A reliable data-based smoothing parameter selection method for circular kernel estimation

Jose Ameijeiras-Alonso

CITMAga, Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain.

Abstract

A new data-based smoothing parameter for circular kernel density (and its derivatives) estimation is proposed. Following the plug-in ideas, unknown quantities on an optimal smoothing parameter are replaced by suitable estimates. This paper provides a circular version of the well-known Sheather and Jones bandwidths (DOI: 10.1111/j.2517-6161.1991.tb01857.x), with direct and solve-the-equation plug-in rules. Theoretical support for our developments, related to the asymptotic mean squared error of the estimator of the density, its derivatives, and its functionals, for circular data, are provided. The proposed selectors are compared with previous data-based smoothing parameters for circular kernel density estimation. This paper also contributes to the study of the optimal kernel for circular data. An illustration of the proposed plug-in rules is also shown using real data on the time of car accidents.

Keywords: Circular data; Directional Statistics; Kernel Density Estimation; Plug-in rule; Sheather and Jones bandwidth.

1 Introduction

Circular data are observations that can be represented on the unit circumference and where periodicity must be taken into account. Classic examples appear when the goal is to model orientations or a periodic phenomenon with a known period. Several applications of circular data can be found, e.g., in Ley and Verdebout (2018) or SenGupta and Arnold (2022). The complicated features that circular data can exhibit on real data applications lead to several new flexible models in the statistical literature. A recent review of flexible parametric circular distributions can be found in Ameijeiras-Alonso and Crujeiras (2022).

1Supported by Grant 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. The author is most grateful to Rosa M. Crujeiras for helpful suggestions and comments on an earlier draft.
When trying to obtain a flexible fit of the density function, an alternative to parametric models is the kernel density estimation. Kernel density estimation for circular data dates back to Beran (1979) and Hall et al. (1987), while the estimator of the density derivatives was studied by Klemelä (2000) and Di Marzio et al. (2011). It is well known that the choice of the smoothing parameter is critical, when using these kernel methods.

In the usual linear inferential framework, where random variables are supported on the Euclidean space, one can find many approaches for selecting the “best” data-driven bandwidth parameter (see, e.g., Jones et al., 1996, for a discussion on this topic). Due to its good performance, one of the most-employed bandwidth selectors is the plug-in bandwidth proposed by Sheather and Jones (1991). The relevance of this plug-in selector is evident from the impressive number of citations that Sheather and Jones (1991) have received, and although other authors have introduced new bandwidth selectors, none of the proposals outperforms, in general, their plug-in selector.

There exists some literature on smoothing parameter selection for circular data, some of these ideas are based on cross-validation (Hall et al., 1987), rule-of-thumb (Taylor, 2008), adaptive-mixture of von Mises (Oliveira et al., 2012), or bootstrap (Di Marzio et al., 2011) techniques. But none of the approaches introduced so far presents an outstanding performance with respect to its competitors, as is the case with the proposal by Sheather and Jones (1991) in the linear case. Hence, the goal of this paper is to provide the needed theory to derive an algorithm that replicates the idea of the two-stage direct and solve-the-equation plug-in bandwidth selectors for circular kernel density estimation. In addition, the developed theory can be also employed when estimating the density derivatives.

Regarding the kernel choice, most of the results for circular density estimation fix the kernel to be the von Mises density function. In this paper, we study the asymptotic results for a more general class of kernels. The developed results allow us to obtain the optimal kernels in the circular estimation context.

This paper is organized as follows. Section 2 is devoted to the definition of the circular kernel density derivative estimate. Also, the key function needed to derive the optimal smoothing parameter is introduced in this section. For a general circular kernel, the asymptotic mean integrated squared error and the optimal smoothing parameter of the derivative estimator are derived in Section 3. Section 4 is devoted to the estimation of density functionals, needed to derive the plug-in rules. In Section 5, we discuss the kernel choice. Section 6 provides all the needed details to compute the two-stage direct and solve-the-equation plug-in smoothing selectors. A simulation study showing that the proposed rules provide a competitive smoothing parameter is given in Section 7. Section 8 uses a real data example to show the applicability of the proposed selector. Some final remarks are provided in Section 9. Section 10 describes the software that implements the proposed smoothing parameters. The proofs of the theoretical results appear in Appendices A and B.
2 Circular kernel density derivative estimation

Given a circular random sample in angles $\Theta_1, \ldots, \Theta_n \in [-\pi, \pi)$, with associated density function $f$, the circular kernel density estimation (see, for example, Oliveira et al., 2012) can be defined as follows

$$\hat{f}_\nu(\theta) = \frac{1}{n} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i), \quad (2.1)$$

where $K_\nu$ is the circular kernel with smoothing parameter $\nu \in [0, 1]$, denoting by $\nu$ the mean resultant length (differently from the previous literature, where $\nu$ usually stands for a concentration parameter). This allows us to establish a theory that will be valid for both, kernels depending on the concentration, such as the von Mises, or on the mean resultant length, such as the wrapped normal (see Section 5 for a formal introduction of these kernels).

The estimator (2.1) can be easily extended, when the objective is estimating the $r$–th derivative of $f$, denoted by $f^{(r)}$. In that case, the estimator of $f^{(r)}$ can be defined as (Di Marzio et al., 2011),

$$\hat{f}^{(r)}_\nu(\theta) = \frac{1}{n} \sum_{i=1}^{n} K^{(r)}_\nu(\theta - \Theta_i). \quad (2.2)$$

The first question is which function can be employed as a kernel? In this case, we will assume that $K$ satisfies the standard conditions that one can find for the canonical linear kernel (Wand and Jones, 1995, Ch. 2). In particular, $K$ will be a circular density, (reflectively) symmetric about zero, unimodal and square-integrable in $[-\pi, \pi)$. Note that these last conditions ensure that the kernel $K$ has the following convergent Fourier series representation (see Mardia and Jupp, 1999, Section 4.2),

$$K_\nu(\theta) = \frac{1}{2\pi} \left( 1 + 2 \sum_{j=1}^{\infty} \alpha_{K,j}(\nu) \cos(j\theta) \right), \quad (2.3)$$

where only the values of $\alpha_{K,j}(\nu) \in [0, 1]$, for $j \in \{1, 2, \ldots\}$, depend on the employed kernel and the smoothing parameter $\nu$.

Secondly, a crucial element in kernel density estimation is the smoothing parameter. Generally, this parameter is taken as a non-random sequence, depending on the sample size, and then as a fixed value for a sample realization. In this work, we will introduce the function $h$ depending on $\nu$,

$$h \equiv h_\nu(\nu) = \frac{\pi^2}{3} + 4 \sum_{j=1}^{\infty} \frac{(-1)^j \alpha_{K,j}(\nu)}{j^2}. \quad (2.4)$$

Then, $h$ can be seen as the bandwidth and the sequence of numbers $h_n \equiv h_\nu(\nu_n)$ must satisfy the following condition $\lim_{n \to \infty} h_n = 0$. 

3
3 Asymptotic results

In this section, we establish the results needed for deriving the asymptotic mean integrated squared error (AMISE) and the AMISE-optimal smoothing parameter. Throughout this section, for a given derivative order \( r \), we will employ the following assumptions to derive the asymptotic results.

(A1) The circular density \( f \) is such that its derivative \( f^{(r+2)} \) is continuous and square-integrable in \([-\pi, \pi)\).

(A2) The kernel \( K \) is a bounded circular density, (reflectively) symmetric about zero, unimodal and its \( r \)-th derivative is square integrable in \([-\pi, \pi)\).

(A3) As \( n \to \infty \), the value \( h_K(\nu_n) = 0 \), \( R_{K;r,2}(\nu_n) = \infty \), and \( n^{-1} R_{K;r,2}(\nu_n) = 0 \). The function \( R_{K;r,t} \), for \( t \in \{1, 2\} \), is defined as,

\[
R_{K;r,t}(\nu) = \begin{cases} 
2(2\pi)^{-1} \left( 1 + 2 \sum_{j=1}^{\infty} \alpha_{K,j}^t(\nu) \right) & \text{if } r = 0, \\
5\pi^{-1} \sum_{j=1}^{\infty} j^{tr} \alpha_{K,j}^t(\nu) & \text{otherwise,} \\
s & \text{if } t = 1 \text{ and } r \mod 4 = 1, \\
-s & \text{if } t = 1 \text{ and } r \mod 4 = 2, \\
1 & \text{otherwise.} 
\end{cases}
\]

Assumptions (A1) and (A2) coincide with those employed in the standard linear case. For a large class of kernels, including the von Mises or the wrapped normal (see Section 5), we will see that Assumption (A3) translates into the standard conditions on the bandwidth, namely, \( h_n = 0 \) and \( nh_n^{(2r+1)/2} = \infty \). The result in Theorem 1 (see Section A of the Appendix for a formal proof) states the AMISE order of the kernel derivative estimator (2.2).

If we also assume the following two extra conditions, we can derive an explicit expression of the AMISE.

(E1) \( \lim_{n \to \infty} \int_{-\pi}^{\pi} \theta^4 K_{\nu_n}(\theta) d\theta = o(h_n) \).

(E2) \( \lim_{n \to \infty} \int_{-\pi}^{\pi} \theta^2 (K_{\nu_n}^{(r)}(\theta))^2 d\theta = o[R_{K;r,2}(\nu_n)] \).

Theorem 1. Under the Assumptions (A1)–(A3), we have

\[
\text{AMISE} \left[ \hat{f}_\nu^{(r)} \right] = O \left( h_K^2(\nu_n) \right) + O \left( n^{-1} R_{K;r,2}(\nu_n) \right)
\]

If we also assume Conditions (E1) and (E2), then the AMISE has the following explicit expression.

\[
\text{AMISE} \left[ \hat{f}_\nu^{(r)} \right] = \frac{1}{4} h_K^2(\nu_n) \int_{-\pi}^{\pi} \left( f^{(r+2)}(\theta) \right)^2 d\theta + \frac{1}{n} R_{K;r,2}(\nu_n)
\]

(3.1)

Theorem 1 states the asymptotic expression of the MISE which depend on the sample size \( n \), the derivative of the density function \( f^{(r+2)} \), the kernel \( K \), and the smoothing parameter \( \nu \). The complete expression of the AMSE is provided in Section A of the Appendix.
Remark 1. Di Marzio et al. (2009, 2011) also analysed the AMISE of the circular kernel estimation obtaining a similar result as in (3.1), replacing $h_K(\nu_n)$ by $(1 - \alpha_{K,2}(\nu_n))/2$. Note that, although not stated there, Condition (E1) must be imposed to derive their asymptotic results. This can be seen in the remainder term of the asymptotic bias in the proof of Theorem 1 in Di Marzio et al. (2009). Assuming Condition (E1), we obtain that $h_K(\nu_n)$ can be approximated by $(1 - \alpha_{K,2}(\nu_n))/2$. Thus, asymptotically, both results coincide, but the expression provided in this paper will allow us to derive an expression of the optimal concentration.

The issue, under the general AMISE expression given in (3.1), is that it is not straightforward to know how to derive an explicit expression of the optimal smoothing parameter, unless a specific kernel, such as the von Mises is chosen (see, e.g., Di Marzio et al., 2011). This problem can be solved when the following extra condition is assumed.

(E3) As $n$ increases, $R_{K;\nu,2}(\nu_n) = Q_{K;\nu,2}h_n^{-(2r+1)/2}$, where $Q_{K;\nu,2}$ is a constant only depending on the kernel $K$ and $r$.

Note that under the Condition (E3), the Assumption (A3) simplifies to $h_n = 0$ and $nh_n^{(2r+1)/2} = \infty$. Using this last assumption, we can observe the classic variance-bias trade-off, where the bias is reduced if $h_n$ is “small” and the variance decreases if $h_n$ is “large”. The optimal smoothing parameter with respect to the AMISE criteria can be obtained using the following corollary of Theorem 1, which is a direct consequence of $h_K(\nu) > 0$ (see Section A of the Appendix). Note that, in the following corollary, we obtain that, under the previous assumptions, the AMISE order coincides with the one obtained in the linear case, $O(n^{-4/(2r+5)})$ (Wand and Jones, 1995, Section 2.5).

Corollary 1. Consider the Assumptions (A1)–(A3) and (E1)–(E3). Then, for the kernel derivative estimator of order $r$ (see (2.2)), we have that, asymptotically, the optimal (AMISE) value of $h_n$ can be obtained from,

$$h_{K;r;\text{AMISE}} = \left( \frac{(2r + 1)Q_{K;r,2}}{n \int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta} \right)^{2/(2r+5)}. \quad (3.2)$$

Under the previous assumptions, the minimal AMISE of $\hat{f}_{\nu}^{(r)}$ is equal to

$$\inf_{0 \leq \nu < 1} \text{AMISE} \left[ \hat{f}_{\nu}^{(r)} \right] = \frac{2r + 5}{8r + 4} \left( \frac{(2r + 1)Q_{K;r,2}}{n \int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta} \right)^{4/(2r+5)} \int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta. \quad (3.3)$$

The optimal concentration $\nu$ is the solution to the equation $h_K(\nu) = h_{K;r;\text{AMISE}}$, see Equation (2.4). Alternatively, to avoid the infinite sum in (2.4), one can also obtain the optimal $\nu$ by solving the equation $\alpha_{K,2}(\nu) = 1 - 2h_{K;r;\text{AMISE}}$ (see Remark 1).

As usual, when using data-based plug-in smoothing parameters, the main problem with employing the optimal $h_{K;r;\text{AMISE}}$ in Corollary 1, is that its value depends on the unknown value of $\int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta$. 

5
A rule-of-thumb smoothing selector can be obtained by replacing \( f \) with a simple and standard density function, such as the von Mises density. One can also follow the Ćwik and Koronacki (1997) approach and replace \( f \) with a mixture model, such as the mixture of von Mises. Both techniques were already proposed and studied in the circular literature when employing the von Mises density as the kernel \( K \) (see Taylor, 2008; Oliveira et al., 2012). Following Sheather and Jones (1991), an alternative is to estimate density functionals related to \( \int_{-\pi}^{\pi} (f^{(s)}(\theta))^2 d\theta \). In the next section, we study how to obtain a kernel estimator of this last quantity, in the circular context.

### 4 Estimation of density functionals

As mentioned in the previous section, an issue with using, in practice, the optimal smoothing parameter (3.2) is its dependence on the unknown quantity \( \int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta \). In this section, we will see how to estimate this last quantity using kernel techniques. For doing so, we first define the functional of the form

\[
\psi_s = \int_{-\pi}^{\pi} f^{(s)}(\theta)f(\theta) d\theta.
\]

Note that under sufficient smoothness assumptions on \( f \) (e.g., the needed conditions to apply integration by parts), we obtain that,

\[
\int_{-\pi}^{\pi} (f^{(s)}(\theta))^2 d\theta = (-1)^s \psi_{2s}.
\]  \((4.1)\)

Since \( \psi_s = E(f^{(s)}(\Theta)) \), the following estimator can be employed to estimate the unknown quantity \( \int_{-\pi}^{\pi} (f^{(r+2)}(\theta))^2 d\theta \) on the optimal smoothing parameter (3.2),

\[
\hat{\psi}_{s;\rho} = \frac{1}{n} \sum_{i=1}^{n}  \hat{f}_\rho^{(s)}(\Theta_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L^{(s)}_{\rho}(\Theta_i - \Theta_j),
\]  \((4.2)\)

where \( L \) and \( \rho \) are a kernel and a concentration parameter, which are possibly different from \( K \) and \( \nu \).

Using the estimator (4.2), a direct plug-in estimator of the smoothing parameter can be obtained from (3.2), replacing the quantity depending on the true \( f \), by its estimator, in the following way,

\[
h_{K;r;\pi;\rho} = \left( \frac{(2r+1)Q_{K;r,2}}{n(-1)^{r+2}\psi_{2r+4;\rho}} \right)^{2/(2r+5)}.
\]  \((4.3)\)

The problem with the direct plug-in estimator (4.3), is that it still depends on a choice of the pilot smoothing parameter \( \rho \). Below, we establish the asymptotic theory to derive the optimal smoothing
parameter of $\hat{\psi}_{2r+4;\rho}$. For obtaining that result, the following condition on the pilot parameter is required.

(A4) As $n \to \infty$, $R_{K;r,1}(\rho_n) = \infty$ and $n^{-3}R_{K;r,1}(\rho_n) = 0$.

Replacing, in Assumptions (A1)–(A3), the kernel $K$ by $L$, the smoothing parameter $\nu_n$ by $\rho_n$, and the order of the derivative $r$ by $s$, we obtain the following AMSE for the estimator $\hat{\psi}_{s;\rho}$ (see Section B of the Appendix for a formal proof).

**Theorem 2.** Under the Assumptions (A1)–(A4), (E1), and (E2) (using $L$, instead of $K$; $\rho_n$, instead $\nu_n$; and $s = r$, an even number); we have

\[
\text{AMSE} \left[ \hat{\psi}_{s;\rho} \right] = \left( n^{-1}R_{L;s,1}(\rho_n) + \frac{1}{2}h_L(\rho_n)\psi_{s+2} \right)^2 + 2n^{-2}\psi_0R_{L;s,2}(\rho_n) + 4n^{-1}\left( \int_{-\pi}^{\pi} \left( f(s)(\theta) \right)^2 f(\theta) d\theta - \psi_s^2 \right). \tag{4.4}
\]

As for Corollary 1, for deriving a simpler expression of the AMSE and the optimal value of $h_n \equiv h_L(\rho_n)$, we will assume the following extra conditions.

(E4) As $n$ increases, $R_{L;s,1}(\rho_n) = Q_{L;s,1}\rho_n^{-(s+1)/2}$, where $Q_{L;s,1}$ is a constant only depending on the kernel $L$ and the even number $s$. The sign of $Q_{L;s,1}$ is equal to the sign of $(-1)^{s+2}$.

(E5) $\lim_{n \to \infty} \int_{-\pi}^{\pi} \theta^4 L_{\rho_n}(\theta) d\theta = o(\rho_n^{5/4})$.

From the AMSE expression in (4.4), we can see that the optimal $h_n$ value, in terms of AMSE, will depend on the relation between $R_{L;s,1}(\rho_n)$, $R_{L;s,2}(\rho_n)$, and $h_L(\rho_n)$. Conditions (E3) and (E4) help to establish this relation, from which the following corollary is derived (see Section B of the Appendix).

**Corollary 2.** Consider the assumptions of Theorem 2, Conditions (E3), and (E4). Then, for the kernel estimator of $\psi_s$ (see (4.2)), we have that, asymptotically, the optimal value, in terms of the AMSE expression, of $h_n$ can be obtained from,

\[
h_{L;s;\text{AMSE}} = \left( -\frac{2Q_{L;s,1}}{n\psi_{s+2}} \right)^{2/(s+3)} . \tag{4.5}
\]

Under the previous assumptions, if Condition (E5) is also assumed, the minimal AMSE for (4.2) is of order $O(\min(5,s+3)/(s+3))$. If $s$ is an even number greater than 2, the minimal AMSE would be equal to

\[
\inf_{0 \leq \rho < 1} \text{AMSE} \left[ \hat{\psi}_{s;\rho} \right] = 2\psi_0Q_{K;r,2} \left( -\frac{\psi_{s+2}}{2Q_{L;s,1}} \right)^{(2s+1)/(s+3)} n^{-5/(s+3)} . \tag{4.6}
\]

If $s = 0$, the minimal AMSE is equal to

\[
\inf_{0 \leq \rho < 1} \text{AMSE} \left[ \hat{\psi}_{s;\rho} \right] = 4 \left( \int_{-\pi}^{\pi} \left( f(s)(\theta) \right)^2 f(\theta) d\theta - \psi_s^2 \right) n^{-1} . \tag{4.7}
\]

When $s = 2$, the minimal AMSE is equal to the sum of the right-hand sides of (4.6) and (4.7).
Note that Condition (E5) is only assumed to derive the same AMSE optimal order as in the linear case. If that condition is not fulfilled, then, when \( s > 0 \), the leading term in the AMSE could be of a larger order (see Section B of the Appendix). Another important consideration is that the sign of \( \psi_{s+2} \) is the same as that of \((-1)^{s/2+1}\) and, using Condition (E4), it also coincides with the sign of \(-Q_{L;8,1}\). Therefore, we always have that \( h_{L;8;\text{AMSE}} \geq 0 \) in (4.5).

Using the quantity \( h_{L;2r+4;\text{AMSE}} \) in (4.5), one can obtain the pilot smoothing parameter \( \rho \) needed to derive the direct plug-in estimator (4.3). The issue, as in the linear case, is that \( h_{L;2r+4;\text{AMSE}} \) still depends on the unknown value of the functional \( \psi_{2r+6} \). We comment on how to overcome this difficulty in Section 6.

5 The kernel choice

Theoretical results in Sections 3 and 4 provide mathematical support for deriving the optimal smoothing parameter. However, we still did not discuss how to obtain the plug-in concentration parameter \( \nu_{r;\text{AMISE}} \) from \( h_{K;r;\text{AMISE}} \). As mentioned in Section 3, we must solve the equation \( h_{K}(\nu) = h_{K;r;\text{AMISE}} \).

We also need to obtain the values of \( Q_{K;r,1} \) and \( Q_{K;r,2} \) in (3.2) and (4.5).

In this section, we study what happens when the “most common” circular models are employed as kernels. For doing so, we restrict our attention to the four standard choices of circular densities (see Mardia and Jupp, 1999, Section 3.5): cardioid, von Mises, wrapped normal, and wrapped Cauchy. Despite all of them satisfy Assumption (A2), the cardioid kernel, \( K_{C;\nu}(\theta) = (1 + 2\nu \cos(\theta))/(2\pi) \), with \( |\nu| < 1/2 \), does not meet Assumption (A3) and will be hence discarded from the analysis. This can be seen by observing that, for any \( \nu \), \( h_{K}(\nu) = \pi^2/3 - 4\nu \geq \pi^2/3 - 2 \). The other three circular densities are studied in Sections 5.1 and 5.2.

Once the behaviour of the standard kernels is studied, a second question to discuss is which is the optimal kernel in the circular context. As in Muller (1984), we study that optimality in terms of the AMISE expression in (3.3). Fixing \( f \) and the sample size \( n \), in Section 5.3, we obtain the circular kernel that minimizes the AMISE.

5.1 von Mises and wrapped normal kernels

First, denoting by \( I_j \) to the modified Bessel function of the first kind and order \( j \), let us consider the density expressions of the von Mises (VM) and the wrapped normal (WN) kernels.
\[ K_{\text{VM},\nu}(\theta) = \frac{\exp(\kappa \cos(\theta))}{(2\pi I_0(\kappa))} = \frac{1}{2\pi} \left( 1 + 2 \sum_{j=1}^{\infty} \frac{I_j(\kappa)}{I_0(\kappa)} \cos(j\theta) \right), \quad \text{where } \nu = \frac{I_1(\kappa)}{I_0(\kappa)}, \text{ being } \kappa \geq 0. \]

\[ K_{\text{WN},\nu}(\theta) = \frac{1}{2\pi} \left( 1 + 2 \sum_{j=1}^{\infty} \nu^j \cos(j\theta) \right), \quad \text{with } \nu \in [0, 1]. \]

For both kernels, if \( \nu_n = 1 \) (equivalently \( \kappa_n = \infty \), for the von Mises kernel), as \( n \to \infty \), then \( h_K(\nu_n) = 0 \). Even more, it is easy to show that, in that setting, Conditions (E1) and (E2) are satisfied. Thus, as \( n \to \infty \), we obtain the following asymptotic approximations.

\[ h_{K_{\text{VM}}}(\nu_n) = \frac{\pi}{2} \left( 1 - \frac{I_2(\kappa_n)}{I_0(\kappa_n)} \right) = \frac{1}{\kappa_n}, \quad \text{i.e., } \kappa_n \overset{\text{VM}}{=} h_n^{-1}. \]  

(5.1)

\[ h_{K_{\text{WN}}}(\nu_n) = \frac{1}{2} \left( 1 - \nu_n^4 \right), \quad \text{i.e., } \nu_n \overset{\text{WN}}{=} (1 - h_n)^{1/4}. \]  

(5.2)

From the previous equalities, we can see that \( \kappa_n \) or \( \nu_n \) are easily derived after computing the optimal/plug-in value of \( h_n \) (see, e.g., (3.2) or (4.3)). Again, considering \( \nu_n = 1 \), it can be seen that Conditions (E3)–(E5) are satisfied for the von Mises and wrapped normal kernels. For both kernels and a non-negative integer number \( r \), the values of the constants in (E3) and (E4) are the following ones,

\[ Q_{K,r,1} = (-1)^{r/2} \frac{r!}{2^{r/2}(r/2)!} \sqrt{2\pi} (r \text{ being even}), \quad Q_{K,r,2} = \frac{(2r)!}{2^{2r+1}r! \pi^{1/2}}. \]  

(5.3)

We can see that the AMSE and AMISE results derived in Sections 3 and 4 can be easily obtained in practice from these last quantities. These optimal asymptotic results will coincide for both, the von Mises and the wrapped normal kernel.

### 5.2 Wrapped Cauchy kernel

The wrapped Cauchy kernel density expression is

\[ K_{\text{WC},\nu}(\theta) = \frac{1 - \nu^2}{2\pi (1 + \nu^2 - 2\nu \cos(\theta))} = \frac{1}{2\pi} \left( 1 + 2 \sum_{j=1}^{\infty} \nu^j \cos(j\theta) \right), \quad \text{with } \nu \in [0, 1]. \]

From the last equality, we derive that \( h_{K_{\text{WC}}}(\nu) = \pi^2/3 + 4\text{Li}_2(-\nu) \), where \( \text{Li}_s \) is the polylogarithm of order \( s \). The last implies \( h_{K_{\text{WC}}}(\nu_n) = 0 \) if \( \nu_n = 1 \) as \( n \to \infty \). Also, the expressions of \( R_{K_{\text{WC},2}}(\nu) \) can be obtained in terms of a polylogarithm,
\[ R_{W_C; r, 2}(\nu) = \begin{cases} (2\pi)^{-1} (1 + 2\text{Li}_0(\nu^2)) & \text{if } r = 0, \\ \pi^{-1} \text{Li}_{-2r}(\nu^2) & \text{otherwise}, \end{cases} = O(h_n^{-2r+1}). \] (5.4)

If we consider a value of \( h_n \) such that \( h_n \) = 0 and \( nh_n^{2r+1} = \infty \), Condition (A3) is satisfied. Therefore, from Theorem 1, if also \( f \) satisfies Assumption (A1), we obtain the following AMISE for the wrapped Cauchy kernel,

\[ \text{AMISE} \left[ \hat{f}^{(r)}_\nu \right] = O \left( h_n^2 \right) + O \left( n^{-1}h_n^{-(2r+1)} \right). \] (5.5)

The AMISE expression (5.5) would be minimized if \( h_n \) is of order \( n^{-1/(2r+3)} \). Using the previous result, we obtain that the AMISE order of the wrapped Cauchy kernel is worse than that obtained with the von Mises or the wrapped normal kernel (see Corollary 1). In particular, its minimal AMISE is equal to

\[ \inf_{0 \leq \nu < 1} \text{AMISE} \left[ \hat{f}^{(r)}_\nu \right] = O \left( n^{-2/(2r+3)} \right). \]

When searching for an explicit expression of the optimal smoothing parameter, one could be tempted to combine (3.1) and (5.4). But note that this cannot be done as Condition (E1) is not verified. Thus, the explicit expression of the asymptotic bias cannot be obtained following the steps in Section A of the Appendix. This means that Equations (A.1) and, therefore, (3.1) are not necessarily true for this particular kernel. On the contrary, Tsuruta and Sagae (2017) were able to obtain an optimal smoothing parameter for the wrapped Cauchy kernel from the results by Di Marzio et al. (2009). Nevertheless, it should be noted that Di Marzio et al. (2009) results must not be employed for this kernel as Condition (E1) is not satisfied (see Remark 1).

### 5.3 Wrapped bounded-support kernels

To find the optimal kernels, consider a wrapped kernel \( K_\nu(\theta) = \sum_{\ell=0}^{\infty} K_{X; \lambda}(\theta + 2\ell \pi) \), whose associated linear density \( K_{X; \lambda} \) has bounded support, i.e., \( K_{X; \lambda}(x) = 0 \) if \( |x| > \lambda \). Then, \( K_\nu(\theta) = K_{X; \lambda}(\theta) \), for all \( \theta \in [-\pi, \pi) \), when \( \lambda < \pi \). In that case, \( h = \int_{-\lambda}^{\lambda} x^2 K_{X; \lambda}(x) dx \) and \( R_{K; r, 2}(\nu) = \int_{-\lambda}^{\lambda} (K_{X; \lambda}(x))^2 dx \).

Consider the asymptotic case for the linear bounded-support kernels, i.e., \( \lambda_n = 0 \). Fixing the unknown linear density function \( f_X \), Muller (1984) gives explicit expressions of the bounded-support kernels minimizing the AMISE when estimating the density function and its derivatives, in the linear case.

Assume now that the wrapped bounded-support kernel satisfies the Assumptions of Corollary 1. Except for the values not depending on the kernel, the minimal AMISE of \( \hat{f}^{(r)}_\nu \) in (3.3) is equal to the minimal AMISE obtained in the linear case. As a consequence, Theorem 2.4 of Muller (1984) can
be employed to see that the optimal kernels in terms of AMISE coincide with the wrapped version of those employed on linear kernel estimation.

The last means that the optimal kernel for circular density estimation is the wrapped Epanechnikov, whose associated linear density is $K_{X;\lambda}(x) = 3(1 - (x/\lambda)^2)/(4\lambda)$. When $\lambda < \pi$, the concentration is $\nu = (3\sin(\lambda) - 3\lambda\cos(\lambda))/\lambda^3$, $h = \lambda^2/5$, and $R_{K;0,2}(\nu) = 3/(5\lambda)$. Thus, for the circular density estimation with the wrapped Epanechnikov, the optimal (AMISE) value of $h_n$ is obtained by taking $Q_{K;0,2} = 3/(5\sqrt{5})$ in (3.2). Given a circular density $f$, independently of the kernel, the minimal AMISE that could be obtained when estimating the circular density is equal to

$$\inf_{0 \leq \nu < 1} \text{AMISE}\left[\hat{f}_\nu\right] = \frac{5}{4} \left(\frac{3}{5\sqrt{5}n} \int_0^{\pi} (f^{(2)}(\theta))^2 d\theta\right)^{4/5} \int_0^{\pi} (f^{(2)}(\theta))^2 d\theta.$$  

This is the AMISE obtained by the wrapped Epanechnikov. For the derivatives of the density, we refer to Muller (1984). There, we can see, e.g., that the wrapped Biweight would be the optimal kernel for the first derivative of $f$.

6 The plug-in smoothing parameters

In this section, we will study how to implement, in practice, the plug-in smoothing parameter $h_{K;r;\rho}$ in (4.3). As mentioned in Section 4, the issue of directly employing (4.3) is that the AMISE-optimal smoothing parameter for $\hat{\psi}_{s;\rho}$ will always depend on an unknown value of $\psi_{s+2}$. A way to overcome this difficulty is to provide an $l$-stage direct plug-in smoothing selector. This procedure consists in estimating $\psi_s$ with $\hat{\psi}_{s+2;\rho}$, in an iterative process, until some point in which we replace $\psi_{2r+2l+4}$ by its value obtained with a simple density (see Section 6.1 for more details). An alternative selector can be computed by noting that the smoothing parameter for $\hat{f}^{(r)}$ can be obtained as a function of the smoothing parameter for $\hat{\psi}_{2r+4}$. This allows us to construct a solve-the-equation rule. In Section 6.2, we discuss in more detail how this selector is derived.

6.1 The two-stage direct plug-in smoothing selector

The procedure to obtain the $l$-stage direct plug-in smoothing selector consists in, at stage 0, using a simple rule of thumb to compute the smoothing parameter of $\hat{\psi}_{s+2;\rho}$. Once this initial step is achieved, from that estimator of $\psi_{2r+4+2l}$, the following functional estimators are derived in an iterative process (see Algorithm 1 for details).

Two decisions remain from this brief explanation: the number of stages $l$ and which reference density should be employed at stage 0. Regarding the number of stages $l$, we suggest employing $l = 2$ for two reasons. First, since $l = 2$ is a common choice in the linear case (see, e.g., Wand and
Jones, 1995, Section 3.6). A second reason was obtained when replicating the simulation study in Section 7, with \( l = 3 \). In that case, similar results were obtained with respect to those obtained with \( l = 2 \), with the added inconvenience of the extra computational time.

As mentioned before, at stage 0, a reference density is needed to compute the smoothing parameter of \( \hat{\psi}_{2l+4;\rho} \). Here, in the same spirit of the original rule of thumb proposed by Taylor (2008), a natural selection would be to replace \( \psi_{2l+4;\rho} \) in (4.5) by the quantity obtained when assuming that the true density follows a von Mises distribution. In the circular case, the issue of employing that simple strategy is that a uniform estimation of the density can be obtained even if the true distribution is not uniform. This occurs, for example, when considering distributions with antipodal symmetry (see also Oliveira et al., 2012, for further discussion on this topic). The consequence would be to have a value of \( \hat{\psi}_{2l+4;\rho} \) close to zero, which derives in a “large” value of the smoothing parameter at the next stage \( \theta_{L;2l+2;\rho} \).

To avoid that last issue, while still having a simple model, one can use as the reference density the following mixture of \( M \) von Mises, all having the same concentration parameter \( \kappa \geq 0 \).

\[
\hat{f}_{M}\text{m}(\theta; M; \mu, \kappa, w) = \frac{1}{(2\pi I_0(\kappa))} \sum_{m=1}^{M} w_m \exp(\kappa \cos(\theta - \mu_m),)
\]

(6.1)

where the parameters \( \mu_m \in [-\pi, \pi) \), \( w_m \in [0, 1] \), for all \( m \in \{1, \ldots, M\} \), and with \( \sum_{m=1}^{M} w_m = 1 \). The value of \( \hat{\psi}_{2l+4;\rho} \) is calculated from the density (6.1), replacing its parameters by their maximum likelihood estimates obtained from the sample. An algorithm providing the maximum likelihood estimates for the density (6.1) was implemented in the R (R Core Team, 2022) library movMF by Hornik and Grün (2014). In practice, following Oliveira et al. (2012), the value of \( M \) can be chosen, using the Akaike Information Criterion (AIC), by comparing the results obtained with the mixtures (6.1) of \( M = 1, \ldots, M_{\text{max}} \) components.

In Algorithm 1, we summarize the steps that are needed to obtain our proposed two-stage direct plug-in smoothing parameter. There, for simplicity, we use \( K = L \), a kernel that satisfies Assumptions (A1)–(A4) and the extra Conditions (E1)–(E4).

**Algorithm 1.** Two-stage direct plug-in smoothing selector.

**Step 1:** Use a rule of thumb to obtain the estimator of \( \psi_{2l+8;\rho} \), \( \hat{\psi}_{2l+8;\rho} \). This can be achieved as follows.

**Step 1.a:** For every \( M = 1, \ldots, M_{\text{max}} \); obtain the maximum likelihood estimators of the parameters in the mixture of \( M \) von Mises, all having the same concentration parameter, i.e., of \( \mu, \kappa, \) and \( w \) in (6.1).

**Step 1.b:** Select the number of components in the mixture, \( M_{\text{AIC}} \), employing the AIC.

**Step 1.c:** Using the density \( \hat{f}_{M}\text{m}(\cdot; M_{\text{AIC}}; \hat{\mu}, \hat{\kappa}, \hat{\rho}) \), compute the value \( \hat{\psi}_{2l+8;\rho} \) (see Equality (4.1)).
Step 2: Estimate $\psi_{2r+6}$ using the estimator $\hat{\psi}_{2r+6;\rho_1}$, where $\rho_1$ is the plug-in concentration parameter relying on $\hat{\psi}_{2r+8;RT}$.

Step 2.a: Obtain $h_{K;2r+6;\rho_1}$ from (4.5), replacing $\psi_{2r+8}$ by $\hat{\psi}_{2r+8;RT}$.

Step 2.b: The value $\rho_1$ is the one that satisfies $h_{K}(\rho_1) = h_{K;2r+6;\rho_1}$, see Equation (2.4).

Step 3: Estimate $\psi_{2r+4}$ using the estimator $\hat{\psi}_{2r+4;\rho_2}$, where $\rho_2$ is the plug-in concentration parameter relying on $\hat{\psi}_{2r+6;\rho_1}$.

Step 4: Compute the AMISE-optimal smoothing parameter for $\hat{f}(r)$, relying on $\hat{\psi}_{2r+4;\rho_2}$.

Step 4.a: Obtain the two-stage direct plug-in smoothing parameter $h_{K;r;\rho_1}$ from (4.3), with the pilot smoothing parameter $\rho_2$.

Step 4.b: The selected concentration parameter $\nu_{DPI}$ is obtained from the value that satisfies $h_{K}(\nu_{DPI}) = h_{K;r;\rho_1}$.

The uniform distribution belongs to all the reference distributions mentioned before. Thus, in practice, the denominator of (4.3) or (4.5) could be equal to zero. If that occurs, we suggest directly returning a value of the concentration parameter $\nu_{DPI} = 0$, which would correspond with the uniform estimation of the density.

Given a finite value of $n$ in (4.3) or (4.5), the equation $h_{K}(\nu) = h_{K;r;\rho_1}$ or its approximation $\alpha_{K;2}(\nu) = 1 - 2h_{K;r;AMISE}$, cannot be solved for "large" values of $h_{K;r;\rho_1}$. The reason is that $h_{K}(\nu)$ and $\alpha_{K;2}(\nu)$ are non-negative and bounded for any value of $\nu$. Since a "large" value of $h_{K}(\nu)$ corresponds to $\nu = 0$, if the equation $h_{K}(\nu) = h_{K;r;AMISE}$ cannot be solved, we also suggest to employ the uniform as the density estimator.

6.2 Solve-the-equation plug-in smoothing selector

An alternative to the previous smoothing selector is the solve-the-equation rule. This rule consists in searching for the smoothing parameter that satisfies

$$h_{K;r;\rho_1} = \left( \frac{(2r + 1)Q_{K;r,2}}{n(1)^{r+2}\psi_{2r+4;\rho_1;\rho_1;AMISE}} \right)^{2/(2r+5)}$$

with $h_{K}(\rho_{K;r;STE}) = \gamma(h_{K;r;STE})$. (6.2)

Now, the smoothing parameter for $\hat{\psi}_{2r+4}$ is a function of the smoothing parameter for $\hat{f}(r)$. We suggest taking the function $\gamma$ by looking at the relation between the optimal smoothing parameter of these two estimators.

$$h_{L;2r+4;AMISE} = \left( (1)^{r+1} \frac{2Q_{L;2r+4,1}}{(2r + 1)Q_{K;r,2}} \right)^{2/(2r+7)} (\psi_{2r+4}/\psi_{2r+6})^{2/(2r+7)} h_{K;r;AMISE}^{(2r+5)/(2r+7)}.$$
Using a plug-in rule, this last relation suggests taking the following function,

\[ \gamma(h) = \left( -1 \right)^{r+1} \frac{2 Q_{L;2r+4,1}}{(2r + 1) Q_{K;2,r}^2} \left( \frac{\hat{\psi}_{2r+4;\hat{\rho}_1}}{\hat{\psi}_{2r+6;\hat{\rho}_2}} \right)^{2/(2r+7)} h^{(2r+5)/(2r+7)}. \]  

(6.3)

The two concentration parameters of the density functional estimators, \( \hat{\psi}_{2r+4} \) and \( \hat{\psi}_{2r+6} \), can be obtained from (4.5). Those smoothing parameters will depend on some other density functionals, \( \psi_{2r+6} \) and \( \psi_{2r+8} \). Following Sheather and Jones (1991), we suggest estimating these two new functionals with a rule of thumb. In Algorithm 2, we summarize the steps that are needed to obtain our proposed solve-the-equation plug-in smoothing parameter. Again, for simplicity, we use \( K = L \), a kernel that satisfies Assumptions (A1)–(A4) and the extra Conditions (E1)–(E4).

Algorithm 2. Solve-the-equation plug-in smoothing selector.

**Step 1:** Use a rule of thumb to obtain the estimator of \( \psi_{2r+6} \) and \( \psi_{2r+8} \), namely \( \hat{\psi}_{2r+6;\hat{\rho}_1} \) and \( \hat{\psi}_{2r+8;\hat{\rho}_2} \). This can be done as in Step 1 of Algorithm 1.

**Step 2:** Estimate \( \psi_{2r+4} \) and \( \psi_{2r+6} \) using the estimators \( \hat{\psi}_{2r+4;\hat{\rho}_1} \) and \( \hat{\psi}_{2r+6;\hat{\rho}_2} \), where \( \rho_1 \) is the plug-in concentration parameter relying on \( \hat{\psi}_{2r+6;\hat{\rho}_1} \), and \( \rho_2 \) the one relying on \( \hat{\psi}_{2r+8;\hat{\rho}_2} \).

**Step 3:** Employ the estimators \( \hat{\psi}_{2r+4;\hat{\rho}_1} \) and \( \hat{\psi}_{2r+6;\hat{\rho}_2} \) to obtain the function \( \gamma \) in (6.3).

**Step 4:** Using the \( \gamma \) function of Step 3, select the value of \( h_{K;\gamma;S} \) by solving Equation (6.2). The solve-the-equation concentration parameter \( \nu_{S} \) is obtained from the value that satisfies \( h_{K}(\nu_{S}) = h_{K;\gamma;S} \).

Note that two extra functional estimators must be estimated in Step 1 of Algorithm 2. For this reason, the obtained smoothing parameter could be considered as a two-stages solve-the-equation plug-in selector. As for the direct rule, the number of stages could be increased to estimate the density functionals. In particular, a rule of thumb can be employed to estimate \( \psi_{2r+10} \) or \( \psi_{2r+12} \) and then, in an iterative process, the values of \( \rho_1 \) and \( \rho_2 \) are obtained (as in Steps 2 and 3 of Algorithm 1).

The simulation study in Section 7 was carried out under the same conditions, using three (relying on \( \hat{\psi}_{2r+8;\hat{\rho}_1} \) and \( \hat{\psi}_{2r+10;\hat{\rho}_2} \)) and four (relying on \( \hat{\psi}_{2r+10;\hat{\rho}_1} \) and \( \hat{\psi}_{2r+12;\hat{\rho}_2} \)) stages. The results in practice were similar to those obtained with Algorithm 2, with the drawback of the extra computational time.

### 7 Simulation study

In this section, we performed a simulation study to analyse the performance of the direct two-stage plug-in concentration parameter (see Algorithm 1) and the solve-the-equation plug-in smoothing selector (see Algorithm 2). We focused only on the density estimation case \( (r = 0 \text{ in (2.2)}) \), with \( K \) being the von Mises kernel. The reason is that their effectiveness can be compared with other
smoothing parameters proposed in the literature. But note that our data-driven smoothing parameters could be employed to estimate any derivative and with other kernels. In particular, in this framework, slightly better results are expected by employing the wrapped Epanechnikov kernel (see Section 5.3).

Regarding the choice of $M_{\text{max}}$ at stage 0 (see Step 1 of Algorithms 1 and 2), we studied two scenarios. First, the reference model for the plug-in smoothing parameters was a simple von Mises ($M_{\text{max}} = 1$). As a second choice, we allowed for mixture models (6.1), with $M_{\text{max}} = 5$, in the reference density. In Tables 1 and 2, we show both results, $M_{\text{max}} = 1 (\nu_{\text{PPI};1})$ and $M_{\text{max}} = 5 (\nu_{\text{PPI};5})$, for the direct plug-in concentration parameter. While for the solve-the-equation plug-in smoothing selector ($\nu_{\text{STE};1}$), we only show the results when $M_{\text{max}} = 1$, given that our empirical experiments reveal that for the shown “small/moderate” sample sizes, most of the time $M_{\text{max}} = 1$ performs better, even in the more complex models, and it requires less computational time.

The results of using a common concentration parameter in the reference density (see (6.1)) at Step 1 were compared with those of the full von Mises mixture (allowing for different concentration parameters in each component). For most of the studied scenarios, better or similar results were obtained using a common value of $\kappa$, while keeping the computational efficiency. For that reason, we only show the results obtained using a common concentration parameter (as described in Algorithms 1 and 2).

We will compare the performance of the proposed concentration parameters with the following three rules for smoothing selection in circular density estimation, all implemented in the R library NPCirc (Oliveira et al., 2014).

- Rule of thumb of Taylor (2008), $\nu_{\text{RT}}$, where to compute $\psi_4$ in (4.3), it is assumed that $f$ follows a von Mises.

- Plug-in rule of Oliveira et al. (2012), $\nu_{\text{PMM}}$, where to compute $\psi_4$ in (4.3), it is assumed that $f$ follows a mixture of $M$ von Mises, with different concentration parameters. The number $M$ is chosen between 1 and 5, according to the AIC.

- Likelihood cross-validation of Hall et al. (1987), $\nu_{\text{LCV}}$. Consider the leave-one-out estimator $\hat{f}_{-i;\nu}(\theta)$, which corresponds with the kernel density estimation (2.1), leaving out the $i$-th observation. Then, the concentration parameter $\nu_{\text{LCV}}$ is obtained from the value of $\nu$ that maximizes $\prod_{i=1}^{n} \hat{f}_{-i;\nu}(\Theta_i)$.

The reasons for not showing the results with the other two well-known rules for obtaining the concentration parameter can be found in Oliveira et al. (2012). Oliveira et al. (2012) mentioned that their empirical experiments show that the likelihood cross-validation rule provides a more stable behaviour than the least-squares cross-validation method (Hall et al., 1987). The reason for not including the bootstrap smoothing selector of Di Marzio et al. (2011) is that it relies on an optimization
algorithm that searches for a local minimum. In that optimization, it is needed to impose the possible range of concentration values and, when \( n \) is not “too large” (\( n = 100 \)), the needed local minimum was not found for several samples. For \( n \) “large” (\( n = 250 \) or \( n = 500 \)), Oliveira et al. (2012) show that their plug-in rule outperforms the bootstrap method except when data is generated from the simplest models (M1–M4 below).

For comparing the performance of the different rules, we have employed the 20 reference models that can be found in Oliveira et al. (2012). They correspond to the uniform (M1), simple unimodal models (symmetric, M2–M5, and asymmetric M6), two-component mixture models (M7–M10), models with more than two components (M11–M16), and more complex models (M17–M20). Their density representations can be found in Tables 1 and 2, and their density expressions appear in Oliveira et al. (2012). From each model, we have generated 1000 random samples of sizes \( n = 50 \) and \( n = 100 \). Given the true density \( f \), for each sample, we have computed the integrated squared error (ISE) as the criterion to analyse the performance of the different estimators.

\[
\text{ISE}(\nu) = \int_{-\pi}^{\pi} \left( \hat{f}_\nu(\theta) - f(\theta) \right)^2 d\theta. \tag{7.1}
\]

As summary measures in Tables 1 (for Models M1–M14) and 2 (for Models M15–M20), for each rule for concentration selection, we show the average ISE and standard deviations, computed over 1000 replicates. The smallest average ISE values are highlighted in bold. On those tables, as a benchmark, we have also included the results with the “gold standard” smoothing parameter \( \nu_{GS} \), which is the value of \( \nu \) minimizing the ISE\((\nu)\) for each generated sample.

In Tables 1 and 2, we can observe that the proposed plug-in rules are the ones providing the smallest or the second smallest average ISE. The obtained average ISE is also close to the benchmark, which is remarkable considering that no-extra information is given about the true density and sample sizes are “small”. The “smallest” dispersion, measured through the standard deviation, is obtained either by the proposed plug-in rules (\( \nu_{DPI;1} \) and \( \nu_{STE;1} \)) or by the rule of thumb (\( \nu_{RT} \)).

In the following, we describe in more detail the behaviour of the two-stage plug-in rule in terms of the average ISE provided in Tables 1 and 2. First, we can observe that the plug-in rule with a simple von Mises at stage 0 (\( \nu_{DPI;1} \)) provides, in general, the smallest average ISE, especially when the sample size is “small” (\( n = 50 \)). The first exception to this general pattern occurs for densities that are well-approximated by a von Mises (M1, M3, and M4), where, as expected, the rule of thumb (\( \nu_{RT} \)) provides a slightly lower average ISE, being \( \nu_{DPI;1} \) the second one providing the best results. The second and more important exception occurs for Models M7, M11, M13, M14, M16, and M20; whose associated densities are multimodal and (almost) \( k \)-fold rotational symmetric. As mentioned in Section 6.1, the “bad” performance of \( \nu_{DPI;1} \) is probably due to the uniform estimation of the density by the von Mises density at stage 0. As commented there, we can solve that issue of the two-stage plug-in rule by using the mixture of von Mises at stage 0 (\( \nu_{DPI;5} \)). For those densities, we can see that \( \nu_{DPI;5} \) provides better results than \( \nu_{DPI;1} \).
### Table 1: Average ISE (×100) and standard deviations (×100, in parentheses) computed from 1000 samples of sample size \( n = 50 \) (first block) and \( n = 100 \) (second block) of Models M1–M14. In bold, smallest (dark) and second smallest (grey) average ISE. Smoothing parameters: concentration of thumb of Taylor (2008) \( (\nu_{\text{ST}}) \), plug-in rule of Oliveira et al. (2012) \( (\nu_{\text{MM}}) \), and likelihood cross-validation of Hall et al. (1987) \( (\nu_{\text{CV}}) \).

| Model | M1 | M2 | M3 | M4 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 |
|-------|----|----|----|----|----|----|----|----|----|-----|-----|-----|-----|-----|
| \( n = 50 \) | | | | | | | | | | | | | | |
| \( \nu_{\text{GS}} \) | 0 (0) | 0.762 (0.644) | 1.67 (1.346) | 0.695 (0.54) | 4.199 (2.499) | 3.258 (1.433) | 1.758 (1.062) | | | | | | | |
| \( \nu_{\text{GPI,1}} \) | 0.381 (0.51) | **1.068 (0.813)** | 2.283 (1.685) | 1.04 (0.753) | 5.495 (3.124) | **3.766 (1.61)** | 2.503 (1.091) | | | | | | | |
| \( \nu_{\text{GPI,5}} \) | 0.473 (0.746) | 1.134 (0.938) | 2.498 (2.289) | 1.317 (1.191) | 5.345 (3.262) | 4.645 (2.616) | 2.177 (1.399) | | | | | | | |
| \( \nu_{\text{STE,1}} \) | 0.847 (0.83) | 1.319 (0.999) | 2.586 (2.074) | 1.319 (1.005) | **5.105 (3.016)** | 3.883 (1.832) | **1.965 (1.166)** | | | | | | | |
| \( \nu_{\text{ST}} \) | **0.064 (0.176)** | 1.141 (0.958) | **1.98 (1.435)** | **0.93 (0.7)** | 10.893 (4.138) | 4.156 (1.469) | 10.471 (0.481) | | | | | | | |
| \( \nu_{\text{MM}} \) | 2.052 (2.807) | 2.688 (2.65) | 3.795 (3.813) | 2.646 (2.743) | 5.51 (3.244) | 5.57 (3.402) | 2.931 (2.67) | | | | | | | |
| \( \nu_{\text{CV}} \) | 0.595 (1.139) | 1.228 (1.041) | 2.45 (2.039) | 1.175 (1.051) | 11.45 (4.434) | 4.201 (1.952) | 2.086 (1.261) | | | | | | | |
| \( n = 100 \) | | | | | | | | | | | | | | |
| \( \nu_{\text{GS}} \) | 0 (0) | 0.467 (0.363) | 1.1 (0.868) | 0.442 (0.308) | 2.648 (1.507) | 2.272 (0.968) | 1.076 (0.572) | | | | | | | |
| \( \nu_{\text{GPI,1}} \) | 0.165 (0.226) | **0.609 (0.414)** | 1.412 (1.012) | 0.583 (0.39) | 3.411 (1.896) | 2.639 (1.054) | 1.437 (0.617) | | | | | | | |
| \( \nu_{\text{GPI,5}} \) | 0.217 (0.359) | 0.636 (0.478) | 1.49 (1.219) | 0.737 (0.558) | **3.138 (1.763)** | 2.874 (1.396) | 1.268 (0.684) | | | | | | | |
| \( \nu_{\text{STE,1}} \) | 0.383 (0.377) | 0.708 (0.5) | 1.512 (1.13) | 0.678 (0.486) | 3.188 (2.162) | **2.614 (1.122)** | **1.172 (0.606)** | | | | | | | |
| \( \nu_{\text{ST}} \) | **0.017 (0.055)** | 0.646 (0.511) | **1.26 (0.89)** | **0.536 (0.364)** | 8.587 (2.629) | 3.173 (0.962) | 10.554 (0.368) | | | | | | | |
| \( \nu_{\text{MM}} \) | 0.643 (1.007) | 1.077 (1.126) | 1.998 (1.789) | 1.118 (1.157) | 3.26 (1.883) | 3.27 (1.738) | 1.449 (1.051) | | | | | | | |
| \( \nu_{\text{CV}} \) | 0.334 (0.698) | 0.73 (0.589) | 1.51 (1.215) | 0.731 (0.632) | 6.827 (2.769) | 2.859 (1.241) | 1.259 (0.69) | | | | | | | |
Table 2: Average ISE (×100) and standard deviations (×100, in parentheses) computed from 1000 samples of sample size \( n = 50 \) (first block) and \( n = 100 \) (second block) of Models M15–M20.

| Model | M15 | M16 | M17 | M18 | M19 | M20 |
|-------|-----|-----|-----|-----|-----|-----|
| \( n = 50 \) |     |     |     |     |     |     |
| \( \nu_{GS} \) | 0.743 (0.201) | 3.644 (1.313) | 5.207 (1.906) | 3.211 (1.067) | 3.03 (1.001) | 4.965 (1.385) |
| \( \nu_{DPI;1} \) | 1.055 (0.519) | 8.128 (0.345) | 6.886 (2.07) | 3.949 (1.204) | 3.589 (1.579) | 6.113 (2.155) |
| \( \nu_{DPI;5} \) | 1.256 (0.968) | 4.942 (2.071) | 6.821 (2.155) | 4.067 (1.204) | 3.589 (1.579) | 6.113 (2.155) |
| \( \nu_{STE;1} \) | 1.404 (0.829) | 4.017 (1.411) | 6.162 (2.224) | 3.815 (1.306) | 3.41 (1.227) | 5.914 (1.364) |
| \( \nu_{RT} \) | 0.903 (0.192) | 7.879 (0.17) | 8.98 (1.667) | 4.212 (0.905) | 4.768 (1.004) | 10.966 (0.11) |
| \( \nu_{MM} \) | 2.781 (2.596) | 5.12 (2.396) | 7.196 (2.689) | 5.089 (2.562) | 4.737 (2.862) | 6.811 (2.565) |
| \( \nu_{CV} \) | 1.364 (0.98) | 3.973 (1.425) | 7.897 (2.241) | 4.11 (0.986) | 3.578 (1.268) | 5.472 (1.529) |
| \( n = 100 \) |     |     |     |     |     |     |
| \( \nu_{GS} \) | 0.63 (0.193) | 2.152 (0.775) | 3.437 (1.259) | 2.159 (0.8) | 2.188 (0.636) | 3.26 (0.888) |
| \( \nu_{DPI;1} \) | 0.768 (0.246) | 7.865 (0.14) | 5.196 (1.59) | 3.014 (0.64) | 2.569 (0.581) | 8.133 (0.14) |
| \( \nu_{DPI;5} \) | 0.86 (0.403) | 2.611 (1.247) | 4.695 (1.767) | 2.982 (0.745) | 2.48 (0.749) | 3.491 (0.977) |
| \( \nu_{STE;1} \) | 0.843 (0.374) | 2.324 (0.805) | 4.36 (1.822) | 2.59 (0.958) | 2.483 (0.62) | 4.067 (0.893) |
| \( \nu_{RT} \) | 0.85 (0.153) | 7.837 (0.045) | 7.93 (1.198) | 3.563 (0.493) | 3.875 (0.639) | 10.961 (0.053) |
| \( \nu_{MM} \) | 1.379 (1.028) | 2.5 (1.008) | 4.39 (1.744) | 3.081 (1.336) | 2.937 (1.237) | 3.893 (1.238) |
| \( \nu_{CV} \) | 0.971 (0.552) | 2.34 (0.849) | 5.695 (1.949) | 2.976 (0.83) | 2.519 (0.686) | 3.514 (0.899) |

On the complex models, those densities that are multimodal (M7, M8, M10–M16, and M18–M20) or “peaked” (M5 and M17), in general, the smallest average ISE is obtained by the solve-the-equation plug-in smoothing selector \( \nu_{STE;1} \), especially when the sample size is “moderate” \( n = 100 \). Even on the strongly (reflectively) asymmetric model (M6), \( \nu_{STE;1} \) provides the smallest average ISE \( (n = 100) \). Thus, the only cases where \( \nu_{STE;1} \) does not provide the smallest average ISE are the models well-approximated by a von Mises density (M1–M4 and M9). For the remaining models, if \( \nu_{STE;1} \) is not the “best” smoothing selector (M5, M15, M19, and M20), it is the rule providing the second smallest average ISE after the direct plug-in rule \( \nu_{DPI} \).

In summary, we have shown that the proposed plug-in rules provide competitive smoothing parameters for “small/moderate” sample sizes. We have also replicated the same simulation study for larger samples sizes \( (n = 250, n = 500, \text{ and } n = 1000) \). For the simple models (the ones well approximated by a von Mises density), \( \nu_{DPI;1} \) and \( \nu_{RT} \) still provide the “best” results. In the remaining models, for \( n = 250 \), we observe a similar pattern as that described for \( n = 100 \), with \( \nu_{STE;1} \) providing the “best” results most of the time. The relative performance (when compared with the other
data-based smoothing parameters) of \( \nu_{\text{DPI};5} \) and \( \nu_{\text{MvM}} \) improves with the sample size. Regarding the solve-the-equation plug-in smoothing selector, as mentioned before, the ISE values with \( M_{\text{max}} = 5 \), for \( n = 50 \) and \( n = 100 \), were not shown since, in most of the cases, \( M_{\text{max}} = 1 \) provided better results. We observed a different behaviour, in the complex models, for larger sample sizes (\( n = 500 \), and \( n = 1000 \)). In those cases, the solve-the-equation plug-in smoothing selector with \( M_{\text{max}} = 5 \) (\( \nu_{\text{STE};5} \)) obtained lower ISE values than those derived with \( M_{\text{max}} = 1 \) (\( \nu_{\text{STE};1} \)). As a result, for most of the non-simple models, \( \nu_{\text{MvM}} \) and \( \nu_{\text{STE};5} \) become the smoothing selectors providing the “best” results when \( n = 500 \) and \( n = 1000 \). In those scenarios, the selector \( \nu_{\text{MvM}} \) behaves better when data is generated by a mixture of von Mises, while \( \nu_{\text{STE};5} \) provides the best results in the mixtures with other (more complex) density models.

As a summary of our findings in this simulation study, our recommendations are as follows.

- For small sample sizes (say \( n = 50 \)), employ the proposed direct plug-in rule with \( M_{\text{max}} = 1 \) (\( \nu_{\text{DPI};1} \)). The only exception to this recommendation is if there could be evidence of the density function being multimodal and k-fold rotational symmetric. In that case, employ the solve-the-equation plug-in smoothing selector with \( M_{\text{max}} = 1 \) (\( \nu_{\text{STE};1} \)).

- For moderate sample sizes (say \( n = 100 \) or \( n = 250 \)), employ the solve-the-equation plug-in smoothing selector with \( M_{\text{max}} = 1 \) (\( \nu_{\text{STE};1} \)). The direct plug-in rule with \( M_{\text{max}} = 1 \) (\( \nu_{\text{DPI};1} \)) is also recommendable if there is evidence of unimodality.

- For large sample sizes (say \( n = 500 \) or \( n = 1000 \)), employ the plug-in rule of Oliveira et al. (2012) (\( \nu_{\text{MvM}} \)). An alternative would be the solve-the-equation plug-in smoothing parameter with \( M_{\text{max}} = 5 \) (\( \nu_{\text{STE};5} \)). When compared with \( \nu_{\text{MvM}} \), the selector \( \nu_{\text{STE};5} \) should provide a slightly better estimation when the density presents a complex shape.

## 8 Real data application

In this section, we revisit the car accident data that can be found in Ameijeiras-Alonso and Crujeiras (2022). This real dataset consists of the time of the day (at a resolution of one minute) at which the car crash happened in El Paso County (Texas, USA) in 2018. A total of 85 observations were recorded on the webpage of the National Highway Traffic Safety Administration of the United States.

In Ameijeiras-Alonso and Crujeiras (2022), several features of the shape of the distribution are analysed with different inferential tools. The conclusion is that the density is unimodal and (reflectively) asymmetric. More specifically among several parametric models studied by Ameijeiras-Alonso and Crujeiras (2022), the conclusion is that the “best” parametric fitting is achieved by the wrapped skew normal density of Pewsey (2000). A representation of the fitted wrapped skew normal density is provided in Figure 1 (left, thick solid line).
This real dataset constitutes a good example where we would recommend employing the proposed direct plug-in smoothing selector, with $M_{\text{max}} = 1$, to estimate the density function as the sample size is “small” ($n = 85$) and unimodality cannot be rejected for this sample. In Figure 1 (left), we show the estimated density function employing this direct plug-in rule (DPI; 1), the proposed solve-the-equation plug-in selector (STE; 1), the rule of thumb of Taylor (2008) (RT), and the plug-in rule of Oliveira et al. (2012) (MvM). We did not include the likelihood cross-validation of Hall et al. (1987) as its density estimation was close to the one obtained with the direct plug-in rule (DPI; 1). If we take the wrapped skew normal density as a reference, we can visually see that the closest kernel estimation is provided by the direct plug-in rule. This can be confirmed by computing the ISE (7.1), replacing $f$ with the fitted wrapped skew normal density. In Table 3, we can observe that the smallest ISE is obtained with the proposed direct plug-in rule.

Finally, another application of the results derived in this paper can be found when the interest is in the density derivative estimate. As before, we will focus on the results obtained with the direct plug-in rule, with $M_{\text{max}} = 1$. In that case, the smoothing parameter for the kernel estimator can be derived with Algorithm 1 (with $r = 1$). The obtained estimator, with the von Mises kernel, for the car accident data is plotted in Figure 1 (right). There, one can see that the derivative estimation is
Table 3: ISE ($\times 100$) computed from the car accident data, taking the fitted wrapped skew normal density as the reference density $f$ in (7.1). Smoothing parameters: rule of thumb of Taylor (2008) ($\nu_{RT}$), plug-in rule of Oliveira et al. (2012) ($\nu_{MvM}$), likelihood cross-validation of Hall et al. (1987) ($\nu_{LCV}$), proposed direct plug-in rule with $M_{\text{max}} = 1$ ($\nu_{DPI;1}$), and proposed solve-the-equation plug-in selector with $M_{\text{max}} = 1$ ($\nu_{STE;1}$).

| Smoothing parameter | $\nu_{RT}$ | $\nu_{MvM}$ | $\nu_{LCV}$ | $\nu_{DPI;1}$ | $\nu_{STE;1}$ |
|---------------------|------------|-------------|-------------|---------------|---------------|
| ISE($\nu$) ($\times 100$) | 1.134 | 1.765 | 0.325 | 0.265 | 0.509 |

9 Concluding remarks

The main contributions of this paper are the new plug-in smoothing parameters for circular kernel density (and its derivatives) estimation. In the past, some papers provided data-driven concentration parameters for estimating the circular density function, but still, a Sheather and Jones (1991) plug-in rule was missing in the circular literature. This paper fills that gap by providing all the theoretical results needed to derive an $l$-stage direct and solve-the-equation plug-in smoothing selectors for a general kernel satisfying some assumptions. The needed constants to obtain the optimal smoothing parameter are given for the “most-popular” circular kernels. Besides that, following the optimality criterion of Muller (1984), this paper also discusses kernel choice. The conclusion is that the optimal kernel for circular density estimation is the wrapped Epanechnikov. This paper also includes a simulation study confirming that the proposed plug-in rules provide competitive smoothing parameters when compared with other proposals available in the statistical literature.

Although not done in this paper, the presented ideas could be extended to the multivariate toroidal setting, using the product kernel of Di Marzio et al. (2011). In that case, the optimal smoothing parameter could be derived if the same kernel and concentration are employed in all the dimensions. Thus, in principle, one could obtain an explicit expression of the optimal smoothing parameter for the kernel estimators of the density derivative and the density functionals. Then, a similar scheme to that provided in Algorithms 1 or 2 could be employed to obtain a two-stage or a solve-the-equation plug-in rule for the toroidal kernel density (and its derivatives) estimator.
10 Availability

The proposed \( l \)-stage direct plug-in rule and the solve-the-equation selector have been added to the R library NPCirc \((\text{Oliveira et al., 2014})\). Given a dataset \( x \), the two-stage plug-in rule, described in Algorithm 1 (with \( M_{\text{max}} = 5 \)), for density estimation, can be obtained with the function \( \text{bw.AA}(x, \text{method}="dpi") \). The solve-the-equation smoothing selector, described in Algorithm 2 (with \( M_{\text{max}} = 1 \)), for density estimation is computed with \( \text{bw.AA}(x) \) or \( \text{bw.AA}(x, \text{method}="ste") \). Thus, \( \text{bw.AA} \) is the circular equivalent of the function \( \text{bw.SJ} \) of the \text{stats} R package.

Some other extra possibilities are available on the function \( \text{bw.AA} \). The kernel \( (K) \) and the derivative order \( (r \text{ in } (2.2)) \) can be selected, respectively, with the options \text{kernel} \ and \text{deriv.order}. If the practitioner prefers to use a different number of stages \( l \), this can be selected with \text{nstage}. At stage 0, this function also allows for modifying the number of components \( M (m) \) in the mixture (6.1), and selecting if a common concentration parameter is employed in all the components \( (\text{commonkappa}) \). Below, we provide the 3-stage direct plug-in rule for the first derivative of the density function. We employed the wrapped normal kernel and a mixture of 4 components, with different concentration parameters, at stage 0. We generated a sample \( x \) of 50 data from Model 18 from Oliveira et al. (2012).

\[
\text{R> library("NPCirc")}
\text{R> x <- rcircmix(n=50,model=18)}
\text{R> bw.AA(x, deriv.order=1, method="dpi", nstage=3, kernel="wrappednormal",}
\text{+ M=4, commonkappa=FALSE)}
\]

For the direct plug-in rule, a slightly more accurate (and more computationally inefficient) concentration parameter could be obtained if the asymptotic approximations are avoided, see, e.g., (5.1) and (5.3), for the von Mises kernel. This can be done if, at the end of every step of Algorithm 1, we compute an optimization routine searching the value of \( \nu \) that minimizes the corresponding MISE/AMISE criterion (3.1) or (4.4). In the optimization routine, we could employ as the initial value the concentration value achieved at each specific step (i.e., the values obtained by (4.5) and (4.3)). This extra substep can be performed, in the \( \text{bw.AA} \) function, if we impose the argument \text{approximate=FALSE}. The simulation study of Section 7 was carried out also with this extra substep and similar results were obtained.

A numerical routine is needed to obtain the value of the smoothing parameter, \( \nu_{\text{STE}} \), satisfying the Equality (6.2) in Step 4 of Algorithm 2. In our simulations and in \( \text{bw.AA}(x, \text{method}='\text{ste}') \), the \text{uniroot} function of the \text{stats} R package was employed. The range over which the value of \( h_{K;r,\text{STE}} \) is searched corresponds with the function arguments \((\text{lower}, \text{upper})\). The convergence tolerance, for searching the smoothing parameter with \text{uniroot}, can be modified with the argument \text{tol}.

In the function \( \text{bw.AA} \), the values of the constants in (E4) and (E5) are only provided for the von
Mises and the wrapped normal kernel. These values can be selected for other kernels with the arguments Q1 (for $Q_{K,1}$) and Q2 (for $Q_{K,2}$).

The kernel density derivative estimate for circular data was also added to the NPCirc package, inside the kern.den.circ function. By default it employs, as the smoothing parameter, the proposed solve-the-equation plug-in selector (with $M_{\text{max}} = 1$), described in Algorithm 2. Alternatively, the concentration parameter ($bw$) can be a real value or a character related to the available data-based smoothing parameters (e.g., if $bw=$"dpi", the direct plug-in rule is employed). The derivative order can be selected with the argument deriv.order. This function generates an object of the class density.circular (see the circular R package, Agostinelli and Lund, 2022). The remaining available options of kern.den.circ are, thus, similar to the ones in the function density.circular. A representation of the kernel density derivative estimator for the sample $x$ could be obtained as follows.

```r
R> fderhat <- kern.den.circ(x, deriv.order=1)
R> plot(fderhat, plot.type="line")
```

The dataset containing the 85 car crashes that happened in El Paso County (Texas, USA) in 2018 is available on the website https://github.com/jose-ameijeiras/Car-crashes-data.

## A Asymptotic results for the kernel derivative estimator

The main objective of this section is to compute the mean squared error (MSE) of the kernel derivative estimator in (2.2), i.e., $E \left[ \hat{f}^{(r)}(\theta) - f^{(r)}(\theta) \right]^2$. For doing so, firstly, we will consider the bias term and, secondly, the variance term. Regarding the expected value of $\hat{f}^{(r)}(\theta)$, we obtain the following result, using Taylor’s theorem, Assumptions (A1) over $f$, and (A2) over $K$.

$$
E \left[ \hat{f}^{(r)}(\theta) \right] = \int_{-\pi}^{\pi} K^{(r)}_\nu (\theta - \varphi) f(\varphi) d\varphi
= \int_{-\pi}^{\pi} K_\nu (\varphi) f^{(r)}(\theta + \varphi) d\varphi
= f^{(r)}(\theta) + \frac{1}{2} \int_{-\pi}^{\pi} \varphi^2 K_\nu (\varphi) f^{(r+2)}(\theta) + O \left( \int_{-\pi}^{\pi} \varphi^4 K_\nu (\varphi) d\varphi \right)
$$

Using the convergent Fourier series representation (2.3) and that $\int_{-\pi}^{\pi} \theta^2 \cos(j\theta) d\theta = 4\pi(-1)^j/j^2$, for any $j$ integer, we obtain that $\int_{-\pi}^{\pi} \varphi^2 K_\nu (\varphi) d\varphi = h_K(\nu)$. Thus, the bias of $\hat{f}^{(r)}(\theta)$ is equal to

$$
E \left[ \hat{f}^{(r)}(\theta) \right] - f^{(r)}(\theta) = \frac{1}{2} h_K(\nu) f^{(r+2)}(\theta) + O \left( \int_{-\pi}^{\pi} \varphi^4 K_\nu (\varphi) d\varphi \right).
$$
Now, if as \( n \to \infty, h_K(\nu_n) = 0 \), i.e., if Assumption (A3) is satisfied, \( \hat{f}^{(r)}(\theta) \) is asymptotically unbiased and its asymptotic bias leading term is of order \( O(h_K(\nu_n)) \). If in addition, we assume Condition (E1), we obtain the following expression of the bias of \( \hat{f}^{(r)}(\theta) \).

\[
\text{Bias} \left[ \hat{f}^{(r)}(\theta) \right] = \frac{1}{2} h_K(\nu_n) f^{(r+2)}(\theta) + o \left[ h_K(\nu_n) \right]. \tag{A.1}
\]

Regarding the variance term, using Assumptions (A1)–(A3), we obtain the following result.

\[
\text{Var} \left[ \hat{f}^{(r)}(\theta) \right] = n^{-1} \int_{-\pi}^{\pi} \left( K^{(r)}(\theta - \varphi) \right)^2 f(\varphi) d\varphi - n^{-1} \left( E \left[ \hat{f}^{(r)}(\theta) \right] \right)^2
\]
\[
= n^{-1} \int_{-\pi}^{\pi} \left( K^{(r)}(\varphi) \right)^2 f(\theta + \varphi) d\varphi - n^{-1} \left( f^{(r)}(\theta) + O(h_K(\nu)) \right)^2
\]
\[
= n^{-1} \int_{-\pi}^{\pi} \left( K^{(r)}(\varphi) \right)^2 d\varphi f(\theta) + O \left( n^{-1} \int_{-\pi}^{\pi} \varphi^2 \left( K^{(r)}(\varphi) \right)^2 d\varphi \right) + O(n^{-1}) \tag{A.2}
\]

As before, employing the convergent Fourier series representation (2.3), we would obtain that \( R_{K;r:2}(\nu) = \int_{-\pi}^{\pi} \left( K^{(r)}(\varphi) \right)^2 d\varphi \). Now, by Assumption (A3), as \( n \to \infty, n^{-1} = o[n^{-1} R_{K;r:2}(\nu_n)] \). Thus, the last term in (A.2), asymptotically vanishes. Then, the leading term of the asymptotic variance is \( O(n^{-1} R_{K;r:2}(\nu_n)) \). If the extra condition (E2) for \( K \) is also assumed, then the variance of \( \hat{f}^{(r)}(\theta) \) is equal to the following quantity.

\[
\text{Var} \left[ \hat{f}^{(r)}(\theta) \right] = n^{-1} R_{K;r:2}(\nu_n) f(\theta) + o \left[ n^{-1} R_{K;r:2}(\nu_n) \right]. \tag{A.3}
\]

Under the previous assumptions, combining (A.1) and (A.3), we derive the AMSE expression.

\[
\text{AMSE} \left[ \hat{f}^{(r)}(\theta) \right] = \frac{1}{4} h_K^2(\nu_n) \left( f^{(r+2)}(\theta) \right)^2 + \frac{1}{n} R_{K;r:2}(\nu_n) f(\theta). \tag{A.4}
\]

If \( f^{(r+2)} \) is integrable, integrating the AMSE expression in (A.4), with respect to \( \theta \), we derive the AMISE expression in (3.1). Now, the expression in (3.2) of the optimal smoothing parameter with respect to the AMISE criterion would be obtained as follows. First, employing the Condition (E3), \( R_{K;r:2}(\nu_n) = Q_{K;r:2} h_n^{(2r+1)/2} \). Replacing that last value in (3.1) and differentiating its expression with respect to \( h_n \), if we set this derivative equal to zero, we obtain as the equation solution the value in (3.2).

**B  Asymptotic results for the kernel density functional**

The first objective of this section is to derive the MSE of \( \hat{\psi}_{s;\rho} \), i.e., \( E \left[ \hat{\psi}_{s;\rho} - \psi_s \right]^2 \). For doing so, first consider the following reparametrization of the kernel estimator given in (4.2),
\[ \hat{\psi}_{s;\rho} = n^{-1} L_\rho^{(s)}(0) + n^{-2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} L_\rho^{(s)}(\Theta_i - \Theta_j). \] (B.1)

Since the first term does not depend on the data, we have only to study in detail the second term on the right-hand side of (B.1). Using Taylor’s theorem and Assumptions (A1) over \( f \) and (A2) over \( L \) (when considering \( s = r \)), we obtain the following result.

\[
\mathbb{E} \left[ L_\rho^{(s)}(\Theta_1 - \Theta_2) \right] = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} L_\rho^{(s)}(\theta - \vartheta) f(\vartheta) f(\theta) d\theta d\vartheta
= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} L_\rho(\Phi) f(\varphi + \theta) f^{(s)}(\vartheta) d\varphi d\theta
= \psi_s + \frac{1}{2} h_L(\rho) \psi_{s+2} + O \left( \int_{-\pi}^{\pi} \theta^4 L_\rho(\theta) d\theta \right). \tag{B.2}
\]

Now, combining (B.1) and (B.2), under Assumption (A3) (using \( L \), instead of \( K \); \( \rho_n \), instead \( \nu_n \)), we obtain that the bias of \( \hat{\psi}_{s;\rho} \) is equal to,

\[
\mathbb{E} \left[ \hat{\psi}_{s;\rho} \right] - \psi_s = n^{-1} L_\rho^{(s)}(0) + (1 - n^{-1}) \mathbb{E} \left[ L_\rho^{(s)}(\Theta_1 - \Theta_2) \right] - \psi_s
= n^{-1} \left( L_\rho^{(s)}(0) + \psi_s \right) + \frac{1}{2} h_L(\rho) \psi_{s+2} + O \left( \int_{-\pi}^{\pi} \theta^4 L_\rho(\theta) d\theta \right) + O \left( n^{-1} h_L(\rho) \right).
\]

Now, using the convergent Fourier series representation (2.3), it is easy to show that \( L_\rho^{(s)}(0) = R_{L;\rho,1}(\rho) \). Therefore, under Assumptions (A3), (A4), and (E1), we obtain that the bias is equal to the following quantity.

\[
\text{Bias} \left[ \hat{\psi}_{s;\rho} \right] = n^{-1} R_{L;\rho,1}(\rho_n) + \frac{1}{2} h_L(\rho_n) \psi_{s+2} + o \left[ n^{-1} R_{L;\rho,1}(\rho_n) + o \left( h_L(\rho_n) \right) \right]. \tag{B.3}
\]

If Condition (E1) is not fulfilled, then we can see that the asymptotic bias is of the same order, but we cannot derive a general explicit expression valid for any kernel \( L \), Bias \[ \hat{\psi}_{s;\rho} \] = \[ n^{-1} R_{L;\rho,1}(\rho_n) + O(h_L(\rho_n)) \).

For the variance, consider again the reparametrization (B.1). Now, since \( L^{(s)} \) is symmetric for \( s \) even, we have the following result (see, e.g., Wand and Jones, 1995, Ch. 3).

\[
\text{Var} \left[ \hat{\psi}_{s;\rho} \right] = 2n^{-3} (n - 1) \text{Var} \left[ L_\rho^{(s)}(\Theta_1 - \Theta_2) \right]
+ 4n^{-3} (n - 1)(n - 2) \text{Cov} \left[ L_\rho^{(s)}(\Theta_1 - \Theta_2), L_\rho^{(s)}(\Theta_2 - \Theta_3) \right]. \tag{B.4}
\]
Now, for computing the above value, first consider the following result, which is a consequence of Taylor’s theorem, Assumption (A1) over $f$, and (A2) over $L$ (when considering $s = r$).

$$
\mathbb{E} \left[ \left( L_{\rho}^{(s)} (\Theta_1 - \Theta_2) \right)^2 \right] = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left( L_{\rho}^{(s)} (\theta - \vartheta) \right)^2 f(\theta) f(\vartheta) d\theta d\vartheta
$$

$$
= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} (L_{\rho} (\varphi))^2 f(\varphi + \vartheta) f^{(s)}(\vartheta) d\varphi d\vartheta
$$

$$
= \psi_0 R_{L,s,2}(\rho) + O \left( \int_{-\pi}^{\pi} \theta^2 L_{\rho}^2 (\theta) d\theta \right). \quad \text{(B.5)}
$$

Note that under Assumptions (A3) and (E2), we obtain that, as $n \to \infty$, $\int_{-\pi}^{\pi} \theta^2 L_{\rho,n}^2 (\theta) d\theta = o(R_{L,s,2}(\nu_n))$, and $n^{-1} R_{L,s,2}(\nu_n) = 0$.

Similarly, the following result will be useful to derive the asymptotic results of the second part of the right-hand side of (B.4).

$$
\mathbb{E} \left[ L_{\rho}^{(s)} (\Theta_1 - \Theta_2) L_{\rho}^{(s)} (\Theta_2 - \Theta_3) \right]
$$

$$
= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} L_{\rho}^{(s)} (\theta - \vartheta) L_{\rho}^{(s)} (\vartheta - \varpi) f(\theta) f(\vartheta) f(\varpi) d\theta d\vartheta d\varpi
$$

$$
= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} L_{\rho} (\varphi) L_{\rho} (\phi) f^{(s)}(\vartheta + \varphi) f(\vartheta) f^{(s)}(\vartheta + \phi) d\varphi d\vartheta d\phi
$$

$$
= \int_{-\pi}^{\pi} \left( f^{(s)}(\vartheta) \right)^2 f(\vartheta) d\vartheta + O \left( h_L(\rho) \right). \quad \text{(B.6)}
$$

The last term in (B.6) asymptotically vanishes using the Assumption (A3) for the kernel $L$. From (B.2), we obtain that $\mathbb{E} \left[ L_{\rho}^{(s)} (\Theta_1 - \Theta_2) \right] = \psi_s + O(h_L(\rho))$. The variance of $\hat{\psi}_{s,\rho}$ can be obtained from (B.4), by combining the results in (B.2), (B.5), and (B.6).

$$
\text{Var} \left[ \hat{\psi}_{s,\rho} \right] = 2n^{-2} \left( \psi_0 R_{L,s,2}(\rho) + O \left( \int_{-\pi}^{\pi} \theta^2 L_{\rho}^2 (\theta) d\theta \right) - \psi_s^2 \right)
$$

$$
+ 4n^{-1} \left( \int_{-\pi}^{\pi} \left( f^{(s)}(\vartheta) \right)^2 f(\vartheta) d\vartheta + O \left( h_L(\rho) \right) - \psi_s^2 \right)
$$

$$
\text{AVar} \left[ \hat{\psi}_{s,\rho} \right] = 2n^{-2} \psi_0 R_{L,s,2}(\rho_n) + 4n^{-1} \left( \int_{-\pi}^{\pi} \left( f^{(s)}(\vartheta) \right)^2 f(\vartheta) d\vartheta - \psi_s^2 \right), \quad \text{(B.7)}
$$

where the last equality is a consequence of Assumptions (A3) and (E2).

Finally, the AMSE expression in (4.4) is obtained by adding the square of the asymptotic bias (see (B.3)) and the asymptotic variance (B.7).
Note that under Conditions (E3) and (E4), we obtain that $R_{L,s,2}(\rho_n) = o(R_{L,s,1}(\rho_n))$. Thus, the term depending on $\rho_n$ in the asymptotic variance (B.7) is a small $o$ term of the squared asymptotic bias. The optimal value of $h_{L,s}$ in (4.5) can be obtained equating to zero the derivative of the square of the asymptotic bias (see (B.3)), using the assumption about the sign in Condition (E4).

When employing the smoothing parameter $h_{L,s}$ in the asymptotic variance (B.7), the main term of the bias vanishes. Then, the minimal AMSE of $\hat{\psi}_{s,\rho}$ is of order

$$\begin{align*}
O \left( \int_{-\pi}^{\pi} \theta^4 L_{\rho_{L,s},\text{AMSE}}(\theta) d\theta \right)^2 + O(n^{-2} h_{L,s,\text{AMSE}}^2) + O(n^{-2} h_{L,s,\text{AMSE}}^{-(2r+1)/2}) + O(n^{-1})
\end{align*}$$

Now, under Condition (E5), we obtain that,

$$\begin{align*}
\text{AMSE} \left[ \hat{\psi}_{s;\rho_{L,s,\text{AMSE}}} \right] &= o(h_{L,s,\text{AMSE}}^{5/2}) + O(n^{-5/(s+3)}) + O(n^{-1}) = O(n^{-5/(s+3)}) + O(n^{-1}).
\end{align*}$$

The minimal AMSE orders in Corollary 2 follow. The explicit expression of the minimal AMSE can be obtained from the asymptotic variance expression in (B.7).

If Condition (E5) is not satisfied, then the minimal AMSE is of the following order,

$$\begin{align*}
\text{AMSE} \left[ \hat{\psi}_{s;\rho_{L,s,\text{AMSE}}} \right] &= O \left( \int_{-\pi}^{\pi} \theta^4 L_{\rho_{L,s,\text{AMSE}}}(\theta) d\theta \right)^2 + O(n^{-1}) = O(n^{-4/(s+3)}) + O(n^{-1}).
\end{align*}$$

References

Agostinelli, C. and U. Lund (2022). $R$ package circular: Circular Statistics. R package version 0.4-95.

Ameijeiras-Alonso, J., A. Benali, R. M. Crujeiras, A. Rodríguez-Casal, and J. M. Pereira (2019). Fire seasonality identification with multimodality tests. *Ann. Appl. Stat.* 13(4), 2120–2139.

Ameijeiras-Alonso, J. and R. M. Crujeiras (2022). Flexible circular modeling: A case study of car accidents. In *Directional Statistics for Innovative Applications*, pp. 93–116. Singapore: Springer.

Beran, R. J. (1979). Exponential models for directional data. *Ann. Stat.* 7(6), 1162–1178.
Cwik, J. and J. Koronacki (1997). A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. *Comput. Stat. Data Anal.* 26(2), 199–218.

Di Marzio, M., A. Panzera, and C. C. Taylor (2009). Local polynomial regression for circular predictors. *Stat. Probab. Lett.* 79(19), 2066–2075.

Di Marzio, M., A. Panzera, and C. C. Taylor (2011). Kernel density estimation on the torus. *J. Stat. Plan. Inference* 141(6), 2156–2173.

Hall, P., G. S. Watson, and J. Cabrera (1987). Kernel density estimation with spherical data. *Biometrika* 74(4), 751–762.

Hornik, K. and B. Grün (2014). movMF: an R package for fitting mixtures of von Mises-Fisher distributions. *J. Stat. Softw.* 58(10), 1–31.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* 91(433), 401–407.

Klemelä, J. (2000). Estimation of densities and derivatives of densities with directional data. *J. Multivar. Anal.* 73(1), 18–40.

Ley, C. and T. Verdebout (Eds.) (2018). *Applied Directional Statistics*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton: CRC Press.

Mardia, K. V. and P. E. Jupp (1999). *Directional Statistics*. Wiley Series in Probability and Statistics. Chichester: Wiley.

Muller, H.-G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Stat.*, 766–774.

Oliveira, M., R. M. Crujeiras, and A. Rodríguez-Casal (2012). A plug-in rule for bandwidth selection in circular density estimation. *Comput. Stat. Data Anal.* 56(12), 3898–3908.

Oliveira, M., R. M. Crujeiras, and A. Rodríguez-Casal (2014). NPCirc: an R package for nonparametric circular methods. *J. Stat. Softw.* 61(9), 1–26.

Pewsey, A. (2000). The wrapped skew-normal distribution on the circle. *Commun. Stat. Theory Methods* 29(11), 2459–2472.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

SenGupta, A. and B. C. Arnold (Eds.) (2022). *Directional Statistics for Innovative Applications*. Forum for Interdisciplinary Mathematics. Singapore: Springer.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 53(3), 683–690.
Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Comput. Stat. Data Anal.* 52(7), 3493–3500.

Tsuruta, Y. and M. Sagae (2017). Asymptotic property of wrapped Cauchy kernel density estimation on the circle. *Bull. Inform. Cybernet.* 49, 1–10.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. Great Britain: Chapman and Hall.