Goodness-of-fit tests for multiple regression with circular response

A. Meilán-Vila[a] and M. Francisco-Fernández[b] and R.M. Crujeiras[c]

[a] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain; [b] Research group MODES, CITIC, Faculty of Computer Science, Department of Mathematics, Universidade da Coruña, A Coruña, Spain; [c] Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

ABSTRACT

Testing procedures for assessing a parametric regression model with a circular response and an $\mathbb{R}^d$-valued covariate are proposed and analyzed in this work. The test statistics are based on a circular distance comparing a (non-smoothed or smoothed) parametric circular regression estimator and a nonparametric one. Two bootstrap procedures for calibrating the tests in practice are also presented. Finite sample performance of the tests in different scenarios is analyzed by simulations and illustrated with real data examples.

KEYWORDS

Model checking; circular data; local linear regression; bootstrap

AMS CLASSIFICATION

62G08; 62G09; 62G10; 62H11

1. Introduction

In many scientific fields, such as oceanography, meteorology or biology, data are angular measurements (sample realizations of a circular variable), which may be accompanied by auxiliary observations of other Euclidean random variables with a possible influence on the circular one. The joint behaviour of these circular and Euclidean variables can be analyzed by considering a linear-circular regression model (circular response and Euclidean covariates), allowing to explain the possible relation between the variables and, at the same time, to make predictions on the variable of interest. In this context, parametric and nonparametric methods can be used to estimate the unknown regression function. Parametric regression approaches were studied, among others, in [1–3]. Alternatively, nonparametric kernel-type estimators of the regression function considering a model with a circular response and a univariate Euclidean covariate were introduced in [4], while the extension to a model with an $\mathbb{R}^d$-valued

CONTACT A. Meilán-Vila. Email: ameilan@est-econ.uc3m.es; andrea.meilan@udc.es. Department of Statistics, Universidad Carlos III de Madrid, Avenida de la Universidad, 30, Leganés, Madrid 28911, Spain; Research group MODES, CITIC, Faculty of Computer Science, Department of Mathematics, Universidade da Coruña, Campus de Elviña s/n, A Coruña 15071, Spain

This is an Accepted Manuscript of an article published by Taylor & Francis in *Journal of Statistical Computation and Simulation,* 92:9, 1941-1963, on 2022, available at:
https://doi.org/10.1080/00949655.2021.2015597

covariate was considered in [5,6]. To compute these kernel-type estimators it is crucial to select a bandwidth parameter (a symmetric $d \times d$ matrix for an $\mathbb{R}^d$-valued covariate) which directly impacts the smoothness of the estimator. If the bandwidth matrix is appropriately chosen, these nonparametric methods provide more flexible and robust estimators than those obtained when using parametric approaches, avoiding misspecification problems. However, if a suitable parametric regression model is assumed, parametric methods usually provide estimators which are more efficient and easier to interpret.

At this point, an important question in this context is to decide if a certain parametric family is appropriate to model the unknown circular regression function. If this assumption holds, a parametric method should be preferably used to estimate it. If not, it would possibly be more convenient to use a nonparametric approach to estimate this function. Both approaches, parametric and nonparametric, have been used to analyze different datasets in the literature. For instance, the classical blue periwinkles dataset, which collects measurements of direction and distance moved by 31 blue periwinkles, was analyzed using parametric methods in [1,2], considering the direction as the response variable and the distance as the covariable. On the other hand, also considering this dataset and this regression model, nonparametric techniques were employed in [4] to estimate the corresponding regression function. Another such example is the sand hopper orientation dataset studied, using parametric methods, in [7]. Following the proposal in [2], these authors considered a projected multivariate linear model (PMLM) to analyze the orientation of two species of sand hoppers as a function of different covariates. This dataset was also explored using nonparametric tools by [5], considering a regression model with a circular response (sand hopper orientation) and two real-valued covariates (temperature and humidity). In this case, the regression function was estimated nonparametrically using a local linear-type estimator. In order to determine if a parametric regression model is a suitable representation of such datasets, goodness-of-fit tests can be designed and analyzed, providing a tool for assessing a general class of parametric linear-circular regression models.

There is a substantial literature on testing parametric regression models involving Euclidean data, for example, [8–14]. See also [15] for a review on this topic. The previous testing procedures are based on measuring differences between a suitable parametric estimator under the null hypothesis and a nonparametric one. Specifically, L_2-norm or supremum-norm tests, among others, can be employed for testing parametric regression models with a Euclidean response and an $\mathbb{R}^d$-valued covariate ($d \geq 1$). In the context of regression models with directional response and directional or Euclidean explanatory variables, the literature on goodness-of-fit tests is relatively scarce. In this setting, in [16], an exploratory tool and a lack-of-fit test for circular-linear regression models (Euclidean response and circular covariates) were proposed. The same problem was studied by [17], using nonparametric methods. The authors proposed a testing procedure based on the weighted squared distance between a nonparametric and a parametric regression estimator, where the nonparametric regression estimator was obtained by a projected local regression on the sphere. Local linear-type estimators have been recently used by [18] in order to propose no-effect and ANCOVA tests for regression models with circular response and/or covariate. However, the problem of assessing a certain class of parametric regression models with circular response and Euclidean covariates (up to the authors' knowledge) has not been considered in the statistical literature yet.

In this work, new approaches for testing a linear-circular parametric regression model are proposed and empirically analyzed. The test statistics employed in these procedures are based on a comparison between a (non-smoothed or smoothed) parametric fit under the null hypothesis and a nonparametric estimator of the circular regression function. More specifically, two different test statistics are considered. In the first one, the parametric estimator of the regression function under the null hypothesis is directly used, while in the second one, a smoothed version of this estimator is employed. Notice that, in this framework, a suitable measure of circular distance must be employed [see 19, Section 1.3.2]. The null hypothesis that the regression function belongs to a certain parametric family is rejected if the distance between both fits exceeds a certain threshold. To perform the parametric estimation, procedures based on least squares or maximum likelihood are used [1,2,20]. For the nonparametric alternative, a local linear-type estimator [4,5] is considered.

For the application in practice of the proposals, the test statistics should be accompanied by a calibration procedure. In this case, this is not based on the asymptotic distribution, given that the convergence to the limit distribution under the null hypothesis will presumably be too slow. Instead, two bootstrap methods are designed and their performance is analyzed and compared employing numerical experiments. Standard resampling procedures adapted to the context of regression models with circular response and Euclidean covariates are used: a parametric circular residual bootstrap (PCB) and a nonparametric circular residual bootstrap (NPCB). The PCB approach consists in using the residuals obtained from the parametric fit in the bootstrap algorithm. If the circular regression function belongs to the parametric family considered in the null hypothesis, then the residuals will tend to be quite similar to the theoretical errors and, therefore, it is expected that the PCB method has a good performance. Following the proposal by [21], the NPCB method aims to increase the power of the test and, for this purpose, the residuals obtained from the nonparametric fit are the ones employed in the bootstrap procedure.

This paper is organized as follows. Section 2 is devoted to present the assumed linear-circular regression model, to introduce the testing problem and to briefly describe the parametric and nonparametric circular regression estimators (in Sections 2.1 and 2.2, respectively) employed in the test statistics. In Section 3, the proposed test statistics are explained. A description of the bootstrap calibration algorithms considered is given in Section 4. Section 5 contains a simulation study for assessing the performance of the tests when using the PCB and NPCB resampling approaches to approximate the sampling distribution of the test statistics. Section 6 illustrates the testing proposals with the blue periwinkle and sand hopper orientation datasets introduced above. Finally, some conclusions and ideas for further research are provided in Section 7.

2. Statistical model

Let $\{(\mathbf{X}_i, \Theta_i)\}_{i=1}^n$ be a random sample from $(\mathbf{X}, \Theta)$, where Θ is a circular random variable taking values on $\mathbb{T} = [0, 2\pi)$, and $\mathbf{X}$ is a random variable with density f and support on $\mathcal{D} \subseteq \mathbb{R}^d$. Assume that the following regression model holds:

$$\Theta_i = [m(\mathbf{X}_i) + \varepsilon_i](\texttt{mod}\,2\pi), \quad i = 1, \ldots, n, \tag{1}$$

where m is a circular regression function, $\varepsilon_i, i = 1, \ldots, n$, is an independent sample of a circular variable ε, with zero mean direction (which is equivalent to assume that $\mathbb{E}[\sin(\varepsilon) \mid \mathbf{X} = \mathbf{x}] = 0$) and finite concentration, and `mod` stands for the modulo operation.

Considering the regression model (1), the aim of this work is to propose and study different testing procedures to assess the suitability of a general class of regression models $\mathcal{M}_{\boldsymbol{\beta}} = \{m_{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}\}$, where $m_{\boldsymbol{\beta}}$ is a certain parametric circular regression function with parameter vector $\boldsymbol{\beta}$. The specific testing problem to be addressed is formulated as:

$$H_0 : m \in \mathcal{M}_{\boldsymbol{\beta}} \qquad \text{vs.} \qquad H_a : m \notin \mathcal{M}_{\boldsymbol{\beta}}. \tag{2}$$

As pointed out in Section 1, the procedure proposed in this work consists in comparing a (non-smoothed or smoothed) parametric fit with a nonparametric estimator of the circular regression function m, measuring the circular distance between both fits and employing this distance as a test statistic. The parametric and nonparametric estimation methods considered in this proposal are described in the following sections.

2.1. Parametric circular regression estimation

As mentioned in Section 1, our proposal requires a parametric estimator of the circular regression function m, once a parametric regression family is set as the null hypothesis. Notice that, for instance, the procedures based on least squares for Euclidean data are not appropriate when the response variable is of circular nature. Minimizing the sum of squared differences between the observed and predicted values may lead to erroneous results, since the squared difference is not an appropriate measure on the circle.

A circular analog to least squares regression for models with a circular response and a set of Euclidean covariates was presented by [20]. Specifically, assuming that the regression model (1) holds and $m \in \mathcal{M}_{\boldsymbol{\beta}}$, a parametric estimator of $m_{\boldsymbol{\beta}}$ is constructed obtaining an estimator of $\boldsymbol{\beta}$, namely $\hat{\boldsymbol{\beta}}$, and computing $m_{\hat{\boldsymbol{\beta}}}$. A parameter estimate of $\boldsymbol{\beta}$ could be obtained by minimizing the sum of the circular distances between the observed and predicted values as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{1 - \cos\left[\Theta_i - m_{\boldsymbol{\beta}}(\mathbf{X}_i)\right]\right\}. \tag{3}$$

Note that the previous estimation proposal does not require any assumption on the conditional distribution of the response over the covariate. Nevertheless, the assumption of a conditional parametric distribution model facilitates the use of maximum likelihood estimation methods. Specifically, if it is assumed that the response variable (conditionally on $\mathbf{X}$) follows a von Mises distribution with mean direction given by $m_{\boldsymbol{\beta}}$ and concentration parameter κ, the maximum likelihood estimator of $m_{\boldsymbol{\beta}}$ maximizes the following expression:

$$\sum_{i=1}^{n} \cos\left[\Theta_i - m_{\boldsymbol{\beta}}(\mathbf{X}_i)\right]. \tag{4}$$

Given the maximum likelihood estimator of $\boldsymbol{\beta}$, the maximum likelihood estimator of κ is given by the solution to $A(\hat{\kappa}) = \frac{1}{n}\sum_{i=1}^{n} \cos[\Theta_i - m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_i)]$, where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, being I_0 and I_1 the modified Bessel functions of the first kind with order zero and one, respectively. As indicated in [20], numerical solutions to $A^{-1}(x)$ can be found in [22].

Notice that the circular least squares estimator given in (3) also maximizes the expression (4) and, therefore, assuming a von Mises distribution, the circular least squares estimator coincides with the maximum likelihood estimator (for further details, see [20]).

Assuming that the response variable follows a von Mises distribution and considering the general class of models for the circular regression function $\mathcal{M}_{\boldsymbol{\beta}} = \{\mu_0 + g(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}), \mu_0 \in [0, 2\pi), \boldsymbol{\beta}_1 \in \mathbb{R}^d\}$, where g is a link function mapping the real line onto the circle, an iteratively reweighted least squares algorithm can be used to compute the maximum likelihood estimators of κ, μ_0 and $\boldsymbol{\beta}_1$ [see 1,20]. The extension of these results to the case of a generic parametric family has not been explicitly considered.

Although the assumption that the response variable follows a von Mises distribution is quite common, other circular distributions can be used in this context. For example, considering a projected normal distribution allows to define general regression models, such as the PMLM [2,7]. This class of models deals with directional observations as projections onto the unit circle of unobserved response vectors in a multivariate linear model. Considering these type of regression models, the estimation of the parameters can be performed using maximum likelihood methods employing iterative procedures. For further details on the estimation approach in this case, we refer to [2].

2.2. Nonparametric circular regression estimation

A nonparametric regression estimator for m in model (1) is presented in this section. The circular regression function m, at a point $\mathbf{x}$, is the conditional mean direction of Θ given $\mathbf{X} = \mathbf{x}$, which can be defined as the minimizer of the risk function $\mathbb{E}\{1 - \cos[\Theta - m(\mathbf{X})] \mid \mathbf{X} = \mathbf{x}\}$. The minimizer of this cosine risk is given by $m(\mathbf{x}) = \operatorname{atan2}[m_1(\mathbf{x}), m_2(\mathbf{x})]$, where $m_1(\mathbf{x}) = \mathbb{E}[\sin(\Theta) \mid \mathbf{X} = \mathbf{x}]$, $m_2(\mathbf{x}) = \mathbb{E}[\cos(\Theta) \mid \mathbf{X} = \mathbf{x}]$, and the function $\operatorname{atan2}(y, x)$ returns the angle between the x-axis and the vector from the origin to (x, y). Therefore, replacing m_1 and m_2 by appropriate estimators, an estimator for m can be directly obtained. In this work, a local linear-type estimator of $m(\mathbf{x})$ is used, defined by considering local linear estimators for $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$. Specifically, the estimator

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \operatorname{atan2}[\hat{m}_{1,\mathbf{H}}(\mathbf{x}), \hat{m}_{2,\mathbf{H}}(\mathbf{x})] \tag{5}$$

is considered, where $\hat{m}_{1,\mathbf{H}}(\mathbf{x})$ and $\hat{m}_{2,\mathbf{H}}(\mathbf{x})$ denote the local linear estimators [23] (with bandwidth matrix $\mathbf{H}$) of $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$, respectively, defined by:

$$\hat{m}_{j,\mathbf{H}}(\mathbf{x}) = \begin{cases} \mathbf{e}_1{}^{\mathrm{T}}(\boldsymbol{\mathcal{X}}_{\mathbf{x}}{}^{\mathrm{T}}\boldsymbol{\mathcal{W}}_{\mathbf{x}}\boldsymbol{\mathcal{X}}_{\mathbf{x}})^{-1}\boldsymbol{\mathcal{X}}_{\mathbf{x}}{}^{\mathrm{T}}\boldsymbol{\mathcal{W}}_{\mathbf{x}}\boldsymbol{\mathcal{S}} & \text{if } j = 1, \\ \\ \mathbf{e}_1{}^{\mathrm{T}}(\boldsymbol{\mathcal{X}}_{\mathbf{x}}{}^{\mathrm{T}}\boldsymbol{\mathcal{W}}_{\mathbf{x}}\boldsymbol{\mathcal{X}}_{\mathbf{x}})^{-1}\boldsymbol{\mathcal{X}}_{\mathbf{x}}{}^{\mathrm{T}}\boldsymbol{\mathcal{W}}_{\mathbf{x}}\boldsymbol{\mathcal{C}} & \text{if } j = 2, \end{cases} \tag{6}$$

where $\mathbf{e}_1$ is a $(d + 1) \times 1$ vector having 1 in the first entry and 0 in all other entries, $\boldsymbol{\mathcal{X}}_{\mathbf{x}}$ is a $n \times (d + 1)$ matrix having $[1, (\mathbf{X}_i - \mathbf{x})^{\mathrm{T}}]$ as its ith row, $\boldsymbol{\mathcal{W}}_{\mathbf{x}} = \operatorname{diag}[K_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{x}), \ldots, K_{\mathbf{H}}(\mathbf{X}_n - \mathbf{x})]$, $\boldsymbol{\mathcal{S}} = [\sin(\Theta_1), \ldots, \sin(\Theta_n)]^{\mathrm{T}}$ and $\boldsymbol{\mathcal{C}} = [\cos(\Theta_1), \ldots, \cos(\Theta_n)]^{\mathrm{T}}$.

Asymptotic properties of estimator (5), considering regression model (1), were studied by [5].

3. The test statistics

In this section, two tests statistics to address the testing problem (2) (that is, to check if the circular regression function belongs to a general class of parametric models) are proposed. The first approach considers a weighted circular distance between the nonparametric and parametric fits:

$$T_n^1 = \int_{\mathcal{D}} \{1 - \cos[\hat{m}_{\mathbf{H}}(\mathbf{x}) - m_{\hat{\boldsymbol{\beta}}}(\mathbf{x})]\} w(\mathbf{x}) d\mathbf{x}, \tag{7}$$

where w is a weight function that helps in mitigating possible boundary effects. The estimator $\hat{m}_{\mathbf{H}}$ is the local linear-type estimator of the circular regression function m, given in (5). As for the parametric estimator $m_{\hat{\boldsymbol{\beta}}}$, in a general setting, the least squares approach described in Section 2.1 can be used to compute (7). As previously mentioned, if a parametric (conditional) distribution model is assumed (e.g. a von Mises model), parametric estimation by maximum likelihood methods is also feasible.

The second test statistic is similar to the first one, but considering a smoothed version of the parametric fit:

$$T_n^2 = \int_{\mathcal{D}} \{1 - \cos[\hat{m}_{\mathbf{H}}(\mathbf{x}) - \hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x})]\} w(\mathbf{x}) d\mathbf{x}, \tag{8}$$

where $\hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}$ is a smoothed version of the parametric estimator $m_{\hat{\boldsymbol{\beta}}}$, which is given by:

$$\hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x}) = \operatorname{atan2}[\hat{m}_{1,\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x}), \hat{m}_{2,\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x})], \tag{9}$$

with

$$\hat{m}_{j,\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x}) = \begin{cases} \mathbf{e}_1^{\mathrm{T}} (\boldsymbol{\mathcal{X}}_{\mathbf{x}}^{\mathrm{T}} \boldsymbol{\mathcal{W}}_{\mathbf{x}} \boldsymbol{\mathcal{X}}_{\mathbf{x}})^{-1} \boldsymbol{\mathcal{X}}_{\mathbf{x}}^{\mathrm{T}} \boldsymbol{\mathcal{W}}_{\mathbf{x}} \hat{\mathbf{S}} & \text{if } j = 1, \\ \mathbf{e}_1^{\mathrm{T}} (\boldsymbol{\mathcal{X}}_{\mathbf{x}}^{\mathrm{T}} \boldsymbol{\mathcal{W}}_{\mathbf{x}} \boldsymbol{\mathcal{X}}_{\mathbf{x}})^{-1} \boldsymbol{\mathcal{X}}_{\mathbf{x}}^{\mathrm{T}} \boldsymbol{\mathcal{W}}_{\mathbf{x}} \hat{\mathbf{C}} & \text{if } j = 2, \end{cases}$$

where

$$\hat{\mathbf{S}} = \{\sin[m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_1)], \ldots, \sin[m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_n)]\}^{\mathrm{T}}$$

and

$$\hat{\mathbf{C}} = \{\cos[m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_1)], \ldots, \cos[m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_n)]\}^{\mathrm{T}}.$$

If the null hypothesis in the testing problem given in (2) holds, then the (non-smoothed or smoothed) parametric fit and the nonparametric circular regression estimator will be similar and, therefore, the value of the test statistics T_n^1 and T_n^2 will be relatively small. Conversely, if the null hypothesis does not hold, the fits will be different and the value of T_n^1 and T_n^2 will be fairly large. So, the null hypothesis will be rejected if the circular distance between both fits exceeds a critical value. Then, to apply these procedures, it is essential to approximate the distribution of the test statistics under the null hypothesis. To tackle this problem, we use bootstrap resampling methods.

Illustration of the tests

For a visual illustration of the performance of the tests (for simplicity, a model with a single covariate, that is, $d = 1$, is initially employed), consider an equally-spaced sample of size $n = 200$, generated in the unit interval following model (1), with regression function (13) and $c = 0$. The random errors ε_i are drawn from a von Mises distribution $vM(0, 10)$. If we want to test if $m(X) \in \mathcal{M}_{1,\boldsymbol{\beta}} = \{\mu_0 + 2\text{atan}(\beta_1 X), \mu_0 \in [0, 2\pi), \beta_1 \in \mathbb{R}\}$ using the test statistics given in (7) and in (8), the local linear estimator, $\hat{m}_h$, given in (5), as well as a parametric fit, $m_{\hat{\boldsymbol{\beta}}}$, and its smoothed version, $\hat{m}_{h,\hat{\boldsymbol{\beta}}}$ (denoting by h the bandwidth parameter when $d = 1$) must be computed. In this case, the estimator obtained from (4) is considered for the parametric fit. The local linear-type estimator and $\hat{m}_{h,\hat{\boldsymbol{\beta}}}$ are computed using a triweight kernel and the optimal bandwidth obtained by minimizing the circular average squared error (CASE), defined as:

$$\text{CASE}[\hat{m}_{\mathbf{H}}(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}\{1 - \cos[m(\mathbf{X}_i) - \hat{m}_{\mathbf{H}}(\mathbf{X}_i)]\}, \tag{10}$$

for $d = 1$ (in this case, $\mathbf{H} = h$ is a real value). Figure 1 shows the linear (top panels) and cylinder (bottom panels) representations of the estimates. The local linear-type regression estimator (left panel), the parametric fit (center panel) and the smoothed version of the parametric fit (right panel) are represented with red lines. The sample points and the circular regression function (black lines) are also included in the plots. All estimates show a very similar behaviour and, therefore, the value of the test statistics T_n^1 and T_n^2 are expected to be presumably small. Consequently, there may be no evidences against the assumption that the circular regression function belongs to the parametric family $\mathcal{M}_{1,\boldsymbol{\beta}}$.

[Figure 1 about here.]

A similar visual experiment considering a regression model with a circular response and two covariates is also presented. A sample of size $n = 400$ is generated on a bidimensional regular grid in the unit square, assuming the linear-circular regression model (1), with regression function (15) and $c = 0$. The random errors ε_i are also drawn from a von Mises distribution $vM(0, 10)$. In this case, as in the previous example, in order to test if $m(\mathbf{X}) \in \mathcal{M}_{2,\boldsymbol{\beta}} = \{\mu_0 + 2\text{atan}(\beta_1 X_1 + \beta_2 X_2), \mu_0 \in [0, 2\pi), \beta_1, \beta_2 \in \mathbb{R}\}$, being $\mathbf{X} = (X_1, X_2)$, using the test statistics given in (7) and in (8), the estimator obtained from (4) is employed for the parametric fit. The local linear-type estimator and the smoothed parametric fit are computed using a multiplicative triweight kernel and an optimal bandwidth obtained by minimizing the CASE, given in (10). Figure 2 shows the theoretical circular regression function (top left panel), the local linear-type regression estimator (top right panel), the parametric fit (bottom left panel) and the smoothed version of the parametric fit (bottom right panel). It can be observed that estimates at top right, bottom left and bottom right panels seem to be very similar and, therefore, analogous conclusions to those given for $d = 1$, but in this case for $d = 2$, can be derived.

[Figure 2 about here.]

Bandwidth selection

The test statistics given in (7) and in (8), respectively, require a $d \times d$ bandwidth matrix $\mathbf{H}$ (or a bandwidth parameter h, if $d = 1$). The selection of the bandwidth in this type of goodness-of-fit problems is non-trivial, since the optimal bandwidth for estimation may not be the optimal one for testing (being not even clear what optimal means). For instance, [24–26] gave some strategies on bandwidth selection in testing problems. As usual in the context of smooth-based goodness-of-fit tests for regression models, the performance of the proposed test statistics is analyzed for a range of bandwidths, in order to evaluate the impact of this parameter in the numerical results.

Weight function

In both tests statistics, the weight function w is used to avoid possible boundary effects that are (or could be) due to the use of a nonparametric estimator. An automatic choice of w can be complex, since that election could depend, among other things, on the bandwidth matrix $\mathbf{H}$. For that reason, ad hoc elections of the weight functions are usually employed.

In this paper, we consider a general class of weight functions trying to get a compromise between (partially) removing the boundary effect and simplicity. Specifically, we use weight functions depending on the sample size n. This is reasonable since the boundary regions depend on the bandwidth employed and an optimal selection of this parameter is also related to the sample size. This type of weight functions are quite common in this and other contexts (for example, in bandwidth selection problems). In Sections 5 and 6, the explicit expressions of the weight functions we used in practice are given. After some empirical tests, we observed that they provided good results in our numerical studies.

4. Calibration in practice

Once a suitable test statistic is available, in order to solve the testing problem (2), a procedure for calibration of critical values is required. This task can be done by means of bootstrap resampling algorithms.

In what follows, a description of two different bootstrap proposals (PCB and NPCB) designed to approximate the distribution (under the null hypothesis) of the tests statistics, given in (7) and in (8), are presented. The main difference between them is the mechanism employed to obtain the residuals. As pointed out in Section 1, the residuals used in PCB come from the parametric regression estimator. On the other hand, in the NPCB algorithm, the residuals employed in the resampling process are obtained from the nonparametric regression estimator. In order to present the PCB and NPCB resampling methods, a generic bootstrap algorithm is described. No matter the method used, $\hat{m}$ denotes the parametric or the nonparametric circular regression estimator.

Algorithm 1

1. Compute the parametric or the nonparametric regression estimates (described in Sections 2.1 and 2.2, respectively), namely $\hat{m}(\mathbf{X}_i)$, $i = 1, \ldots, n$, depending on if a parametric (PCB) or a nonparametric (NPCB) bootstrap procedure is employed.
2. From the residuals $\hat{\varepsilon}_i = [\Theta_i - \hat{m}(\mathbf{X}_i)](\texttt{mod}\, 2\pi)$, $i = 1, \ldots, n$, draw independent bootstrap residuals, $\hat{\varepsilon}_i^*$, $i = 1, \ldots, n$, sampling with replacement from $(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)$. Then, for each $i = 1, \ldots, n$, $P(\hat{\varepsilon}_i^* = \hat{\varepsilon}_j) = 1/n$, $j = 1, \ldots, n$
3. Obtain bootstrap samples $\{(\mathbf{X}_i, \Theta_i^*)\}_{i=1}^n$ with $\Theta_i^* = [m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_i) + \hat{\varepsilon}_i^*](\texttt{mod}\, 2\pi)$, being $m_{\hat{\boldsymbol{\beta}}}(\mathbf{X}_i)$ the parametric regression estimator under H_0.
4. Using the bootstrap sample $\{(\mathbf{X}_i, \Theta_i^*)\}_{i=1}^n$, the bootstrap test statistics T_n^{l*}, with $l = 1, 2$, are computed as in (7) and in (8).
5. Repeat Steps 2-4 a large number of times B.

In Step 1 of the previous algorithm, in the PCB approach, the circular regression function is estimated parametrically, employing one of the procedures described in Section 2.1. Alternatively, the NPCB tries to avoid possible misspecification problems by using more flexible regression estimation methods than those employed in PCB. Then, following the same arguments as in [21] to increase the power of the test, in the NPCB method, the nonparametric circular regression estimator given in (5) is employed in Step 1 of the bootstrap Algorithm 1.

Notice that the empirical distribution of the B bootstrap test statistics can be employed to approximate the distribution of the test statistics T_n^l, for $l = 1, 2$, under the null hypothesis. Denoting by $\{T_{n,1}^{l*}, \ldots, T_{n,B}^{l*}\}$ the sample of the B bootstrap test statistics given in (7), for $l = 1$, and in (8), for $l = 2$, and defining its $(1 - \alpha)$-quantile t_α^{l*}, the null hypothesis in (2) will be rejected if $T_n^l > t_\alpha^{l*}$ (for $l = 1, 2$). Additionally, the p-values of the test statistics can be approximated by:

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}_{\{T_{n,b}^{l*} > T_n^l\}}, \quad l = 1, 2, \tag{11}$$

where $\mathbb{I}_{\{A\}} = 1$ if A is true and 0 otherwise.

5. Simulation study

The finite sample performance of the proposed tests, using the bootstrap approaches described in Algorithm 1 for their calibration, is illustrated in this section with a simulation study, considering a regression model with a single real-valued covariate and also with a bidimensional one. The R code to compute the estimates in this section and the next one are provided in the Supplementary Material.

5.1. Simulation experiment with a single covariate

In order to study empirically the performance of the proposed tests considering a regression model with a circular response and a single real-valued covariate, the parametric regression family

$$\mathcal{M}_{1,\boldsymbol{\beta}} = \{\mu_0 + 2\text{atan}(\beta_1 X), \mu_0 \in [0, 2\pi), \beta_1 \in \mathbb{R}\} \tag{12}$$

is chosen, and for different values of c the regression function

$$m(X) = 2\text{atan}(X) + c\,\text{asin}(2X^5 - 1) \tag{13}$$

is considered. Therefore, the parameter c controls whether the null ($c = 0$) or the alternative ($c \neq 0$) hypotheses hold in problem (2). Values $c = 0$, 1, and 2 are considered in the study. For each value of c, 500 samples of sizes $n = 50, 100$ and 200 are generated on the unit interval, considering an equally-spaced explanatory variable X, following model (1) with regression function (13). The circular errors ε_i are drawn independently from a von Mises distribution $vM(0, \kappa)$, for different values of κ (5, 10 and 15).

To analyze the behaviour of the test statistics, given in (7) and in (8), in the different scenarios, the bootstrap procedures described in Section 4 are applied, using $B = 500$ replications. The non-smoothed or smoothed parametric fits used for constructing (7) and (8) are computed using the estimators obtained from (4) and given in (9), respectively. The nonparametric fit is obtained using the estimator given in (5) with a triweight kernel (empirical experiments with the Epanechnikov kernel yield similar results). We address the bandwidth selection problem by using the same procedure as the one used in [8,9,13,27], among others, applying the tests on a grid of several bandwidths. In order to use a reasonable grid of bandwidths, the optimal bandwidth selected by minimizing the CASE given in (10), for $d = 1$, is calculated for each sample and for each scenario. In this case, the average of the CASE optimal bandwidths are in the interval $[0.2, 0.6]$. Therefore, the values of the bandwidth parameter $h = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65$ are considered to compute both test statistics (7) and (8). The weight function used in both tests is $w(x) = \mathbb{I}_{\{x \in [1/\sqrt{n}, 1-1/\sqrt{n}]\}}$, to avoid possible boundary effects.

5.1.1. *Effect of sample size*

Proportions of rejections of the null hypothesis, for a significance level $\alpha = 0.05$, considering $\kappa = 10$ and different sample sizes are shown in Table 1, when using T_n^1 and T_n^2. If $c = 0$ (null hypothesis), the proportions of rejections are similar to the theoretical level, although they are quite affected by the value of h. The tests preserve the nominal significance level of 5%, since for appropriate values of h, the majority of proportions of rejections under the null hypothesis lie within the intervals $(0, 0.110)$, $(0.007, 0.093)$ and $(0.020, 0.080)$, when $n = 50, 100$ and 200, respectively. For alternative assumptions ($c = 1$ and $c = 2$), as expected, as the sample size increases the proportions of rejections are larger. For all the scenarios, a larger power of both tests is observed as the value of c increases. Notice that, in most of the cases, an increasing power of the tests when the values of h decrease is observed. It should be noted that the NPCB method presents a slightly better performance than the PCB approach. On the other hand, although both test statistics provide a similar behaviour, T_n^2 seems to give slightly better results.

[Table 1 about here.]

5.1.2. *Effect of κ*

The performance of the tests T_n^1 and T_n^2 (for $\alpha = 0.05$) is studied for $n = 200$ and for different values of the concentration parameter κ. Results are included in Table 2. If $c = 0$, the proportions of rejections are similar to the theoretical level when using both bootstrap approaches (PCB and NPCB). For alternative assumptions, as expected, large values of the concentration parameter κ lead to an increase in power, which justifies the correct performance of the bootstrap procedures.

[Table 2 about here.]

5.2. Simulation experiment with several covariates

The extension for regression models with a circular response and two covariates is analyzed in this section. For this purpose, the parametric regression family

$$\mathcal{M}_{2,\boldsymbol{\beta}} = \{\mu_0 + 2\text{atan}(\beta_1 X_1 + \beta_2 X_2), \mu_0 \in [0, 2\pi), \beta_1, \beta_2 \in \mathbb{R}\} \tag{14}$$

is chosen, and for different values of c the regression function

$$m(\mathbf{X}) = 2\text{atan}(-X_1 + X_2) + c\,\text{asin}(2X_1^3 - 1), \tag{15}$$

being $\mathbf{X} = (X_1, X_2)$, is considered. This circular regression function is plotted in Figure 2 (top left panel) considering $c = 0$. For each value of c ($c = 0$, 1, and 2), 500 samples of sizes $n = 100, 225$ and 400 are generated on a bidimensional regular grid in the unit square, following model (1), with regression function (15) and circular errors ε_i drawn from a von Mises distribution $vM(0, \kappa)$, for $\kappa = 5, 10$ and 15. The bootstrap procedures described in Section 4 are applied, using $B = 500$ replications. The non-smoothed or smoothed parametric fits used for constructing (7) and (8) are computed using the estimators obtained from (4) and given in (9), respectively. The nonparametric fit is obtained using the estimator given in (5) with a multiplicative triweight kernel. In order to simplify the calculations, the bandwidth matrix is restricted to a class of diagonal matrices with equal elements. In this case, the diagonal elements of the CASE optimal bandwidths are in the interval $[0.3, 0.8]$. Therefore, diagonal bandwidth matrices $\mathbf{H} = \text{diag}(h, h)$ with different values of h, $h = 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85$, are considered to compute both test statistics (7) and (8). In this case, the weight function used in both tests is $w(\mathbf{x}) = \mathbb{I}_{\{\mathbf{x} \in [1/\sqrt{n}, 1-1/\sqrt{n}] \times [1/\sqrt{n}, 1-1/\sqrt{n}]\}}$.

5.2.1. Effect of sample size

Proportions of rejections of the null hypothesis, for a significance level $\alpha = 0.05$, considering $\kappa = 10$ and different sample sizes are shown in Table 3, when using T_n^1 and T_n^2. It can be observed that using both bootstrap methods (PCB and NPCB), the tests have a reasonable behaviour. If $c = 0$ (null hypothesis), the tests preserve the nominal significance level, since most of the proportions of rejections lie within the intervals $(0.007, 0.093)$, $(0.022, 0.078)$ and $(0.029, 0.071)$, when $n = 100, 225$ and 400, respectively. For alternative assumptions ($c = 1$ and $c = 2$), the NPCB method presents a slightly better performance than the PCB approach. Notice that, in most of the cases, an increasing power of the tests when the values of h increase is observed. Additionally, as in the single covariate case, for all the scenarios a larger power of both tests is obtained as the value of c increases.

[Table 3 about here.]

5.2.2. Effect of κ

The performance of the bootstrap procedures is analyzed for $n = 400$ and for different values of the concentration parameter κ when using T_n^1 and T_n^2, for $\alpha = 0.05$, in Table 4. If $c = 0$, the proportions of rejections are similar to the theoretical level when using both bootstrap approaches (PCB and NPCB). It can be observed that for larger values of the concentration parameter κ, the bandwidth values providing an effective calibration must be smaller. For alternative assumptions, if the value of the concentration parameter κ is larger, an increasing power is obtained.

[Table 4 about here.]

6. Real data examples

The datasets briefly mentioned in Section 1 are used to illustrate the performance in practice of the test statistics T_n^1 and T_n^2, given in (7) and in (8), respectively. Considering the regression model (1) with a single real-valued covariate, the testing procedure is applied to the blue periwinkle dataset. For a bidimensional real-valued covariate, the sand hopper dataset is employed to illustrate the proposed methodology. Based on the simulation study, where both T_n^1 and T_n^2 presented a very similar behaviour, only the test statistic T_n^2 was employed in these illustrations. Moreover, taking into account that NPCB presented a slightly better performance than the PCB in the simulations, only the NPCB resampling approach was used to calibrate the test.

6.1. Blue periwinkle data

The blue periwinkle dataset, mentioned in the Introduction, is described in this section in more detail to illustrate the application of the proposed goodness-of-fit test T_n^2 for a regression model with a single real-valued covariate. These data can be found in Table 1 of [1], and are available in the `R` package `circular` [28].

Directions and distances moved by small blue periwinkles after they had been transplanted downshore from the height at which they normally live are considered. Figure 3 (left panel) shows the observations of this dataset, which were analyzed and modeled by different authors in the literature. In order to study how orientation varies with distance, in [1], a parametric regression model was fitted, considering that the regression function belongs to the parametric family $\mathcal{M}_{1,\boldsymbol{\beta}}$, given in (12). An iteratively reweighted least squares algorithm to perform the maximum likelihood estimation of the parameters was employed. On the other hand, in [2], a parametric approach was also used to model these data. However, in that case a PMLM was assumed, considering linear models on the covariate (distance) for the means of the bivariate normal distribution that is projected. Notice that if a projected normal distribution with identity covariance matrix is assumed, it holds that $\tan(\mu) = \mu_2/\mu_1$, being μ the circular mean direction, and μ_1 and μ_2 the mean components of the bivariate normal distribution that is projected [2,29]. Therefore, using this approach, the following parametric family is considered:

$$\mathcal{M}_{3,\boldsymbol{\beta}} = \{\text{atan2}(\beta_{0,2} + \beta_{1,2}X, \beta_{0,1} + \beta_{1,1}X), \quad \beta_{0,2}, \beta_{1,2}, \beta_{0,1}, \beta_{1,1} \in \mathbb{R}\},$$

where X represents the distance moved by the small blue periwinkles.

This dataset was also explored by [4] using a nonparametric approach. Considering a regression model with a circular response (direction) and a single real-valued covariate (distance), the regression function was estimated using kernel-type methods.

In order to decide if $\mathcal{M}_{1,\boldsymbol{\beta}}$ or $\mathcal{M}_{3,\boldsymbol{\beta}}$ are plausible parametric models for the regression function with this dataset, the test statistic T_n^2 is applied twice considering $B = 500$ replications. In both cases, the parametric fits were computed by maximum likelihood (see Section 2.1). For further details on the estimation procedures, we refer to [1,2]. As for the nonparametric fit, the local linear-type estimator given in (5) with a triweight kernel was considered. As pointed out before, the performance of the test is analyzed in a range of bandwidths.

Figure 3 (left panel) shows the smoothed versions of the parametric fits when considering the parametric families $\mathcal{M}_{1,\boldsymbol{\beta}}$ (dashed line) and $\mathcal{M}_{3,\boldsymbol{\beta}}$ (dotted line), and the nonparametric regression estimator (solid line), using the leave-one-out cross validation (CV) bandwidth (see [4] for further details on bandwidth selection in this context). These curves are compared in the proposed test statistic. Figure 3 (right panel) shows the p-values of the tests for different bandwidths, when considering the parametric families $\mathcal{M}_{1,\boldsymbol{\beta}}$ (dashed line) or $\mathcal{M}_{3,\boldsymbol{\beta}}$ (dotted line) as the null hypothesis, using the significance traces [30]. Taking into account this plot, there are no evidences to reject the null hypothesis in both testing problems. However, it can be observed that for h larger than 15 the p-value decreases considerably when considering the parametric family $\mathcal{M}_{1,\boldsymbol{\beta}}$ in the null hypothesis.

[Figure 3 about here.]

6.2. Sand hopper data

In this section, the testing procedure is applied to the sand hopper dataset, which contains orientations of two species of male and female sand hoppers (*Talorchestia brito* and *Talitrus saltator*). With the purpose of analyzing how sand hopper orientation behaves when other variables are included as covariates (such as azimuth, pressure, temperature, among others), both parametric and nonparametric approaches have been considered in the literature. For instance, following the proposal in [2], [7] used a PMLM to model such data. The authors assumed a projected normal distribution for the scape directions with the corresponding parameters (circular mean and mean resultant vector) depending on the explanatory variables through a linear model. Using nonparametric tools, this dataset (for males and females, being the sample sizes $n = 330$ and $n = 404$, respectively) was also explored by [5], in order to check how orientation behaves when temperature and (relative) humidity are included as covariates. Only observations corresponding to relative humidity values larger than 45% were considered (the corresponding datasets for both sexes are plotted in Figure 4, with males in the left panel and females in the right panel). These authors provided regression estimates using the local linear-type estimator (5).

[Figure 4 about here.]

In order to determine if a parametric multiple regression model is an appropriate representation of these datasets (male and female sand hoppers), it is necessary to carry out a goodness-of-fit test for the selected parametric model. Assuming the parametric model used in [2] for this dataset, and taking into account the arguments in the previous section regarding the PMLM, the following parametric family is considered:

$$\mathcal{M}_{4,\boldsymbol{\beta}} = \{\text{atan2}(\beta_{0,2} + \beta_{1,2}X_1 + \beta_{2,2}X_2, \beta_{0,1} + \beta_{1,1}X_1 + \beta_{2,1}X_2)\},$$

with $\beta_{0,2}, \beta_{1,2}, \beta_{2,2}, \beta_{0,1}, \beta_{1,1}, \beta_{2,1} \in \mathbb{R}$, and $X_1 =$ "temperature" and $X_2 =$ "humidity".

The test statistic T_n^2 is applied with $B = 500$ replications. The parametric fit was computed by maximum likelihood (for further details on the estimation procedure, we refer to [2]). As for the nonparametric fit, the local linear-type estimator given in (5) with a multiplicative triweight kernel was considered. The bandwidth was taken as a diagonal matrix $\mathbf{H} = \text{diag}(h_1, h_2)$, being the values of h_1 and h_2 different. The range of bandwidths was selected taking into account the CV bandwidth matrices, which can be found in [5]. Figure 5 shows the smoothed version of the parametric fit for male (top left panel) and female (bottom left panel), and the nonparametric regression estimators for male (top right panel) and for female (bottom right panel), using the CV bandwidth matrices provided by [5]. The plots corresponding to the left panels are compared with the right panels in the proposed test statistic.

[Figure 5 about here.]

Figure 6 shows the approximated p-values of the test for male (left panel) and female (right panel), using the significance trace. Taking into account this figure, there are no evidences against the circular regression function belonging to the parametric family $\mathcal{M}_{4,\boldsymbol{\beta}}$, for both sexes.

[Figure 6 about here.]

7. Discussion and further research

Novel testing procedures for assessing a parametric circular regression model (with a circular response and an $\mathbb{R}^d$-valued covariate) were proposed and empirically analyzed in this work. The proposed test statistics were constructed by measuring a circular distance between a (non-smoothed or smoothed) parametric fit and a nonparametric estimator of the circular regression function. For the parametric approach, taking into account that the classical least squares regression method is not appropriate when the response variable is of circular nature, a circular analog can be used [1,20]. Other parametric fitting approaches, such as maximum likelihood methods, could be also employed. Regarding the nonparametric fit, although the test statistics were presented and numerically studied for a local linear-type estimator, they may be also defined considering a local polynomial-type estimator of a general order p. Nevertheless, significantly better results than the ones obtained for $p = 1$ (local linear case) are not expected. Moreover, although the multiplicative triweight kernel was considered in practice, some simulations were replicated using the Epanechnikov kernel, and similar results were obtained. In any case, as expected, the effect of the bandwidth in the performance of the tests is clearly more important than the effect of the kernel. For this reason, the tests were applied on a grid of several bandwidths.

Although the asymptotic distribution of the tests, under the null and under local alternatives, is out of the scope of this work, its derivation can follow from using a Taylor approximation of the function $1 - \cos(\Theta)$ by $\Theta^2/2$, for $\Theta \in [0, 2\pi)$ [3]. Using this approach, the expressions $1 - \cos[\hat{m}_{\mathbf{H}}(\mathbf{x}) - m_{\hat{\boldsymbol{\beta}}}(\mathbf{x})]$ and $1 - \cos[\hat{m}_{\mathbf{H}}(\mathbf{x}) - \hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x})]$ in the test statistics T_n^1 and T_n^2, given in (7) and in (8), respectively, can be approximated by

$\frac{1}{2}[\hat{m}_{\mathbf{H}}(\mathbf{x}) - m_{\hat{\boldsymbol{\beta}}}(\mathbf{x})]^2$ and $\frac{1}{2}[\hat{m}_{\mathbf{H}}(\mathbf{x}) - \hat{m}_{\mathbf{H},\hat{\boldsymbol{\beta}}}(\mathbf{x})]^2$, respectively. Consequently, T_n^1 and T_n^2 can be approximated by test statistics similar to the ones used, for example, in [8] or in [13], for regression models with Euclidean response and covariates. Notice that the regression estimators involved in the test statistics T_n^1 and T_n^2 have more complicated expressions than those in [8] or in [13]. Therefore, as intuition suggests, it will be more difficult to calculate close expressions of their asymptotic distributions.

For practical implementation, bootstrap resampling methods were used to calibrate the test. Two procedures have been designed and compared: PCB and NPCB. Both methods are based on computing the residuals and generating independent bootstrap resamples. The main difference between them is the mechanism employed to obtain the residuals. In PCB, the residuals come from the parametric regression estimator. Alternatively, in NPCB, the residuals are obtained from the nonparametric regression estimator. In the majority of scenarios considered in the simulation study, results obtained with NPCB improved those achieved by PCB, especially for alternative assumptions. Moreover, a better behaviour is observed when T_n^2, given in (8), is employed, showing the benefits of using the novel smoothed parametric estimator of the circular regression function defined in (9). The whole simulation study was repeated using a Nadaraya–Watson-type estimator for the nonparametric fits employed to compute the test statistics, given in (7) and in (8). The close expression of the Nadaraya–Watson-type estimator can be found in [5]. In this case, the procedures work fairly well when PCB is employed, while NPCB provides quite poor results. It seems that the tests statistics suffer from boundary problems induced by the use of the Nadaraya–Watson-type estimator, while this issue is overcome employing the local linear-type estimator. When using the Nadaraya–Watson-type estimator, probably a modification of the weight functions w used in the simulation study is required to obtain better results.

Along this work, data generated from the circular regression model are assumed to be independent. However, this assumption does not always hold in practical situations [31–33]. The construction of the proposed test statistics makes possible to easily extend the procedure for more general settings, such as spatially correlated data (or even with spatio-temporal correlation). The estimators described in Section 2.1 could be also employed for the parametric fit. Probably, more accurate results would be obtained if an estimator taking the dependence structure into account was used. However, the problem of estimating parametrically the circular regression function accounting the dependence structure, up to the knowledge of the authors, has not been tackled in the statistical literature. Regarding the nonparametric counterpart, the local linear-type estimator given in (5) could be used. With the purpose of calibrating the tests in a dependence framework, it should be noted that the bootstrap Algorithm 1, which was designed for independent data, should not be used for dependent data, as it does not account for the correlation structure. In order to mimic properly the distribution of the spatial dependence structure of the circular errors in the bootstrap procedure, Step 2 of Algorithm 1 should be modified. A possible approach to deal with this issue is to fit an appropriate spatial circular process to the residuals, such as the wrapped Gaussian spatial process [31], and generate a random sample from the fitted model.

In practice, the numerical studies performed in this work were run in an Intel Core i7-9700K at 3.60Ghz. The procedures were implemented in the statistical environment `R` [34], using functions included in the `npsp` and `CircSpaceTime` packages [35,36]. For regression models with a single real-valued covariate, the computing time for running the whole testing procedure (simulate a sample, compute the test statistics in a range of bandwidths and apply the bootstrap methods considering $B = 500$ replications) for a sample of size $n = 50, 100$ and 200 is around 2, 3 and 5 seconds, respectively, no matter the bootstrap method (PCB or NPCB) used to calibrate the test. For a bidimensional one, the computing times are around 4, 6 and 14 seconds, when $n = 100, 225$ and 400, respectively. As expected, considering a bidimensional covariate is more computationally expensive than using a single covariate.

Acknowledgements

The authors thank Prof. Felicita Scapini and her research team who kindly provided the sand hoppers data that are used in this work. Data were collected within the Project ERB ICI8-CT98-0270 from the European Commission, Directorate General XII Science. The authors also thank the Associate Editor and two anonymous referees for numerous useful comments that significantly improved this article.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research has been partially supported by MINECO (Grants MTM2016-76969-P and MTM2017-82724-R), MICINN (Grant PID2020-113578RB-I00) and by Xunta de Galicia (GrantED481A-2017/361, through the ESF. Grupos de Referencia Competitiva ED431C-2016-015,ED431C-2017-38 and ED431C-2020-14, and Centro de Investigacin del SUG ED431G 2019/0, through the ERDF).

ORCID

A. Meilán-Vila: http://orcid.org/0000-0001-8537-9280
M. Francisco-Fernández: http://orcid.org/0000-0002-9201-5423
R.M. Crujeira: http://orcid.org/0000-0002-3907-8951

References

[1] Fisher NI, Lee AJ. Regression models for an angular response. Biometrics. 1992;48(3):665–677.

[2] Presnell B, Morrison SP, Littell RC. Projected multivariate linear models for directional data. J Am Stat Assoc. 1998;93(443):1068–1077.

[3] Kim S, SenGupta A. Multivariate-multiple circular regression. J Stat Comput Sim. 2017; 87(7):1277–1291.

[4] Di Marzio M, Panzera A, Taylor CC. Non-parametric regression for circular responses. Scand J Stat. 2013;40(2):238–255.

[5] Meilán-Vila A, Francisco-Fernández M, Crujeiras RM, et al. Nonparametric multiple regression estimation for circular response. TEST. 2021;30(3):650–672.

[6] Meilán-Vila A, Crujeiras RM, Francisco-Fernández M. Nonparametric estimation of circular trend surfaces with application to wave directions. Stoch Environ Res Risk Assess. 2021;35(4):923–939.

[7] Scapini F, Aloia A, Bouslama MF, et al. Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, *Talitrus saltator* and *Talorchestia brito*, from an exposed Mediterranean beach. Behav Ecol Sociobiol. 2002;51(5):403–414.

[8] Härdle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Stat. 1993;21:1926–1947.

[9] Alcalá J, Cristóbal J, González-Manteiga W. Goodness-of-fit test for linear models based on local polynomials. Statist Probab Lett. 1999;42:39–46.

[10] Kozek AS. A nonparametric test of fit of a parametric model. J Multivar Anal. 1991; 37(1):66–75.

[11] González-Manteiga W, Vilar-Fernández J. Testing linear regression models using nonparametric regression estimators when errors are non-independent. Comput Stat Data Anal. 1995;20:521–541.

[12] Park C, Kim TY, Ha J, et al. Using a bimodal kernel for a nonparametric regression specification test. Stat Sin. 2015;25:1145–1161.

[13] Meilán-Vila A, Opsomer JD, Francisco-Fernández M, et al. A goodness-of-fit test for regression models with spatially correlated errors. TEST. 2020;29:728–749.

[14] Meilán-Vila A, Fernández-Casal R, Crujeiras RM, et al. A computational validation for nonparametric assessment of spatial trends. Comput Stat. 2021;36:2939–2965.

[15] González-Manteiga W, Crujeiras RM. An updated review of Goodness-of-Fit tests for regression models. TEST. 2013;22:361–411.

[16] Deschepper E, Thas O, Ottoy JP. Tests and diagnostic plots for detecting lack-of-fit for circular-linear regression models. Biometrics. 2008;64(3):912–920.

[17] García-Portugués E, Van Keilegom I, Crujeiras and RM, et al. Testing parametric models in linear-directional regression. Scand Stat. 2016;43(4):1178–1191.

[18] Alonso-Pena M, Ameijeiras-Alonso J, Crujeiras RM. Nonparametric tests for circular regression. J Stat Comput Simul. 2021;91(3):477–500.

[19] Jammalamadaka SR, SenGupta A. Topics in circular statistics. Vol. 5. World Scientific; 2001.

[20] Lund U. Least circular distance regression for directional data. J of Appl Stat. 1999; 26(6):723–733.

[21] González-Manteiga W, Cao R. Testing the hypothesis of a general linear model using nonparametric regression estimation. TEST. 1993;2(1-2):161–188.

[22] Best DJ, Fisher NI. The bias of the maximum likelihood estimators of the von mises-fisher concentration parameters: the bias of the maximum likelihood estimators. Commun Stat-Simul C. 1981;10(5):493–502.

[23] Ruppert D, Wand MP. Multivariate locally weighted least squares regression. Ann Stat. 1994;22:1346–1370.

[24] Fan J, Zhang C, Zhang J. Generalized likelihood ratio statistics and wilks phenomenon. Ann Stat. 2001;:153–193.

[25] Eubank RL, Li CS, Wang S. Testing lack-of-fit of parametric regression models using nonparametric regression techniques. Stat Sin. 2005;15:135–152.

[26] Hart J. Nonparametric smoothing and lack-of-fit tests. Springer Science & Business Media; 2013.

[27] Opsomer J, Francisco-Fernández M. Finding local departures from a parametric model using nonparametric regression. Stat Pap. 2010;51:69.

[28] Lund U, Agostinelli C, Arai H, et al. circular: Circular statistics; 2020. R package version 0.4-93; Available from: http://cran.r-project.org/package=circular.

[29] Wang F, Gelfand AE. Directional data analysis under the general projected normal distribution. Stat Methodol. 2013;10(1):113–127.

[30] Bowman AW, Azzalini A. Applied smoothing techniques for data analysis: the kernel approach with s-plus illustrations. Vol. 18. OUP Oxford; 1997.

[31] Jona-Lasinio G, Gelfand A, Jona-Lasinio M. Spatial analysis of wave direction data using wrapped Gaussian processes. Ann Appl Stat. 2012;6(4):1478–1498.

[32] Lagona F, Picone M, Maruotti A. A hidden Markov model for the analysis of cylindrical time series. Environmetrics. 2015;26(8):534–544.

[33] Mastrantonio G, Gelfand AE, Lasinio GJ. The wrapped skew Gaussian process for analyzing spatio-temporal data. Stoch Env Res Risk A. 2016;30(8):2231–2242.

[34] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: http://www.R-project.org.

[35] Fernández-Casal R. npsp: Nonparametric spatial (geo)statistics; 2021. R package version 0.7-8; Available from: https://rubenfcasal.github.io/npsp.

[36] Jona-Lasinio G, Mastrantonio G, Santoro M. CircSpaceTime: Spatial and spatio-temporal bayesian model for circular data; 2019. R package version 0.9.0; Available from: http://cran.r-project.org/package=CircSpaceTime.

Table 1. Proportions of rejections of the null hypothesis for the parametric family $\mathcal{M}_{1,\beta}$ with different sample sizes and $\kappa = 10$. Significance level: $\alpha = 0.05$.

Test statistic	c	n	Method	$h = 0.15$	$h = 0.25$	$h = 0.35$	$h = 0.45$	$h = 0.55$	$h = 0.65$
T_n^1	0	50	PCB	0.032	0.036	0.034	0.040	0.042	0.042
			NPCB	0.048	0.040	0.044	0.048	0.046	0.050
		100	PCB	0.026	0.026	0.032	0.032	0.036	0.034
			NPCB	0.028	0.028	0.032	0.034	0.036	0.034
		200	PCB	0.024	0.028	0.028	0.034	0.036	0.034
			NPCB	0.026	0.034	0.026	0.036	0.040	0.046
	1	50	PCB	0.100	0.124	0.148	0.162	0.156	0.152
			NPCB	0.142	0.156	0.170	0.184	0.184	0.174
		100	PCB	0.212	0.264	0.300	0.324	0.318	0.304
			NPCB	0.250	0.306	0.344	0.352	0.352	0.336
		200	PCB	0.504	0.604	0.642	0.660	0.668	0.666
			NPCB	0.548	0.636	0.674	0.686	0.692	0.680
	2	50	PCB	0.380	0.506	0.574	0.606	0.618	0.598
			NPCB	0.478	0.582	0.638	0.672	0.678	0.670
		100	PCB	0.856	0.934	0.952	0.958	0.962	0.962
			NPCB	0.896	0.944	0.964	0.972	0.970	0.970
		200	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000
T_n^2	0	50	PCB	0.032	0.032	0.040	0.044	0.042	0.040
			NPCB	0.044	0.042	0.046	0.050	0.052	0.050
		100	PCB	0.026	0.028	0.038	0.032	0.030	0.030
			NPCB	0.026	0.030	0.038	0.036	0.036	0.036
		200	PCB	0.028	0.028	0.028	0.024	0.032	0.038
			NPCB	0.026	0.030	0.030	0.022	0.032	0.038
	1	50	PCB	0.106	0.118	0.144	0.140	0.156	0.146
			NPCB	0.140	0.154	0.168	0.182	0.188	0.180
		100	PCB	0.214	0.260	0.288	0.290	0.290	0.270
			NPCB	0.248	0.298	0.324	0.334	0.336	0.312
		200	PCB	0.502	0.582	0.610	0.618	0.644	0.620
			NPCB	0.536	0.610	0.632	0.650	0.654	0.640
	2	50	PCB	0.380	0.500	0.548	0.570	0.562	0.558
			NPCB	0.476	0.574	0.620	0.626	0.638	0.620
		100	PCB	0.840	0.924	0.944	0.946	0.948	0.944
			NPCB	0.894	0.944	0.962	0.966	0.960	0.958
		200	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	0.998

Table 2. Proportions of rejections of the null hypothesis for the parametric family $\mathcal{M}_{1,\beta}$ with different values of κ and $n = 200$. Significance level: $\alpha = 0.05$.

Test statistic	c	κ	Method	$h = 0.15$	$h = 0.25$	$h = 0.35$	$h = 0.45$	$h = 0.55$	$h = 0.65$
T_n^1	0	5	PCB	0.030	0.030	0.024	0.024	0.028	0.030
			NPCB	0.034	0.030	0.024	0.024	0.028	0.030
		10	PCB	0.024	0.028	0.028	0.034	0.036	0.034
			NPCB	0.026	0.034	0.026	0.036	0.040	0.046
		15	PCB	0.026	0.034	0.030	0.034	0.038	0.040
			NPCB	0.026	0.030	0.028	0.034	0.034	0.036
	1	5	PCB	0.200	0.262	0.282	0.306	0.290	0.278
			NPCB	0.216	0.276	0.306	0.314	0.320	0.294
		10	PCB	0.504	0.604	0.642	0.660	0.668	0.666
			NPCB	0.548	0.636	0.674	0.686	0.692	0.680
		15	PCB	0.764	0.836	0.878	0.880	0.868	0.850
			NPCB	0.784	0.856	0.882	0.882	0.872	0.868
	2	5	PCB	0.872	0.916	0.930	0.942	0.942	0.928
			NPCB	0.884	0.918	0.938	0.946	0.940	0.930
		10	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000
		15	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000
T_n^2	0	5	PCB	0.032	0.028	0.022	0.022	0.028	0.026
			NPCB	0.034	0.028	0.026	0.030	0.030	0.028
		10	PCB	0.028	0.028	0.028	0.024	0.032	0.038
			NPCB	0.026	0.030	0.030	0.022	0.032	0.038
		15	PCB	0.028	0.034	0.038	0.034	0.034	0.034
			NPCB	0.026	0.038	0.036	0.036	0.032	0.034
	1	5	PCB	0.198	0.252	0.264	0.272	0.268	0.250
			NPCB	0.218	0.274	0.280	0.290	0.280	0.264
		10	PCB	0.502	0.582	0.610	0.618	0.644	0.620
			NPCB	0.536	0.610	0.632	0.650	0.654	0.640
		15	PCB	0.752	0.826	0.862	0.868	0.868	0.858
			NPCB	0.782	0.846	0.874	0.874	0.884	0.868
	2	5	PCB	0.870	0.910	0.918	0.932	0.928	0.918
			NPCB	0.884	0.916	0.930	0.942	0.938	0.932
		10	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	0.998
		15	PCB	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000

Table 3. Proportions of rejections of the null hypothesis for the parametric family $\mathcal{M}_{2,\beta}$ with different sample sizes and $\kappa = 10$. Significance level: $\alpha = 0.05$.

Test statistics	c	n	Method	$h = 0.25$	$h = 0.35$	$h = 0.45$	$h = 0.55$	$h = 0.65$	$h = 0.75$	$h = 0.85$
T_n^1	0	100	PCB	0.030	0.034	0.048	0.050	0.062	0.066	0.068
			NPCB	0.044	0.048	0.058	0.062	0.062	0.064	0.068
		225	PCB	0.030	0.032	0.028	0.038	0.042	0.042	0.042
			NPCB	0.024	0.030	0.032	0.038	0.042	0.044	0.044
		400	PCB	0.042	0.040	0.042	0.038	0.030	0.036	0.038
			NPCB	0.034	0.038	0.040	0.030	0.032	0.036	0.036
	1	100	PCB	0.102	0.066	0.034	0.008	0.004	0.004	0.000
			NPCB	0.158	0.106	0.038	0.014	0.004	0.004	0.000
		225	PCB	0.362	0.264	0.140	0.058	0.020	0.008	0.004
			NPCB	0.372	0.280	0.152	0.068	0.022	0.008	0.004
		400	PCB	0.724	0.614	0.396	0.198	0.066	0.030	0.020
			NPCB	0.722	0.616	0.392	0.190	0.076	0.032	0.020
	2	100	PCB	0.574	0.548	0.442	0.302	0.176	0.098	0.070
			NPCB	0.640	0.600	0.478	0.342	0.202	0.114	0.078
		225	PCB	0.992	0.990	0.976	0.916	0.776	0.638	0.472
			NPCB	0.992	0.994	0.980	0.924	0.804	0.664	0.508
		400	PCB	1.000	1.000	1.000	0.998	0.992	0.976	0.932
			NPCB	1.000	1.000	1.000	0.998	0.996	0.984	0.940
T_n^2	0	100	PCB	0.040	0.048	0.060	0.054	0.062	0.050	0.050
			NPCB	0.060	0.064	0.066	0.064	0.066	0.070	0.068
		225	PCB	0.032	0.032	0.030	0.038	0.046	0.044	0.046
			NPCB	0.038	0.034	0.032	0.044	0.048	0.044	0.042
		400	PCB	0.030	0.034	0.040	0.042	0.036	0.038	0.030
			NPCB	0.032	0.032	0.040	0.036	0.030	0.034	0.034
	1	100	PCB	0.132	0.220	0.292	0.332	0.336	0.344	0.342
			NPCB	0.194	0.266	0.332	0.368	0.382	0.386	0.370
		225	PCB	0.398	0.554	0.636	0.672	0.680	0.682	0.672
			NPCB	0.418	0.552	0.640	0.670	0.678	0.676	0.662
		400	PCB	0.944	0.984	0.994	0.994	0.990	0.990	0.988
			NPCB	0.938	0.978	0.994	0.994	0.992	0.988	0.988
	2	100	PCB	0.508	0.736	0.856	0.894	0.904	0.904	0.898
			NPCB	0.556	0.752	0.854	0.898	0.902	0.902	0.902
		225	PCB	0.980	0.996	0.998	1.000	1.000	1.000	1.000
			NPCB	0.980	0.996	0.998	0.998	1.000	1.000	1.000
		400	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4. Proportions of rejections of the null hypothesis for the parametric family $\mathcal{M}_{2,\beta}$ with different values of κ and $n = 400$. Significance level: $\alpha = 0.05$.

Test statistic	c	κ	Method	$h = 0.25$	$h = 0.35$	$h = 0.45$	$h = 0.55$	$h = 0.65$	$h = 0.75$	$h = 0.85$
T_n^1	0	5	PCB	0.026	0.034	0.040	0.038	0.042	0.038	0.038
			NPCB	0.026	0.036	0.038	0.038	0.042	0.038	0.038
		10	PCB	0.042	0.040	0.042	0.038	0.030	0.036	0.038
			NPCB	0.034	0.038	0.040	0.030	0.032	0.036	0.036
		15	PCB	0.038	0.044	0.040	0.038	0.032	0.038	0.038
			NPCB	0.030	0.036	0.038	0.036	0.034	0.040	0.038
	1	5	PCB	0.354	0.268	0.176	0.090	0.046	0.020	0.014
			NPCB	0.360	0.290	0.178	0.094	0.048	0.020	0.012
		10	PCB	0.724	0.614	0.396	0.198	0.066	0.030	0.020
			NPCB	0.722	0.616	0.392	0.190	0.076	0.032	0.020
		15	PCB	0.936	0.802	0.560	0.294	0.140	0.050	0.022
			NPCB	0.922	0.792	0.554	0.302	0.136	0.050	0.026
	2	5	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		10	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		15	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
T_n^2	0	5	PCB	0.030	0.024	0.024	0.026	0.028	0.030	0.026
			NPCB	0.046	0.046	0.050	0.050	0.050	0.054	0.052
		10	PCB	0.030	0.034	0.040	0.042	0.036	0.038	0.030
			NPCB	0.032	0.032	0.040	0.036	0.030	0.034	0.034
		15	PCB	0.034	0.042	0.040	0.044	0.038	0.042	0.040
			NPCB	0.024	0.024	0.038	0.044	0.042	0.036	0.036
	1	5	PCB	0.566	0.684	0.744	0.776	0.786	0.794	0.786
			NPCB	0.574	0.692	0.752	0.776	0.794	0.790	0.780
		10	PCB	0.944	0.984	0.994	0.994	0.990	0.990	0.988
			NPCB	0.938	0.978	0.994	0.994	0.992	0.988	0.988
		15	PCB	0.990	0.998	0.998	0.998	0.998	0.998	0.998
			NPCB	0.990	0.998	0.998	0.998	0.998	0.998	0.998
	2	5	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		10	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		15	PCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000
			NPCB	1.000	1.000	1.000	1.000	1.000	1.000	1.000

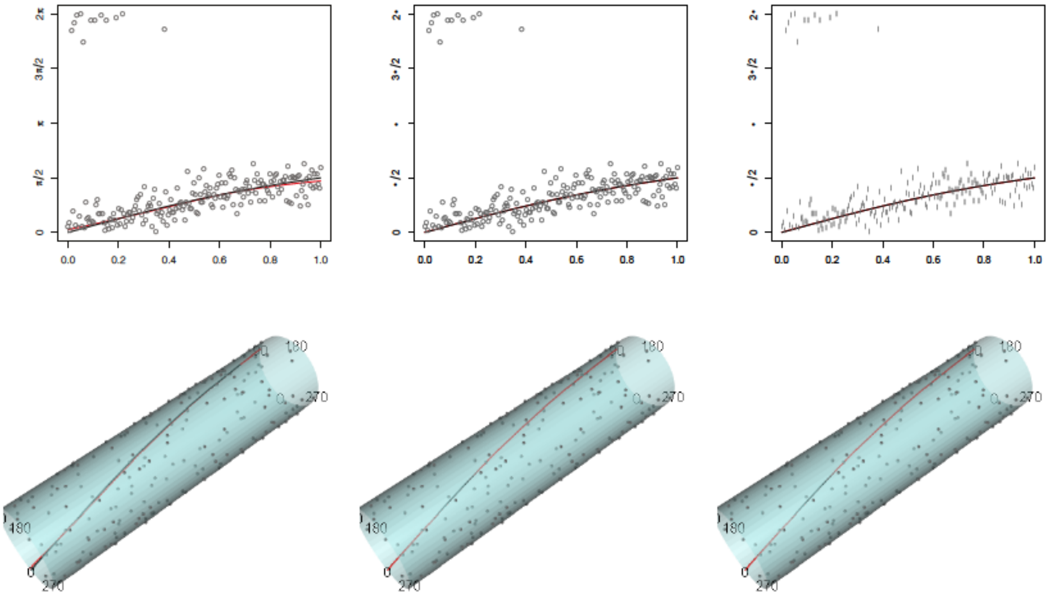

2π
3π/2
π
π/2
0
0.0
0.2
0.4
0.6
0.8
1.0
2•
3•/2
•
•/2
0
0.0
0.2
0.4
0.6
0.8
1.0
2•
3•/2
•
•/2
0
0.0
0.2
0.4
0.6
0.8
1.0
90
180
0
270
180
0
270
90
180
0
270
180
0
270
90
180
0
270
180
0
270

Figure 1.

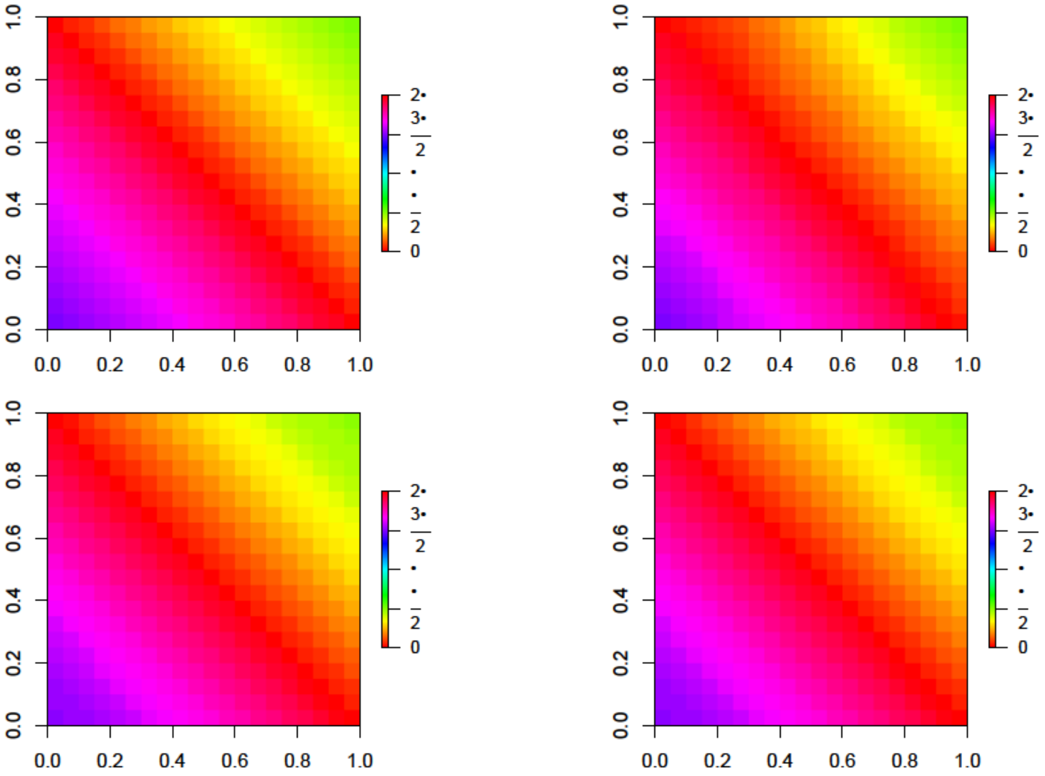


Figure 2.

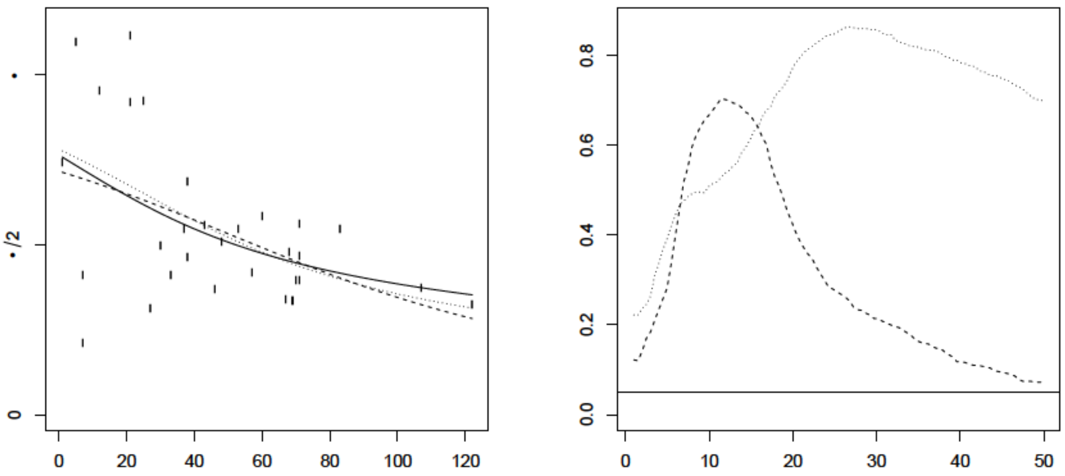


Figure 3.

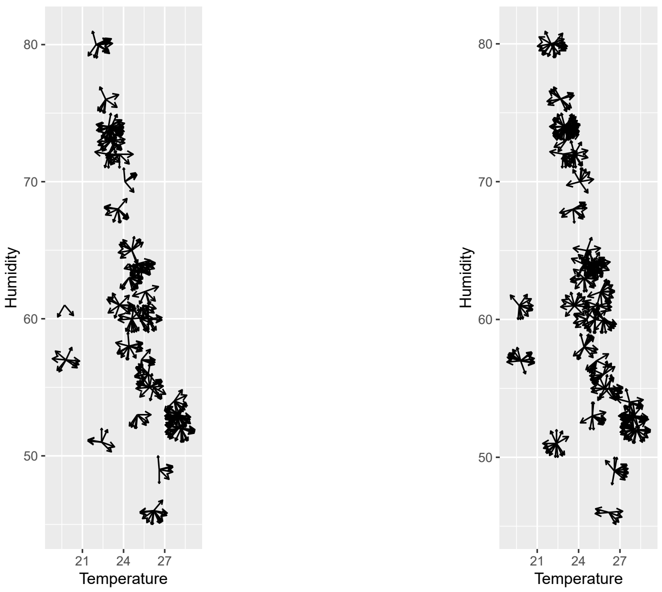
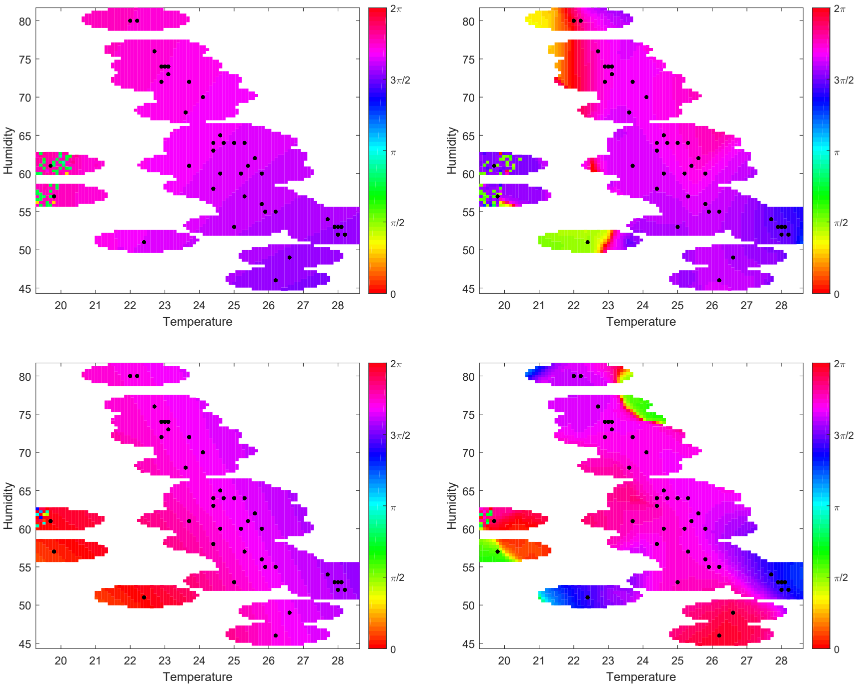
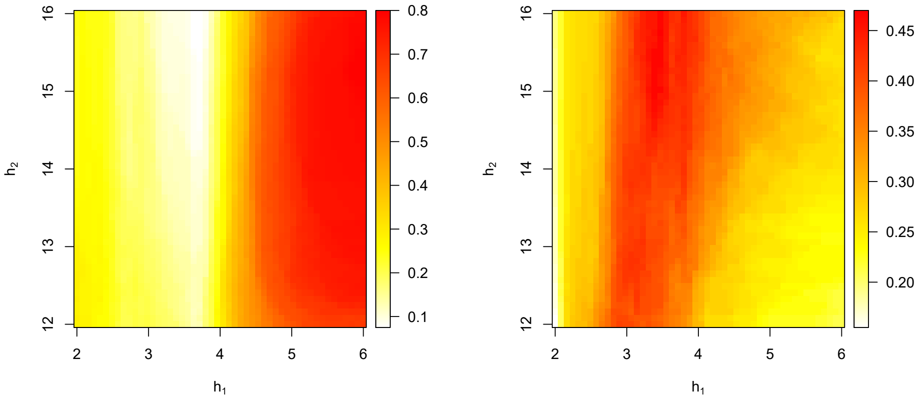

Humidity
80
70
60
50
21
24
27
Temperature
Humidity
80
70
60
50
21
24
27
Temperature

Figure 4.



Humidity
Temperature
80
75
70
65
60
55
50
45
20
21
22
23
24
25
26
27
28
2π
3π/2
π
π/2
0

Figure 5.



h2
12
13
14
15
16
2
3
4
5
6
h1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
h2
12
13
14
15
16
2
3
4
5
6
h1
0.20
0.25
0.30
0.35
0.40
0.45

Figure 6.

Figure captions

Figure 1. Linear (top panels) and cylinder (bottom panels) representations. Red lines: local linear-type regression estimator (left panels), parametric fit (center panels) and smoothed version of the parametric fit (right panels), with sample points and circular regression function (black lines). Equally-spaced sample of size $n = 200$ generated on the unit interval, following model (1), with regression function (13), for $c = 0$, and circular errors ε_i drawn from a $vM(0, 10)$.

Figure 2. Circular regression function (top left panel), local linear-type regression estimator (top right panel), parametric fit (bottom left panel) and smoothed version of the parametric fit (bottom right panel). Sample of size $n = 400$ generated on a bidimensional regular grid in the unit square, following model (1), with regression function (15), for $c = 0$, and circular errors ε_i drawn from a $vM(0, 10)$.

Figure 3. Left panel: Sample of directions and distances moved by periwinkles (circle points), smoothed versions of the parametric fits when considering the parametric families $\mathcal{M}_{1,\boldsymbol{\beta}}$ (dashed line) and $\mathcal{M}_{3,\boldsymbol{\beta}}$ (dotted line), and local linear-type regression estimator (solid line), using the CV bandwidth. Right panel: p-values of the test when considering the parametric family $\mathcal{M}_{1,\boldsymbol{\beta}}$ (dashed line) and $\mathcal{M}_{3,\boldsymbol{\beta}}$ (dotted line) as the null hypothesis for different values of h. Horizontal solid line represents the value 0.05.

Figure 4. Observed orientation of male (left) and female (right) sand hoppers varying with temperature and relative humidity.

Figure 5. Smoothed version of the parametric fit for male (top left panel) and female (bottom left panel), and local linear-type regression estimators for male (top right panel) and for female (bottom right panel), using the CV bandwidth matrices. Horizontal axis: temperature in Celsius degrees. Vertical axis: relative humidity in percentage.

Figure 6. For male (left panel) and female (right panel) sand hopper orientation dataset, p-values of the test for different values of h_1 and h_2, considering the parametric family $\mathcal{M}_{4,\boldsymbol{\beta}}$ as the null hypothesis.