Annals of the Institute of Statistical Mathematics (2020) 72:511–530
https://doi.org/10.1007/s10463-018-0701-x

Theoretical properties of bandwidth selectors for kernel density estimation on the circle

Yasuhito Tsuruta[1] · Masahiko Sagae[2]

Received: 21 June 2017 / Revised: 15 October 2018 / Published online: 26 November 2018

Abstract

We derive the asymptotic properties of the least squares cross-validation (LSCV) selector and the direct plug-in rule (DPI) selector in the kernel density estimation for circular data. The DPI selector has a convergence rate of $O(n^{-5/14})$, although the rate of the LSCV selector is $O(n^{-1/10})$. Our simulation shows that the DPI selector has more stability than the LSCV selector for small and large sample sizes. In other words, the DPI selector outperforms the LSCV selector in theoretical and practical performance.

Keywords Kernel density estimation · Circular data · Smoothing parameter selector · Least squares cross-validation · Direct plug-in rule

1 Introduction

Kernel density estimation is the standard nonparametric method for exploring the structure of circular data. The structure of the kernel density estimator is largely influenced by the value of the smoothing parameter. Therefore, its selection is an important problem in the practical analysis of circular data.

Prior studies proposed automatic selectors of the smoothing parameter for circular data and examined their practical performances in simulations. However, to our knowledge, no study derived theoretical properties for the selectors in the field of circular data analysis.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0701-x) contains supplementary material, which is available to authorized users.

✉ Yasuhito Tsuruta
tsuruta9764@gmail.com

[1] Wakayama Prefectural Office, 1-1 Komatsubaradori, Wakayama-shi, Wakayama-ken 640-8585, Japan

[2] School of Economics, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa 920-1192, Japan

We will explore the least squares cross-validation (LSCV) selector proposed by Hall et al. (1987) and the direct plug-in rule (DPI) selector proposed by Di Marzio et al. (2011). The LSCV selector is a common circular data analysis method due to its simple definition. A few studies researched the properties of the DPI selector for circular data. However, in the studies on selectors for linear data, Wand and Jones (1994) pointed out that the DPI selector has a better performance than the LSCV selector with respect to the convergence rate to the optimal smoothing parameter.

This study derives the theoretical properties of both the LSCV and DPI selectors, including the asymptotic normality and the convergence rate (see Hall and Marron 1987; Scott and Terrell 1987; Sheather and Jones 1991 for previous studies regarding linear data). We obtained the rates from the central limit theorem of a degenerate U-statistic given by Hall (1984). We demonstrate that the convergence rate of the DPI selector is $O(n^{-5/14})$ and that of the LSCV selector is $O(n^{-1/10})$. Numerical experiments show that DPI is much more stable than LSCV, even when sample size n is not large enough.

2 Properties of kernel density estimation

We give the definitions and the asymptotic properties of the kernel density estimators on a circle. A kernel density estimator $\hat{f}_\kappa(\theta)$ of unknown density f based on a random sample $\Theta_1, \ldots, \Theta_n$ is defined as

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

where $K_\kappa(\theta)$ is a symmetric kernel function, and κ is a concentration parameter that acts as a smoothing parameter, and corresponds to $\kappa = h^{-2}$ for a general bandwidth $h > 0$. Our loss function between $\hat{f}_\kappa$ and f is the integrated squared error (ISE) given by $\mathrm{ISE}[\hat{f}_\kappa] := \int_{-\pi}^{\pi}\{\hat{f}_\kappa(\theta) - f(\theta)\}^2 \mathrm{d}\theta$. The risk is the mean integrated squared error (MISE) given by $\mathrm{MISE}[\hat{f}_\kappa] := E_f[\mathrm{ISE}[\hat{f}_\kappa]]$.

We now employ the kernel function for circular data proposed by Hall et al. (1987).

Definition 1 (*Kernel function*) A function $K_\kappa(\theta)$: $[-\pi, \pi) \to \mathbb{R}$ is a kernel function. Let $K_\kappa(\theta)$ denote $K_\kappa(\theta) := C_\kappa^{-1}(L)L_\kappa(\theta)$, where

$$L_\kappa(\theta) := L(\kappa\{1 - \cos(\theta)\}) \tag{1}$$

and $C_\kappa(L) := \int_{-\pi}^{\pi} L_\kappa(\theta)\mathrm{d}\theta$. We define the l-th moment of L as

$$\mu_l(L) := \int_0^\infty L(r) r^{(l-1)/2}\mathrm{d}r,$$

where $l \geq 0$ is even and $r = \kappa\{1 - \cos(\theta)\}$. For even number $p \geq 2$, the function L satisfies the following eight conditions:

(a) The fourth derivative $L^{(4)}(r) := \mathrm{d}^{(4)}L(r)/\mathrm{d}r^4$ is continuous.
(b) If r is large, then $L(r)r^{(p+1)/2} = O(r^{-(p+4)/2})$.
(c) The term $\delta_{2t}(L) := \int_{-\infty}^{\infty} L^2(z^2/2)z^{2t}\mathrm{d}z$ is bounded for $t = 0, 1$.
(d) The moment $\mu_l(L)$ is bounded for $0 \le l \le p+4$, and $\mu_l(L) = \mu_{\kappa,l}(L) + O(\kappa^{-(p+6)/2})$, where $\mu_{\kappa,l}(L) := \int_0^{\kappa} L(r)r^{(l-1)/2}\mathrm{d}r$.
(e) $\lim_{|z|\to\infty} \eta(z)|z|^{3/2} = o(1)$, where $\eta(z) := \int_{-\infty}^{\infty} L(t^2/2)L((t+z)^2/2)\mathrm{d}t$.
(f) $\lim_{|z|\to\infty} \lambda(z)|z|^{3/2} = o(1)$, where $\lambda(L) := \int_{-\infty}^{\infty} L'(t^2/2)L((t+z)^2/2)t^2/2\mathrm{d}t$ is bounded.
(g) The term $\delta_t(S_4^m) := \int_{-\infty}^{\infty} S_4^{2m}(z^2/2)z^{2t}\mathrm{d}z$ is bounded for $t = 1, 2$ and $m = 1, 2$, where $S_4(z^2/2) := 3S^{(2)}(z^2/2) - 6z^2S^{(3)}(z^2/2) + z^4S^{(4)}(z^2/2)$.
(h) For any r, $L(r) \ge 0$.

Conditions (a), (c), and (d) are required to derive MISE$[\hat{f}_\kappa]$; we can replace condition (a) on the assumption that L' is continuous. We use conditions (a)–(f) to prove the theoretical properties of the LSCV selector, and need conditions (a), (c), (d), and (g) to prove the theoretical properties of the DPI selector.

A kernel such as $L(r) = e^{-r}$ satisfies all conditions of Definition 1 and is equivalent to a von Mises (vM) kernel such as $L_\kappa(\theta) = \exp[-\kappa\{1 - \cos(\theta)\}]$. Hall et al. (1987) suggested that smooth and rapidly varying kernels of type (1) are asymptotically equivalent to the kernel of $L(r) = e^{-r}$.

Let $R(g(\theta)\theta^t) := \int_{-\pi}^{\pi} g^2(\theta)\theta^{2t}\mathrm{d}\theta$. Then, we show the following asymptotic MISE.

Theorem 1 *Assume that the following conditions hold*:

(i) $\kappa = \kappa(n)$ *and* $\lim_{n\to\infty} \kappa(n) = \infty$.
(ii) $\lim_{n\to\infty} n^{-1}\kappa^{1/2}(n) = 0$.
(iii) f *is fourth differentiable and* $f^{(s)}$ *is square-integrable for* $s = 1, 2$.

Then, the MISE is given by

$$\mathrm{MISE}[\hat{f}_\kappa] = \mathrm{AMISE}[\hat{f}_\kappa] + o\left(\kappa^{-2} + n^{-1}\kappa^{1/2}\right),$$

where

$$\mathrm{AMISE}[\hat{f}_\kappa] = \frac{\mu_2^2(L)}{\mu_0^2(L)}R(f'')\kappa^{-2} + n^{-1}\kappa^{1/2}d(L), \tag{2}$$

and $d(L) := 2^{-1}\mu_0^2(L)\delta_0(L)$. *The minimizer* κ_* *of* (2) *is given by*

$$\kappa_* = \beta(L)R(f^{(2)})^{2/5}n^{2/5}, \tag{3}$$

where $\beta(L) := [4\mu_2^2(L)/\{\mu_0^2(L)d(L)\}]^{2/5}$. *Then, the convergence rate of the optimal MISE is* $O(n^{-4/5})$.

See Tsuruta and Sagae (2017) for details of Theorem 1. In general, we need to estimate κ_*, which depends on an unknown functional $R(f^{(2)})$. Therefore, we discuss the properties of the selectors of κ_*: the LSCV and DPI selectors from the next section.

3 Bandwidth selectors

3.1 Least squares cross-validation

The motivation of the LSCV selector comes from the minimization of $\mathrm{ISE}[\hat{f}_\kappa] - R(f)$. The LSCV selector $\hat{\kappa}_{\mathrm{CV}}$ is defined as the minimizer of the CV function given by

$$\mathrm{CV}(\kappa) := R(\hat{f}) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(\Theta_i), \tag{4}$$

where $\hat{f}_{-i}(\Theta_i) = (n-1)^{-1} \sum_{j \neq i}^{n} K_\kappa(\theta - \Theta_j)$. Since $\frac{n}{n-1} \to 1$, we can replace (4) by

$$\mathrm{CV}(\kappa) := \frac{R(K_\kappa)}{n} + \frac{2}{n^2} \sum_{i<j} \gamma(y_{ij}), \tag{5}$$

where $y_{ij} := \Theta_i - \Theta_j$ and $\gamma(y) = \int_{-\pi}^{\pi} K_\kappa(w) K_\kappa(w+y) \mathrm{d}w - 2K_\kappa(y)$. We apply the augmented cross-validation given by

$$\overline{\mathrm{CV}}(\kappa) := \mathrm{CV}(\kappa) + \frac{2}{n} \sum_{i} f(\Theta_i) - R(f)$$

for the theoretical analysis. Then, we obtain the variance of $\overline{\mathrm{CV}}(\kappa)$, which has a faster order than that of $\mathrm{CV}(\kappa)$ and is similar to that derived by Scott and Terrell (1987), who indicated that the augmented cross-validation for linear data provides a smaller variance. We derive the expectation and variance of $\overline{\mathrm{CV}}(\kappa)$ as the following theorem:

Theorem 2 *Assume the three conditions of Theorem 1, $R(f^{(4)} f^{1/2}) < \infty$, and $R((f^{(4)})^{1/2} f) < \infty$.*

Then, it follows that

$$E_f[\overline{\mathrm{CV}}(\kappa)] = \mathrm{AMISE}[\hat{f}_\kappa] + o\left(\kappa^{-2} + n^{-1}\kappa^{1/2}\right), \tag{6}$$

and

$$\mathrm{Var}_f[\overline{\mathrm{CV}}(\kappa)] = \frac{2}{n^2} \kappa^{1/2} Q(L) R(f) + o\left(n^{-2}\kappa^{1/2} + n^{-1}\kappa^{-2}\right), \tag{7}$$

where $Q(L) := \int_{-\infty}^{\infty} \{2^{-1}\mu_0^{-2}(L)\eta(z) - 2^{1/2}\mu_0^{-1}(L)L(z^2/2)\}^2 \mathrm{d}z$.

With a strategy similar to Scott and Terrell (1987), Theorem 2 leads to an LSCV selector $\hat{\kappa}_{\mathrm{CV}}$ consistent with the minimizer κ_*.

Corollary 1 *Let $\hat{\kappa}_{\mathrm{CV}} := \mathrm{argmin}_{\kappa \in (a\kappa_*, b\kappa_*)} \mathrm{CV}(\kappa)$ for $0 < a < 1$ and $1 < b$. Then, it holds that*

$$\hat{\kappa}_{\mathrm{CV}}/\kappa_* \xrightarrow{p} 1,$$

as $n \to \infty$.

3.2 Direct plug-in rule

Note that $\psi_r := \int_{-\pi}^{\pi} f^{(r)}(\theta) f(\theta) \mathrm{d}\theta$ and $R(f^{(r)}) = (-1)^r \psi_{2r}$. We now define the DPI estimator as

$$\hat{\kappa}_{\mathrm{PI}} := \beta(L) \hat{\psi}_4(g)^{2/5} n^{2/5},$$

where

$$\hat{\psi}_4(g) := n^{-1} \sum_{i=1}^{n} \hat{f}_g^{(4)}(\Theta_i) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} T_g^{(4)}(\Theta_i - \Theta_j), \tag{8}$$

where $T_g^{(4)}(\theta) := C_\kappa^{-1}(L) S_g^{(4)}(\theta)$, and g and $T_g(\theta) := C_\kappa^{-1}(S) S_g(\theta)$ are a smoothing parameter and a kernel that is possibly different from κ and K_κ, respectively. The main term $S_g^{(4)}(\theta)$ is given by

$$\begin{aligned} S_g^{(4)}(\theta) := &-g \cos(\theta) S_g^{(1)}(\theta) + g^2 \left\{ -4 \sin^2(\theta) + 3 \cos^2(\theta) \right\} S_g^{(2)}(\theta) \\ &+ 6 g^3 \cos(\theta) \sin^2(\theta) S_g^{(3)}(\theta) + g^4 \sin^4(\theta) S_g^{(4)}(\theta). \end{aligned} \tag{9}$$

The asymptotic properties for the mean square error (MSE) of $\hat{\psi}_4$ play an important role in showing the theoretical properties of $\hat{\kappa}_{\mathrm{PI}}$ in the next section. We provide the bias and variance of $\hat{\psi}_4(g)$ in the following theorem.

Theorem 3 *Assume that the following conditions hold*:

(i) $g := g(n)$, $\lim_{n\to\infty} g(n) = \infty$, *and* $\lim_{n\to\infty} n^{-2} g^{9/2}(n) = 0$.
(ii) f *is 6-th differentiable;* ψ_6 *is bounded.*

Then, the bias is given by

$$\mathrm{Bias}_f \left[\hat{\psi}_4(g) \right] = \mathrm{Abias}_f \left[\hat{\psi}_4(g) \right] + O\left(n^{-1} g^{3/2} + g^{-2} \right), \tag{10}$$

where

$$\mathrm{Abias}_f \left[\hat{\psi}_4(g) \right] = 3 n^{-1} g^{5/2} S_g^{(2)}(0) / \left\{ 2^{1/2} \mu_0(S) \right\} + \mu_2(S) \mu_0(S)^{-1} \psi_6 g^{-1}.$$

The variance is given by

$$\mathrm{Var}_f \left[\hat{\psi}_4(g) \right] = 4 n^{-1} \mathrm{Var}_f \left[f^{(4)}(\Theta_i) \right] + 2 G_{1,0}(S_4) \psi_0 n^{-2} g^{9/2} + o\left(n^{-1} + n^{-2} g^{9/2} \right), \tag{11}$$

where $G_{m,t}(S_4) := 2^{-m}\mu_0^{-2m}(S)\delta_t(S_4^m)$*. Select the optimal smoothing parameter* g_* *such that* $\mathrm{Abias}_f[\hat{\psi}(g)] = 0$*. Then,* g_* *is given by*

$$g_* = c\psi_6 n^{2/7}, \tag{12}$$

where $c := -2^{1/2}\mu_2(S)/(3S_g^{(2)}(0))$*. Selecting* g_* *means that the remaining squared bias is* $\mathrm{Bias}_f^2[\hat{\psi}_4(g_*)] = O(n^{-8/7})$ *and the variance is* $\mathrm{Var}_f[\hat{\psi}_4(g_*)] = O(n^{-5/7})$*. Thus, we obtain* $\inf_{g>0}\mathrm{MSE}[\hat{\psi}_4(g)] = O(n^{-5/7})$.

We consider the positive condition of g_*. Since $\psi_6 = -R(f^{(3)})$ shows that this condition is $\mu_2(S)/S_g^{(2)}(0) > 0$, the vM kernel is one suitable kernel satisfying $\mu_2(S)/S_g^{(2)}(0) > 0$.

Estimating g_* also requires estimating an unknown functional ψ_6. We provide the simplest estimator of ψ_6 by employing a reference density that we assume as true density. We proposed the two reference densities: the vM density and a wrapped Cauchy (wC) density. The vM density is defined as

$$f_{\mathrm{vM}}(\theta;\tau) := (2\pi I_0(\tau))^{-1}\exp\{\tau\cos(\theta)\},$$

where $I_p(\tau)$ denotes the modified Bessel function of the first kind and order p, and τ is the concentration parameter. The functional ψ_6 of the vM density is given by

$$\psi_6^{\mathrm{vM}}(\tau) = -\left[4\tau I_1(2\tau) + 30\tau^2 I_2(2\tau) + 15\tau^3 I_3(2\tau)\right] / \left\{16\pi I_0^2(\tau)\right\}.$$

The wC density is defined by

$$f_{\mathrm{wC}}(\theta;\rho) := \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta)},$$

where $\rho \in (0, 1)$ is the concentration parameter. The functional ψ_6 of the wC density is given by

$$\psi_6^{\mathrm{wC}}(\rho) = -\left[\rho^2 + 57\rho^4 + 302\rho^6 + 302\rho^8 + 57\rho^{10} + \rho^{12}\right] / \left\{\pi(1-\rho^2)^7\right\}.$$

We propose the easy and practical algorithm that employs the vM density or the wC density as the reference density for a direct plug-in rule, which we call the "one-step direct plug-in rule".

We provide an algorithm in which the vM density is the reference density.

Algorithm 1 The algorithm uses the following procedure:

Step 1 Calculate Maximum likelihood estimator $\hat{\tau}$ and $\psi_6^{\mathrm{vM}}(\hat{\tau})$.
Step 2 Compute $\hat{g} := [c\psi_6^{\mathrm{vM}}(\hat{\tau})n]^{2/7}$ as the estimator of g_*.
Step 3 Compute $\hat{\kappa}_{\mathrm{PI.vM}} = \beta(L)\hat{\psi}_4(\hat{g})^{2/5}n^{2/5}$.

We provide an algorithm in which the wC density is the reference density.

Algorithm 2 The algorithm uses the following procedure:

```
Step 1  Calculate Maximum likelihood estimator $\hat{\rho}$ and $\psi_6^{\mathrm{wC}}(\hat{\rho})$.
Step 2  Compute $\tilde{g} := [c\psi_6^{\mathrm{wC}}(\hat{\rho})n]^{2/7}$ as the estimator of $g_*$.
Step 3  Compute $\hat{\kappa}_{\mathrm{PI.wC}} = \beta(L)\hat{\psi}_4(\tilde{g})^{2/5}n^{2/5}$.
```

If we admit sacrificing the nonnegativity of kernels, then we find that the convergence rate of the MSE of $\hat{\psi}_4$ is $O(n^{-1})$ after applying the p-th order kernel proposed by Tsuruta and Sagae (2017).

Definition 2 (*p-th order kernel function*) $K_\kappa(\theta)$ is a p-th order kernel if $K_\kappa(\theta)$ satisfies conditions (a)–(g) in Definition 1 and

$$\mu_0(L) \neq 0, \quad \mu_l(L) = 0 \quad l = 2, 4, \ldots, p-2, \quad \text{and} \quad \mu_l(L) \neq 0, \quad l = p.$$

We obtain the following MSE when employing a p-th order kernel.

Corollary 2 *Assume that condition (i) in Theorem 3 holds, f is $(4+p)$th differentiable, and ψ_{4+2t} is bounded for $t = 1, 2, \ldots, p/2$. Then, when we employ a p-th order kernel, the bias is given by*

$$\mathrm{Bias}_f\left[\hat{\psi}_4(g)\right] = \mathrm{Abias}_f\left[\hat{\psi}_4(g)\right] + O\left(n^{-1}g^{3/2} + g^{-(p+2)/2}\right), \tag{13}$$

where

$$\mathrm{Abias}_f\left[\hat{\psi}_4(g)\right] = \frac{3g^{5/2}S_g^{(2)}(0)}{2^{1/2}\mu_0(S)n} + \frac{\mu_p(S)}{\mu_0(S)}\sum_{t=1}^{p/2}\frac{b_{p,2t}\psi_{4+2t}}{(2t)!}g^{-p/2},$$

and $b_{p,2t}$ is the constant (see Lemma 2 in Tsuruta and Sagae 2017 for its definition). The variance is equal to (11). From (13), the optimal parameter g_p is given by

$$g_p = W(S)n^{2/(p+5)},$$

where $W(S) = \left[-\{2^{1/2}\mu_p(S)\sum_{t=1}^{p/2}[\psi_{4+2t}b_{p,2t}/(2t)!]\}/\{3S_g^{(2)}(0)\}\right]^{2/(p+5)}$. Then, we can easily show that $g_2 = g_*$. If the order of the kernel is $p \geq 4$, then we obtain $\inf_{g>0}\mathrm{MSE}[\hat{\psi}_4(g)] = O(n^{-1})$; otherwise, we obtain the result in Theorem 3.

The proof is easily shown by Lemma 2 in Tsuruta and Sagae (2017) the same way as in the proof of Theorem 3. Providing the positive condition of g_p for $p \geq 4$ is difficult because g_p has the sum of some unknown functionals ψ_r. In the practical analysis, we recommend employing the suitable kernels of order $p = 2$ such as the vM kernel to avoid this problem.

4 Theoretical properties for the selectors

From theoretical perspective, we must inspect whether the DPI selector outperforms the LSCV selector. We measure the theoretical performance of selector $\hat{\kappa}$ by the convergence rate of the relative error $\hat{\kappa}/\kappa_* - 1$. We derive the rate through the asymptotically normal distribution:

$$n^{\alpha}(\hat{\kappa}/\kappa_* - 1) \xrightarrow{d} N\left(0, \sigma^2\right),$$

where $\sigma^2 < \infty$ depends only on f and L, but not on n. Theorems 4 and 5 show the asymptotic distributions of LSCV and DPI, respectively.

Theorem 4 *Assume that all conditions of Theorems 1 and 2 hold. Then, it holds that*

$$n^{1/10}(\hat{\kappa}_{\mathrm{CV}}/\kappa_* - 1) \xrightarrow{d} N\left(0, \sigma_{\mathrm{CV}}^2\right), \tag{14}$$

as $n \to \infty$, *where* $\sigma_{\mathrm{CV}}^2 := 50d^{-2}(L)M_{1,0}(L)R(f)\beta^{-1/2}(L)R(f'')^{-1/5}$, *and* $M_{m,t}(L) := \int_{-\infty}^{\infty} m(L)^{2m}z^{2t}\mathrm{d}z$, *where*

$$m(L) := 2^{-1}\mu_0^{-2}(L)\{\eta(z) + \lambda(z) + \lambda(-z)\} - 2^{-1/2}\mu_0^{-1}(L)\left\{L(z^2/2) + L(z^2/2)z^2\right\}.$$

Theorem 5 *Assume that the conditions of Theorem 3 hold. Then, when we employ the suitable second-order kernel, and it holds that*

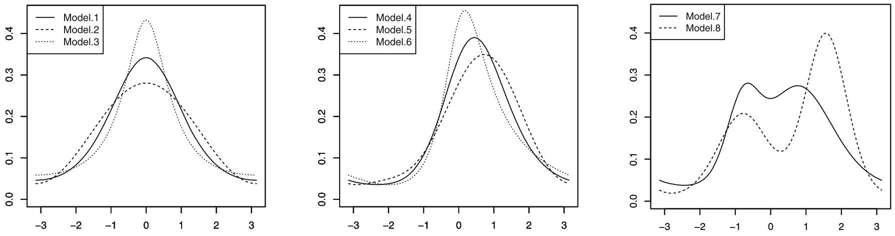
$$n^{5/14}(\hat{\kappa}_{\mathrm{PI}}/\kappa_* - 1) \xrightarrow{d} N\left(0, \sigma_{\mathrm{PI}}^2\right), \tag{15}$$

as $n \to \infty$, *where* $\sigma_{\mathrm{PI}}^2 = 8W^{9/2}(S)G_{1,0}(S_4)\psi_0\psi_4^{-2}/25$.

The convergence rates of $\hat{\kappa}_{\mathrm{CV}}$ and $\hat{\kappa}_{\mathrm{PI}}$ are equivalent to that of the LSCV and DPI selectors on the real line, respectively (Hall and Marron 1987; Scott and Terrell 1987; Sheather and Jones 1991). The rate of $\hat{\kappa}_{\mathrm{PI}}$ is much faster than that of $\hat{\kappa}_{\mathrm{CV}}$. Moreover, $\hat{\kappa}_{\mathrm{PI}}$ is more stable with the smaller order of variance. Therefore, the DPI selector is more appealing with respect to theoretical performance than the LSCV selector.

5 Numerical experiment

Analyzing a real-line small sample often does not indicate the same effect as the theoretical results. Therefore, we perform a simulation to compare the LSCV and DPI selectors with the eight simulation scenarios (models 1–8) in Fig. 1 when we employ the vM kernel. Models 1–3 are well-used distributions: vM, wC, and cardioid distributions. Additionally, models 4–6 (sine skewed vM, sine skewed wC, and sine skewed cardioid distributions) are produced by skewing to the three above distributions, respectively. These six distributions are subclasses of a sine skewed



(a) Models 1–3. **(b)** Models 4–6. **(c)** Models 7–8.

Fig. 1 Models 1–3 are von Mises, cardioid, and wrapped Cauchy densities. Models 4–6 are sine skewed von Mises, sine skewed cardioid, and sine skewed wrapped Cauchy densities. Models 7–8 are mixtures of the two von Mises densities

Jones–Pewsey distribution (Abe and Pewsey 2011). Models 7–8 are mixtures of two vM distributions.

We conduct our simulation with the statistical software R (R Core Team 2018) according to the following procedure:

1. Execute the following six steps for model 1:
   (a) Generate a random sample of size n distributed in model 1.
   (b) Calculate the optimal parameter κ_* applying the density of the model 1 to (3).
   (c) Estimate $\hat{\kappa}_{\mathrm{CV}}$ by `bw.cv.mse.circular`, which is a function in the `circular` (Agostinelli and Lund 2017) library of R.
   (d) Estimate $\hat{\kappa}_{\mathrm{PI.vM}}$ by Algorithm 1 and $\hat{\kappa}_{\mathrm{PI.wC}}$ by Algorithm 2.
   (e) Calculate the three relative errors: $Y_{\mathrm{CV}} = \hat{\kappa}_{\mathrm{CV}}/\kappa_* - 1$, $Y_{\mathrm{PI.vM}} = \hat{\kappa}_{\mathrm{PI.vM}}/\kappa_* - 1$, and $Y_{\mathrm{PI.wC}} = \hat{\kappa}_{\mathrm{PI.wC}}/\kappa_* - 1$.
   (f) Repeat steps (a)–(e) 10,000 times, and give the three sample means and the three sample standard errors of mean of Y_{CV}, $Y_{\mathrm{PI.vM}}$, and $Y_{\mathrm{PI.wC}}$.
2. Execute steps (a)–(f) for models 2–8.

We now discuss the small sample properties of the selectors. Table 1 shows that the DPI.vM selector is the most stable and the DPI.wC selector is the second most stable, but the LSCV selector is highly unstable for all models.

The key to explaining the performance of these selectors is the curvature of f: $R(f'')$, because (3) shows that it determines the value of the optimal parameter κ_* when n and the kernel are fixed. Therefore, each relative magnitude of the optimal parameter in Table 2 corresponds the relative magnitude of the curvature. The DPI.vM selector outperforms the others in the models with the small curvatures (models 1, 2, 4, and 5). The DPI.wC and LSCV selectors perform well when the curvature is large (models 3, 6, 7, and 8). In models 7–8, the LSCV selector has the best performance for $n \geq 500$, and the DPI.wC selector has the best performance for $n \leq 200$. The DPI.vM selector tends to oversmooth, and the DPI.wC and LSCV selectors tend to undersmooth.

Table 1 The means (its standard error) of the relative errors of the three selectors $(\hat{\kappa}/\kappa_* - 1) \times 100$ (κ_* is the optimal parameter) in models 1–8 (and sample sizes n are 50, 100, 200, 500, and 1000) in the simulation in Sect. 5, based on the number of repetitions $N = 10{,}000$

n	Selecter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
50	DPI.vM	11.8 (0.34)	37.3 (0.40)	− 38.3 (0.20)	5.9 (0.30)	18.1 (0.33)	− 34.1 (0.19)	− 37.1 (0.18)	− 51.9 (0.13)
	DPI.wC	115.4 (1.03)	127.7 (1.05)	38.3 (0.73)	125.1 (1.02)	132.4 (1.05)	49.7 (0.71)	0.4 (0.42)	− 14.8 (0.45)
	LSCV	188.2 (5.38)	284.7 (7.68)	60.6 (2.60)	148.9 (4.28)	189.3 (5.19)	57.6 (2.45)	76.4 (3.30)	49.3 (2.21)
100	DPI.vM	7.2 (0.25)	28.9 (0.29)	− 38.1 (0.15)	1.8 (0.22)	12.3 (0.24)	− 33.7 (0.15)	− 39.4 (0.13)	− 49.4 (0.10)
	DPI.wC	89.0 (0.65)	96.1 (0.64)	23.4 (0.45)	96.4 (0.64)	101.5 (0.64)	34.2 (0.45)	− 11.4 (0.25)	− 19.8 (0.29)
	LSCV	139.1 (4.11)	203.8 (5.61)	40.8 (1.97)	105.1 (3.15)	135.0 (3.86)	40.9 (1.87)	45.4 (2.38)	37.2 (1.65)
200	DPI.vM	4.2 (0.18)	23.4 (0.21)	− 36.3 (0.12)	0.2 (0.16)	8.9 (0.18)	− 32.0 (0.11)	− 39.7 (0.10)	− 45.6 (0.08)
	DPI.wC	71.9 (0.44)	77.5 (0.43)	15.4 (0.31)	79.2 (0.44)	83.0 (0.44)	25.0 (0.31)	− 16.7 (0.17)	− 20.4 (0.19)
	LSCV	91.7 (2.87)	147.7 (4.12)	24.2 (1.42)	76.3 (2.38)	97.0 (2.90)	25.1 (1.37)	26.2 (1.78)	25.7 (1.22)
500	DPI.vM	2.5 (0.13)	18.5 (0.14)	− 32.8 (0.09)	− 1.2 (0.12)	6.1 (0.12)	− 28.8 (0.08)	− 38.0 (0.08)	− 39.5 (0.06)
	DPI.wC	56.9 (0.29)	61.4 (0.28)	9.5 (0.20)	62.9 (0.29)	66.3 (0.30)	17.6 (0.20)	− 18.8 (0.12)	− 18.3 (0.12)
	LSCV	58.5 (1.94)	92.0 (2.72)	12.9 (0.96)	46.5 (1.57)	59.6 (1.91)	14.0 (0.91)	11.2 (1.20)	15.2 (0.81)
1000	DPI.vM	1.7 (0.10)	15.7 (0.11)	− 29.8 (0.07)	− 1.2 (0.09)	4.6 (0.10)	− 26.3 (0.07)	− 35.8 (0.07)	− 34.8 (0.04)
	DPI.wC	48.6 (0.22)	52.2 (0.21)	6.8 (0.15)	54.3 (0.23)	56.8 (0.23)	13.4 (0.15)	− 18.8 (0.10)	− 16.1 (0.08)
	LSCV	44.8 (1.49)	67.5 (2.04)	8.8 (0.72)	34.4 (1.16)	44.2 (1.44)	8.9 (0.69)	7.5 (0.92)	11.0 (0.61)

Table 2 The optimal parameter of models 1–8 from the simulation in Sect. 5

n	1	2	3	4	5	6	7	8
50	3.51	2.32	7.51	4.67	3.69	8.24	5.53	8.45
100	4.63	3.06	9.91	6.16	4.87	10.87	7.30	11.15
200	6.11	4.04	13.08	8.12	6.42	14.35	9.63	14.71
500	8.82	5.82	18.87	11.72	9.27	20.7	13.89	21.22
1000	11.64	7.68	24.89	15.46	12.23	27.31	18.33	28.00

The values are the optimal parameter κ_* in models 1–8 (rows) and $n = 50$, 100, 200, 500, and 1000

6 Applications

We now illustrate a real data example showing that a choice among the three selectors leads to different conclusions on a case study and another example in which it does not. Figures 2 and 3 show that the data sets A and B consist of observation values of wind direction acquired every 10 min at Kanazawa University in Japan for a period in which Japan was experiencing a typhoon in 2014. Let the observed values $\theta_i \in [-\pi, \pi)$ take an increasing value clockwise from $-\pi$ (north) on the circle.

We provide kernel density estimations of each data set employing the DPI.vM, $\hat{\kappa}_{\mathrm{PI.vM}}$; DPI.wC, $\hat{\kappa}_{\mathrm{PI.wC}}$; or LSCV $\hat{\kappa}_{\mathrm{CV}}$ selectors, where we estimate these estimators as in the above section. For data set A (Fig. 2), Fig. 4 shows that the DPI.vM selector produces the bimodal density estimation, but the DPI.wC and LSCV estimators provide the too-jagged estimation with the five peaks. They seem to be too-jagged estimation. For data set B (Fig. 3), Fig. 5 indicates that the choice between large concentration parameters give almost the same estimation. This is because the effect of the concentration parameter κ on the vM kernel decreases as κ gets larger because

Fig. 2 Data set A. The rose diagram shows the frequencies of 97 wind directions measured between 5:00 am and 9:00 pm on August 10th

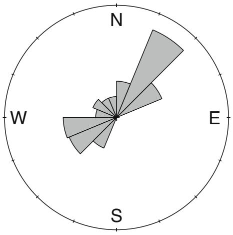
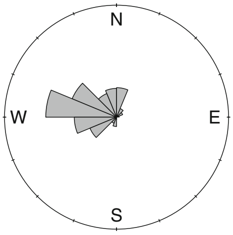


Fig. 3 Data set B. The rose diagrams show the frequencies of 288 wind directions measured from 0:00 am on July 10th to 11:50 pm on July 11th



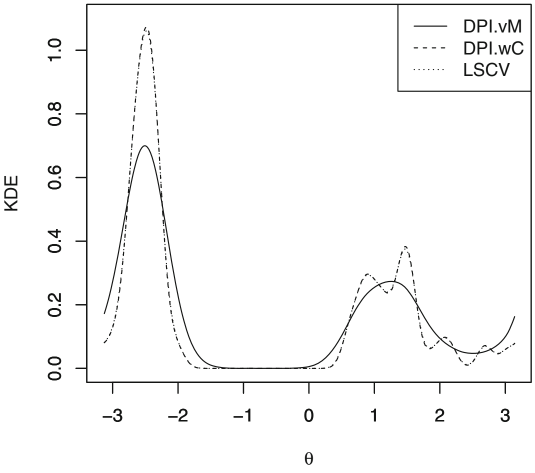


Fig. 4 The kernel density estimations of data set A in Fig. 2. The three lines show three estimations obtained by applying it $\hat{\kappa}_{\text{PI.vM}} = 12.89$, $\hat{\kappa}_{\text{PI.wC}} = 75.28$, and $\hat{\kappa}_{\text{CV}} = 69.34$

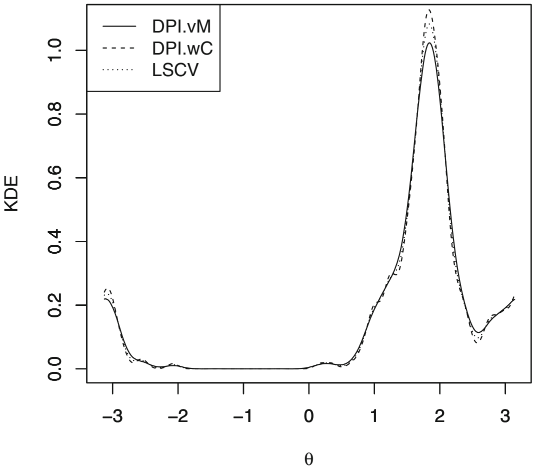


Fig. 5 The kernel density estimations of data set B in Fig. 3. The three lines show three estimations obtained by applying it to $\hat{\kappa}_{\text{PI.vM}} = 55.7$, $\hat{\kappa}_{\text{PI.wC}} = 135.85$, and $\hat{\kappa}_{\text{CV}} = 88.25$

the circular variance of the vM kernel, $1 - I_1(\kappa)/I_0(\kappa)$, approximates $1/(2\kappa)$ (Mardia and Jupp 1999).

7 Discussion

We derived the asymptotic properties for the least squares cross-validation selector and the direct plug-in selector for circular data. The convergence rate of the DPI selector is $O(n^{-5/14})$ and that of the LSCV selector is $O(n^{-1/10})$. The rates are equivalent to the two selectors on the real line. Thus, the theoretical performance of the DPI selector is better than that of the LSCV selector. Our simulation shows that the DPI selector is more stable than the LSCV selector.

We now discuss the properties of the two selectors with comparing them with the recent literatures. Di Marzio et al. (2018) proposed the local likelihood cross-validation (LCV) selector that is a new estimator with a penalized term. The LCV selector outperforms the LSCV selector in Table 3 in Di Marzio et al. (2018), but there is little literature that studied the theoretical aspect of the LCV selector. Its theoretical results such as the consistency may support its numerical result in Di Marzio et al. (2018).

Di Marzio et al. (2017) suggested another derivative estimator $\hat{f}^{(j)}(\theta; p)$. It may produce a functional estimator $\tilde{\psi}_j := n^{-1} \sum_i \hat{f}^{(j)}(\Theta_i; p)$. We think that the derivative and functional estimators are possible to have higher order rate of the MISE, but Di Marzio et al. (2017) did not derive its rate of the MISE of these estimators. If we obtain it, we can compare our functional estimator to these estimators. Investigating properties of $\hat{f}^{(j)}(\theta; p)$ is an interesting study and may provide a new selector that is one of the DPI estimator.

Acknowledgements We would like to thank the reviewers for the helpful comments. This work was partially supported by JSPS KAKENHI Grant Nos. JP16K00043, JP24500339, and JP16H02790.

Appendix A

Proof of Theorem 2. We set $\gamma(y_{ij}) = \gamma_{ij}$ to ease the notation. First, we calculate the expectation of $\overline{\text{CV}}(\kappa)$, given by

$$E_f[\overline{\text{CV}}(\kappa)] = \frac{R(K_\kappa)}{n} + \frac{2}{n^2}\sum_{i<j} E_f[\gamma_{ij}] + \frac{2}{n}\sum_{i} E_f[f(\Theta_i)] - R(f). \tag{16}$$

We set $\gamma_i = E_f[\gamma_{ij}|\Theta_i]$. Then, the conditional expectation γ_i is given by

$$\gamma_i = -f(\Theta_i) + f^{(4)}(\Theta_i)\mu_0^{-2}(L)\mu_2^2(L)\kappa^{-2} + O(\kappa^{-3}). \tag{17}$$

“Appendix B” in ESM presents the details. It follows from (17) that

$$E_f[\gamma_{ij}] = E_f[\gamma_i] = -R(f) + R(f^{(2)})\mu_0^{-2}(L)\mu_2^2(L)\kappa^{-2} + O(\kappa^{-3}). \tag{18}$$

Lemma 1 (Tsuruta and Sagae 2017) *The term $R(K(\theta)\theta^t)$ is equal to*

$$R(K(\theta)\theta^t) := \kappa^{-(2t-1)/2}[d_{2t}(L) + o(1)],$$

where $d_{2t}(L) := 2^{-1}\mu_0^{-2}(L)\delta_{2t}(L)$ and $d(L) := d_0(L)$.

By considering Lemma 1, (18), and $E_f[f(\Theta_i)] = R(f)$, we find that $E_f[\overline{\text{CV}}(\kappa)]$ is equivalent to (6).

We calculate the variance of $\overline{\text{CV}}(\kappa)$. That is,

$$\begin{aligned}\text{Var}_f[\overline{\text{CV}}(\kappa)] \simeq\ & 2n^{-2}\text{Var}_f\left[\gamma_{ij}\right] + 4n^{-1}\text{Var}_f\left[f(\Theta_i)\right] + 4n^{-1}\text{Cov}_f\left[\gamma_{ij}, \gamma_{ik}\right] \\ & + 8n^{-1}\text{Cov}_f\left[\gamma_{ij}, f(\Theta_i)\right],\end{aligned} \tag{19}$$

where $j \neq k$. Let $I_1 := R((f^{(4)})^{1/2} f)$, $I_2 := R(f^{(2)})R(f)$, and $I_{3:} = R(f^{3/2}) - R(f)^2$. Each term on the right-hand side of (19) is given by

$$\mathrm{Var}_f[\gamma_{ij}] = \kappa^{1/2}[Q(L)R(f) + o(1)], \tag{20}$$

$$\mathrm{Var}_f[f(\Theta_i)] = I_3, \tag{21}$$

$$\mathrm{Cov}_f[\gamma_{ij}, \gamma_{ik}] = I_3 - 2\{I_1 - I_2\}\mu_0^{-2}(L)\mu_2^2(L)\kappa^{-2} + o(\kappa^{-2}), \tag{22}$$

and

$$\mathrm{Cov}_f[\gamma_{ij}, f(\Theta_i)] = -I_3 + \{I_1 - I_2\}\mu_0^{-2}(L)\mu_2^2(L)\kappa^{-2} + o(\kappa^{-2}). \tag{23}$$

"Appendix C" in ESM provides the details of (20)–(23). By considering (19)–(23), we find that $\mathrm{Var}_f[\overline{\mathrm{CV}}(\kappa)]$ is equivalent to (7). □

proof of Corollary 1 We set $c := \hat{\kappa}_{\mathrm{CV}}/\kappa_*$. Then, we combine Theorems 1 and 2 and find that

$$\mathrm{AMISE}(c\kappa_*)/\mathrm{MISE}(c\kappa_*) \xrightarrow{p} 1, \tag{24}$$

$$\overline{\mathrm{CV}}(c\kappa_*)/\mathrm{MISE}(c\kappa_*) \xrightarrow{p} 1, \tag{25}$$

and

$$\mathrm{AMISE}(c\kappa_*)/\mathrm{AMISE}(\kappa_*) = \frac{1}{5c^2} + \frac{4c^{1/2}}{5}. \tag{26}$$

(26) is a convex function with a minimum at $c = 1$. Thus, if $c \neq 1$ and n is large, then it follows from combining (24) and (26) that

$$\mathrm{MISE}(c\kappa_*) > \mathrm{MISE}(\kappa_*). \tag{27}$$

Suppose that c does not converge to 1. Recall that it is necessary that $\overline{\mathrm{CV}}(c\kappa_*) \leq \overline{\mathrm{CV}}(\kappa)$ for any κ, because $\hat{\kappa}_{\mathrm{CV}}$ is the minimizer of $\overline{\mathrm{CV}}(\kappa)$. Additionally, if n is large, then $\overline{\mathrm{CV}}(\kappa)$ is a convex function with a minimum at $\kappa = c\kappa_*$, because we find that $\overline{\mathrm{CV}}(\kappa)$ approximates $\mathrm{AMISE}(\kappa)$ from Theorem 2. Therefore, it follows that

$$P(\overline{\mathrm{CV}}(c\kappa_*) < \overline{\mathrm{CV}}(\kappa_*)) \to 1, \tag{28}$$

as $n \to \infty$. From (25) and (28), then it holds that

$$\mathrm{MISE}(c\kappa_*) < \mathrm{MISE}(\kappa_*), \tag{29}$$

as $n \to \infty$. The contradiction between (27) and (29) completes the proof. □

Proof of Theorem 3. Let $U_{ij} = T_g^{(4)}(\Theta_i - \Theta_j)$, and $U_i = E_f[U_{ij}|\Theta_i]$. The expectation of $\hat{\psi}_4(g)$ is given by

$$E_f[\hat{\psi}_4(g)] = n^{-1}T_g^{(4)}(0) + 2n^{-2}\sum_{i<j} E_f[U_{ij}]. \tag{30}$$

It follows from (9) that

$$S_g^{(4)}(0) = 3g^2\left[S_g^{(2)}(0) + O(g^{-1})\right]. \tag{31}$$

Lemma 2 (Tsuruta and Sagae 2017) *The term $C_\kappa(L)$ is given by*

$$C_\kappa(L) = \kappa^{-1/2}2^{1/2}\mu_0(L) + O\left(\kappa^{-3/2}\right).$$

By combining (31) and Lemma 2, we find that the first term on the right side of (30) is equal to

$$n^{-1}T_g^{(4)}(0) = \frac{3g^{5/2}\left[S_g^{(2)}(0) + O(g^{-1})\right]}{2^{1/2}\mu_0(S)n}. \tag{32}$$

Lemma 3 (Tsuruta and Sagae 2017) *We set $\alpha_j(K_\kappa) := \int_{-\pi}^{\pi} K_\kappa(\theta)\theta^j \mathrm{d}\theta$. The terms $\alpha_{2t}(K_\kappa)$ for $t = 1, 2$ are given by*

$$\alpha_2(K_\kappa) = 2\mu_0^{-1}(L)\mu_2(L)\kappa^{-1} + O\left(\kappa^{-2}\right),$$

and $\alpha_4(K_\kappa) = O(\kappa^{-2})$. Lemma 2 in Tsuruta and Sagae (2017) presents the general form of $\alpha_{2t}(K_\kappa)$.

It follows from Lemma 3 that

$$\begin{aligned} U_i &= \int_{-\pi}^{\pi} T_g(\theta_j - \Theta_i)f^{(4)}(\theta_j)\mathrm{d}\theta_j \\ &= f^{(4)}(\Theta_i) + f^{(6)}(\Theta_i)\alpha_2(T_g)/2 + O(\alpha_4(T_g)) \\ &= f^{(4)}(\Theta_i) + f^{(6)}(\Theta_i)\mu_0^{-1}(S)\mu_2(S)g^{-1} + O(g^{-2}). \end{aligned} \tag{33}$$

$E_f[U_{ij}]$ in (30) is given by the expectation of (33) over Θ_i.

$$E_f[U_{ij}] = E_f[U_i] = \psi_4 + \mu_0^{-1}(S)\mu_2(S)\psi_6 g^{-1} + O(g^{-2}). \tag{34}$$

We obtain the bias (13) from combining (30), (32), and (34).

We now derive the variance of $\hat{\psi}_4(g)$. We set $W_{ij} := U_{ij} - U_i - U_j + E_f[U_i]$ and $Z_i := U_i - E_f[U_i]$. Then, we obtain $E_f[W_{ij}] = 0$, $E_f[Z_i] = 0$, and $\mathrm{Cov}_f[Z_i W_{ij}] = 0$. By using W_{ij} and Z_i, we present $\hat{\psi}_4(g) - E_f[\hat{\psi}_4(g)]$ as

$$\hat{\psi}_4(g) - E_f[\hat{\psi}_4(g)] = \frac{2(n-1)}{n^2}\sum_i Z_i + \frac{2}{n^2}\sum_{i<j} W_{ij}. \tag{35}$$

(35) shows that the variance of $\hat{\psi}_4$ is equal to

$$\mathrm{Var}_f[\hat{\psi}_4(g)] = \frac{4(n-1)^2}{n^4}\sum_i \mathrm{Var}_f[Z_i] + \frac{4}{n^4}\sum_{i<j}\mathrm{Var}_f[W_{ij}]. \tag{36}$$

By combining (33) and (34), $\mathrm{Var}_f[Z_i]$ reduces to

$$\begin{aligned}\mathrm{Var}_f[Z_i] &= E_f[U_i^2] - E_f[U_i]^2 \\ &= \mathrm{Var}_f[f^{(4)}(\Theta_i)] + o(1).\end{aligned} \tag{37}$$

By considering (34), $E_f[U_{ij}^2] = g^{9/2}[G_{1,0}(S_4)\psi_0 + o(1)]$, and $E_f[U_i^2] = E_f[U_i]^2 = O(1)$ ("Appendix D" in ESM provides the details of $E_f[U_{ij}^2]$ and $E_f[U_i^2]$), we obtain $\mathrm{Var}_f[W_{ij}]$. That is,

$$\begin{aligned}\mathrm{Var}_f[W_{ij}] &= E_f[U_{ij}^2] - 2E_f[U_i^2] + E_f[U_i]^2 \\ &= g^{9/2}[G_{1,0}(S_4)\psi_0 + o(1)].\end{aligned} \tag{38}$$

We obtain (11) from combining (36) (37), and (38). □

Proof of Theorem 4. If n is large, it follows from Lemma 3 that

$$\mathrm{CV}(\kappa) \simeq \frac{d(L)\kappa^{1/2}}{n} + \frac{2}{n^2}\sum_{i<j}\gamma(y_{ij}). \tag{39}$$

The derivative of (39) is given by

$$\frac{\mathrm{dCV}(\kappa)}{\mathrm{d}\kappa} \simeq \frac{d(L)}{2n\kappa^{1/2}} + \frac{2}{n^2\kappa^{1/2}}\sum_{i<j} V_{ij}, \tag{40}$$

where $V_{ij} := \kappa^{-1/2}[\gamma(y_{ij}) + \rho(y_{ij}) + 3/4\mu_0^{-1}(L)\mu_2(L)\kappa^{-1}\tau(y_{ij})]$, $\phi_\kappa(y_{ij}) := \kappa C_\kappa^{-1}(L)\frac{\mathrm{d}}{\mathrm{d}\kappa}L_\kappa(y_{ij})$, $\rho(y_{ij}) := K_\kappa(y_{ij}) + \int_{-\pi}^{\pi}\{\phi_\kappa(w)K_\kappa(w+y_{ij}) + K_\kappa(w)\phi_\kappa(w+y_{ij})\}\mathrm{d}w - 2\phi_\kappa(y_{ij})$, and $\tau(y_{ij}) := \int_{-\pi}^{\pi}K_\kappa(w)K_\kappa(w+y_{ij})\mathrm{d}w - K_\kappa(y_{ij})$.

"Appendix E" in ESM provides the details. The selector $\hat{\kappa}_{\mathrm{CV}}$ satisfies $\mathrm{dCV}(\kappa)/\mathrm{d}\kappa\ |_{\kappa=\hat{\kappa}_{\mathrm{CV}}} = 0$. This is equivalent to

$$2n^{-2}\sum_{i<j}V_{ij}\Big|_{\kappa=\hat{\kappa}_{\mathrm{CV}}} = -d(L)/(2n). \tag{41}$$

Note that $V_i := E_f[V_{ij}|\Theta_i]$. Then, we set $H_{ij} := V_{ij} - V_i - V_j + E_f[V_i]$ and $X_i := V_i - E_f[V_i]$. Then, we rewrite $2n^{-2}\sum_{i<j}\{V_{ij} - E_f[V_{ij}]\}$ as

$$2n^{-2}\sum_{i<j} V_{ij} - 2n^{-2}\sum_{i<j} E_f[V_{ij}] \simeq 2n^{-1}\sum_{i} X_i + 2n^{-2}\sum_{i<j} H_{ij},$$

where $2n^{-2}\sum_{i<j} H_{ij}$ is the degenerate U-statistic. We obtain the asymptotic normality for $2n^{-1}\sum_i X_i$ from the standard Central Limit Theorem (CLT). That is,

$$\frac{2}{n}\sum_{i} X_i \xrightarrow{d} N\left(0, Bn^{-1}\kappa^{-5}\right), \tag{42}$$

where, $B := 16\mu_2^4(L)\{R(f^{(4)}f^{1/2}) - R(f'')^2\}/\{\mu_0^4(L)\}$. "Appendix F" in ESM presents the details.

We give the definition of a degenerate U-statistic. A U-statistic is defined as $U_n := \sum_{i<j} H_{ij}$, where $H_{ij} := H(\Theta_i, \Theta_j)$ and H_{ij} is symmetric and $E_f[H_{ij}] = 0$. Let the degenerate U-statistic be the U-statistic satisfying $E_f[H_{ij}|\Theta_i] = 0$. The following lemma describes the asymptotic normality of a degenerate U-statistic.

Lemma 4 (Hall 1984) *Assume that H_{ij} is symmetric, and $E_f[H_{ij}|\Theta_i] = 0$, almost surely and $E_f[H_{ij}^2] < \infty$ for each n. We set $G_{ij} := E_f[H_{ii}H_{ij}]$. if*

$$\left\{E_f[G_{ij}^2] + n^{-1}E_f[H_{ij}^4]\right\}/E_f[H_{ij}^2]^2 \to 0, \tag{43}$$

as $n \to \infty$, then,

$$\sum_{1\leq i<j\leq n} H_{ij} \xrightarrow{d} N(0, n^2 E_f[H_{ij}^2]/2).$$

We obtain the asymptotic normality for $2n^{-2}\sum_{i<j} H_{ij}$ from Lemma 4. that is,

$$\frac{2}{n^2}\sum_{i<j} H_{ij} \xrightarrow{d} N(0, 2n^{-2}\kappa^{-1/2}M_{1,0}(L)R(f)). \tag{44}$$

See "Appendix G" in ESM for details. We combine (42) and (44) to derive the asymptotically normal for $2n^{-2}\sum_{i<j} V_{ij}$ as

$$\frac{2}{n^2}\sum_{i<j} V_{ij} \xrightarrow{d} N\left(-2R(f'')\mu_0^{-2}(L)\mu_2^2(L)\kappa^{-5/2}, \sigma_1^2\right), \tag{45}$$

where $\sigma_1^2 := Bn^{-1}\kappa^{-5} + 2n^{-2}\kappa^{-1/2}M_{1,0}(L)R(f)$. We take $\kappa = \hat{\kappa}_{\mathrm{CV}}$ in (45). Then, we replace $\hat{\kappa}_{\mathrm{CV}}$ in the variance to κ_* by Corollary 1. Thus, it follows from combining (41) and (45) that

$$-2R(f'')\mu_0^{-2}(L)\mu_2^2(L)\hat{\kappa}_{\mathrm{CV}}^{-5/2} \xrightarrow{d} N\left(-d(L)/(2n), \sigma_2^2\right), \tag{46}$$

where $\sigma_2^2 := Bn^{-1}\kappa_*^{-5} + 2n^{-2}\kappa_*^{-1/2}M_{1,0}(L)R(f)$. We ignore the first term for the variance of (46), because the convergence rate of the first term is $O(n^{-3})$, and that of the second term is $O(n^{-11/5})$ using $\kappa_* = O(n^{2/5})$. From (3), we obtain $R(f'')\mu_2^2(L)n/(d(L)\mu_0(L)) = \kappa_*^{5/2}$. Thus, (46) reduces to

$$(\hat{\kappa}_{\mathrm{CV}}/\kappa_*)^{-5/2} \xrightarrow{d} N\left(1, 8d(L)^{-2}M_{1,0}(L)R(f)\kappa_*^{1/2}\right). \tag{47}$$

Let $g(x) = x^{-5/2}$. Then, it follows that $g(1) = 1$ and $\{g'(1)\}^2 = 25/4$. We obtain the asymptotic normality for $\hat{\kappa}_{\mathrm{CV}}/\kappa_*$ by applying the delta method to (47). That is,

$$\hat{\kappa}_{\mathrm{CV}}/\kappa_* \xrightarrow{d} N\left(1, 50d(L)^{-2}M_{1,0}(L)R(f)\beta(L)^{-1/2}R(f'')^{-1/5}n^{-1/5}\right). \tag{48}$$

Theorem 4 completes the proof from (48). □

Proof of Theorem 5. The Taylor expansion $\hat{\kappa}_{\mathrm{PI}} = \hat{\kappa}_{\mathrm{PI}}(\hat{\psi}_4(g_*))$ is given by

$$\begin{aligned}\hat{\kappa}_{\mathrm{PI}}\left(\hat{\psi}_4(g_*)\right) &\simeq \beta(L)n^{2/5}\psi_4^{2/5} + \frac{2}{5}\beta(L)n^{2/5}\psi_4^{-3/5}(\hat{\psi}_4(g_*) - \psi_4)\\ &= \kappa_*\left[1 + 2(\hat{\psi}_4(g_*) - \psi_4)/(5\psi_4)\right].\end{aligned} \tag{49}$$

(49) reduces to

$$\hat{\kappa}_{\mathrm{PI}}/\kappa_* - 1 = \frac{2}{5\psi_4}\left(\hat{\psi}_4(g_*) - \psi_4\right). \tag{50}$$

Noting $W_{ij} := U_{ij} - U_i - U_j + E_f[U_i]$, and $Z_i := U_i - E_f[U_i]$, it follows that (35) becomes

$$\hat{\psi}_4(g) - E_f[\hat{\psi}_4(g)] \simeq 2n^{-1}\sum_i Z_i + 2n^{-2}\sum_{i<j} W_{ij}, \tag{51}$$

where $2n^{-2}\sum_{i<j} W_{ij}$ is the degenerate U-statistic. From (37), we obtain the asymptotic normality distribution from the standard CLT. That is,

$$n^{-1/2}\sum_i Z_i \xrightarrow{d} N(0, \mathrm{Var}_f[f(\Theta_i)]). \tag{52}$$

If we choose $g_* = W(S)n^{2/7}$, then applying Lemma A.4 to $2n^{-2}\sum_{i<j} W_{ij}$ gives

$$\frac{2}{n^2}\sum_{i<j} W_{ij} \xrightarrow{d} N\left(0, 2n^{-2}g_*^{9/2}G_{1,0}(S_4)\psi_0\right), \tag{53}$$

as $n \to \infty$. "Appendix H" in ESM presents the details. By combining (52) and (53), we obtain the asymptotic distribution of (51). That is,

$$\hat{\psi}_4(g_*) - E_f\left[\hat{\psi}_4(g_*)\right] \xrightarrow{d} N\left(0, 4n^{-1}\mathrm{Var}_f[f(\Theta_i)] + 2n^{-2}g_*^{9/2}G_{1,0}(S_4)\psi_0\right). \tag{54}$$

Theorem 3 shows that the rate of $\mathrm{Var}_f[\hat{\psi}_4(g^*)]$ is the order $n^{-5/7}$. Thus, (54) reduces to

$$n^{5/14}\left\{\hat{\psi}_4(g_*) - E_f\left[\hat{\psi}_4(g_*)\right]\right\} \xrightarrow{d} N\left(0, 2W^{9/2}(S)G_{1,0}(S_4)\psi_0\right). \tag{55}$$

The main term $\hat{\psi}_4(g_*) - \psi_4$ on the right side for (50) is equivalent to

$$n^{5/14}\left\{\hat{\psi}_4(g_*) - \psi_4\right\} = n^{5/14}\left\{\hat{\psi}_4(g_*) - E_f\left[\hat{\psi}_4(g_*)\right]\right\} - n^{5/14}\mathrm{Bias}_f\left[\hat{\psi}_4(g_*)\right]. \tag{56}$$

We show that $\mathrm{Bias}_f[\hat{\psi}_4(g^*)] = O(n^{-4/7})$ from Corollary 2. Then, we obtain that $n^{5/14}\mathrm{Bias}_f[\hat{\psi}_4(g_*)]$ is $O(n^{-3/14})$. Thus, if n is large, then this term is ignored. Therefore, the asymptotic normal distribution for $n^{5/14}\{\hat{\psi}_4(g_*) - \psi_4\}$ is given by

$$n^{5/14}\{\hat{\psi}_4(g) - \psi_4\} \xrightarrow{d} N(0, 2W^{9/2}(S)G_{1,0}(S_4)\psi_0). \tag{57}$$

Therefore, as $n \to \infty$, Theorem 5 completes the proof from (50) and (57). □

References

Abe, T., Pewsey, A. (2011). Sine-skewed circular distributions. *Statistical Papers*, *52*, 683–707.

Agostinelli, C., Lund, U. (2017). R package 'circular': Circular Statistics (version 0.4-93). https://r-forge.r-project.org/projects/circular/.

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

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C. C. (2017). Nonparametric estimating equations for circular probability density functions and their derivatives. *Electronic Journal of Statistics*, *11*, 4323–4346.

Di Marzio, M., Fensore, S., Panzera, A., Taylor, C. C. (2018). Circular local likelihood. *Test*, 1–25.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, *14*, 1–16.

Hall, P., Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, *74*, 567–581.

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

Mardia, K. V., Jupp, P. E. (1999). *Directional statistics*, p. 40. London: Wiley.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Sheather, S. J., Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*, 683–690.

Tsuruta, Y., Sagae, M. (2017). Higher order kernel density estimation on the circle. *Statistics and Probability Letters*, *131*, 46–50.

Wand, M. P., Jones, M. C. (1994). *Bandwidth selection. Kernel smoothing*, pp. 58–88. USA: Chapman&Hall/CRC.