A general framework for circular local likelihood regression

M. Alonso-Pena, I. Gijbels and R.M. Crujeiras

A general framework for circular local likelihood regression

María Alonso-Pena[1,2]*, Irène Gijbels[3,4] and Rosa M. Crujeiras[2]

[1]ORSTAT, KU Leuven

[2]CITMAga, Universidade de Santiago de Compostela

[3]Department of Mathematics, KU Leuven

[4]Leuven Statistics Research Center (LStat), KU Leuven

Abstract

This paper presents a general framework for the estimation of regression models with circular covariates, where the conditional distribution of the response given the covariate can be specified through a parametric model. The estimation of a conditional characteristic is carried out nonparametrically, by maximizing the circular local likelihood, and the estimator is shown to be asymptotically normal. The problem of selecting the smoothing parameter is also addressed, as well as bias and variance computation. The performance of the estimation method in practice is studied through an extensive simulation study, where we cover the cases of Gaussian, Bernoulli, Poisson and Gamma distributed responses. The generality of our approach is illustrated with several real-data examples from different fields.

Keywords: Circular data, Data-driven smoothing selection, Local likelihood, Nonparametric regression

*M. Alonso-Pena and R.M. Crujeiras acknowledge the support from project PID2020-116587GB-I00, funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia. M. Alonso-Pena and I. Gijbels gratefully acknowledge support from project C16/20/002 of the Research Fund KU Leuven, Belgium. This work was completed while the first author was visiting the Department of Mathematics, KU Leuven, supported by the Xunta de Galicia through the grant ED481A-2019/139 from the Consellería de Educación, Universidade e Formación Profesional. The authors also acknowledge the Supercomputing Center of Galicia (CESGA) for the computational resources.

1 Introduction

Classical statistical techniques are usually devised for modeling data taking values in euclidean spaces. However, with modern measurement tools it is possible to obtain data that, for a complete analysis, require embedding in other spaces beyond the euclidean context (Patrangenaru and Ellingson, 2016). This is the case of circular data, which have received marked attention in recent years (Jammalamadaka and SenGupta, 2001; Pewsey et al., 2013). See, for the more general case of hyperspherical or directional data, Mardia and Jupp (2000) and Ley and Verdebout (2017).

An interesting problem involving circular data is to estimate a regression function when the covariate is of a circular nature. Several parametric models for this setting are described in Jammalamadaka and SenGupta (2001, Ch. 8). However, these parametric models are often either not flexible enough, or include a large number of parameters to estimate. In order to overcome these problems, Di Marzio et al. (2009) proposed a kernel-type estimator of the regression function based on a local sine-polynomial, and its performance in practice was studied by Oliveira et al. (2013). Generalizations for a hyperspherical covariate were proposed by Di Marzio et al. (2014) and García-Portugués et al. (2016). Regarding other regression scenarios involving circular predictors, Di Marzio et al. (2018) proposed a kernel-type logistic regression, focusing on classification purposes.

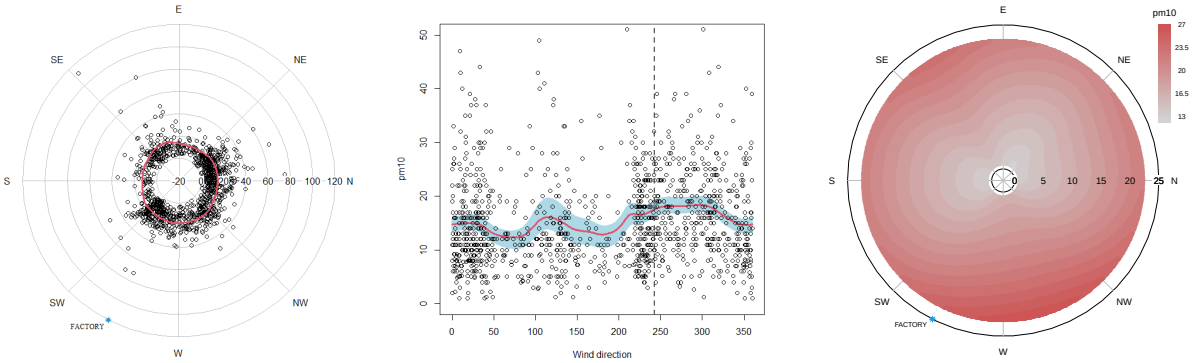


Figure 1: Polar representation of pm10 concentration against wind direction, with estimated regression curve (left). Planar zoomed representation of pm10 against wind direction, with estimated regression curve, 95% point-wise confidence band (center). Estimation of the mean pm10 concentration as a function of the wind direction and wind speed (right). The star (left and right) or vertical dashed line (center) indicate the direction of the factory.

We present a broad methodology to nonparametrically estimate conditional characteristics involving a circular covariate and a general response variable, by maximizing the local log-likelihood weighted by a circular kernel. The idea of maximizing the local kernel

weighted log-likelihood for the estimation of regression curves was studied by Fan et al. (1998) for real-valued variables. Our approach allows to estimate curves representing a conditional characteristic of a general response (which can be discrete or continuous) given a circular covariate. For this, the conditional distribution has to be specified and, then, any conditional characteristic of interest can be estimated via maximum local likelihood, taking into account the periodic behaviour of the covariate. This method englobes, as particular cases, the proposal of Di Marzio et al. (2009) when using a normal likelihood and the kernel logistic method of Di Marzio et al. (2018) if the conditional density is set to a Bernoulli distribution. Additionally, many other types of regression can be performed, such as nonparametric Poisson, binomial or gamma regression.

The asymptotic properties of the circular kernel log-likelihood estimator are explored in this paper, and accurate approximations of the bias and variance of the estimator are derived. These allow the construction of inferential tools, such as confidence intervals. In addition, an automatic criterion for selecting the smoothing parameter is proposed. All the results derived in this manuscript are general in the way that they are valid for a large class of regression settings and for the estimation of general conditional characteristics. In addition, although for Gaussian and Bernoulli particular cases the estimator coincides with estimators already proposed in the literature, the present work sheds more light on these topics, providing asymptotic normality results, approximations for bias and variance and a reliable criterion for selecting the smoothing parameter.

As an example of the broad applicability of the present methodology, the consideration of a gamma conditional distribution allows us to investigate the relationship between pm10 particle concentration and wind direction in the city of Pontevedra, Spain. The dataset is represented in the left panel of Figure 1 and more details about it can be found in Section 6. In this example, it is of special interest to ascertain if the concentration is higher for wind directions around 250 degrees, direction in which there is a possibly contaminating factory. The generality of the proposed methodology can be extended to more complex scenarios, such as partially linear models involving both circular and real-valued covariates. This then allows to broaden our study of the pm10 concentration by including the wind speed as a covariate, as shown in the right panel of Figure 1.

The organization of the manuscript is as follows: Section 2 presents the general local maximum likelihood estimation procedure for circular covariates, presenting some important particular cases and exploring its asymptotic properties. Section 3 shows how to compute both the bias and variance of the estimators. The selection of the smoothing parameter is discussed in Section 4, while the empirical performance of the estimators is studied via simulations in Section 5 for several models, including continuous and discrete responses. Applications to real datasets are shown in Section 6 and extensions of the method

to include more covariates with different nature and to data defined on the hypersphere are discussed in Section 7. Finally, a discussion is provided in Section 8.

2 Local likelihood estimation for circular regression

Let Θ be a continuous circular variable defined on $\mathbb{T} = [0, 2\pi)$ and Y a random variable which can be either a discrete or a real-valued continuous variable. Given a random bivariate sample $\{(\Theta_i, Y_i)\}_{i=1}^n$, we are interested in estimating a generic unknown function g, which may represent, for example, the conditional mean regression function of Y given $\Theta = \theta_0$ or a transformed conditional mean function. For a given $\theta_0 \in \mathbb{T}$, we approximate the function of interest, g, by employing the Taylor-like expansion introduced by Di Marzio et al. (2009) when dealing with kernel regression involving circular predictors. For data points Θ_i in a neighborhood of θ_0, and assuming that the target function g is at least p times continuously differentiable, we have

$$g(\Theta_i) \approx g(\theta_0) + g'(\theta_0)\sin(\Theta_i - \theta_0) + ... + \frac{g^{(p)}(\theta_0)}{p!}\sin^p(\Theta_i - \theta_0), \tag{1}$$

where $g^{(p)}$ denotes the pth derivative of g. This approximation can be expressed as

$$g(\Theta_i) \approx \mathbf{\Theta}_i^\top \boldsymbol{\beta}, \tag{2}$$

where $\mathbf{\Theta}_i = (1, \sin(\Theta_i - \theta_0), \ldots, \sin^p(\Theta_i - \theta_0))^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, with $\beta_\nu = g^{(\nu)}(\theta)/\nu!$, for $\nu = 0, 1, ..., p$. Now, for each observation (Θ_i, Y_i), let $l(g(\Theta_i), Y_i)$ be the log-likelihood function evaluated at $(g(\Theta_i), Y_i)$. If Θ_i belongs to a neighbourhood of θ_0, the approximation in (2) yields that the contribution of (Θ_i, Y_i) to the log-likelihood is $l(\mathbf{\Theta}_i^\top \boldsymbol{\beta}, Y_i)$, weighted by $K_\kappa(\Theta_i - \theta_0)$, where K_κ is a circular kernel function with concentration parameter κ (which acts as the smoothing parameter), tending to infinity as $n \to \infty$. A widely used circular kernel is the von Mises density, $K_\kappa(\theta) = \exp\{\kappa\cos\theta\}/(2\pi I_0(\kappa))$, with $I_0(\kappa)$ the modified Bessel function of the first kind and order zero. Consequently, we can define the local circular kernel weighted log-likelihood as

$$\mathcal{L}_p(\boldsymbol{\beta}; \kappa, \theta_0) = \sum_{i=1}^n l(\mathbf{\Theta}_i^\top \boldsymbol{\beta}, Y_i) K_\kappa(\Theta_i - \theta_0), \tag{3}$$

where the subscript p denotes the degree of the trigonometric polynomial used for the approximation in (1). By maximizing the local log-likelihood in (3) with respect to $\boldsymbol{\beta}$ we obtain the estimations of the local parameters, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_p)^\top$. Then, the estimators of the target function g and its derivatives, at the point θ_0, are given by $\hat{g}^{(\nu)}(\theta_0) = \nu!\hat{\beta}_\nu$, for $\nu = 0, ..., p$, where ν represents the order of the derivative. In practice, an adequate choice

of the order of the sine-polynomial to estimate g is $p = 1$, leading to a local-linear type estimator. Note that the maximization of (3) may not have an explicit solution in some cases, in which numerical methods must be employed in order to obtain the estimators.

The methodology proposed in this section is a general approach which allows to obtain local sine-polynomial estimators for a broad class of regression contexts involving a circular covariate. Apart from including the two particular cases already studied in the literature (normal and Bernoulli), it allows to estimate the transformed regression function in a large class of settings, for example, when having a Poisson or gamma likelihood. In Section 2.1, we will shortly describe two particular cases: the normal and the Poisson distributions. Details on the particular case of the Bernoulli distribution, which was already studied in the context of classification by Di Marzio et al. (2018), are given in the Supplementary Material. In addition, in Section 2.2, the asymptotic properties of the estimator in case the conditional distribution is a member of the exponential family are derived.

2.1 Particular cases: normal & Poisson distributions

As a particular case, consider the scenario where g is the regression function in the model

$$Y = g(\Theta) + \sigma(\Theta)\varepsilon, \quad \text{where } \mathbb{E}(\varepsilon|\Theta = \theta_0) = 0,\ \mathbb{E}(\varepsilon^2|\Theta = \theta_0) = 1, \tag{4}$$

which implies $\mathrm{Var}(\varepsilon|\Theta = \theta_0) = 1$. If the errors are normally distributed, we have that $[Y|\Theta = \theta_0] \sim N(g(\theta_0), \sigma^2(\theta_0))$. Consequently, the local circular kernel weighted log-likelihood for a fixed $\theta_0 \in \mathbb{T}$, $\mathcal{L}_p(\boldsymbol{\beta}; \kappa, \theta_0)$, is given by

$$-\log(\sigma(\theta)\sqrt{2\pi}) \sum_{i=1}^{n} K_\kappa(\Theta_i - \theta_0) - \frac{1}{2\sigma^2(\theta)} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j \sin^j(\Theta_i - \theta_0) \right)^2 K_\kappa(\Theta_i - \theta_0).$$

Maximizing the previous expression with respect to $\boldsymbol{\beta}$ is equivalent to minimizing

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j \sin^j(\Theta_i - \theta_0) \right)^2 K_\kappa(\Theta_i - \theta_0),$$

which corresponds to the local-polynomial least-squares problem studied by Di Marzio et al. (2009) and by Oliveira et al. (2013). Note that, in this case, the only proposal available in practice for the selection of the smoothing parameter is a cross-validation criterion.

Another interesting case arises when Y is a count variable, following a Poisson distribution where the mean parameter depends on the value of Θ. We will consider g as the logarithm of the mean function, $g(\theta_0) = \log[\mathbb{E}(Y|\Theta = \theta_0)]$. Therefore, we have $\mathbb{E}(Y|\Theta = \theta_0) = \exp\{g(\theta_0)\}$. The local log-likelihood is then

$$\mathcal{L}_p(\boldsymbol{\beta}; \kappa, \theta_0) = \sum_{i=1}^{n} \left(Y_i \boldsymbol{\Theta}_i^\top \boldsymbol{\beta} - \exp\left\{ \boldsymbol{\Theta}_i^\top \boldsymbol{\beta} \right\} - \log(Y_i!) \right) K_\kappa(\Theta_i - \theta_0).$$

Since the last term does not depend on $\boldsymbol{\beta}$, the maximization of the previous expression is equivalent to the maximization of

$$\sum_{i=1}^{n}\left(Y_i\sum_{j=0}^{p}\beta_j\sin^j(\Theta_i-\theta_0)-\exp\left\{\sum_{j=0}^{p}\beta_j\sin^j(\Theta_i-\theta_0)\right\}\right)K_\kappa(\Theta_i-\theta_0).$$

2.2 Asymptotic properties in the exponential family case

In the particular cases described above, the conditional densities belong to the exponential family, which is definitely a very important setting. Thus, in this section we derive some asymptotic properties of the circular local likelihood estimator when the conditional distribution is part of the exponential family. The generalization of these results to a broader setting is discussed briefly at the end of the section.

We assume that the conditional distribution belongs to the one-parameter exponential family and that the function of interest $g(\theta)$ is the natural parameter. Then, the contribution to the local likelihood of each observation is given by

$$l[g(\Theta_i),Y_i]=\psi^{-1}\{Y_ig(\Theta_i)-b[g(\Theta_i)]\}+c(Y_i,\psi), \tag{5}$$

where b and c are known functions and ψ is assumed to be a known parameter. We will also denote $l^{(q)}(a,b)=\frac{\partial^q}{\partial a^q}l(a,b)$ and $\rho(\theta)=l^{(2)}[g(\theta),\mu(\theta)]$ where, because of the first Barlett identity, $\mu(\theta)=\mathbb{E}[Y|\Theta=\theta]=b'[g(\theta)]$. In addition, we have $\mathrm{Var}[Y|\Theta=\theta]=\psi b''[g(\theta)]$. The marginal density of Θ will be denoted by f.

In order to study the properties of the estimator, we restrict to the class of kernels

$$K_\kappa(\theta)=c_\kappa(K)K[\kappa(1-\cos\theta)],\quad \text{where } K:[0,\infty)\to[0,\infty) \text{ with} \tag{6}$$

$$\int_0^\infty r^{\frac{j-1}{2}}K^l(r)dr<\infty\quad j\in\mathbb{N}\text{ and }l=1,2,4. \tag{7}$$

The factor $c_\kappa(K)$ is a normalization constant given by

$$c_\kappa(K)^{-1}=\int_0^{2\pi}K[\kappa(1-\cos\theta)]d\theta=\kappa^{-1/2}\lambda_\kappa(K), \tag{8}$$

where $\lambda_\kappa(K)=2\int_0^{2\kappa}r^{-\frac{1}{2}}\left(2-\frac{r}{\kappa}\right)^{-\frac{1}{2}}K(r)dr$. Recall that κ is a sequence depending on n and $\kappa\to\infty$ as $n\to\infty$. Thus, for a large κ we have $c_\kappa(K)^{-1}\sim\kappa^{-1/2}\lambda(K)$, with $\lambda(K)=2^{\frac{1}{2}}\int_0^\infty r^{-\frac{1}{2}}K(r)dr$. Condition (6) is usually assumed in the hyperspherical setting (Hall et al., 1987; Bai et al., 1988; García-Portugués et al., 2013). The von Mises kernel is an example of a kernel satisfying (6). In this case, the normalization constant is given by $c_\kappa(K)=\exp\{\kappa\}/(2\pi I_0(\kappa))$ and $K(r)=\exp\{-r\}$.

Furthermore, let $\boldsymbol{A}$ be a $(p+1)\times(p+1)$ matrix with (i,j)th element given by

$$(\boldsymbol{A})_{ij} = \frac{2^{\frac{i+j-1}{2}}\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}b^*_{i+j-2}(K), \quad b^*_j(K) = \begin{cases} 0 & \text{if } j \text{ is odd,} \\ \int_0^\infty r^{\frac{j-1}{2}}K(r)dr & \text{if } j \text{ is even;} \end{cases}$$

$\boldsymbol{C}$ a $(p+1)\times(p+1)$ matrix with (i,j)th element given by

$$(\boldsymbol{C})_{ij} = \frac{2^{\frac{i+j-1}{2}}b''[g(\theta_0)]f(\theta_0)}{(i-1)!(j-1)!\psi}d^*_{i+j-2}(K), \quad d^*_j(K) = \begin{cases} 0 & \text{if } j \text{ is odd,} \\ \int_0^\infty r^{\frac{j-1}{2}}K^2(r)dr & \text{if } j \text{ is even;} \end{cases}$$

and $\boldsymbol{q}$ a vector of length $(p+1)$ with jth element given by

$$2^{\frac{p+j+1}{2}}n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{-\frac{p}{2}}}{(j-1)!}\rho(\theta_0)f(\theta_0)\frac{g^{(p+1)}(\theta_0)}{(p+1)!}b^*_{p+j}(K).$$

The normalized estimator, given by

$$\widehat{\boldsymbol{\beta}}_N = n^{\frac{1}{2}}\kappa^{-\frac{1}{4}}\left(\hat{\beta}_0 - g(\theta_0), \kappa^{-\frac{1}{2}}[\hat{\beta}_1 - g'(\theta_0)], \ldots, \kappa^{-\frac{p}{2}}[p!\hat{\beta}_p - g^{(p)}(\theta_0)]\right)^\top$$

will be considered. Theorem 1 establishes the asymptotic normality of the estimator.

Theorem 1. *Assume that the function $l[g(\Theta_i), Y_i]$ is given by (5) and that $\kappa \to \infty$, $n\kappa^{-\frac{1}{2}} \to \infty$ as $n \to \infty$. In addition, assume that the following conditions hold:*

C1. $f(\theta_0) > 0$ and the function f is uniformly bounded,

C2. the functions $g^{(p+2)}$, f' and $b^{(3)}$ exist and are continuous,

C3. the matrix $\boldsymbol{A}$ is invertible,

C4. the function K in (6) has exponential decay: $K(r) \leq Be^{-\alpha r}$, with $B, \alpha > 0$.

Then, as $n \to \infty$,

$$\left(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{A}^{-1}\boldsymbol{q}[1+o(1)]\right) \xrightarrow{D} N(0, \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}).$$

The proof of Theorem 1 is given in Section S2.1 of the Supplementary Material.

Remark 1. The asymptotic normality of the estimator can be generalized for conditional densities that are not members of the exponential family by considering suitable regularity assumptions on the log-likelihood function. In the manuscript, however, the result is provided explicitly for the exponential family case given that, on the one hand, this is a very important family regarding applications and, on the other hand, the explicit proof of a more general result would involve more tedious expressions and would be more difficult to follow. Indications of which changes should be made in order to have a more general result are given in Section S2.1 of the Supplementary Material.

Theorem 1 gives expressions of the bias and variance of the estimator in the conditional exponential family case. Note that these expressions, however, depend on unknown quantities. Thus, the next section gives finite-sample approximations of the bias and variance which can be computed in practice and do not rely on the exponential family assumption.

3 Bias and variance of the estimator

For many inferential tasks it is important to compute the bias and variance of the estimator $\hat{\boldsymbol{\beta}}$, obtained after maximizing (3). Estimating these quantities will also be useful in order to select a smoothing parameter. In this section, we follow the approach of Fan et al. (1998) and give approximations of the bias and variance of the estimators presented in Section 2.

3.1 Bias of the estimator

The bias of $\hat{\boldsymbol{\beta}}$ comes from the approximation of the target function by the sine-polynomial in (1). Consequently, the bias can be approximated by computing the difference of two maximum local likelihood fits with different accuracies. Denote the error approximation at Θ_i, resulting from (1), by $\epsilon(\Theta_i) = g(\Theta_i) - \sum_{\nu=0}^{p} \frac{g^{(\nu)}(\theta_0)}{\nu!} \sin^{\nu}(\Theta_i - \theta_0)$. Assume the existence of the $(p+a+1)$th derivative of the function g at the point θ_0 for some $a \in \mathbb{N}$. Then, the error term can be approximated by a further sine-polynomial expansion,

$$\epsilon(\Theta_i) \approx \beta_{p+1} \sin^{p+1}(\Theta_i - \theta_0) + ... + \beta_{p+a} \sin^{p+a}(\Theta_i - \theta_0) = \epsilon_i. \tag{9}$$

The choice of a will affect how well the bias is estimated, but a large value of a will lead to a higher computational time. For simplicity, in practice we restrict to the choice $a = 2$, which gives a good performance when estimating the bias with a feasible computational time. Suppose that the quantities $\epsilon_i,\ i = 1, ..., n$ are known. We could approximate the local log-likelihood in a more precise way as

$$\mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0) = \sum_{i=1}^{n} l(\boldsymbol{\Theta}_i^\top \boldsymbol{\beta} + \epsilon_i, Y_i) K_\kappa(\Theta_i - \theta_0). \tag{10}$$

Denote by $\hat{\boldsymbol{\beta}}^*$ the maximizer of $\mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0)$. The bias of $\hat{\boldsymbol{\beta}}$ can be estimated by $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*$. However, it would not be necessary to compute $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*$, as we will see below. Let $\mathcal{L}_p^{*'}(\boldsymbol{\beta}; \kappa, \theta_0) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0)$, $\quad \mathcal{L}_p^{*''}(\boldsymbol{\beta}; \kappa, \theta_0) = \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0)$ be, respectively, the gradient vector and the Hessian matrix of $\mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0)$. Since $\hat{\boldsymbol{\beta}}^*$ is the maximizer of $\mathcal{L}_p^*(\boldsymbol{\beta}; \kappa, \theta_0)$, we have $\mathcal{L}_p^{*'}(\hat{\boldsymbol{\beta}}^*; \kappa, \theta_0) = 0$ and, hence, by using a Taylor expansion, it holds that

$$0 = \mathcal{L}_p^{*'}(\hat{\boldsymbol{\beta}}^*; \kappa, \theta_0) \approx \mathcal{L}_p^{*'}(\hat{\boldsymbol{\beta}}; \kappa, \theta_0) + \mathcal{L}_p^{*''}(\hat{\boldsymbol{\beta}}; \kappa, \theta_0)(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}).$$

Consequently, we obtain the approximated bias vector of $\hat{\boldsymbol{\beta}} = (\beta_0, \ldots, \beta_p)^\top$, defined as

$$\hat{\boldsymbol{b}}_p(\theta_0; \kappa) = \left[\mathcal{L}_p^{*''}(\hat{\boldsymbol{\beta}}; \kappa, \theta_0)\right]^{-1} \mathcal{L}_p^{*'}(\hat{\boldsymbol{\beta}}; \kappa, \theta_0). \tag{11}$$

Note that the bias in (11) cannot be computed in practice, since it depends on the unknown quantities $\epsilon_1, \ldots, \epsilon_n$. In order to obtain a data-driven approximation of the bias, we proceed as follows. First, a pilot concentration parameter, namely κ^*, is selected. This

pilot concentration is used to fit a sine-polynomial of degree $(p+a)$, obtaining estimates $\hat{\boldsymbol{\beta}}^{(p+a)} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p+a})^\top$. Second, these quantities are substituted into (9) and, thus, we obtain the estimators of $\epsilon_1, \ldots, \epsilon_n$, denoted by $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$. Plugging $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ into (10) we can obtain the estimated bias vector from (11), namely $\hat{\boldsymbol{b}}_p^e(\theta_0; \kappa)$. Then, recalling that $\beta_\nu = g^{(\nu)}(\theta)/\nu!$, the estimated bias of $\hat{g}^{(\nu)}(\theta_0)$ is given by

$$\hat{B}_{p,\nu}(\theta_0; \kappa) = \nu! \boldsymbol{e}_{\nu+1}^\top \hat{\boldsymbol{b}}_p^e(\theta_0; \kappa), \tag{12}$$

where $\boldsymbol{e}_{\nu+1}$ denotes the vector with all entries equal to zero except the one in the $(\nu+1)$th position. The choice of the pilot concentration, κ^*, will be discussed in Section 4.

3.2 Variance of the estimator

Now we proceed to obtain the variance of the estimated vector of local parameters. Since our estimate $\hat{\boldsymbol{\beta}}$ is the maximizer of the circular local likelihood, we have $\mathcal{L}_p'(\hat{\boldsymbol{\beta}}; \kappa, \theta_0) = 0$, and with a Taylor expansion we obtain $0 = \mathcal{L}_p'(\hat{\boldsymbol{\beta}}; \kappa, \theta_0) \approx \mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0) + \mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Therefore, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left[\mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)\right]^{-1} \mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0)$. Now, notice that

$$\begin{aligned}\mathrm{Var}[\hat{\boldsymbol{\beta}}|\Theta_1, \ldots, \Theta_n] &= \mathrm{Var}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\Theta_1, \ldots, \Theta_n] \approx \mathrm{Var}\left[-\left[\mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)\right]^{-1} \mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0)|\Theta_1, \ldots, \Theta_n\right] \\ &\approx \left[\mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)\right]^{-1} \mathrm{Var}\left[\mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0)|\Theta_1, \ldots, \Theta_n\right] \left[\mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)\right]^{-1}.\end{aligned}$$

The matrix $\mathcal{L}_p''(\boldsymbol{\beta}; \kappa, \theta_0)$ can be estimated by $\mathcal{L}_p''(\hat{\boldsymbol{\beta}}; \kappa, \theta_0)$, but $\mathrm{Var}\left[\mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0)|\Theta_1, \ldots, \Theta_n\right]$ is unknown and it is necessary to estimate it. From the definition of the circular local log-likelihood in (3), we have $\mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0) = \sum_{i=1}^n l'(\boldsymbol{\Theta}_i^\top \boldsymbol{\beta}, Y_i) \boldsymbol{\Theta}_i K_\kappa(\Theta_i - \theta_0)$ and, consequently,

$$\mathrm{Var}\left[\mathcal{L}_p'(\boldsymbol{\beta}; \kappa, \theta_0)|\Theta_1, \ldots, \Theta_n\right] = \sum_{i=1}^n \mathrm{Var}\left[l'(\boldsymbol{\Theta}_i^\top \boldsymbol{\beta}, Y_i)|\Theta_1, \ldots, \Theta_n\right] \boldsymbol{\Theta}_i \boldsymbol{\Theta}_i^\top K_\kappa^2(\Theta_i - \theta_0). \tag{13}$$

Now, because of (2), the expression in (13) can be approximated by

$$\mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right] \sum_{i=1}^n \boldsymbol{\Theta}_i \boldsymbol{\Theta}_i^\top K_\kappa^2(\Theta_i - \theta_0) = \mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right] \boldsymbol{\Gamma}_n,$$

where $\boldsymbol{\Gamma}_n$ is a $(p+1) \times (p+1)$ matrix with (i,j)th element given by $\gamma_{n,i+j-2}$ and $\gamma_{n,j} = \sum_{i=1}^n \sin^j(\Theta_i - \theta_0) K_\kappa^2(\Theta_i - \theta_0)$. Then, we have

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}|\Theta_1, \ldots, \Theta_n] \approx \mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right] \left[\mathcal{L}_p''(\hat{\boldsymbol{\beta}}; \kappa, \theta_0)\right]^{-1} \boldsymbol{\Gamma}_n \left[\mathcal{L}_p''(\hat{\boldsymbol{\beta}}; \kappa, \theta_0)\right]^{-1} = \Xi_p(\theta_0; \kappa).$$

For the estimation of $\mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right]$, we distinguish two cases:

A. $\mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right] = v[g(\theta_0)]$ for some known function v, as it happens for the Bernoulli or Poisson likelihoods. In this case we estimate it as $v[\hat{g}(\theta_0)]$.

B. When the form in A is not available, we use a pilot estimator $\hat{\boldsymbol{\beta}}^{(p+a)}$ obtained by fitting a local polynomial of degree $p+a$, with a pilot concentration κ^*, as in the bias calculations. Then, $\mathrm{Var}\left[l'(g(\theta_0), Y)|\Theta = \theta_0\right]$ is estimated by

$$\frac{\sum_{i=1}^{n}[l'(\tilde{\boldsymbol{\Theta}}_i^\top \hat{\boldsymbol{\beta}}^{(p+a)}, Y_i)]^2 K_{\kappa^*}(\Theta_i - \theta_0)}{\sum_{i=1}^{n} K_{\kappa^*}(\Theta_i - \theta_0)},$$

with $\tilde{\boldsymbol{\Theta}}_i^\top = (1, \sin(\Theta_i - \theta_0), \ldots, \sin^{p+a}(\Theta_i - \theta_0))$.

We will use $\hat{V}_{p,\nu}(\theta_0; \kappa)$ to denote the variance of $\hat{g}^{(\nu)}(\theta_0)$ constructed with a sine-polynomial of degree p, *i.e.*,

$$\hat{V}_{p,\nu}(\theta_0; \kappa) = \nu!^2 \boldsymbol{e}_{\nu+1}^\top \Xi_p(\theta_0; \kappa) \boldsymbol{e}_{\nu+1}. \tag{14}$$

4 Selection of the smoothing parameter

As in all kernel methods, the selection of the smoothing parameter is of great importance, since it substantially affects the performance of the estimator. However, when employing circular kernels, the role of the smoothing (concentration) parameter is opposite to the role of the bandwidth when using *linear* kernels. When the concentration κ is very small, the estimation procedure leads to a global fit of a sine-polynomial of degree p, whereas if κ is very large, the estimation results in the interpolation of the data. Thus, it is necessary to select a smoothing parameter which correctly balances the bias and variance of the estimator, and therefore minimizes the MSE.

In the particular case of the normal likelihood (least-squares regression), Di Marzio et al. (2009) derived an expression for the optimal smoothing parameter minimizing the asymptotic MSE of the estimator when $p = 1$ and $\nu = 0$, and specifying the von Mises density as the kernel. In the hyperspherical setting, where it is assumed that the predictor lies on a hypersphere of arbitrary dimension, García-Portugués (2014) derived an optimal expression for the concentration minimizing the MSE which, in the particular case of the circumference and a von Mises kernel, is equivalent to the optimal parameter obtained by Di Marzio et al. (2009). Note that, however, in order to select a smoothing parameter in practice, the only proposals available in the literature are a rule of thumb based on a preliminary parametric estimator (García-Portugués, 2014) and a cross-validation method implemented by Oliveira et al. (2013).

In this paper, a new selection rule for the concentration parameter, not only for the least-squares setting but also for the general likelihood scenario, is proposed. In order to automatically select a smoothing parameter for the estimation of $g^{(\nu)}(\theta_0)$, we can minimize an estimation of the integrated version of the Mean Squared Error (MSE):

$$\hat{\kappa}_{p,\nu} = \operatorname*{arg\,min}_{\kappa>0} \int_0^{2\pi} \widehat{\mathrm{MSE}}_{p,\nu}(\alpha; \kappa) d\alpha, \tag{15}$$

where $\widehat{\text{MSE}}_{p,\nu}(\theta_0;\kappa)$ denotes an estimator of the MSE of $g^{(\nu)}(\theta_0)$ constructed with the concentration parameter κ. This quantity can be obtained by approximating the bias and variance of the estimator as described in Sections 3.1 and 3.2, respectively, obtaining the bias and variance estimates given by (12) and (14). Then, the estimated MSE is given by $\widehat{\text{MSE}}_{p,\nu}(\theta_0;\kappa) = \hat{B}^2_{p,\nu}(\theta_0;\kappa) + \hat{V}_{p,\nu}(\theta_0;\kappa)$. Note that in order to estimate the MSE it is necessary to first select a pilot smoothing parameter, κ^*, and fit locally a sine-polynomial of degree $p+a$. Thus, we will refer to this smoothing parameter selection method as the refined rule, since first we have to select a preliminary smoothing parameter.

In what follows, we will discuss the selection of the pilot concentration parameter. Although the role of the smoothing parameter is reversed when employing circular kernels, a similar approach to that of Fan et al. (1998) could be used to select the pilot concentration. The main idea is to come up with a criterion which, when minimized, leads to an approximated optimal smoothing parameter. In Section 4.1 we study the problem of obtaining a pilot concentration in the least-squares case. The general case is discussed in Section 4.2.

4.1 Selection of the pilot concentration: least-squares case

In this section, we consider the least-squares scenario exposed in Section 2. Let the relationship between the variables Θ and Y be modeled as in (4). As shown in Section 2, the function g and its derivatives can be estimated by minimizing the least-squares function weighted by a circular kernel, which is equivalent to maximizing the local circular kernel weighted log-likelihood function when assuming a Normal likelihood. In this case, the estimator can be explicitly expressed as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\Theta}^\top \boldsymbol{W} \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^\top \boldsymbol{W} \boldsymbol{Y}, \tag{16}$$

where $\boldsymbol{Y}$ is the vector of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{W} = \{\text{diag}(K_\kappa(\Theta_i - \theta_0))\}_{i=1,\ldots,n}$ and $\boldsymbol{\Theta}$ is a $n \times (p+1)$ matrix with (i,j)th element given by $\sin^j(\Theta_i - \theta_0)$. Following Fan and Gijbels (1995), the Circular Residual Squares Criterion (CRSC) is defined as

$$\text{CRSC}(\theta_0;\kappa) = \hat{\sigma}^2(\theta_0)\left(1 + \frac{p+1}{N}\right),$$

where

$$\hat{\sigma}^2(\theta_0) = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 K_\kappa(\Theta_i - \theta_0)}{\text{tr}(\boldsymbol{W}) - \text{tr}((\boldsymbol{\Theta}^\top \boldsymbol{W} \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^\top \boldsymbol{W}^2 \boldsymbol{\Theta})}, \tag{17}$$

and N^{-1} being the first diagonal element of the matrix $(\boldsymbol{\Theta}^\top \boldsymbol{W} \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^\top \boldsymbol{W}^2 \boldsymbol{\Theta} (\boldsymbol{\Theta}^\top \boldsymbol{W} \boldsymbol{\Theta})^{-1} = \boldsymbol{S}_n^{-1} \boldsymbol{\Gamma}_n \boldsymbol{S}_n^{-1}$. Note that, here, the notation $\boldsymbol{W}\boldsymbol{W} = \boldsymbol{W}^2$ is used. The quantities $\hat{Y}_i$, with $i = 1, \ldots, n$, denote the fitted values obtained after fitting a pth order sine-polynomial locally. When κ is very small, the bias of the estimator will be large and so will be $\hat{\sigma}^2(\theta_0)$,

obtaining a large CRSC$(\theta_0;\kappa)$. On the contrary, if κ is too small, the variance will be large and, hence, N^{-1} will also be large, resulting in a large CRSC$(\theta_0;\kappa)$.

Proposition 1 gives the conditional expectation of the CRSC quantity. The following notation will be used. Let $\widetilde{b}_j^*(K) = b_j^*(K)/\int_0^\infty r^{-\frac{1}{2}}K(r)dr$ and $\widetilde{d}_j^*(K) = d_j^*(K)/\left(\int_0^\infty r^{-\frac{1}{2}}K(r)dr\right)^2$. Further, let $\boldsymbol{B}$ and $\boldsymbol{D}$ be the $(p+1)\times(p+1)$ matrices having, respectively, the (i,j)th element given by $\widetilde{b}_{i+j-2}^*(K)$ and $\widetilde{d}_{i+j-2}^*(K)$. By $\boldsymbol{c}_p$ we denote the vector $\left(\widetilde{b}_{p+1}^*(K),\ldots,\widetilde{b}_{2p+1}^*(K)\right)^\top$.

Proposition 1. *Assume that the circular kernel satisfies (6) and (7) and that $f>0$, the density function of Θ, is continuously differentiable. Additionally, assume that $Var[Y|\Theta=\theta]=\sigma^2(\theta)>0$ exists and is continuous at $\theta=\theta_0$. If $\kappa\to\infty$ and $n\kappa^{-\frac{1}{2}}\to\infty$, then the expression of $\mathbb{E}[CRSC(\theta_0;\kappa)|\Theta_1,\ldots,\Theta_n]$ is given by*

$$\sigma^2(\theta_0)+C_p\beta_{p+1}^2 2^{p+1}\kappa^{-(p+1)}+(p+1)\frac{\sigma^2(\theta_0)a_0\kappa^{\frac{1}{2}}}{2^{\frac{1}{2}}nf(\theta_0)}+o_P\left(\frac{1}{\kappa^{p+1}}+\frac{\kappa^{\frac{1}{2}}}{n}\right),$$

where $C_p=\widetilde{b}_{2p+2}^*(K)-\boldsymbol{c}_p^\top\boldsymbol{B}\boldsymbol{c}_p$ and a_0 is the first diagonal element of the matrix $\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}$.

The proof is given in Appendix A. Given the previous result, it follows that the minimizer of $\mathbb{E}[\text{CRSC}(\theta_0;\kappa)|\Theta_1,\cdots,\Theta_n]$ with respect to κ is approximately equal to

$$\kappa_0(\theta_0)=\left(\frac{2^{\frac{2p+5}{2}}C_p\beta_{p+1}^2 nf(\theta_0)}{a_0\sigma^2(\theta_0)}\right)^{\frac{2}{2p+3}}. \tag{18}$$

It can be seen that the expression of $\kappa_0(\theta_0)$ shares some similarities with the optimal κ minimizing the asymptotic mean squared error of $\hat{g}^{(\nu)}(\theta_0)$ which, for odd $(p-\nu)$, is given by

$$\kappa_{\text{opt},p,\nu}(\theta_0)=\left(\frac{(p+1-\nu)nf(\theta_0)\beta_{p+1}^2[\boldsymbol{e}_{\nu+1}^\top\boldsymbol{B}^{-1}\boldsymbol{c}_p]^2 2^{\frac{2p+5}{2}}}{(1+2\nu)a_\nu\sigma^2(\theta_0)}\right)^{\frac{2}{2p+3}}, \tag{19}$$

where a_ν is the $(\nu+1)$th diagonal element of the matrix $\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}$. The derivation of (19) is given in Section S2.3 of the Supplementary Material.

Remark 2. Equation (19) is a generalization of the optimal parameter obtained by Di Marzio et al. (2009), who studied the case $p=1$ and $\nu=0$, with the von Mises kernel. This is easy to see by noting that, when $p=1$, $\nu=0$ and K_κ is the von Mises kernel, $a_\nu=2^{1/2}\pi^{1/2}$ and $\boldsymbol{e}_{\nu+1}^\top\boldsymbol{B}^{-1}\boldsymbol{c}_p=1/2$. In this particular case, (19) also coincides with the optimal parameter obtained by García-Portugués (2014) in the hyperspherical setting.

Comparing equations (18) and (19), it is easy to see that $\kappa_{\text{opt},p,\nu}(\theta_0)=\xi_{p,\nu}(K)\kappa_0(\theta_0)$, where $\xi_{p,\nu}(K)$ only depends on p, ν and K and is given by

$$\xi_{p,\nu}(K)=\left(\frac{(p+1-\nu)a_0[\boldsymbol{e}_{\nu+1}^\top\boldsymbol{B}^{-1}\boldsymbol{c}_p]^2}{(1+2\nu)a_\nu C_p}\right)^{\frac{2}{2p+3}}.$$

Thus, a simple approach for selecting a global pilot concentration parameter is as follows. First, we obtain the value of the concentration minimizing the integrated CRSC: $\hat{\kappa}_p = \arg\min_{\kappa>0} \int_0^{2\pi} \mathrm{CRSC}(\alpha;\kappa)d\alpha$. Afterwards, we select the pilot concentration $\hat{\kappa}_{p,\nu}^{\mathrm{CRSC}}$ as $\hat{\kappa}_{p,\nu}^{\mathrm{CRSC}} = \xi_{p,\nu}(K)\hat{\kappa}_p$.

4.2 Selection of the pilot concentration: general case

In the general circular local likelihood problem, we may distinguish two options to select the pilot concentration parameter. First, if the target function g is a transformed mean function, *i.e.*, $g(\theta_0) = T(\mu(\theta_0))$ with $\mu(\theta_0) = \mathbb{E}[Y|\Theta = \theta_0]$, we can still use the CRSC criterion, but substituting $\hat{Y}_i = T^{-1}(\boldsymbol{\Theta}_i^\top \hat{\boldsymbol{\beta}})$ in (17).

Another possibility is to use an extended version of the CRSC, namely ECRSC, regarding the local likelihood problem as an iterative local least-squares problem, as in Fan et al. (1998). In the following we give some details about this extended criterion.

Consider the Fisher scoring method for updating the vector of estimated parameters $\hat{\boldsymbol{\beta}}$ in which, for a current value $\boldsymbol{\beta}_c$, we update

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_c - \left[\mathbb{E}[\mathcal{L}_p^{''}(\boldsymbol{\beta}_c;\kappa,\theta_0)|\Theta_1,\ldots,\Theta_n]\right]^{-1} \mathcal{L}_p'(\boldsymbol{\beta}_c;\kappa,\theta_0). \tag{20}$$

Now,
$$\begin{aligned}\mathbb{E}[\mathcal{L}_p^{''}(\boldsymbol{\beta}_c;\kappa,\theta_0)|\Theta_1,\ldots,\Theta_n] &= \sum_{i=1}^n \mathbb{E}[l''(\boldsymbol{\Theta}_i^\top\boldsymbol{\beta}_c, Y_i)]\boldsymbol{\Theta}_i\boldsymbol{\Theta}_i^\top K_\kappa(\Theta_i-\theta_0)\\ &\approx \mathbb{E}[l''(g(\theta_0),Y)|\Theta=\theta_0]\sum_{i=1}^n \boldsymbol{\Theta}_i\boldsymbol{\Theta}_i^\top K_\kappa(\Theta_i-\theta_0),\end{aligned}$$

given that the expectation $\mathbb{E}[l''(g(\cdot),Y)]$ is continuous. Plugging this expression into equation (20), we have that the updating rule with the approximated expectation is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}_c - \left[\mathbb{E}[l''(g(\theta_0),Y)|\Theta=\theta_0]\sum_{i=1}^n \boldsymbol{\Theta}_i\boldsymbol{\Theta}_i^\top K_\kappa(\Theta_i-\theta_0)\right]^{-1}\sum_{i=1}^n \boldsymbol{\Theta}_i l'(\boldsymbol{\Theta}_i^\top\boldsymbol{\beta}_c,Y_i)K_\kappa(\Theta_i-\theta_0)\\ &= \left[\sum_{i=1}^n \boldsymbol{\Theta}_i\boldsymbol{\Theta}_i^\top K_\kappa(\Theta_i-\theta_0)\right]^{-1}\sum_{i=1}^n Z_i\boldsymbol{\Theta}_i K_\kappa(\Theta_i-\theta_0),\end{aligned}$$

where $Z_i = \boldsymbol{\Theta}_i^\top\boldsymbol{\beta}_c - l'(\boldsymbol{\Theta}_i^\top\boldsymbol{\beta}_c,Y_i)/\mathbb{E}[l''(g(\theta_0),Y)|\Theta=\theta_0]$ and the conditional expectation in the expression of Z_i can be computed with the value of $\boldsymbol{\beta}_c$. Thus, at the end of the iteration process the estimator $\hat{\boldsymbol{\beta}}$ is obtained by regressing Z_i over Θ_i using the local sine-polynomial of order p. The ECRSC criterion is defined then as $\mathrm{ECRSC}(\theta_0;\kappa) = \hat{\sigma}_*^2(\theta_0)\left[1+\frac{p+1}{N}\right]$, where $\hat{\sigma}_*^2(\theta_0)$ is computed as in equation (17) but using the variable Z_i. More details on the justification of this can be found in Fan et al. (1998). The concentration selector based on the ECRSC criterion can be obtained by first computing

$$\hat{\kappa}_p^* = \arg\min_{\kappa>0} \int_0^{2\pi} \mathrm{ECRSC}(\alpha;\kappa)d\alpha \tag{21}$$

and then evaluating $\hat{\kappa}_{p,\nu}^{\text{ECRSC}} = \xi_{p,\nu}(K)\hat{\kappa}_p^*$. In the case where $g(\theta) = T(\mu(\theta))$, the ECRSC criterion will be approximately the same as the CRSC criterion, but including weights $[T'(\mu(\alpha))]^{-2}$ in equation (21).

5 Simulation experiments

In this section, we study the empirical performance of the estimator presented in Section 2, as well as the behavior of the concentration selection methods introduced in Section 4. The code for all the methods can be found as Supplementary Material. We consider responses from normal, Bernoulli, Poisson and gamma models, described in Table 1 and represented in Figure 2. For each model, we simulate $B = 500$ replications of the data, and estimate the target function g with the local sine-polynomial estimator ($p = 1$, $\nu = 0$). Sample sizes are $n = 70, 100, 250, 500, 1500$ and the covariate, Θ, is drawn from a circular uniform distribution. The concentration parameter was selected by the refined rule in Section 4 (see equation (15)), where the pilot estimator was constructed with a local sine-polynomial of degree 3 and the pilot concentration parameter was selected by the CRSC criterion (in the normal case) and the ECRSC criterion (in the other cases). For comparison purposes, we also compute the estimators obtained by selecting the smoothing parameter directly with the CRSC/ECRSC criterion and with a cross-validation method. The quality of the estimators was obtained by approximating the Integrated Squared Error (ISE) as

$$\frac{\int_0^{2\pi}[\hat{g}_{(b)}(\theta) - g(\theta)]^2 d\theta}{\int_0^{2\pi} g(\theta)^2 d\theta}, \tag{22}$$

where $\hat{g}_{(b)}(\theta)$ represents the estimator of $g(\theta)$ for the bth replication of the data and the integrals are approximated numerically by Simpson's rule.

For the Bernoulli, Poisson and gamma models, the estimator involves an iterative solution, which may not even exist for very large concentrations. We avoid these situations by only considering values of the concentration parameter for which the estimators exist.

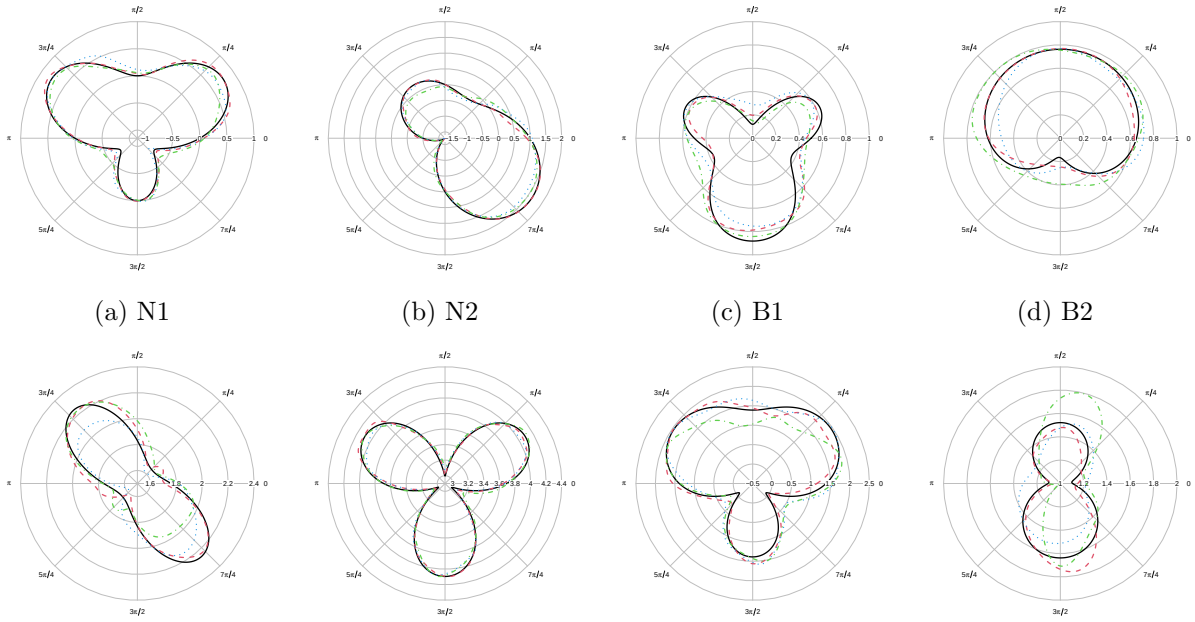
Table 2 shows the average approximated ISE for both models and all methods. It can be seen that, as expected and for all scenarios, the average approximated ISE diminishes as the sample size increases. Regarding the selection of the smoothing parameter, although the orders of magnitude are usually the same, the refined rule obtains the lowers values for all models with Normal, Poisson and gamma distributions, for all samples sizes. The only scenario where the refined rule does not completely outperform the cross-validation method is the case with a binary response. For model B1, results are slightly better when employing the cross-validation rule, except for the largest sample size ($n = 1500$), where the refined rule gives a moderately lower value of the average approximated ISE. For B2, cross-validation gives better results only for $n = 70$ and $n = 100$, but for larger sample sizes the refined rule seems the best alternative.

Table 1: Description of the simulated models.

Model	$(Y\|\Theta=\theta)$	Model elements	
Normal (N)	$\mathrm{N}\left(\mu(\theta),\sigma^2\right)$	N1:	$g(\theta)=\sin(2\theta)\cos(\theta)$;
	$g(\theta)=\mu(\theta)$	N2:	$g(\theta)=1.75\cos(\theta-\pi)\sin(\theta)+\cos(\theta)$;
			$\sigma=0.35$, $\sigma=0.5$
Bernoulli (B)	$\mathrm{Bernoulli}\left(p(\theta)\right)$	B1:	$g(\theta)=2\sin(\theta)\cos(2\theta)$
	$g(\theta)=\mathrm{logit}(p(\theta))$	B2:	$g(\theta)=\log\left(1.6+1.5\sin(\theta)+0.1\exp\{\cos(\theta)\}\right)$
Poisson (P)	$\mathrm{Poisson}\left(\mu(\theta)\right)$	P1:	$\mu(\theta)=5+\exp\left\{1.5\sin(2\theta-3)\right\}$
	$g(\theta)=\log(\mu(\theta))$	P2:	$\mu(\theta)=40+20\sin(3\theta)$
Gamma (G)	$\mathrm{Gamma}\left(\alpha,\beta(\theta)\right)$	G1:	$\mu(\theta)=4+4\sin(2\theta)\cos(\theta)$;
	$\mu(\theta)=\alpha\beta^{-2}(\theta)$	G2:	$\mu(\theta)=5+2\cos\left(3+1.5\sin(\theta)\right)$;
	$g(\theta)=\log(\mu(\theta))$		$\alpha=2$, $\alpha=1$

Table 2: Monte Carlo averages of the numerically approximated ISE obtained with the CRSC/ECRSC rule, the refined rule and cross-validation for all models.

		Model 1			Model 2		
	n	(E)CRSC	Refined	CV	(E)CRSC	Refined	CV
N	70	$1.17\cdot10^{-1}$	$9.86\cdot10^{-2}$	$1.01\cdot10^{-1}$	$5.80\cdot10^{-2}$	$4.21\cdot10^{-2}$	$4.71\cdot10^{-2}$
	100	$7.58\cdot10^{-2}$	$6.98\cdot10^{-2}$	$7.06\cdot10^{-2}$	$3.94\cdot10^{-2}$	$2.96\cdot10^{-2}$	$3.34\cdot10^{-2}$
	250	$3.18\cdot10^{-2}$	$3.13\cdot10^{-2}$	$3.16\cdot10^{-2}$	$1.44\cdot10^{-2}$	$1.34\cdot10^{-2}$	$1.44\cdot10^{-2}$
	500	$1.76\cdot10^{-2}$	$1.73\cdot10^{-2}$	$1.77\cdot10^{-2}$	$8.06\cdot10^{-3}$	$7.54\cdot10^{-3}$	$8.18\cdot10^{-3}$
	1500	$6.99\cdot10^{-3}$	$6.84\cdot10^{-3}$	$6.95\cdot10^{-3}$	$3.30\cdot10^{-3}$	$3.15\cdot10^{-3}$	$3.30\cdot10^{-3}$
B	70	$1.32\cdot10^{-1}$	$9.21\cdot10^{-2}$	$9.12\cdot10^{-2}$	$6.86\cdot10^{-2}$	$4.85\cdot10^{-2}$	$4.36\cdot10^{-2}$
	100	$8.78\cdot10^{-2}$	$7.31\cdot10^{-2}$	$6.95\cdot10^{-2}$	$4.49\cdot10^{-2}$	$3.50\cdot10^{-2}$	$3.25\cdot10^{-2}$
	250	$3.18\cdot10^{-2}$	$2.97\cdot10^{-2}$	$2.91\cdot10^{-2}$	$1.58\cdot10^{-2}$	$1.32\cdot10^{-2}$	$1.50\cdot10^{-2}$
	500	$1.77\cdot10^{-2}$	$1.69\cdot10^{-2}$	$1.66\cdot10^{-2}$	$8.08\cdot10^{-3}$	$7.32\cdot10^{-3}$	$8.28\cdot10^{-3}$
	1500	$7.06\cdot10^{-3}$	$6.62\cdot10^{-3}$	$6.68\cdot10^{-3}$	$3.09\cdot10^{-3}$	$2.92\cdot10^{-3}$	$3.19\cdot10^{-3}$
P	70	$6.00\cdot10^{-3}$	$4.63\cdot10^{-3}$	$5.69\cdot10^{-3}$	$6.84\cdot10^{-4}$	$6.75\cdot10^{-4}$	$7.15\cdot10^{-4}$
	100	$4.04\cdot10^{-3}$	$3.46\cdot10^{-3}$	$4.09\cdot10^{-3}$	$4.69\cdot10^{-4}$	$4.66\cdot10^{-4}$	$4.81\cdot10^{-4}$
	250	$1.78\cdot10^{-3}$	$1.67\cdot10^{-3}$	$1.87\cdot10^{-3}$	$2.02\cdot10^{-4}$	$2.00\cdot10^{-4}$	$2.05\cdot10^{-4}$
	500	$9.76\cdot10^{-4}$	$9.32\cdot10^{-4}$	$1.02\cdot10^{-3}$	$1.10\cdot10^{-4}$	$1.09\cdot10^{-4}$	$1.11\cdot10^{-4}$
	1500	$3.98\cdot10^{-4}$	$3.87\cdot10^{-4}$	$4.11\cdot10^{-4}$	$4.45\cdot10^{-5}$	$4.40\cdot10^{-5}$	$4.48\cdot10^{-5}$
G	70	$1.05\cdot10^{-1}$	$9.05\cdot10^{-2}$	$1.09\cdot10^{-1}$	$2.61\cdot10^{-2}$	$2.01\cdot10^{-2}$	$2.55\cdot10^{-2}$
	100	$7.35\cdot10^{-2}$	$6.13\cdot10^{-2}$	$8.19\cdot10^{-2}$	$1.99\cdot10^{-2}$	$1.50\cdot10^{-2}$	$1.79\cdot10^{-2}$
	250	$3.54\cdot10^{-2}$	$2.75\cdot10^{-2}$	$3.67\cdot10^{-2}$	$9.22\cdot10^{-3}$	$6.53\cdot10^{-3}$	$7.88\cdot10^{-3}$
	500	$2.01\cdot10^{-2}$	$1.53\cdot10^{-2}$	$1.95\cdot10^{-2}$	$5.01\cdot10^{-3}$	$3.72\cdot10^{-3}$	$4.38\cdot10^{-3}$
	1500	$8.00\cdot10^{-3}$	$6.27\cdot10^{-3}$	$7.80\cdot10^{-3}$	$1.92\cdot10^{-3}$	$1.50\cdot10^{-3}$	$1.67\cdot10^{-3}$



(a) N1 (b) N2 (c) B1 (d) B2

(e) P1 (f) P2 (g) G1 (h) G2

Figure 2: Radial representations of the true mean functions (continuous line) in all models, with representative estimates when $n = 250$ (5th ISE percentile with dashed line, 50th ISE percentile with dotted line and 95th ISE percentile with dotted-dashed line). The smoothing parameters were selected by the refined rule.

The distribution of the approximated ISE values and the smoothing parameters selected by each method are further analyzed in the Supplementary Material, where boxplots of the approximated ISE and kernel density estimates of the selected parameters are provided (for each model, sample size and concentration parameter selection method). In summary, the distribution of the concentrations selected by the refined rule is usually more concentrated, while (E)CRSC and cross-validation often select concentrations which are too large. This leads to a better performance (in general) of the refined rule in term of approximated ISE.

In order to graphically assess the quality of the estimators when selecting the concentration parameter with the refined rule, Figure 2 shows the true target functions of all models, along with three representatives of the estimators corresponding to the 5th, 50th and 95th percentiles of the approximated ISE, where the sample size was fixed to $n = 250$.

6 Real data examples

In this section we illustrate the practical use of the circular local likelihood estimator and the refined concentration selection rule with different real datasets. In all examples we estimate the target function by fitting a local sine-polynomial of degree $p = 1$, where the

concentration parameter is obtained with the refined rule selector. The bias and variance are estimated with a sine-polynomial of degree 3, and the pilot smoothing parameter is selected by the CRSC/ECRSC criterion. Employing the asymptotic normal distribution of the local likelihood estimator (see Theorem 1), we also compute point-wise confidence bands for the estimated functions, with a confidence level of 95%.

For the construction of the confidence intervals, the bias and variance approximations obtained in Section 3 could be employed, with the asymptotic normality result in Section 2.2. Then, point-wise confidence bands for the estimated target functions, with approximately $1-\alpha$ coverage, could be computed as $\hat{g}(\theta_0) - \hat{B}_{p,\nu}(\theta_0;\kappa) \pm \Phi(1-\alpha/2)\hat{V}_{p,\nu}(\theta_0;\kappa)^{1/2}$, where $\hat{B}_{p,\nu}(\theta_0;\kappa)$ and $\hat{V}_{p,\nu}(\theta_0;\kappa)$ are, respectively, the approximations of the bias and variance of the estimator, obtained in Section 3 and Φ denotes the quantile of the Gaussian distribution. However, both $\hat{B}_{p,\nu}(\theta_0;\kappa)$ and $\hat{V}_{p,\nu}(\theta_0;\kappa)$ may change abruptly, since they are constructed by using higher order derivatives, and the estimation of these may be unstable. Therefore, for the construction of confidence intervals, we employ local weighted averages of the bias and variance approximations, obtained by using kernel weights. Let

$$\hat{B}^k_{p,\nu}(\theta_0;\kappa) = \int_0^{2\pi} \hat{B}_{p,\nu}(\theta;\kappa)K_\kappa(\theta-\theta_0)d\theta \quad \text{and} \quad \hat{V}^k_{p,\nu}(\theta_0;\kappa) = \int_0^{2\pi} \hat{V}_{p,\nu}(\theta;\kappa)K_\kappa(\theta-\theta_0)d\theta$$

be the kernel weighted averages of the bias and variance approximations, respectively. The use of the weighted quantities prevent them from abrupt change along the covariate range. Then, we construct the point-wise confidence bands for the estimated functions with, approximately, $1-\alpha$ coverage probability as

$$\hat{g}(\theta_0) - \hat{B}^k_{p,\nu}(\theta_0;\kappa) \pm \Phi(1-\alpha/2)\hat{V}^k_{p,\nu}(\theta_0;\kappa)^{1/2}.$$

It is necessary to note, though, that the point-wise character of these confidence bands means that the approximately $1-\alpha$ coverage is only for a given value θ_0. In order to obtain a band for which the true target function lies inside with probability $1-\alpha$, one should consider the use of simultaneous confidence bands, although it is known that this type of bands are quite conservative. More details on this topic are discussed in Section 8.

Human motor resonance data: normal response. An example of normal data is the human motor resonance dataset obtained by Puglisi et al. (2017). In this experiment, subjects were requested to observe a movement of a rhythmic hand flexion-extension in front of them. For each angular position of the hand, the H-reflex technique was used to quantitatively measure the resonance response (see Puglisi et al., 2017, for details). Our goal is to explore the relationship between the angular position of the hand (circular predictor variable) and the H-reflex amplitude (real-valued response variable). The dataset, which is composed of $n = 70$ observations, is depicted in the left panels of Figure 3 with

the estimated regression function. Note that the H-reflex amplitude increases when the angular position ranges from $\pi/2$ to $5\pi/4$ and decreases between $5\pi/4$ and 2π.

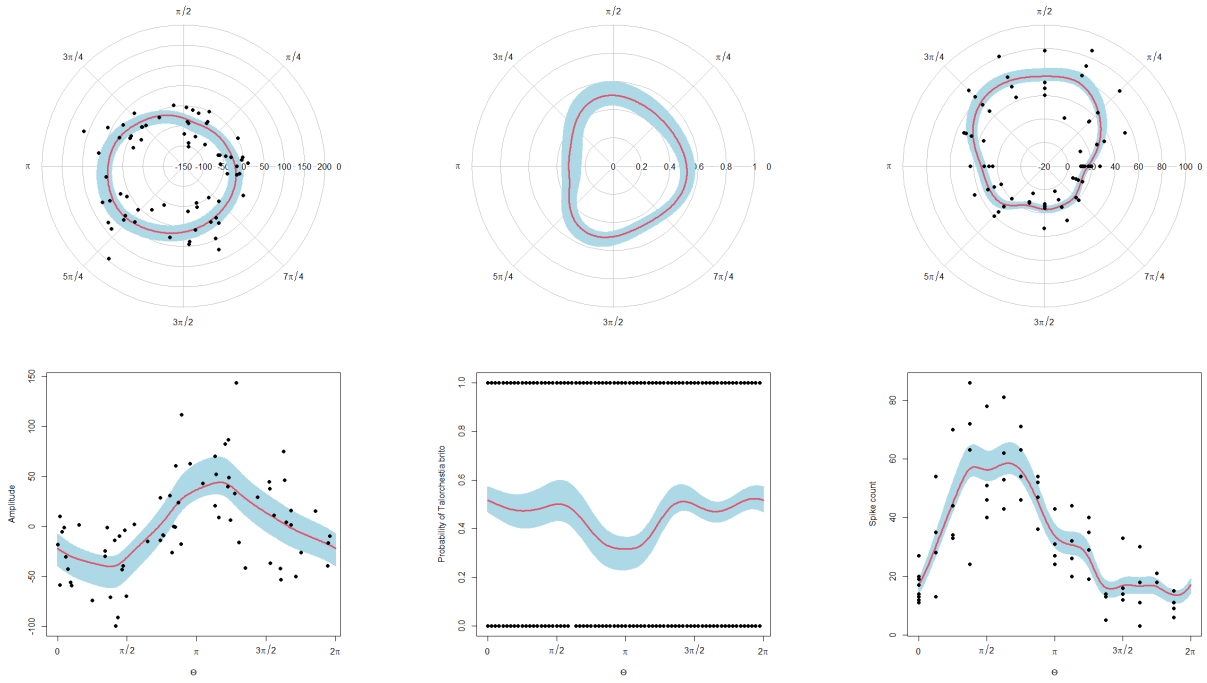


Figure 3: Radial and planar representations of the human resonance data with the nonparametric estimator of the regression function (left), the sandhopper data with the estimated probability of belonging to the *Talorchestia brito* species (center) and the spike count data with the nonparametric estimator of the mean function. The concentration parameter was selected by the refined rule in all cases and point-wise 95% confidence bands are displayed.

Sandhoppers data: Bernoulli response. The sandhoppers data corresponds to an experimental study carried out by Scapini et al. (2002) in which the aim was to investigate how sandhoppers from two different species (*Talitrus saltator* and *Talorchestia brito*) behaved when released in the sand. Here, we focus on estimating the probability of the animals belonging to each of the two species according to the direction in which they escape, by using the estimator presented in Section 2 in the particular case of a Bernoulli likelihood. The sample size is $n = 1644$, where 867 animals belonged to the *Talitrus saltator* species and 777 to the *Talorchestia brito* species. We estimated the logit function in equation (S1) of the Supplementary Material and represented the estimated probability of belonging to the *Talorchestia brito* species in the central panels of Figure 3. The estimated probability is around 0.5 for most of the escape directions, but there is a reasonable reduction of the probability of belonging to the *Talorchestia brito* species for escape directions around π.

The point-wise 95% confidence band suggests that this reduction in the probability is not just an artifact due to the randomness of the data.

Spike count data: Poisson model. We illustrate the estimator in the case of a Poisson likelihood with the spike count dataset, obtained from an experiment with an anesthetized and paralyzed adult male monkey (*Macaca nemestrina*), who was presented with a visual stimuli consisting on moving dots with different moving directions. The experiment is fully described in Kohn and Movshon (2003). The data consists of the total number of spikes per trial and the stimulus direction in a V5/MT cell and the sample size is $n = 68$.

Radial and planar representations of the dataset are shown in the right panels of Figure 3. To get more insight into the relationship between the variables, we maximize the circular local likelihood function to obtain the estimate of the transformed mean function $g(\theta) = \log(\mu(\theta))$. The estimated mean function, $\hat{\mu}(\theta) = \exp\{\hat{g}(\theta)\}$, is represented in the right panels of Figure 3. According to the estimator, the mean count of spikes is larger when the stimulus direction is approximately $\pi/2$, and a lower expected number of spikes is estimated when the stimulus direction is opposite from that, approximately at $3\pi/2$.

Pm10 particles data: gamma model. Now we consider the pm10 particle dataset introduced in Section 1, where the response variable presents an asymmetric conditional distribution. The dataset consists of measurements of pm10 particles, expressed in $\mu g/m^3$, and recordings of the wind direction (with 0 degrees representing the north direction and a clockwise sense of rotation) and wind speed, expressed in km/hour, in a meteorological station in Pontevedra, Spain. Data on pm10 particles is measured hourly, while meteorological data is measured on a 10-minute base, and the period of study is the year 2019. In order to account for temporal dependence, we have used a subsample of the data, with observations taken every six hours. In this section we only consider the wind direction as the predictor (an extension accounting for the effect of the wind speed is discussed in Section 7). We aim to study how the pm10 particles change with the direction of the wind. One must note that there is a pulp factory approximately 2.5 km south-west from the meteorological station and, hence, it would be of interest to determine if the concentration of pm10 particles is actually higher when the wind blows from the factory's direction. For this aim, we have also removed observations in which the wind speed was lower than 1 km/hour, since winds with speed between 0 and 1 km/hour are considered as periods of calm according to the Beaufort scale. The final sample size is $n = 1156$.

In the left panel of Figure 1 we have represented the pm10 measurements against the wind direction. By assuming a gamma likelihood, we have estimated the logarithm of the conditional mean function with the estimator presented in Section 2. The estimator of the

mean is represented in the left panel of Figure 1, where the approximate direction of the wind that blows from the pulp factory is represented with a star. Since the values of the response variable distort the representation of the mean function, the top right panel of Figure 1 shows a zoomed planar representation of the data and the estimated mean, with the 95% point-wise confidence band. It can be observed that the mean concentration of pm10 particles changes with the direction of the wind and it seems that the concentration is higher when the wind blows from the South-West/West direction.

7 Extensions

7.1 Partially linear and additive models

Throughout this work, we have considered the estimation of regression functions (or transformed regression functions) with just a single circular predictor. However, in many practical situations, several covariates (possibly of different nature) may influence a response variable. Consider, for example, the pm10 particles dataset represented in Figure 1. In order to study the concentration of pm10 particles, it seems reasonable to consider not only the wind direction as the covariate, but also the wind speed. Furthermore, these covariates could enter the model in a nonparametric form (for example maximizing the local log-likelihood as in Fan et al. (1998)) or in a parametric (possibly linear) way.

Consider a more general model with circular and real-valued covariates, with some of them entering the model parametrically and others having a nonparametric effect on the response. For the parametric part, assume that real-valued covariates enter the model linearly, while circular covariates can also enter the model linearly by means of their sine and cosine components. Let g be the target function to estimate, which depends on real-valued covariates $X_1, \ldots, X_k, X_{k+1}, \ldots, X_r$ and circular covariates $\Theta_1, \ldots, \Theta_j, \Theta_{j+1}, \ldots, \Theta_s$. Assume that $X_1, \ldots, X_k$ have a linear effect on the response, $\Theta_1, \ldots, \Theta_j$ also have a linear effect on the response through their sine and cosine components and $X_{k+1}, \ldots, X_r, \Theta_{j+1}, \ldots, \Theta_s$ have an unknown (nonparametric) effect on the response. Then, we can model the target function g as

$$
\begin{aligned}
&g(X_1, \ldots, X_k, X_{k+1}, \ldots, X_r, \Theta_1, \ldots, \Theta_j, \Theta_{j+1}, \ldots, \Theta_s) \\
&= \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k + \gamma_{11} \cos(\Theta_1) + \gamma_{12} \sin(\Theta_1) + \ldots + \gamma_{j1} \cos(\Theta_j) + \gamma_{j2} \sin(\Theta_j) \\
&+ \eta_{k+1}(X_{k+1}) + \ldots + \eta_r(X_r) + \rho_{j+1}(\Theta_{j+1}) + \ldots + \rho_s(\Theta_s),
\end{aligned}
$$

where $\alpha_0, \ldots, \alpha_k, \gamma_{11}, \gamma_{12}, \ldots, \gamma_{j1}, \gamma_{j2}$ are the parameters corresponding to the parametric part and $\eta_{k+1}, \ldots, \eta_r, \rho_{j+1}, \ldots, \rho_s$ represent unknown functions.

Model estimation can be carried out through a backfitting algorithm: first, the global parameters are estimated by maximizing the log-likelihood. By including the estimated

global parameters on the model, one of the nonparametric functions is estimated by maximizing the local log-likelihood (with the approach of Fan et al. (1998) if the covariate is real-valued or with the method of Section 2 if the covariate is circular). Now, the new estimated function is included in the model and the next nonparametric function is estimated as before, until one has estimated all smooth functions. The previous steps are repeated until convergence. This approach entails the selection of one bandwidth parameter for each real-valued covariate entering the model nonparametrically as well as one concentration parameter for each circular covariate with a nonparametric effect.

For the pm10 particles dataset, we have assumed a gamma likelihood and a partially linear model for the logarithm of the mean function:

$$g(X, \Theta) = \log(\mu(X, \Theta)) = \alpha_0 + \alpha_1 X + \rho(\Theta),$$

where X denotes wind speed and Θ denotes wind direction. The estimated average pm10 concentration is represented in the bottom left panel of Figure 1, where the angle represents the wind direction, the distance to the center of the circle represents the wind speed and the colour indicates the estimated mean concentration of pm10 particles (as indicated in the legend). As it can be seen, larger values of the wind speed lead to an increase in the concentration of pm10 particles. The largest concentration is obtained for wind directions coming from the South-West/West direction, where the pulp factory is located.

7.2 Hyperspherical covariates

The methodology proposed in this work can be extended to account for hyperspherical covariates, which can be seen as a generalization of circular variables. Let $\boldsymbol{X}$ be a hyperspherical random variable supported on $\mathbb{S}^d = \{\boldsymbol{x} \in \mathbb{R}^{d+1} : \boldsymbol{x}^\top \boldsymbol{x} = 1\}$. Consider the regression setting where Y is a real-valued variable, discrete or continuous, and $\boldsymbol{X}$ is a hyperspherical covariate. The estimation of a (conditional) characteristic of interest, $g(\boldsymbol{x})$ can be performed by constructing a hyperspherical kernel weighted local likelihood estimator by following the approach presented in this paper and taking into account the following remarks.

Given a sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, the Taylor-like expansion employed in (2) can be substituted by a projected Taylor expansion, obtained after performing a radial projection of the target function from the d-dimensional sphere to $\mathbb{R}^{d+1}$ as in García-Portugués et al. (2016). The construction of the kernel weighted log-likelihood can be done by employing a spherical kernel $L_\kappa(\boldsymbol{x}, \boldsymbol{X}_i)$, such as the von Mises-Fisher kernel. Note that condition (6) must be replaced by $L_\kappa(\boldsymbol{x}, \boldsymbol{X}_i) = k_{\kappa,d}(L) L\left(\kappa(1 - \boldsymbol{x}^\top \boldsymbol{X}_i)\right)$, where κ is the concentration parameter, $k_{\kappa,d}(L)$ is a normalization constant and L satisfies (7).

Taking into account the hyperspherical kernel, an asymptotic normality result analogous to Theorem 1 can be obtained where, in order to compute the expectations and variances of the quantities involved in the proof (see Section S2.1 of the Supplementary Material), the spherical change of variables in Lemma 2 of García-Portugués et al. (2013) and in Efthimiou and Frye (2014, pgs. 91-93) are employed. The approximation of the bias and variance of the estimator can be attained in an analogue way as in Section 3, which enables the construction of confidence intervals. A similar approach as the one in Section 4 can be derived to select the smoothing parameter.

8 Discussion

In this work, we have considered a general setting for kernel regression where the covariate presents a circular nature and the conditional distribution of the response variable given the covariate is a known and arbitrary parametric model. An estimator for a general conditional characteristic of interest is obtained by first approximating the target function by a Taylor-like sine-polynomial and, second, by maximizing the circular kernel weighted local likelihood. Particular cases of this estimator include the classical least-squares regression estimator or the logistic regression estimator (for circular covariate), already considered in the literature, but also encompass many other settings such as Poisson, gamma or beta regression, for example. The consideration of more complex settings with different types of covariates, including hyperspherical ones, has been also discussed.

Accurate approximations for the bias and variance of this general estimator were given, which allows to construct different inferential tools, such as pointwise confidence bands. In addition, an automatic rule to select the smoothing parameter, based on the approximated bias and variance, was provided. Simulation experiments for the Gaussian, Bernoulli, Poisson and gamma conditional distributions were conducted in order to ascertain the behaviour of the estimator with finite sample sizes and to compare the new selector to other alternatives. The empirical study showed that the performance of the refined rule to select the concentration parameter is much more satisfactory than cross-validation, except for the logistic case, where the refined rule is preferred only for large sample sizes.

Regarding the construction of the estimator, the sine-polynomial in (1) is used to approximate the target function locally. However, any periodic local model could be employed in its place, such as one with different phase and frequency parameters. Since the estimator's performance depends heavily on the selection of the smoothing parameter, any appropriate local model would obtain a satisfactory behaviour, provided it comes along with a good selection of the concentration parameter. The main benefits regarding the use of (1) are that it arises naturally through a Taylor expansion and that it also allows for the estimation of the target function's derivatives, apart from a simple and inexpensive

computation.

As mentioned above, the approximation of the bias and variance allows the computation of confidence intervals. Nevertheless, when employing the pointwise bands in Figures 1 and 3 it is necessary to take into account that the coverage $1-\alpha$ is only for a given θ_0. In order to assure that the target function is contained in the band for all the values of the covariate, one should consider simultaneous confidence bands. Such bands could be constructed by, for example, considering different results regarding the asymptotic distribution or through a bootstrap procedure (see, e.g., Hall et al., 2000; Li et al., 2010). It should be noted that, usually, simultaneous confidence bands tend to be quite conservative, achieving covering rates much lower than the desired coverage level.

On another note, the extension to partially linear models in Section 7.1 could be modified to account for other (partially) parametric models. For instance, in the case of real-valued covariates, some of the variables could enter the model parametrically through a quadratic or a polynomial effect, while the effect of the circular variables could be expressed through more sine and cosine components with different frequencies, as in truncated Fourier series. This truncated Fourier series approach is usually employed as a parametric regression model for circular covariates known as the extended cosine model (see, for example, Pewsey et al., 2013, Ch. 8). This is usually considered as an analogue to a polynomial regression model for circular covariates, and the use of too many terms would lead to overfitting issues.

Finally, it also seems interesting to note that the truncated Fourier series approach, or extended cosine model, resembles the Fourier basis approach employed for nonparametric regression (see, e.g., Rice and Rosenblatt, 1981). In fact, the idea behind the construction of both methodologies is the same, although in the latter a penalization on the curvature of the target function is added, with the penalization parameter controlling the smoothness of the estimator. This evidences that the kernel method presented in this paper is not the only form of nonparametric regression for circular covariates, and other methods, such as periodic splines, could be considered.

SUPPLEMENTARY MATERIAL

Supplementary proofs and simulation results Document containing technical proofs and complementary simulation results (PDF file)

Code and datasets R scripts with code and workflow and data files (zipped folder)

Acknowledgments

The authors thank the Associate Editor and three anonymous reviewers for their helpful comments, which considerably improved the quality of the paper.

References

Bai, Z. D., Rao, C. R., and Zhao, L. C. (1988). Kernel estimators of density function of directional data. *J. Multivar. Anal.*, 27:24–39.

Di Marzio, M., Fensore, S., and Panzera, A. (2018). Nonparametric classification for circular data. In Ley, C. and Verdebout, T., editors, *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman & Hall/CRC, New York.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statist. Probab. Lett.*, 798:2066–2075.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2014). Nonparametric regression for spherical data. *J. Am. Stat. Assoc.*, 109:748–763.

Efthimiou, V. and Frye, C. (2014). *Spherical Harmonics In p Dimensions.* World Scientific.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *J. R. Statist. Soc. B*, 60(3):591–608.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc. B*, 57:371–394.

García-Portugués, E. (2014). *Nonparametric Inference with Directional and Linear Data.* PhD thesis, Departamento de Estatística e Investigación Operativa, Facultade de Matemáticas, Universidade de Santiago de Compostela, Spain.

García-Portugués, E., Crujeiras, R. M., and González-Manteiga, W. (2013). Kernel density estimation for directional-linear data. *J. Multivar. Anal.*, 121(1):152–275.

García-Portugués, E., Van Keilegom, I., Crujeiras, R. M., and González-Manteiga, W. (2016). Testing parametric models in linear-directional regression.. *Scand. J. Stat.*, 43:1178–1191.

Hall, P., Lee, S. M.-S., and Young, G. A. (2000). Importance of interpolation when constructing double-bootstrap confidence intervals. *J. R. Statist. Soc. B*, 62:479–491.

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

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

Kohn, A. and Movshon, J. A. (2003). Neuronal adaptation to visual motion in area mt of the macaque. *Neuron.*, 39(4):681–691.

Ley, C. and Verdebout, T. (2017). *Modern Directional Statistics*. CRC Press, Boca Ratón.

Li, J., Zhang, C., Doksum, K. A., and Nordheim, E. V. (2010). Simultaneous confidence intervals for semiparametric logistics regression and confidence regions for the multi-dimensional effective dose. *Stat. Sin.*, 20:637–659.

Mardia, K. and Jupp, P. (2000). *Directional Statistics.* John Wiley, Chichester.

Oliveira, M., Crujeiras, R., and Rodríguez-Casal, A. (2013). Nonparametric circular methods for exploring environmental data. *Environ. Ecol. Stat.*, 20:1–17.

Patrangenaru, V. and Ellingson, L. (2016). *Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis*. CRC Press, Boca Raton.

Pewsey, A., Neuhäuser, M., and Ruxton, G. D. (2013). *Circular Statistics in R.* Oxford University Press, Oxford.

Puglisi, G., Leonetti, A., Landau, A., Fornia, L., Cerri, G., and Borroni, P. (2017). The role of attention in human motor resonance. *PLOS ONE*, 12(5):e0177457.

Rice, J. and Rosenblatt, M. (1981). Integrated mean squared error of a smoothing spline. *J. Approx. Theory*, 33(4):353–369.

Scapini, F., Aloia, A., Bouslama, M., Chelazzi, L., Colombini, I., El Gtari, M., Fallaci, M., and Marchetti, G. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed mediterranean beach. *Behav. Ecol.*, 51:403–414.

A Proof of Proposition 2

We give the derivation of the result obtained in Proposition 1. We make use of some preliminary results which we summarise in the following lemma.

Lemma 1. *Given a sample of circular data $\Theta_1, \ldots, \Theta_n$ with twice continuously density f with $f(\theta_0) > 0$ and a kernel K_κ satisfying (6) and (7), we have*

a) $s_{n,j} = \sum_{i=1}^n \sin^j(\Theta_i - \theta_0) K_\kappa(\Theta_i - \theta_0) = nf(\theta_0) 2^{\frac{j}{2}} \kappa^{-\frac{j}{2}} [\widetilde{b}_j^*(K) + o_P(1)]$.

b) $\gamma_{n,j} = \sum_{i=1}^n \sin^j(\Theta_i - \theta_0) K_\kappa^2(\Theta_i - \theta_0) = nf(\theta_0) 2^{\frac{j-1}{2}} \kappa^{-\frac{j-1}{2}} [\widetilde{d}_j^*(K) + o_P(1)]$.

The proof is given in Section S2 of the Supplementary Material.

Proof of Proposition 1. The proof is based on the proof of Theorem 1 in Fan and Gijbels (1995), taking into account the results in Lemma 1. By denoting $d_n = \text{tr}[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top \boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top \boldsymbol{W}]$, we can express $\hat{\sigma}^2(\theta_0)$ as

$$\begin{aligned}\hat{\sigma}^2(\theta_0) &= d_n^{-1} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 K_\kappa(\Theta_i - \theta_0) = d_n^{-1} (\boldsymbol{Y} - \boldsymbol{\Theta}\hat{\boldsymbol{\beta}})^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{\Theta}\hat{\boldsymbol{\beta}}) \\ &= d_n^{-1} \boldsymbol{Y}^\top [\boldsymbol{W} - \boldsymbol{W}\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top \boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top \boldsymbol{W}] \boldsymbol{Y}.\end{aligned}$$

Now, because of the continuity of $\sigma^2(\theta)$, we have

$$\mathbb{E}[\hat{\sigma}^2(\theta_0)|\Theta_1,\ldots,\Theta_n] = d_n^{-1}\boldsymbol{m}^\top[\boldsymbol{W}-\boldsymbol{W}\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top\boldsymbol{W}]\boldsymbol{m}+\sigma^2(\theta_0), \tag{23}$$

where $\boldsymbol{m} = (m(\Theta_1),\ldots,m(\Theta_n))^\top$. Further, we can approximate $\boldsymbol{r} = \boldsymbol{m} - \boldsymbol{\Theta}\boldsymbol{\beta}$ by the vector with elements

$$r(\Theta_i) = m(\Theta_i) - \sum_{j=0}^{p}\beta_j\sin^j(\Theta_i-\theta_0) = \beta_{p+1}\sin^{p+1}(\Theta_i-\theta_0) + O_P\left(\kappa^{-\frac{p+2}{2}}\right). \tag{24}$$

In addition,

$$\begin{aligned} d_n &= \mathrm{tr}[\boldsymbol{W}-\boldsymbol{W}\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top\boldsymbol{W}] \\ &= s_{n,0} - \mathrm{tr}[(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top\boldsymbol{W}^2\boldsymbol{\Theta}] = nf(\theta_0) + O_P\left(\kappa^{\frac{1}{2}}\right), \end{aligned} \tag{25}$$

where we have used Lemma 1. Then, starting from (23) and applying (24) and (25),

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2(\theta_0)|\Theta_1,\ldots,\Theta_n] &= d_n^{-1}[s_{n,2p+2} - \boldsymbol{c}_n^\top\boldsymbol{S}_n^{-1}\boldsymbol{c}_n]\beta_{p+1}^2 + \sigma^2(\theta_0) + o_P\left(\kappa^{-(p+1)}\right) \\ &= C_p\beta_{p+1}^2 2^{p+1}\kappa^{-(p+1)} + \sigma^2(\theta_0) + o_P\left(\kappa^{-(p+1)}\right), \end{aligned} \tag{26}$$

where $\boldsymbol{c}_n = (s_{n,p+1},\ldots,s_{n,2p+1})^\top$. On the other hand, we have

$$N^{-1} = \boldsymbol{e}_1^\top(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top\boldsymbol{W}^2\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{e}_1 = \boldsymbol{e}_1^\top\boldsymbol{S}_n^{-1}\boldsymbol{\Gamma}_n\boldsymbol{S}_n^{-1}\boldsymbol{e}_1.$$

By recalling Lemma 1, it is easy to see that

$$\boldsymbol{S}_n = nf(\theta_0)[\boldsymbol{L}\boldsymbol{B}\boldsymbol{L} + o_P(\boldsymbol{L}\boldsymbol{1}\boldsymbol{L})], \quad \text{and} \quad \boldsymbol{\Gamma}_n = 2^{-\frac{1}{2}}nf(\theta_0)\kappa^{\frac{1}{2}}[\boldsymbol{L}\boldsymbol{D}\boldsymbol{L} + o_P(\boldsymbol{L}\boldsymbol{1}\boldsymbol{L})], \tag{27}$$

where $\boldsymbol{L} = \mathrm{diag}\left\{1, 2^{\frac{1}{2}}\kappa^{-\frac{1}{2}},\ldots,2^{\frac{p}{2}}\kappa^{-\frac{p}{2}}\right\}$ and $\boldsymbol{1}$ denotes the $(p+1)\times(p+1)$ matrix with all elements equal to 1. Thus,

$$\begin{aligned} N^{-1} &= \frac{\kappa^{\frac{1}{2}}}{2^{\frac{1}{2}}nf(\theta_0)}\boldsymbol{e}_1^\top\boldsymbol{L}^{-1}\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{L}^{-1}\boldsymbol{e}_1 + o_P\left(n^{-1}\kappa^{\frac{1}{2}}\right) \\ &= \frac{\kappa^{\frac{1}{2}}}{2^{\frac{1}{2}}nf(\theta_0)}\boldsymbol{e}_1^\top\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{e}_1 + o_P\left(n^{-1}\kappa^{\frac{1}{2}}\right) = \frac{a_0\kappa^{\frac{1}{2}}}{2^{\frac{1}{2}}nf(\theta_0)} + o_P\left(n^{-1}\kappa^{\frac{1}{2}}\right). \end{aligned} \tag{28}$$

The combination of (26) and (28) leads to

$$\mathbb{E}[\mathrm{CRSC}(\theta_0;\kappa)|\Theta_1,\cdots,\Theta_n] = C_p\beta_{p+1}^2 2^{p+1}\kappa^{-(p+1)} + \sigma^2(\theta_0) + (p+1)\frac{\sigma^2(\theta_0)a_0\kappa^{\frac{1}{2}}}{2^{\frac{1}{2}}nf(\theta_0)} + o_P\left(\frac{1}{\kappa^{p+1}} + \frac{\kappa^{\frac{1}{2}}}{n}\right).$$

□

Supplementary Material to *A general framework for circular local likelihood regression*

Abstract

This Supplementary Material for the paper "A general framework for circular local likelihood regression" provides further results complementing the main text. Section S1 gives details on the estimator presented in Section 2 of the main manuscript when the variable of interest follows a Bernoulli distribution. Section S2 provides technical proofs: Subsection S2.1 gives the proof of Theorem 1, in addition to some discussion on the extension of the proof to an even more general setting; Subsection S2.2 contains the proof of Lemma 1, needed for the derivation of Proposition 1 in the main manuscript; Subsection S2.3 contains the derivation of the concentration parameter in the least-squares case. Finally, Section S3 provides additional graphical results of the simulation experiments carried out in Section 5 of the main text, including a detailed description of the results.

Keywords: Circular data, Data-driven smoothing selection, Local likelihood, Nonparametric regression

S1 Particular case: Bernoulli distribution

Section 2.1 of the main text contains details on the circular local likelihood estimator when the response variable follows Normal and Poisson distributions. Now, we consider the case where the variable of interest, Y, is a binary variable, with a Bernoulli conditional distribution, *i.e.*, $[Y|\Theta = \theta_0] \sim \text{Bernoulli}(p(\theta_0))$. Consider the target function

$$g(\theta_0) = \text{logit}(p(\theta_0)) = \log\left(\frac{p(\theta_0)}{1-p(\theta_0)}\right). \tag{S1}$$

Then, we have

$$l(g(\theta_0), y) = yg(\theta_0) - \log\left(1 + \exp\{g(\theta_0)\}\right)$$

and, thus, the local circular kernel log-likelihood is given by

$$\mathcal{L}_p(\boldsymbol{\beta}; \kappa, \theta_0) = \sum_{i=1}^{n} \left(Y_i \boldsymbol{\Theta}_i^\top \boldsymbol{\beta} - \log\left(1 + \exp\{\boldsymbol{\Theta}_i^\top \boldsymbol{\beta}\}\right)\right) K_\kappa(\Theta_i - \theta_0),$$

leading to a nonparametric logistic regression model with a circular covariate, studied by Di Marzio et al. (2018) in the context of nonparametric classification.

S2 Technical proofs and derivations

We will make extensive use of the Taylor-like sine expansion employed by Di Marzio et al. (2009), which is based on the fact that, for small values of α, we have $\sin\alpha \approx \alpha$, and then, for a function v,

$$v(\theta+\alpha) = v(\theta) + v'(\theta)\sin\alpha + \ldots + \frac{1}{k!}v^{(k)}(\theta)\sin^k\alpha + O\left(\sin^{k+1}\alpha\right). \tag{S2}$$

S2.1 Proof of Theorem 1

Proof of Theorem 1. The scheme of this proof is based on the proofs of Lemma 1 and Lemma 2 in Fan et al. (1995). The following notation will be used:

$$\bar{g}(\theta_0, \Theta_i) = g(\theta_0) + g'(\theta_0)\sin(\Theta_i - \theta_0) + \ldots + \frac{g^{(p)}(\theta_0)}{p!}\sin^p(\Theta_i - \theta_0)$$

and

$$\boldsymbol{B}_i = \begin{pmatrix} 1 \\ \sin(\Theta_i - \theta_0)\kappa^{\frac{1}{2}} \\ \vdots \\ \sin^p(\Theta_i - \theta_0)\kappa^{\frac{p}{2}}/p! \end{pmatrix}.$$

Recall that the normalized estimator is given by

$$\widehat{\boldsymbol{\beta}}_N = n^{\frac{1}{2}}\kappa^{-\frac{1}{4}}\left(\hat{\beta}_0 - g(\theta_0), \kappa^{-\frac{1}{2}}[\hat{\beta}_1 - g'(\theta_0)], \ldots, \kappa^{-\frac{p}{2}}[p!\hat{\beta}_p - g^{(p)}(\theta_0)]\right)^\top.$$

It is clear that $\widehat{\boldsymbol{\beta}}_N$ maximizes

$$\ell(\boldsymbol{\beta}_N) = \sum_{i=1}^{n}\left\{l[\bar{g}(\theta_0,\Theta_i) + n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\boldsymbol{\beta}_N^\top\boldsymbol{B}_i, Y_i] - l[\bar{g}(\theta_0,\Theta_i), Y_i]\right\}K[\kappa\{1-\cos(\Theta_i-\theta_0)\}]$$

with respect to $\boldsymbol{\beta}_N$. After a Taylor expansion on the log-likelihood function,

$$\begin{aligned}
\ell(\boldsymbol{\beta}_N) &= n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\sum_{i=1}^{n} l'[\bar{g}(\theta_0,\Theta_i), Y_i]\boldsymbol{\beta}_N^\top\boldsymbol{B}_i K[\kappa\{1-\cos(\Theta_i-\theta_0)\}] \\
&\quad + \frac{n^{-1}\kappa^{\frac{1}{2}}}{2}\sum_{i=1}^{n} l^{(2)}[\bar{g}(\theta_0,\Theta_i), Y_i](\boldsymbol{\beta}_N^\top\boldsymbol{B}_i)^2 K[\kappa\{1-\cos(\Theta_i-\theta_0)\}] \\
&\quad + \frac{n^{-\frac{3}{2}}\kappa^{\frac{3}{4}}}{6}\sum_{i=1}^{n} l^{(3)}(g_a, Y_i)(\boldsymbol{\beta}_N^\top\boldsymbol{B}_i)^3 K[\kappa\{1-\cos(\Theta_i-\theta_0)\}] \\
&= \boldsymbol{W}_n^\top\boldsymbol{\beta}_N^\top + \frac{1}{2}\boldsymbol{\beta}_N^\top\boldsymbol{A}_n\boldsymbol{\beta}_N + \frac{n^{-\frac{3}{2}}\kappa^{\frac{3}{4}}}{6}\sum_{i=1}^{n} l^{(3)}(g_a, Y_i)(\boldsymbol{\beta}_N^\top\boldsymbol{B}_i)^3 K[\kappa\{1-\cos(\Theta_i-\theta_0)\}],
\end{aligned} \tag{S3}$$

where g_a is a random value between $\bar{g}(\theta_0,\Theta_i)$ and $\bar{g}(\theta_0,\Theta_i) + n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\boldsymbol{\beta}_N^\top\boldsymbol{B}_i$,

$$\boldsymbol{W}_n = n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\sum_{i=1}^{n} l'[\bar{g}(\theta_0,\Theta_i), Y_i]\boldsymbol{B}_i K[\kappa\{1-\cos(\Theta_i-\theta_0)\}]$$

and

$$\boldsymbol{A}_n = n^{-1}\kappa^{\frac{1}{2}}\sum_{i=1}^{n} l^{(2)}[\bar{g}(\theta_0,\Theta_i), Y_i]\boldsymbol{B}_i\boldsymbol{B}_i^\top K[\kappa\{1-\cos(\Theta_i-\theta_0)\}].$$

In the following it will be shown that, as $n\to\infty$,

$$\boldsymbol{A}_n = \boldsymbol{A} + o_P(1), \tag{S4}$$

with $\boldsymbol{A}$ defined in the main text. It holds that $(\boldsymbol{A}_n)_{ij} = \mathbb{E}[(\boldsymbol{A}_n)_{ij}] + O_P\left\{[\mathrm{Var}(\boldsymbol{A}_n)]_{ij}^{1/2}\right\}$. For the expectation part, employing the law of iterated expectation and noting that $l^{(r)}[g(\theta), y]$ is linear in y yields

$$\mathbb{E}[(\boldsymbol{A}_n)_{ij}] = \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\mathbb{E}\left\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), \mu(\Theta_1)]\sin^{i+j-2}(\Theta_1-\theta_0)K[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right\}.$$

A Taylor expansion gives

$$l^{(2)}[\bar{g}(\theta_0,\Theta_1), \mu(\Theta_1)] = l^{(2)}[g(\Theta_1), \mu(\Theta_1)] + [\bar{g}(\theta_0,\Theta_1) - g(\Theta_1)]l^{(3)}[g_b, \mu(\Theta_1)],$$

where g_b is between $\bar{g}(\theta_0, \Theta_1)$ and $g(\Theta_1)$. Consequently,

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{A}_n)_{ij}] \\
&= \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\mathbb{E}\left\{l^{(2)}[g(\Theta_1), \mu(\Theta_1)]\sin^{i+j-2}(\Theta_1-\theta_0)K[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right\} \\
&+ \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\mathbb{E}\left\{[\bar{g}(\theta_0, \Theta_1) - g(\Theta_1)]l^{(3)}[g_b, \mu(\Theta_1)]\sin^{i+j-2}(\Theta_1-\theta_0)K[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right\} \\
&= (A) + (B).
\end{aligned} \tag{S5}$$

Now, the term (A) can be computed as

$$\begin{aligned}
(A) &= \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\int_0^{2\pi} l^{(2)}[g(\alpha), \mu(\alpha)]\sin^{i+j-2}(\alpha-\theta_0)K[\kappa\{1-\cos(\alpha-\theta_0)\}]f(\alpha)d\alpha \\
&= \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\int_0^{\pi} l^{(2)}[g(\theta_0+\varphi), \mu(\theta_0+\varphi)]\sin^{i+j-2}(\varphi)K[\kappa\{1-\cos\varphi\}]f(\theta_0+\varphi)d\varphi \\
&+ \frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\int_{\pi}^{2\pi} l^{(2)}[g(\theta_0+\varphi), \mu(\theta_0+\varphi)]\sin^{i+j-2}(\varphi)K[\kappa\{1-\cos\varphi\}]f(\theta_0+\varphi)d\varphi,
\end{aligned}$$

where the change of variables $\varphi = \alpha - \theta_0$ was employed. Next, the change of variables $r = \kappa(1-\cos\varphi)$ is used. For the first integral, $\varphi \in [0, \pi]$, giving $\varphi = \arccos(1 - r/\kappa)$ and $d\varphi = \kappa^{-1}\left(\frac{2r}{\kappa} - \frac{r^2}{\kappa^2}\right)^{-1/2} dr$. On the other hand, for the second integral $\varphi \in [\pi, 2\pi]$, and thus, $\varphi = -\arccos(1 - r/\kappa)$ and $d\varphi = -\kappa^{-1}\left(\frac{2r}{\kappa} - \frac{r^2}{\kappa^2}\right)^{-1/2} dr$. Consequently,

$$\begin{aligned}
(A) &= \frac{\kappa^{-\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\int_0^{2\kappa} \rho\left[\theta_0 + \arccos\left(1-\frac{r}{\kappa}\right)\right] f\left[\theta_0 + \arccos\left(1-\frac{r}{\kappa}\right)\right] \\
&\times \left(\frac{2r}{\kappa} - \frac{r^2}{\kappa^2}\right)^{\frac{i+j-3}{2}} K(r)dr \\
&+ \frac{\kappa^{-\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\int_0^{2\kappa} \rho\left[\theta_0 - \arccos\left(1-\frac{r}{\kappa}\right)\right] f\left[\theta_0 - \arccos\left(1-\frac{r}{\kappa}\right)\right] \\
&\times \left[-\left(\frac{2r}{\kappa} - \frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{i+j-2} \left(\frac{2r}{\kappa} - \frac{r^2}{\kappa^2}\right)^{-\frac{1}{2}} K(r)dr.
\end{aligned}$$

It is now necessary to distinguish two separate cases: $i+j-2$ even or odd. In the situation

where $i+j-2$ is even, it holds that

$$\begin{aligned}(A) &= \frac{2\kappa^{-\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{i+j-3}{2}}K(r)dr[1+o(1)] \\ &= \frac{2\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}\int_0^{2\kappa} r^{\frac{i+j-3}{2}}\left(2-\frac{r}{\kappa}\right)^{\frac{i+j-3}{2}}K(r)dr[1+o(1)] \\ &= \frac{2\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}\left[2^{\frac{i+j-3}{2}}\int_0^{\infty} r^{\frac{i+j-3}{2}}K(r)dr+o(1)\right][1+o(1)] \\ &= \frac{2^{\frac{i+j-1}{2}}\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}b_{i+j-2}(K)+o(1),\end{aligned}$$

where $b_j(K) = \int_0^{\infty} r^{\frac{j-1}{2}}K(r)dr < \infty$ (because of the assumption in equation (7) of the main text). Analogous arguments show that, when $i+j-2$ is odd, $(A) = o(1)$. Thus, it is possible to write

$$(A) = \frac{2^{\frac{i+j-1}{2}}\rho(\theta_0)f(\theta_0)}{(i-1)!(j-1)!}b^*_{i+j-2}(K)+o(1),$$

with

$$b^*_j(K) = \begin{cases} 0 & \text{if } j \text{ is odd,} \\ b_j(K) & \text{if } j \text{ is even.}\end{cases}$$

Now, because of assumption $C2$, the term $l^{(3)}[g_b, \mu(\Theta_1)]$ in the second addend of (S5) is bounded. Thus, by noting that $\bar{g}(\theta_0, \Theta_1) - g(\Theta_1) = o_P\left(\kappa^{-\frac{p}{2}}\right)$, similar arguments to the ones above show that the term (B) in (S5) is $o(1)$.

Next, the variance of $\boldsymbol{A}_n$ is considered. It holds that

$$\begin{aligned}&\mathrm{Var}[(\boldsymbol{A}_n)_{ij}] \\ &= \mathrm{Var}\left(\frac{n^{-1}\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\sum_{i=1}^{n} l^{(2)}[\bar{g}(\theta_0,\Theta_i), Y_i]\sin^{i+j-2}(\Theta_i-\theta_0)K[\kappa\{1-\cos(\Theta_i-\theta_0)\}]\right) \\ &\leq \frac{n^{-1}\ \kappa^{i+j-1}}{[(i-1)!]^2[(j-1)!]^2}\mathbb{E}\left(\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), Y_1]\}^2\sin^{2(i+j-2)}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right) \\ &= \frac{n^{-1}\ \kappa^{i+j-1}}{[(i-1)!]^2[(j-1)!]^{!2}}\mathbb{E}\left[\sin^{2(i+j-2)}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\mathbb{E}\left(\left\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), Y_1]\right\}^2\middle|\Theta_1\right)\right] \\ &= \frac{n^{-1}\ \kappa^{i+j-1}}{[(i-1)!]^2[(j-1)!]^2}\mathbb{E}\left[\left\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), \mu(\Theta_1)]\right\}^2\sin^{2(i+j-2)}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right] \\ &+ \frac{n^{-1}\ \kappa^{i+j-1}}{[(i-1)!]^2[(j-1)!]^2}\mathbb{E}\left[\mathrm{Var}\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), Y_1]|\Theta_1\}\sin^{2(i+j-2)}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right],\end{aligned}$$

where the linearity of $l^{(r)}[g(\theta), y]$ in y was employed again. Note that $\mathrm{Var}\{l^{(2)}[\bar{g}(\theta_0,\Theta_1), Y_1]|\Theta_1\}$ is a bounded term because of assumption $C2$. Then, analogous arguments to those used for the expectation of $\boldsymbol{A}_n$ yield $\mathrm{Var}[(\boldsymbol{A}_n)_{ij}] = O\left(n^{-1}\kappa^{\frac{1}{2}}\right)$, implying the result in (S4).

In addition, the third term in (S3) is $O_P\left(n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\right)$. This is shown by first noting that the random vector $\boldsymbol{B}_i$ is bounded (because of assumption C4) and afterwards noticing that the expectation of the absolute value of last term in (S3) is bounded by

$$O\left(n\kappa^{\frac{3}{4}}\mathbb{E}\left[|l^{(3)}(g_a, Y_1)K[\kappa\{1-\cos(\Theta_1-\theta_0)\}]|\right]\right),$$

which is $O\left(n^{-\frac{1}{2}}\kappa^{\frac{1}{4}}\right)$ by taking into account equation (5) of the main text and employing arguments similar to the ones above. Therefore, it holds

$$\ell(\boldsymbol{\beta}_N) = \boldsymbol{W}_n^\top \boldsymbol{\beta}_N^\top + \frac{1}{2}\boldsymbol{\beta}_N^\top \boldsymbol{A}\boldsymbol{\beta}_N + o_P(1).$$

Therefore, employing the quadratic approximation lemma (Fan and Gijbels, 1996, page 210) yields that if $\boldsymbol{W}_n$ is a stochastically bounded sequence of random vectors, the maximizer of $\ell(\boldsymbol{\beta}_N)$, namely $\widehat{\boldsymbol{\beta}_N}$, is given by

$$\widehat{\boldsymbol{\beta}_N} = \boldsymbol{A}^{-1}\boldsymbol{W}_n + o_P(1).$$

Thus, by proving the asymptotic normality of $\boldsymbol{W}_n$, one can verify the asymptotic normality of $\widehat{\boldsymbol{\beta}_N}$. First, the expectation of the jth element of $\boldsymbol{W}_n$ is computed by taking into account the linearity of $l^{(r)}[g(\theta), y]$ in y:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{W}_n)_j] &= n^{\frac{1}{2}}\kappa^{\frac{1}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}\int_0^{2\kappa} l'[\bar{g}(\theta_0,\alpha),\mu(\alpha)]\sin^{j-1}(\alpha-\theta_0))K[\kappa\{1-\cos(\alpha-\theta_0)\}]f(\alpha)d\alpha \\
&= n^{\frac{1}{2}}\kappa^{\frac{1}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}\int_0^{2\kappa} l'[\bar{g}(\theta_0,\theta_0+\varphi),\mu(\theta_0+\varphi)]\sin^{j-1}\varphi)K[\kappa\{1-\cos\varphi\}]f(\theta_0+\varphi)d\varphi,
\end{aligned}$$

where the change of variables $\varphi = \alpha - \theta_0$ was employed. Next, splitting the integral and performing the change of variables $r = \kappa[1-\cos\varphi]$ on each part gives

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{W}_n)_j] = n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}&\int_0^{2\kappa} l'\left\{\bar{g}[\theta_0,\theta_0+\arccos(1-r/\kappa)],\mu[\theta_0+\arccos(1-r/\kappa)]\right\} \\
&\times\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-2}{2}}K(r)f[\theta_0+\arccos(1-r/\kappa)]dr \\
+\, n^{1/2}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}&\int_0^{2\kappa} l'\left\{\bar{g}[\theta_0,\theta_0-\arccos(1-r/\kappa)],\mu[\theta_0-\arccos(1-r/\kappa)]\right\} \\
&\times\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{j-1}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{-\frac{1}{2}}K(r)f[\theta_0-\arccos(1-r/\kappa)]dr.
\end{aligned}$$

Now, note that, by performing a Taylor expansion around $g[\theta_0 \pm \arccos(1-r/\kappa)]$, we have

$$\begin{aligned}
&l'\left\{\bar{g}[\theta_0,\theta_0\pm\arccos(1-r/\kappa)],\mu[\theta_0\pm\arccos(1-r/\kappa)]\right\} = \rho[\theta_0\pm\arccos(1-r/\kappa)] \\
&\times\left\{\bar{g}[\theta_0,\theta_0\pm\arccos(1-r/\kappa)]-g[\theta_0\pm\arccos(1-r/\kappa)]\right\},
\end{aligned}$$

where it was employed that $l'[g(\theta),\mu(\theta)]=0\ \forall\ \theta$. In addition,

$$\begin{aligned}&\bar{g}[\theta_0,\theta_0\pm\arccos(1-r/\kappa)]-g[\theta_0\pm\arccos(1-r/\kappa)]\\&=-\left[\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\sin^{p+1}[\pm\arccos(1-r/\kappa)]+o\left(\kappa^{-\frac{p+1}{2}}\right)\right],\end{aligned}$$

which gives

$$\begin{aligned}&l'\left\{\bar{g}[\theta_0,\theta_0\pm\arccos(1-r/\kappa)],\mu[\theta_0\pm\arccos(1-r/\kappa)]\right\}=-\rho[\theta_0\pm\arccos(1-r/\kappa)]\\&\times\left[\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\sin^{p+1}[\pm\arccos(1-r/\kappa)]+o\left(\kappa^{-\frac{p+1}{2}}\right)\right].\end{aligned}$$

Then, it holds

$$\begin{aligned}\mathbb{E}[(\boldsymbol{W}_n)_j]&=\left(n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\int_0^{2\kappa}\rho[\theta_0+\arccos(1-r/\kappa)]\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{p+j-1}{2}}\right.\\&\left.\times K(r)f[\theta_0+\arccos(1-r/\kappa)]dr\right)\left[1+o\left(\kappa^{-\frac{p+1}{2}}\right)\right]\\&+\left(n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\int_0^{2\kappa}\rho[\theta_0-\arccos(1-r/\kappa)]\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{p+j}\right.\\&\left.\times\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{-\frac{1}{2}}K(r)f[\theta_0-\arccos(1-r/\kappa)]dr\right)\left[1+o\left(\kappa^{-\frac{p+1}{2}}\right)\right].\end{aligned}$$

It will now be assumed that p is odd (although derivations for an even value of p can be obtained in an equivalent way). If p is odd and j is even, then

$$\begin{aligned}\mathbb{E}[(\boldsymbol{W}_n)_j]&=\left(2n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!}\rho(\theta_0)f(\theta_0)\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{p+j-1}{2}}K(r)dr\right)\\&\times\left[1+o\left(\kappa^{-\frac{p+1}{2}}\right)\right]\\&=\left(2n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{\frac{j-1}{2}}}{(j-1)!\kappa^{\frac{p+j-1}{2}}}\rho(\theta_0)f(\theta_0)\frac{g^{(p+1)}(\theta_0)}{(p+1)!}\left[2^{\frac{p+j-1}{2}}\int_0^{\infty}r^{\frac{p+j-1}{2}}K(r)dr+o(1)\right]\right)\\&\times\left[1+o\left(\kappa^{-\frac{p+1}{2}}\right)\right]\\&=2^{\frac{p+j+1}{2}}n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{-\frac{p}{2}}}{(j-1)!}\rho(\theta_0)f(\theta_0)\frac{g^{(p+1)}(\theta_0)}{(p+1)!}b_{p+j}(K)[1+o(1)]\end{aligned}$$

where it was used that $\rho[\theta_0\pm\arccos(1-r/\kappa)]=\rho(\theta_0)+o(1)$ and $f[\theta_0\pm\arccos(1-r/\kappa)]=f(\theta_0)+o(1)$. Similar derivations show that if j is odd, $\mathbb{E}[(\boldsymbol{W}_n)_j]=o(1)$ and, thus, for a general j it can be written that

$$\mathbb{E}[(\boldsymbol{W}_n)_j]=2^{\frac{p+j+1}{2}}n^{\frac{1}{2}}\kappa^{-\frac{3}{4}}\frac{\kappa^{-\frac{p}{2}}}{(j-1)!}\rho(\theta_0)f(\theta_0)\frac{g^{(p+1)}(\theta_0)}{(p+1)!}b^*_{p+j}(K)[1+o(1)].$$

Now, the variance of $\boldsymbol{W}_n$ is computed. It holds that

$$\mathrm{Var}(\boldsymbol{W}_n) = \kappa^{\frac{1}{2}}\mathbb{E}\left(l'^2[\bar{g}(\theta_0,\Theta_1),Y_1]\boldsymbol{B}_1\boldsymbol{B}_1^\top K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right) + O\left(\kappa^{-\frac{2p+3}{2}}\right). \quad \text{(S6)}$$

Then, the (i,j)th element of the first term in (S6) is given by

$$\frac{\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!}\mathbb{E}\left[\sin^{i+j-2}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\mathbb{E}\left(l'^2[\bar{g}(\theta_0,\Theta_1),Y_1]|\Theta_1\right)\right]$$

and, taking into account equation (5) of the main text gives

$$\begin{aligned}
&[\mathrm{Var}(\boldsymbol{W}_n)]_{ij}\\
&= \frac{\mathrm{Var}(Y|\Theta=\theta_0)\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!\psi^2}\mathbb{E}\left(\sin^{i+j-2}(\Theta_1-\theta_0)K^2[\kappa\{1-\cos(\Theta_1-\theta_0)\}]\right) + o(1)\\
&= \frac{\mathrm{Var}(Y|\Theta=\theta_0)\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!\psi^2}\int_0^{2\pi}\sin^{i+j-2}(\alpha-\theta_0)K^2[\kappa\{1-\cos(\alpha-\theta_0)\}]f(\alpha)d\alpha + o(1)\\
&= \frac{\mathrm{Var}(Y|\Theta=\theta_0)\kappa^{\frac{1}{2}}\kappa^{\frac{i+j-2}{2}}}{(i-1)!(j-1)!\psi^2}\int_0^{2\pi}\sin^{i+j-2}\varphi K^2[\kappa\{1-\cos\varphi\}]f(\theta_0+\varphi)d\varphi + o(1).
\end{aligned}$$

Finally, splitting the integral into two and performing the change of variables $r = \kappa(1-\cos\varphi)$ yields

$$[\mathrm{Var}(\boldsymbol{W}_n)]_{ij} = \frac{2^{\frac{i+j-1}{2}}b''[g(\theta_0)]f(\theta_0)}{(i-1)!(j-1)!\psi}d^*_{i+j-2}(K) + o(1),$$

where

$$d^*_j(K) = \begin{cases} 0 & \text{if } j \text{ is odd,}\\ \int_0^\infty r^{\frac{j-1}{2}}K^2(r)dr & \text{if } j \text{ is even.}\end{cases}$$

It only remains to prove that

$$[\mathrm{Var}(\boldsymbol{W}_n)]^{-1/2}[\boldsymbol{W}_n - \mathbb{E}(\boldsymbol{W}_n)] \to^D N(0, I_{p+1}).$$

For that, the Cramér-Wold device is employed: it is enough to prove that, for any unit vector $\boldsymbol{c}$,

$$[\boldsymbol{c}^\top\mathrm{Var}(\boldsymbol{W}_n)\boldsymbol{c}]^{-1/2}[\boldsymbol{c}^\top\boldsymbol{W}_n - \boldsymbol{c}^\top\mathbb{E}(\boldsymbol{W}_n)\boldsymbol{c}] \xrightarrow{D} N(0,1).$$

The previous statement can be proved by checking Lyapounov's condition. It holds that

$$\boldsymbol{c}^\top\boldsymbol{W}_n = \frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^n \kappa^{\frac{1}{4}}l'[\bar{g}(\theta_0,\Theta_i),Y_i]\boldsymbol{c}^\top\boldsymbol{B}_i K[\kappa\{1-\cos(\Theta_i-\theta_0)\}] \equiv \frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^n V_{i,n},$$

where the dependence of $V_{i,n}$ on n is through κ. It is then enough to prove that, for some $\delta > 0$,

$$\lim_{n\to\infty}\frac{\mathbb{E}\left[|V_{i,n}-\mathbb{E}(V_{i,n})|^{2+\delta}\right]}{n^{\frac{\delta}{2}}\mathrm{Var}(V_{i,n})^{1+\frac{\delta}{2}}} = 0. \quad \text{(S7)}$$

In addition, it can be proved that $\mathbb{E}\left[|V_{i,n} - \mathbb{E}(V_{i,n})|^{2+\delta}\right] = O[\mathbb{E}(|V_{i,n}|^{2+\delta})]$ and

$$\mathbb{E}(|V_{i,n}|^{2+\delta}) = \mathbb{E}\left[\left|\kappa^{\frac{1}{4}} l'[\bar{g}(\theta_0, \Theta_i), Y_i]\boldsymbol{c}^\top \boldsymbol{B}_i K[\kappa\{1 - \cos(\Theta_i - \theta_0)\}]\right|^{2+\delta}\right].$$

Moreover, a Taylor expansion gives

$$l'[\bar{g}(\theta_0, \Theta_i), Y_i] = l'[g(\Theta_i), Y_i] + [\bar{g}(\theta_0, \Theta_i) - g(\Theta_i)]l'(g_c, Y_i),$$

where g_c is a random value between $\bar{g}(\theta_0, \Theta_i)$ and $g(\Theta_i)$. Now, taking into account the form of the log-likelihood in equation (5) of the main text and the first Barlett identity, it holds that

$$l'[g(\Theta_i), Y_i] = \frac{1}{\psi}[Y_i - \mu(\Theta_i)].$$

Then, $O\left[\mathbb{E}(|V_{i,n}|^{2+\delta})\right]$ is given by

$$O\left[\mathbb{E}\left(\left|\kappa^{\frac{1}{4}} \frac{1}{\psi}[Y_i - \mu(\Theta_i)]c_1 K[\kappa\{1 - \cos(\Theta_i - \theta_0)\}]\right|^{2+\delta}\right)\right][1 + o(1)] = O\left(\kappa^{\frac{\delta}{4}}\right),$$

by noting that $\mathbb{E}[|Y - \mu(\Theta)|^{2+\delta}] < \infty$ because of assumption C1. In addition, it holds that $\mathrm{Var}(V_{i,n}) \leq \mathbb{E}(V_{i,n}^2) = O(1)$. Thus, the quotient in (S7) is $O\left(n^{-\frac{\delta}{2}}\kappa^{\frac{\delta}{4}}\right)$ and hence converges to zero when $n\kappa^{-\frac{1}{2}} \to \infty$ as $n \to \infty$. □

Remark S1. As stated in Section 2.2 of the main text, the assumptions of Theorem 1 can be relaxed so that the asymptotic normality result is valid for conditional densities not belonging to the exponential family. First, one should note that the assumption on the third derivative of the function b (part of assumption C2) is placed to guarantee the continuity of $l^{(3)}$. Thus, in the case that the log-likelihood does not have the form in equation (5) of the main text, the continuity of $l^{(3)}$ should be assumed. In addition, the likelihood should satisfy Bartlett's first and second identities: the first identity is used to show that $\mathbb{E}[(\boldsymbol{W}_n)_j] < \infty \ \forall\ j = 1, \ldots, p+1$ and to verify Lyapunov's condition; the second identity is employed to show that $\mathrm{Var}[(\boldsymbol{A}_n)_{ij}] \to 0$ as $n \to \infty$. Lastly, a key step in the previous proof is that, when equation (5) of the main text is verified, $l^{(r)}[g(\theta), y]$ is linear in y. Note, however, that this is needed to compute the expressions of the bias and variance of the estimator. Without this assumption, the asymptotic normality result would hold, but the calculation of general expressions for the bias and variance of the estimator would be much more tedious, and should be done for each specific log-likelihood. The finiteness of such expressions is, however, guaranteed by the continuity of the third partial derivative of the log-likelihood.

S2.2 Proof of Lemma 1

Proof of Lemma 1. We first obtain the proof of the statement in a) and afterwards give the derivations for statement b).

Statement a)

We have

$$s_{n,j} = \mathbb{E}(s_{n,j}) + O_P\left(\sqrt{\mathrm{Var}(s_{n,j})}\right).$$

For the expectation part,

$$\mathbb{E}(s_{n,j}) = n\int_0^{2\pi} \sin^j(\alpha-\theta_0)K_\kappa(\alpha-\theta_0)f(\alpha)d\alpha = nc_\kappa(K)\int_0^{2\pi} \sin^j(\varphi)K[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi,$$

where the second equality was obtained by applying the change of variables $\varphi = \alpha - \theta_0$ and using equation (6) of the main text. Now, by splitting the integral, we have

$$\mathbb{E}(s_{n,j}) = nc_\kappa(K)\left[\int_0^{\pi} \sin^j(\varphi)K[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi + \int_\pi^{2\pi} \sin^j(\varphi)K[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi\right]. \tag{S8}$$

For the first integral in (S8), we apply the change of variables $r = \kappa(1-\cos\varphi)$ noting that, since $\varphi \in [0,\pi]$, we have $\varphi = \arccos(1-r/\kappa)$, and thus obtaining, for the first integral in (S8):

$$\int_0^{\pi} \sin^j(\varphi)K[\kappa(1-\cos\varphi)]d\varphi = \frac{1}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}} K(r)f(\theta_0+\arccos(1-r/\kappa))dr.$$

Applying the expansion in (S2), we have that $f(\theta_0+\arccos(1-r/\kappa)) = f(\theta_0)+o(1)$ and, then, the first integral in (S8) can be expressed as

$$\frac{1}{\kappa}f(\theta_0)\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}} K(r)dr[1+o(1)]. \tag{S9}$$

Now we move on to the second integral in (S8). By employing again the change of variables $r = \kappa(1-\cos\varphi)$, but taking into account that now $\varphi \in [\pi, 2\pi]$ and, thus, having $\varphi = -\arccos(1-r/\kappa)$, we obtain

$$\int_\pi^{2\pi} \sin^j(\varphi)K[\kappa(1-\cos\varphi)]d\varphi = -\frac{1}{\kappa}\int_0^{2\kappa}\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{j-1} K(r)f\left(\theta_0-\arccos(1-r/\kappa)\right)dr. \tag{S10}$$

Note that we have two different scenarios depending on j being odd or even. When j is even ($j-1$ is odd), the right term in (S10) can be expressed as

$$\frac{1}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}} K(r)f(\theta_0-\arccos(1-r/\kappa))dr = \frac{1}{\kappa}f(\theta_0)\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}} K(r)dr[1+o(1)]$$

and, then, equation (S8) for an even j becomes

$$\begin{aligned}
\mathbb{E}(s_{n,j}) &= nc_\kappa(K)\frac{2}{\kappa}f(\theta_0)\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}}K(r)dr[1+o(1)] \\
&= nc_\kappa(K)2\kappa^{-\frac{j+1}{2}}f(\theta_0)\int_0^{2\kappa}r^{\frac{j-1}{2}}\left(2-\frac{r}{\kappa}\right)^{\frac{j-1}{2}}K(r)dr[1+o(1)] \\
&= n2\kappa^{-\frac{j}{2}}f(\theta_0)\left[\lambda(K)^{-1}2^{\frac{j-1}{2}}\int_0^{\infty}r^{\frac{j-1}{2}}K(r)dr+o(1)\right][1+o(1)] \\
&= nf(\theta_0)2^{\frac{j}{2}}\kappa^{-\frac{j}{2}}\widetilde{b}_j(K)+o\left(n\kappa^{-\frac{j}{2}}\right),
\end{aligned} \tag{S11}$$

where $\widetilde{b}_j(K)$ is given by

$$\widetilde{b}_j(K)=\frac{\int_0^{\infty}r^{\frac{j-1}{2}}K(r)dr}{\int_0^{\infty}r^{-\frac{1}{2}}K(r)dr}.$$

Note that in the third equality of (S11) we have used equation (8) of the main text, the approximation $c_\kappa(K)^{-1}\sim\kappa^{-1/2}\lambda(K)$ and the fact that

$$\lim_{\kappa\to\infty}\int_0^{2\kappa}r^{\frac{j-1}{2}}\left(2-\frac{r}{\kappa}\right)^{\frac{j-1}{2}}K(r)dr=2^{\frac{j-1}{2}}\int_0^{\infty}r^{\frac{j-1}{2}}K(r)dr, \tag{S12}$$

for which the justification is analogous to the proof of Lemma 1 in García-Portugués et al. (2013) by using assumption in equation (7) of the main text.

On the other hand, when j is odd ($j-1$ is even), the right term in (S10) is given by

$$\begin{aligned}
&-\frac{1}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}}K(r)f(\theta_0-\arccos(1-r/\kappa))dr \\
&=-\frac{1}{\kappa}f(\theta_0)\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}}K(r)dr[1+o(1)],
\end{aligned}$$

which coincides with the expression in (S9) but with opposite sign. Then, we have that for an odd j,

$$\mathbb{E}(s_{n,j})=o\left(n\kappa^{-\frac{j}{2}}\right).$$

In order to have a more compact expression, for a general j we can write

$$\mathbb{E}(s_{n,j})=nf(\theta_0)2^{\frac{j}{2}}\kappa^{-\frac{j}{2}}[\widetilde{b}_j^*(K)+o(1)], \tag{S13}$$

where

$$\widetilde{b}_j^*(K)=\begin{cases}0 & \text{if } j \text{ is odd,}\\ \widetilde{b}_j(K) & \text{if } j \text{ is even.}\end{cases}$$

Now, for the variance of $s_{n,j}$ we have, by analogous computations,

$$\begin{aligned}
\mathrm{Var}(s_{n,j}) \leq & n\mathbb{E}\left[\sin^{2j}(\Theta_1-\theta_0)K_\kappa^2(\Theta_1-\theta_0)\right] \\
= & n\int_0^{2\pi}\sin^{2j}(\alpha-\theta_0)K_\kappa^2(\alpha-\theta_0)f(\alpha)d\alpha \\
= & nc_\kappa^2(K)\int_0^{2\pi}\sin^{2j}(\varphi)K^2[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi \\
= & nc_\kappa^2(K)\frac{1}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{2j-1}{2}}K^2(r)f(\theta_0+\arccos(1-r/k))dr \\
& - nc_\kappa^2(K)\frac{1}{\kappa}\int_0^{2\kappa}\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{2j-1}K^2(r)f(\theta_0-\arccos(1-r/k))dr \\
= & O\left(n\kappa^{-\frac{2j-1}{2}}\right).
\end{aligned} \tag{S14}$$

Note that we have used equation (8) of the main text, the approximation $c_\kappa(K)^{-1} \sim \kappa^{-1/2}\lambda(K)$ and the assumption in equation (7) of the main text. Finally, putting (S13) and (S14) together, we obtain

$$\begin{aligned}
s_{n,j} = & nf(\theta_0)2^{\frac{j}{2}}\kappa^{-\frac{j}{2}}\left[\widetilde{b}_j^*(K)+o(1)\right]+O_P\left(n^{\frac{1}{2}}\kappa^{-\frac{2j-1}{4}}\right) \\
= & nf(\theta_0)2^{\frac{j}{2}}\kappa^{-\frac{j}{2}}\left[\widetilde{b}_j^*(K)+o(1)+O_P\left(\frac{\kappa^{1/4}}{n^{1/2}}\right)\right] \\
= & nf(\theta_0)2^{\frac{j}{2}}\kappa^{-\frac{j}{2}}[\widetilde{b}_j^*(K)+o_P(1)].
\end{aligned}$$

Statement b)

By using the same changes of variables as in the previous calculations, we have that the expectation of $\gamma_{n,j}$ can be expressed as

$$\begin{aligned}
\mathbb{E}(\gamma_{n,j}) = & n\int_0^{2\pi}\sin^{j}(\alpha-\theta_0)K_\kappa^2(\alpha-\theta_0)f(\alpha)d\alpha \\
= & nc_\kappa^2(K)\int_0^{2\pi}\sin^{j}(\varphi)K^2[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi \\
= & nc_\kappa^2(K)\int_0^{\pi}\sin^{j}(\varphi)K^2[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi \\
& + nc_\kappa^2(K)\int_\pi^{2\pi}\sin^{j}(\varphi)K^2[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi \\
= & n\frac{c_\kappa^2(K)}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}}K^2(r)f(\theta_0+\arccos(1-r/\kappa))dr \\
& - n\frac{c_\kappa^2(K)}{\kappa}\int_0^{2\kappa}\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{j-1}K^2(r)f(\theta_0-\arccos(1-r/\kappa))dr.
\end{aligned}$$

Now, if j is even, by applying the expansion in (S2), the expectation of $\gamma_{n,j}$ is given by

$$\begin{aligned}\mathbb{E}(\gamma_{n,j}) =& 2nf(\theta_0)\frac{c_\kappa^2(K)}{\kappa^{\frac{j+1}{2}}}\int_0^{2\kappa} r^{\frac{j-1}{2}}\left(2-\frac{r}{\kappa}\right)^{\frac{j-1}{2}} K^2(r)dr[1+o(1)] \\ =& 2nf(\theta_0)\kappa^{-\frac{j-1}{2}}\left[\lambda(K)^{-2}2^{\frac{j-1}{2}}\int_0^{\infty} r^{\frac{j-1}{2}}K^2(r)dr + o(1)\right][1+o(1)] \\ =& nf(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}\widetilde{d}_j(K) + o\left(n\kappa^{-\frac{j-1}{2}}\right),\end{aligned} \tag{S15}$$

where

$$\widetilde{d}_j(K) = \frac{\int_0^\infty r^{\frac{j-1}{2}}K^2(r)dr}{\left(\int_0^\infty r^{-\frac{1}{2}}K(r)dr\right)^2}.$$

Note that in the second equality of (S15) we have used equation (8) of the main text and (S12), and the quantities $\widetilde{d}_j(K)$ exist due to assumption (7) in the main text.

On the other hand, when j is odd we have that, by analogous derivations,

$$\mathbb{E}(\gamma_{n,j}) = n\frac{c_\kappa^2(K)}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{j-1}{2}} K^2(r)dro(1) = o\left(n\kappa^{-\frac{j-1}{2}}\right).$$

Then, for a general j, we can write

$$\mathbb{E}(\gamma_{n,j}) = nf(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}[\widetilde{d}_j^*(K)+o(1)], \quad \text{with} \quad \widetilde{d}_j^*(K) = \begin{cases} 0 & \text{if } j \text{ is odd} \\ \widetilde{d}_j(K) & \text{if } j \text{ is even.}\end{cases}$$

Further, with analogous derivations, the variance of $\gamma_{n,j}$ can be expressed as

$$\begin{aligned}\mathrm{Var}(\gamma_{n,j}) \le& \mathbb{E}\left[\sin^{2j}(\Theta_1-\theta_0)K_\kappa^4(\Theta_1-\theta_0)\right] \\ =& nc_\kappa^4(K)\int_0^{2\pi}\sin^{2j}(\alpha-\theta_0)K^4[\kappa(1-\cos(\alpha-\theta_0))]f(\alpha)d\alpha \\ =& nc_\kappa^4(K)\int_0^{2\pi}\sin^{2j}(\varphi)K^4[\kappa(1-\cos\varphi)]f(\theta_0+\varphi)d\varphi \\ =& n\frac{c_\kappa^4(K)}{\kappa}\int_0^{2\kappa}\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{2j-1}{2}} K^4(r)f(\theta_0+\arccos(1-r/\kappa))dr \\ & - n\frac{c_\kappa^4(K)}{\kappa}\int_0^{2\kappa}\left[-\left(\frac{2r}{\kappa}-\frac{r^2}{\kappa^2}\right)^{\frac{1}{2}}\right]^{2j-1} K^4(r)f(\theta_0-\arccos(1-r/\kappa))dr \\ =& O\left(n\kappa^{-\frac{2j-3}{2}}\right).\end{aligned}$$

Lastly, we have

$$\begin{aligned}\gamma_{n,j} &= \mathbb{E}(\gamma_{n,j}) + O_P\left(\sqrt{\mathrm{Var}(\gamma_{n,j})}\right) \\ &= nf(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}[\widetilde{d}_j^*(K) + o(1)] + O_P\left(n^{\frac{1}{2}}\kappa^{-\frac{2j-3}{4}}\right) \\ &= nf(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}\left[\widetilde{d}_j^*(K) + o(1) + O_P\left(\frac{\kappa^{\frac{1}{4}}}{n^{\frac{1}{2}}}\right)\right] \\ &= nf(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}\left[\widetilde{d}_j^*(K) + o_P(1)\right].\end{aligned}$$

□

S2.3 Derivation of the optimal smoothing parameter in the least-squares case

In this section we derive the expression of the optimal smoothing parameter in the least-squares case, given in equation (19) of the main text. We start by computing the bias of estimator in equation (16) of the main manuscript, which is given by

$$\mathrm{Bias}[\hat{\boldsymbol{\beta}}|\Theta_1,\ldots,\Theta_n] = (\boldsymbol{\Theta}^\top \boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top \boldsymbol{W}\boldsymbol{r},$$

with

$$\boldsymbol{r} = \left[\beta_{p+1}\sin^{p+1}(\Theta_i - \theta_0) + o_P\left(\sin^{p+1}(\Theta_i - \theta_0)\right)\right]_{i=1,\ldots,n}.$$

Then, we have

$$\mathrm{Bias}[\hat{\boldsymbol{\beta}}|\Theta_1,\ldots,\Theta_n] = \boldsymbol{S}_n^{-1}\left[\beta_{p+1}\boldsymbol{c}_n + o_P\left(n\kappa^{-\frac{p+1}{2}}\right)\right],$$

where $\boldsymbol{c}_n = (s_{n,p+1},\ldots,s_{n,2p+1})^\top$. Now, by using the expression of $\boldsymbol{S}_n$ in equation (27) of the main text and Lemma 1, we have

$$\mathrm{Bias}[\hat{\boldsymbol{\beta}}|\Theta_1,\ldots,\Theta_n] = \beta_{p+1}\boldsymbol{L}^{-1}\boldsymbol{B}^{-1}\boldsymbol{c}_p 2^{\frac{p+1}{2}}\kappa^{-\frac{p+1}{2}}[1 + o_P(1)].$$

Thus, since $\hat{g}^{(\nu)}(\theta_0) = \nu!\boldsymbol{e}_{\nu+1}^\top\hat{\boldsymbol{\beta}}$, we know

$$\mathrm{Bias}[\hat{g}^{(\nu)}(\theta_0)|\Theta_1,\ldots,\Theta_n] = \nu!\beta_{p+1}\boldsymbol{e}_{\nu+1}^\top\boldsymbol{B}^{-1}\boldsymbol{c}_p 2^{\frac{p+1-\nu}{2}}\kappa^{-\frac{p+1-\nu}{2}} + o_P\left(\kappa^{-\frac{p+1-\nu}{2}}\right).$$

On the other hand, the variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}|\Theta_1,\ldots,\Theta_n] = (\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^\top\boldsymbol{W}\mathrm{Var}(\boldsymbol{Y})\boldsymbol{W}\boldsymbol{\Theta}(\boldsymbol{\Theta}^\top\boldsymbol{W}\boldsymbol{\Theta})^{-1} = \boldsymbol{S}_n^{-1}\boldsymbol{\Theta}^\top\boldsymbol{\Sigma}\boldsymbol{\Theta}\boldsymbol{S}_n^{-1},$$

where $\boldsymbol{\Sigma} = \mathrm{diag}\{K_\kappa^2(\Theta_i - \theta_0)\sigma^2(\Theta_i)\}$. Note that the $(i,j)th$ element of $\boldsymbol{\Theta}^\top\boldsymbol{\Sigma}\boldsymbol{\Theta}$ is given by $\delta_{n,i+j-2}$ where

$$\delta_{n,j} = \sum_{i=0}^{n}\sin^j(\Theta_i - \theta_0)K_\kappa^2(\Theta_i - \theta_0)\sigma^2(\Theta_i).$$

Using arguments analogue to the proof of statement b) in Lemma 1, we can write

$$\delta_{n,j} = nf(\theta_0)\sigma^2(\theta_0)2^{\frac{j-1}{2}}\kappa^{-\frac{j-1}{2}}[d_j^*(K) + o_P(1)]$$

and, thus,

$$\boldsymbol{\Theta}^\top \boldsymbol{\Sigma} \boldsymbol{\Theta} = nf(\theta_0)\sigma^2(\theta_0)2^{-\frac{1}{2}}\kappa^{\frac{1}{2}}[\boldsymbol{L}\boldsymbol{D}\boldsymbol{L} + o_P(\boldsymbol{L}\boldsymbol{1}\boldsymbol{L})].$$

By using the previous equation in addition to the expression of $\boldsymbol{S}_n$ in equation (27) of the main text, we have

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}|\Theta_1, \ldots, \Theta_n] = \frac{\sigma^2(\theta_0)2^{-\frac{1}{2}}\kappa^{\frac{1}{2}}}{nf(\theta_0)}[\boldsymbol{L}^{-1}\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{L}^{-1} + o_P(\boldsymbol{L}^{-1}\boldsymbol{1}\boldsymbol{L}^{-1})].$$

Now, recalling that $\hat{g}^{(\nu)}(\theta_0) = \nu!\boldsymbol{e}_{\nu+1}^\top\hat{\boldsymbol{\beta}}$, we obtain

$$\begin{aligned}\mathrm{Var}[\hat{g}^{(\nu)}(\theta_0)|\Theta_1, \ldots, \Theta_n] &= \frac{\nu!^2\sigma^2(\theta_0)2^{-\frac{1+2\nu}{2}}\kappa^{\frac{1+2\nu}{2}}}{nf(\theta_0)}\boldsymbol{e}_{\nu+1}^\top\boldsymbol{L}^{-1}\boldsymbol{B}^{-1}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{L}^{-1}\boldsymbol{e}_{\nu+1} + o_P\left(n^{-1}\kappa^{\frac{1+2\nu}{2}}\right) \\ &= \frac{\nu!^2\sigma^2(\theta_0)2^{-\frac{1+2\nu}{2}}\kappa^{\frac{1+2\nu}{2}}}{nf(\theta_0)}a_\nu + o_P\left(n^{-1}\kappa^{\frac{1+2\nu}{2}}\right).\end{aligned}$$

Consequently, the Mean Squared Error of the estimator in equation (12) of the main text can be expressed as

$$\begin{aligned}\mathrm{MSE}[\hat{g}^{(\nu)}(\theta_0)|\Theta_1, \ldots, \Theta_n] &= \nu!^2\beta_{p+1}^2[\boldsymbol{e}_{\nu+1}^\top\boldsymbol{B}^{-1}\boldsymbol{c}_p]^2 2^{p+1-\nu}\kappa^{-(p+1-\nu)} \\ &\quad + \frac{\nu!^2\sigma^2(\theta_0)2^{-\frac{1+2\nu}{2}}\kappa^{\frac{1+2\nu}{2}}}{nf(\theta_0)}a_\nu + o_P\left(\kappa^{-\frac{p+1-\nu}{2}} + n^{-1}\kappa^{\frac{1+2\nu}{2}}\right).\end{aligned}$$

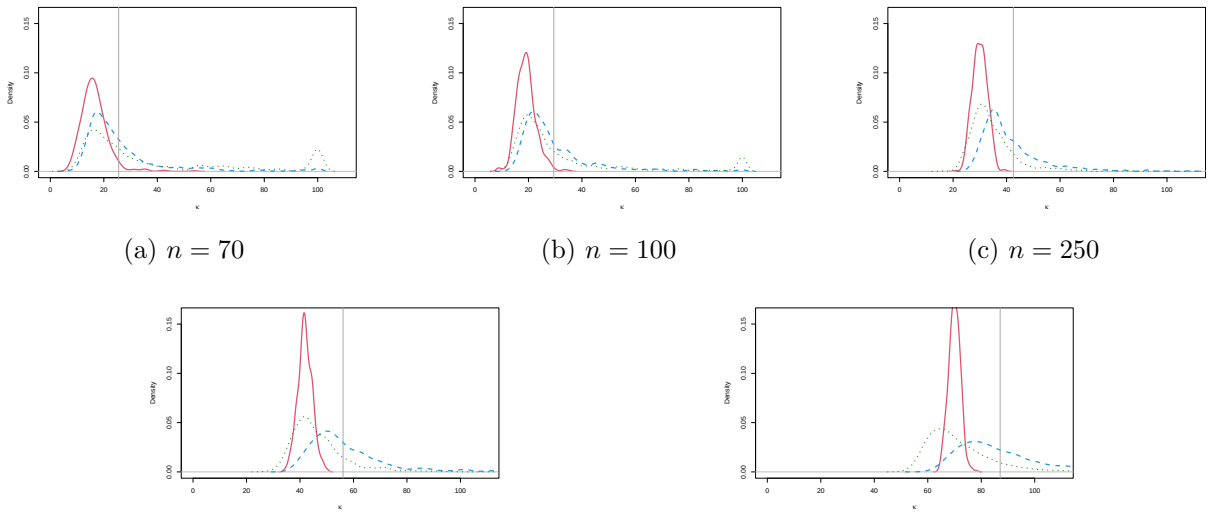
Finally, the concentration which minimizes the asymptotic version of the MSE is given by the expression in equation (19) of the main manuscript.

S3 Additional simulation results

In this section, additional material supporting the findings of the simulation study in the main paper is presented. Four cases are distinguished, where the conditional likelihood is given by the normal, Bernoulli, Poisson and gamma distributions. The code for all methods can be found at `https://anonymous.4open.science/r/CircLocalLikelihood-2424` and also as supplementary material.

Normal likelihood. First, we focus on the concentration parameters selected by each method. Figures S1 and S2 show kernel density estimators of the selected smoothing parameters when using the refined rule, the CRSC criterion and the cross-validation method in models N1 and N2, for the different sample sizes. The optimal concentration parameters

given by equation (19) of the main text were also computed and represented as a vertical line. We observe that, for model N1, the parameters selected by the refined rule are usually smaller than those obtained by the CRSC rule or the cross-validation criterion, and also than the optimal concentration minimizing the MISE of the estimator. For model N2, the parameters obtained by the refined rule are usually larger than the ones obtained by the other two methods, and are fairly close to the optimal concentration. However, in all scenarios, the distribution of the parameters selected by the refined rule is highly peaked and symmetric, while being skewed for the other two methods. Note that sometimes the CRSC and cross-validation selectors lead to high concentrations (occasionally selecting the maximum possible value) and this behavior is not observed with the refined rule. As expected, the selected parameters are generally larger when increasing the sample size.

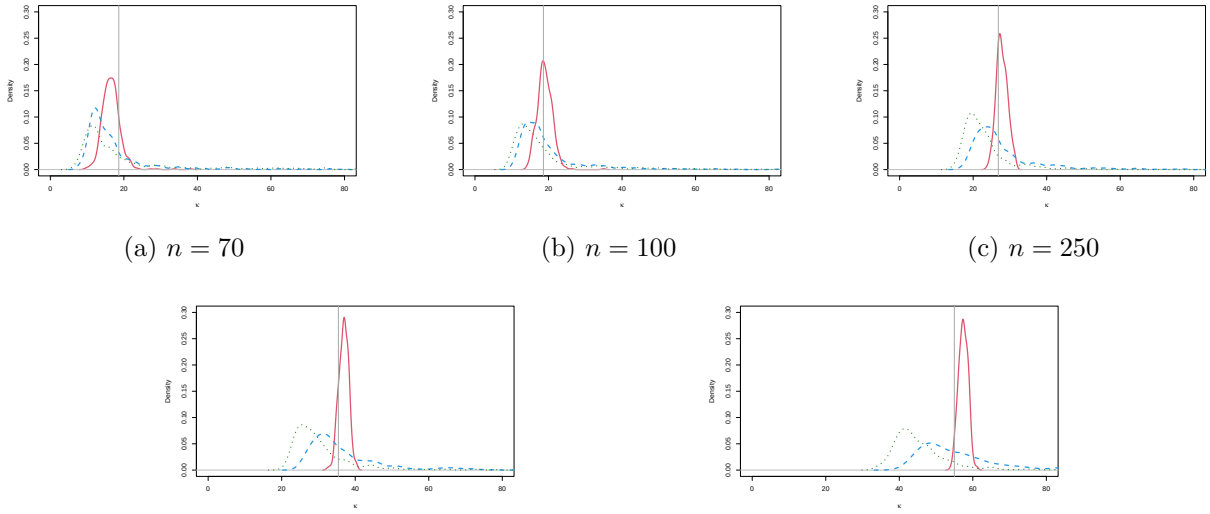


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S1: Kernel density estimators of the obtained values of κ for model N1 with therefined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line). Grey vertical line represents the optimal concentration parameters.

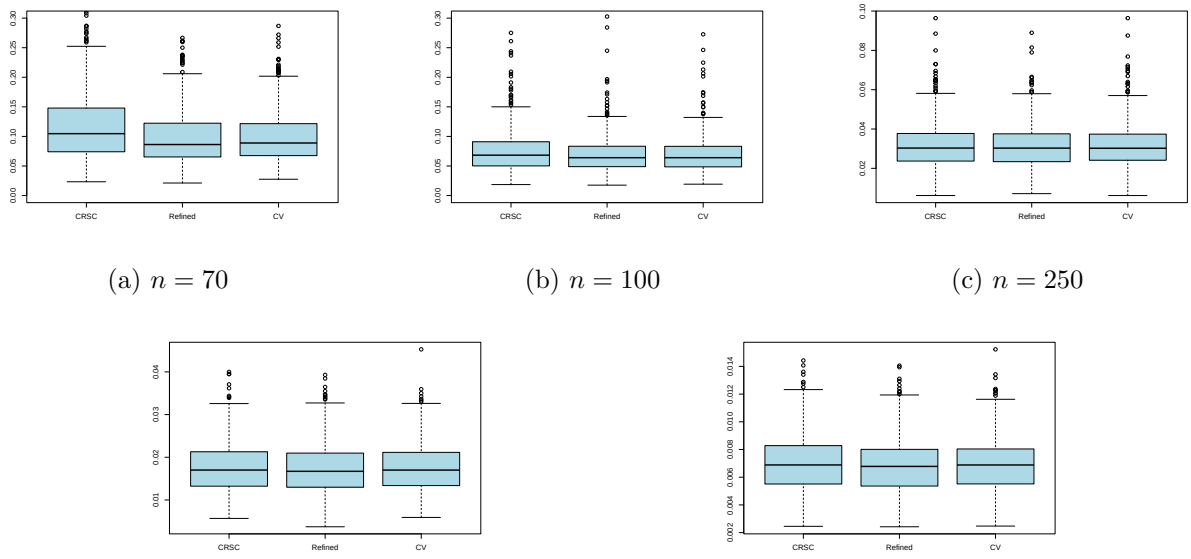
Regarding the performance of each criterion in terms of the approximated ISE, computed as in equation (22) of the main text, Figures S3 and S4 show boxplots of the approximated ISE for models N1 and N2, the different sample sizes and for each concentration selection method. It can be observed that for model N1 the distribution of the approximated ISE is similar for the three methods, while for model N2 it seems that the values of the approximated ISE are moderately lower for the refined rule for all sample sizes, while the difference is more noticeable for the lowest sample sizes.



(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S2: Kernel density estimators of the obtained values of κ for model N2 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line). Grey vertical line represents the optimal concentration parameters.

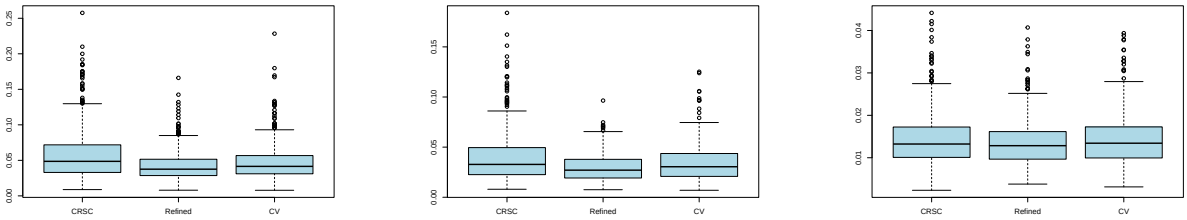


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S3: Boxplots of the estimated ISE for model N1 with the CRSC, refined rule and cross-validation.

Bernoulli likelihood. Regarding the concentration parameters selected by each method, Figures S5 and S6 show kernel density estimators of the obtained smoothing parameters.



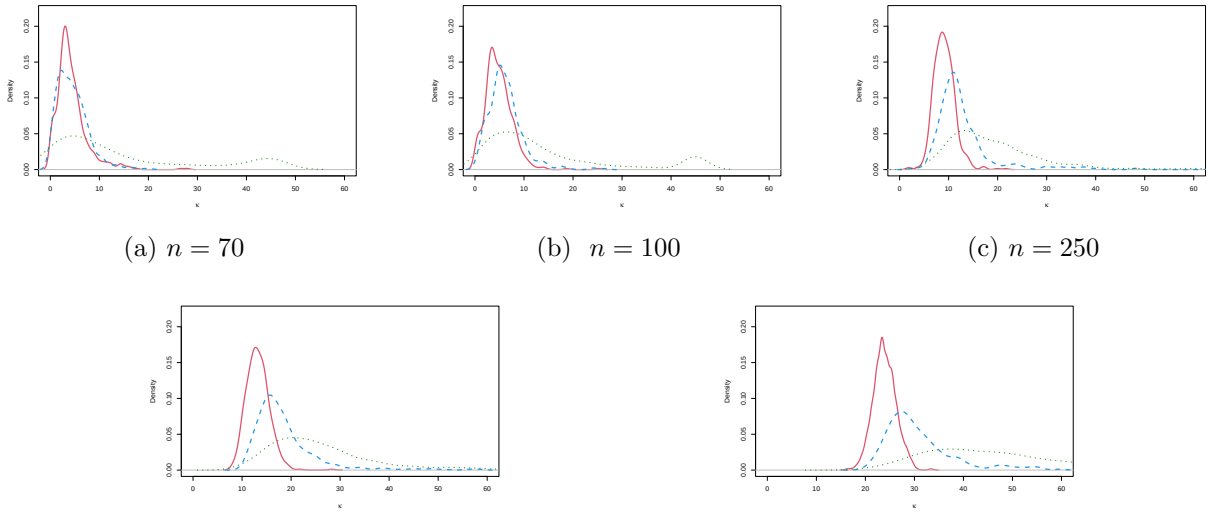
(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S4: Boxplots of the estimated ISE for model N2 with the CRSC, refined rule and cross-validation.

In this case, there is not a closed form available for the optimal concentration. For model B1, the behavior of the cross-validation method and the refined rule is quite similar for lower sample sizes, while the distribution of the parameters selected by the ECRSC is less concentrated, often selecting the maximum possible value. Surprisingly, when the sample size increases, the estimated distribution of the selected parameters is less peaked, specially for the cross-validation method. For model B2, the parameters selected by the ECRSC and refined rules are centered around the same values, while the concentrations obtained by cross-validation are usually smaller. While the latter method produces less peaked densities when increasing the sample size, the density of the parameters selected by the refined rule is more peaked for larger n.

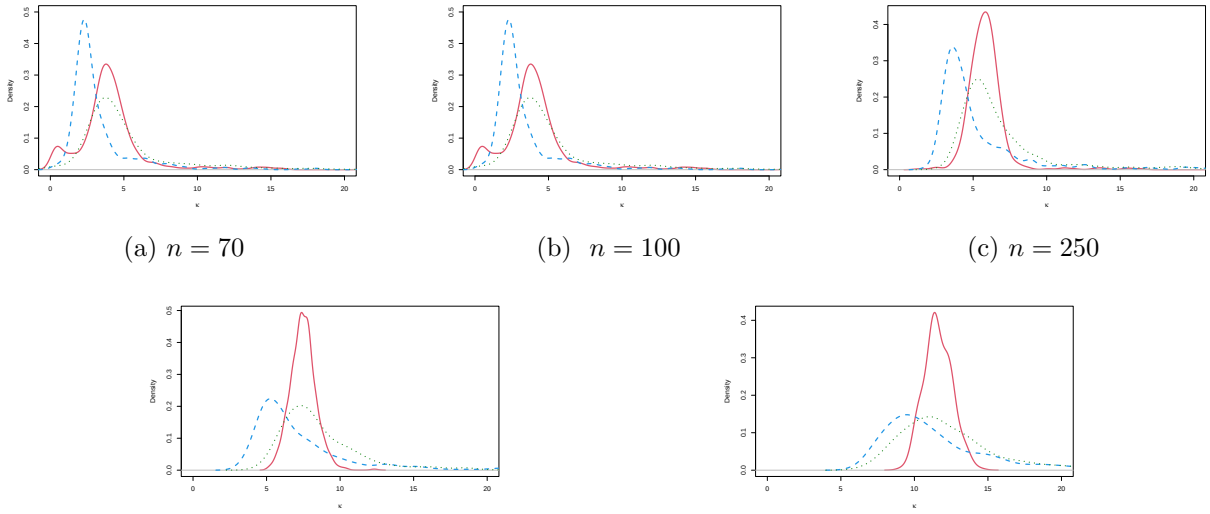
Now we move on to the performance of the three methods in terms of the approximated ISE, for which boxplots are represented in Figures S7 and S8. For model B1, it is observed that cross-validation and the refined rule obtain very similar values of the approximated ISE, specially for large sample sizes, with cross-validation slightly outperforming the other two methodswhen the sample size is small. Although the behavior is similar in model B2, the refined rule obtains smaller values of the approximated ISE when $n = 250$, $n = 500$ and $n = 1500$. In all cases, the largest values of the approximated ISE are obtained with the ECRSC method.



(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S5: Kernel density estimators of the obtained values of κ for model B1 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

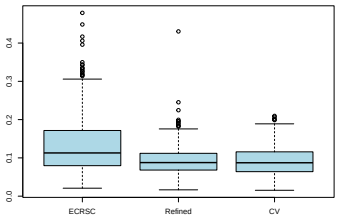


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

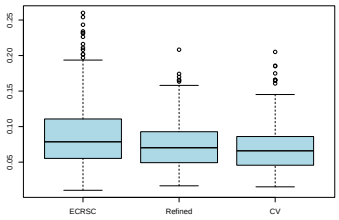
(d) $n = 500$ (e) $n = 1500$

Figure S6: Kernel density estimators of the obtained values of κ for model B2 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

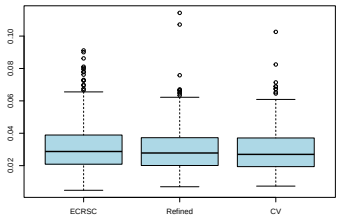
Poisson likelihood. Regarding the Poisson scenario, the kernel density estimators of the selected κ with each selection method are shown in Figures S9 and S10. The estimated



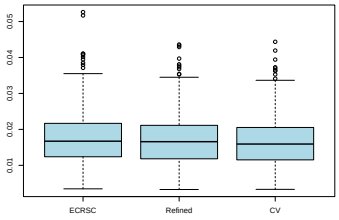
(a) $n = 70$



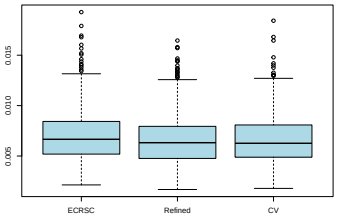
(b) $n = 100$



(c) $n = 250$

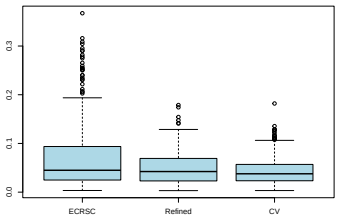


(d) $n = 500$

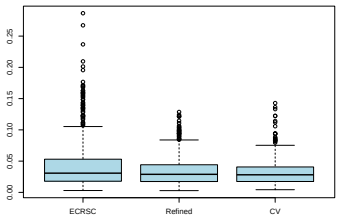


(e) $n = 1500$

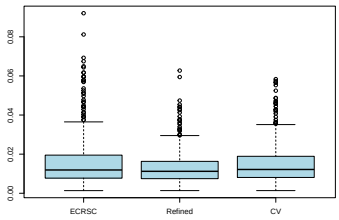
Figure S7: Boxplots of the estimated ISE for model B1 with the ECRSC, refined rule and cross-validation.



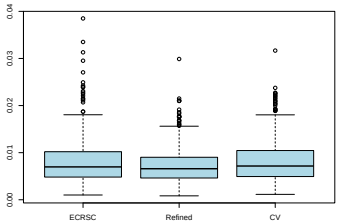
(a) $n = 70$



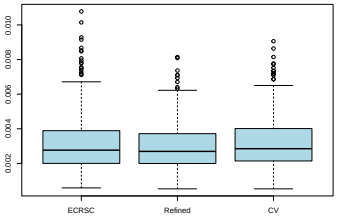
(b) $n = 100$



(c) $n = 250$



(d) $n = 500$

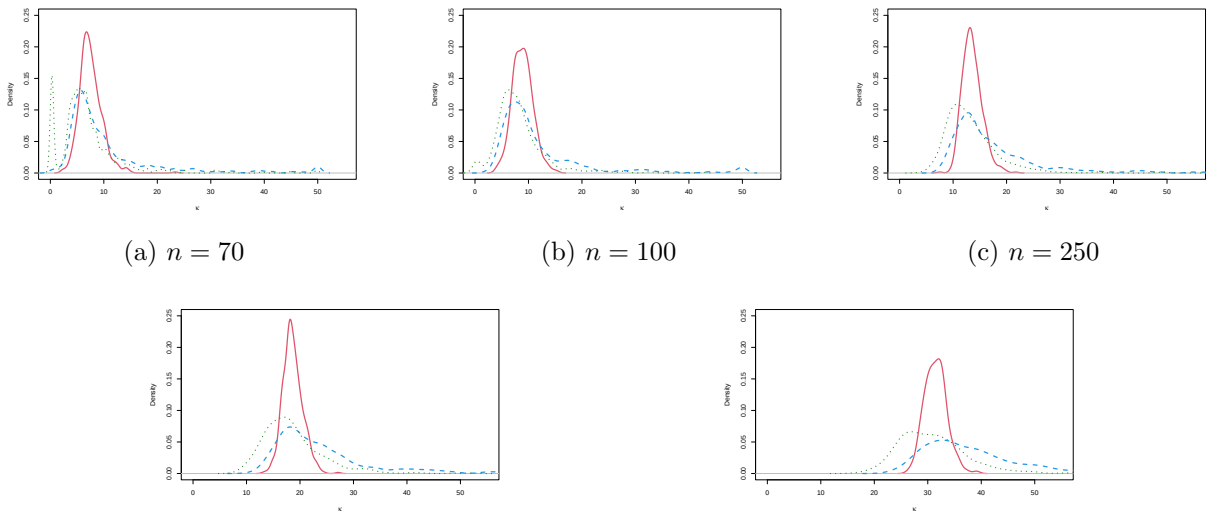


(e) $n = 1500$

Figure S8: Boxplots of the estimated ISE for model B2 with the ECRSC, refined rule and cross-validation.

distribution of the obtained parameters is similar to the ones achieved in the normal likelihood case, since the estimated density obtained for the refined rule is generally symmetric

and highly peaked, while the other two are skewed, selecting sometimes concentration parameters which are overly large.



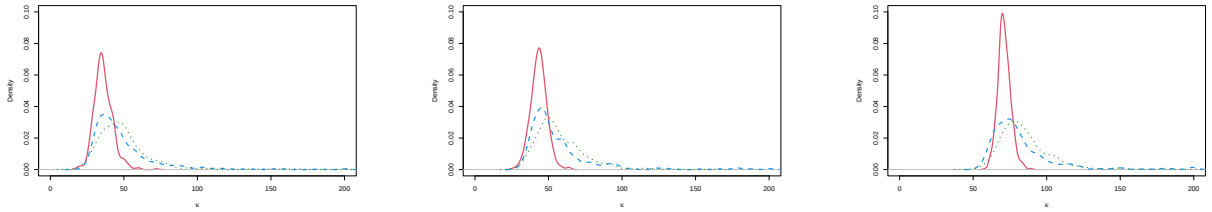
(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S9: Kernel density estimators of the obtained values of κ for model P1 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

Boxplots of the approximated ISE obtained with each method are displayed in Figures S11 and S12. For model P1, it can be observed that the approximated ISEs obtained with the refined rule are slightly smaller than the ones obtained with the ECRSC and the cross-validation method, for all sample sizes. On the other hand, for model P2, it seems that the three methods achieve a similar performance in terms of ISE, although the refined rule visibly outperfomrs the other two methods for $n = 70$.

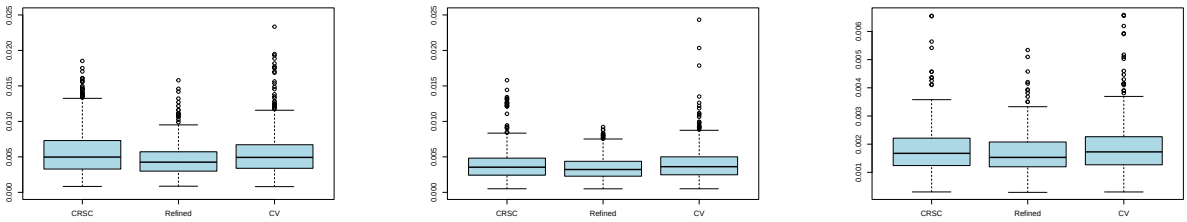
Gamma likelihood. The kernel density estimators of the selected concentration parameters with each method are shown in Figures S13 and S14. For model G1, the shape of the estimated density for the cross-validation method is more peaked than in the previous scenarios, although it is still quite asymmetric, selecting sometimes values of the smoothing parameter which are too large. On the other hand, the values of the concentration obtained by the ECRSC rule are highly variable, and even more when increasing the sample size. On the contrary, values computed with the refined rule are reasonably concentrated, and usually larger than those selected by cross-validation. Regarding results for model G2, the selection of κ seems more complicated when $n = 70$ and $n = 100$, with the ECRSC rule and the cross-validation method frequently selecting values of the concentration very



(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S10: Kernel density estimators of the obtained values of κ for model P2 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

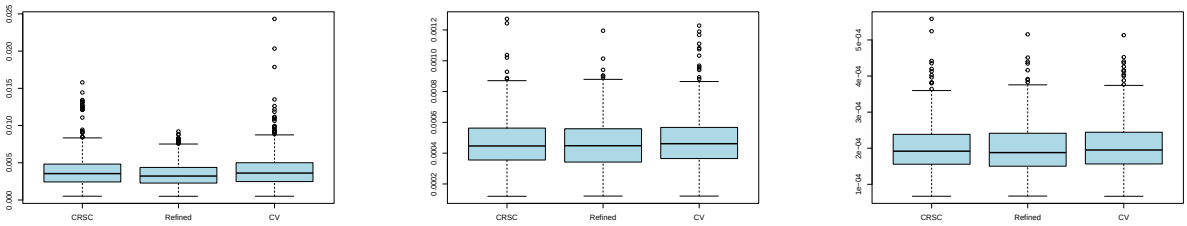


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S11: Boxplots of the approximated ISE for model P1 with the ECRSC, refined rule and cross-validation.

close to zero. Although the refined rule also leads to very small values of the concentration sometimes, this behavior is not shown as often as with the other two methods. As usual,

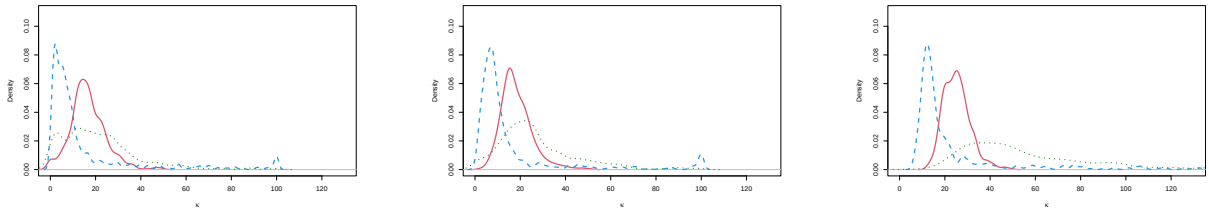


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S12: Boxplots of the approximated ISE for model P2 with the ECRSC, refined rule and cross-validation.

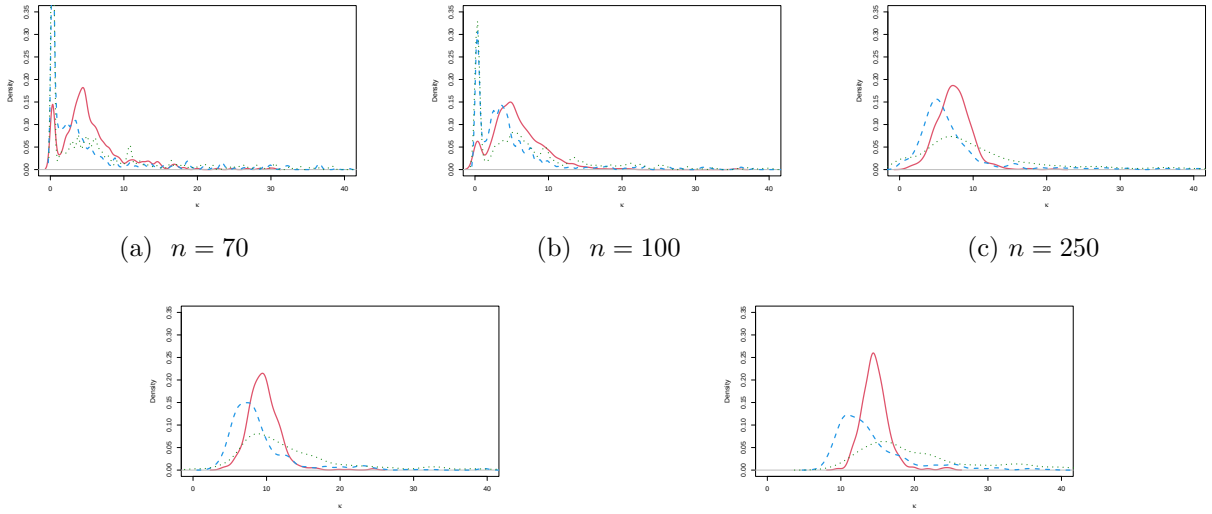
both the cross-validation method and the ECRSC rule sometimes select exceedingly large concentration parameters.



(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

(d) $n = 500$ (e) $n = 1500$

Figure S13: Kernel density estimators of the obtained values of κ for model G1 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

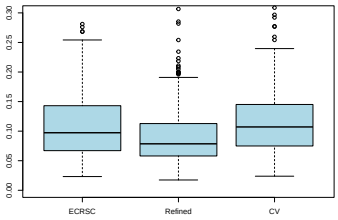


(a) $n = 70$ (b) $n = 100$ (c) $n = 250$

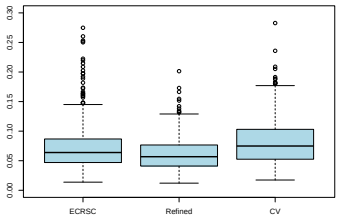
(d) $n = 500$ (e) $n = 1500$

Figure S14: Kernel density estimators of the obtained values of κ for model G2 with the refined rule (red, continuous line), ECRSC (green, dotted line) and cross-validation (blue, dashed line).

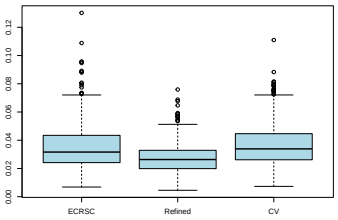
Boxplots of the approximated ISE for each method are represented in Figures S15 and S16. For both models and all sample sizes, it seems that values of the approximated ISE are usually smaller when employing the refined rule and the ECRSC.



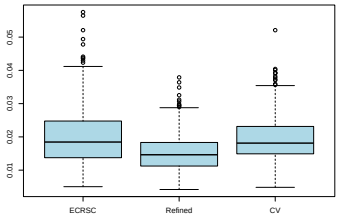
(a) $n = 70$



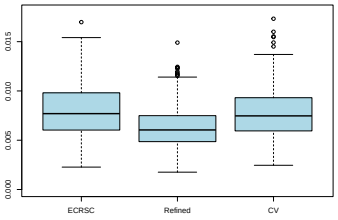
(b) $n = 100$



(c) $n = 250$

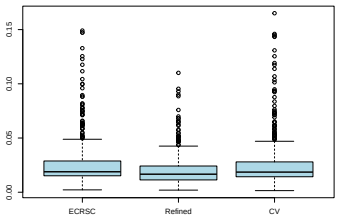


(d) $n = 500$

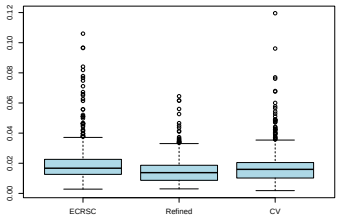


(e) $n = 1500$

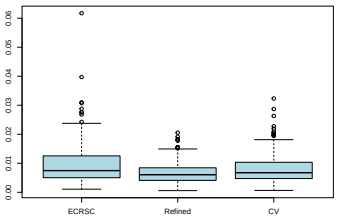
Figure S15: Boxplots of the estimated ISE for model G1 with the ECRSC, refined rule and cross-validation.



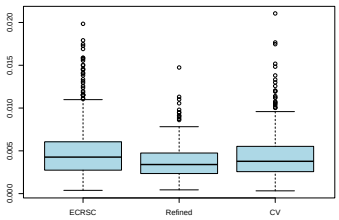
(a) $n = 70$



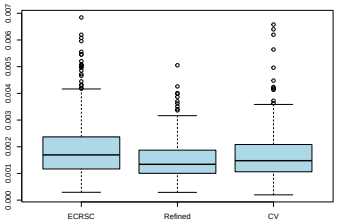
(b) $n = 100$



(c) $n = 250$



(d) $n = 500$



(e) $n = 1500$

Figure S16: Boxplots of the estimated ISE for model G2 with the ECRSC, refined rule and cross-validation.

References

Di Marzio, M., S. Fensore, and A. Panzera (2018). Nonparametric classification for circular data. In C. Ley and T. Verdebout (Eds.), *Applied Directional Statistics: Modern Methods and Case Studies*. New York: Chapman & Hall/CRC.

Di Marzio, M., A. Panzera, and C. C. Taylor (2009). Local polynomial regression for circular predictors. *Statist. Probab. Lett. 798*, 2066–2075.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasilikelihood functions. *J. Am. Statist. Ass. 90*, 141–150.

García-Portugués, E., R. M. Crujeiras, and W. González-Manteiga (2013). Kernel density estimation for directional-linear data. *J. Multivar. Anal. 121*(1), 152–275.