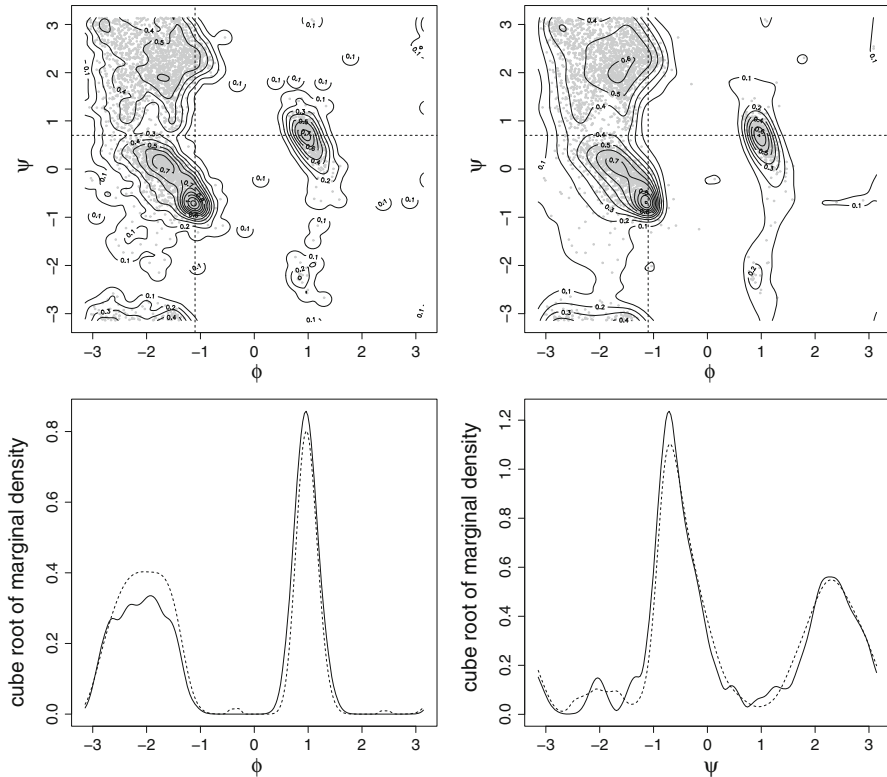
Fig. 5 Top: transformed (cube-root) contour plots of the bivariate density estimates for amino acid N (asparagine) using $\mathcal{P}_0$ (left) and Q_0 (right). Bottom: profile densities for $\mathcal{P}_0$ (continuous) and Q_0 (dashed) corresponding to $\psi = 0.7$ (left) and $\phi = -1.1$ (right)

A promising development could lie in replacing our sin-polynomial expansion by a proper, flexible circular parametric family, namely the distributions based on non-negative trigonometric sums introduced by Fernández-Durán ([2004](#)) and Fernández-Durán and Gregorio-Domínguez ([2016](#)). This would give a fully parametric method for a null concentration of the kernel, becoming more nonparametric with increasing concentration. The main difficulty would be to find the global maximum of the likelihood functions, which have *many* local maxima. Kernel weighting would still be used to make the method nonparametric. The formal asymptotic theory would need to be studied, and this task appears less straightforward.

Appendix

Proof of Theorem 1 Letting $g = \log f$, from Eq. (3) we have

$$
\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{a}}] - \tilde{\boldsymbol{a}} = \tilde{\mathbf{J}}^{-1} \Bigg(& \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{A}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\theta}_i - \boldsymbol{\theta})\right] \\
& - \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \exp\left(\tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) d\boldsymbol{\alpha} \Bigg) \\
= \tilde{\mathbf{J}}^{-1} & \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \\
& \times \left[\exp(g(\boldsymbol{\alpha})) - \exp\left(\tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right)\right] d\boldsymbol{\alpha}.
\end{aligned}
$$

Observe that

$$
\exp(g(\boldsymbol{\alpha})) - \exp\left(\tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right) = \exp(g(\boldsymbol{\alpha}))\left[1 - \exp\left(\tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta}) - g(\boldsymbol{\alpha})\right)\right]
$$
$$
\approx f(\boldsymbol{\alpha})\left[g(\boldsymbol{\alpha}) - \tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta})\right]. \tag{15}
$$

Hence, when p is odd, using

$$
g(\boldsymbol{\alpha}) - \tilde{\mathcal{P}}_p(\boldsymbol{\alpha} - \boldsymbol{\theta}) = (S'_{\boldsymbol{\alpha}-\boldsymbol{\theta}})^{\otimes(p+1)} \frac{\tilde{\boldsymbol{a}}_{p+1}}{(p+1)!} + o\left(\sin^{p+1}\left(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}\right)\right), \tag{16}
$$

and $f(\boldsymbol{\alpha}) = f(\boldsymbol{\theta}) + O\left(\sin\left(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}\right)\right)$ in (15), due to assumption (a), we get

$$
\mathrm{E}[\hat{\boldsymbol{a}}] - \tilde{\boldsymbol{a}} \approx \tilde{\mathbf{J}}^{-1} \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha} - \boldsymbol{\theta}) f(\boldsymbol{\theta}) \frac{(S'_{\boldsymbol{\alpha}-\boldsymbol{\theta}})^{\otimes p+1} \tilde{\boldsymbol{a}}_{p+1}}{(p+1)!} d\boldsymbol{\alpha}
$$

Then the bias result follows by approximating (15) along with $f(\boldsymbol{\alpha}) = f(\boldsymbol{\theta}) + S'_{\boldsymbol{\alpha}-\boldsymbol{\theta}} D_f(\boldsymbol{\theta}) + o\left(\sin\left(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}\right)\right)$.

For the variance, again from Eq. (3) we see that $\mathrm{Var}[\hat{\boldsymbol{a}}] \approx \tilde{\mathbf{J}}^{-1}\mathbf{V}\tilde{\mathbf{J}}^{-1}$, where $\mathbf{V}$ stands for the covariance matrix of the LHS of system (1). We have

$$
\begin{aligned}
\mathbf{V} = \frac{1}{n} & \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta})\mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta})' K^2_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha}) f(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \\
& - \frac{1}{n} \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) K_{\kappa_1,\ldots,\kappa_d}(\boldsymbol{\alpha}) f(\boldsymbol{\alpha}) d\boldsymbol{\alpha}
\end{aligned}
$$

$$\times \int_{[-\pi,\pi)^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta})' K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha}) f(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}.$$

The second term on the RHS is $O\left(n^{-1}\mathbf{1}\mathbf{1}'\left\{\int K_{\kappa_1,\dots,\kappa_d}(\boldsymbol{\alpha}) \sin^{p+1}\left(\boldsymbol{\alpha}^{(1)}\right) \mathrm{d}\boldsymbol{\alpha}\right\}^2\right)$, and so it is negligible under hypothesis (a). After a change of variable, the first-order expansion for $f(\boldsymbol{\alpha})$ around $\boldsymbol{\theta}$ leads to the variance under assumption (b).

References

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T (2008) A generative, probabilistic model of local protein structure. PNAS 105:8932–8937

Delicado P (2006) Local likelihood density estimation based on smooth truncation. Biometrika 93:472–480

Di Marzio M, Panzera A, Taylor CC (2009) Local polynomial regression for circular predictors. Stat Probab Lett 79:2066–2075

Di Marzio M, Panzera A, Taylor CC (2011) Kernel density estimation on the torus. J Stat Plan Inference 141:2156–2173

Di Marzio M, Panzera A, Taylor CC (2013) Non-parametric regression for circular responses. Scand J Stat 40:238–255

Di Marzio M, Fensore S, Panzera A, Taylor CC (2016) Practical performance of local likelihood for circular density estimation. J Stat Comput Simul 86:2560–2572

Eguchi S, Copas JB (1998) A class of local likelihood methods and near-parametric asymptotics. J R Stat Soc B 60:709–724

Fernández-Durán JJ (2004) Circular distributions based on nonnegative trigonometric sums. Biometrics 60:499–503

Fernández-Durán JJ, Gregorio-Domínguez MM (2016) CircNNTSR: an R package for the statistical analysis of circular, multivariate circular, and spherical data using nonnegative trigonometric sums. J Stat Softw 70:1–19

Fisher NI (1993) Statistical analysis of circular data. Cambridge University Press, Cambridge

Gill J, Hangartner D (2010) Circular data in political science and how to handle it. Polit Anal 18:316–336

Glad I, Hjort NL, Ushakov NG (2003) Correction of density estimators that are not densities. Scand J Stat 30:415–427

Hall P, Watson GS, Cabrera J (1987) Kernel density estimation with spherical data. Biometrika 74:751–762

Hjort NL, Jones MC (1996) Locally parametric nonparametric density estimation. Ann Stat 24:1619–1647

Lennox KP, Dahl DB, Vannucci M, Day R, Tsai JW (2010) A Dirichlet process mixture of hidden Markov models for protein structure prediction. Ann Appl Stat 4:916–942

Loader CR (1996a) Local likelihood density estimation. Ann Stat 24:1602–1618

Loader CR (1996b) Local regression and likelihood. Springer, London

Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins Struct Funct Bioinf 50:437–450

Mardia KV, Taylor CC, Subramaniam GK (2007) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. Biometrics 63:505–512

Singh H, Hnizdo V, Demchuk E (2002) Probabilistic model for two dependent circular variables. Biometrika 89:719–723

Taylor CC, Mardia KV, Di Marzio M, Panzera A (2012) Validating protein structure using kernel density estimates. J Appl Stat 39:2379–2388

Tibshirani R, Hastie T (1987) Local likelihood estimation. J Am Stat Assoc 82:559–567