NONPARAMETRIC CIRCULAR METHODS FOR DENSITY AND REGRESSION

María Oliveira Pérez
Departamento de Estatística e Investigación Operativa
Universidade de Santiago de Compostela

PhD Dissertation

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Todas as cousas
son imposibles,
mentres o parecen.
Concepción Arenal

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Prof. Dra. Rosa María Crujeiras Casais e Prof. Dr. Alberto Rodríguez Casal, do Departamento de Estatística e Investigación Operativa da Universidade de Santiago de Compostela, informan que a memoria titulada:

NONPARAMETRIC CIRCULAR METHODS FOR DENSITY AND REGRESSION

foi realizada baixo a súa dirección por Dona María Oliveira Pérez, estimando que a interesada se atopa en condicións de optar ao Grao de Doutor en Estatística e Investigación Operativa, polo que solicitan que sexa admitida a trámite para a súa lectura e defensa pública na Facultade de Matemáticas da Universidade de Santiago de Compostela.

En Santiago de Compostela, a 11 de setembro de 2013.

Os directores:

Prof. Dra. Rosa María Crujeiras Casais

Prof. Dr. Alberto Rodríguez Casal

A doutoranda: María Oliveira Pérez

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

Introduction

This essay is focused on the analysis of circular data. Circular data are a particular case of directional data where observations are directions in two dimensions. As its name suggests, circular data can be represented as points on the circumference of a unit circle centered at the origin and are usually expressed as angles. Hence, in order to define a circular observation, an initial direction and a sense of rotation must be chosen. The data analysis cannot depend on these ad hoc choices. Moreover, circular data have a periodic structure. All these features make them different from the usual linear data and so, it seems obvious that such data cannot be treated in the same way. Some general references on circular statistics, or more general directional data analysis, are Mardia (1972), Batschelet (1981), Fisher (1993), Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001).

In recent years, there has been an increasing interest in directional statistics since this kind of data appears in a large variety of disciplines such as biology (in the study of orientation of animals), meteorology (when analyzing wind direction) or environmental sciences (in the study of directions of ocean currents), among others. Two of the most fundamental problems in statistics are knowing how a circular random variable is distributed (density estimation) and how it is related with other variable (regression estimation). In our case, how can be it used to model the behaviour of a scalar response.

Density and regression estimation can be approached from a parametric or from a nonparametric perspective. The parametric approach assumes that data are drawn from a known parametric model and the problem is reduced to estimate its parameters. The nonpametric approach does not rely on such somewhat restrictive assumptions and "let the data speak for themselves". In most of the applied papers dealing with circular data, it is only considered the use of circular descriptive techniques providing graphical displays of the data and classical parametric inferential tools (see, e.g., Aradóttir et al., 1997; Mooney et al., 2003; Corcoran et al., 2009). Thus, our interest is focused on the analysis of circular data from a nonparametric perspective. Concretely, the main aim of this dissertation is to propose and analyze nonparametric circular methods for density and regression estimation.

In this context, nonparametric density estimation was approached by Beran (1979), Hall et al. (1987) and Bai et al. (1988) who studied the circular kernel density estimator for the general

case of directional data. In the regression setting, nonparametric methods involving circular data have been studied by Di Marzio et al. (2009), Qin et al. (2011a,b) and Di Marzio et al. (2012) who proposed kernel smoothers. Periodic smoothing splines introduced by Cogburn and Davis (1974) provide an alternative to kernel estimators when the predictor is a circular variable and the response is linear.

A critical issue in any nonparametric procedure is the smoothing parameter selection. Although, some procedures have been proposed by Hall et al. (1987), Taylor (2008) and Di Marzio et al. (2011) for circular kernel density estimation, the main contribution in this setting will be the introduction of a new smoothing parameter selector that performs well in very different and complex distributional scenarios. For regression estimation with a linear response and a circular covariate, cross–validation rules are suggested in order to select the smoothing parameter, both for kernel and smoothing spline estimators.

Another important problem in the use of smoothing methods is whether or not observed features in the smoothed curve, such as peaks and valleys are really there, as opposed to being artifacts of the natural sampling variability. For linear data, the SiZer method developed by Chaudhuri and Marron (1999) both for density and regression estimation allows for the assessment of statistically significant features in a smoothed curve and moreover, it avoids the problem of selecting a smoothing parameter. However, nothing similar exists for circular data. With this goal, a new method namely CircSiZer, conveniently adapted to the circular nature of the variables will be introduced in this dissertation.

The manuscript is organized in the following way:

Chapter 1 is devoted to the introduction of circular data, revising some circular distributions and studying the parameter estimation for the von Mises distribution and for mixtures of these distributions. In this chapter, the datasets that motivate and illustrate the methods proposed along the manustript are described.

Chapter 2 is focused on nonparametric curve estimation. The estimation of the density function is addressed in the first part where the kernel density estimator for circular data is introduced, revising and proposing different methods for selecting the smoothing parameter and checking their behaviour in a simulation study. The second part of this chapter is focused on nonparametric regression estimation for a circular explanatory variable and a linear response. This problem is approached by using kernel smoothers and periodic smoothing splines, which are compared in a simulation study. In both settings, density and regression, the techniques proposed are illustrated with classical data examples and applied to analyze some real datasets.

In Chapter 3, in order to assess the significance of the features observed in the smoothed curves, both for density and regression, a SiZer (Significative Zero crossing of the derivative) technique is developed for circular data, namely CircSiZer. The performance of CircSiZer is illustrated with simulated data and finally, the method is used for analyzing some real datasets.

In Chapter 4, a new library in R, namely `NPCirc`, which implements the nonparametric kernel

methods for density and regression estimation for circular data studied in the previous chapters is described. The library includes most of the self–programmed code which has been implemented for applying the proposed methods in practice.

Finally, the manuscript includes four appendices. In Appendix A, the circular density models used in the simulation study in Chapter 2 and for illustration throughout the manuscript are defined. Appendices B and C give technical details on kernel regression smoothers and periodic smoothing splines, which complement Chapter 2 and 3. Appendix D describes the functions in the `NPCirc` library, giving instructions about their usage and arguments and illustrating them with examples.

I would like to thank my advisors, Prof. Rosa M. Crujeiras and Prof. Alberto Rodríguez Casal for their work and support during these years. I also wish to thank Prof. Augusto Pérez–Alberti from the Department of Physical Geography of the University of Santiago de Compostela and Dr. Kenneth A. Mann from the Upstate Medical University (New York) for kindly providing some real datasets that motivate part of the work done in this thesis.

This work has been supported by Project MTM2008–03010 from the Spanish Ministry of Science and Innovation, and by the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences), from Belgian Science Policy.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Circular models and data

1.1 Introduction

The analysis of circular data appears in many applied fields, such as biology (Batschelet, 1981), ecology (Jammalamadaka and Lund, 2006), meteorology (Bowers et al., 2000), sociology (Brunsdon and Corcoran, 2005), medicine (Mooney et al., 2003) or biomechanics (Mann et al., 2003), among others.

A circular observation can be defined as a point on a circle of unit radius, or a unit vector in two dimensions (i.e., a direction in the plane). Hence, once an initial direction and orientation of the circle have been chosen, each circular observation can be specified by the angle from the initial direction to the point on the circle corresponding to the observation. More generally, a unit vector in a d–dimensional space ($d \geq 2$) is called a directional data or a spherical data.

Directional data in general, and circular data in particular, have special features both in terms of models and in terms of their statistical treatment. For instance, the numeric representation of a circular obsevation, as an angle or a unit vector, is not necessarily unique since it depends of the initial direction and the sense of rotation. Because of this, another feature of circular data is that there is not a natural ordering of the observations. Moreover, circular data are periodic, i.e., if θ is an angle in the interval $[0, 2\pi)$ then θ can be also represented by $(\theta + 2\pi k)$ for any integer k. Thus, the methods for dealing with circular data must take into account these features and so, standard real–line methods are not appropriate for analyzing this kind of data.

Circular data analysis has been approached from parametric and nonparametric perspectives, existing a broad literature on parametric methods, both for density and regression. Comprehensive reviews such as Mardia (1972), Fisher (1993), Mardia and Jupp (2000), Jammalamadaka and SenGupta (2001) and Lee (2010) present parametric density models such as the von Mises, the cardiod, or the wrapped distributions, among others, testing procedures for assessing uniformity, such as Rayleigh's, Kuiper's, Rao's Spacing or Watson's tests, jointly with different correlation coefficients and parametric regression models for two circular variables or a circular and a linear

variable.

Some background on circular parametric models and some real data examples are presented in this chapter. Section 1.2 is devoted to the introduction the concept of circular distribution and give a brief overview on the most important circular parametric distribution families, such as the classical von Mises distribution, the cardioid distribution, some wrapped distributions and mixtures of circular distributions. Parameter estimation will be studied for the von Mises distribution and mixture of them. For mixtures, the associated Expectation Maximization (EM) algorithm for carrying out maximum likelihood estimation will be detailed. Finally, in Section 1.3, some real datasets that motivate this dissertation will be described. Moreover, these datasets will be considered along the manuscript for illustration purposes.

1.2 Circular parametric distributions

A circular distribution is a probability measure supported on the unit circle. Each point in the circumference represents a direction and so, a circular distribution assigns probabilities to different directions. As for linear data, the distribution of a circular variable can be absolutely continuous. In this case, one way of specifying a distribution on the unit circle is by means of its density function. From now on, any circular random variable Θ will be measured in radians and its support will be the interval $[0, 2\pi)$. Hence, the probability density function is a function f defined for each angle $\theta \in [0, 2\pi)$ satisfying the conditions:

(i) $f(\theta) \geq 0$, $\theta \in [0, 2\pi)$;

(ii) $\int_0^{2\pi} f(\theta)d\theta = 1$;

(iii) $f(\theta) = f(\theta + 2\pi k)$, $\forall \theta \in [0, 2\pi)$ and any integer k, i.e., f is periodic with period 2π;

(see Jammalamadaka and SenGupta, 2001, Section 2.1).

Any such function describes a probability distribution on the circle. Let θ_1 and θ_2 be fixed angles with $0 \leq \theta_1 \leq \theta \leq \theta_2 \leq 2\pi$, then

$$\mathbb{P}\left[\theta_1 \leq \Theta \leq \theta_2\right] = \int_{\theta_1}^{\theta_2} f(\theta)d\theta.$$

As for linear variables, the characteristic function determines the distribution. The value of the characteristic function $\varphi(t) = \mathbb{E}(e^{it\Theta})$ at an integer r is also called the r–th trigonometric moment of Θ which is given by:

$$\mathbb{E}(e^{ir\Theta}) = \int_0^{2\pi} e^{ir\theta} f(\theta)d\theta = \int_0^{2\pi} \cos(r\theta)f(\theta)d\theta + i\int_0^{2\pi} \sin(r\theta)f(\theta)d\theta, \quad r = 0, \pm 1, \pm 2, \ldots$$

where i denotes the imaginary unit. Concretely, the first trigonometric moment, expressed in polar coordinates, is:

$$\mathbb{E}(e^{i\theta}) = \rho e^{i\mu},$$

where ρ is the mean resultant length and μ is the mean direction.

The most simple absolutely continuous distribution on the circle is the circular uniform distribution which assigns the same probability to all the directions. When a distribution is not uniform, this may be concentrated around one or more directions. In that case, the distribution is said unimodal or multimodal, respectively.

An appropriate measure of the mean direction for a set of directions which are unimodal is obtained by treating the data as unit vectors and computing the direction of their average resultant vector. Let $\theta_1, \ldots, \theta_n$ be a set of circular observations given in terms of angles. The sample mean direction $\bar{\theta}$ is then given by

$$\bar{\theta} = \arg\left\{ \frac{1}{n}\sum_{j=1}^{n} \cos\theta_j + i\frac{1}{n}\sum_{j=1}^{n} \sin\theta_j \right\}.$$

where arg denotes the function which returns the argument of a complex number. More explicitly, using the notation $C = \sum_{j=1}^{n} \cos\theta_j$ and $S = \sum_{j=1}^{n} \sin\theta_j$, the sample mean direction is given by

$$\bar{\theta} = \arctan^*\left(\frac{S}{C}\right) = \begin{cases} \arctan(S/C) & \text{if} \quad C > 0, S \geq 0 \\ \pi/2 & \text{if} \quad C = 0, S > 0 \\ \arctan(S/C) + \pi & \text{if} \quad C < 0 \\ \arctan(S/C) + 2\pi & \text{if} \quad C \geq 0, S < 0 \\ \text{undefined} & \text{if} \quad C = 0, S = 0 \end{cases} \tag{1.1}$$

where the inverse tangent function arctan takes values in $[-\pi/2, \pi/2]$ and so, arctan* takes values in $[0, 2\pi)$.

Moreover, the length of the average resultant vector, denoted by $\bar{R}$, provides a measure of concentration of the data. If all angles are identical, then $\bar{R} = 1$ and if data are widely dispersed then $\bar{R}$ will be almost 0.

Circular uniform distribution

This distribution has a constant density

$$f(\theta) = \frac{1}{2\pi}, \quad 0 \leq \theta < 2\pi,$$

i.e., all directions are equally likely.

Some unimodal distributions are the cardioid, von Mises and some wrapped distributions such as the wrapped normal, the wrapped Cauchy and the wrapped skew–normal distribution. These models are characterized by at least two parameters, one defining the location or reference direction and other defining the dispersion about that location.

Cardioid distribution

This distribution was introduced by (Jeffreys, 1961, p. 328). The cardioid distribution, $C(\mu, \rho)$ is a perturbation of the uniform density by a cosine funtion whose density function is:

$$f(\theta; \mu, \rho) = \frac{1}{2\pi}\left(1 + 2\rho\cos(\theta - \mu)\right), \quad 0 \leq \theta < 2\pi, |\rho| \leq \frac{1}{2}.$$

The mean resultant length of $C(\mu, \rho)$ is ρ and (if $\rho > 0$) the mean direction is μ. When $\rho = 0$, the cardioid distribution reduces to the circular uniform distribution. This is a symmetric and unimodal distribution around μ.

von Mises distribution

The von Mises distribution is also known as circular normal distribution since it was derived in a way analogous to the derivation of the normal distribution in the real line. von Mises (1918) asked about the existence of a circular model verifying that: for any sample of circular data, the maximum likelihood estimator of the location parameter is equal to the sample mean. Hence, the von Mises distribution was obtained.

The von Mises distribution with parameters μ and κ, $vM(\mu, \kappa)$, is a symmetric and unimodal distribution with density

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)}, \quad 0 \leq \theta < 2\pi, \tag{1.2}$$

where I_0 denotes the modified Bessel function of the first kind and order zero which is defined as follows

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta} d\theta.$$

Here, $I_0(\kappa)$ is a normalizing constant. The parameter μ $(0 \leq \mu < 2\pi)$ is the mean direction and where the mode is located. The parameter κ $(\kappa \geq 0)$ is known as the concentration parameter since the ratio of the density at the mode μ to the density at the antimode $(\mu + \pi)$, where the minimum density is reached, is given by $e^{2\kappa}$ and so, as κ increases the distribution becomes more concentrated around μ. When $\kappa = 0$, the distribution reduces to the circular uniform distribution.

Wrapped distributions

Circular distributions can be obtained by wrapping linear distributions onto the circle of unit radius. If X is a random variable on the real line, this variable may be transformed to a circular random variable Θ by reducing its modulo 2π, i.e.,

$$\Theta = X (\text{mod } 2\pi).$$

Hence, if X has density function g, the density function of Θ is given by

$$f(\theta) = \sum_{k=-\infty}^{\infty} g(\theta + 2\pi k), \quad 0 \leq \theta < 2\pi,$$

(see Jammalamadaka and SenGupta, 2001, p. 31).

The wrapped normal and wrapped Cauchy distributions, described below, are some examples of the distributions that can be obtained in this way.

Wrapped normal distribution

The wrapped normal distribution, $WN(\mu, \rho)$, is obtained by wrapping the $N(\mu, \sigma^2)$ distribution onto the circle, where $\rho = e^{-\sigma^2/2}$. Its probability density function is given by

$$f(\theta; \mu, \rho) = \frac{1}{2\pi}\left(1 + 2\sum_{k=1}^{\infty} \rho^{k^2} \cos p(\theta - \mu)\right), \quad 0 \leq \theta < 2\pi,$$

where μ $(0 \leq \mu < 2\pi)$ is the mean direction and ρ is the mean resultant length. Note that, when ρ goes to zero the distribution becomes less concentrated aroud the mean direction. This distribution is unimodal and symmetric about its mode μ.

Wrapped Cauchy distribution

The wrapped Cauchy distribution, $WC(\mu, \rho)$, is obtained by wrapping the Cauchy distribution on the real line with density

$$g(x; \mu, \sigma^2) = \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (x - \mu)^2}, \quad -\infty < x < \infty$$

onto the circle. Its density function has the following expression

$$f(\theta; \mu, \rho) = \frac{1}{2\pi}\frac{1 - \rho^2}{1 + \rho^2 - 2\rho\cos(\theta - \mu)}, \quad 0 \leq \theta < 2\pi,$$

where $0 \leq \mu < 2\pi$ is the mean direction and $\rho = e^{-\sigma}$ is the mean resultant length. It is also a unimodal and symmetric distribution around its mode μ.

Figure 1.1 shows the von Mises, cardioid, wrapped normal and wrapped Cauchy densities with the same values for the parameters μ and ρ. Among them, the wrapped Cauchy distribution is noted by its peaked mode.

However, not all circular distributions are unimodal and symmetric. Pewsey (2000) defined a unimodal and asymmetric distribution, namely the wrapped skew–normal distribution.

Wrapped skew–normal distribution

The wrapped skew–normal distribution is a skewed distribution characterized by a location parameter μ $(0 \leq \mu < 2\pi)$, a scale parameter κ $(\kappa > 0)$ and a skewness parameter λ $(\lambda \geq 0)$. Its density function is given by

$$f(\theta; \mu, \kappa, \lambda) = \frac{2}{\kappa}\sum_{r=-\infty}^{\infty} \phi\left(\frac{\theta - 2\pi r - \mu}{\kappa}\right) \Phi\left(\lambda\left(\frac{\theta + 2\pi r - \mu}{\kappa}\right)\right),$$

where ϕ and Φ denote the standard normal density and distribution functions, respectively. This distribution will be denoted by $WSN(\mu, \kappa, \lambda)$. The asymmetric shape of this distribution can be seen in Figure 1.2, for different values of λ.

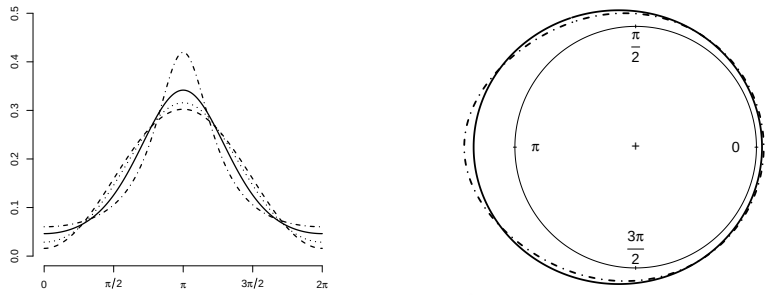

Figure 1.1: Left panel: linear representation of the density functions of $vM(\mu, A^{-1}(\rho))$ where A^{-1} denotes the inverse of function $A(\cdot) = I_1(\cdot)/I_0(\cdot)$ (solid line), $C(\mu, \rho)$ (dashed line), $WN(\mu, \rho)$ (dotted line) and $WC(\mu, \rho)$ (dotted–dashed line) with $\mu = \pi$ and $\rho = 0.45$. Right panel: circular representation of $vM(\mu, A^{-1}(\rho))$ (solid line) and $WC(\mu, \rho)$ (dotted–dashed line).

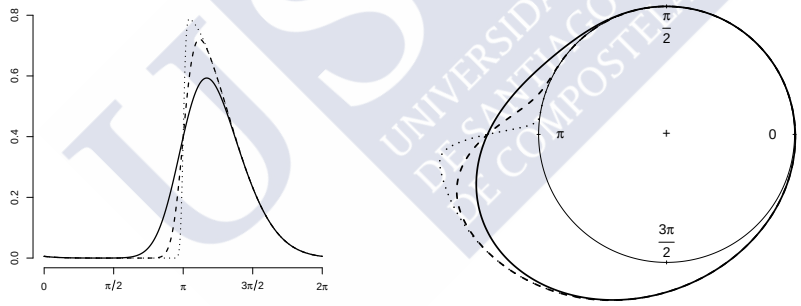

Figure 1.2: Linear (left panel) and circular (right panel) representations of a wrapped skew–normal distribution $WSN(\pi, 1, \lambda)$ with $\lambda = 2$ (solid line), $\lambda = 5$ (dashed line) and $\lambda = 20$ (dotted line).

Further details on these and other distribution models can be found in Fisher (1993), Jammalamadaka and SenGupta (2001), Mardia and Jupp (2000) and Pewsey (2000).

Although being widely used, the von Mises model and the other models presented may not be flexible enough to capture the underlying structure of multimodal, highly peaked or skewed distributions. Some new parametric models for handling these features have been presented by Abe and Pewsey (2011), who introduced circular models with two diametrically opposed modes, or Jones and Pewsey (2012), who proposed the inverse Batschelet distribution, accounting for skewness and high kurtosis (far from the nicely bell–shaped von Mises distributions). A more flexible model involving mixtures of von Mises distributions was used by Mooney et al. (2003).

The consideration of mixtures of parametric models may offer a route to capture complex structures, allowing multimodality and/or asymmetry.

Mixtures

A finite mixture of M circular distributions f_m, $m = 1, \ldots, M$, has density:

$$f(\theta) = \sum_{m=1}^{M} p_m f_m(\theta), \quad 0 \leq \theta < 2\pi,$$

where p_m are positive quantities that sum one ($p_m > 0$ and $\sum_{m=1}^{M} p_m = 1$) and the $f_m(\theta)$ are circular densities. The quantities $p_1, \ldots, p_m$ are known as weights or mixing proportions and the $f_m(\theta)$ are called the component densities of the mixture.

A particular case is the mixture of M von Mises distributions whose density is:

$$f(\theta; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) = \sum_{m=1}^{M} p_m f_m(\theta; \mu_m, \kappa_m), \quad 0 \leq \theta < 2\pi, \tag{1.3}$$

where $\boldsymbol{p} = (p_1, \ldots, p_M)$ with $p_m > 0$ and $\sum_{m=1}^{M} p_m = 1$ are the weights of the component densities, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M) \in [0, 2\pi)^M$ is the vector of circular means and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_M) \in (\mathbb{R}^+)^M$ is the vector of concentrations; f_m denotes the density function of a von Mises distribution $vM(\mu_m, \kappa_m)$, for $m = 1, \ldots, M$.

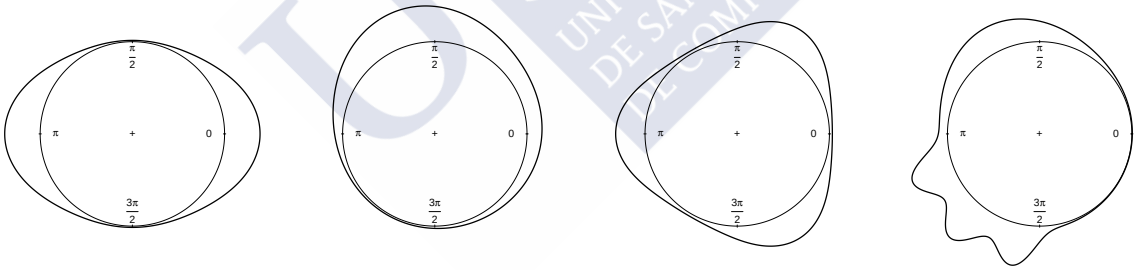

Figure 1.3: Mixtures of von Mises distributions with different number of components. These models correspond to models M7, M9, M11 and M19 defined in Appendix A.

Figure 1.3 shows four mixtures of von Mises distributions with different number of components, which present multimodality and asymmetry. For example, bimodality arises in the first model setting the mixture proportion to one–half and combining two confronted distributions whereas asymmetry is induced in the second model by considering different weights. The third plot shows a mixture of three von Mises distributions with equally spaced modes and the same concentration parameter. Finally, the last plot is an asymmetric model which is a mixture of five von Mises distributions and only shows four modes. These models will be used in the next chapter and the specific formulae is given in the Appendix A (models M7, M9, M11 and M19).

1.2.1 Parameter estimation for a von Mises distribution

In this section, the parameters of a von Mises distribution will be estimated by the method of moments and maximum likelihood. Although they are two classical techniques, the estimation procedure in the setting of circular data will be detailed since apart from its usefulness as parametric estimation methods, they will be needed in Chapter 2 for constructing smoothing parameter selectors in a nonparametric setting.

Let $\Theta_1, \Theta_2, \ldots, \Theta_n \in [0, 2\pi)$ be a random sample from $vM(\mu, \kappa)$.

Estimation by the method of moments

The modified Bessel function of the first kind and order r is defined as

$$I_r(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(r\theta) e^{\kappa \cos\theta} d\theta, \quad r = 0, 1, 2, \ldots$$

Since

$$\frac{1}{2\pi} \int_0^{2\pi} \sin(r\theta) e^{\kappa \cos\theta} d\theta = 0,$$

the moment of order r of the von Mises probability density is given by

$$\begin{aligned} \frac{1}{2\pi I_0(\kappa)} \int_0^{2\pi} e^{ir\theta} e^{\kappa \cos(\theta-\mu)} d\theta &= \frac{1}{2\pi I_0(\kappa)} \int_0^{2\pi} e^{ir(\omega+\mu)} e^{\kappa \cos\omega} d\omega \\ &= \frac{e^{ir\mu}}{2\pi I_0(\kappa)} \int_0^{2\pi} [\cos(r\omega) + i \sin(r\omega)] \, e^{\kappa \cos\omega} d\omega \\ &= \frac{I_r(\kappa)}{I_0(\kappa)} e^{ir\mu} = \frac{I_r(\kappa)}{I_0(\kappa)} [\cos(r\mu) + i \sin(r\mu)] . \end{aligned}$$

The r-th order sample moment is given by

$$\frac{1}{n} \sum_{j=1}^{n} e^{ir\Theta_j} = \frac{1}{n} \sum_{j=1}^{n} \cos(r\Theta_j) + i \frac{1}{n} \sum_{j=1}^{n} \sin(r\Theta_j).$$

By equating population moments with sample moments for $r = 1$, the equations that define the estimators $\tilde{\mu}$ and $\tilde{\kappa}$ of μ and κ by the method of moments are obtained:

$$A(\tilde{\kappa}) \cos\tilde{\mu} = \frac{1}{n} \sum_{j=1}^{n} \cos\Theta_j, \tag{1.4}$$

$$A(\tilde{\kappa}) \sin\tilde{\mu} = \frac{1}{n} \sum_{j=1}^{n} \sin\Theta_j, \tag{1.5}$$

where $A(\cdot) = I_1(\cdot)/I_0(\cdot)$.

If $\sum_{j=1}^{n} \cos\Theta_j \neq 0$, by dividing (1.4) by (1.5) results

$$\tilde{\mu} = \arctan^*\left(\frac{\sum_{j=1}^{n} \sin\Theta_j}{\sum_{j=1}^{n} \cos\Theta_j}\right), \tag{1.6}$$

where $\arctan^*$ is defined in equation (1.1). The method of moments estimator for μ is the direction of the sample mean direction.

By multipliying (1.4) by $\cos\tilde{\mu}$ and (1.5) by $\sin\tilde{\mu}$ and adding, the equation that defines the method of moments estimator for κ is given by:

$$A(\tilde{\kappa}) = \frac{1}{n}\sum_{j=1}^{n} \cos(\Theta_j - \tilde{\mu}). \tag{1.7}$$

So, as long as $\sum_{j=1}^{n} \cos\Theta_j \neq 0$, the estimator of κ, namely $\tilde{\kappa}$, is obtained by solving (1.7).

Estimation by maximum likelihood

The maximum likelihood estimators for μ and κ will be those values that maximize the likelihood function based on the observations $\Theta_1, \ldots, \Theta_n$, i.e.,

$$L(\mu, \kappa|\Theta_1, \ldots, \Theta_n) = \prod_{j=1}^{n} f(\Theta_j; \mu, \kappa) = \frac{1}{[2\pi I_0(\kappa)]^n} e^{\sum_{j=1}^{n} \kappa \cos(\Theta_j - \mu)},$$

or the log–likelihood function

$$\log L(\mu, \kappa|\Theta_1, \ldots, \Theta_n) = -n\log(2\pi I_0(\kappa)) + \kappa\sum_{j=1}^{n} \cos(\Theta_j - \mu). \tag{1.8}$$

From the above expression (1.8), by computing the partial derivatives with respect to μ and κ, respectively and equating to zero, the equations that define the maximum likelihood estimators for μ and κ, $\hat{\mu}$ and $\hat{\kappa}$, are:

$$\hat{\kappa}\sum_{j=1}^{n} \sin(\Theta_j - \hat{\mu}) = 0, \tag{1.9}$$

$$A(\hat{\kappa}) = \frac{1}{n}\sum_{j=1}^{n} \cos(\Theta_j - \hat{\mu}). \tag{1.10}$$

If $\sum_{j=1}^{n} \cos\Theta_j \neq 0$, the maximum likelihood estimator for μ is obtained from (1.9):

$$\hat{\mu} = \arctan^*\left(\frac{\sum_{j=1}^{n} \sin\Theta_j}{\sum_{j=1}^{n} \cos\Theta_j}\right)$$

and $\hat{\kappa}$ is obtained from the solution of equation (1.10). Note that, if $\hat{\kappa} = 0$, equation (1.9) is verified and a value for $\hat{\mu}$ can be obtained from the solution of (1.10). However, it can be shown that this solution $(\hat{\mu}, \hat{\kappa} = 0)$ is not a maximum of (1.8).

Thus, the maximum likelihood estimators for the parameters of a von Mises distribution are equal to the ones obtained by the method of moments (see equations (1.6) and (1.7)), i.e., $\tilde{\mu} = \hat{\mu}$ y $\tilde{\kappa} = \hat{\kappa}$, with probability one.

Figure 1.4 shows the boxplots of the differences in absolute value between the estimated and true values of the parameters of a $vM(\pi, 1)$. Parameters are estimated by maximum likelihood by using 1000 random samples of size $n = 100$ and $n = 500$ from that distribution. As it was expected, differences are smaller for the largest sample size. In order to obtain a better visualization, outliers are not plotted in the figure.

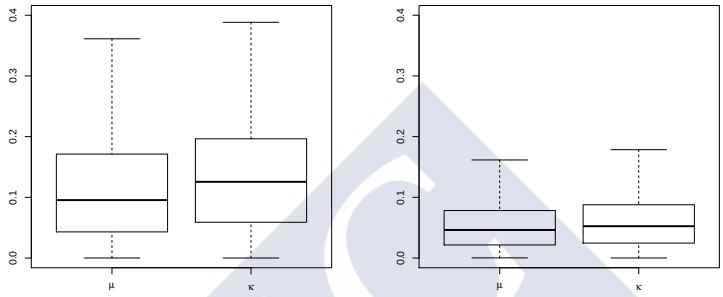

Figure 1.4: Boxplots of the differences in absolute value between the estimated and true values of the parameters of a von Mises distribution $vM(\pi, 1)$. Results were obtained from 1000 random samples of size $n = 100$ (left panel) and $n = 500$ (right panel).

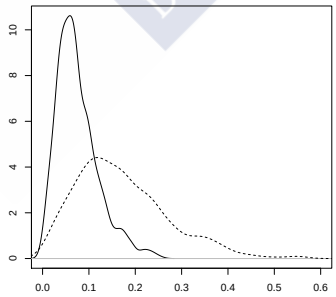

Figure 1.5: Density function[1] of the estimates of the concentration parameter of a von Mises distribution based on 1000 random samples of size $n = 100$ (dashed line) and $n = 500$ (solid line) from a mixture of two von Mises distribution (see model M7 in Appendix A for the specific formulae).

[1]The density functions were obtained by using the kernel density estimator for linear data defined at the beginning of Section 2.2 and taking as smoothing parameter the value selected by the rule of thumb of Silverman (1986, p. 47).

In this case, the model is correctly specified and the errors in the parameter estimation are small, so the parametric estimation of the density function should be right. But, what happens when the model is not well specified? For example, consider that the observations come from a mixture of two equally weighted von Mises distribution with diametrically opposed means and the same concentration parameter (such as model M7 defined in Appendix A) but, it is wrongly assumed that the underlying model is a von Mises distribution. When the concentration parameter of a von Mises distribution is estimated based on observations coming from model M7, it is observed that it takes values close to zero as shown in Figure 1.5 for 1000 samples of size $n = 100$ (dashed line) and $n = 500$ (solid line). Therefore, if the underlying model is parametrically estimated, assuming a von Mises model, the estimate is close to a circular uniform (concentration parameter close to 0), which is far from the real distribution. Before going into a purely nonparametric approach for density estimation, a parametric method for estimating mixtures of von Mises will be introduced in the next section.

1.2.2 Parameter estimation for a mixture of von Mises distributions

Let $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$ be a random sample of angles from a mixture of M von Mises distributions, as the one presented in (1.3). In order to estimate the parameters of the mixture, the log–likelihood function of the sample is computed

$$\log(L(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}|\Theta_1, \ldots, \Theta_n)) = \log\left(\prod_{i=1}^{n} f(\Theta_i; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})\right) = \sum_{i=1}^{n} \log\left(\sum_{m=1}^{M} p_m f_m(\Theta_i; \mu_m, \kappa_m)\right). \quad (1.11)$$

However, the log–likelihood function has a complex expression (it involves the logarithm of a sum) which is difficult to optimize. The main problem lies in the fact that it is not known which density component generates each observation. Assuming that this information is available, i.e., given a vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ such that Z_i takes the value m if Θ_i was generated by the m–th mixture component, then the log–likelihood would be

$$\log(L(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}|\Theta_1, \ldots, \Theta_n, \boldsymbol{Z})) = \sum_{i=1}^{n} \log\left(p_{Z_i} f_{Z_i}\left(\Theta_i; \mu_{Z_i}, \kappa_{Z_i}\right)\right), \quad (1.12)$$

which has an expression less complicated than (1.11).

For the sake of simplicity, consider the particular case of a mixture of two von Mises, $M = 2$. In this case, (1.12) becomes

$$\log(L(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}|\Theta_1, \ldots, \Theta_n, \boldsymbol{Z})) = \sum_{i;\ Z_i=1} \log\left(p_1 f_1\left(\Theta_i; \mu_1, \kappa_1\right)\right) + \sum_{i;\ Z_i=2} \log\left(p_2 f_2\left(\Theta_i; \mu_2, \kappa_2\right)\right). \quad (1.13)$$

From (1.13), by computing the partial derivatives with respect to μ_1 and κ_1 and equating them to zero, the equations that define the maximum likelihood estimators $\hat{\mu}_1$ and $\hat{\kappa}_1$ of μ_1 and κ_1 respectively, are obtained:

$$\hat{\kappa}_1 \sum_{i;\ Z_i=1} \sin\left(\Theta_i - \hat{\mu}_1\right) = 0, \quad (1.14)$$

$$A(\hat{\kappa}_1) = \frac{1}{n_1} \sum_{i;\ Z_i=1} \cos(\Theta_i - \hat{\mu}_1) = 0, \tag{1.15}$$

where n_1 is the cardinality of the set $\{i :\ Z_i = 1,\ i = 1, \ldots, n\}$. Note that, if only the data of the first component of the mixture are considered then, these equations are the same to those defined in (1.9) and (1.10).

If $\sum_{i;\ Z_i=1} \cos\Theta_i \neq 0$, from (1.14),

$$\hat{\mu}_1 = \arctan^* \left(\frac{\sum_{i;\ Z_i=1} \sin\Theta_i}{\sum_{i;\ Z_i=1} \cos\Theta_i} \right) \tag{1.16}$$

and $\hat{\kappa}_1$ is obtained from the solution of equation (1.15). Maximum likelihood estimators for μ_2 and κ_2 are obtained in the same way.

Since $p_1 + p_2 = 1$, equation (1.13) can be written in terms of p, where $p_1 = p$ and $p_2 = 1 - p$. Taking the derivative with respect to p and setting it equals to zero, the maximum likelihood estimator of p is

$$\hat{p} = \hat{p}_1 = n_1/n.$$

Therefore, the maximum likelihood estimate of the parameters of a mixture of two von Mises distributions is easy to obtain if it is known which density component generates each sample data Θ_i, i.e., if the values of the variable $\boldsymbol{Z}$ are known. However this information is often unknown.

In this latter case, when the sample data is incomplete, the EM algorithm provides a method for estimating the parameters of the mixture by maximum likelihood. The EM algorithm (Dempster et al., 1977) is an iterative proccess which applies two steps alternatively until the convergence:

- E–step: Compute the expected value of the complete data log–likelihood with respect to the conditional distribution of the non observed variable. In our context, the variable that indicates which density component generates each observation is the non observed variable.
- M–step: Estimate the parameters by maximizing the expectation computed in the E–step.

Given values for $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$ and $\boldsymbol{p}$ and following this procedure, if $p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})$ denotes the conditional distribution of the variable Z_i, i.e.,

$$p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) = \frac{p_m f_m(\Theta_i; \mu_m, \kappa_m)}{\sum_{l=1}^{M} p_l f_l(\Theta_i; \mu_l, \kappa_l)}$$

then, the expectation of (1.12) can be written as:

$$\sum_{i=1}^{n} \mathbb{E}_p \left[\log(p_{Z_i} f_{Z_i}(\Theta_i; \mu_{Z_i}, \kappa_{Z_i}))\right] = \sum_{i=1}^{n} \sum_{m=1}^{M} \log\left(p_m f_m(\Theta_i; \mu_m, \kappa_m)\right) p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) =$$

$$\sum_{m=1}^{M} \sum_{i=1}^{n} (\log p_m)\, p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) + \sum_{m=1}^{M} \sum_{i=1}^{n} \log f_m(\Theta_i; \mu_m, \kappa_m) p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}). \tag{1.17}$$

The next step, the M–step, consists on the maximization of the expectaction (1.17) with respect to the parameters $(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})$. Banerjee et al. (2005) proved that the mean direction estimator is

$$\hat{\mu}_m = \arctan^* \left(\frac{\sum_{i=1}^n \sin \Theta_i p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})}{\sum_{i=1}^n \cos \Theta_i p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})} \right), \quad m = 1, \ldots, M,$$

which is a weighted version of the estimator in (1.16). The estimator of the concentration parameter is obtained from the solution of equation

$$A(\hat{\kappa}_m) = \frac{\sum_{i=1}^n p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) \cos(\Theta_i - \hat{\mu}_m)}{\sum_{i=1}^n p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})}, \quad m = 1, \ldots, M,$$

and the estimator of p_m has the following expression

$$\hat{p}_m = \frac{1}{n} \sum_{i=1}^n p\left(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}\right), \quad m = 1, \ldots, M.$$

E–step and M–step are repeated iteratively until the likelihood converges. Each iteration is guaranteed to increase the log–likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood (Bilmes, 1998).

Initialization of the EM algorithm

The EM algorithm needs to be initialized, i.e., it is required initial values of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$ and $\boldsymbol{p}$. In order to obtain such starting values, there exist several approaches such as soft–assignment schemes and hard–assignment schemes. Hard–assignments consist in assigning each observation to one component of the mixture in such way that the probability of each observation of belonging to a certain component is 0 or 1. However, in this work, soft–assignments are considered. Let $\text{Sim}(\omega_1, \omega_2) = \cos(\omega_1 - \omega_2)$, be a measure of similartiy between two angles ω_1 and ω_2. The soft–assignments consists in:

1. Take the mean direction of the sample as the global centroid of the data.

2. Starting with this first centroid, if the number of components in the mixture is M then $(M-1)$ more centroids are computed. In order to compute the m-th centroid $(m = 2, \ldots, M)$, the similarity between each point in the sample and each one of the $(m-1)$ centroids computed before is calculated. For each point in the sample, the maximum value of similarity between that point and each centroid is taken. The m-th centroid will be the sample point with a smaller similarity.

3. Once the M centroids are computed, the dissimilarity (dissimilarty=1-similarity) between each sample point and each centroid is computed.

4. Two cases are distinguished in this step. For each sample point:

- If the sample point is equal to some centroid, i.e., the dissimilarity between the sample point and the centroid is equal to zero then, it is assigned probability one to the corresponding centroid and probability zero to the remaining centroids.
- If the sample point is not equal to any centroid, then the probability of belonging to the group represented by each centroid is representative is computed. That probability is proportional to the inverse of the dissimilarity. Hence, as smaller is the dissimilarity between one sample point and one centroid, larger is the probability of belonging to that group.

This procedure is equivalent to initialize the conditional distribution $p(m|\Theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p})$, $m = 1, \ldots, M$.

In order to illustrate the performance of the EM algorithm for estimating the parameters of a mixture of von Mises distributions, 1000 random samples of size $n = 100$ from the first model in Figure 1.3 were generated. Figure 1.6 shows the boxplots of the differences in absolute value between the estimated and true values of the parameters. Figure 1.6 shows that the differences for the mean directions (left panel) and for the proportions (right panel) are small, and they are sligthly larger for the concentration parameters (center panel).

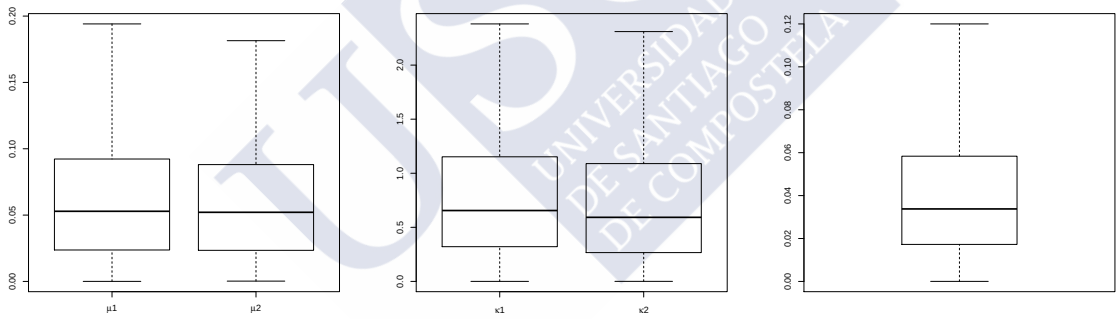

Figure 1.6: Boxplots of the differences in absolute value between the estimated parameters using the EM algorithm and the true values for a mixture of two von Mises distributions, $1/2 \cdot vM(0, 4) + 1/2 \cdot vM(\pi, 4)$. Results were obtained from 1000 random samples of size $n = 100$. Left panel: boxplots for the mean directions. Center panel: boxplots for the concentration parameters. Rigth panel: boxplot for the proportion.

Note that in the mixture considered both density components have the same proportion and so, the differences for the mean direction and for the concentration parameters are almost equal for both densities. However, if a mixture of two von Mises distribution in different proportion is considered then, the errors for the component in larger proportion are smaller, as shown in Figure 1.7, where 1000 random samples of size $n = 100$ from the second mixture in Figure 1.3 have been considered.

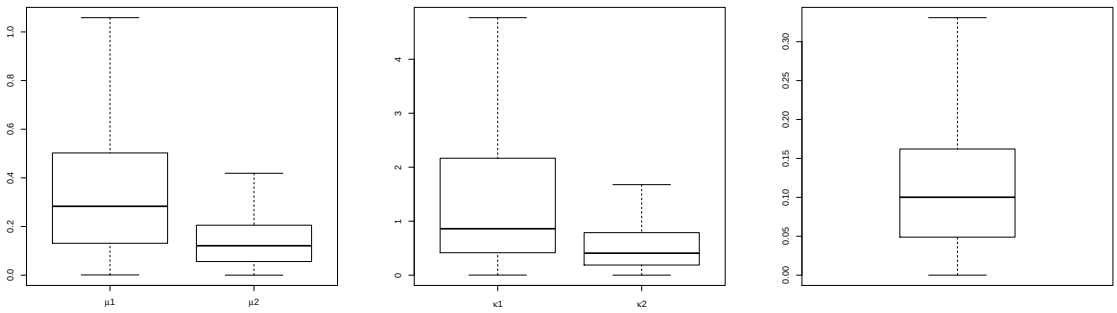

Figure 1.7: Boxplots of the differences (in absolute value) between the estimated parameters using the EM algorithm and the true values for a mixture of two von Mises distributions, $1/4 \cdot vM(0,2) + 3/4 \cdot vM(\pi/\sqrt{3},2)$. Results obtained from 1000 random samples of size $n = 100$. Left panel: boxplots for the mean directions. Center panel: boxplots for the concentration parameters. Right panel: boxplot for the proportion.

Number of components selection

In order to apply the EM algorithm, the number of components in the mixture must be chosen. If a large number of components (i.e., a larger number of parameters) is considered then overfitting may occur, whereas the opposite effect may occur if that number is small. Determining the number of components M can be seen as a model selection problem which can be approached by considering some kind of information criteria, such as the Akaike Information Criterion (AIC). AIC (Akaike, 1974) is a criterion for choosing the best model among a set of admissible models. AIC takes into account the model complexity by means of the number of parameters in the model and the model fit by means of the likelihood function. It has the following expression

$$\text{AIC} = -2\log(L) + 2d,$$

where d is the number of parameters in the model and L is the maximized value of the likelihood function for the estimated model. According to this criterion, one model is better than another if it has a smaller AIC value. So, given a set of models, the best model using the AIC criterion, is one with the lowest AIC.

In the scenarios considered above, it has been assumed that the number of distributions in the mixture is known. However, in most cases, this information is unknown. AIC can be used for selecting the mixture of von Mises distribution that fits the data best. In order to illustrate how this method performs in practice, five models have been considered: the von Mises distribution $vM(\pi,1)$ and the four mixtures represented in Figure 1.3. The specific formulae of these models can be seen in the Appendix A (models M2, M7, M9, M11 and M19, respectively). For each

distribution, 1000 samples of size $n = 100$ and $n = 500$ were generated and the number of times that the AIC selects $M = 1, 2, 3, 4$ and 5 has been computed. Results are shown in Tables 1.1 and 1.2. For both sample sizes and for the models with three or less components (M2, M7, M9 and M11), it can be seen that AIC selects the right number of components in most cases. However, for the model with more than three components (M19) and sample size $n = 100$, AIC tends to select $M = 2$ since the number of parameters of the model is too large (14 parameters) in comparison with the sample size. For sample size $n = 500$, in most cases AIC selects $M = 4$ which corresponds to the number of modes of the model. Note that, in this latter model the AIC criterion does not tend to select the exact number of components ($M = 5$) because the proportion of one of the components in the model is small and moreover, its concentration parameter is also small and so, this component hardly affects to the model. Therefore, in this particular case, a mixture of four von Mises distribution provides a good approximation.

$n = 100$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
Model with 1 component (M2)	814	99	53	26	8
Model with 2 components (M7)	0	807	111	60	22
Model with 2 components (M9)	318	540	75	40	27
Model with 3 components (M11)	0	0	834	118	48
Model with 5 components (M19)	0	439	293	184	84

Table 1.1: Number of times that AIC has selected $M = 1, 2, 3, 4$ and 5 in a von Mises distribution (M2), mixture of two von Mises (M7 and M9), mixture of three von Mises (M11) and mixture of four von Mises (M19). For each model, results were obtained from 1000 samples of sample size $n = 100$.

$n = 500$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
Model with 1 component (M2)	943	42	11	4	0
Model with 2 components (M7)	0	913	67	13	7
Model with 2 components (M9)	2	939	39	11	9
Model with 3 components (M11)	0	0	920	66	14
Model with 5 components (M19)	0	51	299	414	236

Table 1.2: Number of times that AIC has selected $M = 1, 2, 3, 4$ and 5 in a von Mises distribution (M2), mixture of two von Mises (M7 and M9), mixture of three von Mises (M11) and mixture of five von Mises (M19). For each model, results were obtained from 1000 samples of sample size $n = 500$.

1.3 Real datasets

In this section, several real datasets will be introduced. Three of them are original datasets which motivated the development of some techniques shown in this dissertation and so, they will be analized as part of this work. The others, corresponding to cross–beds angles and animal orientation data, are classical datasets and they will be used purely for illustrative purposes. A description of all of them is given below:

- **Temperature cycle changes.**

  The International Polar Year addresses as one of the main subjects the quantification and understanding of the environmental change in the polar regions. In particular, monitoring the retreat of glaciers is in the scope of this project. Within this project, measurement stations were placed in periglacial Monte Alvear (Tierra del Fuego, Argentina) (see Figure 1.8), recording temperatures hourly at different depths. The ocurrence of changes in cycles of temperature (from frosting to defrosting and viceversa) are important for the analysis of the mobility in the glacier's surface. The hours when a cycle change has occurred constitute a sample of circular data, coming from an unknown circular distribution, that must be estimated in order to determine the cycle change behaviour.

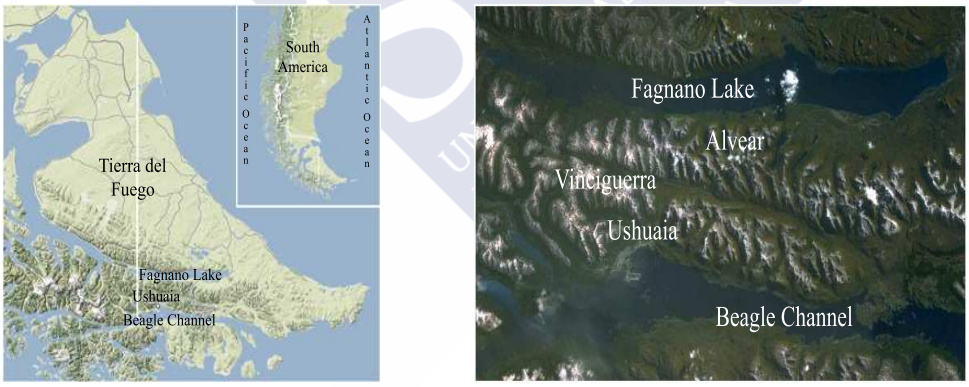

Figure 1.8: Ushuaia region (Tierra del Fuego, Argentina). Monte Alvear and Vinciguerra Glacier.

  The available data for studying the cycle change behaviour consist of 350 observations which correspond to the hours when the temperature (measured in °C) at ground level changed from positive to negative and viceversa (see Figure 1.9) from February 2008 to December 2009 in periglacial Monte Alvear. This dataset will be analized in Sections 2.2.3 and 3.6.

  These data has been kindly provided by Prof. Augusto Pérez Alberti from the Department of Physical Geography of the University of Santiago de Compostela.

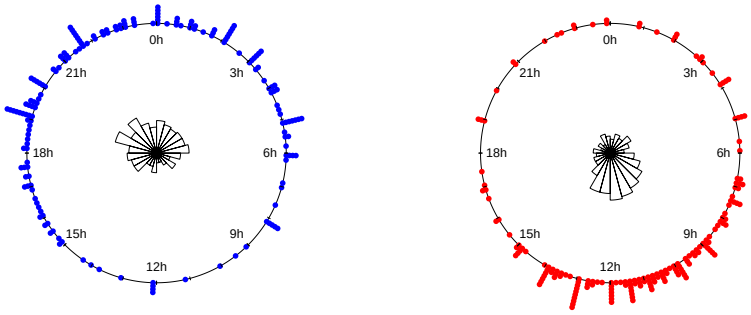

Figure 1.9: Circular plots and rose diagrams of data of hours when the temperature changes from positive to negative (left panel) and viceversa (right panel).

- **Wind speed and wind direction.**

  The Atlantic coast of Galicia (NW Spain) has suffered two major ship accidents which caused serious environmental and ecological damages: the burning of a cargo ship named Casón in 1987, and the oil spill of the Prestige tanker in 2002. The strong winds played a decisive role in both accidents. In the first one, the strong winds caused a displacement of the cargo and the corrosive and toxic chemical flamable products trasported by Casón exploded and burned while in the Prestige accident, the highly variable and strong winds caused the sinking of the tanker.

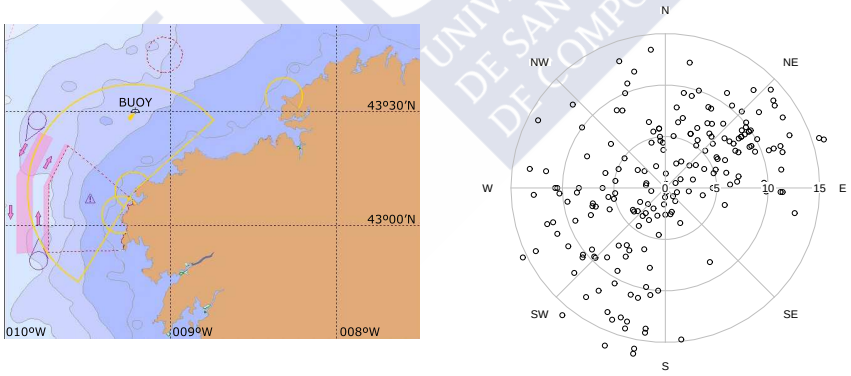

Figure 1.10: Left panel: Atlantic coast of Galicia (NW Spain). The plot shows the marine traffic control area (arrows indicate the directions that ships must follow), whithin the influence area of two major lighthouses (white lines). The buoy registering the data is located NE from the traffic control area at longitude -0.210E and latitude 43.500N. Right panel: wind direction is represented over the circumference in clockwise sense, starting form N and wind speed is represented along the radius in m/s.

Motivated by these facts, one question of interest is whether the wind speed may be influenced by the wind direction. A buoy (with a diameter of 1.8 m and a height of 6.5 m)

anchored in the area at longitude -0.210E and latitude 43.500N (see Figure 1.10, left panel) provides hourly collected wind speed and wind direction. Wind measurements regarding direction and speed are recorded every ten minutes, and hourly averaged, at a height of 3 m above sea level. The buoy is far away from the coastline so that the measurements are not influenced by local effects.

Data for studying the relation between wind speed and wind direction consists of hourly observations of wind direction (measured in degrees from North direction) and wind speed (in m/s) in winter season (from November to February) from 2003 until 2012. Data were freely downloaded from the Spanish Portuary Authority (`http://www.puertos.es`) in July of 2012. Figure 1.10 (right panel) shows the measurements of wind direction and wind speed. This plot correspond to about 200 observations out of the total data, where observations were taken with a lag period of 95 h for avoiding the dependence present between consecutive measurements in the time series. Since wind direction is a circular variable and wind speed is a scalar variable, the methods for studying the relation between these variables must be take into account the nature of both variables. This data set will be considered in Sections 2.3.4 and 3.6.

- **Cracks in cemented femoral components.**

  This real dataset, kindly provided by Dr. Kenneth A. Mann from the Upstate Medical University (New York), concerns angular positions of cracks in the cement mantle in a hip implant. These data, described in more detail in Mann et al. (2003), are obtained from an in vitro fatigue study for investigating the distribution of fatigue cracks around cemented femoral components in total hip replacements.

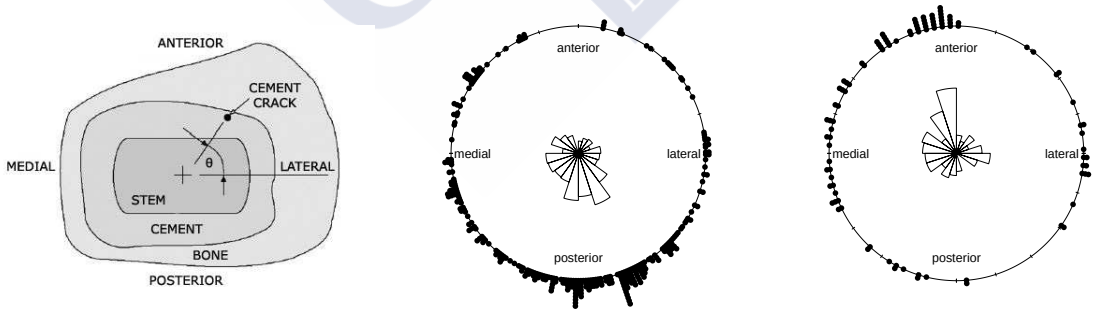

Figure 1.11: Rigth panel: a counterclockwise definition of angular position of crack was used with a zero angle representing the lateral direction. Center and right panels: circular plots and rose diagrams of the data of angular positions of cracks for one cemented implant for proximal and distal region, respectively.

Each femur is loaded using a stair climbing apparatus and after loading, it is sectioned in 10 mm intervals from the level of the implant collar to the distal tip of the stem. For

each section, angular position of the cracks relative to the center of the stem section were documented. A counterclockwise definition of angular position of crack was used with a zero angle representing the lateral direction as shown in Figure 1.11 (left panel). In Figure 1.11 (center and right panels) the angular positions of cracks for proximal (sections at 10–50 mm) and distal (sections at 80–110 mm) regions are represented. The number of data in each region is 322 and 99, respectively.

Studies to improve understanding of the mechanical aspects of cemented implant loosening were carried out showing that the distribution of the angular positions of the cracks around cemented femoral components is not uniform (see Mann et al., 2003). It is of interest to know if there exists some predominant direction of crack. This will be studied in Section 3.6.

Apart from the previous datasets, some of which had not been previously studied, and certainly none of them had been analyzed with nonparametric methods, for illustration purposes, some classical datasets will be also considered. Specifically, the following examples will be used in Section 2.2.3.

- **Cross–beds (I).**

  This classical dataset corresponds to azimuths of cross–beds in the Kamthi river (India). Originally analized by SenGupta and Rao (1966) and included in Table 1.5 in Mardia (1972), the dataset collects 580 azimuths (measured in degrees) of layers lying oblique to principal accumulation surface along the river, being these layers known as cross–beds. A photo of cross–beds is shown in Figure 1.12 (left panel) and data are shown in Figure 1.12 (right panel).

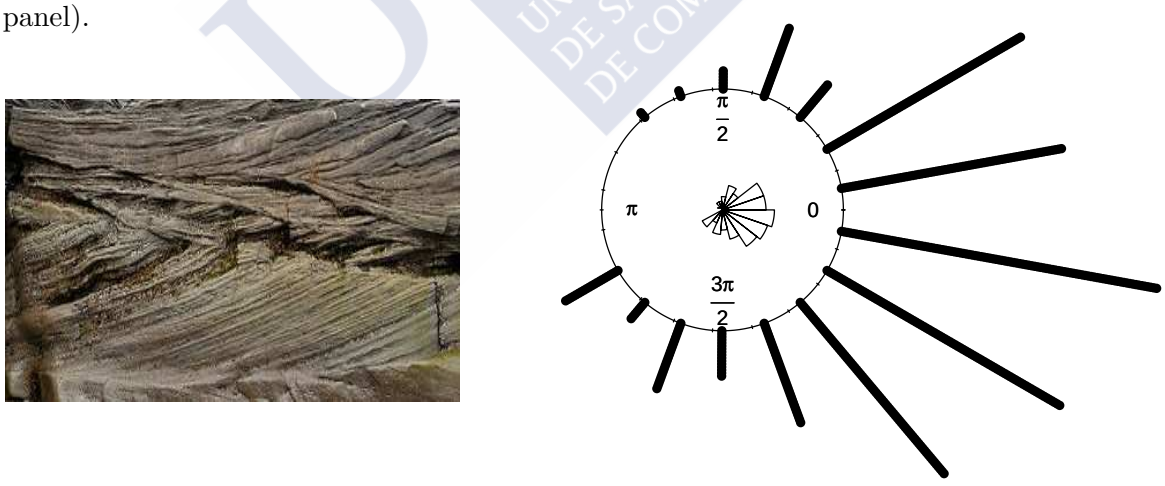

Figure 1.12: Left panel: photo showing cross–beds[2]. Right panel: circular plot and rose diagram of data of azimuhts of cross–beds in the Kamthi river.

[2]Rygel, M.C. (2006) Through cross–bedding in the Waddens Cove Formation (Pennsylvanian), Sidney Basin, Nova Scotia. File from the Wikimedia Commons.
http://commons.wikimedia.org/wiki/File:Trough_xbed_mcr1.JPG.

- **Cross–beds (II).**

  This dataset, presented in Fisher (1993), includes 104 measurements of Chaudwan Zam large bedforms from Himalayan molasse in Pakistan which are represented in Figure 1.13.

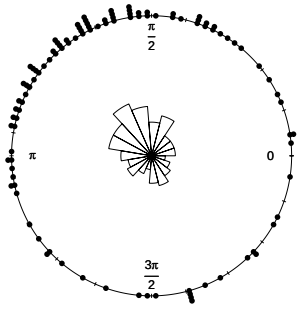

Figure 1.13: Circular plot and rose diagram of cross–bed measurements from Himalayan molasse in Pakistan.

- **Dragonfly orientation.**

  Animal orientation is another classical example of circular data. This dataset, presented in Batschelet (1981), contains the orientation of 213 dragonflies with respect to the Sun's azimuth. As it can be seen in Figure 1.14 (right panel), this is a clear example of bimodal circular distribution. This dataset was also studied by Pewsey (2004), who applied a test for circular reflective symmetry.

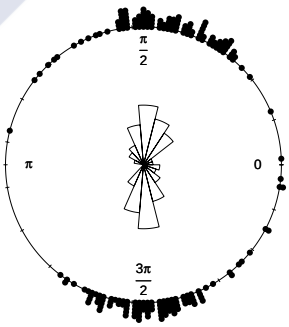

Figure 1.14: Left panel: image of a dragongly[3]. Right panel: circular plot and rose diagram of dragonflies orientations data.

[3]Image of a dragonfly taken with a PackshotCreator photo studio by Creative Tools AB. Date: 27 January 2010. Source: CreativeTools.se - PackshotCreator - Dragonfly top view. Author: Creative Tools from Halmstad, Sweden. Watermark removed by Ainali.
http://commons.wikimedia.org/wiki/File:CreativeTools.se_-_PackshotCreator_-_Dragonfly_top_view.jpg.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 2

Nonparametric curve estimation for circular data

2.1 Introduction

Nonparametric estimation methods have turned up as an alternative approach to the parametric techniques, both inferentially and as a descriptive tool. In the circular setting, nonparametric density estimation was approached by Fisher (1989), who proposed an adaptation to circular data of linear data methods in Silverman (1986) using a quartic kernel function and Beran (1979) and Hall et al. (1987), who proposed a kernel density estimation procedure for the general case of spherical data, following the ideas of the classical kernel density estimator for linear data (Parzen, 1962; Rosenblatt, 1956). Although asymptotic properties of this latter estimator were further studied by Bai et al. (1988) and Klemelä (2000), these works do not provide a solution for the most critical issue from a practical point of view: smoothing parameter selection. The use of cross–validation smoothing parameters is suggested by Hall et al. (1987) in the spherical context, for the particular case of circular data, Taylor (2008) derived a *rule of thumb* for smoothing parameter selection in circular kernel density estimation and Di Marzio et al. (2011) introduced a bootstrap method for data lying on a d–dimensional torus.

Regression estimation involving circular variables, as response or as covariate, is indeed an interesting problem. In the available literature, most efforts have been focused on the development of parametric models. For instance, Presnell et al. (1998) and the references therein dealt with a circular response and linear covariates; SenGupta and Ugwuowo (2006) proposed some asymmetric models for environmental applications accounting for the circular nature of the covariate, and Downs and Mardia (2002) and Kato et al. (2008), among others, addressed the regression with circular response and covariates. Regression estimation avoiding the assumption of a specific parametric shape for the regression curve was addressed by Di Marzio et al. (2009) who extended least squares local polynomial to the case of d–dimensional circular predictors and real–valued

responses; Qin et al. (2011a,b) who extended nonparametric models to the case when there is one circular predictor and one or more linear predictors and the response is real–valued, and more recently Di Marzio et al. (2012) proposed a nonparametric estimator for the regression function when the response is circular and the covariate is circular or linear. Periodic smoothing splines, as considered in Wahba (1990) and Wood (2006) among others, are an alternative form of smoothing when the covariate is periodic and the response is linear.

The goal of this chapter is to introduce a new procedure for selecting the smoothing parameter in circular kernel density estimation that allows estimating circular densities, accounting for asymmetry and/or multimodality. In the regression setting, a review of the nonparametric methods for a scalar response and a circular covariate will be provided.

This chapter is organized as follows. Section 2.2 is devoted to the introduction of the kernel density estimator for circular data, revising different techniques for selecting the smoothing parameter and introducing a new method. The performance of the described procedures is checked in a simulation study, considering a wide class of circular density families, involving multimodality, peakdness and skewness. The methods are also illustrated with the three classical datasets and the real dataset corresponding to temperature cycle changes. Section 2.3 is devoted to nonparametric regression estimation for a circular explanatory variable and a linear response, focusing on the adaptation of the Nadaraya–Watson and Local Linear estimators to the circular nature of the covariate and on periodic smoothing splines. The performance of circular kernel regression estimators and periodic smoothing splines estimator is explored in some simulated examples and they are applied to study the relation between the wind speed and wind direction in the Atlantic coast of Galicia. The contents of this chapter can be seen in Oliveira et al. (2012a,b, 2013c).

2.2 Nonparametric circular kernel density estimation

Before introducing the circular kernel density estimator, the classical kernel estimator for a density function will be reviewed. Denote by $X_1, \ldots, X_n$ a random sample from a scalar random variable X with density g. At each fixed point $x \in \mathbb{R}$, the kernel estimator of $g(x)$ is defined as:

$$\hat{g}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} L\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where $h > 0$ is the bandwidth or smoothing parameter and L is a kernel function, usually the standard normal density, or any other unimodal and symmetric around zero density function. The estimator in (2.1) can be written as follows:

$$\hat{g}(x; h) = \frac{1}{n} \sum_{i=1}^{n} L_h(x - X_i), \tag{2.2}$$

where L_h is the h–rescaled kernel function L, $L_h(\cdot) = \frac{1}{h} L\left(\frac{\cdot}{h}\right)$. In the particular case of L being the standard normal, the kernel estimator in (2.2) can be interpreted as a mixture of n normally distributed random variables, centered in the sample points and with standard deviation h.

Since this estimator does not provide a periodic estimate of the density function, its usage is not appropriate for estimating the density function of a sample of circular data. However, bearing the idea of its construction in mind, the kernel estimator in (2.2) can be generalized to circular data.

Given a random sample of angles $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$ from some unknown circular density f, the circular kernel density estimator of f is defined as:

$$\hat{f}(\theta; \nu) = \frac{1}{n} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i), \quad 0 \leq \theta < 2\pi, \tag{2.3}$$

where K_ν is a circular kernel function with concentration parameter $\nu > 0$ (see, e.g., Di Marzio et al., 2009). As a circular kernel, the von Mises distribution can be considered. With this specific kernel, the density estimator is given by:

$$\hat{f}(\theta; \nu) = \frac{1}{n(2\pi)I_0(\nu)} \sum_{i=1}^{n} e^{\nu \cos(\theta - \Theta_i)}, \quad 0 \leq \theta < 2\pi, \tag{2.4}$$

which can be seen as a mixture of n von Mises distributions, centered in the data sample Θ_i and with common concentration parameter ν. Throughout this dissertation, the circular kernel density estimator with von Mises kernel defined in (2.4) will be considered, unless otherwise indicated.

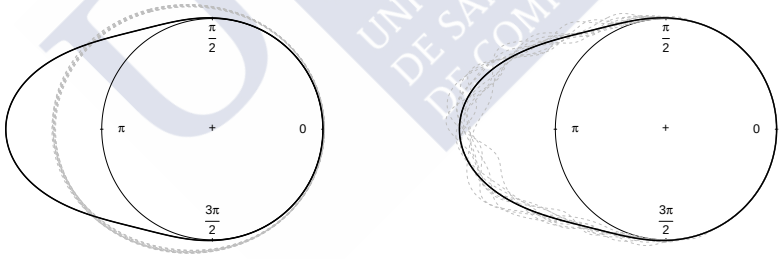

Figure 2.1: Circular kernel density estimates (gray lines) with $\nu = 2$ (left panel) and $\nu = 100$ (right panel) for 10 random samples of size 200 from a $vM(\pi, 5)$ (black line).

A critical issue when applying this estimator in practice is the choice of the smoothing parameter ν which determines the degree of smoothing. The effect of the smoothing parameter can be seen in Figure 2.1, large values of ν lead to highly variable (undersmoothed) estimators, i.e., estimators with small bias and large variance, whereas small values of ν imply low concentration of the kernel, providing oversmoothed estimators for the circular density, i.e., estimators with large bias and small variance. For that reason the study of automatic smoothing parameter selection procedures constitutes one of the most relevant problems in nonparametric density estimation.

2.2.1 Smoothing parameter selectors

There are various approaches to the smoothing parameter selection problem. In this section, the methods proposed in the literature will be reviewed and a new method will be introduced.

As for linear data, most commonly used techniques for selecting the smoothing parameter are based on the minimization of some error criteria that quantify the accuracy of the kernel density estimator, i.e., how well the estimator approximates the true density. The mean integrated squared error (MISE),

$$\begin{aligned}\text{MISE}\,(\nu) &= \mathbb{E}\left[\int_0^{2\pi}\left(\hat{f}(\theta;\nu)-f(\theta)\right)^2 d\theta\right] = \int_0^{2\pi}\mathbb{E}\left[\left(\hat{f}(\theta;\nu)-f(\theta)\right)^2\right] d\theta \\ &= \int_0^{2\pi}\left[\mathbb{E}\left(\hat{f}(\theta;\nu)\right)-f(\theta)\right]^2 d\theta + \int_0^{2\pi}\mathbb{E}\left[\hat{f}(\theta;\nu)-\mathbb{E}\left(\hat{f}(\theta;\nu)\right)\right]^2 d\theta \\ &= \int_0^{2\pi}\left[\text{bias}\left(\hat{f}(\theta;\nu)\right)\right]^2 d\theta + \int_0^{2\pi}\text{var}\left(\hat{f}(\theta;\nu)\right) d\theta,\end{aligned}$$

is one of these criteria but, in practice, often its asymptotic expression (AMISE) is used for selecting the smoothing parameter, which may also be written in terms of the aymptotic bias and the variance of the estimator. Precisely, as noted before, a main challenge in nonparametric density estimation is the bias–variance trade–off. Therefore, selecting ν by minimizing either MISE or AMISE amounts to balancing bias and variance at the same time.

In the circular setting, the asymptotic expression for the MISE (AMISE) was derived by Di Marzio et al. (2009). For the circular kernel estimator (2.4), if f'' is continuous and square–integrable, the AMISE(ν) when $\nu \to \infty$ and $\sqrt{\nu}n^{-1} \to 0$ is given by:

$$\text{AMISE}(\nu) = \left\{\frac{1}{16}\left[1-\frac{I_2(\nu)}{I_0(\nu)}\right]^2 \int_0^{2\pi}\left[f''(\theta)\right]^2 d\theta + \frac{I_0(2\nu)}{2n\pi\left(I_0(\nu)\right)^2}\right\}, \tag{2.5}$$

where f'' denotes the second–order derivative of the target density to be estimated, which measures the curvature of f. Densities with marked modes will give a larger value of its integral, whereas the lowest value is achieved by a circular uniform model.

A *rule of thumb*, adapting the idea of Silverman (1986) for bandwidth selection in kernel linear density estimation, was proposed by Taylor (2008). Assuming that the data follow a von Mises distribution with concentration parameter κ, the AMISE is given by

$$\text{AMISE}(\nu) = \frac{3\kappa^2 I_2(2\kappa)}{32\pi\nu^2 I_0(\kappa)^2} + \frac{\nu^{1/2}}{2n\pi^{1/2}}. \tag{2.6}$$

Hence, the value of the smoothing parameter minimizing the AMISE (2.6) can be estimated by

$$\hat{\nu}_{RT} = \left[\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\pi^{1/2} I_0(\hat{\kappa})^2}\right]^{2/5}, \tag{2.7}$$

where $\hat{\kappa}$ is the concentration parameter estimator obtained by maximum likelihood. This selector performs satisfactorily in fitting unimodal symmetric distributions, without highly peaked modes

but its behaviour can be dramatically misleading in the presence of antipodal modes and/or skewed distributions (see Section 2.2.2). A very simple example of this situation arises when mixing two population with opposite centers but in the same proportion and with the same concentration parameters. The maximum likelihood estimate $\hat{\kappa}$ will return a value close to zero, which corresponds with a circular uniform distribution. Consequently, a small value for $\hat{\nu}_{RT}$ will be obtained resulting in an oversmoothed kernel estimator for the circular density.

The poor performance of the rule of thumb is sometimes due to the non robust estimation by maximum likelihood of the concentration parameter κ, so a possible modification of (2.7) consists in the following procedure:

Step 1. Select $\alpha \in (0,1)$ and find the shortest arc containing $\alpha \cdot 100\%$ of the sample data.

Step 2. Obtain the estimated $\hat{\kappa}$ in such way that $\int f(\theta, \mu_{arc}, \hat{\kappa})d\theta = \alpha$ where μ_{arc} is the midpoint of the arc computed in Step 1. The intregral is computed over the arc selected in Step 1.

Step 3. Replace in (2.7) the value of $\hat{\kappa}$ computed in the previous step to obtain $\hat{\nu}_{RT}^{R}$.

The performance of this procedure in relation with the rule proposed by Taylor (2008) can be seen in Oliveira et al. (2013c), where it is shown that, in some scenarios, the selector $\hat{\nu}_{RT}^{R}$ can improve the results of $\hat{\nu}_{RT}$, if α is properly chosen.

An alternative route, also based in the AMISE minimization, would be to plug–in a more flexible distribution family as a reference density in the AMISE formula (2.5). That is the idea of the new rule proposed in this dissertation which is introduced below.

The new proposal: the plug–in rule

The new method, based on the ideas of Ćwik and Koronacki (1997) for the multivariate setting, consists on considering a mixture of von Mises distribution as reference. The proposed plug–in smoothing parameter selector, $\hat{\nu}_{PI}$, is obtained as follows:

Step 1. Select the number of mixture components M for the reference distribution.

Step 2. Estimate the AMISE in (2.5) as follows:

- Step 2.1. Estimate the parameters in the von Mises mixture (1.3), (μ_m, κ_m, p_m), for $m = 1, \ldots, M$.
- Step 2.2. Compute the integral $\int (\hat{f}''(\theta))^2 d\theta$ where $\hat{f}''$ is the second derivative of the density function of a mixture of M von Mises distribution with the parameters estimated in the previous step.
- Step 2.3. Plug–in the quantity above in (2.5) to get $\widehat{\text{AMISE}}(\nu)$.

Step 3. Minimize $\widehat{\text{AMISE}}(\nu)$ and obtain $\hat{\nu}_{PI}$.

For Step 1, the selection of the number of mixture components in the reference distribution can be done by AIC (see Section 1.2.2), considering mixtures with different number of components. In that case the selector will be denoted by $\hat{\nu}_{PI}^{AIC}$, otherwise, if M is selected a priori, the selector will be denoted by $\hat{\nu}_{PI}^{M}$ where M will indicate the number of components in the mixture. Maximum likelihood estimation via EM algorithm, as described in Section 1.2.2, is used for Step 2.1. The integral in Step 2.2 can be efficiently computed numerically, by quadrature methods such as the Simpson's rule. In Step 3, an optimization method can be used in order to minimize the $\widehat{\text{AMISE}}$.

These types of plug–in rules are not the only alternative to smoothing parameter selection, and some other data–driven procedures were already proposed by Hall et al. (1987) using cross–validation ideas. Least squares cross–validation (LSCV) is based on minimizing the integrated squared error (ISE):

$$\begin{aligned} \text{ISE}(\nu;\Theta_1,\ldots,\Theta_n) &= \int_0^{2\pi} \left(\hat{f}(\theta;\nu) - f(\theta)\right)^2 d\theta \\ &= \int_0^{2\pi} \hat{f}^2(\theta;\nu)d\theta - 2\int_0^{2\pi} \hat{f}(\theta;\nu)f(\theta)d\theta + \int_0^{2\pi} f^2(\theta)d\theta, \end{aligned} \tag{2.8}$$

Since the third term does not depend on ν, the minimization of (2.8) involves only the first two addends, being the first one known as a function of ν. The integral in the second term in (2.8), which depends on the unknown density f, can be approximated by $n^{-1}\sum_{i=1}^{n}\hat{f}^{-i}(\Theta_i;\nu)$, where $\hat{f}^{-i}$ is the circular kernel density estimator obtained by leaving out the i-th observation. Hence, the LSCV smoothing parameter is obtained as the value of ν that minimizes:

$$\text{LSCV}(\nu) = \int_0^{2\pi} \hat{f}^2(\theta;\nu)d\theta - \frac{2}{n}\sum_{i=1}^{n}\hat{f}^{-i}(\Theta_i;\nu). \tag{2.9}$$

The likelihood cross–validation smoothing parameter $\hat{\nu}_{LCV}$ is obtained by maximizing:

$$\text{LCV}(\nu) = \prod_{i=1}^{n}\hat{f}^{-i}(\Theta_i;\nu). \tag{2.10}$$

The performance of the cross–validation selectors, including an adaptation of the biased cross–validation rule (Scott and Terrell, 1987), was studied by Di Marzio et al. (2011) for selecting the smoothing parameter in kernel density estimation for data lying on a d–dimensional torus, concluding that the likelihood cross–validation method appears asymptotically the most stable.

Di Marzio et al. (2011) introduced a bootstrap procedure for selecting the smoothing parameter by adapting the proposal of Taylor (1989) for linear data. This method consists on selecting ν to minimize the bootstrap MISE

$$\int_0^{2\pi} \mathbb{E}_B\left[\hat{f}^*(\theta;\nu) - \hat{f}(\theta;\nu)\right]^2 d\theta, \tag{2.11}$$

where $\mathbb{E}_B$ denotes the bootstrap expectation with respect to random samples $\Theta_1^*,\ldots,\Theta_n^*$ generated from $\hat{f}(\theta;\nu)$. If a von Mises kernel is used, as it is the case, then the integrand of (2.11) has a

closed expression:

$$\mathbb{E}_B\left[\hat{f}^*(\theta;\nu)-\hat{f}(\theta;\nu)\right]^2 = (2\pi n I_0(\nu))^{-2} I_0(\nu)^{-1}\sum_{l=1}^{n} I_0\left(\nu(5+4\cos(\theta-\Theta_l))^{1/2}\right) + \left(\mathbb{E}_B\left[\hat{f}^*(\theta;\nu)\right]-\hat{f}(\theta;\nu)\right)^2 - n^{-1}\left(\mathbb{E}_B\left[\hat{f}^*(\theta;\nu)\right]\right)^2 \quad (2.12)$$

where the bootstrap expected value for the kernel estimator is given by

$$\mathbb{E}_B\left[\hat{f}^*(\theta,\nu)\right] = \frac{1}{n(2\pi)I_0(\nu)}\sum_{i=1}^{n}\frac{I_0(2\nu\cos((\theta-\Theta_i)/2))}{I_0(\nu)}.$$

Note that (2.12) is zero for $\nu = 0$ and so, the target function (2.11) has a global minimum at $\nu = 0$, leading to a uniform estimate, no matter the true underlying model. In practice, this will lead to search for a local minimum, which may pose a problem specially for small samples. The value of the smoothing parameter selected by using this method will be denoted by $\hat{\nu}_{boot}$.

2.2.2 Simulation study

The effectiveness of the new selector, the plug–in rule, for selecting the smoothing parameter described in the previous section has been compared with the rule of thumb defined in (2.7), leasts squares cross–validation rule (2.9), likelihood cross–validation rule (2.10) and bootstrap method (2.12) through Monte Carlo experiments. A variety of circular distributions (von Mises, cardioid, various wrapped distributions and mixtures of them) displaying multimodality, skewness and/or peakedness have been tried (see Figure 2.2 for plots and Appendix A for specific formulae). For illustration purposes, the models have been classified in four groups, according to their complexity:

Simple models: circular uniform (M1); von Mises (M2); wrapped normal (M3); cardioid (M4); wrapped Cauchy (M5) and wrapped skew–normal (M6).

Two components models: von Mises mixtures (M7, M8 and M9); mixture of von Mises and wrapped Cauchy (M10).

Models with more than two components: von Mises mixtures with three components (M11, M12 and M13); von Mises mixture with four components (M14); mixture of wrapped Cauchy, wrapped normal, von Mises and wrapped skew–normal (M15); von Mises mixture with five components (M16).

Other complex models: mixture of cardioid and wrapped Cauchy (M17); mixture of von Mises (M18 and M19); mixture of two wrapped skew–normal and two wrapped Cauchy (M20).

Note that *Simple models* include unimodal models from von Mises distributions, with the circular uniform as a particular case. The wrapped Cauchy shows a highly peaked mode, whereas an asymmetric model is obtained with the wrapped skew–normal, as shown in Pewsey (2006). The *Two components models* collect different mixtures of two von Mises distributions (with antipodal modes and combining different weights and centers) and a mixture of a von Mises and a wrapped

Cauchy, which results in a distribution with two modes with different concentrations. In *Models with more than two components*, there are mixtures of three, four and five equally spaced and equally weighted von Mises distributions. Other situations with mild modes such as model M15 are also considered. Finally, *Other complex models* are also included in the study. Although the distributions in this group are generated by mixtures of two or more models, the appearance may show a single mode, as in M17.

For each distribution model, 1000 random samples of sizes $n = 100,\ 250,\ 500$ and 1000 were generated. In Tables 2.3, 2.4, 2.5 and 2.6, the average integrated squared errors of the circular kernel density estimator (2.4), considering different smoothing parameter selectors are shown. For each selector, the average ISE over the 1000 replicates will be denoted, for the sake of simplicity, by MISE($\hat{\nu}_\bullet$). Specifically, the performance of the new plug–in rule $\hat{\nu}_{PI}$ will be compared with the rule of thumb $\hat{\nu}_{RT}$, the likelihood and least squares cross–validation smoothing parameters $\hat{\nu}_{LCV}$ and $\hat{\nu}_{LSCV}$ respectively, and the bootstrap smoothing parameter $\hat{\nu}_{boot}$. As a benchmark, the minimum average ISE has been computed for a broad grid of smoothing parameters, denoted in the tables by MISE(ν_0). The simulations have been carried out in R (see R Development Core Team, 2012) using the self programmed functions implemented in the `NPCirc` package (see Chapter 4 and Appendix D).

Step 1 in the algorithm for computing the plug–in smoothing parameter requires the selection of the number of components of the mixture for the reference distribution. Note that the rule of thumb proposed in Taylor (2008) corresponds to $M = 1$. The procedure with fixed M has been applied in all the scenarios, obtaining $\hat{\nu}_{PI}^{M}$, for $M = 2, 3, 4$ and 5 and observing that even with a fixed value $M = 2$, the plug–in rule gives better results than $M = 1$. Just for illustrating our conclusions, the average ISE values for $\hat{\nu}_{RT}$ and $\hat{\nu}_{PI}^{M}$ with $M = 2, 3, 4$ and 5 can be seen in Tables 2.1 and 2.2, for $n = 100$ and $n = 1000$, respectively. In general, small values of M are suitable for simple models and models with two components for any sample size, and large values of M are a good choice for complex models and moderate and large sample sizes. Hence, fixing the number of mixtures M does not produce satisfactory results in all the simulation scenarios. The AIC criterion (described in Section 1.2.2) provides a data driven procedure for selecting M, so Step 1 in the algorithm is done as follows: AIC is computed for mixtures of $M = 2, 3, 4, 5$ von Mises distributions and the selected number of mixtures M for the reference distribution is the one minimizing the AIC.

For sample size $n = 100$ (see Table 2.3), the plug–in rule $\hat{\nu}_{PI}^{AIC}$ is competitive with the other selectors, although the likelihood cross–validation rule provides better results except for models M5 and M17 which present highly peaked modes. Results using the least squares cross–validation rule are not so good as using likelihood cross–validation rule. It should be noted that the AIC criterion tends to select a large value for M, which may be damaging in some simple models compared with the results for $\hat{\nu}_{PI}^{M=2}$ shown in Table 2.1. Therefore, for this sample size one should not try AIC with a large number of mixtures in the reference distribution. Besides, for small sample sizes, it may not be realistic to attempt to estimate too complicated models (see MISE(ν_0)

in Table 2.3 for M17 to M20). Results were also obtained considering the bootstrap selector for $n = 100$. The poor results for some models provided by this method are due to the fact that, as already noticed in the previous section, only the global minimum at $\nu = 0$ for (2.12) is found, not being able to attain a local minimum for most of the generated samples. This problem was already noticed by Di Marzio et al. (2011).

As it was noted above, whereas chosing $M = 2$ seems fair for small samples, including a complex reference distribution, i.e., large M, is reasonable for large enough datasets. The AIC criterion succeeds in selecting a suitable M for all the considered scenarios. For moderate and large sample sizes ($n = 250$, 500, 1000), results with the AIC selection equal or even outperform the best $\hat{\nu}_{PI}^{M}$. The strength of the new proposal can be seen in Tables 2.4, 2.5 and 2.6.

In more detail, for *Simple models* and $n = 250$, 500 and 1000, the rule of thumb and bootstrap method show the best results for models M1 to M4, but in any case the other methods also provide good results. However, the behaviour shown by the rule of thumb in models M5 and M6 is quite poor, compared with the plug–in selector which is the best in these cases. The least squares cross-validation rule provides similar results to those obtained with the plug–in rule and the likelihood cross–validation rule does not provide good results for model M5 (highly peaked mode), although it behaves better than the rule of thumb. For $n = 250$, the bootstrap method does not provide good results in models M5 and M6. Despite for $n = 500$ and $n = 1000$ its behaviour improves, it does not outperform the plug–in rule. Note that M5 is the wrapped Cauchy distribution and M6 is the wrapped skew–normal, confirming the adecquate performance of the plug–in rule for estimating highly peaked and asymmetric distributions.

In the *Two components models*, for $n = 250$, 500 and 1000 , the performance of $\hat{\nu}_{RT}$ is extremely poor for model M7 (antipodal modes), and is also far from satisfactorily for models M8 and M10. The plug–in rule $\hat{\nu}_{PI}^{AIC}$ provides good results for all the models in this group (compared with the optimal MISE, MISE(ν_0)), whereas $\hat{\nu}_{LCV}$ and $\hat{\nu}_{boot}$ seem to be fair competitors except for model M10 for which $\hat{\nu}_{LSCV}$ provides better results.

For models with *More than two components*, the rule of thumb seems not consistent (except perhaps for model M15, which is *almost* flat). The bootstrap selector $\hat{\nu}_{boot}$ behaves poorly for $n = 250$ and although it is better for $n = 500$ and 1000, it does not reach $\hat{\nu}_{PI}^{AIC}$ and $\hat{\nu}_{LCV}$. The plug–in rule and the likelihood cross–validation rule behave similarly. Results for $\hat{\nu}_{LSCV}$ are very similar to, but not quite as good as using $\hat{\nu}_{LCV}$.

For *Other complex models* and $n = 250$, 500 and 1000, the plug–in rule and cross–validation rules performs similarly, except for model M17 for which $\hat{\nu}_{LCV}$ provide worse results. The bootstrap method does not provide good results for these models for any sample size.

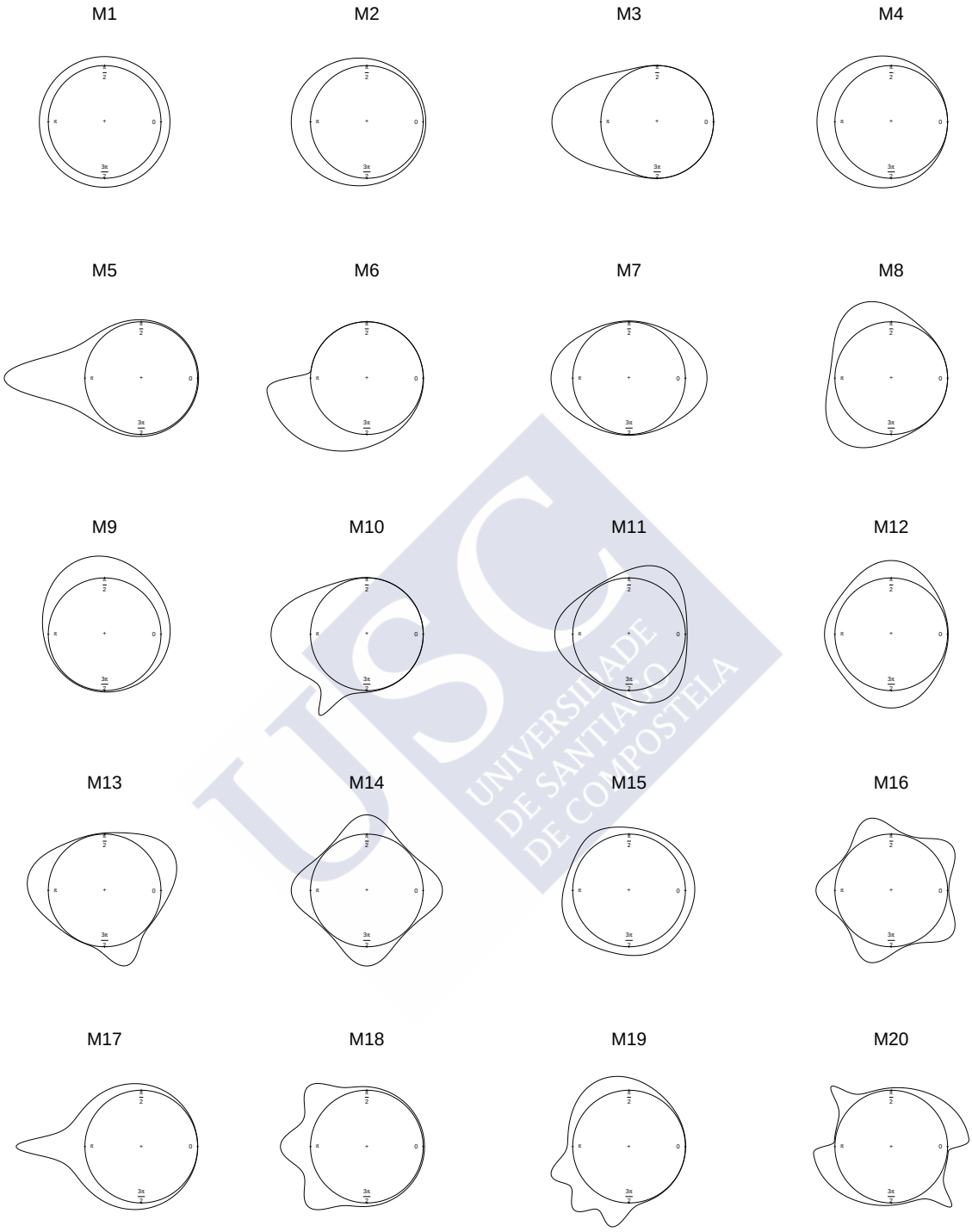

Figure 2.2: Circular density models. M1-M6: simple models. M7-M10: two components models. M11-M16: models with three or more components. M17-M20: complex models.

$n = 100$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{M=2}$)	MISE($\hat{\nu}_{PI}^{M=3}$)	MISE($\hat{\nu}_{PI}^{M=4}$)	MISE($\hat{\nu}_{PI}^{M=5}$)
M1	0.0000	0.0197 (0.0561)	0.2558 (0.4639)	0.7362 (0.7665)	1.2175 (0.9589)	1.5597 (1.0746)
M2	0.5143	0.6677 (0.5171)	0.7488 (0.7795)	0.8948 (0.7626)	1.2632 (0.9980)	1.5962 (1.1849)
M3	1.1861	1.2481 (0.9033)	1.4939 (1.1914)	1.8727 (1.5225)	2.1598 (1.6892)	2.5710 (1.8995)
M4	0.4673	0.5195 (0.3706)	0.6795 (0.5743)	0.9965 (0.8639)	1.3401 (1.0761)	1.7590 (1.2991)
M5	2.7539	8.5089 (2.5804)	3.2430 (1.7832)	3.3060 (1.8041)	3.4369 (1.9805)	3.9859 (2.4719)
M6	2.3628	3.1577 (0.9371)	2.6280 (1.2056)	2.8852 (1.3383)	3.2703 (1.4240)	3.4551 (1.4615)
M7	1.1182	10.5622 (0.3741)	1.1795 (0.6582)	1.4993 (1.2908)	1.8082 (1.6620)	2.3984 (2.3100)
M8	1.2326	3.7176 (0.6908)	1.3031 (0.7086)	1.6127 (1.0454)	1.9020 (1.2553)	2.4057 (1.3751)
M9	0.6766	0.8320 (0.5467)	0.8451 (0.6139)	1.0514 (0.7896)	1.3511 (0.9737)	1.6821 (1.1484)
M10	2.3203	2.8659 (0.7894)	2.7317 (1.0735)	2.9159 (1.2413)	3.0347 (1.2872)	3.1319 (1.3129)
M11	1.3573	6.4858 (0.0188)	2.1984 (1.7288)	1.4751 (0.6856)	1.8364 (1.1071)	2.1494 (1.2780)
M12	1.0051	4.2671 (0.5160)	1.1053 (0.5799)	1.2990 (0.7522)	1.5974 (1.0729)	1.8819 (1.2169)
M13	1.7569	10.8771 (0.1257)	1.9361 (0.8015)	1.9901 (0.8490)	2.3760 (1.5334)	2.9217 (2.2857)
M14	1.8510	8.1840 (0.0510)	7.0262 (2.0491)	2.2744 (1.3453)	2.0392 (0.8299)	2.6959 (1.6725)
M15	0.7091	0.8400 (0.1507)	0.9652 (0.4238)	1.1375 (0.6909)	1.4059 (0.8796)	1.6926 (1.0328)
M16	2.1505	7.8356 (0.0502)	7.4690 (1.1866)	4.3266 (2.4983)	2.5389 (1.1013)	2.4512 (0.9729)
M17	3.4817	7.9040 (1.2024)	5.9841 (1.9124)	4.3443 (1.5192)	4.2346 (1.5348)	4.2770 (1.6335)
M18	2.2067	3.5795 (0.4979)	3.2344 (0.8189)	2.7847 (1.1375)	2.8780 (1.2712)	2.9221 (1.3419)
M19	2.3756	3.8709 (0.6246)	2.4580 (0.6342)	2.6701 (0.9434)	2.8739 (1.0713)	3.2066 (1.1322)
M20	3.2890	10.9610 (0.0545)	6.4831 (1.3001)	3.9311 (1.4886)	3.8228 (1.4249)	4.1138 (1.8071)

Table 2.1: Average integrated squared error for different smoothing parameter selectors, MISE (×100), and standard deviations (×100, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb) and $\hat{\nu}_{PI}^{M}$ (plug–in rule) with $M = 2, 3, 4$ and 5. MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 100$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

$n = 1000$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{M=2}$)	MISE($\hat{\nu}_{PI}^{M=3}$)	MISE($\hat{\nu}_{PI}^{M=4}$)	MISE($\hat{\nu}_{PI}^{M=5}$)
M1	0.0000	0.0003 (0.0011)	0.0136 (0.0266)	0.0437 (0.0407)	0.0691 (0.0470)	0.0816 (0.0517)
M2	0.0955	0.1088 (0.0673)	0.0991 (0.0585)	0.1001 (0.0582)	0.1040 (0.0597)	0.1104 (0.0642)
M3	0.2257	0.2291 (0.1294)	0.2317 (0.1328)	0.2346 (0.1337)	0.2378 (0.1360)	0.2467 (0.1426)
M4	0.0836	0.0854 (0.0474)	0.0918 (0.0507)	0.0961 (0.0542)	0.0999 (0.0584)	0.1087 (0.0623)
M5	0.5001	2.9980 (0.5902)	0.5349 (0.2859)	0.5255 (0.2718)	0.5267 (0.2739)	0.5234 (0.2698)
M6	0.5317	1.2071 (0.1992)	0.5888 (0.2011)	0.5466 (0.2003)	0.5431 (0.1973)	0.5442 (0.1983)
M7	0.1964	10.7433 (0.0836)	0.1985 (0.0890)	0.2000 (0.0894)	0.2009 (0.0903)	0.2020 (0.0912)
M8	0.2245	1.0836 (0.1748)	0.2268 (0.1037)	0.2282 (0.1040)	0.2284 (0.1046)	0.2299 (0.1048)
M9	0.1272	0.1546 (0.0896)	0.1308 (0.0739)	0.1335 (0.0751)	0.1354 (0.0754)	0.1390 (0.0769)
M10	0.5276	1.2004 (0.1777)	0.5815 (0.2365)	0.5500 (0.1963)	0.5439 (0.1862)	0.5447 (0.1870)
M11	0.2312	6.4800 (0.0016)	2.3859 (2.7829)	0.2334 (0.0867)	0.2350 (0.0867)	0.2358 (0.0867)
M12	0.1789	2.9699 (0.3453)	0.1876 (0.0837)	0.1836 (0.0785)	0.1847 (0.0785)	0.1855 (0.0788)
M13	0.3117	10.9019 (0.0604)	0.3754 (0.1146)	0.3185 (0.1050)	0.3196 (0.1052)	0.3204 (0.1054)
M14	0.3144	8.1666 (0.0009)	7.8797 (0.5790)	0.5111 (1.1235)	0.3559 (0.5104)	0.3198 (0.1012)
M15	0.1609	0.7009 (0.0942)	0.4727 (0.1722)	0.2051 (0.1038)	0.1720 (0.0666)	0.1711 (0.0644)
M16	0.3723	7.8192 (0.0011)	7.7280 (0.2077)	6.6255 (1.3787)	0.3768 (0.1110)	0.3761 (0.1100)
M17	0.7022	4.8753 (0.3789)	3.4968 (1.3066)	0.7757 (0.3226)	0.7512 (0.2867)	0.7391 (0.2697)
M18	0.4019	2.0780 (0.1616)	1.5967 (0.5630)	0.5234 (0.3101)	0.4207 (0.1510)	0.4141 (0.1476)
M19	0.4763	2.1306 (0.1383)	0.9899 (0.1271)	0.7714 (0.2731)	0.6087 (0.1950)	0.5470 (0.1877)
M20	0.6786	10.9911 (0.0167)	3.7033 (0.2338)	0.8481 (0.4578)	0.7498 (0.1751)	0.7189 (0.1623)

Table 2.2: Average integrated squared error for different smoothing parameter selectors, MISE ($\times 100$), and standard deviations ($\times 100$, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb) and $\hat{\nu}_{PI}^{M}$ (plug–in rule) with $M = 2, 3, 4$ and 5. MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 1000$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

$n = 100$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{AIC}$)	MISE($\hat{\nu}_{LCV}$)	MISE($\hat{\nu}_{LSCV}$)	MISE($\hat{\nu}_{boot}$)
M1	0.0000	0.0197 (0.0561)	0.7369 (1.1351)	0.3394 (0.6413)	0.3832 (0.7658)	0.0002 (0.0002)
M2	0.5143	0.6677 (0.5171)	1.0893 (1.1894)	0.7429 (0.6460)	0.9935 (1.0801)	1.7484 (2.4292)
M3	1.1861	1.2481 (0.9033)	1.9162 (1.6936)	1.4902 (1.1241)	1.8803 (1.5438)	1.4199 (1.0033)
M4	0.4673	0.5195 (0.3706)	1.1967 (1.2703)	0.7639 (0.6916)	0.9029 (1.0062)	0.712 (1.2764)
M5	2.7539	8.5089 (2.5804)	3.2726 (1.7542)	6.7155 (2.9400)	3.4708 (1.9075)	22.0763 (21.3347)
M6	2.3628	3.1577 (0.9371)	3.1992 (1.5703)	2.8330 (1.1939)	2.9799 (1.4405)	4.5217 (1.3819)
M7	1.1182	10.5622 (0.3741)	1.4558 (1.0851)	1.2516 (0.7082)	1.4852 (1.0313)	3.1809 (3.8574)
M8	1.2326	3.7176 (0.6908)	1.6459 (1.2467)	1.4515 (0.8559)	1.7113 (1.2717)	11.6473 (4.1182)
M9	0.6766	0.8320 (0.5467)	1.1381 (1.0364)	0.8404 (0.5753)	1.1122 (1.0398)	1.4682 (2.1964)
M10	2.3203	2.8659 (0.7894)	3.0615 (1.4278)	3.0427 (1.0538)	2.8872 (1.3191)	3.8685 (0.8533)
M11	1.3573	6.4858 (0.0188)	1.7697 (1.1089)	1.5068 (0.6908)	1.675 (0.9997)	6.4807 (0.0002)
M12	1.0051	4.2671 (0.5160)	1.5295 (1.1269)	1.1671 (0.6155)	1.3687 (0.9119)	5.2076 (0.1748)
M13	1.7569	10.8771 (0.1257)	2.2131 (1.1394)	1.8851 (0.7855)	2.1091 (1.0592)	10.9275 (0.4629)
M14	1.8510	8.1840 (0.0510)	2.2464 (1.1066)	1.9885 (0.8044)	2.1127 (0.9735)	8.1666 (0.0002)
M15	0.7091	0.8400 (0.1507)	1.3980 (0.9839)	0.9765 (0.5455)	1.0440 (0.6943)	1.0256 (0.0132)
M16	2.1505	7.8356 (0.0502)	2.4498 (0.9670)	2.2858 (0.8196)	2.4396 (0.9824)	7.8191 (0.0002)
M17	3.4817	7.9040 (1.2024)	4.2581 (1.6030)	5.6499 (1.8984)	4.0658 (1.5774)	9.9210 (1.9547)
M18	2.2067	3.5795 (0.4979)	3.0838 (1.3216)	2.9230 (0.8912)	2.7037 (1.2062)	3.8837 (0.5016)
M19	2.3756	3.8709 (0.6246)	2.9164 (1.1040)	2.5375 (0.6532)	2.7516 (0.9851)	9.5571 (4.0343)
M20	3.2890	10.9610 (0.0545)	3.8827 (1.2087)	3.4914 (0.8968)	3.6366 (1.0675)	10.9997 (0.0005)

Table 2.3: Average integrated squared error for different smoothing parameter selectors, MISE (×100), and standard deviations (×100, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb), $\hat{\nu}_{PI}^{AIC}$ (plug–in rule), $\hat{\nu}_{LCV}$ (likelihood cross–validation), $\hat{\nu}_{LSCV}$ (least squares cross–validation), $\hat{\nu}_{boot}$ (bootstrap). MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 100$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

$n = 250$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{AIC}$)	MISE($\hat{\nu}_{LCV}$)	MISE($\hat{\nu}_{LSCV}$)	MISE($\hat{\nu}_{boot}$)
M1	0.0000	0.0034 (0.0096)	0.1440 (0.2684)	0.1534 (0.3053)	0.1587 (0.3194)	0.0001 (0.0001)
M2	0.2768	0.3323 (0.2314)	0.3582 (0.2873)	0.3649 (0.2728)	0.4330 (0.3835)	0.3141 (0.2112)
M3	0.5876	0.6073 (0.4053)	0.7257 (0.5684)	0.7125 (0.5141)	0.8496 (0.6459)	0.6476 (0.4458)
M4	0.2321	0.2438 (0.1565)	0.3951 (0.3391)	0.3491 (0.2559)	0.3781 (0.3580)	0.2420 (0.1539)
M5	1.4468	5.7994 (1.5296)	1.6164 (0.8346)	2.9328 (1.3887)	1.6958 (0.8684)	2.2145 (1.2529)
M6	1.3609	2.1685 (0.4964)	1.5752 (0.6529)	1.5816 (0.6085)	1.5482 (0.6305)	2.5303 (0.7047)
M7	0.5536	10.6636 (0.2268)	0.6165 (0.3616)	0.6095 (0.3033)	0.7019 (0.4228)	0.6041 (0.2849)
M8	0.6398	2.4037 (0.4340)	0.7001 (0.3946)	0.7338 (0.3838)	0.7930 (0.4610)	1.0706 (1.3377)
M9	0.3541	0.4289 (0.2566)	0.4146 (0.2870)	0.4331 (0.2807)	0.5177 (0.4197)	0.4394 (0.2550)
M10	1.3336	2.0175 (0.4318)	1.5544 (0.6039)	2.0287 (0.6732)	1.5464 (0.5979)	2.7877 (0.5087)
M11	0.6765	6.4797 (0.0016)	0.7382 (0.3499)	0.7397 (0.3218)	0.7862 (0.3833)	1.7834 (2.2475)
M12	0.5142	3.9052 (0.4825)	0.5947 (0.3329)	0.5752 (0.2898)	0.6174 (0.3476)	3.2204 (2.2725)
M13	0.8661	10.8967 (0.0924)	0.9432 (0.3843)	0.9178 (0.3611)	0.9840 (0.4258)	1.4211 (1.6912)
M14	0.9124	8.1694 (0.0089)	0.9698 (0.3660)	0.9608 (0.3501)	1.0110 (0.4005)	2.7165 (3.0938)
M15	0.4300	0.8007 (0.1312)	0.6259 (0.2864)	0.5488 (0.2620)	0.5577 (0.2845)	1.0246 (0.026)
M16	1.1045	7.8228 (0.0130)	1.1513 (0.3596)	1.1570 (0.3646)	1.2033 (0.4188)	6.7337 (2.4918)
M17	1.8911	6.6158 (0.7782)	2.1440 (0.8353)	3.2693 (1.2456)	2.0687 (0.7471)	8.5558 (0.7698)
M18	1.1218	2.9481 (0.2920)	1.3367 (0.5989)	1.4394 (0.5904)	1.2852 (0.5241)	3.2811 (0.2608)
M19	1.3044	2.9910 (0.3509)	1.5192 (0.4665)	1.5805 (0.4206)	1.4597 (0.4813)	3.0670 (1.3432)
M20	1.8094	10.9769 (0.0348)	1.9165 (0.4900)	1.9039 (0.4657)	1.9192 (0.4914)	10.9995 (0.0005)

Table 2.4: Average integrated squared error for different smoothing parameter selectors, MISE (×100), and standard deviations (×100, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb), $\hat{\nu}_{PI}^{AIC}$ (plug–in rule), $\hat{\nu}_{LCV}$ (likelihood cross–validation), $\hat{\nu}_{LSCV}$ (least squares cross–validation), $\hat{\nu}_{boot}$ (bootstrap). MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 250$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

$n = 500$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{AIC}$)	MISE($\hat{\nu}_{LCV}$)	MISE($\hat{\nu}_{LSCV}$)	MISE($\hat{\nu}_{boot}$)
M1	0.0000	0.0012 (0.0038)	0.0537 (0.1085)	0.0751 (0.1471)	0.0726 (0.1439)	0.0000 (0.0000)
M2	0.1604	0.1823 (0.1252)	0.1829 (0.1290)	0.2052 (0.1540)	0.2404 (0.2097)	0.1724 (0.1165)
M3	0.3764	0.3842 (0.2462)	0.4130 (0.2870)	0.4245 (0.2775)	0.5006 (0.3461)	0.4004 (0.2665)
M4	0.1435	0.1480 (0.0851)	0.2032 (0.1329)	0.2119 (0.1446)	0.2125 (0.1744)	0.1471 (0.0843)
M5	0.8673	4.2662 (0.9643)	0.9372 (0.4761)	1.5791 (0.7807)	0.9695 (0.4870)	1.0834 (0.5883)
M6	0.8385	1.5981 (0.3096)	0.9169 (0.3413)	0.9722 (0.3721)	0.9171 (0.3355)	1.5791 (0.4490)
M7	0.3318	10.7199 (0.1344)	0.3411 (0.1639)	0.3537 (0.1671)	0.3872 (0.2043)	0.3513 (0.1684)
M8	0.3733	1.6375 (0.2731)	0.3872 (0.1971)	0.4121 (0.2004)	0.4370 (0.2326)	0.4024 (0.1985)
M9	0.2112	0.2551 (0.1442)	0.2285 (0.1339)	0.2492 (0.1553)	0.2860 (0.2085)	0.2419 (0.1398)
M10	0.8309	1.5683 (0.2643)	0.8955 (0.2987)	1.3761 (0.4601)	0.9228 (0.3131)	2.1135 (0.4220)
M11	0.3924	6.4797 (0.0021)	0.4044 (0.1689)	0.4210 (0.1758)	0.4395 (0.1956)	0.4189 (0.1729)
M12	0.3008	3.4879 (0.4202)	0.3174 (0.1451)	0.3317 (0.1635)	0.3546 (0.1875)	0.4480 (0.5568)
M13	0.5299	10.8991 (0.0795)	0.5511 (0.1990)	0.5542 (0.2044)	0.5876 (0.2278)	0.6253 (0.2075)
M14	0.5346	8.1672 (0.0021)	0.5471 (0.1882)	0.5599 (0.1949)	0.5808 (0.2121)	0.5725 (0.1959)
M15	0.2650	0.7602 (0.1150)	0.3349 (0.1826)	0.3229 (0.1545)	0.3277 (0.1650)	1.0076 (0.0968)
M16	0.6461	7.8198 (0.0026)	0.6588 (0.1987)	0.6703 (0.2021)	0.6904 (0.2163)	0.7052 (0.2123)
M17	1.1389	5.6896 (0.5330)	1.2276 (0.4999)	1.9657 (0.7296)	1.1992 (0.4414)	7.6512 (1.1003)
M18	0.6744	2.5206 (0.2028)	0.7232 (0.2923)	0.7969 (0.3014)	0.7370 (0.2780)	2.9546 (0.1814)
M19	0.7965	2.5166 (0.2266)	0.9011 (0.2924)	0.9979 (0.2871)	0.8660 (0.2712)	1.8160 (0.2813)
M20	1.1327	10.9853 (0.0253)	1.1746 (0.2936)	1.1842 (0.2885)	1.1738 (0.2943)	9.9138 (2.9293)

Table 2.5: Average integrated squared error for different smoothing parameter selectors, MISE (×100), and standard deviations (×100, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb), $\hat{\nu}_{PI}^{AIC}$ (plug–in rule), $\hat{\nu}_{LCV}$ (likelihood cross–validation), $\hat{\nu}_{LSCV}$ (least squares cross–validation), $\hat{\nu}_{boot}$ (bootstrap). MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 500$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

$n = 1000$	MISE(ν_0)	MISE($\hat{\nu}_{RT}$)	MISE($\hat{\nu}_{PI}^{AIC}$)	MISE($\hat{\nu}_{LCV}$)	MISE($\hat{\nu}_{LSCV}$)	MISE($\hat{\nu}_{boot}$)
M1	0.0000	0.0003 (0.0011)	0.0216 (0.0418)	0.0378 (0.0826)	0.0387 (0.0828)	0.0000 (0.0000)
M2	0.0955	0.1088 (0.0673)	0.1014 (0.0622)	0.1151 (0.0778)	0.1244 (0.0918)	0.1015 (0.0619)
M3	0.2257	0.2291 (0.1294)	0.2353 (0.1373)	0.2515 (0.1473)	0.2783 (0.1656)	0.2343 (0.1367)
M4	0.0836	0.0854 (0.0474)	0.1047 (0.0610)	0.1225 (0.0792)	0.1145 (0.0906)	0.0850 (0.0472)
M5	0.5001	2.9980 (0.5902)	0.5269 (0.2731)	0.8427 (0.4314)	0.5325 (0.2762)	0.5677 (0.3136)
M6	0.5317	1.2071 (0.1992)	0.5435 (0.1977)	0.6122 (0.2388)	0.5592 (0.1980)	0.9493 (0.2725)
M7	0.1964	10.7433 (0.0836)	0.1992 (0.0902)	0.2101 (0.0965)	0.2252 (0.1105)	0.2036 (0.0931)
M8	0.2245	1.0836 (0.1748)	0.2279 (0.1048)	0.2438 (0.1106)	0.2561 (0.1260)	0.2343 (0.1094)
M9	0.1272	0.1546 (0.0896)	0.1320 (0.0753)	0.1443 (0.0858)	0.1621 (0.1046)	0.1396 (0.0843)
M10	0.5276	1.2004 (0.1777)	0.5425 (0.1865)	0.8809 (0.2732)	0.5612 (0.1932)	1.3528 (0.3940)
M11	0.2312	6.4800 (0.0016)	0.2342 (0.0870)	0.2435 (0.0921)	0.2520 (0.0998)	0.2391 (0.0909)
M12	0.1789	2.9699 (0.3453)	0.1844 (0.0788)	0.1951 (0.0895)	0.2048 (0.0986)	0.1990 (0.0922)
M13	0.3117	10.9019 (0.0604)	0.3190 (0.1053)	0.3237 (0.1095)	0.3372 (0.1160)	0.3448 (0.1127)
M14	0.3144	8.1666 (0.0009)	0.3184 (0.1012)	0.3273 (0.1058)	0.3362 (0.1115)	0.3256 (0.1052)
M15	0.1609	0.7009 (0.0942)	0.1736 (0.0713)	0.1847 (0.0808)	0.1864 (0.0833)	0.8469 (0.2525)
M16	0.3723	7.8192 (0.0011)	0.3761 (0.1100)	0.3843 (0.1137)	0.3903 (0.1156)	0.3894 (0.1147)
M17	0.7022	4.8753 (0.3789)	0.7440 (0.2764)	1.2008 (0.4393)	0.7354 (0.2616)	4.5347 (2.7871)
M18	0.4019	2.0780 (0.1616)	0.4142 (0.1468)	0.4589 (0.1732)	0.4309 (0.1567)	2.1778 (1.0097)
M19	0.4763	2.1306 (0.1383)	0.5302 (0.1841)	0.5805 (0.1865)	0.5024 (0.1595)	1.4249 (0.1528)
M20	0.6786	10.9911 (0.0167)	0.7190 (0.1621)	0.7003 (0.1553)	0.6912 (0.1554)	1.4263 (0.9852)

Table 2.6: Average integrated squared error for different smoothing parameter selectors, MISE (×100), and standard deviations (×100, in parentheses). Smoothing parameter selectors: $\hat{\nu}_{RT}$ (rule of thumb), $\hat{\nu}_{PI}^{AIC}$ (plug–in rule), $\hat{\nu}_{LCV}$ (likelihood cross–validation), $\hat{\nu}_{LSCV}$ (least squares cross–validation), $\hat{\nu}_{boot}$ (bootstrap). MISE(ν_0): benchmark average integrated squared error. Sample size: $n = 1000$. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

As commented in Section 2.2.1, and from what is seen in the results for model M7, one of the problems of the rule of thumb in the presence of antipodal modes is that it tends to provide uniform estimates for the circular density, which corresponds to a null concentration parameter in the von Mises family. A natural question arises: what would happen if a different parametric family, not including the uniform distribution, is used as a reference? It has been also checked by simulations, considering the same models as the ones presented here (see Table 2.7), that setting a wrapped Cauchy as reference density in the minimization of the AMISE error in (2.5) provides better results in some models such as M5 and M17 which have a highly peaked mode, but far from the new plug–in proposal. The parameters of the wrapped Cauchy distribution were estimated by maximum likelihood (see Jammalamadaka and SenGupta, 2001, Section 4.2.1).

	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
M1	0.0024 (0.0616)	0.0001 (0.0001)	0.0000 (0.0000)	0.0000 (0.0000)
M2	0.6562 (0.5120)	0.3243 (0.1854)	0.1864 (0.1066)	0.1069 (0.0551)
M3	1.8990 (1.1691)	0.9149 (0.5097)	0.5567 (0.2869)	0.3339 (0.1533)
M4	0.5942 (0.4208)	0.2854 (0.1725)	0.1753 (0.0952)	0.1012 (0.0527)
M5	3.1003 (1.5813)	1.5552 (0.7661)	0.9045 (0.4327)	0.5081 (0.2508)
M6	2.4522 (1.0591)	1.4330 (0.5265)	0.9219 (0.2932)	0.6249 (0.1839)
M7	9.7027 (2.5315)	10.0167 (2.1393)	10.3725 (1.5812)	10.4199 (1.4425)
M8	3.3625 (1.1122)	2.0125 (0.4197)	1.3394 (0.2475)	0.8594 (0.1479)
M9	0.7981 (0.4916)	0.3910 (0.2280)	0.2308 (0.1236)	0.1359 (0.0726)
M10	2.4127 (0.9716)	1.4253 (0.4726)	0.9691 (0.2650)	0.6679 (0.1716)
M11	6.4794 (0.0326)	6.4805 (0.0033)	6.4806 (0.0000)	6.4805 (0.0000)
M12	5.0614 (0.5999)	5.0203 (0.6724)	4.6847 (0.9792)	2.7246 (1.0780)
M13	10.8255 (0.6926)	10.9218 (0.3012)	10.9275 (0.2990)	10.9331 (0.1997)
M14	8.1678 (0.0307)	8.1664 (0.0001)	8.1664 (0.0000)	8.1664 (0.0000)
M15	1.0220 (0.0595)	1.0011 (0.1044)	0.9477 (0.1763)	0.7012 (0.2409)
M16	7.8199 (0.0203)	7.8191 (0.0037)	7.8189 (0.0000)	7.8189 (0.0000)
M17	4.7801 (1.5212)	3.3422 (0.9723)	2.4902 (0.6261)	1.8998 (0.4132)
M18	2.9987 (0.6180)	2.0874 (0.3707)	1.5399 (0.2531)	1.0971 (0.1864)
M19	3.2194 (0.9824)	2.3438 (0.4143)	1.9523 (0.2441)	1.6371 (0.1522)
M20	10.9687 (0.2115)	10.9953 (0.0662)	10.9995 (0.0004)	10.9988 (0.0215)

Table 2.7: Average integrated squared error for the smoothing parameter selector obtained from the minimization of the AMISE error in (2.5) setting a wrapped Cauchy as reference density, and standard deviations ($\times 100$, in parentheses). Sample size: $n = 100, 250, 500$ and 1000. Models M1–M20 distributed by complexity: M1–M6 (simple models); M7–M10 (two components models); M11–M16 (models with more than two components); M17–M20 (other complex models).

2.2.3 Real data analysis

In this section, the performance of the circular kernel density estimator with different smoothing parameter selectors is illustrated. For that purpose, the classical datasets introduced in Section 1.3 will be used, since they present asymmetric, symmetric unimodal and bimodal distributions, re-

spectively. In addition, an application on the dataset concerning changes in temperature cycles will be also provided.

Example 1. Cross–beds azimuths (I). A circular kernel density estimation has been computed for the dataset of azimuths of cross–beds in the Kamthi river, introduced in Section 1.3, which contains 580 azimuths of layers lying oblique to the principal accumulation surface along the river. Four different smothing parameter selectors: rule of thumb, $\hat{\nu}_{RT}$, likelihood cross–validation, $\hat{\nu}_{LCV}$, plug–in rule, $\hat{\nu}_{PI}$ and bootstrap, $\hat{\nu}_{boot}$, have been considered. The estimator with the least squares cross–validation selector is not shown because, in this case, it behaves as the likelihood cross–validation selector. In Figure 2.3, it can be seen that the estimators with the rule of thumb, the plug–in and bootstrap smoothing parameters perform similarly, fitting a unimodal distribution with negative (anticlockwise) asymmetry. However, the likelihood cross–validation criterion provides a too large smoothing parameter, resulting in an undersmoothed fitted density. In this case, the number of selected mixtures by AIC was $M = 2$.

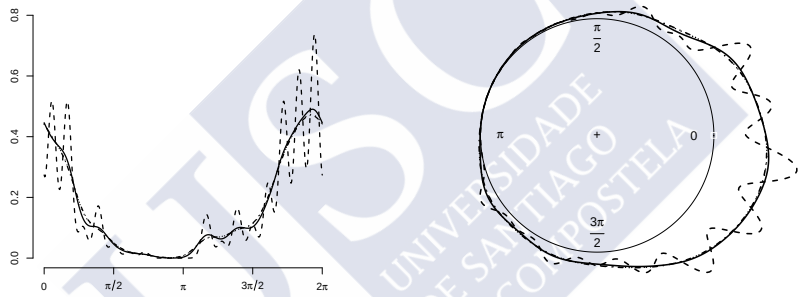

Figure 2.3: Linear (left panel) and circular (right panel) representation of the circular kernel density estimators for the azimuths. Solid line: plug–in selector, $\hat{\nu}_{PI}^{AIC}$. Dashed line: likelihood cross–validation rule, $\hat{\nu}_{LCV}$. Dotted line: rule of thumb, $\hat{\nu}_{RT}$. Dotted–dashed line: bootstrap selector, $\hat{\nu}_{boot}$.

Example 2. Cross–beds (II). Now the circular kernel density estimator is applied to the dataset of cross–beds measurements from Himalayan molasse in Pakistan described in Section 1.3. Although in the same practical situation as in Example 1, this dataset collects just 104 measurements. The circular kernel density estimator has been computed using the different smoothing parameter selectors. In Figure 2.4, it can be seen that the rule of thumb, likelihood cross–validation and plug–in selectors provide similar fitted curves. In this case, $\hat{\nu}_{LSCV}$ provides also a similar result (not shown). However, the bootstrap method behaves poorly, providing a uniform estimate. As noted in the previous section, this flat estimate for the underlying distribution is due to the fact that a local minimum for (2.12) does not exist for this particular sample.

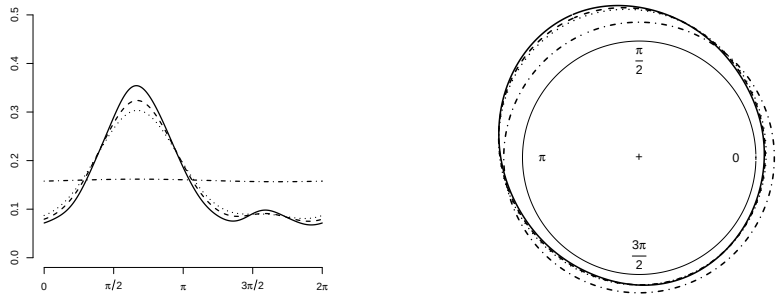

Figure 2.4: Linear (left panel) and circular (right panel) representation of the circular kernel density estimators for the cross–beds data. Solid line: plug–in selector, $\hat{\nu}_{PI}^{AIC}$. Dashed line: likelihood cross–validation rule, $\hat{\nu}_{LCV}$. Dotted line: rule of thumb, $\hat{\nu}_{RT}$. Dot-dashed line: bootstrap selector, $\hat{\nu}_{boot}$.

Example 3. Dragonflies orientation. The circular kernel density estimator and the smoothing parameter selection methods are illustrated with the classical data of orientation of 214 dragonflies with respect to the sun's azimuth introduced in Section 1.3. As it can be seen already from the circular plot in Figure 1.14, this is a clear example of bimodal circular distribution. In a situation like this one (opposite modes), the rule of thumb behaves quite poorly, as shown in Figure 2.5. Likelihood cross–validation, plug–in and bootstrap selectors provide similar fitted curves, showing two modes. The AIC criterion selects $M = 4$ mixtures for the reference distribution.

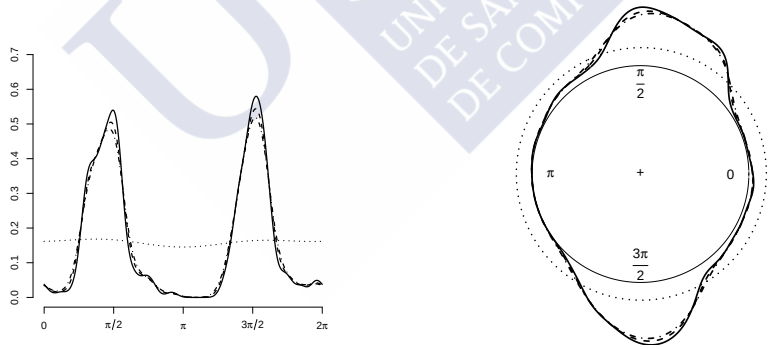

Figure 2.5: Linear (left panel) and circular (right panel) representation of the circular kernel density estimators for the dragonflies orientations data. Solid line: plug–in selector, $\hat{\nu}_{PI}^{AIC}$. Dashed line: likelihood cross–validation rule, $\hat{\nu}_{LCV}$. Dotted line: rule of thumb, $\hat{\nu}_{RT}$. Dotted–dashed line: bootstrap selector, $\hat{\nu}_{boot}$.

Example 4. Temperature cycle changes. Finally, the circular kernel density estimator is applied to explore the distribution of changes in cycles of temperatures at ground level in Monte Alvear. The results, using different smoothing parameter selectors, are shown in Figure 2.6. The rule of thumb ($\hat{\nu}_{RT}$) and the bootstrap method ($\hat{\nu}_{boot}$) provide an oversmoothed estimate, close

to the circular uniform distribution. However, both the plug–in rule and the likelihood cross–validation rule provide similar estimates, identifying a mode around 11 a.m. which indicates that the cycle changes occurs mainly at midday hours. This mode corresponds to the temperature changes from negative to positive since they are more concentrated around the midday whereas the changes from positive to negative occur mainly during the other hours of the day, as it can be observed in Figure 1.9. The mode can be seen in the linear representation (left panel) but, in the circular representation it is almost imperceptible. So, a question arises: is this mode really significant? In Chapter 3, this question will be answered.

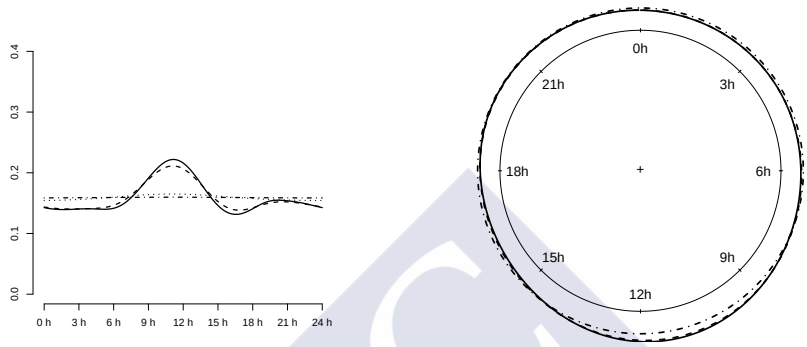

Figure 2.6: Linear (left panel) and circular (right panel) representation of the circular kernel density estimators for the dragonflies orientations data. Solid line: plug–in selector, $\hat{\nu}_{PI}^{AIC}$. Dashed line: likelihood cross–validation rule, $\hat{\nu}_{LCV}$. Dotted line: rule of thumb, $\hat{\nu}_{RT}$. Dotted–dashed line: bootstrap selector, $\hat{\nu}_{boot}$.

2.3 Nonparametric circular–linear regression estimation

In this section nonparametric regression estimators for a circular explanatory variable and a linear response will be studied. Two types of nonparametric smoothers will be considered: kernel estimators introduced by Di Marzio et al. (2009) and periodic smoothing splines introduced by Cogburn and Davis (1974).

2.3.1 Kernel smoothers

Let $\{(\Theta_i, Y_i),\ i = 1, \ldots, n\}$ denote a random sample from a circular random variable Θ and a linear random variable Y, respectively. The relation between these variables may be modelled by

$$Y_i = f(\Theta_i) + \sigma(\Theta_i)\varepsilon_i,\ \ i = 1, \ldots, n \tag{2.13}$$

where f denotes the regression function, $\sigma^2(\cdot)$ is the conditional variance of Y given Θ and ε_i are real–valued random variables with zero mean and unit variance. From now on, it is asumed that the error is uncorrelated with the covariate. In order to obtain a smooth nonparametric estimator

for the regression function f, a local polynomial fit could be applied, similarly to Fan and Gijbels (1996) for linear random variables. Based on this idea, Di Marzio et al. (2009) considers

$$\beta_0 + \beta_1 \sin(\cdot - \theta),$$

providing a local trigonometric polynomial fit. The parameters β_0 and β_1 can be estimated by a local least squares procedure:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i) \left[Y_i - (a + b\sin(\theta - \Theta_i))\right]^2, \tag{2.14}$$

where K_ν denotes the density of a $vM(0,\nu)$. The circular Local Linear estimator for $f(\theta)$ is given by $\hat{f}_{CLL}(\theta;\nu) = \hat{\beta}_0$ and the estimator of its derivative $f'(\theta)$ is given by $\hat{f}'_{CLL}(\theta;\nu) = \hat{\beta}_1$.

To obtain the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, one may proceed as follows. Consider the matrix

$$\boldsymbol{\Theta} = \begin{pmatrix} 1 & \sin(\Theta_1 - \theta) \\ \vdots & \vdots \\ 1 & \sin(\Theta_n - \theta) \end{pmatrix},$$

$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ the response vector, $W = \operatorname{diag}\{K_\nu(\theta - \Theta_1), \ldots, K_\nu(\theta - \Theta_n)\}$ a diagonal weight matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$ the parameter vector (where t denotes the transpose vector). Then, the minimization problem (2.14) can be written as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} (\boldsymbol{Y} - \boldsymbol{\Theta}\boldsymbol{\beta})^t W (\boldsymbol{Y} - \boldsymbol{\Theta}\boldsymbol{\beta}).$$

Assuming the non–singularity of $\boldsymbol{\Theta}^t W \boldsymbol{\Theta}$, standard weighted least squares theory yields $\hat{\boldsymbol{\beta}} = (\boldsymbol{\Theta}^t W \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^t W \boldsymbol{Y}$. In particular, $\hat{f}_{CLL}(\theta;\nu) = \boldsymbol{e_1}^t (\boldsymbol{\Theta}^t W \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^t W \boldsymbol{Y}$ where $\boldsymbol{e_1} = (1,0)^t$.

Since the Local Linear estimator is a linear estimator, the value of the estimator at the design points, $\hat{\boldsymbol{f}}_{CLL}$, or at any grid of locations $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ denoted by $\hat{\boldsymbol{f}}_{CLL,\boldsymbol{\theta}}$, can be obtained in the following way, respectively:

$$\hat{\boldsymbol{f}}_{CLL} = L_\nu \boldsymbol{Y}$$

$$\hat{\boldsymbol{f}}_{CLL,\boldsymbol{\theta}} = L_{\nu,\boldsymbol{\theta}} \boldsymbol{Y}$$

where L_ν is an $(n \times n)$ matrix (the hat matrix) and $L_{\nu,\boldsymbol{\theta}}$ is an $(N \times n)$ matrix. Both matrices are defined in Appendix B.

If the regression function at θ is locally approximated by a constant instead of using a trigonometric polynomial, the Nadaraya–Watson estimator for circular–linear data is obtained:

$$\hat{f}_{CNW}(\theta;\nu) = \frac{\sum_{i=1}^{n} Y_i K_\nu(\theta - \Theta_i)}{\sum_{i=1}^{n} K_\nu(\theta - \Theta_i)}. \tag{2.15}$$

Nadaraya–Watson estimator is also a linear smoother and so, it may be written using matrix notation.

Smoothing parameter selection

The concentration parameter ν in the kernel function $K_\nu(\cdot)$ controls the degree of smoothing of Nadaraya–Watson and Local Linear estimators. Large values of ν lead to undersmoothed estimations of the regression curve, tending to an interpolation of the data. On the other hand, small values of ν result in a global averaging, oversmoothing the local features in the data.

As for density estimation, choosing the smoothing parameter is of crucial importance in regression estimation. A simple and widely used procedure for smoothing parameter selection in the linear regression setting is least squares cross–validation, which may be extended to the case of circular kernel regression. This method chooses ν as the value minimizing

$$CV(\nu) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{f}^{-i}(\Theta_i;\nu)\right]^2, \tag{2.16}$$

where $\hat{f}^{-i}$ is the leave-one-out estimator for the regression function. If the Local Linear estimator is considered then $\hat{f}^{-i} \equiv \hat{f}^{-i}_{CLL}$ whereas $\hat{f}^{-i} \equiv \hat{f}^{-i}_{CNW}$ for the Nadaraya–Watson estimator.

2.3.2 Periodic smoothing splines

Periodic smoothing splines, introduced by Cogburn and Davis (1974), are a variant of the usual smoothing spline estimators. Such estimators are useful when the mean response function is assumed to be smooth and periodic on an interval. When the covariate is periodic with period $T = 2\pi$, periodic smoothing splines offer an alternative to circular kernel smoothers described in the previous section. Since results that will be shown along this section are valid for any periodic covariate, the independent variable will be denoted by X (instead of Θ).

Let $\{(X_i, Y_i), i = 1, \ldots, n\} \in [0, T) \times \mathbb{R}$ be a random sample from (X, Y) where X is a periodic random variable with period T (the distribution of $(X + T)$ coincides with the distribution of X) and Y is a linear random variable. Assume that data are sorted across the covariate and there is no repeated data. Consider again the regression model

$$Y_i = f(X_i) + \sigma(X_i)\varepsilon_i,\ \ i = 1, \ldots, n, \tag{2.17}$$

where f is an unknown regression function that must be estimated (it is only known that f is a smooth periodic function with period T), $\sigma^2(\cdot)$ is the conditional variance of Y given X and ε_i are real–valued random variables with zero mean and unit variance.

The goal of smoothing spline estimators for the regression function is to fit a dataset using a function that reflects the key features of the data but at the same time, retains some degree of smoothness. A natural measure of smoothness associated with a function g is $\int g''(x)^2 dx$ while a standard measure of goodness of fit to the data is the residual sum of squares. Thus, an overall assessment of the quality of a candidate estimator g is provided by the penalized least squares

criterion:

$$S(g) = \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx, \tag{2.18}$$

for some $\lambda > 0$. The result of minimizing (2.18) over the class of twice continuously differentiable periodic functions with period T is the periodic smoothing splines estimator, $\hat{f}_\lambda$, of the regression function f.

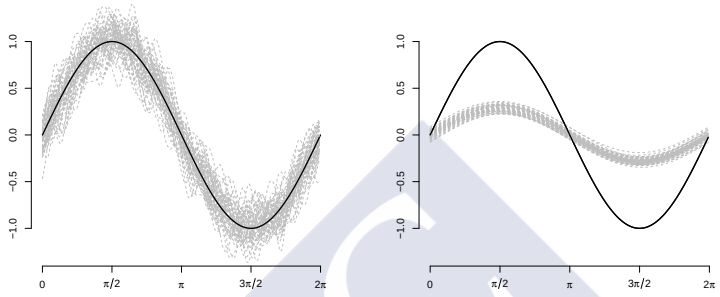

Figure 2.7: Fitted periodic smoothing splines (gray curves) with $\lambda = 0.001$ (left panel) and $\lambda = 100$ (right panel) from 50 simulated random samples of size $n = 250$ from model (2.17) where the design points are equally spaced in the interval $[0, 2\pi)$, $f(x) = \sin(x)$ and the errors are normally distributed errors with variance $\sigma^2 = 0.25$.

The periodic smoothing spline estimator, $\hat{f}_\lambda$, also depends on a smoothing parameter λ which controls the degree of smoothing of the estimator. When λ is large, a premium is being placed on smoothness and potential estimators with large second derivatives are penalized. Conversely, a small value of λ corresponds to more emphasis on goodness of fit, with $\lambda = 0$ giving an estimator that interpolates the data, i.e., small values of λ tend to provide undersmoothed estimators. The effect of λ can be seen in Figure 2.7.

It will be shown that, for $\lambda > 0$, the function $\hat{f}_\lambda$ that minimizes $S(g)$ in (2.18) is neccesarily a periodic cubic spline on $[X_1, X_{n+1}]$ with knots at sampling points X_i, $i = 1, \ldots, (n+1)$, where $X_{n+1} = X_1 + T$. This considerably simplifies the problem of finding $\hat{f}_\lambda$ since the space of periodic cubic splines can be parametrized. So, first of all, the definition of a periodic cubic spline will be given.

Let $t_1, \ldots, t_{m+1}$ be $(m+1)$ $(m \geq 2)$ real numbers with $t_1 < t_2 < \ldots < t_{m+1}$. A cubic spline on $[t_1, t_{m+1}]$ with knots at $t_1, \ldots, t_{m+1}$, is a function s that coincides with a third–order polynomial s_i on each subinterval $[t_i, t_{i+1}]$, $i = 1, \ldots, m$, and such that $s \in \mathcal{C}^2[t_1, t_{m+1}]$, where $\mathcal{C}^2[t_1, t_{m+1}]$ denotes the space of all functions on $[t_1, t_{m+1}]$ that have continuous first and second order derivatives. In other words, s is a cubic spline on $[t_1, t_{m+1}]$ if, for each $i = 1, \ldots, m$, there

exist real numbers a_i, b_i, c_i, d_i (the "spline coefficients" of s) such that, for every t in $[t_i, t_{i+1}]$,

$$s(t) = s_i(t) = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3, \quad i = 1, \ldots, m.$$

Furthermore, the continuity of s, s' and s'' on $[t_1, t_{m+1}]$ implies that at each interior knot t_i, with $i = 2, \ldots, m$,

$$s_{i-1}(t_i) = s_i(t_i); \quad s'_{i-1}(t_i) = s'_i(t_i); \quad s''_{i-1}(t_i) = s''_i(t_i).$$

The cubic spline s is said to be periodic, with period $T = t_{m+1} - t_1$, if it satisfies the following additional conditions:

$$s(t_{m+1}) = s(t_1); \quad s'(t_{m+1}) = s'(t_1); \quad s''(t_{m+1}) = s''(t_1).$$

For the particular case of a spline on the circle, t_1 and t_{m+1} can be taken as 0 and 2π respectively.

In order to prove that the solution of (2.18) is a periodic cubic spline, let $g \in \mathcal{C}^2[X_1, X_{n+1}]$ be any periodic curve in $[X_1, X_{n+1}]$ that is not a periodic cubic spline with knots at the X_i, $i = 1, \ldots, (n + 1)$. Hence, by Theorem 3.24 in Nürnberger (1989), there exists a unique periodic spline, s_g, that interpolates the values $\{X_i, g(X_i)\}$ for $i = 1, \ldots, (n + 1)$. Since by definition $s_g(X_i) = g(X_i)$ for all i, it follows that $\sum_{i=1}^n (Y_i - s_g(X_i))^2 = \sum_{i=1}^n (Y_i - g(X_i))^2$. Because of the optimality properties of the periodic cubic spline interpolant (see Nürnberger, 1989, Theorem 3.25), $\int_{X_1}^{X_{n+1}} (s''_g)^2 < \int_{X_1}^{X_{n+1}} (g'')^2$, and hence, for $\lambda > 0$, $S(s_g) < S(g)$. This means that, unless g itself is a periodic cubic spline (in which case the equality holds in the previous inequalites), a periodic cubic spline which attains a smaller values of the penalized sum of squares can be found. Therefore, it follows that the minimizer of (2.18), $\hat{f}_\lambda$, must be a periodic cubic spline.

Since $\hat{f}_\lambda$ is a periodic cubic spline, the problem of minimizing the penalized sum of squares (2.18) over the class of twice continuosly differentiable periodic functions with period T reduces to minimize (2.18) over a finite dimensional class of functions, the periodic cubic splines with knots at X_i.

Now, an explicit expression for the periodic smoothing spline estimator $\hat{f}_\lambda$ will be obtained.

Let g be a periodic cubic spline on $[X_1, X_{n+1}]$ with knots at X_i, $i = 1, \ldots, n + 1$. Denoting $g_i = g(X_i)$ and $\gamma_i = g''(X_i)$ $i = 1, \ldots, n$, the conditions that the spline must be continuous to second derivative at each knot, and that $g(X_1)$ must match $g(X_{n+1})$, up to second derivative, are equivalent to

$$R\boldsymbol{\gamma} = Q\boldsymbol{g}, \tag{2.19}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^t$ and $\boldsymbol{g} = (g_1, \ldots, g_n)^t$ and the matrices R and Q, which are given in terms of the distances betweeen succesive knots. The specific expressions of matrices R and Q are derived in Appendix C. These matrices coincide with matrices $\tilde{B}$ and $\tilde{D}$ in Wood (2006, pp. 150–151).

The roughness penalty term can be expressed as

$$\int_0^T g''(x)^2 dx = \int_{X_1}^{X_{n+1}} g''(x)^2 dx = \boldsymbol{g} Q R^{-1} Q \boldsymbol{g},$$

as it can be seen in Appendix C or Wood (2006, p. 151).

Hence, the penalized sum squares (2.18) can be rewritten as

$$S(g) = (\boldsymbol{Y} - \boldsymbol{g})^t(\boldsymbol{Y} - \boldsymbol{g}) + \lambda \boldsymbol{g}^t K \boldsymbol{g}$$

where $K = QR^{-1}Q$.

The necessary condition $\frac{dS}{dg_i} = 0,\ i = 1, \ldots, n$ for a minimum of the above expression leads to a linear system

$$(I_n + \lambda K)\boldsymbol{g} = \boldsymbol{Y}, \tag{2.20}$$

where I_n denotes the identity matrix of dimension n. Since λK is non–negative definite, the matrix $(I_n + \lambda K)$ is strictly positive definite. Thus, it follows that (2.18) has a unique minimum given by the solution of (2.20). Therefore, the periodic cubic spline estimator $\hat{f}_\lambda$ evaluated in the sample points, $\hat{\boldsymbol{f}}_\lambda = (\hat{f}_\lambda(X_1), \ldots, \hat{f}_\lambda(X_n))^t$, is given by

$$\hat{\boldsymbol{f}}_\lambda = (I_n + \lambda K)^{-1}\boldsymbol{Y} = A_\lambda \boldsymbol{Y}, \tag{2.21}$$

where $A_\lambda = (I_n + \lambda K)^{-1}$ is the hat matrix. Note that the smoothing spline estimator is also a linear smoother.

Following Green and Silverman (1994, pp. 22–23), the value of the estimator and its derivative can be obtained for any point $x \in [X_1, X_{n+1})$. Let $\boldsymbol{x} = (x_1, \ldots, x_N)^t$ be a grid of locations with $x_i \in [X_1, X_{n+1})$. Then, by the computations done in Appendix C and bearing in mind equations (2.19) and (2.21), it holds that

$$\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = \left[C - DR^{-1}Q\right]\hat{\boldsymbol{f}}_\lambda = MA_\lambda \boldsymbol{Y}, \tag{2.22}$$

$$\hat{\boldsymbol{f}}'_{\lambda,\boldsymbol{x}} = \left[\tilde{C} - \tilde{D}R^{-1}Q\right]\hat{\boldsymbol{f}}_\lambda = \tilde{M}A_\lambda \boldsymbol{Y}, \tag{2.23}$$

where $\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = (\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_N))^t$ and $\hat{\boldsymbol{f}'}_{\lambda,\boldsymbol{x}} = (\hat{f}'_\lambda(x_1), \ldots, \hat{f}'_\lambda(x_N))^t$. Matrices C, D, $\tilde{C}$ and $\tilde{D}$ are defined as in Appendix C.

Weighted smoothing

When there are ties among the original design points, i.e., at the point X_i, independent observations $Y_{ij},\ j = 1, \ldots, m_i$ are taken all with mean $g(X_i)$ then, the penalized sum of squares (2.18) is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{ij} - g(X_i))^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx. \tag{2.24}$$

The problem of minimizing (2.24) is shown to be equivalent to minimizing the penalized weighted sum of squares

$$\sum_{i=1}^{n} m_i(\overline{Y}_i - g(X_i))^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx$$

where $\overline{Y}_i$ is the average of the observations at X_i. This is a particular case of weighted smoothing where the square residuals $(Y_{ij} - g(X_i))^2$ are weighted by the number of observations at X_i, m_i.

Suppose that $w_1, \ldots, w_n$ are strictly positive weights such that $\sum_{i=1}^n w_i = n$, each w_i associated with (X_i, Y_i), $i = 1, \ldots, n$, respectively. In this case, the estimator of f, namely $\hat{f}_\lambda^W$, is obtained from the minimization of the penalized weighted sum of squares

$$S_W(g) = \sum_{i=1}^{n} w_i \left[Y_i - g(X_i)\right]^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx.$$

It can be shown that the penalized weighted sum of squares $S_W(g)$ is uniquely minimized over the class of twice continuously differentiable periodic functions with period T by the periodic cubic spline with knots at the desing points X_i given by

$$\hat{\boldsymbol{f}}_\lambda^W = (W + \lambda K)^{-1} W \boldsymbol{Y} = A_\lambda^W \boldsymbol{Y},$$

with $A_\lambda^W = (W + \lambda K)^{-1} W$ where W is a diagonal matrix with diagonal elements w_i, $i = 1, \ldots, n$. The proof is exactly parallel to the argument set out before, replacing the residual sum of squares by the weighted residual sum of squares. In order to compute the value of the estimator and its derivative in $\boldsymbol{x}$, the matrix A_λ in equations (2.22) and (2.23) will be replaced by the matrix A_λ^W.

Binning

For handling large datasets, unequally spaced designs and/or datasets with repeated measures, the binned implementation may be used. The key idea of binned implementation is to reduce the number of evaluations of the estimator, based on the fact that many of these evaluations are nearly the same.

Binning consists in dividing the entire range of data points into some equally spaced bins and distributing data into bins. Simple binning and linear binning are described in Fan and Marron (1994) for linear data. These procedures can be easily adapted to circular data, taking into account that the distance between observations θ_1 and θ_2 from a circular variable with period T is computed as $\min\{|\theta_1 - \theta_2|, T - |\theta_1 - \theta_2|\}$.

Let $t_1 < \ldots < t_m$ be m points equally spaced on $[0, T)$ such that $t_j = (j-1)\Delta$ for $j = 1, \ldots, m$ where $\Delta = T/m$. These points are known as bin centers.

Simple binning consists in replacing each X_i by the nearest bin center t_j. Hence, the weight that observation X_i $(i = 1, \ldots, n)$ gives to the grid point t_j $(j = 1, \ldots, m)$ is

$$n_{i,j}^{simple} = \mathbb{I}_{\{\min\{|X_i - t_j|, T - |X_i - t_j|\} < \Delta/2\}},$$

where $\mathbb{I}$ denotes the indicator function.

The idea behind linear binning is to split the unit mass of each data observation between the two closest bin centers. The fraction assigned to each side is taken to be proportional to the

distance from X_i to the nearest bin center on the opposite side. Hence, the weight that observation X_i $(i = 1, \ldots, n)$ gives to the grid point t_j $(j = 1, \ldots, m)$ is

$$n_{i,j}^{linear} = \left(1 - \frac{\min\{|X_i - t_j|, T - |X_i - t_j|\}}{\Delta}\right)_+,$$

where the subscript + denotes the positive part.

Thus, the original data $\{(X_i, Y_i), i = 1, \ldots, n\}$ are summarized by the binned data

$$\left\{(t_j, \overline{Y_j}, n_j), j = 1, \ldots, m\right\},$$

where $\overline{Y_j} = \sum_{i=1}^{n} n_{i,j} Y_i / n_j$ being $n_j = \sum_{i=1}^{n} n_{i,j}$ and $n_{i,j}$ corresponds to $n_{i,j}^{simple}$ for simple binning and to $n_{i,j}^{linear}$ for linear binning.

Once the binned data are computed, the periodic smoothing spline will be calculated as before, but applied to the data $\left\{(t_j, \overline{Y_j}), j = 1, \ldots, m\right\}$ with weights n_j. In this case, the weighted formulation is required.

Smoothing parameter selection

As noted before, the periodic smoothing spline estimator depends on the smoothing parameter λ. There are several methods for selecting a value for λ such as the cross–validation method. This method selects the value of $\lambda \geq 0$ that minimizes the cross–validation function

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{f}_\lambda^{-i}(X_i)\right)^2$$

or

$$CV(\lambda) = \sum_{i=1}^{n} w_i \left(Y_i - (\hat{f}_\lambda^W)^{-i}(X_i)\right)^2$$

for the weighted case (see, e.g., Green and Silverman, 1994) or when a binned implementation is used. Here, $\hat{f}_\lambda^{-i}$ and $(\hat{f}_\lambda^W)^{-i}$ denote the periodic smoothing spline estimator computed after removing the observation (X_i, Y_i) from the data.

2.3.3 Simulation study

In this section, the performance of kernel and periodic smoothing spline estimators for circular–linear regression will be studied through a simulation study. For that purpose, a variety of scenarios have been simulated, combining two sampling designs and four regression models, with different degrees of complexity. Specifically, for the circular covariate, two sampling schemes have been considered:

- Design 1. $\Theta_1, \ldots, \Theta_n$ are equally spaced in the interval $[0, 2\pi)$.

- Design 2. $\Theta_1, \ldots, \Theta_n$ are the quantiles of a $vM(\pi, 1)$.

Design 1 corresponds to a fixed design setting, whereas Design 2 provides a covariate distribution with a mode at π.

The errors ε_i in (2.13) and (2.17) are generated independently from a normal distribution with zero mean and variance 0.5. Samples of size $n = 50, 250$ and 500 have been generated according to model (2.13) with four different regression functions:

- Model 1. $f_1(\theta) = \sin(\theta)$.
- Model 2. $f_2(\theta) = \cos(\theta)$.
- Model 3. $f_3(\theta) = \sin\left(\frac{3}{2}\left(\theta - \frac{\pi}{2}\right)\right) + \frac{2\sqrt{2}}{3}\cos\left(\frac{\theta}{3}\right)$.
- Model 4. $f_4(\theta) = \sin(\theta - 1.2\pi) + 3e^{-10\left(15\frac{(\theta-\pi)}{2\pi}\right)^2}$.

Models 1 and 2 represent simple sinusoidal trends in the circle and Models 3 and 4 represent more complex trend structures, as shown in Figure 2.8.

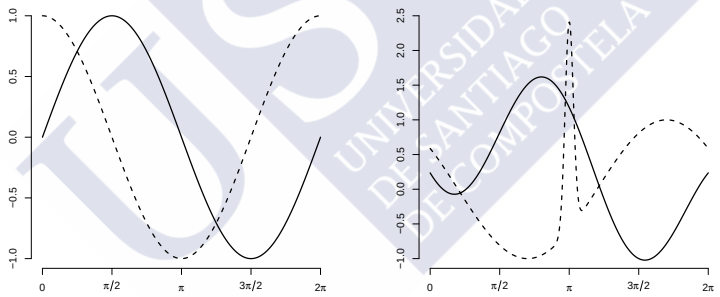

Figure 2.8: Linear representation of trends for regression model (2.13). Left panel: Model 1 (solid line) and Model 2 (dashed line). Right panel: Model 3 (solid line) and Model 4 (dashed line).

For checking the global performance of each estimator for f, the integrated squared error ISE (defined as in (2.8)) is computed. For all the estimators, smoothing parameters are selected by cross–validation. Table 2.8 shows the average ISE out of 1000 replicates. From these results, it is observed that Nadaraya–Watson and Local Linear estimators perform similarly for all the regression models considered and for both designs whereas, in general, the periodic smoothing spline estimator provide better results in all the scenarios. It can be seen that results for Models 1 and 2 are very similar and the errors for Model 4, which is the most complicated model, are larger. Moreover, all the estimators provide better results for Design 1. As it was expected, the performance of the estimators improves when the sample size is increased.

Model	n	$\hat{f}_{CNW}$	$\hat{f}_{CLL}$	$\hat{f}_{\lambda}$	$\hat{f}_{CNW}$	$\hat{f}_{CLL}$	$\hat{f}_{\lambda}$
1	50	0.3796 (0.2614)	0.3761 (0.2631)	0.2763 (0.2361)	0.5472 (0.4454)	0.5413 (0.4858)	0.3432 (0.3243)
	250	0.0901 (0.0573)	0.0898 (0.0574)	0.0605 (0.0467)	0.1620 (0.1128)	0.1517 (0.1331)	0.0870 (0.0783)
	500	0.0504 (0.0293)	0.0503 (0.0292)	0.0328 (0.0243)	0.0874 (0.0544)	0.0791 (0.0587)	0.0452 (0.0357)
2	50	0.4219 (0.3667)	0.4244 (0.3654)	0.2704 (0.2056)	0.5977 (0.4370)	0.6177 (0.4950)	0.3955 (0.3507)
	250	0.0910 (0.0593)	0.0913 (0.0592)	0.0617 (0.0462)	0.1778 (0.1150)	0.1487 (0.1271)	0.0948 (0.0827)
	500	0.0521 (0.0303)	0.0522 (0.0303)	0.0345 (0.0249)	0.0941 (0.0519)	0.0769 (0.0518)	0.0492 (0.0375)
3	50	0.4992 (0.3553)	0.5014 (0.3551)	0.4246 (0.2644)	0.7352 (0.4496)	0.6819 (0.4704)	0.6120 (0.3899)
	250	0.1198 (0.0588)	0.1200 (0.0590)	0.1017 (0.0530)	0.2163 (0.1252)	0.2027 (0.1385)	0.1674 (0.1028)
	500	0.0696 (0.0318)	0.0696 (0.0318)	0.0584 (0.0292)	0.1196 (0.0569)	0.1105 (0.0554)	0.1062 (0.0582)
4	50	1.2275 (0.3123)	1.2279 (0.3122)	1.3265 (0.2792)	1.5460 (0.4955)	1.5028 (0.5114)	1.3874 (0.4564)
	250	0.3581 (0.0821)	0.3582 (0.0822)	0.3815 (0.0888)	0.5837 (0.1747)	0.6441 (0.2059)	0.5032 (0.1922)
	500	0.2143 (0.0437)	0.2144 (0.0437)	0.2168 (0.0483)	0.3139 (0.0792)	0.3249 (0.0837)	0.2796 (0.0882)

Table 2.8: Average ISE for different regression estimators and sampling designs. Design 1 (left part) and Design 2 (right part). $\hat{f}_{CNW}$: Nadaraya–Watson estimator; $\hat{f}_{CLL}$: Local Linear estimator; $\hat{f}_{\lambda}$: Periodic smoothing spline estimator.

2.3.4 Real data analysis

Example. Exploring wind paterns. In order to explore the relation between the wind speed and the wind direction during winter season in the Atlantic coast of Galicia, the Nadaraya–Watson and Local Linear estimators and the periodic smoothing spline estimator are applied to the data of wind speed and wind direction introduced in Section 1.3. In Figure 2.9, the circular and linear representation of Nadaraya–Watson, Local Linear and periodic smoothing spline estimators for the real data are shown. Smoothing parameter selection is done using the cross–validation criterion.

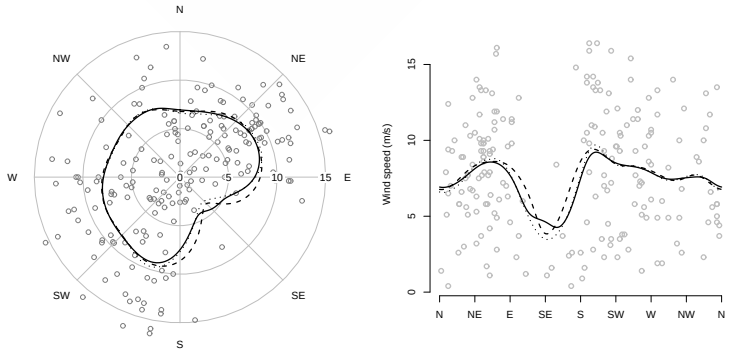

Figure 2.9: Circular (left panel) and linear (right panel) representations of the Nadaraya–Watson estimator (dashed curve), Local linear estimator (solid curve) and periodic smoothing spline estimator (dotted curve) with cross–validation smoothing parameter for wind speed (m/s) with respect to wind direction.

From these plots, it is observed that wind speed is slightly higher for winds coming from NE ans SW whereas the valley around SE indicates that winds coming from S are calm. Since the modes around NE and SW are not very marked, one may think that there is no effect of the wind direction over the wind speed. In Chapter 3, a new method for the exploratory analysis of circular data will allow to assess if the effect of wind direction over wind speed is significant.

Chapter 3

Assessment of significant features in nonparametric curve estimates

3.1 Introduction

Both for density and regression estimation, the smoothing parameter controls the global appearance of the estimator and its dependence on the sample, in such a way that an unsuitable choice of this value may provide a misleading estimate of the density or the regression curve. Hence, the assessment of the statistical significance of the observed features through the smoothed curve should be required for not compromising the extracted conclusions. An approach to circumvent the choice of the smoothing parameter, and still be able to assess global structure features in the curve, is given by the SiZer (SIgnificant ZERo crossings of derivatives) method developed by Chaudhuri and Marron (1999) for the analysis of linear data.

The original SiZer is a visualization method based on nonparametric curve estimates which addresses the question of which features observed in a smoothed curve are really present, or represent an important underlying structure, and not simply artifacts of the sampling noise from a scale–space perspective. In the nonparametric curve estimation context, the scale–space framework is given by a family of smoothers indexed by the smoothing parameter. SiZer considers a wide range of smoothing values, which avoids the problem of smoothing parameter selection, whilst peaks and troughs are identified by finding the regions of significant gradient (zero crossings of the derivative), presenting this information in a simple visual way by the SiZer map.

SiZer was originally designed for kernel smoothers and adapted later to smoothing splines estimators for the regression setting by Marron and Zhang (2005). Several extensions of SiZer have been proposed in the statistical literature, making it possible to transfer this graphical tool to a variety of contexts such as local likelihood (Li and Marron, 2005), dependent data (Rondonotti et al., 2007) and survival data (Marron and de Uña Álvarez, 2004), among others. Moreover, SiZer for linear variables has been successfully applied in many different scientific fields. For example,

Rudge (2008) used this method to find peaks in geochemical distributions; Sonderegger et al. (2009) considered SiZer to detect threshold in ecological data and Rydén (2010) applied SiZer to determine a possible increasing trend in hurricane activity in the North Atlantic.

In the special setting of circular data, both for density and regression estimation, the adaptation of SiZer ideas must take into account the circular nature of the variable. This particular scenario involves, specifically: (1) the assessment of the variability in the derivatives of the estimators, both for density and regression, through the computation of standard deviations and appropriate quantiles; (2) the development of a suitable visualization device to facilitate the practitioner the output interpretation. Bearing these premises in mind, the SiZer ideas can be fitted to the circular data setting yielding the CircSiZer plot presented in this dissertation.

This chapter is organized as follows. Section 3.2 gives a short introduction to the CircSiZer. Section 3.3 is devoted to the development of the CircSiZer method, both for density and regression. The construction of the CircSiZer map and its interpretation is given in Section 3.4. In Section 3.5, the performance of the CircSiZer is investigated with simulated examples. In Section 3.6, CircSiZer is used for exploring the crack distribution in cemented femoral components and describing the wind direction and the relation between wind speed and wind direction in the Galician coast during winter season. The contents of this chapter can be found in Oliveira et al. (2013a,b,f).

3.2 CircSiZer: SiZer for circular data

Before going into detail, the idea of CircSiZer will be illustrated with a simulated dataset in the density setting. Consider a random sample of size 250 from model M10 (see Appendix A and Figure 2.2). Figure 3.1 (left panel) shows the family of smoothers (gray curves) for a wide range of values of the smoothing parameter. Some estimates are very different from the theoretical curve (black line) whereas other estimates show the same structure but, none of them is very good at attaining the goal of recovering the original density. However, instead of trying to recover the original density, the goal of CircSiZer is to determine which observed features in the gray curves are important underlying structure and which are sampling artifacts. For that purpose, since features like peaks and valleys of a curve can be characterized in terms of zero crossings of its derivative, CircSiZer focuses on finding regions where the gray curves are significantly increasing/decreasing and displaying this information in a color map as the one shown in Figure 3.1 (right panel).

Blue color indicates that the slope of the curve is significantly increasing, red color indicates that it is significantly decreasing, purple color indicates that the slope is not significantly different from zero and gray color is used to indicate that data are too sparse for determining the behaviour of the curve. Thus, taking the sense marked by the arrow as the positive sense of rotation, a significant peak can be identified when a blue region is followed by a red region, and a significant trough by the reverse, i.e., when a red region is followed by a blue region. Since each ring in the CircSiZer map corresponds to each value of the smoothing parameter in the family of gray curves,

in the CircSiZer displayed in Figure 3.1 it can be seen that the two modes of the model M10 are correctly identified for large values of the smoothing parameter (small rings in the CircSiZer map).

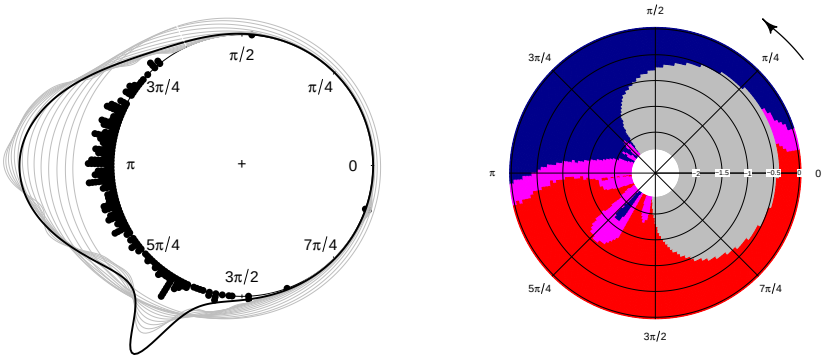

Figure 3.1: Left panel: family of kernel density estimates (gray curves) indexed by the smoothing parameter for a sample of 250 data (points over the circle) from model M10 (solid line). Right panel: CircSiZer map for kernel density estimator. For reading CircSiZer, take counterclockwise sense of rotation (marked by the arrow in the upper–right corner). Values of ν, which are transformed to $-\log_{10}$ scale, are indicated along the radius.

In the next sections, the development and construction of the CircSiZer map is given in detail. Along these sections, f will denote both the density and the regression function. The estimator of f will be denoted by $\hat{f}(\theta;\tau)$ where τ is the smoothing parameter. Note that τ will correspond with ν if kernel smoothers are considered and with λ for spline smoothers. Thus, the effect of the smooting parameter τ will depend on the kind of smoother since the effect of ν and λ is not the same. While small (large) values of ν will provide oversmoothed (undersmoothed) estimators, small (large) values of λ will provide undersmoothed (oversmoothed) estimators.

3.3 Development of CircSiZer

As noted before, features like peaks and valleys of a smooth curve can be characterized in terms of zero crossings of the derivative. Hence, the significance of such features can be judged from statistical significance of zero crossings or equivalenty by the sign changes of derivatives. With CircSiZer, significance is based on confidence intervals for the derivative of the smoothed underlying curve (density or regression).

The usual inferential approach in nonparametric statistics places the spotlight on the true underlying curve f by doing inference on it, in particular, based on confidence bands. A crucial problem in nonparametric estimation is that $f(\theta;\tau) \equiv \mathbb{E}(\hat{f}(\theta;\tau))$ is not necessarily equal to $f(\theta)$, due to the inherent bias of smooth estimates. In the particular case of kernel estimators, the bias is large specially for small values of τ (see Figure 3.2, left panel). Moreover, the smoothed curve or the expected curve for a certain τ, namely $f(\cdot;\tau)$, can be very different from f. The bias can

be reduced by taking large values of τ, but in this case the estimator is highly variable, depending strongly on the data sample (see Figure 3.2 right panel). However if τ is within a reasonable range, $f(\cdot;\tau)$ shows the same valley–peaks structure as f and the variability of the estimators is not large (see Figure 3.2, center panel).

Thus, Chaudhuri and Marron (1999) avoided the bias–variance trade–off problem by adopting the scale–space ideas which naturally lead to make inference on the smoothed curve $f(\cdot;\tau)$ rather than on the curve f.

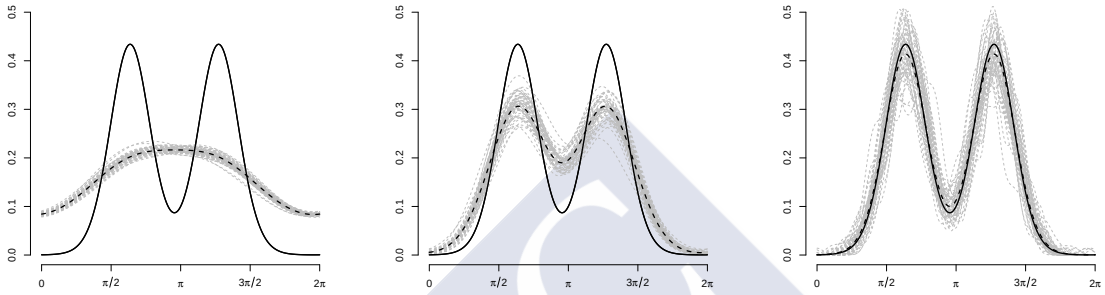

Figure 3.2: Each panel shows the density function of model M8 (solid line) defined in Appendix A, the expected curve for a certain value ν (dashed line) and kernel density estimates for 50 random samples from model M8 (gray lines) for a given value of ν. Left panel: $\nu = 1$. Center panel: $\nu = 5$. Right panel: $\nu = 50$.

Therefore, in order to assess the significance of features such as peaks and valleys, instead of constructing confidence intervals for $f'(\theta)$, CircSiZer seeks confidence intervals for the scale–space version $f'(\theta;\tau) \equiv \mathbb{E}(\hat{f}'(\theta;\tau))$ for which $\hat{f}'(\theta;\tau)$ is an unbiased estimator at each location θ and scale τ.

In the density setting, $\hat{f}'(\theta;\tau)$ is obtained by deriving the expression of the circular kernel density estimator given in (2.4). In the regression setting, the estimator of the derivative of the regression function is given by $\hat{\beta}_1$ for the Local Linear estimator (see Section 2.3.1 and Appendix B), it is obtained by deriving (2.15) for the Nadaraya–Watson estimator and, for the periodic smoothing spline estimator, its computation is detailed in Section 2.3.2.

In general, confidence intervals are of the form

$$\left(\hat{f}'(\theta;\tau) - q^{(1-\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)), \hat{f}'(\theta;\tau) - q^{(\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))\right), \tag{3.1}$$

where $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are appropriate quantiles (see Section 3.3.1) and $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))$ is an estimator of the standard deviation of $\hat{f}'(\theta;\tau)$ (see Section 3.3.2 for its calculation).

3.3.1 Computation of the quantiles

Quantiles in (3.1) may be computed in order to provide pointwise or simultaneous confidence intervals for the derivative of the smooth curve. Moreover, in each case, two approaches for computing the quantiles may be considered: a first approach based on a normal approximation and another approach using bootstrap techniques. In the subsequent sections, the different possibilities are detailed.

Pointwise normal confidence intervals

The simplest approach to confidence interval construction uses a normal approximation, i.e., it is assumed that

$$\frac{\hat{f}'(\theta;\tau) - f'(\theta;\tau)}{\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))} \sim N(0,1).$$

Hence, given a significance level α, an approximate $(1-\alpha)$ confidence inteval is obtained by taking $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ equal to the $(1-\alpha/2)$ and $\alpha/2$ quantiles of the standard normal distribution, respectively. Note that an estimation of the estandard deviation is required. Its computation will be detailed in the Section 3.3.2.

Pointwise bootstrap confidence intervals

Through the use of bootstrap, accurate intervals without imposing normal assumptions can be obtained. A possible way to get such intervals, namely the "bootstrap–t" approach (see Efron and Tibshirani, 1993, Chapter 12), is detailed below. Given a significance level α and for a fixed value of $\tau > 0$ and with θ varying in the interval $[0, 2\pi)$, the following algorithm is considered:

Step 1. Generate B bootstrap samples, i.e., random samples drawn with replacement from the data.

Step 2. For $b = 1, \ldots, B$ compute

$$Z_b^*(\theta;\tau) = \frac{\hat{f}'(\theta;\tau)^{*b} - \hat{f}'(\theta;\tau)}{\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)^{*b})}, \quad b = 1, \ldots, B, \tag{3.2}$$

where $\hat{f}'(\theta;\tau)^{*b}$ is the value of $\hat{f}'(\theta;\tau)$ for the b bootstrap sample and $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)^{*b})$ is an estimator of the standard deviation of $\hat{f}'(\theta;\tau)^{*b}$ (see Section 3.3.2 for its calculation).

Step 3. Quantiles $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ in (3.1) are given by the $(1-\alpha/2)$ and $\alpha/2$ sample quantiles of $Z_1^*(\theta;\tau), \ldots, Z_B^*(\theta;\tau)$, respectively.

Both approaches discussed above provide pointwise confidence intervals with the drawback that many features may be flagged as "significant". An alternative is to approximate simultaneous (across $[0, 2\pi)$, the entire range of Θ) confidence intervals as already proposed in Chaudhuri and Marron (1999) which is the goal of the following approaches.

Simultaneous normal confidence intervals

Chaudhuri and Marron (1999) proposed taking as quantiles for (3.1):

$$q^{(1-\alpha/2)} = -q^{(\alpha/2)} = \Phi^{-1}\left\{\frac{1+(1-\alpha)^{1/m(\tau)}}{2}\right\},$$

where Φ^{-1} is the inverse of the standard normal distribution function, α is the given significance level and $m(\tau)$ is the number of independents blocks of average size available from a dataset of size n. Specifically, $m(\tau)$ is defined as

$$m(\tau) = \frac{n}{\text{avg}_{\theta\in\mathcal{D}_\tau} ESS(\theta;\tau)},$$

where $\text{avg}_{\theta\in\mathcal{D}_\tau} ESS(\theta;\tau)$ denotes the average value of the Effective Sample Size $ESS(\theta;\tau)$ on the set $\mathcal{D}_\tau = \{\theta : ESS(\theta;\tau) \geq 5\}$.

For density estimation or regression estimation with kernel smoothers, the Effective Sample Size is given by

$$ESS(\theta;\tau) = ESS(\theta;\nu) = \frac{\sum_{i=1}^{n} K_\nu(\theta-\Theta_i)}{K_\nu(0)},$$

and for regression estimation with periodic smoothing splines, following Marron and Zhang (2005),

$$ESS(\theta;\tau) = ESS(\theta;\lambda) = \frac{n\ n_\theta}{\text{tr}(A_\lambda)},$$

where A_λ is the hat matrix (see Section 2.3.2) and tr is the trace. For a grid of points $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)$, $\boldsymbol{n_\theta} = (n_{\theta_1},\ldots,n_{\theta_N})^t$ is obtained as $\boldsymbol{n_\theta} = MA_\lambda \boldsymbol{n}$, with $\boldsymbol{n} = (n_1,\ldots,n_n)^t$ where n_i is the number of observations at the design point Θ_i.

Simultaneous bootstrap confidence intervals

As before, simultaneous confidence intervals without imposing Gaussian assumptions can be obtained by a bootstrap strategy. The algorithm is as follows:

Step 1. Generate B bootstrap samples, i.e., random samples drawn with replacement from the data.

Step 2. For $b = 1,\ldots,B$ compute

$$Z^{*b}_{inf} = \inf_{\theta\in\mathcal{D}_\tau} Z^*_b(\theta;\tau)$$

$$Z^{*b}_{sup} = \sup_{\theta\in\mathcal{D}_\tau} Z^*_b(\theta;\tau)$$

where $Z^*_b(\theta,\tau)$ is given by (3.2) and D_τ is defined as before.

Step 3. Quantile $q^{(1-\alpha/2)}$ is given the $(1-\alpha/2)$ sample quantile of $Z^{*1}_{sup},\ldots,Z^{*B}_{sup}$ and $q^{(\alpha/2)}$ is given by the $\alpha/2$ sample quantile of $Z^{*1}_{inf},\ldots,Z^{*B}_{inf}$.

Note that, in order to compute the (pointwise or simultaneous) quantiles based on bootstrap techniques, random samples drawn with replacement from the data are generated in Step 1 and so, in the regression setting, bootstrap confidence intervals are computed for random design, i.e., when the covariate values are random.

3.3.2 Estimation of the standard deviation

For the computation of confidence intervals (3.1), it is necessary to derive an expression for $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))$, and also for its bootstrap version $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)^{*b})$ involved in (3.2) when bootstrap confidence intervals are obtained. The computation procedure, for density and regression, is detailed below.

Density estimation

The main idea behind the calculation of $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))$, in the context of density estimation, is that the derivative of the circular kernel density estimator defined in (2.3) is a weighted average of the derivative of the kernel function at different locations. So, following Chaudhuri and Marron (1999) for the linear case, its variance may be estimated by

$$\begin{aligned}\widehat{\mathrm{var}}\left(\hat{f}'(\theta;\tau)\right) &= \widehat{\mathrm{var}}\left(\hat{f}'(\theta;\nu)\right) = \widehat{\mathrm{var}}\left(n^{-1}\textstyle\sum_{i=1}^{n} K'_\nu(\theta-\Theta_i)\right)\\ &= n^{-1}s^2\left(K'_\nu(\theta-\Theta_1),\ldots,K'_\nu(\theta-\Theta_n)\right), \quad 0\le\theta<2\pi,\end{aligned}$$

where s^2 is the usual sample variance of n data, which in this context is formed by the derivative of the kernel centered at each sample value Θ_i, with $i=1,\ldots,n$.

Regression estimation

For the regression setting, consider the model (2.13). For any linear smoother, as the ones studied in Chapter 2, the estimator of the derivative of the regression function evaluated in a grid $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)^t$ can be written as

$$\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}} = H\boldsymbol{Y}$$

where H is an $(N\times n)$ matrix and $\boldsymbol{Y}$ is the response vector. For the Local Linear smoother $H=\tilde{L}_{\nu,\boldsymbol{\theta}}$ (see Appendix B) and $H=\tilde{M}A_\lambda$ for periodic smoothing splines (see equation (2.23) and Appendix C), where A_λ is the hat matrix and $\tilde{M}$ is an $(N\times n)$ coefficient matrix which allows to obtain the value of the derivative of the estimator in the grid $\boldsymbol{\theta}$.

Suppose that the covariate values Θ_i are fixed (fixed design), so the matrix H is not random since it only depends on the data points $\Theta_1,\ldots,\Theta_n$ and on the grid points $\boldsymbol{\theta}$. Therefore, the variance–covariance matrix of $\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}}$ can be computed as follows

$$\mathrm{var}(\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}}) = H\Sigma H^t,$$

where $\Sigma = \text{diag}\left\{\sigma^2(\Theta_1), \sigma^2(\Theta_2), \ldots, \sigma^2(\Theta_n)\right\}$ (see Sheather, 2009, Chapter 6). If homoscedasticity is assumed, meaning that $\sigma^2 = \text{var}(\varepsilon_i)$ does not vary with θ, then following the ideas of Rice (1984) $\sigma^2(\Theta_i) = \sigma^2$ $(i = 1, \ldots, n)$ can be estimated by

$$\widehat{\sigma^2} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2.$$

For random design, the standard deviation will be approximated by bootstrap. For a given $\tau > 0$ and with θ varying in $[0, 2\pi)$, the standard deviation of $\hat{f}'(\theta; \tau)$ is estimated following the next steps:

Step 1. Generate B bootstrap samples, i.e., random samples drawn with replacement from the data.

Step 2. For each bootstrap sample, calculate $\hat{f}'^{*b}(\theta; \tau)$, with $b = 1, \ldots, B$.

Step 3. Estimate the standard deviation by the sample standard deviation of the B replicates:

$$\widehat{\text{sd}}\left(\hat{f}'(\theta; \tau)\right) = \left[s^2\left(\hat{f}'^{*1}(\theta; \tau), \ldots, \hat{f}'^{*B}(\theta; \tau)\right)\right]^{1/2}.$$

Note that the computation of bootstrap confidence intervals requires, in this case, the estimation of the standard deviation of $\hat{f}'(\theta; \tau)^{*b}$ for $b = 1, \ldots, B$. Then, a nested bootstrap is needed, increasing notably the computational cost.

3.4 CircSiZer map

Although the procedure for obtaining the confidence intervals must be carefully adapted to the circular setting, as shown along the previous section, the construction and interpretation of the CircSiZer map is fairly simple, as detailed below.

For constructing the CircSiZer map a grid of angles equally spaced in the interval $[0, 2\pi)$ and a grid of smoothing parameters are considered. According to both grids, the "skeleton" of the CircSiZer map is constructed as shown in Figure 3.3, where each ring corresponds to a value of the smoothing parameter and where each ring is divided according to the grid of angles. In this way, each pixel in the CircSiZer map is identified by a pair (θ, τ) with θ in the grid of angles and $\tau > 0$ in the grid of smoothing parameters. For assigning the colour of each pixel, a confidence interval for $f'(\theta; \nu)$, at each pair (θ, τ), is computed. Then:

- If the confidence interval is above zero, indicating that the smoothed curve is significantly increasing then, the corresponding pixel in the CircSiZer map is coloured blue.
- If the confidence interval is below zero, indicating that the smoothed curve is significantly decreasing then, the corresponding pixel in the CircSiZer map is coloured red.

- If the confidence interval contains zero, meaning that the derivative of the curve is not significantly different from zero then, the corresponding pixel in the CircSiZer map is coloured purple.

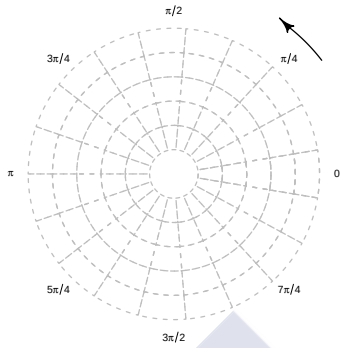

Figure 3.3: "Skeleton" of the CircSiZer map. Each ring makes reference to a value of the smoothing parameter τ and each cell in the ring makes reference to angle $\theta \in [0, 2\pi)$. The arrow in the upper–right corner indicates the sense of rotation for reading the CircSiZer map.

To determine the gray areas (not enough data), the same rule as in Chaudhuri and Marron (1999) is considered: for each (θ, τ) the estimated effective sample size (ESS) is calculated and those regions where $\text{ESS}(\theta; \tau) < 5$ are shaded gray.

In the CircSiZer map, the values of the smoothing parameter τ, in $-\log_{10}$ scale when $\tau \equiv \nu$ and in $\log_{10}$ scale when $\tau \equiv \lambda$, will be indicated along the radius. In this way, small rings will refer to mild smoothing, whereas large grids corresponds to strong smoothing. Moreover, an arrow in the upper–right corner will indicate the sense of rotation for reading the CircSiZer map.

3.5 Performance of CircSiZer

In this section, the performance of CircSiZer map constructed with the different confidence intervals discussed in Section 3.3 will be investigated in the density setting (Section 3.5.1) and in the regression setting (Section 3.5.2). Throughout these sections, statistical significance will be assessed with a significance level $\alpha = 0.05$.

3.5.1 Density setting

Firstly, in the density setting, the performance of pointwise confidence intervals will be analyzed. As introduced in previous sections, the CircSiZer map may be obtained by considered pointwise or simultaneous confidence intervals. However, when using the first alternative, it should be beared in mind that the bands obtained are usually narrower. Hence, the problem of constructing the

CircSiZer map with pointwise confidence intervals is that some features may be identified as significant when they are not. The pointwise coverages of normal and bootstrap confidence intervals have been studied for the circular uniform model (model M1, see Appendix A), considering different values of the smoothing parameter. Results obtained from 1000 samples of size $n = 250$ from this model are summarized in Figure 3.4. These plots show, by using a palette of colors from green to white (from negative to positive), the differences between the empirical coverage of pointwise confidence intervals and the nominal confidence level (0.95) for a grid of smoothing parameters (represented along the ordinate axis in $-\log$ scale). In these plots, it is observed that normal and bootstrap confidence intervals behave similarly for small values of the smoothing parameter ($-\log \nu$ between -1 and 0). This fact is also reflected in Table 3.1 which shows, for each value of the smoothing parameter, the number of modes detected by the CircSiZer maps using normal and boostrap pointwise confidence intervals.

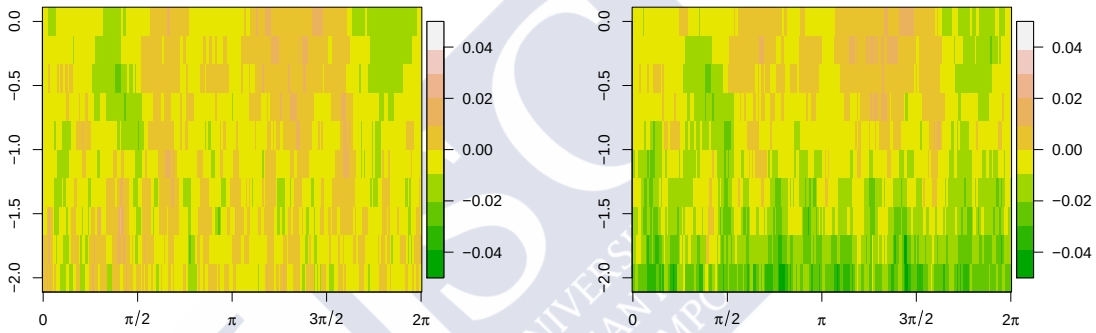

Figure 3.4: Left panel: differences between the coverage of pointwise normal confidence intervals and the nominal confidence level. Right panel: differences between the coverage of pointwise bootstrap confidence intervals and the nominal confidence level. Results were obtained from 1000 replicates of sample size $n = 250$ from M1, the uniform model.

The pointwise coverages of normal and bootstrap confidence intervals have been also studied for model M2, which corresponds to a von Mises centered at π and contentration parameter 1 (see Appendix A). As before, 1000 samples of size $n = 250$ from this model have been considered and results were summarized using the same representation (see Figure 3.5). For large smoothing parameters, the right panel (corresponding to bootstrap confidence intervals) shows two green regions around 0 and 2π indicating that the coverage of the pointwise bootstrap confidence intervals are smaller than the nominal confidence level (0.95). Specifically, these green patches appear in zones with sparse data, affecting the coverage of pointwise bootstrap confidence intervals. However, this behaviour is not observed in the left panel which corresponds to normal confidence intervals. In view of Table 3.2, it seems that the lack of data for some combinations of location and scale (angle and smoothing parameter), spurious modes will be flagged more frequently by the bootstrap method than by the normal approximation. This is also observed in Table 3.1. So,

from now on, when a pointwise CircSiZer map is plotted the normal approximation is considered for constructing the confidence intervals. In addition, given that no resampling is required, there are also some benefits in terms of computational burden.

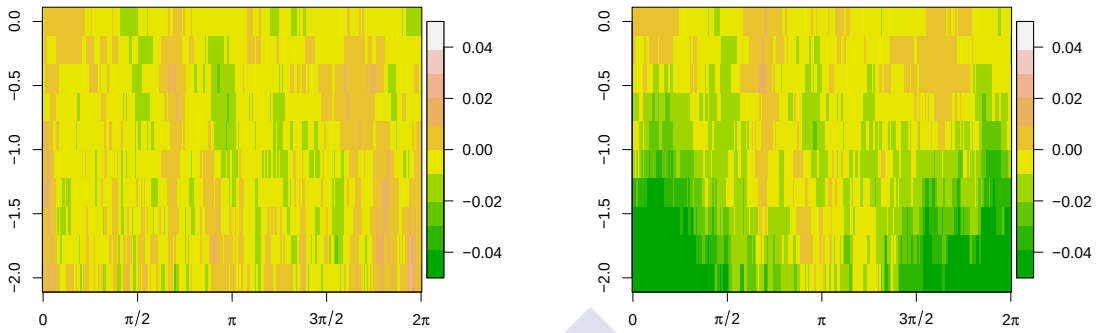

Figure 3.5: Left panel: differences between the coverage of pointwise normal confidence intervals and the nominal confidence level. Right panel: differences between the coverage of pointwise bootstrap confidence intervals and the nominal confidence level. Results were obtained from 1000 replicates of sample size $n = 250$ from model M2.

Model M1	No. modes \ $-\log_{10}(\nu)$	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
Bootstrap	0	86	206	342	463	598	712	775	842	882	897
	1	358	247	480	445	353	268	215	152	118	103
	2	362	243	153	86	47	20	10	6	0	0
	3	149	74	24	6	2	0	0	0	0	0
	4	37	10	1	0	0	0	0	0	0	0
	5	7	0	0	0	0	0	0	0	0	0
	6	1	0	0	0	0	0	0	0	0	0
Normal approximation	0	212	330	444	554	643	740	794	851	888	899
	1	480	482	449	384	324	246	197	145	112	101
	2	240	154	98	61	31	14	9	4	0	0
	3	56	32	8	1	2	0	0	0	0	0
	4	11	2	1	0	0	0	0	0	0	0
	5	1	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0

Table 3.1: Number of modes flagged by CircSiZer map with pointwise bootstrap and normal confidence intervals and number of times that these modes are identified. Results were obtained from 1000 replicates of sample size $n = 250$ from the uniform model. Shady cells show the number of times that the right number of modes of $f(\cdot, \nu)$ is detected.

Although augmented when using bootstrap confidence intervals, for large values of the smoothing parameters, more artificial modes are identified as significant by both methods (see Tables 3.1

and 3.2). Therefore, not very large values of this parameter should be used in the construction of the pointwise CircSiZer map.

Model M2	No. modes \ $-\log_{10}(\nu)$	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
Bootstrap	0	99	83	49	12	2	0	0	0	0	0
	1	533	662	793	928	990	1000	1000	1000	1000	1000
	2	295	218	149	59	8	0	0	0	0	0
	3	62	35	8	1	0	0	0	0	0	0
	4	9	2	1	0	0	0	0	0	0	0
	5	2	0	0	0	0	0	0	0	0	0
Normal approximation	0	220	159	104	22	4	0	0	0	0	0
	1	601	706	815	945	994	1000	1000	1000	1000	1000
	2	153	124	79	33	2	0	0	0	0	0
	3	25	11	2	0	0	0	0	0	0	0
	4	1	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0

Table 3.2: Number of modes flagged by CircSiZer map with pointwise bootstrap and normal confidence intervals and number of times that these modes are identified. Results were obtained from 1000 replicates of sample size $n = 250$ from model M2. Shady cells show the number of times that the right number of modes of $f(\cdot, \nu)$ is detected.

Now, the coverage of normal and bootstrap simultaneous confidence intervals will be compared. To this end, 1000 random samples of size $n = 250$ from models M1, M2, M10, M18 and M20 from Chapter 2 (see also Appendix A) are considered. These models include simple models such as the uniform and the von Mises models, a bimodal model which shows a highly peaked mode and two complex models which have three and four modes, respectively. For constructing the CircSiZer maps, a grid of 250 angles equally spaced in the interval $[0, 2\pi)$ and a grid of 10 equally spaced values of the smoothing parameter (in $-\log_{10}$ scale) between -2 and 0 are considered. For each simulation of bootstrap confidence intervals, 1000 bootstrap samples will be generated.

For each scenario, a way of summarizing the information provided by the 1000 CircSiZer maps is to report the percentage of times that, over the 1000 replicates, the simultaneous bootstrap confidence interval and simultaneous normal confidence interval contain the curve $f'(\cdot; \nu)$ for the grid of the smoothing parameters considered. Results are summarized in Figure 3.6. In these plots, the curves are piecewise linear, with nodes at each value of the smoothing parameter considered which are represented along the abscissa axis in $-\log_{10}$ scale. Ideally, all of these values should be close to $(1 - \alpha)100\% = 95\%$. In Figure 3.6, it can ben seen that the coverage of simultaneous bootstrap confidence intervals (represented by a dashed line) is closer to the nominal value (represented by a dotted line) for the five models and for all the smoothing parameters, whereas the coverage of simultaneous normal confidence interval (represented by a solid line) is about 80%. Moreover, Figure 3.7 shows (for model M18) that the coverage of simultaneous normal

confidence intervals does not improve for larger sample sizes. Hence, from now on, the bootstrap procedure will be considered for constructing the simultaneous CircSiZer map, unless otherwise specified.

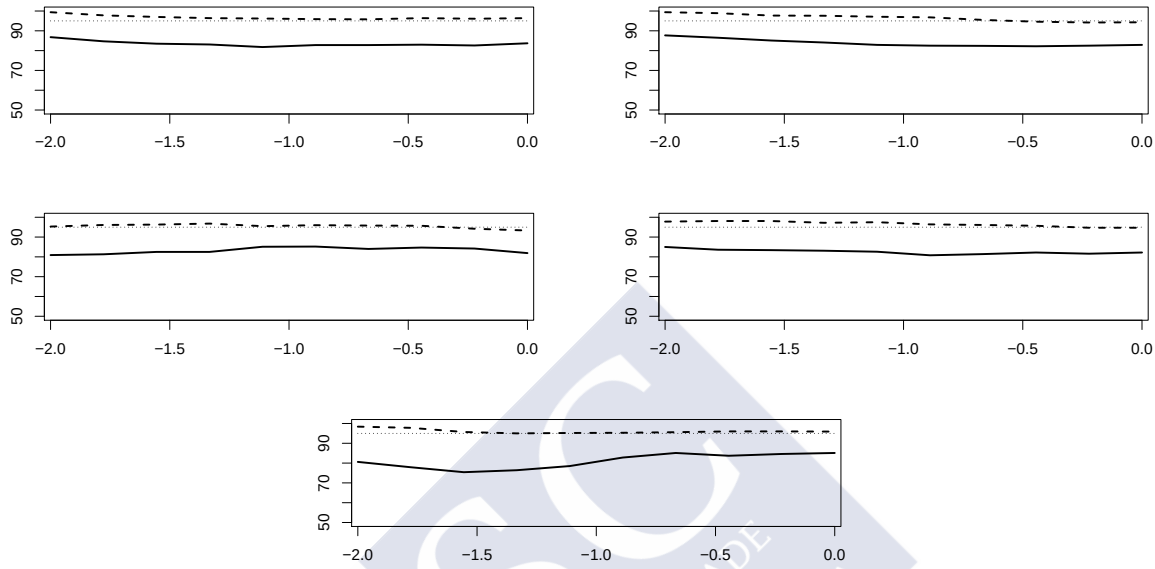

Figure 3.6: Percentage of times over 1000 replicates of sample size $n = 250$ that the simultaneous bootstrap confidence interval (dashed line) and simultaneous normal confidence interval (solid line) contain the expected curve for certain values of ν (represented along the coordinates axis in $-\log(\nu)$ scale). Dotted horizontal line represents the nominal confidence level. From left to right and from top to bottom: models M1, M2, M10, M18 and M20, respectively.

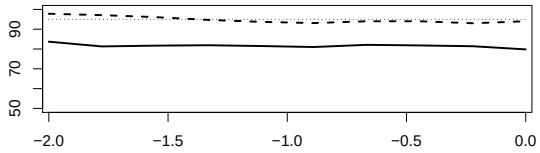

Figure 3.7: Percentage of times over 1000 replicates of sample size $n = 1000$ that the simultaneous bootstrap confidence interval (dashed line) and simultaneous normal confidence interval (solid line) contain the expected curve for certain values of ν (represented along the coordinates axis in $-\log(\nu)$ scale). Results for model M18. Dotted horizontal line represents the nominal confidence level.

As an exploratory tool for data analysis, CircSiZer may be also used for detecting modes. Table 3.3 shows the number of modes (blue–red patterns) detected by the simultaneous CircSiZer

map for each value of the smoothing parameter and for each model considered. Note that, even using simultaneous confidence intervals, some artificial structure may be flagged. For example, for the uniform model (M1) no trend should be observed, regardless the smoothing parameter. However, Table 3.3 shows that some modes may be identified as significant in this case. This is due to the fact that these intervals are simultaneous over $\theta \in [0, 2\pi)$ but not over ν and so, for each fixed value of the smoothing parameter, significant structures may be identified 5% of the times. From this table, it is also observed that, for complex models such as models M10, M18 and M20, simultaneous CircSiZer map does not behaves well at attaining the goal of detecting the modes present in the model. For example, for model M18, the trimodal structure of the model only is identified by one or two out of 1000 CircSiZer maps. As the sample size increases, simultaneous CircSiZer map behaves better as shown in Table 3.4 for model M18. However, for the exploratory analysis of a dataset, it may be interesting to explore the simultaneous CircSiZer map and the pointwise CircSiZer map, specially when the sample size is not very large.

Model	No. modes \ $-\log_{10}(\nu)$	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
M1	0	1000	999	997	999	997	996	997	998	996	992
	1	0	1	3	1	3	4	3	2	4	8
M2	0	1000	1000	994	941	640	115	2	0	0	0
	1	0	0	6	59	360	885	998	1000	1000	10000
M10	0	544	172	5	0	0	0	0	0	0	0
	1	445	785	954	998	1000	1000	1000	1000	1000	1000
	2	11	43	41	2	0	0	0	0	0	0
M18	0	793	481	184	35	2	0	0	0	0	0
	1	198	481	754	953	998	1000	1000	1000	1000	1000
	2	8	37	62	12	0	0	0	0	0	0
	3	1	1	0	0	0	0	0	0	0	0
M20	0	397	80	4	1	0	0	0	7	34	246
	1	453	369	94	12	5	7	31	239	507	708
	2	147	513	605	161	129	933	969	754	459	46
	3	3	38	273	462	463	59	0	0	0	0
	4	0	0	24	364	403	1	0	0	0	0

Table 3.3: Number of modes flagged by CircSiZer map with simultaneous bootstrap confidence intervals and number of times that these modes are identified. Results were obtained from 1000 replicates of sample size $n = 250$ from models M1, M2, M10, M18 and M20. Shady cells show the number of times that the right number of modes of $f(\cdot, \nu)$ is detected.

For example, consider a sample of size $n = 250$ from model M20. Simultaneous CircSiZer map shown in Figure 3.8 (center panel) does not allow to identify the cuatrimodal structure of the model and it can be seen that only two modes are identified for small values of the smoothing parameter whereas for larger values of this parameter only the mode around $3\pi/4$ is detected. However, the pointwise CircSiZer map shown in Figure 3.8 (right panel) allows to identify the

four modes.

Model	No. modes \ $-\log_{10}(\nu)$	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
M18	0	6	0	0	0	0	0	0	0	0	0
	1	396	96	76	530	999	1000	1000	1000	1000	1000
	2	439	393	398	399	1	0	0	0	0	0
	3	159	511	526	71	0	0	0	0	0	0

Table 3.4: Number of modes flagged by CircSiZer map with simultaneous bootstrap confidence intervals and number of times that these modes are identified. Results were obtained from 1000 replicates of sample size $n = 1000$ from model M18. Shady cells show the number of times that the right number of modes of $f(\cdot, \nu)$ is detected.

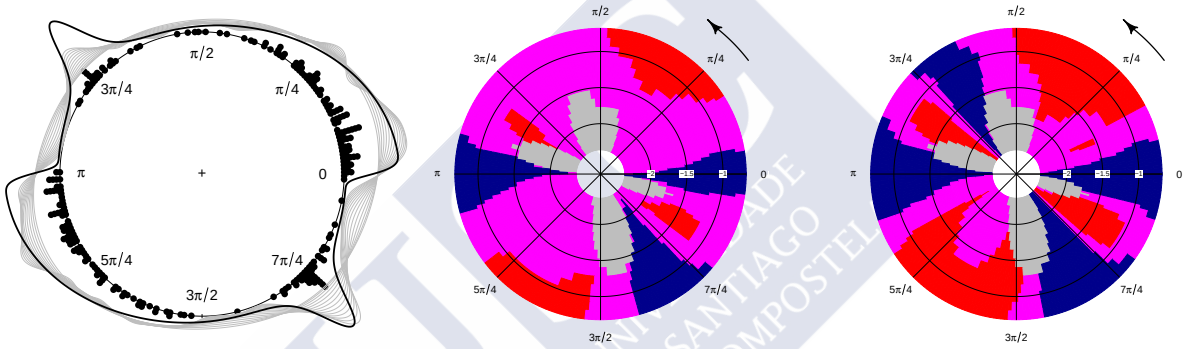

Figure 3.8: Left panels: family of kernel density estimates (gray lines) indexed by the smoothing parameter for a sample of size $n = 250$ (points over the circle) from model M20 (black line). Center and right panels: simultaneous and pointwise CircSiZer map for kernel density estimates.

As noted before, the problem of using pointwise confidence intervals is that some features may be identified as significant when they are not, specially for large values of the smoothing parameter. For illustration purposes, a random sample of size $n = 250$ from model M2 has been considered. In this case, the simultaneous CircSiZer map identifies the unimodal structure of the model perfectly (see Figure 3.9, center panel). However, the pointwise CircSiZer map (see Figure 3.9, right panel) identifies, for large values of the smoothing parameter, some features which are not really present in the underlying density indicating that two modes are significant. Therefore, it is recommend not consider very large values of the smoothing parameter.

3.5.2 Regression setting

For the regression setting, the coverage of simultaneous normal confidence intervals has been studied for simulated data from model (2.13) with $f(\theta) = 0$ in the case of equally spaced design

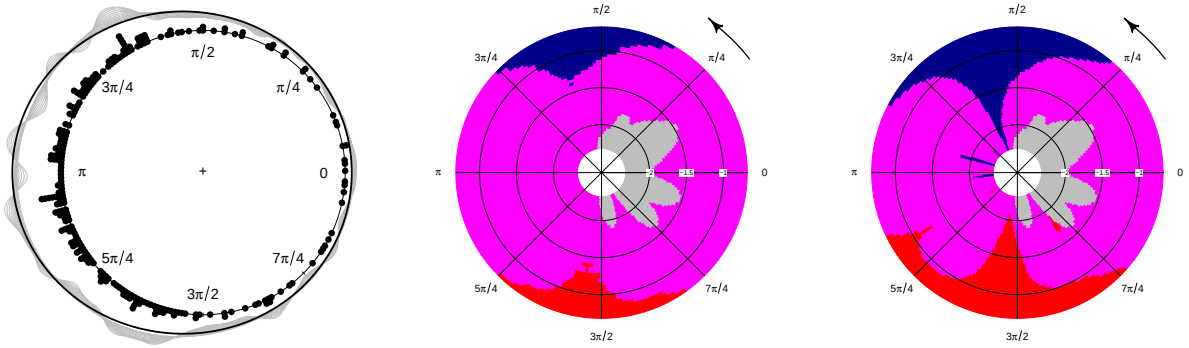

Figure 3.9: Left panel: family of kernel density estimates (gray lines) indexed by the smoothing parameter for a sample of size $n = 250$ (points over the circle) from model M2 (black line). Center and right panel: simultaneous and pointwise CircSiZer map for kernel density estimates.

regression and normally distributed errors with zero mean and unit variance. For doing the simulation study, 1000 random samples of size $n = 250$ have been generated. Using the same representation than in Figure 3.6 for summarizing the results, Figure 3.10 shows that the coverage of simultaneous normal confidence intervals is smaller than the nominal value, both for the Local Linear estimator (left panel) and the periodic smoothing spline estimator (right panel). As for the density setting, the coverage of simultaneous normal confidence intervals is about 80%.

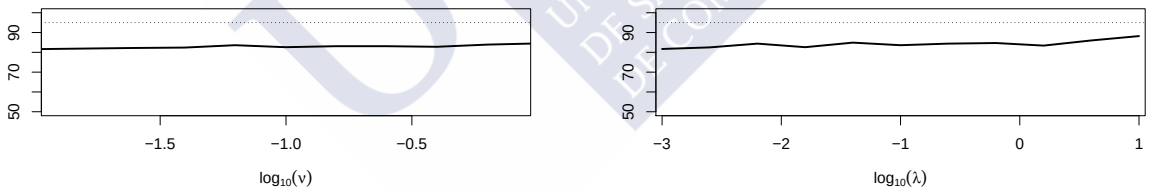

Figure 3.10: Percentage of times over 1000 replicates of sample size $n = 250$ that the simultaneous normal confidence interval (solid line) contains the expected curve for certain values of the smoothing parameter. Left panel: Local Linear estimator. Right panel: periodic smoothing spline estimator. Design points are equally spaced in the interval $[0, 2\pi)$ and responses are generated according to model (2.13) with $f(\theta) = 0$ and normally distributed errors with zero mean and unit variance. Dotted horizontal line represents the nominal confidence level.

For random design, the coverage of simultaneous bootstrap intervals has not been computed since it involves the variance estimation for each bootstrap sample which would entail a very high computational cost. However, it is expected that its performance is similar as in the density setting.

For illustrating the performance of simultaneous bootstrap CircSiZer map in the regression

setting, a sample of 250 observations has been considered, where the design points are uniformly distributed in the interval $[0, 2\pi)$ and the responses are generated according to the model (2.13), with $f(\theta) = \sin(\theta)$ and normally distributed errors with variance $\sigma^2 = 0.5$. Figure 3.11 shows the corresponding CircSiZer maps obtained by using the Local Linear estimator (center panel) and the smoothing spline estimator (righ panel). In both figures, the unimodal structure of the sine function in $[0, 2\pi)$ is clearly brought out by the CircSiZer maps.

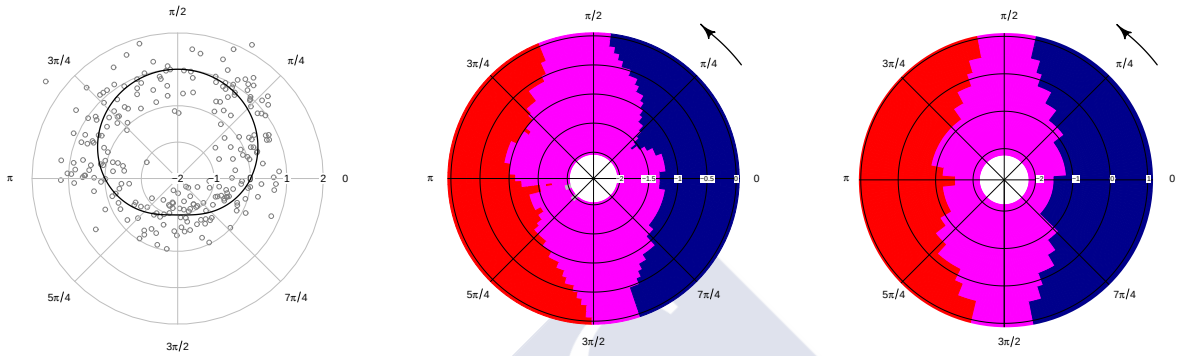

Figure 3.11: Left panel: sample of size $n = 250$ (gray points) from model (2.13), where the design points are uniformly distributed in the interval $[0, 2\pi)$, $f(\theta) = \sin(\theta)$ (black line) and errors are normally distributed with zero mean and variance $\sigma^2 = 0.5$. Center panel: simultaneous CircSiZer map for kernel regression estimates using the Local Linear estimator. Right panel: simultaneous CircSiZer map for periodic smoothing splines estimates.

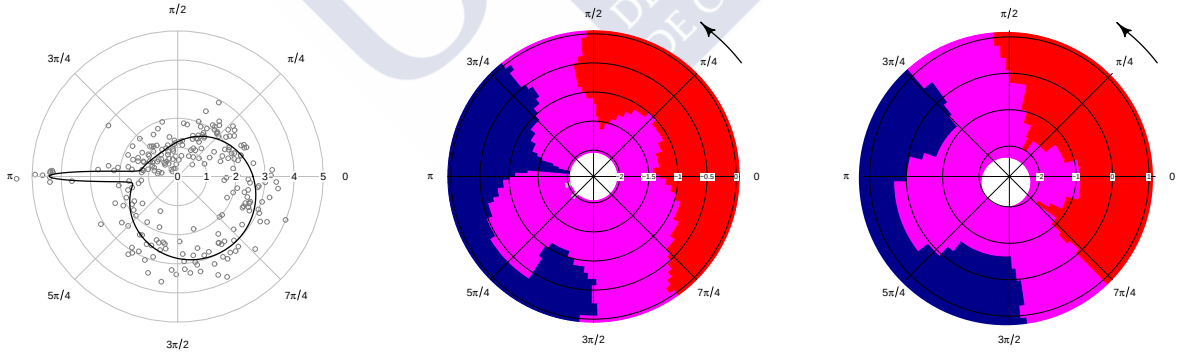

Figure 3.12: Left panel: sample of size $n = 250$ (gray points) from model (2.13), where the design points are uniformly distributed in the interval $[0, 2\pi)$, $f(\theta) = 2+\sin(\theta-1.2\pi)+3e^{-10(15(\theta-\pi)/(2\pi))^2}$ (black line) and errors are normally distributed with zero mean and variance $\sigma^2 = 0.5$. Center panel: simultaneous CircSiZer map for kernel regression estimates using the Local Linear estimator. Right panel: simultaneous CircSiZer map for periodic smoothing splines estimates.

Now, consider a sample of 250 observations where the design points are uniformly distributed in the interval $[0, 2\pi)$ and the responses are generated according to the model (2.13), with

$f(\theta) = \sin(\theta - 1.2\pi) + 3e^{-10(15(\theta-\pi)/(2\pi))^2}$ and normally distributed errors with variance $\sigma^2 = 0.5$. From Figure 3.12 (left panel), it is clear that the regression model to estimate may present some challenges given the highly peaked mode centered in π and another less concentrated mode about $7\pi/4$. For the sample size considered, simultaneous CircSiZer map using the Local Linear estimator (see Figure 3.12, center panel) does not allow to identify the two modes, only the mode around $7\pi/4$ is detected. The same behaviour is observed using the periodic smoothing spline estimator (see Figure 3.12, right panel). However, if the sample size is increased ($n = 500$) the simultaneous CircSiZer map flags the two modes as significant, as shown in Figure 3.13 (center panel).

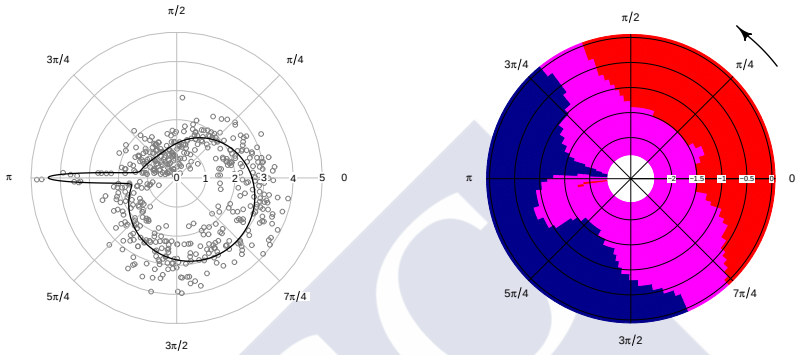

Figure 3.13: Left panel: sample of size $n = 500$ (gray points) from model (2.13), where the design points are uniformly distributed in the interval $[0, 2\pi)$, $f(\theta) = 2+\sin(\theta-1.2\pi)+3e^{-10(15(\theta-\pi)/(2\pi))^2}$ (black line) and errors are normally distributed with zero mean and variance $\sigma^2 = 0.5$. Right panel: simultaneous CircSiZer map for kernel regression estimates using the Local Linear estimator.

As in the density setting, the exploration of pointwise CircSiZer maps may help to identify the structure of the model.

3.6 Real data analysis

In this section, CircSiZer is applied to the analysis of three datasets introduced in Section 1.3: temperature cycle data, cracks in cemented femoral components data and wind speed and wind direction data.

Example 1. Temperature cycle changes. In Section 2.2.3, data from hourly temperature cycle changes, that is, temperatures changing from positive to negative or viceversa, were analyzed. In this case, a final question was posed: is the mode shown in Figure 2.6 significant? The corresponding simultaneous CircSiZer map shown in Figure 3.14 indicates that the mode around 11 a.m. is significant. This means that changes from positive to negative temperatures may happen at any hour along the day but the changes from negative to positive are concentrated around midday.

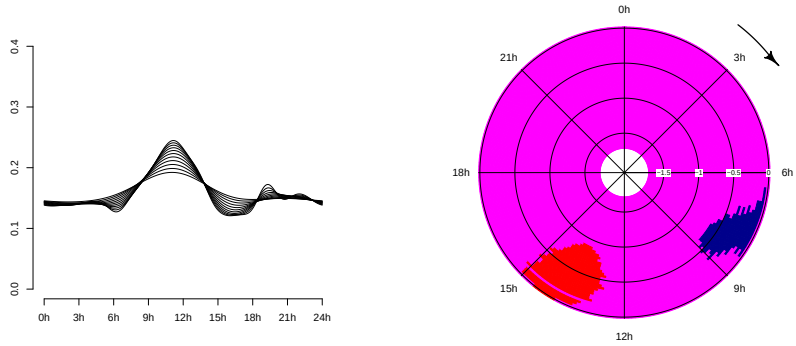

Figure 3.14: Left panel: family of kernel density estimates indexed by the smoothing parameter for data of changes in cycles of temperature. Right panel: simultaneous CircSiZer map for kernel density estimates.

Example 2. Exploring crack distribution in cemented femoral components. The performance of CircSiZer is illustrated by the analysis of the real dataset concerning angular positions of cracks in the cement mantle in a hip implant. The goal of the study is to determine if there exists a preferred direction for cracks in the cement mantle. As in Mann et al. (2003) the crack distribution for the proximal (sections at 10–50 mm) and distal regions (sections at 80–110 mm) of the cement mantle are analyzed separately. The corresponding simultaneous and pointwise CircSiZer maps for the angular position of the cracks for one femur are shown in Figure 3.15 for proximal (top panels) and distal (bottom panels) regions. Simultaneous CircSiZer maps (center panels) shows that the crack distribution around the cement mantle is not uniform in both cases, and a preferred direction exists in each case being cracks concentrated around the posterior direction (270°) for proximal regions and around the anterior–medial direction (110°) of the mantle for distal regions. The same preferred directions are flagged by the pointwise CircSiZer maps (left panels). In this latter case, other modes are also identified as significant but it is probably due to the pointwise nature of the intervals.

Example 3. Exploring wind patterns. The practical usefulness of the proposed CircSiZer map is now illustrated by the analysis of the real dataset concerning wind direction and speed in the Atlantic coast of Galicia (NW Spain). As it was described in Section 1.3, the dataset consists of hourly observations of wind direction (in degrees) and wind speed (in m/s) in winter season (from November to February), from 2003 until 2012. The main aim of these data is to describe the wind pattern in the Galician coast during winter season, focusing on the most significative wind directions and their relation with wind speed.

In order to avoid the dependence present between consecutive measurements in the time series, the autocorrelation functions were studied. Observations taken with a lag period of 95 hours can be considered as uncorrelated, providing a final dataset with about 200 values. With this lag

period, all the day hours are represented in the sample.

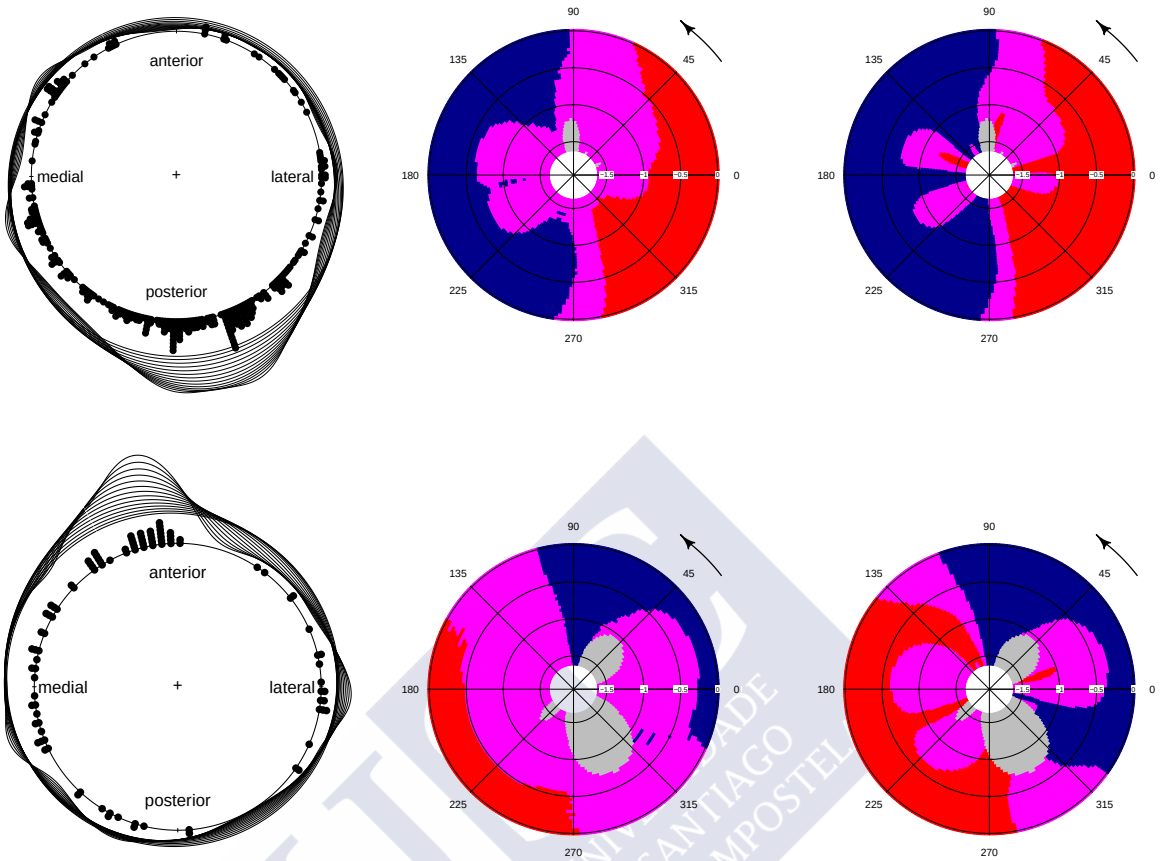

Figure 3.15: Left panels: family of kernel density estimates indexed by the smoothing parameter for data of angular positions of craks in proximal (top panel) and distal (bottom panel) regions. Center panels: simultaneous CircSiZer maps for kernel density estimates in proximal (top panel) and distal (bottom panel) regions. Right panels: pointwise CircSiZer maps for kernel density estimates in proximal (top panel) and distal (bottom panel) regions.

Figure 3.16 shows the CircSiZer maps for wind directions. In this case, both the simultaneous and pointwise CircSiZer maps distinguish two significant modes which indicate that winds in winter period come mostly from NE and SW. Moreover, they show that winds coming from SE are not frequent at all, being this fact reflected by the absence of data in the SE sector (gray shaded area).

CircSiZer maps for exploring the relation between wind speed as a response and wind direction as a covariate, are shown in Figure 3.17. In this case, the simultaneous bootstrap CircSiZer map (center panel) only shows a small blue region in the SE–S sector. However, CircSizer map with pointwise normal confidence intervals (right panel) identifies (even for very small values of the smoothing parameter) a blue region in the N–NE sector followed by a red region in the E–SE sector which means that the wind speed increases with northeasterly winds and begins to decrease

with easterly winds. Then a gray region around SE indicates that winds from this direction are not frequent and then, a blue region is flagged indicating that the wind speed increases with southerly winds.

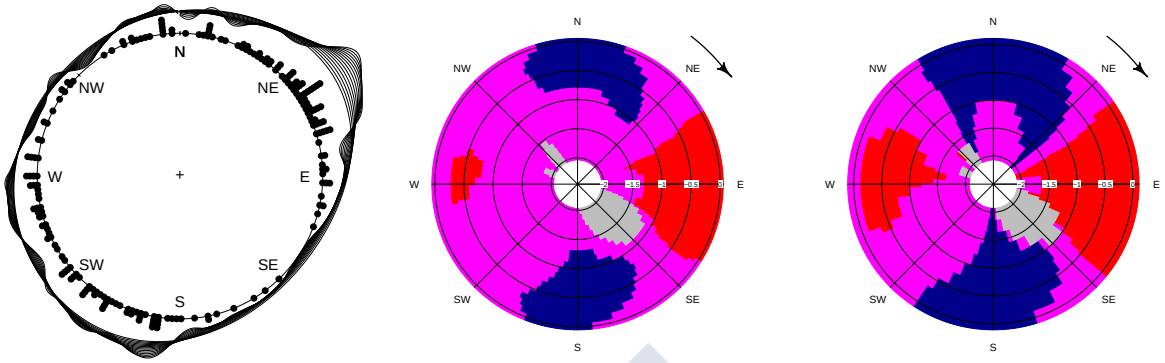

Figure 3.16: Left panel: family of kernel density estimates indexed by the smoothing parameter for data of wind direction. Center and right panel: simultaneous and pointwise CircSiZer map for kernel density estimates.

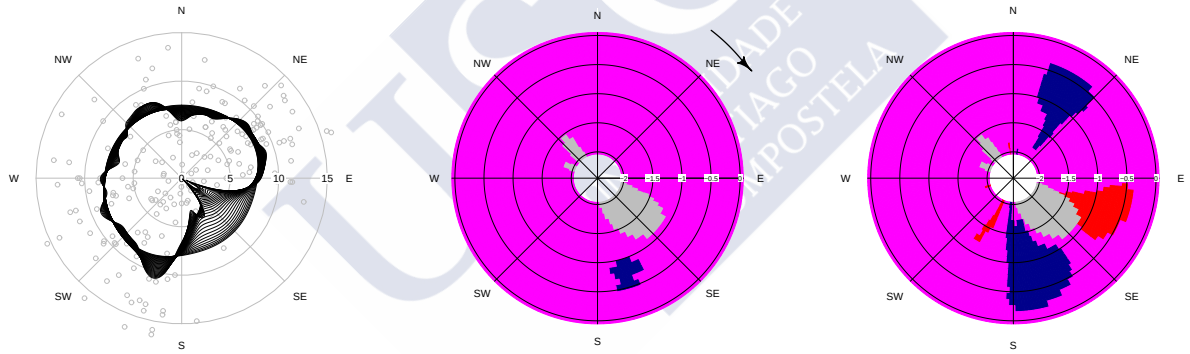

Figure 3.17: Left panel: family of kernel regression estimates indexed by the smoothing parameter for wind speed (m/s) with respect wind direction. Center and right panel: simultaneous and pointwise CircSiZer map for kernel regression estimates.

It should be also noticed that the results shown in the CircSiZer maps reflect to some extent the buoy location at the north–western corner of the Iberian Peninsula and the coastline orientation, roughly SW–NE. For that reason, in the CircSizer map for the density (Figure 3.16), a red area around E followed by a gray area around SE appears in the CircSiZer maps, indicating that winds from that directions are limited. The blue area at S indicates an increment in southerly winds. In addition, a recent study by Sousa et al. (2013) points out that S–SW winds dominate in autumn–winter and N–NE winds are more frequent in spring–summer but, it is also frequent that typically summer patterns appear in winter season, and viceversa. Hence, the two modes at SW and NE detected by the CircSizer maps in Figure 3.16 are justified.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

Software: NPCirc package

4.1 Introduction

As mentionted before, circular data appear in a large variety of disciplines and so, software is required so that practitioners can analyze their own datasets. For R (R Development Core Team, 2012) users, there are some packages for working with circular data, such as: `CircStats` package (Lund and Agostinelli, 2012), based on the book "*Topics in circular Statistics*" by Jammalamadaka and SenGupta (2001), which provides methods for the descriptive and inferential statistical analysis of directional data; `circular` package (Lund and Agostinelli, 2011) which is an extension of the `CircStats` package and provides functions for the statistical analysis (descriptive statistics, circular models, tests), graphical representation and some circular datasets; `CircNNTSR` package (Fernández-Durán and Gregorio-Domínguez, 2012) which implements functions for constructing circular distributions based on nonnegative trigonometric sums, estimating parameters and plotting the constructed densities; package `isocir` (Barragán and Fernández, 2012) which provides a set of routines for analyzing angular data subjected to order constraints on a unit circle; package `movMF` (Hornik and Grün, 2012) allows to draw random samples from mixtures of von Mises distributions and to proceed with parameter estimation, by using an EM algorithm.

From the parametric perspective, packages `circular`, `CircStats` and `movMF` allow to compute the density function and do random generation of mixtures of von Mises distributions but, it is not possible to do the same with mixtures of different circular distributions. From the nonparametric perspective, a specific function for kernel density estimation for circular data, and three functions for selecting the smoothing parameter (cross–validation rules proposed by Hall et al. (1987) and rule of thumb introduced by Taylor (2008)), have been already included in package `circular`. Apart from this, there is no other function for nonparametric regression estimation and smoothing parameter selection.

With the goal of complementing the available packages for circular data analysis and providing to the R users a comprehensive set of functions for nonparametric density and regression analysis

with circular data, the package `NPCirc` has been built. The simulation studies and real data analysis done along the previous chapters have been carried out in the R computing environment using this package. Note that, R routines related with periodic smoothing splines are not still included in the package but they will be included in future updates. The `NPCirc` package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=NPCirc` (Oliveira et al., 2013e).

This chapter provides an overview of the contents of the package. All functions included in the package are described, detailing its usage, and illustrated with some examples in Section 4.2. The contents of this chapter can be found in Oliveira et al. (2013d).

4.2 Description and illustration of the NPCirc package

In this package, the circular kernel density estimator with an up–to–date collection of smoothing parameter selection procedures described in Section 2.2 are included. For the regression setting, Nadaraya–Watson and Local Linear estimators for a linear response and a circular covariate jointly with the least squares cross–validations rule for smoothing parameter choice, introduced in Section 2.3, have been implemented. CircSiZer maps for density and regression using pointwise bootstrap confidence intervals presented in Chapter 3, can be obtained. `NPCirc` package also contains functions for generating data and obtaining densities of a variety of circular models, and mixtures of them. Specifically, the collection of circular models presented in Chapter 2, can be directly generated. Morever the package `NPCirc` includes most of the datasets introduced in Chapter 1 and analyzed along this thesis.

The list of functions and datasets available in `NPCirc` with a brief explanation of each of them can be seen in Table 4.1, and the complete documentation of the package, including the description of the functions, its arguments and its usage, is available in Appendix D.

Throughout this section, random samples will be generated by fixing `set.seed(1)`, so the results can be reproduced by the user.

4.2.1 Functions `dcircmix` and `rcircmix`

Function `dcircmix` allows to compute the density function of a circular distribution (circular uniform, von Mises, cardioid, wrapped Cauchy, wrapped normal, wrapped skew-normal) or the density of a mixture of these distributions. Function `rcircmix` allows for random generation from a circular distribution or from a mixture of circular distributions. Both functions have an argument called `model` which allows to specify a model among the ones considered in Chapter 2 and Appendix A. For example, the density function of model M20 in a grid of 250 points between 0 and 2π can be obtained by:

Dataset	**Description**
`cross.beds1`	Cross–beds azimuths (I)
`cross.beds2`	Cross–beds (II)
`cycle.changes`	Cycle changes
`dragonfly`	Orientation of dragonflies
`speed.wind, speed.wind2`	Wind speed and wind direction data
`temp.wind`	Temperature and wind direction data
Function	**Description**
`circsizer.density`	CircSiZer map for density
`circsizer.regression`	CircSiZer map for regression
`dcircmix`	Density function of mixtures of circular distributions
`rcircmix`	Random generation from mixtures of circular distributions
`kern.den.circ`	Nonparametric circular kernel density estimation
`kern.reg.circ`	Nonparametric circular kernel regression estimation
`nu.CV`	Cross–validation for density estimation
`nu.LSCV.reg`	Least squares cross–validation for circular–linear regression estimation
`nu.boot`	Bootstrap method for density estimation
`nu.pi`	Plug–in rule for density estimation
`nu.rt`	Rule of thumb for density estimation

Table 4.1: Summary of `NPCirc` package contents. Top part: data sets included in the package. Bottom part: available functions.

```
R> t <- seq(0, 2*pi, length=500)
R> f20 <- dcircmix(x=t, model=20)
```

and 100 random deviates from the same model can be obtained by:

```
R> data20 <- rcircmix(n=100, model=20)
```

Both the model density curve and the random sample are plotted in Figure 4.1 (left panel). Funtions from `circular` package allow to plot the sample over a circle and the circular density as shown in the following lines

```
R> plot(circular(data20), shrink=1.2, stack=TRUE)
R> lines(circular(t), f20)
```

which provide Figure 4.1 (left panel).

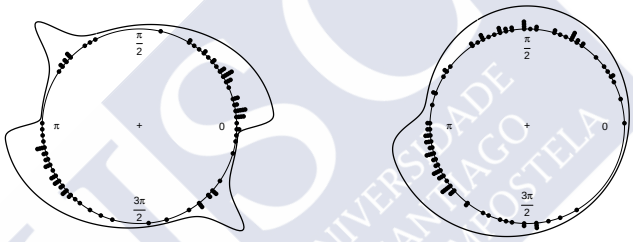

Figure 4.1: Left panel: density function (solid line) of model M20 and random sample of size $n = 100$ (dots on the circle) from the same model. Right panel: density function (solid line) of the mixture model $0.5 \cdot vM(0, 5) + 0.5 \cdot WSN(\pi, 1, 10)$ and random sample of size n=100 (dots on the circle) from the same model.

Apart from the predefined models from Chapter 2, the density function or random sample from a circular mixture model can be obtained by using the same functions by specifying the distributions that participate in the mixture through argument `dist` and the parameters of each distribution by means of argument `param`. For example, a mixture in the same proportion of a von Mises $WN(\pi/2, 0.7)$ and a wrapped skew-normal $WSN(\pi, 1, 10)$ can be obtained with the code:

```
R> fmix <- dcircmix(x=t, model=NULL, dist=c("wn","wsn"),
+  param=list(p=c(0.5,0.5), mu=c(pi/2,pi), con=c(0.7,1), sk=c(0,10)))
```

and random deviates from the same model can be obtained by:

```
R> datamix <- rcircmix(100, model=NULL, dist=c("wn","wsn"),
+  param=list(p=c(0.5,0.5), mu=c(pi/2,pi), con=c(0.7,1), sk=c(0,10)))
```

The corresponding density function and random sample are shown in Figure 4.1 (right panel).

For generating samples of a wrapped skew–normal distribution, the function `rsn` from package `sn` is used. This function allows to generate random numbers from the skew–normal distribution.

4.2.2 Functions for density estimation

Function `kern.den.circ` computes the circular kernel density estimator defined in (2.4), with von Mises kernel. This function computes the circular kernel density function estimator for a sample of angles in radians between 0 and 2π specified by argument `x`, over the points specified by argument `t`, with a smoothing parameter selected by the user, included in argument `nu`. The output of this function is a vector with the kernel density estimated values at `t`.

From a sample of 500 data from model M11, the circular kernel density estimator can be obtained as follows:

```
R> data11 <- rcircmix(500, model=11)
R> est11 <- kern.den.circ(x=data11, t=t, nu=40)
```

The graphical display of the estimator is shown in Figure 4.2. Circular and linear representations are displayed in left and right panels, respectively. The solid line is the true underlying density and the dashed line is the kernel density estimator.

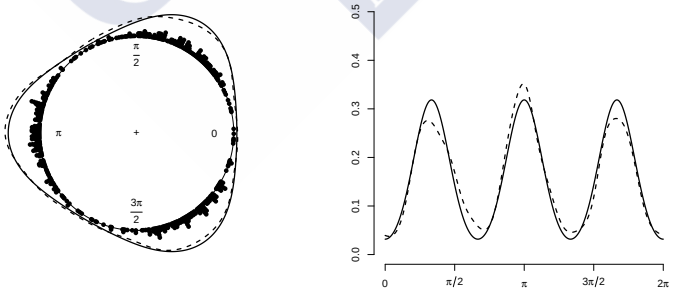

Figure 4.2: Circular (left panel) and linear (right panel) representation of the circular kernel density estimator with $\nu = 40$ (dashed line) from a sample of size 500 from model M11 and true density (solid line).

The value of the parameter `nu` in function `kern.den.circ` can be fixed by the user, as in the example above, or selected by some of the rules defined in Section 2.2.1. The available procedures for choosing the smoothing parameter will be described below. The main argument in

all the functions is the sample of angles in radians (between 0 and 2π) from which the smoothing parameter is to be computed, denoted by `x`.

Function `nu.boot` implements the bootstrap procedure proposed by Di Marzio et al. (2011). The minimum of bootstrap MISE is obtained by using the `optimize` function, which searches the minimum in the interval specified by arguments `lower` and `upper` (default values are 0 and 100, respectively) and with accuracy specified by `tol` (default `tol=0.1`). The integral in (2.11) is approximated by the Simpson's rule by a sum of `np=500` terms.

```
R> nu.boot(x=data11)

[1] 29.70744
```

Cross–validation smoothing parameters for density estimation are computed by function `nu.CV`. The cross–validation rule to be used, LSCV or LCV, will be specified by argument `method`, taking LCV as default. When the LSCV smoothing parameter is computed, the integral term in equation (2.9) is calculated using the Simpson's rule (through an internal function) and so, the argument `np` will be used. As before, the minimum/maximum is searched with `optimize` according to arguments `lower`, `upper` and `tol`.

```
R> nu.CV(x=data11, method="LCV")

[1] 32.61996

R> nu.CV(x=data11, method="LSCV")

[1] 32.55328
```

Function `nu.pi` implements the new plug–in rule proposed in Section 2.2.1. Two options are available: fix the number of components in the mixture (denoted by M in equation (1.3)) by specifying argument `M`:

```
R> nu.pi(x=data11, M=3)

[1] 41.50809
```

or select the number of components by AIC (default option):

```
R> nu.pi(x=data11, outM=TRUE)

[1] 41.50809  3.00000
```

Argument `outM=TRUE` indicates that the function also returns the number of components in the mixture. Again, the integral term is approximated by the Simpson's rule and the minimum is searched by using the function `optimize` from package `stats`.

In practice, computational problems may disable the AIC output in the implementation of the plug–in rule. These difficulties may appear in the implementation of the EM algorithm (which is available in the R package `movMF`) in Step 2.1 and/or from the numerical approximation of the integral in Step 2.2, which may not be finite. In this situation, the number of mixtures for the reference distribution is chosen as the one that provides the minimum valid AIC. Just when no results can be obtained for the different values of M, the rule of thumb proposed by Taylor (2008) is chosen. It should be noticed that in our simulation study in Chapter 2, this situation only occurred for model M1 (for $n = 100$, 1 out of 1000 samples), M3 (for $n = 100$, 7 out of 1000 samples needed $M = 1$), M6 (for $n = 100$, 1 out of 1000 samples) and M10 (for $n = 100$, 40 out of 1000 samples; for $n = 250$, 1 out of 1000 samples). It does not seem to be an issue for large sample size.

Finally, the selector proposed by Taylor (2008) for density estimation is computed by function `nu.rt`. The concentration parameter can be estimated by maximum likelihood (`robust=FALSE`):

```
R> nu.rt(x=data11, robust=FALSE)

[1] 0.2054091
```

or by the robustified procedure described in Section 2.2.1, by setting `robust=TRUE`. In this case, the argument `alpha` must be also specified:

```
R> nu.rt(x=data11, robust=TRUE, alpha=0.5)

[1] 1.124611
```

The CircSiZer map for density estimation using pointwise bootstrap confidence intervals is provided by `circsizer.density`. The main arguments in this function are `x`, the angle data sample and `NU`, a grid of positive smoothing parameters. Other arguments can be fixed: `ngrid`, integer indicating the number of equally spaced angles between 0 and 2π where the estimator is evaluated (default to `ngrid=250`); `alpha`, the significance level for assessing increasing/deacreasing patterns (default to `alpha=0.05`); and `B`, the number of bootstrap samples to estimate the standard deviation of $\hat{f}'(\theta;\nu)$ (default to `B=500`). In order to edit the graph, additional arguments can be passed to this function. The CircSiZer map in Figure 4.3 is obtained with the next code lines:

```
R> data14 <- rcircmix(250, model=14)
R> circsizer.density(data14, NU=seq(0,100,by=5), type=3, raw.data=TRUE,
+  log.scale=TRUE,zero=0,clockwise=FALSE)
```

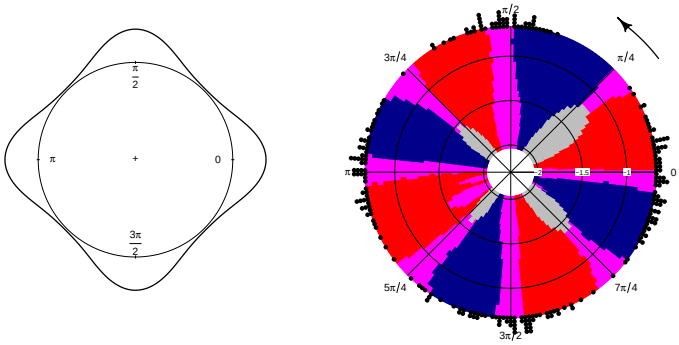

Figure 4.3: Density for model M14 (left panel) and CircSiZer map for kernel density estimator (right panel) based on 250 simulated data (dots over the circle). Peaks and valleys are identified by clockwise blue–red and red–blue patterns, respectively.

In the CircSiZer map, values of the smoothing parameter are indicated along the radius, transformed to $-\log_{10}$ scale for `log.scale=TRUE` (default option). Moreover, the arrow indicating the sense of rotation for reading the CircSiZer map is plotted according to the argument `clockwise`.

This function makes use of packages `plotrix` and `shape` for plotting the CircSiZer map.

4.2.3 Functions for regression estimation

For regression estimation, with circular covariate and linear response, `NPCirc` includes function `kern.reg.circ` which allows to compute the Local Linear and Nadaraya–Watson estimators, a function `nu.LSCV` to compute the least squares cross–validation smoothing parameters and another function `circsizer.regression` which allows to obtain the CircSiZer map using pointwise bootstrap confidence intervals for the regression setting.

The function `kern.reg.circ` implements the local linear estimator and the Nadaraya–Watson estimator for circular–linear data (circular covariate and linear response), as described in Section 2.3.1. The arguments in this function are: `x`, the sample of angles in radians (between 0 and 2π) for the circular covariate; `y`, the sample values for the dependent linear variable; `t`, the vector of angles (in radians) where to evaluate the estimator; `nu`, the value of the smoothing parameter to be used; `method`, the character string giving the estimator to be used. This must be one of `"LL"` for local linear estimator or `"NW"` for Nadaraya–Watson estimator.

The value of `nu` can be set manually or can be obtained by calling the function `nu.LSCV.reg` which provides the least squares cross-validation smoothing parameter for the Nadaraya–Watson and Local Linear estimators from equation (2.16). The arguments `x`, `y` and `method` of this function have the same meaning as those for function `kern.reg.circ`.

Functions `kern.reg.circ` and `nu.LSCV.reg` are illustrated with `wind.speed` dataset corresponding to the measurements of wind speed and wind direction in the Atlantic coast of Galicia.

The Nadaraya–Watson and Local Linear estimators for a regression model of wind speed over wind direction are shown in Figure 4.4 (left panel), in solid and dashed lines respectively. Estimators are obtained with the code:

```
R> data("speed.wind2")
R> dir <- rad(speed.wind2$Direction)
R> speed <- speed.wind2$Speed
R> t <- seq(0, 2*pi, length=200)
R> nuNW <- nu.LSCV.reg(x=dir, y=speed, method="NW")
R> nuLL <- nu.LSCV.reg(x=dir, y=speed, method="LL")
R> estLL <- kern.reg.circ(x=dir, y=speed, t=t, nu=nuLL, method="NW")
R> estNW <- kern.reg.circ(x=dir, y=speed, t=t, nu=nuNW, method="NW")
```

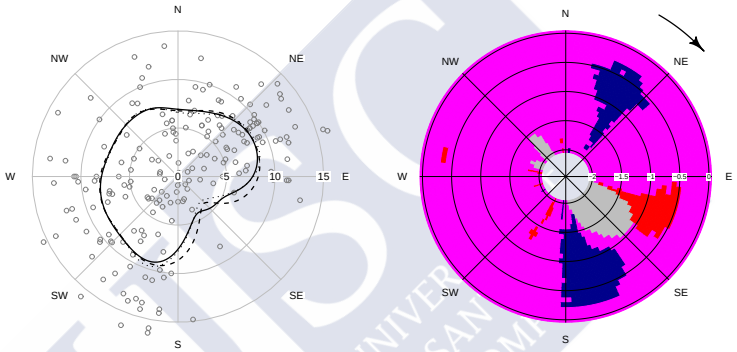

Figure 4.4: Left: Nadaraya–Watson estimator (dashed curve) and Local Linear estimator (solid curve). Right: CircSiZer map for circular–linear regression for wind speed (m/s) with respect to wind direction.

The function `circsizer.regression` provides the CircSiZer map for regression considering a circular covariate and a linear response and based on Local Linear estimator. The first arguments for this function are `x`, the sample of angles in radians (between 0 and 2π) for the circular covariate and `y`, the sample of angles for the dependent linear variable `y`. The remaining arguments are the same as for function `circsizer.density`.

Figure 4.4 (right panel) shows the CircSiZer map for exploring the relation between wind speed as a response and wind direction as a covariate, obtained with the code:

```
R> circsizer.regression(dir, vel, NU=seq(10,60,by=5), type=1)
```

In the CircSiZer map, it can be seen that wind speed increases when wind direction comes from NE and S–SW and winds from SE are not frequent at all, being this fact reflected by the gray coloured area.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Conclusions

This dissertation has been focused on nonparametric methods for the analysis of circular data, both for density and regression. Within this framework, the specific goals that have been achieved are the following:

- The study of nonparametric techniques for the estimation of density and regression curves for circular data. Concretely, the smoothing parameter selection problem in circular density estimation and the comparison of three linear smoothers for regression estimation, for circular covariate and linear response. This goal corresponds to the contents in Chapter 2.
- The construction of an exploratory tool, namely CircSiZer, for the assessment of significant features in curve estimates, for density and regression. The development, performance and practical usage of CircSiZer is collected in Chapter 3.
- The implementation of the proposed methods in the R computing environment, gathered in `NPCirc` package. A detailed description of the library contents and utilities is included in Chapter 4.

As mentioned before, nonparametric curve estimation for circular data has been studied in Chapter 2, focusing on the estimation of the density function of a sample of circular data and on the estimation of the regression function when the covariate is circular and the response is scalar.

In the density setting, the circular kernel density estimator and several methods for selecting the smoothing parameter have been reviewed. The main contribution in this context has been the proposal of a new method for choosing the smoothing parameter, namely the plug–in rule (Oliveira et al., 2012b). Through a simulation study, the new selector has been compared with other methods already proposed in the literature, such as the cross–validation rules proposed by Hall et al. (1987), the rule of thumb proposed by Taylor (2008) and the bootstrap procedure proposed by Di Marzio et al. (2011). Simulation results showed that the new plug–in rule equals or even outperforms the other existing selectors. In addition, the different selectors have been also applied for analyzing some real datasets.

Nonparametric regression estimators for a circular explanatory variable and a linear response have also been studied. Two types of nonparametric smoothers have been considered: kernel and

spline smoothers. Specifically, adaptations of the classical Nadaraya–Watson and Local Linear estimators to the circular nature of the covariate and the periodic smoothing spline estimator have been reviewed. The three smoothers considered have been compared in a simulation study and applied to a real dataset. Results from the simulation study have shown that Nadaraya–Watson and Local Linear estimators performs similarly and periodic smoothing spline estimator provide better results in terms of average integrated squared error.

In Chapter 3, a new exploratory method, namely CircSiZer, which avoids the problem of selecting the smoothing parameter and, allows to know which observed features in the smoothed curve (density or regression) are statistically significant and which features are simply artifacts of the sampling noise has been presented. CircSiZer assesses the significance of features by constructing confidence intervals for the derivative of the smoothed curve, $f'(\theta;\tau)$, being f the density or regression curve and τ the smoothing parameter. Different proposals for obtaining the quantiles and the estimation of the standard deviation of $\hat{f}'(\theta;\tau)$ have been studied. Quantiles were computed in order to provide pointwise and simultaneous confidence intervals for $f'(\theta;\tau)$, and in both cases, two approaches were considered: a first approach based on a normal approximation and another approach using bootstrap techniques.

It has been also detailed how the information provided by the confidence intervals is displayed in a circular color map, the CircSiZer map, in such a way that, different colors allow to indentify increasing and decreasing patterns in the smoothed curve for different values of the smoothing parameter.

For the density setting, the coverage of simultaneous confidence intervals based on the normal approximation and the ones based on bootstrap has been compared through a simulation study. The study reflects that the coverage of bootstrap simultaneous confidence intervals is close to the nominal level $(1-\alpha)$ whereas the coverage of normal simultaneous confidence intervals is below this value. Also from a simulation study, it was observed that the use of bootstrap simultaneous confidence intervals may present difficulties for attaining the goal of detecting the modes of the underlying model. In this sense, the CircSiZer map with pointwise confidence intervals may be helpful for identifying the modes presented by the model. However, the interpretation of pointwise CircSiZer map must be done carefully because it may flags spurious modes as significant.

In the regression setting, the coverage of simultaneous confidence intevals based on the normal approximation has been studied for the case of fixed design. Results showed that the coverage of these intervals is below the nominal value and so, unfortunately, in the context of fixed desing, none of the proposals for constructing the confidence intervals is appropriate, so further research is required in order to solve this issue.

For random design, the performance of CircSiZer with simultaneous bootstrap confidence intervals has been checked in some simulated and real data examples. The main drawback of this proposal is the high computational cost for obtaining the CircSiZer map. For linear data, Hannig

and Marron (2006) improve the coverage of simultaneous confidence intervals based on normal assumptions. An approach similar to the one described in the previous reference could be adapted in the circular setting.

Finally, most of the techniques introduced througthout this manuscript have been implemented using the statistical software R in the `NPCirc` package. Chapter 4 has been devoted to the description and illustration of the functions included in the package. Among them, a function for kernel density estimation for circular data with an up–to–date collection of smoothing parameter selection procedures: the cross–validation rules introduced by Hall et al. (1987), the rule of thumb proposed by Taylor (2008), the bootstrap method described in Di Marzio et al. (2009) and the new plug–in rule proposed in this dissertation. For nonparametric regression estimation, the Nadaraya–Watson and Local Linear estimators for a linear response and a circular covariate jointly with the least squares cross–validations rule for the selection of the smoothing parameter, have been included. As commented in Chapter 4, R routines related with periodic smoothing splines are not still included in the package but they will be included in future updates. Both for density and regression, the CircSiZer method with pointwise bootstrap confidence intervals have been also implemented. In the current version, the `NPCirc` package only allows to plot the CircSiZer map constructed with bootstrap pointwise confidence intervals. The other possibilities described in Section 3.3 will be also implemented in future versions of the library.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix A

Simulated models

The specific formulae of those models considered in the simulation study carried out in Chapter 2 (see Figure 2.2) are given here.

Simple models:

M1: Circular uniform.

M2: von Mises: $vM(\pi, 1)$.

M3: Wrapped normal: $WN(\pi, 0.9)$.

M4: cardioid: $C(\pi, 0.5)$.

M5: Wrapped Cauchy: $WC(\pi, 0.8)$.

M6: Wrapped skew–normal: $WSN(\pi, 1, 20)$.

Two components models:

M7: Mixture of two von Mises $1/2 \cdot vM(0, 4) + 1/2 \cdot vM(\pi, 4)$.

M8: Mixture of two von Mises $1/2 \cdot vM(2, 5) + 1/2 \cdot vM(4, 5)$.

M9: Mixture of two von Mises $1/4 \cdot vM(0, 2) + 3/4 \cdot vM(\pi/\sqrt{3}, 2)$.

M10: Mixture of von Mises and wrapped Cauchy $4/5 \cdot vM(\pi, 5) + 1/5 \cdot WC(4\pi/3, 0.9)$.

Models with more than two components:

M11: Mixture of three von Mises $1/3 \cdot vM(\pi/3, 6) + 1/3 \cdot vM(\pi, 6) + 1/3 \cdot vM(5\pi/3, 6)$.

M12: Mixture of three von Mises $2/5 \cdot vM(\pi/2, 4) + 1/5 \cdot vM(\pi, 5) + 2/5 \cdot vM(3\pi/2, 4)$.

M13: Mixture of three von Mises $2/5 \cdot vM(0.5, 6) + 2/5 \cdot vM(3, 6) + 1/5 \cdot vM(5, 24)$.

M14: Mixture of four von Mises $1/4 \cdot vM(0, 12) + 1/4 \cdot vM(\pi/2, 12) + 1/4 \cdot vM(\pi, 12) + 1/4 \cdot vM(3\pi/2, 12)$.

M15: Mixture of wrapped Cauchy, wrapped normal, von Mises and wrapped skew-normal $3/10 \cdot WC(\pi - 1, 0.6) + 1/4 \cdot WN(\pi + 0.5, 0.9) + 1/4 \cdot vM(\pi + 2, 3) + 1/5 \cdot WSN(6, 1, 3)$.

M16: Mixture of five von Mises $1/5 \cdot vM(\pi/5, 18) + 1/5 \cdot vM(3\pi/5, 18) + 1/5 \cdot vM(\pi, 18) + 1/5 \cdot vM(7\pi/5, 18) + 1/5 \cdot vM(9\pi/5, 18)$.

Other complex models:

M17: Mixture of cardioid and wrapped Cauchy $2/3 \cdot C(\pi, 0.5) + 1/3 \cdot WC(\pi, 0.9)$.

M18: Mixture of four von Mises $1/2 \cdot vM(\pi, 1) + 1/6 \cdot vM(\pi - 0.8, 30) + 1/6 \cdot vM(\pi, 30) + 1/6 \cdot vM(\pi + 0.8, 30)$.

M19: Mixture of five von Mises $4/9 \cdot vM(2, 3) + 5/36 \cdot vM(4, 3) + 5/36 \cdot vM(3.5, 50) + 5/36 \cdot vM(4, 50) + 5/36 \cdot vM(4.5, 50)$.

M20: Mixture of two wrapped skew-normal and two wrapped Cauchy $1/3 \cdot WSN(0, 0.7, 20) + 1/3 \cdot WSN(\pi, 0.7, 20) + 1/6 \cdot WC(3\pi/4, 0.9) + 1/6 \cdot WC(7\pi/4, 0.9)$.

Appendix B

Kernel smoothers

In the regression setting (consider model (2.13)), the Local Linear estimator is a linear smoother, i.e, there exists a vector $\boldsymbol{l}(\theta) = (l_1(\theta), \ldots, l_n(\theta))^t$ such that it can be written in the following way

$$\hat{f}_{CLL}(\theta;\nu) = \sum_{j=1}^{n} l_j(\theta) Y_j,$$

where

$$l_j(\theta) = \frac{b_j(\theta)}{\sum_{k=1}^{n} b_k(\theta)}$$

with

$$\begin{aligned} b_j(\theta) &= K_\nu(\Theta_j - \theta)\left(S_{n,2}(\theta) - \sin(\Theta_j - \theta) S_{n,1}(\theta)\right) \\ S_{n,r}(\theta) &= \sum_{j=1}^{n} K_\nu(\Theta_j - \theta)\left(\sin(\Theta_j - \theta)\right)^r, \ r = 0, 1, 2. \end{aligned}$$

Hence, if $\hat{\boldsymbol{f}}_{CLL} = (\hat{f}_{CLL}(\Theta_1;\nu), \ldots, \hat{f}_{CLL}(\Theta_n;\nu))^t$ denotes the vector of fitted values at the design points and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$, it follows that

$$\hat{\boldsymbol{f}}_{CLL} = L_\nu \boldsymbol{Y}$$

where L_ν is an $(n \times n)$ matrix whose i–th row is $l(\Theta_i)^t$. Thus, the (i,j) element of L_ν is $L_{ij} = l_j(\Theta_i)$, $i, j \in \{1, \ldots, n\}$. The matrix L_ν is known as the hat matrix.

Similarly, if $\hat{\boldsymbol{f}}'_{CLL} = (\hat{f}'_{CLL}(\Theta_1;\nu), \ldots, \hat{f}'_{CLL}(\Theta_n;\nu))^t$ denotes the vector of values of the estimator of the derivative at the design points, it then follows that

$$\hat{\boldsymbol{f}}'_{CLL} = \tilde{L}_\nu \boldsymbol{Y}$$

where $\tilde{L}_\nu$ is an $(n \times n)$ matrix whose (i,j) element is given by $\tilde{L}_{ij} = \tilde{l}_j(\Theta_i)$, $i, j \in \{1, \ldots, n\}$ where

$$\tilde{l}_j(\theta) = \frac{\tilde{b}_j(\theta)}{\sum_{k=1}^{n} \tilde{b}_k(\theta)}$$

with

$$\tilde{b}_j(\theta) = K_\nu(\Theta_j - \theta)\left(S_{n,1}(\theta) - \sin(\Theta_j - \theta)S_{n,0}(\theta)\right).$$

The value of the estimator and its derivative can be obtained for a grid of locations $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^t$ in the interval $[0, 2\pi)$ just by matrix–vector operations:

$$\hat{\boldsymbol{f}}_{CLL,\boldsymbol{\theta}} = L_{\nu,\boldsymbol{\theta}}\boldsymbol{Y}$$

$$\hat{\boldsymbol{f}}'_{CLL,\boldsymbol{\theta}} = \tilde{L}_{\nu,\boldsymbol{\theta}}\boldsymbol{Y}$$

where $L_{\nu,\boldsymbol{\theta}}$ and $\tilde{L}_{\nu,\boldsymbol{\theta}}$ are $(N \times n)$ matrix whose (i,j) element is given by $l_j(\theta_i)$ and $\tilde{l}_j(\theta_i)$, $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, n\}$, respectively.

Appendix C

Periodic cubic splines

Details behind our technical calculations in Section 2.3.2 are given here. This appendix is divided in two parts: the first one is devoted to specify a periodic cubic spline by giving its value and second derivative at each of the knots, and the second one is devoted to give a matrix notation for its integrated squared second derivative.

The value–second derivative representation

Let s be an arbitrary periodic cubic spline on $[t_1, t_{m+1}]$ with splines coefficients a_i, b_i, c_i, d_i, $i = 1, \ldots, m$. Then, for t in $[t_i, t_{i+1}]$,

$$\begin{aligned} s(t) = \quad & s_i(t) = \quad a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3, \\ s'(t) = \quad & s_i'(t) = \quad b_i + 2c_i(t - t_i) + 3d_i(t - t_i)^2, \\ s''(t) = \quad & s_i''(t) = \quad c_i + 6d_i(t - t_i). \end{aligned}$$

From the continuity of s' at the interior knots

$$s_{i-1}'(t_i) = s_i'(t_i), \ i = 2, \ldots, m,$$

and from the periodicity of s' on $[t_1, t_{m+1}]$

$$s'(t_{m+1}) = s'(t_1) \Rightarrow s_m'(t_{m+1}) = s_1'(t_1),$$

it follows that,

$$\begin{aligned} b_m + 2c_m(t_{m+1} - t_m) + 3d_m(t_{m+1} - t_m)^2 &= b_1 \\ b_{i-1} + 2c_{i-1}(t_i - t_{i-1}) + 3d_{i-1}(t_i - t_{i-1})^2 &= b_i, \ i = 2, \ldots, m. \end{aligned}$$

By denoting $h_i = t_{i+1} - t_i$, $i = 1, \ldots, m$

$$\begin{aligned} b_m + 2c_m h_m + 3d_m h_m^2 &= b_1 \\ b_i + 2c_i h_i + 3d_i h_i^2 &= b_{i+1}, \ i = 1, \ldots, m-1, \end{aligned}$$

or equivalently,

$$2c_ih_i + 3d_ih_i^2 = b_{i+1} - b_i, \quad i = 1, \ldots, m \tag{C.1}$$

where $b_{m+1} = b_1$. By Remark 3.22 in Nürnberger (1989), if s is a periodic cubic spline with knots $t_1, \ldots, t_{m+1}$, there exists an extension $\tilde{s} : \mathbb{R} \to \mathbb{R}$ of s such that $\tilde{s}$ has two continuous derivatives and $\tilde{s}$ is a periodic function with period $T = t_{m+1} - t_1$. So, from now, any subscript i will be indicating i mod m where mod denotes the modulo operator.

From the continuity of s at the interior knots

$$s_{i-1}(t_i) = s_i(t_i), \ i = 2, \ldots, m,$$

and the periodicity of s on $[t_1, t_{m+1}]$

$$s(t_{m+1}) = s(t_1) \Rightarrow s_m(t_{m+1}) = s_1(t_1),$$

it follows that,

$$\begin{aligned} a_m + b_m(t_{m+1} - t_m) + c_m(t_{m+1} - t_m)^2 + d_m(t_{m+1} - t_m)^3 &= a_1 \\ a_{i-1} + b_{i-1}(t_i - t_{i-1}) + c_{i-1}(t_i - t_{i-1})^2 + d_{i-1}(t_i - t_{i-1})^3 &= a_i, \ i = 2, \ldots, m \end{aligned}$$

and then,

$$\begin{aligned} a_m + b_m h_m + c_m h_m^2 + d_m h_m^3 &= a_1 \\ a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 &= a_{i+1}, \ i = 1, \ldots, m-1 \end{aligned}$$

or equivalently,

$$b_i = (a_{i+1} - a_i)/h_i - c_i h_i - d_i h_i^2, \ i = 1, \ldots, m. \tag{C.2}$$

From the continuity of s'' at the interior knots

$$s''_{i-1}(t_i) = s''_i(t_i), \quad i = 2, \ldots, m,$$

and the periodicity of s'' on $[t_1, t_{m+1}]$, it follows that

$$s''(t_{m+1}) = s''(t_1) \Rightarrow s''_m(t_{m+1}) = s''_1(t_1),$$

it follows that

$$\begin{aligned} 2c_m + 6d_m(t_{m+1} - t_m) &= 2c_1 \\ 2c_{i-1} + 6d_{i-1}(t_i - t_{i-1}) &= 2c_i, \ i = 2, \ldots, m, \end{aligned}$$

and then,

$$\begin{aligned} 2c_m + 6d_m h_m &= 2c_1, \\ 2c_i + 6d_i h_i &= 2c_{i+1},\ i = 1, \ldots, m-1, \end{aligned}$$

or equivalently,

$$d_i = (c_{i+1} - c_i)/(3h_i), \quad i = 1, \ldots, m. \tag{C.3}$$

From (C.2) and (C.3),

$$b_i = \frac{a_{i+1} - a_i}{h_i} - \frac{1}{3}(c_{i+1} + 2c_i)h_i, \quad i = 1, \ldots, m. \tag{C.4}$$

Replacing (C.3) and (C.4) in (C.1), for $i = 1, \ldots, m$, it holds that:

$$\frac{1}{3}c_i h_i + \frac{2}{3}c_{i+1}(h_i + h_{i+1}) + \frac{1}{3}c_{i+2}h_{i+1} = \frac{a_{i+2}}{h_{i+1}} + a_{i+1}\left(-\frac{1}{h_{i+1}} - \frac{1}{h_i}\right) + \frac{a_i}{h_i}. \tag{C.5}$$

Let be $s_i = s(t_i)$ and $\gamma_i = s''(t_i)$ for $i = 1, \ldots, m$. From the explicit representation of s and s'', $s(t_i) = a_i$ and $s''(t_i) = 2c_i$ for $i = 1, \ldots, m$. Then, for $i = 1, \ldots, m$, equation (C.5) can be written as

$$\gamma_i \frac{h_i}{6} + \gamma_{i+1}\frac{h_i + h_{i+1}}{3} + \gamma_{i+2}\frac{h_{i+1}}{6} = \frac{s_{i+2}}{h_{i+1}} + s_{i+1}\left(-\frac{1}{h_{i+1}} - \frac{1}{h_i}\right) + \frac{s_i}{h_i}. \tag{C.6}$$

Equations (C.6) constitute a system of m linear equations which can be written in matrix notation as follows:

$$R\boldsymbol{\gamma} = Q\boldsymbol{s}, \tag{C.7}$$

where R and Q are symmetric, cyclic–tridiagonal matrices of order m, and $\boldsymbol{s} = (s_1, \ldots, s_m)^t$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^t$. Thus, a periodic cubic spline is completely specified by giving its value and second derivative at each of the knots t_i, $i = 1, \ldots, m$.

The non zero entries of R and Q are expressed in terms of the distances between succesive knots:

$$\begin{array}{ll} r_{1,1} = (h_m + h_1)/3 & q_{1,1} = -1/h_1 - 1/h_m \\ r_{1,m} = r_{m,1} = h_m/6 & q_{1,m} = q_{m,1} = 1/h_m \\ r_{ii} = (h_{i-1} + h_i)/3,\ i = 2, \ldots, m & q_{i,i} = -1/h_{i-1} - 1/h_i,\ i = 2, \ldots, m \\ r_{i-1,i} = r_{i,i-1} = h_{i-1}/6,\ i = 2, \ldots, m & q_{i-1,i} = q_{i,i-1} = 1/h_{i-1},\ i = 2, \ldots, m \end{array}$$

The vectors $\boldsymbol{s}$ and $\boldsymbol{\gamma}$ specify the curve s completely, and it is possible to give explicit formulae in terms of $\boldsymbol{s}$ and $\boldsymbol{\gamma}$ for the value and derivatives of s at any point t. Following Green and Silverman (1994, pp. 22–23), the next expressions allow to compute the value of s and its derivative at any point $t \in [t_1, t_{m+1})$. Let be $h_i(t) = t - t_i$, if $t_i \le t \le t_{i+1}$, $i = 1, \ldots, m$,

$$s(t) = \frac{h_i(t)s_{i+1} - h_{i+1}(t)s_i}{h_i} + \frac{h_i(t)h_{i+1}(t)}{6}\left\{\left(1 + \frac{h_i(t)}{h_i}\right)\gamma_{i+1} + \left(1 - \frac{h_{i+1}(t)}{h_i}\right)\gamma_i\right\}. \tag{C.8}$$

Taking derivatives in (C.8) with respect to t,

$$s'(t) = \frac{s_{i+1} - s_i}{h_i} + \frac{h_i(t)h_{i+1}(t)(\gamma_{i+1} - \gamma_i)}{6h_i} + \frac{(h_i(t) + h_{i+1}(t))}{6}\left\{\left(1 + \frac{h_i(t)}{h_i}\right)\gamma_{i+1} + \left(1 - \frac{h_{i+1}(t)}{h_i}\right)\gamma_i\right\}. \tag{C.9}$$

From (C.8) and (C.9), it follows that $s(t)$ and $s'(t)$ can be written as linear combinations of $\boldsymbol{s}$ and $\boldsymbol{\gamma}$. Given a grid of locations $\boldsymbol{x} = (x_1, \ldots, x_N)^t$ with $x_i \in [t_1, t_{m+1})$ for all i then,

$$\begin{aligned} \boldsymbol{s_x} &= C\boldsymbol{s} - D\boldsymbol{\gamma}, \\ \boldsymbol{s'_x} &= \tilde{C}\boldsymbol{s} - \tilde{D}\boldsymbol{\gamma}, \end{aligned}$$

or equivalently, taking into account (C.7),

$$\begin{aligned} \boldsymbol{s_x} &= \left[C - DR^{-1}Q\right]\boldsymbol{s} = M\boldsymbol{s}, \\ \boldsymbol{s'_x} &= \left[\tilde{C} - \tilde{D}R^{-1}Q\right]\boldsymbol{s} = \tilde{M}\boldsymbol{s}, \end{aligned}$$

where $\boldsymbol{s_x} = (s(x_1), \ldots, s(x_N))^t$, $\boldsymbol{s'_x} = (s'(x_1), \ldots, s'(x_N))^t$ and C, D, $\tilde{C}$ and $\tilde{D}$ are $N \times m$ coefficient matrices defined as is given below:

For each x_i, with $i = 1, \ldots, N$:

- If $t_j \le x_i \le t_{j+1}$ for some $j = 1, \ldots, m-1$, then

$$\begin{aligned} C_{i,j} &= 1 - \delta_{ij} & D_{i,j} &= \delta_{ij}(1 - \delta_{ij})(2 - \delta_{ij})h_j^2/6 \\ C_{i,j+1} &= \delta_{ij} & D_{i,j+1} &= \delta_{ij}(1 - \delta_{ij}^2)h_j^2/6 \end{aligned}$$

$$\begin{aligned} \tilde{C}_{i,j} &= -1/h_j & \tilde{D}_{i,j} &= \left(2 - 6\delta_{ij} + 3\delta_{ij}^2\right)h_j/6 \\ \tilde{C}_{i,j+1} &= 1/h_j & \tilde{D}_{i,j+1} &= \left(1 - 3\delta_{ij}^2\right)h_j/6, \end{aligned}$$

  where $\delta_{ij} = h_j(x_i)/h_j$ being $h_j(x_i) = x_i - t_j$.

- If $t_m \le x_i \le t_{m+1}$ then,

$$\begin{aligned} C_{i,m} &= 1 - \delta_{im} & D_{i,m} &= \delta_{im}(1 - \delta_{im})(2 - \delta_{im})h_m^2/6 \\ C_{i,1} &= \delta_{im} & D_{i,1} &= \delta_{im}(1 - \delta_{im}^2)h_{m-1}^2/6 \end{aligned}$$

$$\begin{aligned} \tilde{C}_{i,m} &= -1/h_m & \tilde{D}_{i,m} &= \left(2 - 6\delta_{im} + 3\delta_{im}^2\right)h_m/6 \\ \tilde{C}_{i,1} &= 1/h_m & \tilde{D}_{i,m} &= \left(1 - 3\delta_{im}^2\right)h_m/6, \end{aligned}$$

  where $\delta_{im} = h_m(x_i)/h_m$ being $h_m(x_i) = x_i - t_m$.

Now, it will be shown that the matrices are well defined. For any $x_i \in [t_j, t_{j+1}]$, $i = 1, \ldots, N$ and $j = 1, \ldots, m$,

$$\begin{aligned}
C_{i,j+1} &= \delta_{ij} = \frac{h_j(x_i)}{h_j}, \\
C_{i,j} &= 1 - \delta_{ij} = \left(1 - \frac{h_j(x_i)}{h_j}\right) = \frac{h_j - h_j(x_i)}{h_j} = -\frac{h_{j+1}(x_i)}{h_j}, \\
D_{i,j} &= \frac{\delta_{ij}(1-\delta_{ij})(2-\delta_{ij})h_j^2}{6} = \frac{1}{6}\frac{h_j(x_i)}{h_j}\left(-\frac{h_{j+1}(x_i)}{h_j}\right)\left(1 - \frac{h_{j+1}(x_i)}{h_j}\right)h_j^2 \\
&= -\frac{h_j(x_i)h_{j+1}(x_i)}{6}\left(1 - \frac{h_{j+1}(x_i)}{h_j}\right), \\
D_{i,j+1} &= \frac{\delta_{ij}(1-\delta_{ij}^2)h_j^2}{6} = \frac{1}{6}\frac{h_j(x_i)}{h_j}\left(1 - \left(\frac{h_j(x_i)}{h_j}\right)^2\right)h_j^2 \\
&= \frac{1}{6}\frac{h_j(x_i)}{h_j}\left(1 - \frac{h_j(x_i)}{h_j}\right)\left(1 + \frac{h_j(x_i)}{h_j}\right)h_j^2 = \frac{1}{6}\frac{h_j(x_i)}{h_j}\frac{h_{j+1}(x_i)}{h_j}\left(1 + \frac{h_j(x_i)}{h_j}\right)h_j^2 \\
&= \frac{h_j(x_i)h_{j+1}(x_i)}{6}\left(1 + \frac{h_j(x_i)}{h_j}\right),
\end{aligned}$$

where if $x_i \in [t_m, t_{m+1}]$, $C_{i,m+1}$ and $D_{i,m+1}$ denote $C_{i,1}$ and $D_{i,1}$, respectively. Then

$$\begin{aligned}
&C_{i,j}s_j + C_{i,j+1}s_{j+1} - D_{i,j}\gamma_j - D_{i,j+1}\gamma_{j+1} \\
&= -\frac{h_{j+1}(x_i)}{h_j}s_j + \frac{h_j(x_i)}{h_j}s_{j+1} + \frac{h_j(x_i)h_{j+1}(x_i)}{6}\left(1 - \frac{h_{j+1}(x_i)}{h_j}\right)\gamma_j \\
&-\frac{h_j(x_i)h_{j+1}(x_i)}{6}\left(1 + \frac{h_j(x_i)}{h_j}\right)\gamma_{j+1} \\
&= \frac{h_j(x_i)s_{j+1} - h_{j+1}(x_i)s_j}{h_j} + \frac{h_j(x_i)h_{j+1}(x_i)}{6}\left[\left(1 - \frac{h_{j+1}(x_i)}{h_j}\right)\gamma_j + \left(1 + \frac{h_j(x_i)}{h_j}\right)\gamma_{j+1}\right],
\end{aligned}$$

that is equal to (C.8) for $t = x_i$.

The entries of $\tilde{C}$ and $\tilde{D}$ are obtained by deriving with respect to x_i the entries of the matrices C and D. Computing $\tilde{C}_{i,j}s_j + \tilde{C}_{i,j+1}s_{j+1} - \tilde{D}_{i,j}\gamma_j - \tilde{D}_{i,j+1}\gamma_{j+1}$ results expression (C.9).

Expression for the roughness penalty

A natural way of measuring the roughness of a twice–differentiable curve is to compute its

integrated squared second derivative. Thus, from the explicit representation of s'',

$$\begin{aligned}
\int_{t_1}^{t_{m+1}} (s''(t))^2 dt &= \sum_{i=1}^{m} \int_{t_i}^{t_{i+1}} (s''(t))^2 dt \\
&= \sum_{i=1}^{m} \int_{t_i}^{t_{i+1}} (2c_i + 6d_i(t-t_i))^2 dt \\
&= \sum_{i=1}^{m} \int_{t_i}^{t_{i+1}} (4c_i^2 + 24c_i d_i (t-t_i) + 36 d_i^2 (t-t_i)^2) dt \\
&= \sum_{i=1}^{m} 4 \left[c_i^2 (t_{i+1} - t_i) + 6 c_i d_i \int_{t_i}^{t_{i+1}} (t-t_i) dt + 9 d_i^2 \int_{t_i}^{t_{i+1}} (t-t_i)^2 dt \right] \\
&= \sum_{i=1}^{m} 4 \left[c_i^2 (t_{i+1} - t_i) + 6 c_i d_i \frac{(t_{i+1}-t_i)^2}{2} + 9 d_i^2 \frac{(t_{i+1}-t_i)^3}{3} \right].
\end{aligned}$$

Using (C.3) and $h_i = t_{i+1} - t_i$ for $i = 1 \ldots, m$,

$$\int_{t_1}^{t_m} (s''(t))^2 dt = \sum_{i=1}^{m-1} 4/3 h_i (c_i^2 + c_i c_{i+1} + c_{i+1}^2) + 4/3 h_m (c_m^2 + c_m c_1 + c_1^2).$$

Since $c_i = \gamma_i / 2$,

$$\int_{t_1}^{t_m} (s''(t))^2 dt = \sum_{i=1}^{m-1} h_i/3 (\gamma_i^2 + \gamma_i \gamma_{i+1} + \gamma_{i+1}^2) + h_{m-1}/3 (\gamma_{m-1}^2 + \gamma_{m-1}\gamma_1 + \gamma_1^2) = \boldsymbol{\gamma}^t R \boldsymbol{\gamma},$$

and using (C.7),

$$\int_0^T (s''(t))^2 dt = \boldsymbol{\gamma}^t R \boldsymbol{\gamma} = \boldsymbol{s} Q R^{-1} Q \boldsymbol{s} = \boldsymbol{s} K \boldsymbol{s},$$

where $K = Q R^{-1} Q$.

Appendix D

The NPCirc package

Title Nonparametric Circular Methods

Version 1.0.0

Date 2012-12-24

Author María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez-Casal

Maintainer María Oliveira <maria.oliveira@usc.es>

Depends R($\geq$ 2.11.0), circular, movMF, plotrix, shape, sn

Description This package implements nonparametric smoothing methods for circular data

License GPL–2

R topics documented:

`circsizer.density` *CircSiZer map for density*

Description

This function plots the CircSiZer map for circular density estimation based on circular kernel methods, as described in Oliveira et al. (2013). The CircSiZer is an extension of SiZer proposed by Chaudhuri and Marron (1999) to circular data.

Usage

```
circsizer.density(x, NU, ngrid=250, alpha=0.05, B=500, type=3,
raw.data=FALSE, log.scale=TRUE, zero=pi/2, clockwise=TRUE, title=NULL,
labels=NULL, label.pos=NULL, rad.pos=NULL)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the estimate is to be computed.
`NU`	Vector of smoothing parameters. Values of `NU` must be positive. `NU` will be coerced to be equally spaced. Length of vector `NU` must be at least 2.
`ngrid`	Integer indicating the number of equally spaced angles between 0 and 2π where the estimator is evaluated. Default `ngrid=250`.
`alpha`	Significance level for the CircSiZer map. Default `alpha=0.05`.
`B`	Integer indicating the number of bootstrap samples to estimate the standard deviation of the derivative estimator. Default `B=500`.
`type`	Number indicating the labels to display in the plot: 1 (directions), 2 (hours), 3 (angles in radians), 4 (angles in degrees) or 5 (months). Default `type=3`.
`raw.data`	Logical, if `TRUE`, points indicated by `x` are stacked on the perimeter of the circle. Default is `FALSE`.
`log.scale`	Logical, if `TRUE`, the CircSiZer map is plotted in the scale $-\log_{10}(NU)$. Default is `TRUE`. See Details.
`zero`	Where to place the starting (zero) point. Defaults to the North position.

`clockwise`	Whether to interpret positive positions as clockwise from the starting point. The default is clockwise (`clockwise=TRUE`).
`title`	Title for the plot.
`labels`	Character or expression vector of labels to be placed at the `label.pos`. `label.pos` must also be supplied.
`label.pos`	Vector indicating the position (between 0 and 2π) at which the labels are to be drawn.
`rad.pos`	Vector (between 0 and 2π) with the drawing position for the radius.

Details

With CircSiZer, significance features (peaks and valleys) in the data are seeked via the construction of confidence intervals for the scale-space version of the smoothed derivative curve, as it is described in Oliveira et al. (2013). Thus, for a given point and a given value of the smoothing parameter, the curve is significantly increasing (decreasing) if the confidence interval is above (below) 0 and if the confidence interval contains 0, the curve for that value of the smoothing parameter and at that point does not have a statistically significant slope. This information is displayed in a circular color map, the CircSiZer map, in such a way that, at a given point, the performance of the estimated curve is represented by a color ring with radius proportional to the value of the smoothing parameter.

Differents colors allow to indentify peaks and valleys. Blue color indicates locations where the curve is significantly increasing; red color shows where it is significantly decreasing and purple indicates where it is not significantly different from zero. Gray color correspons to those regions where there is not enough data to make statements about significance. Thus, at a given bandwidth, a significant peak can be identified when a region of significant positive gradient is followed by a region of significant negative gradient (i.e. blue-red pattern), and a significant trough by the reverse (red-blue pattern), taking clockwise as the positive sense of rotation.

If `log.scale=TRUE` then, the values of the considered smoothing parameters `NU` are transformed to $-\log_{10}$ scale, i.e, a sequence of equally spaced smoothing parameters according to the parameters `-log10(max(NU))`, `-log10(min(NU))` and `length(NU)` is used. Hence, small values of this parameter corresponds with larger rings and large values corresponds with smaller rings. Whereas if `log.scale=FALSE`, small values of this parameter corresponds with smaller rings and large values corresponds with larger rings.

The NAs will be automatically removed.

Value

CircSiZer map for density.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves, *Journal of the American Statistical Association*, **94**, 807–823.

Oliveira, M., Crujeiras, R.M. and Rodríguez–Casal (2013) CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-013-0249-0.

Examples

```
# set.seed(2012)
# x <- rcircmix(100,model=7)
# circsizer.density(x,NU=seq(0,50,length=12),type=4,zero=0,clockwise=FALSE)
```

---

circsizer.regression

CircSiZer map for regression

Description

This function plots the CircSiZer map for circular regression estimation based on circular kernel methods, as described in Oliveira et al. (2013). The CircSiZer is an extension of SiZer proposed by Chaudhuri and Marron (1999) to circular data.

Usage

```
circsizer.regression(x, y, NU, ngrid=150, alpha=0.05, B=500, B2=250,
type=3, log.scale=TRUE, zero=pi/2, clockwise=TRUE, title=NULL,
labels=NULL, label.pos=NULL, rad.pos=NULL)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) for the circular covariate.
`y`	Sample of angles for the dependent linear variable. This must be same length as `x`.
`NU`	Vector of smoothing parameters. Values of `NU` must be positive. `NU` will be coerced to be equally spaced. Length of vector `NU` must be at least 2.
`ngrid`	Integer indicating the number of equally spaced angles between 0 and 2π where the estimator is evaluated. Default `ngrid=150`.

`alpha`	Significance level for the CircSiZer map. Default `alpha=0.05`.
`B`	Integer indicating the number of bootstrap samples to estimate the standard deviation of the derivative estimator. Default `B=500`.
`B2`	Integer indicating the number of bootstrap samples to compute the denominator in Step 2 of algorithm described in Oliveira et al. (2013). Default `B=250`.
`type`	Number indicating the labels to display in the plot: 1 (directions), 2 (hours), 3 (angles in radians), 4 (angles in degrees) or 5 (months). Default `type=3`.
`log.scale`	Logical, if `TRUE`, the CircSiZer map is plotted in the scale $-\log_{10}(NU)$. Default is `TRUE`.
`zero`	Where to place the starting (zero) point. Defaults to the North position.
`clockwise`	Whether to interpret positive positions as clockwise from the starting point. The default is clockwise (`clockwise=TRUE`).
`title`	Title for the plot.
`labels`	Character or expression vector of labels to be placed at the `label.pos`. `label.pos` must also be supplied.
`label.pos`	Vector indicating the position (between 0 and 2π) at which the labels are to be drawn.
`rad.pos`	Vector (between 0 and 2π) with the drawing position for the radius.

Details

See Details Section of `circsizer.density`. The NAs will be automatically removed.

Value

CircSiZer map for regression.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves, *Journal of the American Statistical Association*, **94**, 807–823.

Oliveira, M., Crujeiras, R.M. and Rodríguez–Casal (2013) CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-013-0249-0.

Examples

```
# Not run: the code works but it is slow
# set.seed(2012)
# n <- 100
# x <- seq(0,2*pi,length=n)
# y <- sin(x)+sqrt(0.5)*rnorm(n)
# circsizer.regression(x,y,NU=seq(10,60,by=5),title="CircSiZer for regression")
```

---

cross.beds1 *Cross-beds azimuths (I)*

Description

This dataset corresponds to azimuths of cross-beds in the Kamthi river (India). Originally analized by SenGupta and Rao (1966) and included in Table 1.5 in Mardia (1972), the dataset collects 580 azimuths of layers lying oblique to principal accumulation surface along the river, being these layers known as cross-beds.

Usage

```
data(cross.beds1)
```

Format

A single-column data frame with 580 observations in radians.

Details

Data were originally recorded in degrees.

Source

Mardia, K.V. (1972) *Statistics of Directional Data.* Academic Press, New York.

SenGupta, S. and Rao, J.S. (1966) Statistical analysis of cross–bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Godavari Valley. *Sankhya: The Indian Journal of Statistics, Series B*, **28**, 165–174.

Examples

```
data(cross.beds1)
```

`cross.beds2` *Cross-beds (II)*

Description

A dataset of cross-beds measurements from Himalayan molasse in Pakistan presented in Fisher (1993). This dataset collects 104 measurements of Chaudan Zam large bedforms.

Usage

```
data(cross.beds2)
```

Format

A single-column data frame with 104 observations in radians.

Details

Data were originally recorded in degrees.

Source

Fisher, N.I. (1993) *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, U.K.

Examples

```
data(cross.beds2)
```

`cycle.changes` *Cycle changes*

Description

The data consists on the changes in cycles of temperatures at ground level in periglacial Monte Alvear (Argentina). The dataset includes 350 observations which correspond to the hours where the temperature changes from positive to negative and viceversa from February 2008 to December 2009.

Usage

```
data(cycle.changes)
```

Format

A data frame with 350 observations on two variables: `change`, which indicates if the temperature changed from positive to negative (-1) or viceversa (1) and `hour`, which indicates the hour (in radians) when the cycle change occured.

Details

Analysis of cycle changes in temperatures for another locations can be seen in Oliveira et al. (2013).

Source

The authors want to acknowledge Prof. Augusto Pérez–Alberti for providing the data, collected within the Project POL2006–09071 from the Spanish Ministry of Education and Science.

References

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

Examples

```
data(cycle.changes)

thaw <- (cycle.changes[,1]==1)
frosting <- (cycle.changes[,1]==-1)

plot(circular(cycle.changes[frosting,2]), shrink=1.08, col=4, stack=TRUE,
zero=pi/2, rotation="clock", axes=FALSE, main="Frosting")
axis.circular(at=circular(seq(0, 7/4*pi, pi/4)),
labels=c("0h","3h","6h","9h","12h","15h","18h","21h"),
zero=pi/2, rotation="clock")

plot(circular(cycle.changes[thaw,2]), shrink=1.08, col=2, stack=TRUE,
zero=pi/2, rotation="clock", axes=FALSE, main="Thaw")
axis.circular(at=circular(seq(0, 7/4*pi, pi/4)),
labels=c("0h","3h","6h","9h","12h","15h","18h","21h"),
zero=pi/2, rotation="clock")
```

`dcircmix` *Mixtures of circular distributions*

Description

Density and random generation functions for a circular distribution or a mixture of circular distributions allowing the following components: circular uniform, von Mises, cardioid, wrapped Cauchy, wrapped normal, wrapped skew-normal.

Usage

```
dcircmix(x, model=NULL, dist=NULL, param=NULL)
rcircmix(n, model=NULL, dist=NULL, param=NULL)
```

Arguments

`x`	Vector of angles in radians to compute the density.
`n`	Number of observations to generate.
`model`	Number between 1 and 20, corresponding with a model defined in Oliveira et al. (2012). See Details.
`dist`	Vector of strings with the distributions that participate in the mixture: `"unif"`, `"vm"`, `"car"`, `"wc"`, `"wn"`, `"wsn"`.
`param`	List with three or four objects. The first object will be a vector containing the proportion of each distribution in the mixture, the second object will be a vector containg the location parameters and the third object will be a vector containing the concentration parameters. If the wrapped skew-normal distribution participates in the mixture, a fourth object will be introduced in the list, a vector containing the skewness parameter. In this case, the values of the skewness parameter for the rest of distributions in the mixture will be zero. The length of each object in the list must be equal to the length of argument `dist`. See Details and Examples.

Details

Models from Oliveira et al. (2012) are described below:

M1: Circular uniform.

M2: von Mises: $vM(\pi, 1)$.

M3: Wrapped normal: $WN(\pi, 0.9)$.

M4: cardioid: $C(\pi, 0.5)$.

M5: Wrapped Cauchy: $WC(\pi, 0.8)$.

M6: Wrapped skew–normal: $WSN(\pi, 1, 20)$.

M7: Mixture of two von Mises $1/2vM(0, 4) + 1/2vM(\pi, 4)$.

M8: Mixture of two von Mises $1/2vM(2,5) + 1/2vM(4,5)$.

M9: Mixture of two von Mises $1/4vM(0,2) + 3/4vM(\pi/\sqrt{3},2)$.

M10: Mixture of von Mises and wrapped Cauchy $4/5vM(\pi,5) + 1/5WC(4\pi/3,0.9)$.

M11: Mixture of three von Mises $1/3vM(\pi/3,6) + 1/3vM(\pi,6) + 1/3vM(5\pi/3,6)$.

M12: Mixture of three von Mises $2/5vM(\pi/2,4) + 1/5vM(\pi,5) + 2/5vM(3\pi/2,4)$.

M13: Mixture of three von Mises $2/5vM(0.5,6) + 2/5vM(3,6) + 1/5vM(5,24)$.

M14: Mixture of four von Mises $1/4vM(0,12)+1/4vM(\pi/2,12)+1/4vM(\pi,12)+1/4vM(3\pi/2,12)$.

M15: Mixture of wrapped Cauchy, wrapped normal, von Mises and wrapped skew-normal $3/10WC(\pi-1,0.6) + 1/4WN(\pi+0.5,0.9) + 1/4vM(\pi+2,3) + 1/5WSN(6,1,3)$.

M16: Mixture of five von Mises $1/5vM(\pi/5,18)+1/5vM(3\pi/5,18)+1/5vM(\pi,18)+1/5vM(7\pi/5,18)+1/5vM(9\pi/5,18)$.

M17: Mixture of cardioid and wrapped Cauchy $2/3C(\pi,0.5) + 1/3WC(\pi,0.9)$.

M18: Mixture of four von Mises $1/2vM(\pi,1)+1/6vM(\pi-0.8,30)+1/6vM(\pi,30)+1/vM(\pi+0.8,30)$.

M19: Mixture of five von Mises $4/9vM(2,3)+5/36vM(4,3)+5/36vM(3.5,50)+5/36vM(4,50)+5/36vM(4.5,50)$.

M20: Mixture of two wrapped skew-normal and two wrapped Cauchy $1/3WSN(0,0.7,20) + 1/3WSN(\pi,0.7,20) + 1/6WC(3\pi/4,0.9) + 1/6WC(7\pi/4,0.9)$.

When the wrapped skew-normal distribution participates in the mixture, the argument `param` for function `dcircmix` can be a list with fifth objects. The fifth object would be the number of terms to be used in approximating the density function of the wrapped skew normal distribution. By default the number of terms used is 20.

Value

`dcircmix` gives the density and `rcircmix` generates random deviates.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

Examples

```
set.seed(2012)
# Linear representation of models M1-M20, each one in a separate window
for (i in 1:20){
dev.new()
f <- function(x) dcircmix(x, model=i)
curve(f, from=0, to=2*pi, main=i)
}

# Circular representation of model vM(0,1)
vM <- function(x) dcircmix(x, model=NULL, dist="vm", param=list(p=1, mu=0, con=1))
curve.circular(vM, n=1000, xlim=c(-1.65,1.65), main="vM(0,1)")
# Random generation from a vM(0,1)
datavM <- rcircmix(50, model=NULL, dist="vm", param=list(p=1, mu=0, con=1))
points(circular(datavM))

# Circular representation of model M18
f18 <- function(x) dcircmix(x, model=18)
curve.circular(f18, n=1000, xlim=c(-1.65,1.65), main="Model 18")
# Random generation from model M8
data18 <- rcircmix(50, model=18)
points(circular(data18))

# Density function and random generation from a mixture of a von Mises and
# a wrapped skew-normal
f <- function(x) dcircmix(x, model=NULL, dist=c("vm","wsn"),
param=list(p=c(0.5,0.5), mu=c(0,pi), con=c(1,1), sk=c(0,10), l=10))
curve.circular(f, n=500, xlim=c(-1.65,1.65))
data <- rcircmix(100, model=NULL, dist=c("vm","wsn"),
param=list(p=c(0.5,0.5), mu=c(0,pi), con=c(1,1), sk=c(0,10)))
points(circular(data))

# Density function and random generation from a mixture of two von Mises and
# two wrapped Cauchy
f <- function(x) dcircmix(x, model=NULL, dist=c("vm","vm","wc","wc"),
param=list(p=c(0.3,0.3,0.2,0.2), mu=c(0,pi,pi/2,3*pi/2), con=c(5,5,0.9,0.9)))
curve.circular(f, n=1000, xlim=c(-1.65,1.65))
data <- rcircmix(100, model=NULL, dist=c("vm","vm","wc","wc"),
param=list(p=c(0.3,0.3,0.2,0.2), mu=c(0,pi,pi/2,3*pi/2), con=c(5,5,0.9,0.9)))
points(circular(data))
```

---

dragonfly *Orientations of dragonflies*

Description

The data, presented in Batschelet (1981), consists on the orientation of 214 dragonflies with respect to the sun's azimuth.

Usage

```
data(dragonfly)
```

Format

A single-column data frame with 214 observations in radians.

Details

Data were originally recorded in degrees.

Source

Batschelet, E. (1981) *Circular Statistics in Biology.* Academic Press, New York.

Examples

```
data(dragonfly)
plot(circular(dragonfly), shrink=1.3)
t <- seq(0,2*pi,length=500)
dens <- kern.den.circ(dragonfly$orientation, t, nu=10)
lines(circular(t), dens)
```

`kern.den.circ` *Nonparametric circular kernel density estimation*

Description

This function computes circular kernel estimates with the given bandwidth, taking the von Mises distribution as circular kernel.

Usage

```
kern.den.circ(x, t=NULL, nu, from=0, to=2*pi, len=250)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the estimate is to be computed.
`t`	Vector of angles in radians where to evaluate the estimator. If `NULL` equally spaced points are used according to the parameters `from`, `to` and `len`.
`nu`	Smoothing parameter to be used. The value of `nu` can be chosen by using the functions `nu.rt`, `nu.CV`, `nu.pi` and `nu.boot`.
`from, to`	Left and right-most points of the grid at which the density is to be estimated.
`len`	Number of equally spaced points at which the density is to be estimated.

Details

The NAs will be automatically removed.

Value

Numeric vector of the same length of `t` with the kernel density estimated values at `t`.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

Taylor, C.C. (2008) Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493–3500.

See Also

`nu.rt`, `nu.CV`, `nu.pi`, `nu.boot`

Examples

```
##  Estimating the density function of a sample of circular data
set.seed(2012)
n <- 100
x <- rcircmix(n, model=14)
t <- seq(0,2*pi,length=500)
est <- kern.den.circ(x, t, nu=50)
plot(t, dcircmix(t,model=14), ylim=c(0,0.4), type="l", lwd=2,
main="Linear representation")
```

```
lines(t, est, col=2)
plot(circular(x), shrink=1.3, main="Circular representation")
lines(circular(t), dcircmix(t,model=14), lwd=2)
lines(circular(t), est, col=2)
```

`kern.reg.circ` *Nonparametric circular kernel regression estimation*

Description

This function implements the Nadaraya-Watson estimator and the Local-Linear estimator for circular-linear data (circular covariate and linear response), as described in Di Marzio et al. (2009) and Oliveira et al. (2013), taking the von Mises distribution as kernel.

Usage

```
kern.reg.circ(x, y, t=NULL, nu, method="LL", tol=300, from=0, to=2*pi,
len=250)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) for the circular covariate.
`y`	Sample of angles for the dependent linear variable. This must be same length as `x`.
`t`	Vector of angles (in radians) where to evaluate the estimator. If `NULL` equally spaced points are used according to the parameters `from`, `to` and `len`.
`nu`	Smoothing parameter to be used. The value of `nu` can be chosen by using the function `nu.LSCV.reg`
`method`	Character string giving the estimator to be used. This must be one of `"LL"` for Local-Linear estimator or `"NW"` for Nadaraya-Watson estimator. Default `method="LL"`.
`tol`	Tolerance parameter to avoid overflow when `nu` is larger than `tol`. Default is `tol=300`.
`from, to`	Left and right-most points of the grid at which the density is to be estimated.
`len`	Number of equally spaced points at which the density is to be estimated.

Details

See Section 3 in Oliveira et al. (2013). See Di Marzio et al. (2009). The NAs will be automatically removed.

Value

Numeric vector of the same length of `t` with the values of the estimate at the evaluation points.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

Di Marzio, M., Panzera A. and Taylor, C. C. (2009) Local polynomial regression for circular predictors. *Statistics and Probability Letters*, **79**, 2066–2075.

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

See Also

`nu.LSCV.reg`

Examples

```
data(speed.wind2)
dir <- rad(speed.wind2$Direction)
vel <- speed.wind2$Speed
nas <- which(is.na(vel))
dir <- dir[-nas]
vel <- vel[-nas]
t <- seq(0,2*pi,length=200)
estLL <- kern.reg.circ(dir, vel, t=t, nu=30)
estNW <- kern.reg.circ(dir, vel, t=t, nu=30, method="NW")

plot(dir, vel, xlab="direction", ylab="speed (m/s)", axes=FALSE)
lines(t, estLL, col=2)
lines(t, estNW, col=3)
axis(1, at=circular(seq(0,2*pi,by=pi/4)),
labels=c("N","NE","E","SE","S","SW","W","NW","N"))
legend("topleft", c("LL","NW"), lty=1, col=2:3)
axis(2)
```

`nu.boot` *Bootstrap method*

Description

This function implements the bootstrap procedure proposed by Di Marzio et al. (2011) for selecting the smoothing parameter for density estimation taking the von Mises density as kernel.

Usage

```
nu.boot(x, lower=0, upper=100, np=500, tol=0.1)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the smoothing parameter is to be computed.
`lower, upper`	`lower` and `upper` boundary of the interval to be used in the search for the value of the smoothing parameter. Default `lower=0` and `upper=100`.
`np`	Number of points where to evaluate the estimator for numerical integration. Default `np=500`.
`tol`	Convergence tolerance for `optimize`.

Details

This method is based on the proposal of Taylor (1989) for linear data. See also Oliveira et al. (2012). The NAs will be automatically removed.

Value

Value of the smoothing parameter.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

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

Taylor, C.C. (1989) Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, **76**, 705–712.

See Also

`kern.den.circ`, `nu.rt`, `nu.CV`, `nu.pi`

Examples

```
set.seed(2012)
n <- 100
x <- rcircmix(n, model=17)
nu.boot(x, lower=0, upper=20)
```

`nu.CV` *Cross-validation for density estimation*

Description

This function gives the least squares cross-validation smoothing parameter or the likelihood cross-validation smoothing parameter for density estimation.

Usage

```
nu.CV(x, method="LCV", lower=0, upper=100, tol=0.1, np=500)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the smoothing parameter is to be computed.
`method`	Character string giving the cross-validation rule to be used. This must be one of `"LCV"` or `"LSCV"`. Default `method="LCV"`.
`lower, upper`	`lower` and `upper` boundary of the interval to be used in the search for the value of the smoothing parameter. Default `lower=0` and `upper=100`.
`tol`	Convergence tolerance for `optimize`. Default `tol=0.1`.
`np`	Number of points where to evaluate the estimator for numerical integration when `method="LSCV"`. Default `np=500`.

Details

The LCV smoothing parameter is obtained as the value of ν that maximizes the logarithm of the likelihood cross-validation function (8) in Oliveira et al. (2013). The LSCV smoothing parameter is obtained as the value of ν that minimizes expression (7) in Oliveira et al. (2013). See also Hall et al. (1987) and Oliveira et al. (2012). The NAs will be automatically removed.

Value

Value of the smoothing parameter.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

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

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

See Also

`kern.den.circ`, `nu.rt`, `nu.CV`, `nu.boot`

Examples

```
set.seed(2012)
n <- 100
x <- rcircmix(n, model=11)
nu.CV(x, method="LCV", lower=0, upper=20)
nu.CV(x, method="LSCV", lower=0, upper=20)
```

NPCirc-package *Nonparametric circular methods.*

Description

This package implements nonparametric kernel methods for density and regression estimation for circular data.

Details

Package:	NPCirc
Type:	Package
Version:	1.0.0

Date:	2012-12-24
License:	GPL-2
LazyLoad:	yes

This package incorporates the function `kern.den.circ` which computes the kernel circular density estimator. For choosing the smoothing parameter different functions are available: `nu.rt`, `nu.CV`, `nu.pi`, and `nu.boot`. For circular-linear regression (circular covariate and linear response), Nadaraya-Watson and Local-Linear smoothers, are also available in function `kern.reg.circ`. The function `nu.LSCV.reg` computes the least squares cross-validation bandwidth for those estimators. Functions `circsizer.density` and `circsizer.regression` provides CircSiZer maps for kernel density estimation and regression estimation, respectively. Functions `dcircmix` and `rcircmix` compute the density function and generate random samples of a circular distribution or a mixture of circular distributions, allowing for different components such as the circular uniform, von Mises, cardioid, wrapped Cauchy, wrapped normal and wrapped skew-normal. Finally, some data sets are provided. Missing data are allowed. Registries with missing data are simply removed.

For a complete list of functions, use `library(help="NPCirc")`.

Acknowledgements

This work has been supported by Project MTM2008-03010 from the Spanish Ministry of Science and Innovation IAP network (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences (StUDyS)) from Belgian Science Policy. The authors want to acknowledge Prof. Arthur Pewsey for facilitating data examples and for his comments.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal Mantainer: María Oliveira `<maria.oliveira@usc.es>`

References

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

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

Oliveira, M., Crujeiras, R.M. and Rodríguez–Casal (2013) CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-013-0249-0.

`nu.LSCV.reg` *Least squares cross-validation for circular-linear regression estimation*

Description

This function provides the least squares cross-validation (LSCV) smoothing parameter for the Nadaraya-Watson and Local-Linear estimators when the covariate is circular and the response variable is linear.

Usage

```
nu.LSCV.reg(x, y, method="LL", lower=0, upper=100, tol=0.1)
```

Arguments

`x`	Vector of data in radians (between 0 and 2π) for the circular covariate.
`y`	Vector of data for the dependent linear variable. This must be same length as `x`.
`method`	Character string giving the estimator to be used. This must be one of `"LL"` or `"NW"`. Default `method="LL"`.
`lower, upper`	`lower` and `upper` boundary of the interval to be used in the search for the value of the smoothing parameter. Default `lower=0` and `upper=100`.
`tol`	Convergence tolerance for `optimize`. Default `tol=0.1`.

Details

The LSCV smoothing parameter is obtained as the value of ν that minimizes expression (13) in Oliveira et al. (2013). The NAs will be automatically removed.

Value

Value of the smoothing parameter.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

See Also

`kern.reg.circ`

Examples

```
set.seed(2012)
n <- 100
x <- seq(0,2*pi,length=n)
y <- sin(x)+0.2*rnorm(n)
nu.LSCV.reg(x, y, method="LL", lower=1, upper=20)
nu.LSCV.reg(x, y, method="NW", lower=1, upper=20)
```

`nu.pi` *Plug–in rule*

Description

This function implements the von Mises scale plug-in rule for the smoothing parameter for density estimation when the number of components in the mixture is selected by Akaike Information Criterion (AIC) which selects the best model between a mixture of 2-5 von Mises distributions.

Usage

```
nu.pi(x, M=NULL, lower=0, upper=100, np=500, tol=0.1, outM=FALSE)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the smoothing parameter is to be computed.
`M`	Integer indicating the number of components in the mixture. If `M=1`, the rule of thumb is carried out with κ estimated by maximum likelihood. If `M=NULL`, AIC will be used.
`lower, upper`	`lower` and `upper` boundary of the interval to be used in the search for the value of the smoothing parameter. Default `lower=0` and `upper=100`.
`np`	Number of points where to evaluate the estimator for numerical integration. Default `np=500`.
`tol`	Convergence tolerance for `optimize`. Default `tol=0.1`.
`outM`	Logical; if `TRUE` the function also returns the number of components in the mixture. Default, `outM=FALSE`.

Details

The value of the smoothing parameter is chosen by minimizing the asymptotic mean integrated squared error (AMISE) derived by Di Marzio et al. (2009) assuming that the data follow a mixture of von Mises distributions. The number of componentes in the mixture can be fixed by the user, by specifying the argument `M` or selected by using AIC (default option) as described in Oliveira et al. (2012). The NAs will be automatically removed.

Value

Vector with the value of the smoothing parameter and the number of components in the mixture (if specified).

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

See Also

`kern.den.circ`, `nu.rt`, `nu.CV`, `nu.boot`

Examples

```
set.seed(2012)
n <- 100
x <- rcircmix(n,model=18)
nu.pi(x, M=3)
nu.pi(x, outM=TRUE)  # Using AIC
```

`nu.rt` *Rule of thumb*

Description

This function implements the selector proposed by Taylor (2008) for density estimation, based on an estimation of the concentration parameter of a von Mises distribution. The concentration parameter can be estimated by maximum likelihood or by a robustified procedure as described in Oliveira et al. (2013).

Usage

```
nu.rt(x, robust=FALSE, alpha=0.5)
```

Arguments

`x`	Sample of angles in radians (between 0 and 2π) from which the smoothing parameter is to be computed.
`robust`	Logical, if `robust=FALSE` the parameter κ is estimated by maximum likelihood, if `TRUE` it is estimated as described in Oliveira et al. (2012b). Default `robust=FALSE`.
`alpha`	Arc probability when `robust=TRUE`. Default is `alpha=0.5`. See Details.

Details

When `robust=TRUE`, the parameter κ is estimated as follows:

1. Select $\alpha \in (0, 1)$ and find the shortest arc containing $\alpha \cdot 100\%$ of the sample data.

2. Obtain the estimated $\hat{\kappa}$ in such way that the probability of a von Mises centered in the midpoint of the arc is `alpha`.

The NAs will be automatically removed.

See also Oliveira et al. (2012).

Value

Value of the smoothing parameter.

Author(s)

María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez–Casal

References

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

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

Taylor, C.C. (2008) Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493–3500.

See Also

`kern.den.circ`, `nu.CV`, `nu.pi`, `nu.boot`

Examples

```
set.seed(2012)
n <- 100
x <- rcircmix(n,model=7)
nu.rt(x)
nu.rt(x, robust=TRUE)
```

---

speed.wind *Wind speed and wind direction data*

Description

This dataset consists of hourly observations of wind direction and wind speed in winter season (from November to February) from 2003 until 2012 in the Atlantic coast of Galicia (NW–Spain). Data are registered by a buoy located at longitude -0.210E and latitude 43.500N in the Atlantic Ocean. The dataset `speed.wind2`, analized in Oliveira et al. (2013), is a subset of `speed.wind` which is obtained by taking the observations with a lag period of 95 hours.

Usage

```
data(speed.wind)
data(speed.wind2)
```

Format

`speed.wind` is a data frame with 19488 observations on six variables: day, month, year, hour, wind speed (in m/s) and wind direction (in degrees). `speed.wind2` is a subset with 200 observations.

Details

Data contains NAs. There is no data in November 2005, December 2005, January 2006, February 2006, February 2007, February 2009 and November 2009. Months of November 2004, December 2004, January 2007, December 2009 are not fully observed.

Source

Data can be freely downloaded from the Spanish Portuary Authority (http://www.puertos.es).

References

Oliveira, M., Crujeiras, R.M. and Rodríguez–Casal (2013) CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, DOI: 10.1007/s10651-013-0249-0.

Examples

```
data(speed.wind2)

t <- seq(0,2*pi,length=500)
dir <- rad(speed.wind2$Direction)

# Density
plot(circular(dir), stack=TRUE, zero=pi/2, rotation="clock", axes=FALSE)
axis.circular(at=circular(seq(0,2*pi,by=pi/4)),
labels=c("N","NE","E","SE","S","SW","W","NW","N"), zero=pi/2,
rotation="clock")
rose.diag(circular(dir), bins=16, add=TRUE, axes=FALSE, zero=pi/2,
rotation="clock")

rose.diag(circular(dir), bins=16, shrink=1.1, axes=FALSE, zero=pi/2,
rotation="clock")
axis.circular(at=circular(seq(0,7*pi/4,by=pi/4)),
labels=c("N","NE","E","SE","S","SW","W","NW"), zero=pi/2, rotation="clock")
lines(circular(t), kern.den.circ(dir,t,nu=1), lwd=2, lty=2, zero=pi/2,
rotation="clock")
lines(circular(t), kern.den.circ(dir,t,nu=10), lwd=2, lty=1, zero=pi/2,
rotation="clock")
lines(circular(t), kern.den.circ(dir,t,nu=60), lwd=2, lty=3, zero=pi/2,
rotation="clock")

# Regression
vel <- speed.wind2$Speed
nas <- which(is.na(vel))
dir <- dir[-nas]
vel <- vel[-nas]
radial.plot(vel, dir, rp.type="s", start=pi/2, clockwise=TRUE,
point.col="gray", radial.lim=c(0,15), label.pos=seq(0,7*pi/4,by=pi/4),
labels=c("N","NE","E","SE","S","SW","W","NW"))
radial.plot(as.numeric(kern.reg.circ(dir,vel,t,nu=1,method="LL")), t,
rp.type="p", add=TRUE, start=pi/2, clockwise=TRUE, radial.lim=c(0,15),
lwd=2, lty=2)
radial.plot(as.numeric(kern.reg.circ(dir,vel,t,nu=10,method="LL")),t,
rp.type="p", add=TRUE, start=pi/2, clockwise=TRUE, radial.lim=c(0,15),
lwd=2, lty=1)
radial.plot(as.numeric(kern.reg.circ(dir,vel,t,nu=60,method="LL")),t,
rp.type="p", add=TRUE, start=pi/2, clockwise=TRUE, radial.lim=c(0,15),
lwd=2, lty=3)
```

`temp.wind` *Temperature and wind direction data*

Description

The data, analized in Oliveira et al. (2013), consists of observations of temperature and wind direction during the austral summer season 2008-2009 (from November 2008 to March 2009) in Vinciguerra (Tierra del Fuego, Argentina).

Usage

```
data(temp.wind)
```

Format

A data frame with 3648 observations on four variables: `Date`, `Time`, `Temperature` (in degrees Celsius) and `Direction` (in degrees).

Details

Data contains NAs.

Source

The authors want to acknowledge Prof. Augusto Pérez–Alberti for providing the data, collected within the Project POL2006–09071 from the Spanish Ministry of Education and Science.

References

Oliveira, M., Crujeiras R.M. and Rodríguez–Casal, A. (2013) Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

Examples

```
data(temp.wind)

winddir <- rad(temp.wind[,4]) # wind direction in radians
temp <- temp.wind[,3]
nas <- which(is.na(winddir))
winddir <- winddir[-nas]
temp <- temp[-nas]

# value of the smoothing parameter selected using the function nu.LSCV.reg
```

```
# with method="LL"
nu <- 3.41
t <- seq(0,2*pi,length=100)
est <- kern.reg.circ(winddir, temp, t, nu=nu, method="LL")

# Circular representation
radial.plot(temp, winddir, rp.type="s", labels=c("N","NE","E","SE","S","SW",
"W","NW"), start=pi/2, clockwise=TRUE, label.pos=seq(0,7*pi/4,by=pi/4),
lwd=2, point.col="grey50", radial.lim=c(-10,15))
radial.plot(as.vector(est), rp.type="p", start=pi/2, clockwise=TRUE, lwd=2,
radial.lim=c(-10,15), add=TRUE)

# Linear representation
plot(t, est, type="l", xlab="", ylab="Temperatute (žC)", axes=FALSE)
axis(1, labels=c("N","NE","E","SE","S","SW","W","NW","N"),
at=seq(0,2*pi,by=pi/4))
axis(2)
```

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Summary

In many applied fields such as biology, metoerology, ecology or medicine, measurements from the process under study are directions. Circular data are a particular case of directional data where the observations are directions in two dimensions which can be expressed as angles. Due to the circular nature of this kind of data, which implies fixing a reference point and a sense of rotation in order to define a circular observation, the statistical methods for the analysis of linear data are not appropriate for the analysis of circular data. Some general references on circular statistics are Mardia (1972), Batschelet (1981), Fisher (1993), Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001).

Density estimation with circular data and regression estimation with a scalar response and a circular covariate are two usual problems in a large variety of disciplines. Both problems can be approached from a parametric or a nonparametric perspective. In most of the applied papers dealing with circular data, parametric methods are used. Thus, the main goal of this work is focused on nonparametric techniques for circular data, putting special emphasis in the smoothing parameter selection problem and in the assessment of which features observed in a smoothed curve are significant.

This thesis has been structured in four chapters. For each of them, a brief summary is given below, with specific references to the main achievements.

Chapter 1: Circular models and data. In this chapter the concept of circular distribution and the most important circular parametric distribution families are introduced. Specifically, von Mises, cardioid and some wrapped distributions are reviewed. The method of moments and the maximum likelihood method for estimating the parameters of a von Mises distribution are detailed. The EM algorithm for estimating the parameters of a finite mixture of von Mises distributions is also specified. Although this dissertation is focused in nonparametric methods, the introduction of some parametric techniques is required since they will be needed in the following chapters. In the last section of this chapter, several datasets which will be analized by using the techniques presented along this thesis are described.

The von Mises distribution, $vM(\mu, \kappa)$, is a unimodal and symmetric distribution characterized

by two parameters a mean direction $\mu \in [0, 2\pi)$ and a concentration parameter $\kappa \geq 0$. Its density function is

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad 0 \leq \theta < 2\pi,$$

where I_r denotes the modified Bessel function of the first kind and order r.

More flexible models, allowing for multimodality and/or asymmetry, can be obtained by mixing a finite number of circular distributions. A particular case is the mixture of M von Mises distributions $vM(\mu_m, \kappa_m)$, $m = 1, \ldots, M$, whose density is

$$f(\theta; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) = \sum_{m=1}^{M} p_m f_m(\theta; \mu_m, \kappa_m), \quad 0 \leq \theta < 2\pi,$$

where $\boldsymbol{p} = (p_1, \ldots, p_M)$ with $p_m > 0$ and $\sum_{m=1}^{M} p_m = 1$ are the weights of the component densities, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M) \in [0, 2\pi)^M$ is the vector of circular means, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_M) \in (\mathbb{R}^+)^M$ is the vector of concentrations and f_m denotes the density function of a von Mises distribution $vM(\mu_m, \kappa_m)$ for $m = 1, \ldots, M$. The main problem for estimating the parameters of a mixture of von Mises distributions lies in the fact that it is not known which density component generates each observation. In this case, when the information provided by the sample is imcomplete, the EM algorithm (see Banerjee et al., 2005) allows to estimate its parameters.

Chapter 2: Nonparametric curve estimation for circular data. This chapter is devoted to the nonparametric density estimation and nonparametric regression estimation for a circular explanatory variable and a linear response. In the density setting, the kernel density estimator for circular data is introduced. This estimator depends on a smoothing parameter which controls the global aspect of the curve. In this chapter, different techniques for selecting the smoothing parameter are reviewed. The main contribution is the introduction of a new smoothing parameter selector that allows to estimate complex circular densities, accounting for asymmetry and/or multimodality. In the regression setting, a review of the methods for a circular covariate and a scalar response is provided, both for kernel methods and spline smoothers. Specifically, the adaptation of the Nadaraya–Watson and Local Linear estimators to the circular nature of the covariate and the periodic smoothing spline estimator are presented. The performance of the three estimators is compared in a simulation study.

Given a random sample of angles $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$ from a circular random variable Θ with unknown density f, the circular kernel density estimator of f is defined as:

$$\hat{f}(\theta; \nu) = \sum_{i=1}^{n} \frac{1}{n} K_\nu(\theta - \Theta_i), \quad 0 \leq \theta < 2\pi,$$

where K_ν is a circular kernel function with concentration parameter $\nu > 0$. As a circular kernel, the von Mises distribution with concentration parameter ν is considered. With this specific kernel,

the circular kernel density estimator has the following expression:

$$\hat{f}(\theta;\nu) = \frac{1}{n2\pi I_0(\nu)} \sum_{i=1}^{n} e^{\nu \cos(\theta - \Theta_i)}, \quad 0 \leq \theta < 2\pi.$$

A critical issue when applying this estimator in practice is the choice of the smoothing parameter ν, since large values of this parameter will lead to undersmoothed estimators and small values of ν will provide oversmoothed estimators. Usually, the value of the smoothing parameter is selected in order to minimize some error criterion, such as the mean integrated squared error (MISE, $\text{MISE}(\nu) = \mathbb{E}(\int(\hat{f} - f)^2)$, where $\hat{f}$ is the nonparametric estimator that depends on ν). For the circular kernel density estimator defined above, the asymptotic expression for the MISE (AMISE), when $\nu \to \infty$ and $\sqrt{\nu} n^{-1} \to 0$, is given by

$$\text{AMISE}(\nu) = \left\{ \frac{1}{16} \left[1 - \frac{I_2(\nu)}{I_0(\nu)} \right]^2 \int_0^{2\pi} \left(f''(\theta) \right)^2 d\theta + \frac{I_0(2\nu)}{2n\pi \left(I_0(\nu) \right)^2} \right\},$$

where f'' denotes the second–order derivative of the target density to be estimated (see Di Marzio et al., 2009).

The new proposal for selecting the smoothing parameter consists in estimating the integral $\int_0^{2\pi} \left(f''(\theta) \right)^2 d\theta$, which appears in the AMISE expression, taking a finite mixture of von Mises distributions as reference density. Hence, the plug–in selector (see Oliveira et al., 2013b) is obtained as follows:

Step 1. Select the number of mixture components M for the reference distribution, for example, by using the Akaike Information Criterion.

Step 2. Estimate the AMISE as follows:

- Step 2.1. Estimate the parameters in the von Mises mixture, $(\mu_m, \kappa_m, \alpha_m)$, para $m = 1, \ldots, M$, by using the EM algorithm.
- Step 2.2. Compute the integral $\int (\hat{f}''(\theta))^2 d\theta$, where $\hat{f}''$ is the second derivative of the density function of a mixture of M von Mises distributions with the parameters estimated in the previous step.
- Step 2.3. Plug–in the quantity above in the AMISE expression to get $\widehat{\text{AMISE}}(\nu)$.

Step 3. Minimize $\widehat{\text{AMISE}}(\nu)$ and obtain $\hat{\nu}_{PI}$.

The performance of the new plug–in selector is compared through a simulation study with the cross–validation rules introduced by Hall et al. (1987), the rule of thumb proposed by Taylor (2008) and the bootstrap method proposed by Di Marzio et al. (2011). Results obtained from the simulation study showed that the new plug-in rule procedure for smoothing parameter selection in circular density estimation behaves satisfactorily for all the simulation scenarios, equalizing or even

outperforming the other methods. The good performance of the plug–in selector is also observed in the real data examples, corresponding to cross–beds azimuths and orientation of dragonflies.

The second part of this chapter is focused on nonparametric regression estimation for a circular explanatory variable and a linear response. In this setting, two types of smoothers are studied: kernel smoothers and spline smoothers.

Let $\{(\Theta_i, Y_i),\ i = 1, \ldots, n\}$ be a random sample from a circular random variable Θ and a linear random variable Y, respectively. The relation between these variables may be modelled by

$$Y_i = f(\Theta_i) + \sigma(\Theta_i)\varepsilon_i,\ i = 1, \ldots, n$$

where f denotes the unknown regression function, $\sigma^2(\cdot)$ is the conditional variance of Y given Θ and ε_i are real–valued random variables with zero mean and unit variance.

Following Di Marzio et al. (2009), the Local Lineal estimator for $f(\theta)$ is given by $\hat{f}_{CLL}(\theta;\nu) = \hat{\beta}_0$ where

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i)\left[Y_i - (a + b\sin(\theta - \Theta_i))\right]^2,$$

where K_ν denotes a $vM(0,\nu)$.

If the regression function at θ is locally approximated by a constant instead of using a trigonometric polynomial, the Nadaraya–Watson estimator for circular–linear data is obtained:

$$\hat{f}_{CNW}(\theta;\nu) = \frac{\sum_{i=1}^{n} Y_i K_\nu(\theta - \Theta_i)}{\sum_{i=1}^{n} K_\nu(\theta - \Theta_i)}.$$

As for density estimation, choosing the smoothing parameter is of crucial importance in regression analysis. A simple and widely used procedure for smoothing selection in the regression setting is the least squares cross–validation rule, which chooses ν as the value minimizing

$$CV(\nu) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{f}^{-i}(\Theta_i;\nu)\right]^2,$$

where $\hat{f}^{-i}$ denotes the estimator of f computed leaving out the pair (Θ_i, Y_i).

Periodic smoothing splines, introduced by Cogburn and Davis (1974), offer an alternative to the Nadaraya–Watson and Local Linear estimators and moreover, this kind of smoother is valid when the covariate is any periodic random variable with period T (the distribution of $(X + T)$ coincides with the distribution of X), in particular, when $T = 2\pi$.

Let $\{(X_i, Y_i), i = 1, \ldots, n\} \in [0, T) \times \mathbb{R}$ be a random sample from a periodic random variable X with period T and a linear random variable Y. Assume that data are sorted across the covariate and there is no repeated data. Consider again the noparametric regression model

$$Y_i = f(X_i) + \sigma(X_i)\varepsilon_i,\ i = 1, \ldots, n,$$

where f is an unknown regression function that must be estimated, ε_i are random variables with zero mean and unit variance and $\sigma^2(\cdot)$ is the conditional variance of Y given X.

The smoothing spline estimator of the regression function f, $\hat{f}_\lambda$, is obtained by finding the smooth function that minimizes the penalized least squares criterion:

$$S(g) = \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx,$$

over the class of twice continuously differentiable periodic functions, g, with period T, for some $\lambda > 0$. The parameter λ is known as the smoothing parameter. In this case, small values of λ will provide undersmoothed estimators whereas large values of this parameter will lead to oversmoothed estimators.

It is shown that, for $\lambda > 0$, $\hat{f}_\lambda$ is neccesarily a periodic cubic spline un spline on $[X_1, X_{n+1}]$ with knots at sampling points X_i, $i = 1, \ldots, (n+1)$, where $X_{n+1} = X_1 + T$. Fixed a value of the smoothing parameter λ, the estimator evaluated in the sample points, $\hat{\boldsymbol{f}}_\lambda = (\hat{f}_\lambda(X_1), \ldots, \hat{f}_\lambda(X_n))^t$, is given by:

$$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ and $A_\lambda = (I_n + \lambda K)^{-1}$ with $K = QR^{-1}Q$ (matrices Q and R are defined in Appendix C). The value of the estimator can be obtained for any point $x \in [X_1, X_{n+1})$. Hence, for a grid of locations $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ with $x_i \in [X_1, X_{n+1})$, the value of the estimator on those points, $\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = (\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_N))^t$, is given by

$$\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = M A_\lambda \boldsymbol{Y},$$

where M is an $(N \times n)$ coefficient matrix defined as in Appendix C. As for kernel estimators, the smoothing parameter λ can be selected by cross-validation.

The performance of kernel and periodic smoothing splines estimators is compared through a simulation study, where the smoothing parameter has been selected by cross–validation. In the simulation study, it is observed that Nadaraya–Watson and Local Linear estimators perform similarly for all the regression models considered whereas, the periodic smoothing spline estimator provides better results in all the scenarios in terms of integrated squared error. Moreover, the estimators are applied to analyze the relation between the wind direction and wind speed in the Atlantic coast of Galicia.

Chapter 3. Assessment of significant features in nonparametric curve estimates. This chapter introduces a new nonparametric technique for the exploratory analysis of circular data, namely CircSiZer, which allows to know which observed features in the smoothed curve (density or regression) are statistically significant and which ones are simply artifacts of the sampling noise.

CircSiZer (Oliveira et al. 2013) is an adaptation to circular data of the SiZer method proposed originally by Chaudhuri and Marron (1999) to circular data. The idea of SiZer methods is to

provide a graphical tool that shows the increasing/decreasings patterns of a smooth curve. SiZer methods considers a wide range of smoothing parameters and so, the smoothing parameter selection is avoided. Thus, for each value of the smoothing parameter, CircSiZer address the question of which features, like peaks and valleys, are really present, i.e., which ones are really significant. CircSiZer assess the significance of features by constructing confidence intervals for the derivative of the smoothed curve $f'(\theta;\tau) \equiv \mathbb{E}(\hat{f}'(\theta;\tau))$, where f denotes the density or regression function and τ is the smoothing parameter ($\tau \equiv \nu$ for kernel estimators and $\tau \equiv \lambda$ for the smoothing spline estimator).

So, at each pair $(\theta;\tau)$ with $\theta \in [0, 2\pi)$ and $\tau > 0$, the curve at a smoothing level τ is significantly increasing (decreasing) if the confidence interval is above (below) 0 and if the confidence interval contains 0, the curve at smoothing level τ and at point θ does not have a statistically significant slope. This information is displayed in a circular color map, the CircSiZer map, in such a way that, at a given τ, the performance of the estimated curve in the interval $[0, 2\pi)$ is represented by a color ring where differents colors will allow to indentify the increasing and decreasing regions.

Confidence intervals are of the form

$$\left(\hat{f}'(\theta;\tau) - q^{(1-\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)), \hat{f}'(\theta;\tau) - q^{(\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))\right),$$

where $\hat{f}'(\theta;\nu)$ is the estimator of the derivative of the density or regression function, $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are appropriate quantiles and $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))$ is an estimator of the standard deviation of $\hat{f}'(\theta;\tau)$.

Four alternatives of computing the quantiles $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are studied, two approaches are based on a normal approximation (see Chaudhuri e Marron, 1999):

(i) $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are the quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of the standard normal distribution.

(ii) $q^{(1-\alpha/2)} = -q^{(\alpha/2)} = \Phi^{-1}\left\{\frac{1+(1-\alpha)^{1/m(\tau)}}{2}\right\}$ where Φ^{-1} is the inverse of the standard normal distribution function and $m(\tau) = n/\mathrm{avg}_{\theta\in\mathcal{D}_\tau} ESS(\theta;\tau)$ with $ESS(\theta;\tau)$ the Effective Sample Size for each $(\theta;\tau)$ and $\mathcal{D}_\tau = \{\theta : ESS(\theta;\tau) \geq 5\}$.

And two approaches are based on bootstrap techniques:

(iii) $q^{(1-\alpha/2)}$ e $q^{(\alpha/2)}$ are the sample quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of $Z_1^*(\theta;\tau), \ldots, Z_B^*(\theta;\tau)$ where

$$Z_b^*(\theta;\tau) = \frac{\hat{f}'(\theta;\tau)^{*b} - \hat{f}'(\theta;\tau)}{\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)^{*b})},\ b = 1, \ldots, B$$

is the standardized version of the derivative of the estimator computed for the b–th bootstrap sample drawn with replacement from the data.

(iv) $q^{(1-\alpha/2)}$ is the quantile of order $(1-\alpha/2)$ of $Z_{sup}^{*1}, \ldots, Z_{sup}^{*B}$ and $q^{(\alpha/2)}$ is the quantile of order $\alpha/2$ of $Z_{inf}^{*1}, \ldots, Z_{inf}^{*B}$ where

$$Z_{inf}^{*b} = \inf_{\theta \in \mathcal{D}_\tau} Z_b^*(\theta, \tau) \text{ e } Z_{sup}^{*b} = \sup_{\theta \in \mathcal{D}_\tau} Z_b^*(\theta, \tau), \quad b = 1, \ldots, B$$

Whereas (i) and (iii) provide pointwise confidence intervals, (ii) and (iv) provide simultaneous confidence intervals.

In order to obtain the confidence intervals, an estimator of the standard deviation of $\hat{f}'(\theta;\tau)$ is required. For density estimation, an estimation of the variance is provided by

$$\widehat{\text{var}}\left(\hat{f}'(\theta;\tau)\right) = n^{-1}s^2\left(K'_\nu(\theta - \Theta_1), \ldots, K'_\nu(\theta - \Theta_n)\right), \quad 0 \leq \theta < 2\pi,$$

where s^2 is the usual sample variance of n data. For regression estimation, since the proposed estimators are linear, the value of $\hat{f}'(\theta;\tau)$ in a grid $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_2)$, may be written as $\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}} = H\boldsymbol{Y}$ where H is an $(N \times n)$ coefficient matrix and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ is the response vector. Thus, for fixed design,

$$\text{var}(\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}}) = H\Sigma H^t,$$

where $\Sigma = \text{diag}\left\{\sigma^2(\Theta_1), \sigma^2(\Theta_2), \ldots, \sigma^2(\Theta_n)\right\}$. If homoscedasticity is assumed then, $\sigma^2(\Theta_i) = \sigma^2$ $(i = 1, \ldots, n)$ may be estimated by using the estimator due to Rice (1984). For random design, $\widehat{\text{sd}}\left(\hat{f}'(\theta;\tau)\right)$ is estimated by bootstrap:

$$\widehat{\text{sd}}\left(\hat{f}'(\theta;\tau)\right) = \left[s^2\left(\hat{f}'^{*1}(\theta;\tau), \ldots, \hat{f}'^{*B}(\theta;\tau)\right)\right]^{1/2},$$

where $\hat{f}'^{*b}(\theta;\tau)$ is the derivative of the estimator computed for the b–th bootstrap sample drawn with replacement from the data.

For the density setting, through a simulation study, the coverage of pointwise confidence intervals based on the normal approximation and the ones based on bootstrap have been compared. From the results, it is observed that, for large values of the smoothing parameter, both intervals tend to identify some artificial features as significant, this fact is enhanced for bootstrap intervals. For small values of this parameter, both approaches behave similarly, in terms of coverage and number of modes flagged as significant.

Another simulation study compares the coverage of simultaneous confidence intervals based on the normal approximation and the ones based on bootstrap. The study allowed to observe that the coverage of bootstrap simultaneous confidence intervals is close to the nominal level $(1 - \alpha)$ whereas the coverage of normal simultaneous confidence intervals is below the target value. It also was observed that the use of bootstrap simultaneous confidence intervals may present difficulties for attaining the goal of detecting the modes of the model. In this sense, it was seen that CircSiZer map with pointwise confidence intervals may help to identify the modes presented by the model however, the interpretation of pointwise CircSiZer map must be done carefully since it may flags spurious modes as significant.

In the regression setting, the coverage of simultaneous bootstrap confidence intevals based on the normal approximation has been studied for the case of fixed design. Results showed that the

coverage of these intervals is below the nominal value. The same behaviour has been observed for the Local Linear estimator and for the periodic smoothing spline estimator. For random design, the performance of CircSiZer with simultaneous bootstrap confidence intervals has been checked in some simulated data. In this case, both estimators perform similarly.

Finally, the practical usefulness of the proposed CircSiZer map is illustrated by the analysis of some real datasets where the goal is to study when certain changes in temperature occur, in what directions craks in cemented femoral components appear or which is the relation between the wind direction and the wind speed.

Chapter 4: Software: NPCirc package. The last chapter is devoted to the presentation of the `NPCirc` package for R, which implements the estimators and methods described in the previous chapters and which is intended to provide R users with a comprehensive set of functions for nonparametric density and regression analysis with circular data.

Specifically, it implements the circular kernel density estimator and the differents methods for selecting the smoothing parameter described in Chapter 2. In the regression setting, for linear response and scalar covariate, `NPCirc` contains a function for nonparametric estimation of the regression curve by Nadaraya–Watson and Local Linear, as described in Chapter 2. Both for density and regression, the CircSiZer method with pointwise bootstrap confidence intervals is also available. The library also includes one function which allows to compute the density function of a circular distribution (von Mises, cardioid, wrapped Cauchy, wrapped normal and wrapped skew–normal) or the density of a mixture of these distributions and, another function which allows for random generation from a circular distribution or from a mixture of circular distributions. The `NPCirc` package also includes the datasets analyzed along the manuscript.

Appendix A. This appendix includes the specific formulae of those circular models considered in the simulation study carried out in Chapter 2 and used for illustration purposes throughout the manuscript.

Appendices B and C. These appendices give technical details on kernel regression smoothers and periodic smoothing splines, which complement Chapters 2 and 3.

Appendix D. This appendix describes the functions in the `NPCirc` library, giving instructions about their usage and arguments and illustrating them with examples.

Acknowledgements

I would like to thank my advisors, Prof. Rosa M. Crujeiras and Prof. Alberto Rodríguez Casal for their work and support during these years. I also wish to thank Prof. Augusto Pérez Alberti from the Department of Physical Geography of the University of Santiago de Compostela and Dr. Kenneth A. Mann from the Upstate Medical University (New York) for kindly providing some real datasets that motivate part of the work done in this thesis.

This work has been supported by Project MTM2008–03010 from the Spanish Ministry of Science and Innovation, and by the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences), from Belgian Science Policy.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Resumo en galego

En numerosos campos aplicados, como a bioloxía, a meteoroloxía, a ecoloxía ou a medicina, as medidas obtidas dos procesos de interese son direccións. Os datos circulares, dos que nos ocupamos nesta tese, son un caso particular de datos direccionais onde as observacións son direccións en dúas dimensións que adoitan expresarse mediante ángulos. Debido á natureza circular deste tipo de datos, que conleva a necesidade de fixar un punto de referencia e un sentido de rotación para definir unha observación circular, as técnicas estatísticas utilizadas para a análise de datos lineais non resultan adecuadas para a análise de datos circulares. Neste eido, algunhas referencias xerais son os libros de Mardia (1972), Batschelet (1981), Fisher (1993), Mardia e Jupp (2000) e Jammalamadaka e SenGupta (2001).

A estimación da función de densidade da que provén unha mostra de datos circulares, así como a estimación da función de regresión cando a variable explicativa é circular e a variable resposta é escalar, son dous problemas recurrentes en diferentes contextos. Ambos os dous poden enfocarse dende unha perspectiva paramétrica ou nonparamétrica. Na maioría dos traballos de carácter aplicado que se atopan na literatura sobre datos circulares utilízanse técnicas paramétricas. Así, o obxectivo principal deste traballo é afondar nas técnicas nonparamétricas para datos circulares, facendo especial fincapé no problema que atangue á selección do parámetro de suavizado e á determinación das características significativas observadas na curva suavizada.

A continuación inclúese un breve resumo de cada un dos capítulos que constitúen esta tese doutoral, facendo mención aos principais avances obtidos en cada un deles.

Capítulo 1: Modelos e datos circulares. Neste capítulo introdúcese o concepto de distribución circular e revísanse as principais familias paramétricas de distribucións circulares: von Mises, cardioide e varias distribucións enroladas. Detállase a estimación de parámetros polo método dos momentos e por máxima verosimilitude para a distribución von Mises e a estimación por máxima verosimilitude dos parámetros dunha mixtura de tales distribucións, utilizando neste último caso o algoritmo EM. Se ben esta tese se centra en métodos nonparamétricos, a introdución de técnicas paramétricas será necesaria como ferramenta para o desenvolvemento de capítulos posteriores. Na última sección deste capítulo descríbense varios conxuntos de datos reais que serán analizados utilizando as diferentes técnicas presentadas ao longo da tese.

A distribución von Mises, $vM(\mu, \kappa)$, é unha distribución unimodal e simétrica caracterizada por dous parámetros, unha dirección media $\mu \in [0, 2\pi)$ e un parámetro de concentración $\kappa \geq 0$. A súa función de densidade é

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad 0 \leq \theta < 2\pi,$$

onde I_r denota a función de Bessel modificada de primeiro tipo e orde r.

Poden obterse modelos circulares máis complexos presentando multimodalidade e/ou asimetría mediante mixturas finitas de distribucións circulares. Un caso particular son as mixturas de M distribucións von Mises $vM(\mu_m, \kappa_m)$, $m = 1, \ldots, M$, cuxa densidade é

$$f(\theta; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{p}) = \sum_{m=1}^{M} p_m f_m(\theta; \mu_m, \kappa_m), \quad 0 \leq \theta < 2\pi,$$

onde $\boldsymbol{p} = (p_1, \ldots, p_M)$ con $p_m > 0$ e $\sum_{m=1}^{M} p_m = 1$ é o vector de pesos de cada distribución, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M) \in [0, 2\pi)^M$ é o vector de dirección medias, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_M) \in (\mathbb{R}^+)^M$ é o vector de parámetros de concentración e f_m denota a función de densidade dunha distribución $vM(\mu_m, \kappa_m)$, para $m = 1, \ldots, M$. O principal problema á hora de estimar os parámetros dunha mixtura de tales distribucións reside en que non se adoita coñecer a que compoñente da mixtura pertence cada dato. Neste caso, cando a información da mostra está incompleta, o algoritmo EM (véxase Banerjee et al., 2005) permite a estimación por máxima verosimilitud dos parámetros.

Capítulo 2: Estimación nonparamétrica de curvas para datos circulares. Este capítulo adícase á estimación nonparamétrica da función de densidade con datos circulares e da regresión con covariable circular e resposta escalar. No contexto da densidade, introdúcese o estimador tipo núcleo circular como estimador nonparamétrico da densidade circular. Este estimador depende dun parámetro de suavizado que controla o seu aspecto global. Neste capítulo revísanse os procedementos existentes para a selección do parámetro de suavizado do que depende dito estimador. A principal achega neste ámbito é a proposta dun novo selector do parámetro de suavizado que permite estimar nonparametricamente densidades circulares cunha estrutura complexa, como pode ser a asimetría e/ou multimodalidade. No contexto da regresión, faise unha revisión dos métodos existentes para o caso dunha covariable circular e unha variable resposta escalar, tanto das técnicas tipo núcleo como das técnicas baseadas en suavizadores tipo spline. Concretamente, preséntanse os estimadores tipo núcleo de Nadaraya–Watson e Local Lineal convenientemente adaptados á natureza circular da covariable e os splines de suavizado periódicos. O comportamento dos tres estimadores compárase nun estudo de simulación. A continuación, descríbense brevemente as achegas neste capítulo.

Dada unha mostra aleatoria simple $\Theta_1, \Theta_2, \ldots, \Theta_n \in [0, 2\pi)$ dunha variable circular Θ con función de densidade descoñecida f, defínese o estimador tipo núcleo da densidade circular f como:

$$\hat{f}(\theta; \nu) = \sum_{i=1}^{n} \frac{1}{n} K_\nu(\theta - \Theta_i), \quad 0 \leq \theta < 2\pi,$$

onde K_ν é unha función núcleo circular con parámetro de concentración $\nu > 0$. Como núcleo circular, considérase a función de densidade dunha distribución von Mises con parámetro de concentración ν. Con este núcleo específico, o estimador da densidade circular ten a seguinte expresión:

$$\hat{f}(\theta;\nu) = \frac{1}{n2\pi I_0(\nu)} \sum_{i=1}^{n} e^{\nu\cos(\theta-\Theta_i)}, \quad 0 \leq \theta < 2\pi.$$

O problema principal do estimador tipo núcleo circular reside na elección do parámetro de suavizado ν, xa que valores grandes deste parámetro proporcionarán estimadores infrasuavizados e valores pequenos proporcionarán estimadores sobresuavizados. O valor de dito parámetro adoita calcularse de maneira que se minimice algún criterio de erro, como o erro cadrático medio integrado (MISE, $\mathrm{MISE}(\nu) = \mathbb{E}(\int(\hat{f} - f)^2)$, onde $\hat{f}$ é o estimador nonparamétrico que depende de ν). Na práctica, utilízase a versión as asintótica do MISE (AMISE), que para o estimador tipo núcleo circular, cando $\nu \to \infty$ e $\sqrt{\nu}n^{-1} \to 0$, vén dada por

$$\mathrm{AMISE}(\nu) = \left\{ \frac{1}{16}\left[1 - \frac{I_2(\nu)}{I_0(\nu)}\right]^2 \int_0^{2\pi} \left(f''(\theta)\right)^2 d\theta + \frac{I_0(2\nu)}{2n\pi\left(I_0(\nu)\right)^2} \right\},$$

onde f'' denota a derivada segunda da densidade descoñecida (véxase Di Marzio et al., 2009).

O novo selector proposto consiste en estimar a integral $\int_0^{2\pi} \left(f''(\theta)\right)^2 d\theta$, que aparece na expresión do AMISE, tomando como densidade de referencia unha mixtura finita de distribucións von Mises. Deste xeito, o selector plug–in que se propón (véxase Oliveira et al., 2013b) obtense como segue:

Paso 1. Seleccionar o número de compoñentes M da mixtura para a distribución de referencia, por exemplo, utilizando o criterio de información de Akaike.

Paso 2. Estimar o AMISE como segue:

- Paso 2.1. Estimar os parámetros da mixtura de distribucións von Mises, $(\mu_m, \kappa_m, \alpha_m)$, para $m = 1, \ldots, M$, mediante o algoritmo EM.
- Paso 2.2. Calcular a integral $\int(\hat{f}''(\theta))^2 d\theta$, onde $\hat{f}''$ denota a derivada segunda da función de densidade dunha mixtura de M distribucións von Mises cos parámetros estimados no paso anterior.
- Paso 2.3. Substituír esta cantidade na expresión do AMISE para obter $\widehat{\mathrm{AMISE}}(\nu)$.

Paso 3. Minimizar $\widehat{\mathrm{AMISE}}(\nu)$ con respecto a ν e obter $\hat{\nu}_{PI}$.

O comportamento do selector plug–in compárase coas regras de validación cruzada introducidas por Hall et al. (1987), a regra proposta por Taylor (2008) e o método bootstrap proposto en Di Marzio et al. (2011) nun estudo de simulación. Os resultados de dito estudo amosan que a nova regra plug–in se comporta satisfactoriamente en todos os escenarios considerados igualando ou incluso superando aos outros métodos. O bo comportamento do selector tamén se observa na

aplicación a tres conxuntos de datos clásicos, relativos á orientación de estratos cruzados e a orientación de libélulas.

A segunda parte deste capítulo adícase a estimación nonparamétrica da función de regresión cando a variable resposta é lineal e a covariable é circular. Neste contexto, estúdanse dous tipos de suavizadores: estimadores tipo núcleo e estimadores tipo spline.

Sexa $\{(\Theta_i, Y_i),\ i = 1, \ldots, n\}$ unha mostra aleatoria da variable aleatoria bidimensional (Θ, Y), onde Θ é unha variable circular e Y é unha variable lineal. Dende agora, asúmese que o erro e a covariable son incorrelados. A relación entre estas dúas variables pode escribirse da forma

$$Y_i = f(\Theta_i) + \sigma(\Theta_i)\varepsilon_i,\ i = 1, \ldots, n$$

onde f é a función de regresión que supoñemos descoñecida, $\sigma^2(\cdot)$ é a varianza condicional de Y dada Θ e ε_i son variables aleatorias con media cero e varianza un. Seguindo a Di Marzio et al. (2009), o estimador Local Lineal para $f(\theta)$ vén dado por $\hat{f}_{CLL}(\theta;\nu) = \hat{\beta}_0$ onde

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i)\left[Y_i - (a + b\sin(\theta - \Theta_i))\right]^2,$$

sendo K_ν a función de densidade da distribución $vM(0, \nu)$.

Se a función de regresión se aproxima localmente por unha constante en lugar de usar un polinomio trigonométrico, obtense o estimador de Nadaraya–Watson para covariable circular que vén dado por

$$\hat{f}_{CNW}(\theta;\nu) = \frac{\sum_{i=1}^{n} Y_i K_\nu(\theta - \Theta_i)}{\sum_{i=1}^{n} K_\nu(\theta - \Theta_i)}.$$

Ao igual que na estimación da función de densidade, a selección do parámetro de suavizado en regresión é de crucial importancia. A regra de validación cruzada por mínimos cadrados selecciona ν de maneira que se minimice

$$CV(\nu) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{f}^{-i}(\Theta_i;\nu)\right]^2,$$

onde $\hat{f}^{-i}$ denota o estimador da función de regresión construido a partir da mostra orixinal despois de eliminar o par (Θ_i, Y_i).

Unha alternativa aos estimadores tipo núcleo de Nadaraya–Watson e Local–Lineal son os splines de suavizado periódicos, introducidos por Cogburn e Davis (1974), que ademais, son válidos para a estimación da función de regresión cando a covariable é calquera variable periódica de período T (a distribución de $(X + T)$ coincide coa distribución de X), en particular, cando $T = 2\pi$.

Sexa $\{(X_i, Y_i), i = 1, \ldots, n\} \in [0, T) \times \mathbb{R}$ unha mostra aleatoria da variable aleatoria bidimensional (X, Y) onde X é unha variable periódica con período T e Y é unha variable linear. Asúmese que os datos están ordeados segundo a covariable e non hai datos repetidos. Considerando novamente o modelo de regresión nonparamétrico

$$Y_i = f(X_i) + \sigma(X_i)\varepsilon_i,\ i = 1, \ldots, n,$$

onde f é a función de regresión descoñecida que debe estimarse, ε_i son variables aleatorias con media cero e $\sigma^2(\cdot)$ é a varianza condicional de Y dada X.

O estimador tipo spline da función de regresión, $\hat{f}_\lambda$, obtense como a función que minimiza a seguinte suma de cadrados penalizada:

$$S(g) = \sum_{i=1}^{n} [Y_i - g(X_i)]^2 + \lambda \int_0^T \left[g''(x)\right]^2 dx,$$

sobre a clase de funcións periódicas, g, con período T e dúas veces diferenciables, para algún $\lambda > 0$. O parámetro λ coñécese como parámetro de suavizado. Neste caso, valores pequenos de λ proporcionarán estimadores infrasuavizados, mentres que valores grandes deste parámetro darán lugar a estimadores sobresuavizados.

Demóstrase que, para $\lambda > 0$, $\hat{f}_\lambda$ é necesariamente un spline cúbico periódico en $[X_1, X_{n+1}]$ con nodos nos puntos da mostra X_i, $i = 1, \ldots, (n+1)$, onde $X_{n+1} = X_1 + T$. Fixado un valor do parámetro de suavizado λ, o estimador avaliado nos puntos mostrais $\hat{\boldsymbol{f}}_\lambda = (\hat{f}_\lambda(X_1), \ldots, \hat{f}_\lambda(X_n))^t$ vén dado por:

$$\hat{\boldsymbol{f}}_\lambda = A_\lambda \boldsymbol{Y}$$

onde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ e $A_\lambda = (I_n + \lambda K)^{-1}$ con $K = QR^{-1}Q$ (as matrices Q e R están definidas no Apéndice C). O valor do estimador pode obterse para calquera punto $x \in [X_1, X_{n+1})$. Así para unha grella $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ con $x_i \in [X_1, X_{n+1})$, o estimador en ditos puntos, $\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = (\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_N))^t$, vén dado por

$$\hat{\boldsymbol{f}}_{\lambda,\boldsymbol{x}} = M A_\lambda \boldsymbol{Y},$$

onde M é unha matriz $(N \times n)$ definida como no Apéndice C. Ao igual que para os estimadores tipo núcleo, o parámetro de suavizado λ pode seleccionarse por validación cruzada.

O comportamento dos estimadores tipo núcleo e do estimador baseado en splines de suavizado periódicos compárase nun breve estudo de simulación, onde os respectivos parámetros de suavizado son seleccionados mediante validación cruzada por mínimos cadrados. No estudo de simulación obsérvase que os estimadores de Nadaraya–Watson e Local–Lineal se comportan de maneira moi semellante mentres que o estimador tipo spline produce mellores resultados en termos do erro cadrático integrado. Ademais, os estimadores aplícanse a un conxunto de datos reais sobre a dirección e velocidade do vento na costa atlántica de Galicia.

Capítulo 3. Determinación de características significativas na estimación non-paramétrica de curvas. Este capítulo presenta unha nova técnica nonparamétrica para a análise exploratoria de datos circulares, denominada CircSiZer, que permite coñecer que características observadas na curva suavizada (densidade ou regresión), son estatísticamente significativas e cales se poden atribuír á variabilidade dos datos.

O CircSiZer (Oliveira et al., 2013a) é unha adaptación a datos circulares do método SiZer proposto orixinalmente por Chaudhuri e Marron (1999). A idea dos métodos SiZer é proporcionar

unha ferramenta gráfica que amose os patróns de crecemento e decrecemento significativo da curva suavizada. Ademais, as técnicas SiZer consideran un amplo rango de parámetros de suavizado evitando así o problema da súa selección. Deste xeito, para cada valor do parámetro de suavizado, abordan a cuestión de que características, tales como picos e vales, están realmente presente nos datos, é dicir, cales son estatisticamente significativas e cales son debidas á variabilidade mostral. A metodoloxía CircSiZer aborda esta cuestión construindo intervalos de confianza para a derivada da curva suavizada $f'(\theta;\tau) \equiv \mathbb{E}(\hat{f}'(\theta;\tau))$, onde f representa aquí a función de densidade ou de regresión segundo o contexto e τ representa o parámetro de suavizado ($\tau \equiv \nu$ para os estimadores tipo núcleo e $\tau \equiv \lambda$ para os estimadores tipo spline).

Así, dado un par (θ,τ) con $\theta \in [0,2\pi)$ e $\tau > 0$, a curva suavizada $f(\theta;\tau)$ é significativamente crecente (decrecente) se o intervalo de confianza está por enriba (debaixo) de cero e, se o intervalo de confianza contén ao cero, non se pode determinar crecemento nin decrecemento da curva. A información porporcionada polos intervalos de confianza amósase nun mapa circular de tal forma que, para cada nivel de suavizado τ, o comportamento da curva suavizada no intervalo $[0,2\pi)$ se representa por un anel de cores, onde as diferentes cores permiten identificar as zonas de crecemento e decrecemento da mesma.

Os intervalos de confianza para $f'(\theta;\tau)$ defínense da forma

$$\left(\hat{f}'(\theta;\tau) - q^{(1-\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)), \hat{f}'(\theta;\tau) - q^{(\alpha/2)} \cdot \widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))\right),$$

onde $\hat{f}'(\theta;\nu)$ é a derivada do estimador da función de densidade ou da regresión, $q^{(1-\alpha/2)}$ e $q^{(\alpha/2)}$ son cuantiles e $\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau))$ é un estimador da desviación típica de $\hat{f}'(\theta;\tau)$.

Estúdanse catro alternativas para o cálculo dos cuantiles $q^{(1-\alpha/2)}$ e $q^{(\alpha/2)}$, dúas delas baseadas na aproximación normal (véxase Chaudhuri e Marron, 1999):

(i) $q^{(1-\alpha/2)}$ e $q^{(\alpha/2)}$ son os cuantiles de orde $(1-\alpha/2)$ e $\alpha/2$ da distribución normal estándar.

(ii) $q^{(1-\alpha/2)} = -q^{(\alpha/2)} = \Phi^{-1}\left\{\frac{1+(1-\alpha)^{1/m(\tau)}}{2}\right\}$ onde Φ^{-1} é a inversa da función de distribución da normal estándar e $m(\tau) = n/\mathrm{avg}_{\theta\in\mathcal{D}_\tau} ESS(\theta;\tau)$ sendo $ESS(\theta;\tau)$ o tamaño mostral efectivo definido para cada $(\theta;\tau)$ e $\mathcal{D}_\tau = \{\theta : ESS(\theta;\tau) \geq 5\}$.

E outras dúas baseadas en técnicas bootstrap:

(iii) $q^{(1-\alpha/2)}$ e $q^{(\alpha/2)}$ son os cuantiles mostrais de orde $(1-\alpha/2)$ e $\alpha/2$ de $Z_1^*(\theta;\tau),\ldots,Z_B^*(\theta;\tau)$ onde

$$Z_b^*(\theta;\tau) = \frac{\hat{f}'(\theta;\tau)^{*b} - \hat{f}'(\theta;\tau)}{\widehat{\mathrm{sd}}(\hat{f}'(\theta;\tau)^{*b})},\ b = 1,\ldots,B$$

é a versión estandarizada do estimador da derivada calculado para a b–ésima mostra bootstrap extraída dos datos con reemprazamento.

(iv) $q^{(1-\alpha/2)}$ é o cuantil mostral de orde $(1-\alpha/2)$ de $Z_{sup}^{*1},\ldots,Z_{sup}^{*B}$ e $q^{(\alpha/2)}$ é o cuantil mostral de orde $\alpha/2$ de $Z_{inf}^{*1},\ldots,Z_{inf}^{*B}$ onde

$$Z_{inf}^{*b} = \inf_{\theta \in \mathcal{D}_\tau} Z_b^*(\theta, \tau) \text{ e } Z_{sup}^{*b} = \sup_{\theta \in \mathcal{D}_\tau} Z_b^*(\theta, \tau), \quad b = 1, \ldots, B$$

Mentres que (i) e (iii) proporcionan intervalos de confianza puntuais, (ii) e (iv) proporcionan intervalos de confianza simultáneos.

Para o cálculo dos intervalos de confianza tamén é preciso estimar a desviación típica de $\hat{f}'(\theta;\tau)$. No contexto da densidade, a varianza pode estimarse mediante

$$\widehat{\text{var}}\left(\hat{f}'(\theta;\tau)\right) = n^{-1}s^2\left(K'_\nu(\theta - \Theta_1), \ldots, K'_\nu(\theta - \Theta_n)\right), \quad 0 \leq \theta < 2\pi,$$

onde s^2 denota a varianza mostral de n datos. No contexto da regresión, para a estimación da desviación típica faise uso de que os estimadores propostos son estimadores lineales e polo tanto o valor de $\hat{f}'(\theta;\tau)$ nunha grella $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_2)$, se pode escribir como $\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}} = H\boldsymbol{Y}$ onde H é unha matriz de coeficientes de dimensión $(N \times n)$ e $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t$ é o vector de respostas. Así, para deseño fixo,

$$\text{var}(\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}}) = H\Sigma H^t,$$

onde $\Sigma = \text{diag}\left\{\sigma^2(\Theta_1), \sigma^2(\Theta_2), \ldots, \sigma^2(\Theta_n)\right\}$. No caso de que se asuma homocedasticidade, entón $\sigma^2(\Theta_i) = \sigma^2$ $(i = 1, \ldots, n)$ pódese estimar co estimador proposto por Rice (1984). Para deseño aleatorio, $\widehat{\text{sd}}\left(\hat{f}'(\theta;\tau)\right)$ é estimada por bootstrap mediante

$$\widehat{\text{sd}}\left(\hat{f}'(\theta;\tau)\right) = \left[s^2\left(\hat{f}'^{*1}(\theta;\tau), \ldots, \hat{f}'^{*B}(\theta;\tau)\right)\right]^{1/2}.$$

onde $\hat{f}'^{*b}(\theta;\tau)$ denota o valor do estimador para a b–ésima mostra bootsrap, extraída dos datos con reemprazamento.

No contexto da densidade, realízase un estudo de simulación para comparar os intervalos de cofianza puntuais baseados na aproximación normal e en técnicas bootstrap. De dito estudo conclúese que, para valores elevados do parámetro de suavizado, ámbolos dous intervalos tenden a identificar como significativas algunhas características que non o son, sendo este feito máis acusado para os intervalos bootstrap. Para valores pequenos deste parámetro, as dúas aproximacións se comportan de maneira similar, tanto en termos de cobertura como do número de modas detectadas como significativas.

Outro estudo de simulación compara a cobertura dos intervalos de confianza simultáneos baseados na aproximación normal cos intervalos baseados na aproximación bootstrap. Este estudo permite observar que as coberturas empíricas dos intervalos de confianza bootstrap están máis proximas ao nivel nominal $(1 - \alpha)$ que as coberturas dos intervalos de confianza baseados na aproximación normal. Tamén se observa que os intervalos de confianza simultáneos poden ter dificultades para atopar as modas que presenta o modelo se o tamaño mostral non é suficientemente grande. Neste sentido, mediante a aplicación a varios conxuntos de datos simulados, obsérvase que os intervalos de confianza puntuais poden axudar a detectar as modas.

No contexto da regresión, estúdase a cobertura dos intervalos de confianza bootstrap simultáneos no caso de deseño fixo. Os resultados amosan que a cobertura empírica destes intervalos está por debaixo do valor nominal. O mesmo comportamento obsérvase tanto para o estimador Local Lineal como para o estimador spline de suavizado periódico. Para deseño aleatorio, o comportamento do CircSiZer con intervalos de confianza simultáneos baseados en técnicas bootstrap compróbase coa aplicación a conxuntos de datos simulados. Neste caso, os resultados con ambos estimadores tamén son semellantes.

Finalmente, a utilidade práctica da metodoloxía proposta ilústrase coa análise de varios conxuntos de datos reais, cos que se pretende estudar cando se producen certos cambios nas temperaturas, en que direccións se producen fracturas en implantes de cadeira cementados ou cal é a relación entre a dirección e a velocidade do vento.

Capítulo 4: Software: o paquete NPCirc. O derradeiro capítulo adícase á presentación da libraría `NPCirc` para o paquete estatístico R, que implementa os distintos estimadores e métodos descritos nos capítulos anteriores e cuxo obxectivo é proporcionar aos usuarios un amplo conxunto de métodos nonparamétricos para a densidade e a regresión con datos circulares e que ademais, complemente os paquetes xa existentes para a análise deste tipo de datos. En concreto, `NPCirc` contén o estimador tipo núcleo da densidade circular con núcleo von Mises, xunto cos diferentes métodos para elixir o parámetro de suavizado descritos no Capítulo 2. No contexto de regresión, para resposta escalar e covariable circular, a libraría dispón dunha función para a estimación nonparamétrica da función de regresión mediante os estimadores de Nadaraya–Watson e Local Lineal adaptados á natureza circular da covariable, tamén descritos no Capítulo 2. A metodoloxía CircSiZer baseada nos estimadores tipo núcleo introducida no Capítulo 3, tanto en densidade como en regresión, tamén está dispoñible. A libraría tamén inclúe funcións que permiten xerar observacións e calcular a función de densidade de varias distribucións circulares (von Mises, cardioid, wrapped Cauchy, wrapped normal e wrapped skew–normal) ou de mixturas destas distribucións. Ademais, o paquete `NPCirc` tamén recolle os conxuntos de datos que son analizados ao longo do manuscrito.

Apéndice A. Neste apéndice están definidos os vinte modelos de densidades circulares usados no estudo de simulación do Capítulo 2 e ao longo de todo o manuscrito para a ilustración das técnicas.

Apéndices B e C. Neste apéndices inclúense detalles técnicos sobre o estimador Local Lineal e os splines de suavizado periódicos estudados no Capítulo 2.

Apéndice D. Este apéndice describe as funcións da libraría `NPCirc`, detallando o seu uso e argumentos e ilustrando o seu funcionamento con exemplos.

Agradecementos

Quero agradecerlle aos profesores Rosa M. Crujeiras e Alberto Rodríguez Casal o seu apoio e dedicación durante a realización desta tese. Tamén quero darlle as grazas aos profesores Augusto Pérez Alberti e Kenneth A. Mann por facilitar algúns conxuntos de datos que motivan parte do traballo realizado nesta tese.

Este traballo foi financiado polo Ministerio de Ciencia e Innovación, a través do proxecto MTM2008–03010 e pola rede IAP StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences), do goberno belga.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliography

Abe, T. and Pewsey, A. (2011). Symmetric circular models through duplication and cosine perturbation. *Computational Statistics & Data Analysis*, 55:3271–3282.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Aradóttir, A. L., Robertson, A., and Moore, E. (1997). Circular statistical analysis of birch colonization and the directional growth response of birch and black cottonwood in south Iceland. *Agricultural and Forest Meteorology*, 84:179–186.

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

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

Barragán, S. and Fernández, M. A. (2012). *isocir: Isotonic inference for circular data*. R package version 1.1–1. Available from: http://www.cran.r-project.org/package=isocir.

Batschelet, E. (1981). *Circular Statistics in Biology*. Academic Press, New York.

Beran, R. (1979). Exponential models for directional data. *The Annals of Statistics*, 7:1162–1178.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report. International Computer Science Institute. University of Berkley.*

Bowers, J. A., Morton, I. D., and Mould, G. I. (2000). Directional statistics of the wind and waves. *Applied Ocean Research*, 22:13–30.

Brunsdon, C. and Corcoran, J. (2005). Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems*, 30:300–319.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94:807–823.

Cogburn, R. and Davis, H. T. (1974). Periodic splines and spectral estimation. *The Annals of Statistics*, 2:1108–1126.

Corcoran, K., Chhetri, P., and R., S. (2009). Using circular statistics to explore the geography of the journey to work. *Papers in Regional Science*, 88:119–132.

Ćwik, J. and Koronacki, J. (1997). A combined adaptive–mixtures/plug–in estimator of multivariate probability densities. *Computational & Data Analysis*, 26:199–218.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Di Marzio, M., Panzera, A., and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, 79:2066–2075.

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

Di Marzio, M., Panzera, A., and Taylor, C. C. (2012). Non–parametric regression for circular responses. *Scandinavian Journal of Statistics*, 40:238–255.

Downs, T. D. and Mardia, K. V. (2002). Circular regression. *Biometrika*, 89:683–697.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

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

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3:35–56.

Fernández-Durán, J. J. and Gregorio-Domínguez, M. M. (2012). *CircNNTSR: An R package for the statistical analysis of circular data using nonnegative trigonometric sums (NNTS) models*. R package version 2.0. Available from: http://www.cran.r-project.org/package=CircNNTSR.

Fisher, N. I. (1989). Smoothing a sample of circular data. *Journal of Structural Geology*, 11:775–778.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, U.K.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

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

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for sizer. *Journal of the American Statistical Association*, 101:484–499.

Hornik, K. and Grün, B. (2012). *movMF: Mixtures of von Mises–Fisher distributions*. R package version 0.1–0. Available from: http://www.cran.r-project.org/package=movMF.

Jammalamadaka, S. R. and Lund, U. J. (2006). The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics*, 13:287–298.

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

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford.

Jones, M. C. and Pewsey, A. (2012). Inverse Batschelet distributions for circular data. *Biometrics*, 68:183–193.

Kato, S., Shimizu, K., and Shieh, G. (2008). A circular–circular regression model. *Statistica Sinica*, 18:633–645.

Klemelä, J. (2000). Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, 73:18–40.

Lee, A. (2010). Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:477–486.

Li, R. and Marron, J. S. (2005). Local likelihood SiZer map. *Sankyha: The Indian Journal of Statistics*, 67:476–498.

Lund, U. and Agostinelli, R. (2011). *circular: Circular Statistics*. R package version 0.4–3. Available from: http://www.cran.r-project.org/package=circular.

Lund, U. and Agostinelli, R. (2012). *CircStats: Circular Statistics, from "Topics in circular Statistics" (2001)*. R package version 0.2–4. Available from: http://www.cran.r-project.org/package=CircStats.

Mann, K. A., Gupta, S., Race, A., Miller, M. A., and Cleary, R. J. (2003). Application of circular statistics in the study of crack istribution around cemented femoral components. *Journal of Biomechanics*, 36:1231–1234.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press, New York.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, New York.

Marron, J. S. and de Uña Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. *Journal of Statistics and Planning Inference*, 12:149–161.

Marron, J. S. and Zhang, J. T. (2005). SiZer for smoothing splines. *Computational Statistics*, 20:481–502.

Mooney, J. A., Helms, P. J., and Jollife, I. T. (2003). Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, 41:505–513.

Nürnberger, G. (1989). *Approximation by Spline Functions*. Springer–Verlag, Berlin.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2012a). Nonparametric circular density estimation for temperature cycles. In *Proceedings of 27th International Workshop on Statistical Modelling*, pages 257–262, Prague.

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

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013a). CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*. DOI: 10.1007/s10651-013-0249-0.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013b). CircSiZer for exploring circular data. In *Proceedings of 28th International Workshop on Statistical Modelling*, volume 1, pages 313–318, Palermo.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013c). Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, 20:1–17.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013d). NPCirc: an R package for nonparametric circular methods. Manuscript submitted for publication.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013e). *NPCirc: Nonparametric circular methods*. R package version 1.0.0. Available from: http://www.cran.r-project.org/package=NPCirc.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013f). Study of crack distribution around cemented femoral components using nonparametric circular methods. In *Libro de resúmenes de la XIV Conferencia Española de Biometría*, pages 57–60, Ciudad Real.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Pewsey, A. (2000). The wrapped skew–normal distribution on the circle. *Communications in Statistics - Theory and Methods*, 29:2459–2472.

Pewsey, A. (2004). Testing for circular reflective symmetry about a known median axis. *Journal of Applied Statistics*, 31:575–585.

Pewsey, A. (2006). Modelling asymmetrically distributed circular data using the wrapped skew–normal distribution. *Environmental and Ecological Statistics*, 13:257–269.

Presnell, B., Morrison, S. P., and Littel, R. C. (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, 93:1068–1077.

Qin, X., Zhang, J. S., and Yan, X. D. (2011a). Local linear least squares kernel regression for linear and circular predictors. *Communications in Statistics–Theory and Methods*, 40:3812–3823.

Qin, X., Zhang, J. S., and Yan, X. D. (2011b). A nonparametric circular–linear multivariate regression model with a rule–of–thumb bandwidth selector. *Computers & Mathematics with Applications*, 62:3048–3055.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from: http://www.r-project.org/.

Rice, S. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12:1215–1230.

Rondonotti, V., Marron, J. S., and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289.

Rosenblatt, M. (1956). Remarks on some nonparametric estimate of a density function. *The Annals of Mathematical Statistics*, 27:832–837.

Rudge, J. F. (2008). Finding peaks in geochemical distributions: A re-examination of the helium-continental crust correlation. *Earth and Planetary Science Letters*, 274:179–188.

Rydén, J. (2010). Exploring possibly increasing trend of hurricane activity by a SiZer approach. *Environmental and Ecological Statistics*, 17:125–132.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross–validation in density estimation. *Journal of the American Statistical Association*, 89:807–817.

SenGupta, A. and Ugwuowo, F. I. (2006). Asymmetric circular–linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics*, 13:299–309.

SenGupta, S. and Rao, J. S. (1966). Statistical analysis of cross–bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Godavari Valley. *Sankhya: The Indian Journal of Statistics, Series B*, 28:165–174.

Sheather, S. J. (2009). *A modern Approach to Regression with R*. Springer, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Sonderegger, D., Wang, H., William, C., and Noon, B. (2009). Using SiZer to detect thresholds in ecological data. *Frontiers in Ecology and the Environment*, 7:190–195.

Sousa, M. C., Alvarez, I., Vaz, N., Gomez-Gesteira, M., and Dias, J. M. (2013). Assessment of wind pattern accuracy from the QuikSCAT satellite and the WRF model along the Galician coast (Northwest Iberian Peninsula). *Monthly Weather Review*, 141:742–753.

Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika*, 76:705–712.

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

von Mises, R. (1918). über die "ganzzahligkeit" der atomgewichte und verwandte fragen. *Physikal. Z.*, 19:490–500.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial Mathematics, Philadelphia.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall, Boca Raton.