Master's Thesis

Nonparametric ANCOVA for cylindrical and toroidal data

María Alonso Pena

Máster en Técnicas Estatísticas

Academic year 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: ANCOVA non paramétrico para datos cilíndricos e toroidais

Título en español: ANCOVA no paramétrico para datos cilíndricos y toroidales

English title: Nonparametric ANCOVA for cylindrical and toroidal data

Modalidad: Modalidad A

Autora: María Alonso Pena, Universidad de Santiago de Compostela

Directores: Rosa M. Crujeiras Casais, Universidade de Santiago de Compostela ; Jose Ameijeiras Alonso, KU Leuven

Breve resumen del trabajo:

El objetivo principal de este trabajo es extender los modelos de regresión ANCOVA no paramétricos al caso en el que alguna de las variables (o ambas) son de naturaleza circular. Se presentan métodos para contrastar la igualdad y el paralelismo de las curvas de regresión en cada uno de los escenarios posibles. Los métodos propuestos son analizados mediante un estudio de simulación. Además, se utilizan dos conjuntos de datos reales para ilustrar las nuevas propuestas, el primero referente a datos pertenecientes a la industria automovilística, y el segundo perteneciente al estudio de la orientacion de animales.

Otras observaciones:

Este TFM es una propuesta de estudiante presentada por María Alonso Pena.

Doña Rosa M. Crujeiras Casais, profesora titular del área de estadística e investigación operativa (Departamento de Estadística, Análisis Matemático y Optimización) de la Universidade de Santiago de Compostela y don Jose Ameijeiras Alonso, investigador postdoctoral del área de estadística (Departamento de Matemáticas) de la KU Leuven informan que el Trabajo Fin de Máster titulado

Nonparametric ANCOVA for cylindrical and toroidal data

fue realizado bajo su dirección por doña María Alonso Pena para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 3 de Julio de 2019.

La directora: El director:

Doña Rosa M. Crujeiras Casais Don Jose Ameijeiras Alonso

La autora:

Doña María Alonso Pena

Acknowledgments

To begin with, I would like to thank Instituto de Matemáticas da USC (IMAT) for supporting this project through the scholarship *Beca de colaboración en investigación do Instituto de Matemáticas 2017-2018 (Modalidade Máster)* under *Plan de Lanzamento da Área de Matemáticas da USC 2017-2019*. Secondly, I also want to thank professor Felicita Scapini from the University of Florence for kindly providing one of the datasets employed in this manuscript, which was obtained under the European Project ERB ICI8-CT98-0270 (Bases for the Integrated Sustainable Management of Mediterranean Sensitive Coastal System). The Supercomputing Center of Galicia (CESGA) is also acknowledged for providing the computational resources which allowed to perform many of the simulations achieved in this project. Lastly, I want to thank my directors, Rosa and Jose, for outstandingly guiding me through the elaboration of this MSc Thesis.

Contents

Abstract

This work presents new proposals for nonparametric ANCOVA models in regression contexts where the predictor variable, the response variable or both are of a circular nature. Testing tools for assessing equality and parallelism of the regression curves in those scenarios are provided. In addition, the problem of determining the significance of the predictor variable in regression settings with circular variables is analyzed, and new approaches for assessing the significance of the covariate are also given. The performance of the proposed methods is analyzed in an extensive simulation study, in which the calibration and power of the tests are investigated. To conclude, the novel techniques are applied to real data.

Resumo en galego

Neste traballo preséntanse propostas de modelos ANCOVA non paramétricos nos contextos de regresión onde a variable explicativa, a resposta, ou ambas teñen natureza circular. Proporciónanse ferramentas para contrastar a igualdade e o paralelismo das curvas de regresión en ditos escenarios. Ademais, tamén se analiza o problema de determinar a significación da variable explicativa no marco da regresión con variables circulares, e apórtanse propostas de test de significación. Estes métodos analízanse mediante un estudo de simulación exhaustivo, nos que se investiga o calibrado e a potencia dos test. Por último, as novas técnicas son aplicadas a datos reais.

Resumen en español

En este trabajo se presentan propuestas de modelos ANCOVA no paramétricos en contextos de regresión donde la variable explicativa, la respuesta o ambas tienen naturaleza circular. Se proporcionan herramientas para contrastar la igualdad y el paralelismo de las curvas de regresión en dichos escenarios. Además, se analiza también el problema de determinar la significación de la variable explicativa en el marco de la regresión con variables circulares, y se aportan propuestas de test de significación. Estos métodos son analizados mediante un estudio de simulación exhaustivo, en los que se investiga el calibrado y la potencia de los test. Por último, las nuevas técnicas son aplicadas a datos reales.

Preface

Circular statistics is a branch of statistics which involves angles and directions as observations. One may think that this kind of data can be treated with regular statistical techniques, but hideous mistakes would be made when doing so. A simple example to see what can go wrong can be found in the sample mean. Consider an experiment where the interest lies on studying the direction in which a group of sandhoppers move in the sand from a given starting point, where 0^o represents the north direction. If it was observed that three sandhoppers moved in directions 3^o, 355^o and 5^o, respectively, one can trivially say that the animals move north. However, the mean of those angles is 121^o, which would mean that the sandhoppers move in the south-west direction. This toy example gives an idea of why different methods must be applied to circular data.

Classical theory on circular statistics has been around for decades, but its usage in practice has been limited because of the lack of accurate circular observations. Advances on technology have made it possible to precisely record these type of data, increasing the interest of circular statistics in recent years. Consequently, in the last decade many advances on the circular statistics field have been achieved, including nonparametric methods for this kind of observations, specifically for regression.

In the context of regression, when the data belongs to different groups it is useful to determine the existence of differences on the regression functions for each group. This can be done through an ANCVOA model. The primary objective of this project is to come up with nonparametric ANCOVA models for regression with circular variables (predictors, responses or both). In addition, the problem of determining the significance of the predictor variable in a circular regression model is also approached.

The distribution of the manuscript is as follows: Chapter 1 gives a review on ANCOVA focusing on nonparametric alternatives. Chapter 2 presents some background on circular data, including the most widely used parametric models for density and regression. Chapter 3 is devoted to the introduction of the nonparametric proposals of significance tests and analysis of covariance models. Chapter 4 contains a simulation study which analyzes the proposals given in this MSc Thesis. Moreover Chapter 5 gives an illustration of the new methods with real data. Lastly, Chapter 6 contains the principal conclusions of the project, as well as some possible extensions.

Chapter 1

A review on nonparametric ANCOVA

Regression models are meant to investigate the relationship of dependence between a response variable and an explanatory variable. Once this relationship is modeled, the regression analysis can be used to predict future values of the response given the predictor. However, an interesting problem arises when wondering if the predictor variable actually has any effect on the responses.

On the other hand, different kinds of regression models emerge when considering a discrete covariate in addition to the predictor variable, dividing the sample between several groups. Such models are known as ANalysis of COVAriance models (ANCOVA), and can be thought as an extension of the one-way ANalysis Of VAriance (ANOVA) when a continuous covariate is included.

Inference with regression models is usually done under a parametric (generally linear) perspective, and methods for testing the signification of the predictor variable or for assessing the differences of the regression curves between several groups can be easily obtained under the linear regression model. However, more flexible models can be considered by using nonparametric regression. Bowman and Azzalini (1997) derived a nonparametric significance test to investigate the effect of the predictor on the response variable. In addition, Young and Bowman (1995) proposed an analysis of covariance model where the regression function does not take any specific parametric form.

In this chapter, some background on parametric and nonparametric regression models will be presented. Section 1.1 focuses on the linear setting, and a significance test, as well as tests for the ANOVA and ANCOVA models are reviewed. Section 2 is devoted to the nonparametric context. First, a brief introduction to kernel methods for regression will be provided. Afterwards, the nonparametric no-effect test and ANCOVA models will be studied and illustrated with real data examples. A brief simulation study investigating calibration and power of the tests will also be included.

1.1 The linear regression model

Regression models are used to represent the dependence of a response variable (Y) regarding one or several explanatory variables (X). The regression function is usually defined as

$$m(x) = \mathbb{E}(Y|X = x),$$

the expected value of Y when X takes the specific value x. Assuming a linear regression function m is a simple and commonly used hypothesis. Under this assumption, and having p explanatory variables $X_1, X_2, ..., X_p$, the regression function is written as

$$m(x) = \gamma + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p,$$

where $\gamma, \beta_1, ..., \beta_p$ are real-valued regression parameters. Given the observations $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_j = (X_{j1}, X_{j2}, ..., X_{jp})'$, $j = 1, ..., n$, with n being the sample size, the linear regression model can be expressed as

$$Y_j = \gamma + \beta_1 X_{j1} + \beta_2 X_{j2} + ... + \beta_p X_{jp} + \varepsilon_j, \tag{1.1}$$

where ε_j is the random error and $\mathbb{E}(\varepsilon | X = x) = 0$. It is usually assumed that the errors ε_j are independent with distribution $N(0, \sigma)$.

In the rest of the section, for simplicity, only one continuous predictor variable, X, will be considered, although ideas could be easily extended to the multivariate case. Then, the resulting model is

$$Y_j = \gamma + \beta X_j + \varepsilon_j, \quad j \in \{1, ..., n\}, \ \gamma, \beta \in \mathbb{R}. \tag{1.2}$$

In order to fit model (1.2), the parameters can be estimated through the least squares method, obtaining the estimates

$$\hat{\gamma} = \bar{Y} - \frac{S_{XY}}{S_X^2}\bar{X} \quad \text{and} \quad \hat{\beta} = \frac{S_{XY}}{S_X^2}, \tag{1.3}$$

where S_{XY} is the sample covariance between X and Y, S_X^2 is the sample variance of the explanatory variable and $\bar{X}$ and $\bar{Y}$ are, respectively, the sample mean of the predictors and the responses. Faraway (2004) and Sheather (2009) give a complete review of linear regression models, including inference methods on the parameters.

A discrete explanatory variable taking I attributes can also be added to the regression model (1.2). Then, every observation belongs to one of the I groups, with n_i being the number of observations in each group and $n = \sum_{i=1}^{I} n_i$ the total sample size. Therefore, the regression model can be formulated as

$$Y_{ij} = \gamma_i + \beta_i X_{ij} + \varepsilon_{ij}, \quad i \in \{1, ..., I\}, \ j \in \{1, ..., n_i\}. \tag{1.4}$$

In this section a significance test (also known as no-effect test) for model (1.2) will be surveyed. In addition, the concepts of ANOVA and ANCOVA will be studied. In the ANOVA model, the test of equality of means will be presented, while in the ANCOVA context three different tests will be shown: two equality tests (one for the model without interaction and one for the model with interaction) and a parallelism test. A more detailed discussion on ANOVA and ANCOVA models, among other topics, can be found in Maxwell and Delaney (2003, Chapters 3 and 9). All the tests will be shown in practice using classical data given in Fisher (1936) and collected by Anderson (1935). The dataset is called `iris` and it is available in package `datasets` (R Core Team, 2018). It contains, among other information, the sepal length and width (measured in centimeters) for 50 flowers from each of 3 species of iris: Iris setosa, versicolor, and virginica.

1.1.1 Significance test

In many situations, after having fitted a linear model of the type (1.2), it could be unclear if the predictor variable is actually significant or, if on the contrary, the regression function is just a horizontal line. In the later case, the regression would not give any relevant information about the response variable. In order to ascertain this, a significance test can be carried out with the following hypotheses:

$$H_0 : Y_j = \gamma + \varepsilon_j, \quad \forall\ j \in \{1, ..., n\}$$
$$H_1 : Y_j = \gamma + \beta X_j + \varepsilon_j, \quad \beta \neq 0, \ \forall\ j \in \{1, ..., n\}.$$

This test is also called no-effect test since under the null hypothesis the predictor variable has no effect on the responses. Note that in the case of only one predictor variable this test is equivalent to contrast the null hypothesis $\beta = 0$. A suitable test statistic is

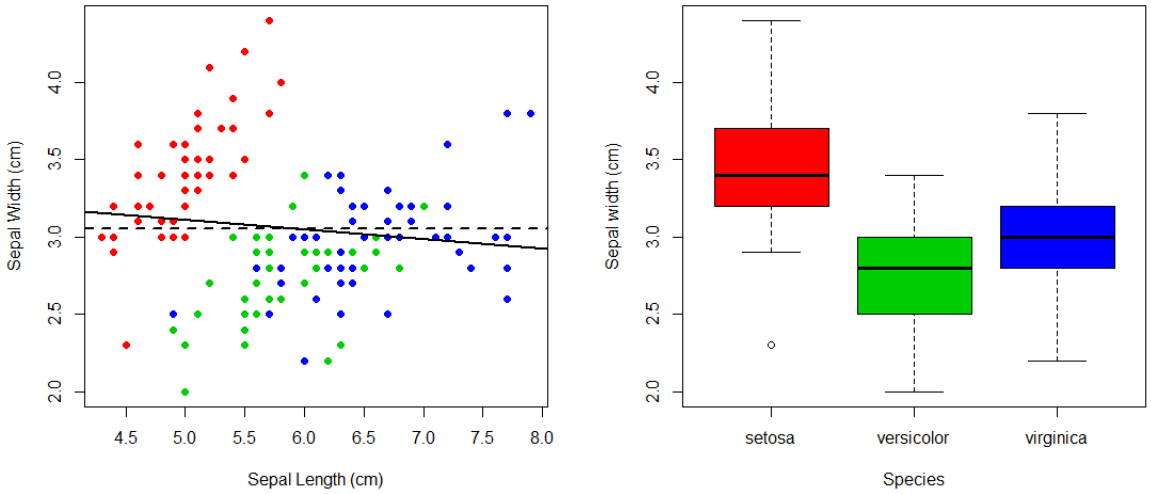
$$\frac{RSS_0 - RSS}{RSS/(n-2)} \sim F_{1,n-2},$$

where $F_{n,m}$ denotes the F-Snedecor distribution with n and m degrees of freedom and the residual sums of squares under H_0 and H_1 are, respectively,

$$RSS_0 = \sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma})^2 \quad \text{and} \quad RSS = \sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma} - \hat{\beta}X_j)^2.$$

In RSS_0, the estimator of γ is the sample mean of the responses, $\bar{Y}$, while in RSS the estimators $\hat{\gamma}$ and $\hat{\beta}$ correspond to the ones obtained by the least squares method presented in (1.3). The number of degrees of freedom under H_0 is $(n-1)$, since only the mean value of the responses is being estimated. Under H_1, two parameters are being estimated, hence the number of degrees of freedom is $(n-2)$. Therefore, the numerator of the statistic has a total of 1 degree of freedom.

To illustrate the previous method, the test is now applied to the iris data. The left image in Figure 1.1 shows a scatter plot of the response variable, sepal width, over the predictor variable, sepal length, with the estimation of the regression function under both hypotheses. The linear model was fitted and the obtained estimation of the line's slope was -0.061, while the intercept estimation was 3.418. On the other hand, under the null hypothesis the value of the responses does not depend on the sepal length, and the expectation of the sepal width is estimated as the sample mean, which is 3.057. After applying the test, the obtained p-value was 0.1519, therefore there is no evidence to reject the null hypothesis of no effect of the predictor for the usual significance levels (0.10, 0.05 and 0.01). However, when focusing on the different types of iris, it seems noticeable that differences arise between the three classes. The right panel in Figure 1.1 displays boxplots for the variable sepal width corresponding to each type of iris, indicating that the mean value in each group could be different.



(a) (b)

Figure 1.1: (a) Scatter plot of sepal width over sepal length with the estimated regression lines under H_0 (dashed line) and under H_1 (whole line). (b) Boxplots for the sepal width data over the type of iris.

1.1.2 ANOVA model. Test for the equality of means

Consider the model in (1.4) with $\beta_i = 0 \; \forall \; i \in \{1, ..., I\}$, which means there is not a continuous predictor but there is a discrete one. Such model is known as ANOVA and it compares the mean values of the response variable across the I groups. The ANOVA model is then expressed as

$$Y_{ij} = \gamma_i + \varepsilon_{ij}, \quad i \in \{1, ..., I\}, \; j \in \{1, ..., n_i\},$$

where γ_i is the mean value of each group. The parameters can be estimated as the local means for each group:

$$\hat{\gamma}_i = \bar{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_I} Y_{ij}, \quad \forall i \in \{1, ..., I\}.$$

The primary goal of the ANOVA model is to test the equality of means and the F test can be used for this aim. The hypotheses are

$$\begin{aligned} &H_0 : \gamma_1 = ... = \gamma_I, \\ &H_1 : \gamma_i \neq \gamma_k, \quad \text{for some } i, k \in \{1, ..., I\}. \end{aligned}$$

The test statistic under H_0 is

$$\frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{Y}_{i\bullet} - \bar{Y})^2 / (I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2 / (n-I)} \sim F_{I-1, n-I},$$

where

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=i}^{n_I} Y_{ij}.$$

As exposed in the previous section, according to Figure 1.1b, one could think that not all three species have the same mean for the sepal width. In order to study this matter, the F test shown above is run, obtaining a p-value much lower than .01 and concluding, for this significance level, that the means for the three species are not the same.

1.1.3 ANCOVA model without interaction. Test of equality

In the previous section it was assumed that there was just a factor variable affecting the response. When both a continuous and a discrete predictor have an influence on the response, the resulting regression model is known as ANCOVA. It compares the values of the response variable across several groups in the presence of a covariate effect. Depending on the formulation, it will be an ANCOVA model with or without interaction. This section will focus on the model without interaction, which, assuming that the covariate effect is linear, corresponds to the model in (1.4) with $\beta_i = \beta_k \; \forall \; i, k \in \{1, ..., I\}$. The reason why this model is known as a non-interaction model is that the effects of both the discrete and the continuous variables are being considered, but they are simply summing up to each other, not interacting with the other.

The parameters γ_i must be estimated for each group, obtaining estimates $\hat{\gamma}_i$, $i \in \{1, ..., I\}$, while the estimate of the slope, $\hat{\beta}$ is the same for all groups. These estimates are obtained using partitioned regression (see Maxwell and Delaney, 2003, Chapter 9). Now, the interest lies on testing whether the regression lines are the same for all groups. Thus, the hypotheses considered are

$$\begin{aligned} &H_0 : Y_{ij} = \gamma + \beta X_{ij} + \varepsilon_{ij}, \quad \forall \; i \in \{1, ..., I\}, \\ &H_1 : Y_{ij} = \gamma_i + \beta X_{ij} + \varepsilon_{ij}, \quad \gamma_i \neq \gamma_k \text{ for some } i, k \in \{1, ..., I\}. \end{aligned}$$

The corresponding test statistic is

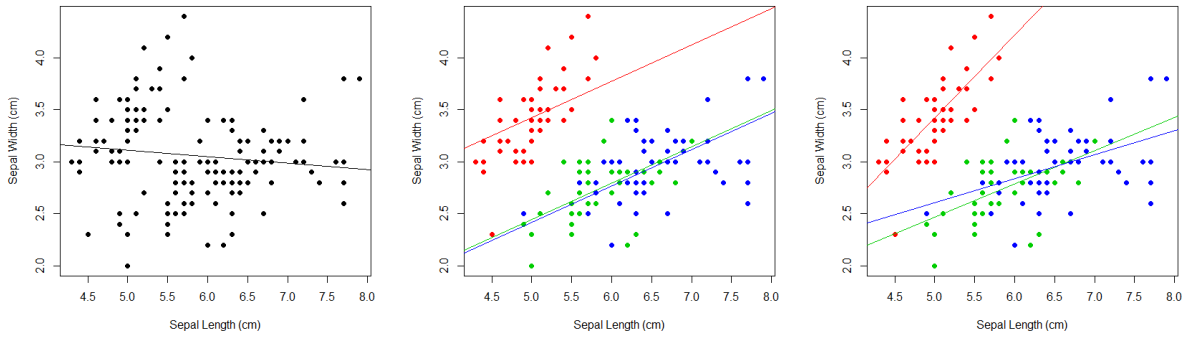
$$\frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-I-1)} \sim F_{I-1, n-I-1},$$

where the residual sums of squares under the null and the alternative hypotheses are

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma} - \hat{\beta}X_{ij})^2 \quad \text{and} \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}X_{ij})^2.$$

Under H_0 the estimators of the parameters are the ones given in (1.3), whereas under the alternative hypothesis the estimators are obtained with the partitioned regression technique mentioned above. The number of degrees of freedom under H_0 is $(n-2)$, as two parameters are being estimated. Under H_1, there are $(n-I-1)$ degrees of freedom because γ_i is estimated for each group in this case, so the number of degrees of freedom in the numerator of the statistic is $(I-1)$.

The test described above will be now applied to the iris data. Figure 1.2 shows three scatter plots of sepal width over sepal length. Figure 1.2a shows the regression lines obtained after fitting the model for all the data, as in H_0. Figure 1.2b shows the three regression lines obtained after fitting the non-interaction model (corresponding to H_1), with the same slopes but different intercepts for each group. The p-value obtained for the test is smaller than 0.001, so the null hypotheses is rejected (for the usual significance levels), leading to the conclusion that not all the intercepts are equal in the three groups.



(a) (b) (c)

Figure 1.2: Scatter plots of sepal width over sepal length. (a) Regression line obtained after fitting a simple regression model. (b) Dividing the population in the three classes and fitting the model without interaction. (c) Dividing the population in the three classes and fitting the model with interaction.

1.1.4 ANCOVA model with interaction. Test of equality

As it was noted before, the non-interaction model assumes that the effects for the continuous and the discrete predictors do not interact with each other. When this happens, the interaction model, corresponding to model (1.4), is considered. Thus, assuming interaction between the variables means that a change on the group will not only modify the intercept but also the slope.

As in the previous case, it is of interest to study the equality of the regression lines. For such test, the hypotheses are

$$H_0 : Y_{ij} = \gamma + \beta X_{ij} + \varepsilon_{ij}, \quad \forall\, i \in \{1, ..., I\},$$
$$H_1 : Y_{ij} = \gamma_i + \beta_i X_{ij} + \varepsilon_{ij}, \quad \gamma_i \neq \gamma_k \text{ or } \beta_i \neq \beta_k \text{ for some } i, k \in \{1, ..., I\}.$$

As for the estimation of the parameters, under the null hypothesis the estimates $\hat{\gamma}$ and $\hat{\beta}$ are obtained as in (1.3). The estimation under the alternative hypothesis is also simple, since it is only necessary

to fit I separate regression lines, one for each group, to estimate γ_i and β_i, $i \in \{1, ..., I\}$. A suitable test statistic for the test is

$$\frac{(RSS_0 - RSS)/(2I-2)}{RSS/(n-2I)} \sim F_{2I-2,n-2I},$$

where the residual sums of squares are

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma} - \hat{\beta}X_{ij})^2 \quad \text{and} \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}_i X_{ij})^2.$$

The degrees of freedom under the alternative hypothesis are $(n-2I)$, hence, the difference $(RSS_0 - RSS)$ has $(2I-2)$ degrees of freedom.

Hereafter, the test is applied to the iris data. The null hypothesis is the same as in the previous test (hence, the fitted model under the null is shown in Figure 1.2a), while the model fitted under the alternative hypothesis is represented in Figure 1.2c, fitting separate regression models for the different species. Unlike in the previous section with the non-interaction model, now it is allowed for each group to have different slopes, as well as different intercepts. The test for the interaction model is applied, obtaining a p-value smaller than the usual significance levels (.1, .05 and .01), which leads to the rejection of the null hypothesis: the regression lines are not equal for all three groups.

1.1.5 Test of parallelism

After having reviewed the non-interaction and the interaction model, it is natural to set out the question of whether there is an interaction between the two predictors. This issue can be explored by considering a formal test with the following hypotheses:

$$H_0 : Y_{ij} = \gamma_i + \beta X_{ij} + \varepsilon_{ij}, \quad \forall\ i \in \{1, ..., I\},$$
$$H_1 : Y_{ij} = \gamma_i + \beta_i X_{ij} + \varepsilon_{ij}, \quad \beta_i \neq \beta_k \text{ for some } i, k \in \{1, ..., I\}.$$

This test is also known as test of parallelism, since under the null hypothesis the regression lines are parallel. The corresponding statistic is

$$\frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-2I)} \sim F_{I-1,n-2I},$$

where

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}X_{ij})^2 \quad \text{and} \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}_i X_{ij})^2.$$

The parameters in RSS_0 are estimated with partitioned regression (as in Section 1.1.3 under H_1) and in RSS, I separate regression lines are fitted using the estimations in (1.3) for each line. Under H_0 there are $(n-I-1)$ degrees of freedom while under H_1 the number of degrees of freedom is $(n-2I)$. As a result, $(RSS_0 - RSS)$ has $(I-1)$ degrees of freedom.

Back to the iris data, the existence of interaction can now be determined. The model under the null hypothesis (without interaction) was represented in Figure 1.2b, while the model in the alternative hypothesis (with interaction) was represented in Figure 1.2c. The test is applied and yields a critical value close to zero, and thus rejecting the null hypothesis. Then, it is concluded that the regression lines do not have the same slope for the three groups, i.e., they are not parallel.

1.2 Nonparametric regression

In many situations the linear assumption described in (1.1) is not correct, since the relation between the variables is not necessarily linear. A more general model can be expressed as

$$Y_j = m(X_j) + \varepsilon_j, \quad j \in \{1, ..., n\}, \tag{1.5}$$

for any function m not necessarily linear. In the nonparametric context, for simplicity, just one predictor variable will be employed. The errors ε_j are independent and $N(0, \sigma)$. In some cases m can be a nonlinear parametric function, so nonlinear regression may be used to estimate m. Nevertheless, assumptions on the parametric shape of m could lead to a misspecification problem. Also, the estimation for the parameters might be difficult, especially when the number of parameters is large. Since numerical methods are needed for obtaining the estimators, one should also confront the problems inherent to optimization algorithms, such as starting points selections, tolerance thresholds and/or convergence.

An alternative route to avoid these difficulties is nonparametric regression, in which the only assumption needed is for m to be a sufficiently smooth function. Although many estimators have been proposed, kernel methods are a widely used alternative. Nadaraya (1964) and Watson (1964) proposed estimating m as

$$\hat{m}_h(x) = \frac{\sum_{j=1}^{n} K_h(X_j - x)Y_j}{\sum_{j=1}^{n} K_h(X_j - x)},$$

where $K_h(\cdot) = K(\cdot/h)/h$, with K being a kernel function (usually a symmetric around zero density) and h is known as the smoothing parameter. Generally, the standard normal density is used as a kernel.

For any point x from the support of X, the Nadaraya-Watson estimator is a weighted average of the responses, where the weights depend on the kernel function, assigning higher weights to the observations closer to x. On top of this interpretation, the Nadaraya-Watson estimator can also be thought as the result of fitting horizontal lines locally (in a neighborhood of x of length $2h$), assigning weights to each observation. Fan (1992) extended this concept resulting on the local linear estimator, which instead of fitting horizontal lines fits straight lines of the form $\beta_0 + \beta_1(\cdot - x)$. Again, each observation in a neighborhood of x is assigned a weight, given by the function K_h, which depends on the distance to x. The parameters β_0 and β_1 are estimated via weighted local least squares. The local linear estimator is given by $\hat{\beta}_0 = \hat{m}_h(x)$ (see Wand and Jones, 1995, Section 5.2). It can be proven that under some conditions the estimator[1] $\hat{m}(x)$ has asymptotically zero bias and variance (see Wand and Jones, 1995, Section 5.3 and Fan and Gijbels, 1996, Section 3.2). The selection of h is of great importance, since small values of h lead to less bias, but larger variance, while larger values of the smoothing parameter give rise to more bias and less variance. In addition, if $h \to 0$ the curve estimation tends to the interpolation of the data. On the contrary, as $h \to \infty$, $\hat{m}$ tends to a straight line, the same one as in least squares linear regression. An optimal value for h in terms of the Mean Integrated Squared Error (MISE) can be calculated:

$$h_{opt} = \left(\frac{R(K)\sigma^2}{\mu_2^2(K) \int [m''(x)]^2 f_X(x) dx} \right)^{1/5} n^{-1/5}, \tag{1.6}$$

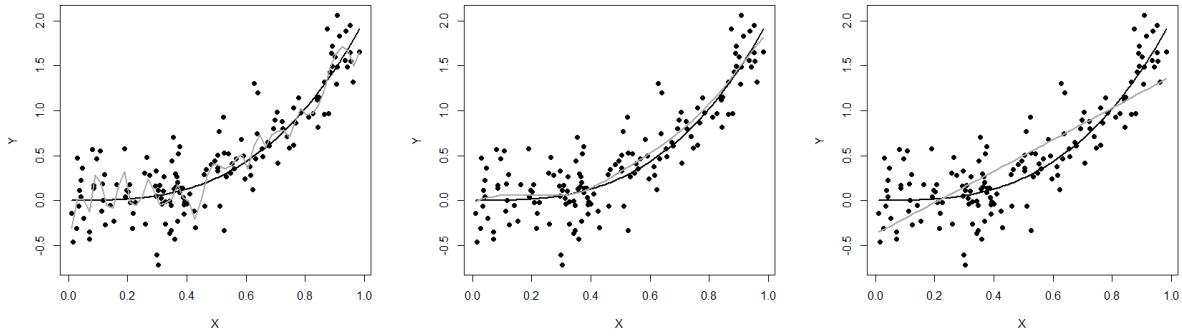
where f_X is the density function corresponding to the predictor variable X, $R(K) = \int K(z)dz$ and $\mu_2(K) = \int z^2 K(z)dz$. However, an optimal bandwidth cannot be used in practice, since it depends on the unknown quantities $m''(x)$ and $f_X(x)$. Several methods have been proposed in order to obtain a usable smoothing parameter in practice. Some of the most popular methods are cross-validation (proposed by Bowman (1984) in the estimation of the density function context), corrected AIC (Hurvich *et al.*, 1988) and the *plug-in* bandwidth (Ruppert *et al.*, 1995). As an example of the performance of the local linear estimator, Figure 1.3 shows simulated data from the model

$$Y_j = 2X_j^3 + \varepsilon, \quad j \in \{1, ..., 150\},$$

where the ε_j were drawn from a normal distribution with zero mean and standard deviation .25. The regression function was estimated with the local linear estimator using different values for the

[1]The local linear estimator depends on the smoothing parameter h and it is usually denoted as $\hat{m}_h$ but, from now on, in this Master thesis it will be denoted as $\hat{m}$ for simplicity.

smoothing parameter. In Figure 1.3a a small value of h was used, obtaining an undersmoothed estimation of the regression function. On the contrary, in Figure 1.3c a large value of h was selected, getting a straight line as the estimation of the cubic function. Lastly, in Figure 1.3b the bandwidth used was the one obtained through the cross-validation method, resulting in the estimation being close to the actual curve.



(a) Model A (b) Model B (c) Model C

Figure 1.3: Representations of 150 realizations of simulated data from a cubic model along with the true regression curves (black) and the estimations of the regression function (grey) using a small bandwidth in (a), the one selected by cross-validation in (b) and a large bandwidth in (c).

The pursued data-driven character of kernel methods makes it difficult to ascertain which features of the estimation correspond to the underlying regression function and which ones are just sample noise. As a graphical tool, Chaudhuri and Marron (1999) introduced the SiZer map, which allows to visualize the regions where the regression (or density) curve is significantly increasing or decreasing over a range of smoothing parameters. Therefore, the SiZer map can be a useful tool in order to determine the significance of the predictor variable. Nevertheless, it is important to formally asses the shape of the curve through hypothesis testing. For such purposes, Bowman and Azzalini (1997, Chapter 5) introduce a no-effect test in the nonparametric setting, which tries to analyze the significance of the predictor variable. This is the nonparametric alternative to the proposal in Section 1.1.1.

Additionally, when considering regression model (1.5) a discrete predictor variable may be also taken into account. In this setting, Young and Bowman (1995) propose an analysis of covariance model where the covariate effect is assumed only to be smooth. Two tests are considered: a test of equality, and a test of parallelism. The test of equality tries to analyze whether all groups follow the same regression function or not, while the test of parallelism tests if the curves are parallel, i.e., if they have the same shape but are separated by a constant shift.

Details on the three tests (no effect, equality and parallelism) will be presented in the next sections and illustrated with real data. In addition, a simulation study will be conducted to analyze the performance of the significance test and the simulation study in the original work of Young and Bowman (1995) will be partially replicated.

1.2.1 Significance test

When trying to examine the evidence of the response and explanatory variables being related, two competing models are considered. Thus, for the significance test the next hypotheses statement is used:

$$H_0 : Y_j = \gamma + \varepsilon_j, \quad \forall\, j \in \{1, ..., n\}$$
$$H_1 : Y_j = m(X_j) + \varepsilon_j, \quad m(X_j) \neq \gamma \text{ for some } j \in \{1, ..., n\}.$$

The errors ε_j are independent and $N(0,\sigma^2)$ distributed, and also independent from X. Under the null hypothesis γ is estimated as the sample mean of the responses and under H_1 the regression function is estimated with the local linear estimator. In order to construct the test statistic, Bowman and Azzalini (1997) proposed the following:

$$L_1 = \frac{RSS_0 - RSS}{RSS}, \tag{1.7}$$

with RSS_0 being the residual sum of squares under H_0 and RSS being the residual sum of squares under H_1:

$$RSS_0 = \sum_{j=1}^{n}(Y_j - \hat{\gamma})^2, \quad \text{and} \quad RSS = \sum_{j=1}^{n}(Y_j - \hat{m}(X_j))^2.$$

The statistic is a ratio between the difference of residual sums of squares and the residual sum of squares under H_1, and because of this the effect of the error variance σ^2 is scaled out. Now, the distribution of L_1 under the null hypothesis must be obtained. For such aim it should be noted that the two residual sums of squares can be expressed in vector-matrix notation:

$$RSS_0 = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{L})'(\boldsymbol{I}-\boldsymbol{L})\boldsymbol{Y} \overset{2}{=} \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{L})\boldsymbol{Y},$$

$$RSS_1 = \boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{S})'(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{Y},$$

with $\boldsymbol{Y}$ being the vector of the responses, $\boldsymbol{L}$ being a $n \times n$ matrix with n^{-1} in all its components and $\boldsymbol{S}$ being the smoothing matrix, i.e., a $n \times n$ matrix composed of known weights. Therefore, L_1 can be rewritten as

$$L_1 = \frac{\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}$$

where $\boldsymbol{A} = (\boldsymbol{I}-\boldsymbol{S})'(\boldsymbol{I}-\boldsymbol{S})$ and $\boldsymbol{B} = \boldsymbol{I}-\boldsymbol{L}-\boldsymbol{A}$. Now, the correspondent p-value for the test would be obtained as

$$\mathbb{P}\left(\frac{\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}} > Obs\right) = \mathbb{P}\left(\boldsymbol{Y}'(\boldsymbol{B}-Obs\boldsymbol{A})\boldsymbol{Y} > 0\right),$$

with Obs being the observed value of the statistic. Note that $\boldsymbol{Y}'(\boldsymbol{B}-Obs\boldsymbol{A})\boldsymbol{Y}$ is a quadratic form in normal variables where the matrix $\boldsymbol{B}-Obs\boldsymbol{A}$ is symmetric. Johnson and Kotz (1972) give a summary of general results about this type of variables, but they are more easily applied when the normal variables have zero expectation, which is not the case, since under the null hypothesis $\mathbb{E}(Y_j) = \gamma$. However, under H_0, $\boldsymbol{Y} = \boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\gamma}$ is a vector containing γ in all its components and $\boldsymbol{\varepsilon}$ is the vector containing all the errors. Thus,

$$\boldsymbol{Y}'(\boldsymbol{B}-Obs\boldsymbol{A})\boldsymbol{Y} = \boldsymbol{\gamma}'(\boldsymbol{B}-Obs\boldsymbol{A})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}'(\boldsymbol{B}-Obs\boldsymbol{A})\boldsymbol{\varepsilon},$$

and because of the construction of the matrices $\boldsymbol{B}$ and $\boldsymbol{A}$, the first term of the sum in the previous equation disappears[3]. Then,

$$p = \mathbb{P}(\boldsymbol{\varepsilon}'(\boldsymbol{B}-\boldsymbol{A}\cdot Obs)\boldsymbol{\varepsilon} > 0).$$

Now, $\boldsymbol{\varepsilon}'(\boldsymbol{B}-\boldsymbol{A}\cdot Obs)\boldsymbol{\varepsilon}$ is a quadratic form in normal variables of the type $\boldsymbol{z}'\boldsymbol{C}\boldsymbol{z}$, where $\mathbb{E}(\boldsymbol{z}) = 0$ and $\boldsymbol{C}$ is an $n \times n$ symmetric matrix. Although the results in Johnson and Kotz (1972) allow to calculate the probability p exactly trough numerical methods, a computational efficient way of obtaining it is approximating p by replacing the real distribution of $\boldsymbol{\varepsilon}'(\boldsymbol{Q}-\boldsymbol{B}\cdot Obs)\boldsymbol{\varepsilon}$ by another more convenient distribution with the same first three moments. Johnson and Kotz (1972, Chapter 29) show that the

[2]The proof of the identity becomes trivial when noting that $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{L}$.

[3]The proof easily follows from the fact that each row of the matrix $\boldsymbol{S}$ sums 1. Then, it is trivial that $\boldsymbol{S}\boldsymbol{\gamma} = \boldsymbol{\gamma}$, and thus $\boldsymbol{\gamma}'\boldsymbol{A}\boldsymbol{\gamma} = (\boldsymbol{\gamma}-\boldsymbol{S}\boldsymbol{\gamma})'(\boldsymbol{\gamma}-\boldsymbol{S}\boldsymbol{\gamma}) = \boldsymbol{0}$. In addition, $\boldsymbol{\gamma}'\boldsymbol{L}\boldsymbol{\gamma} = \boldsymbol{\gamma}'\boldsymbol{\gamma}$ leading to $\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma} = \boldsymbol{0}$.

s^{th} cumulant[4] of a quadratic form in normal variables $\boldsymbol{z'Cz}$ with $\mathbb{E}(\boldsymbol{z}) = \boldsymbol{0}$ and $\boldsymbol{C}$ a being symmetric matrix is given by

$$\nu_s = 2^{s-1}(s-1)!\text{tr}(\boldsymbol{VC})^s,$$

where tr denotes de trace operator and $\boldsymbol{V} = \text{Cov}(\boldsymbol{z}, \boldsymbol{z})$. In many problems involving quadratic forms, a shifted and scaled χ^2 distribution (i.e., a distribution $a\chi^2_b + c$, where a and c are, respectively, the shift and scale parameters and b is the number of degrees of freedom) was found to be a very good approximation (Solomon and Stephens, 1977 and Buckley and Eagleson, 1988). Hence, the first three moments of $\boldsymbol{\varepsilon'}(\boldsymbol{B} - \boldsymbol{A} \cdot Obs)\boldsymbol{\varepsilon}$ are matched to the first three moments of an $a\chi^2_b + c$ distribution, and then the parameters a, b and c are calculated as follows,

$$a = |\nu_3|/(4\nu_2), \quad b = (8\nu_2^3)/\nu_3^2, \quad c = \nu_1 - ab. \tag{1.8}$$

The p-value of interest can be approximated as $(1 - q)$, where q is the probability lying below the point $-c/a$ in a χ^2 distribution with b degrees of freedom. Note that the parameters a, b, and c are calculated with the cumulants, which depend on the $\boldsymbol{B}$ and $\boldsymbol{A}$ matrices, and on the observed value of the statistic. Therefore, these parameters depend on the data.

It is important to remark that for the implementation of the no-effect test it is necessary to select a smoothing parameter. The outcome of the test is very influenced by the bandwidth, and in practice, the test is usually carried out over a range of smoothing parameters. Nevertheless, the performance of the test will be studied in Section 1.2.4 considering several smoothing parameters.

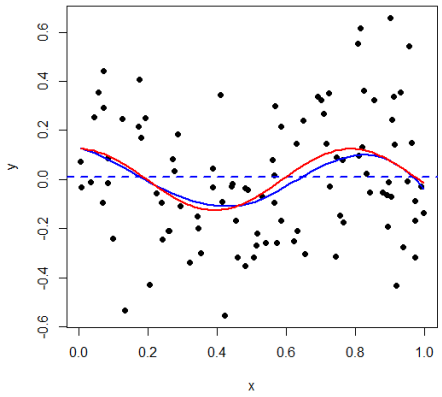


Figure 1.4: Scatter plot of simulated data with the true regression curve (red), the estimated regression curve through the local-linear estimator (blue, whole line) and the estimated regression function under the no effect hypothesis (blue, dashed line).

As an illustration, Figure 1.4 shows a representation of simulated data drawn from the model

$$Y_j = \frac{\cos(8X_j)}{8} + \varepsilon_j, \quad j \in \{1, ..., 100\},$$

where the errors ε_j are drawn from a Normal distribution with zero mean and standard deviation .25 and the predictors X_j are drawn from a $U(0, 1)$ distribution. The true regression curve is represented in the plot, as well as the local linear estimator and the estimation of the curve under the no effect

[4]The cumulants of a probability distribution are the Taylor coefficients at the origin of the cumulant generating function, which is the natural logarithm of the moment generating function. It is known that the first three cumulants match the first three central moments.

hypothesis, which corresponds to the mean of the responses. The significance test was applied to the simulated data over a range of smoothing parameters, and the results are displayed on Figure 1.5. For the nominal level $\alpha = .05$, there are evidences to reject the null hypothesis of no effect for the most part of the bandwidths considered. In addition, the smoothing parameters for which the test does not reject H_0 are implausible in practice, either to small or too large. Therefore, it can be concluded that the predictor variable is significant.

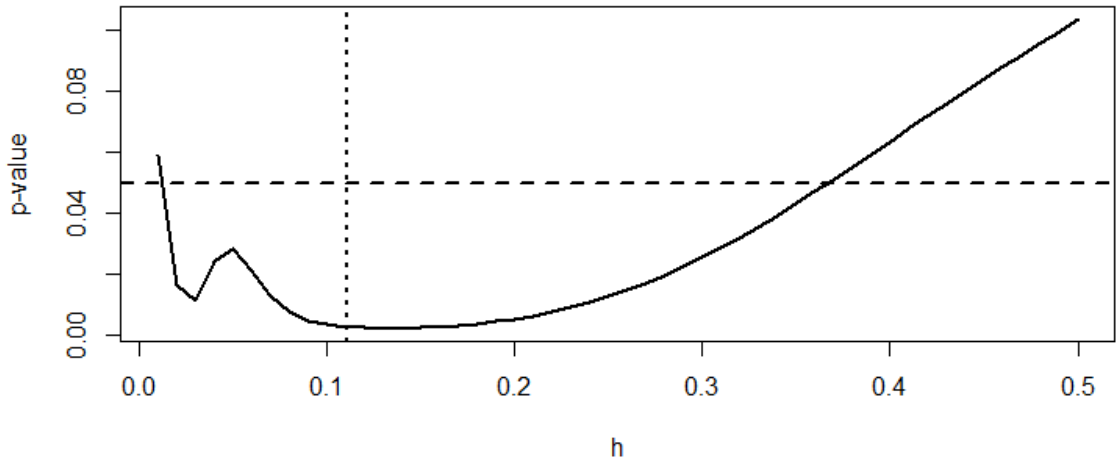


Figure 1.5: Trace of the significance test applied to the onion data. Dotted vertical line representing the bandwidth selected by cross-validation. Dashed horizontal line representing the p-value .05.

1.2.2 ANCOVA model. Test of equality

In an ANCOVA regression context, studying whether the regression relationship is the same for all groups is done trough an equality test. In the nonparametric setting the model is written as

$$Y_{ij} = m_i(X_{ij}) + \varepsilon_{ij}, \quad i \in \{1, ..., I\}, \quad j \in \{1, ..., n_i\},$$

for any smooth functions m_i and where the errors ε_{ij} are independent and $N(0, \sigma)$. The equality test is carried out with the following hypotheses statement:

$$\begin{aligned} &H_0 : Y_{ij} = m(X_{ij}) + \varepsilon_{ij}, \quad \forall\, i \in \{1, ..., I\}, \\ &H_1 : Y_{ij} = m_i(X_{ij}) + \varepsilon_{ij}, \quad m_i(\cdot) \neq m_k(\cdot) \text{ for some } i, k \in \{1, ..., I\}. \end{aligned}$$

Young and Bowman (1995) proposed the following test statistic for the equality test:

$$L_2 = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{m}_i(X_{ij}) - \hat{m}(X_{ij})]^2, \tag{1.9}$$

where $\hat{m}$ and $\hat{m}_i$ are the local linear estimators of m and m_i respectively. To estimate the variance in each group a first-difference approach, proposed by Rice (1984), can be used:

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i - 1)} \sum_{j=1}^{n_i - 1} [Y_{i[j+1]} - Y_{i[j]}]^2, \tag{1.10}$$

where $Y_{i[j]}$ denotes the value of Y corresponding to $X_{i[j]}$, with $X_{i[j]}$ being the j^{th} smallest value of X in the i^{th} group. Then, the global variance estimator is

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i-1)\hat{\sigma}_i^2.$$

Bowman and Azzalini (1997) noted that the differences $Y_{i[j+1]} - Y_{i[j]}$ are influenced by the shape of the underlying regression function, so this variance estimator will be inflated. The authors use an alternative method proposed by Gasser *et al.* (1986). They define the so-called pseudo-residuals as

$$\tilde{\varepsilon}_{i[j]} = \frac{X_{i[j+1]} - X_{i[j]}}{X_{i[j+1]} - X_{i[j-1]}}Y_{i[j-1]} + \frac{X_{i[j]} - X_{i[j-1]}}{X_{i[j+1]} - X_{i[j-1]}}Y_{i[j+1]} - Y_{i[j]}, \quad i \in \{1,...,I\},\ j \in \{2,...,n_i-1\}.$$

These pseudo-residuals measure the difference between each Y_{ij} and the line joining its two immediate neighbors, and they can be written as $\tilde{\varepsilon}_{i[j]} = a_{i[j]}Y_{i[j-1]} + b_{i[j+1]}Y_{i[j+1]} - Y_{i[j]}$, where the expressions for $a_{i[j]}$ and $b_{i[j]}$ can be directly obtained from the previous formula. Given that

$$\mathbb{E}(\tilde{\varepsilon}_{i[j]}^2) = (a_{i[j]}^2 + b_{i[j]}^2 + 1)\sigma^2 + O(n^{-2}),$$

the variance estimator for the ith group is defined as

$$\hat{\sigma}_i^2 = \frac{1}{n_i-2}\sum_{j=2}^{n_i-1}\frac{1}{c_{i[j]}^2}\tilde{\varepsilon}_{i[j]}^2, \tag{1.11}$$

where $c_{i[j]}^2 = a_{i[j]}^2 + b_{i[j]}^2 + 1$, $i \in \{1,...,I\}$, $j \in \{2,...,n_i-1\}$. Therefore, the global variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i-2)\hat{\sigma}_i^2.$$

In order to obtain the distribution of the test statistic L_2 it is important to note that both the numerator and the denominator can be expressed as quadratic forms in the data. For the ith group, the vector of fitted values $\{\hat{m}(X_{ij})\}_{j\in\{1,...,n_i\}}$ can be written in vector-matrix notation as $\hat{\boldsymbol{m}}_i = \boldsymbol{S}_i\boldsymbol{Y}_i$, where $\boldsymbol{Y}_i$ is the vector with the n_i observations of the response variable corresponding to the ith group and $\boldsymbol{S}_i$ is a $n_i \times n_i$ matrix of known weights. The vector with all of the fitted values, denoted by $\hat{\boldsymbol{m}}$, can be expressed as $\hat{\boldsymbol{m}} = \boldsymbol{S_d}\boldsymbol{Y}$, where $\boldsymbol{Y}$ denotes the vector with all of the observations from the response variable and $\boldsymbol{S_d}$ is a $n \times n$ block matrix, where each block corresponds to one of the I groups. Under the null hypothesis, where it is assumed that there is only one curve for all the groups, the vector of fitted values can be written as $\hat{\boldsymbol{m}} = \boldsymbol{S}\boldsymbol{Y}$, where $\boldsymbol{S}$ is a different $n \times n$ matrix of weights (the one corresponding to a global fit). Therefore, the numerator of (1.9) can be expressed as $\boldsymbol{Y'}[\boldsymbol{S_d} - \boldsymbol{S}]'[\boldsymbol{S_d} - \boldsymbol{S}]\boldsymbol{Y}$. For simplicity, $[\boldsymbol{S_d} - \boldsymbol{S}]'[\boldsymbol{S_d} - \boldsymbol{S}]$ will be denoted as $\boldsymbol{Q}$. Furthermore, independently of which one of the above variance estimators is considered, $\hat{\sigma}^2$ can be written as $\hat{\sigma}^2 = \boldsymbol{Y'}\boldsymbol{B}\boldsymbol{Y}$, where $\boldsymbol{B}$ is a $n \times n$ block matrix. If estimator (1.10) is used, the $\boldsymbol{B}$ matrix is composed of I $n_i \times n_i$ blocks of the form

$$(2n-2I)^{-1}\begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \dots & \dots & \dots & & \\ & & & \dots & \dots & \dots & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}.$$

On the other hand, if estimator (1.11) is applied, $\boldsymbol{B} = \boldsymbol{A'A}$, where $\boldsymbol{A}$ is a matrix composed of I $n_i \times n_i$ blocks, where the ith block is

$$(n-I)^{-1/2}\begin{bmatrix} 0 & 0 & & & & & & & 0 \\ a_{i[2]}/c_{i[2]} & -1/c_{i[2]} & b_{i[2]}/c_{i[2]} & & & & & & \\ & a_{i[3]}/c_{i[3]} & -1/c_{i[3]} & b_{i[3]}/c_{i[3]} & & & & & \\ & & \dots & \dots & \dots & & & & \\ & & & \dots & \dots & \dots & & & \\ & & & & a_{i[n_i-1]}/c_{i[n_i-1]} & -1/c_{i[n_i-1]} & b_{i[n_i-1]}/c_{i[n_i-1]} & & \\ 0 & & & & & & 0 & 0 & \end{bmatrix}.$$

Consequently, a p-value for the test can be obtained as follows:

$$p = \mathbb{P}\left(\frac{\boldsymbol{Y'QY}}{\boldsymbol{Y'BY}} > Obs\right),$$

where Obs is the observed value of the test statistic. Unfortunately, calculating the distribution of this statistic is complicated because, as in the significance test, the expectation of the $\boldsymbol{Y}$ under H_0 is different from zero, so the results in Johnson and Kotz (1972) cannot be applied. However, Young and Bowman (1995) argue that the distribution of L_2 is almost equivalent to the distribution of $\boldsymbol{\varepsilon'Q\varepsilon}/\boldsymbol{\varepsilon'B\varepsilon}$. A clarification on this follows.

In the first place, it is necessary to note that the bias of estimator $\hat{m}$ at any point x is given by

$$\text{bias}[\hat{m}(x)|X_1, ..., X_n] \approx \frac{1}{2}h^2 m''(x)\mu_2(K),$$

while the bias for the estimator $\hat{m}_i$ at an arbitrary point x is

$$\text{bias}[\hat{m}_i(x)|X_1, ..., X_n] \approx \frac{1}{2}h^2 m_i''(x)\mu_2(K).$$

Therefore, bias at any point x is controlled principally by the smoothing parameter h and the shape of the true curve. Then, if the curves are identical or at least have the same shape, both biases will be

identical if a common smoothing parameter is used (see Bowman and Azzalini, 1997, Section 6.4 for details). Consequently, when substituting $\boldsymbol{Y}$ by $\boldsymbol{m}+\boldsymbol{\varepsilon}$ in the numerator of L_2, the first term,

$$\boldsymbol{m}'\boldsymbol{Q}\boldsymbol{m} = (\boldsymbol{S_d}\boldsymbol{m} - \boldsymbol{S}\boldsymbol{m})'(\boldsymbol{S_d}\boldsymbol{m} - \boldsymbol{S}\boldsymbol{m}),$$

disappears asymptotically, since the means of $\hat{m}$ and $\hat{m}_i$ are approximately equal because of the bias properties shown above. On the other hand, when substituting $\boldsymbol{Y}$ by $\boldsymbol{m}+\boldsymbol{\varepsilon}$, the denominator of L_2 can be expressed as

$$\hat{\sigma} = \boldsymbol{Y'BY} = \boldsymbol{m'Bm} + \boldsymbol{\varepsilon'B\varepsilon}.$$

Because of the construction of the $\boldsymbol{B}$ matrices, the first term is very small relative to the second term, and it can be ignored. Therefore, the p-value calculation is (almost) equivalent to

$$p = \mathbb{P}\left(\frac{\boldsymbol{\varepsilon'Q\varepsilon}}{\boldsymbol{\varepsilon'B\varepsilon}} > Obs\right) = \mathbb{P}\left(\boldsymbol{\varepsilon'}(\boldsymbol{Q} - \boldsymbol{B}\cdot Obs)\boldsymbol{\varepsilon} > 0\right).$$

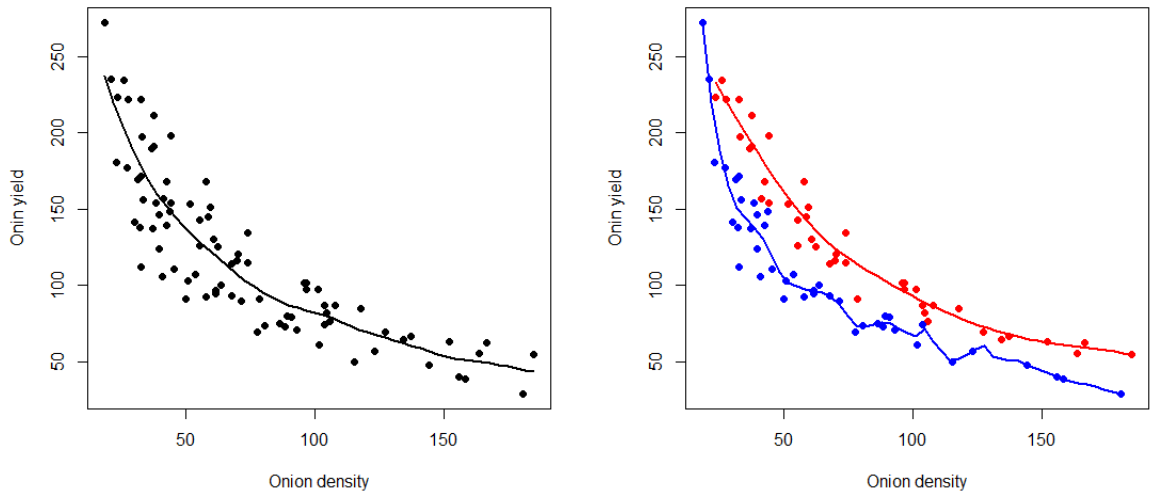
Now, $\boldsymbol{\varepsilon'}(\boldsymbol{Q} - \boldsymbol{B}\cdot Obs)\boldsymbol{\varepsilon}$ is a quadratic form in normal variables of the type $\boldsymbol{z'Cz}$, where $\mathbb{E}(\boldsymbol{z}) = 0$ and $\boldsymbol{C}$ is an $n \times n$ symmetric matrix. Therefore, the first three cumulants of the distribution can be calculated as

$$\nu_s = 2^{s-1}(s-1)!\mathrm{tr}(\boldsymbol{Q - B\cdot Obs})^s, \quad s = 1, 2, 3.$$

Finally, as in the no-effect test, the distribution is approximated to a shifted and scaled χ^2, where the parameters are calculated as a function of the cumulants as in equation (1.8).

As an illustration, the nonparametric equality test is applied to the data given in Ratkowsky (1983), available at R's package `agridat` (Wright, 2018). The dataset's name is `ratkowsy.onions` and it contains 84 sets of observations of white Spanish onion yields for different densities at two South Australian locations: Purnong and Virginia. Figure 1.6a shows a plot of the data, including the estimation of the regression curve via local linear regression using the cross-validation criterion for selecting the smoothing parameter. Figure 1.6b shows the same data, now adjusting the regression curve for each location.

The plots show that the regression functions might be different for each location, so it is necessary to use the test of equality. The test was applied obtaining a p-value of $3\cdot 10^{-4}$, rejecting the hypothesis of identical curves for the nominal level $\alpha = .05$. Since the outcome of the test is influenced by the smoothing parameter, it is recommended to run the test over a range of smoothing parameters. Figure 1.7 shows the p-values obtained in the equality test over a sequence of smoothing parameters. All values lie below .05, which clearly shows that the regression curves are different for this significance level.



(a) (b)

Figure 1.6: (a) Scatter plot of onion yield over onion density along with the estimated regression curve. (b) Same scatter plot as (a) with the data corresponding to Purnong represented in blue and data corresponding to Virgina represented in red. The estimated regression curves for each group are also shown.

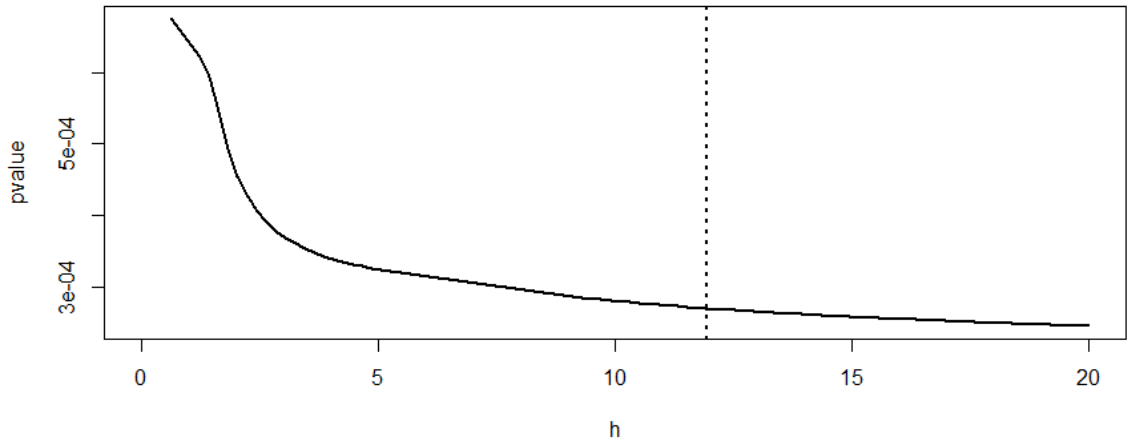


Figure 1.7: Trace of the equality test applied to the onion data. Dotted vertical line representing the bandwidth selected with the cross-validation criterion using all the data.

1.2.3 ANCOVA model. Test of parallelism

There may be cases where the regression curves for each of the groups are not the same but differ of each other just by constants. This assumption may be analyzed through the test of parallelism. In this case, the hypotheses considered are

$$H_0 : Y_{ij} = \gamma_i + m(X_{ij}) + \varepsilon_{ij}, \quad \gamma_1 = 0, \ \forall\ i \in \{1, ..., I\},$$
$$H_1 : Y_{ij} = m_i(X_{ij}) + \varepsilon_{ij}, \quad m_i(\cdot) \neq m_k(\cdot) + \gamma \text{ for some } i, k \in \{1, ..., I\} \text{ and } \forall\ \gamma \in \mathbb{R}.$$

In order to fit the model under the null hypothesis, the γ_i term must be estimated. For this aim, the model under the null hypothesis can be written in vector-matrix notation:

$$\boldsymbol{Y} = \boldsymbol{D\gamma} + \boldsymbol{m} + \boldsymbol{\varepsilon}, \tag{1.12}$$

where $\boldsymbol{\gamma}$ denotes the vector of parameters $(\gamma_2, ..., \gamma_I)'$ and $\boldsymbol{D}$ is a design matrix consisting on 0s and 1s. If the vector $\boldsymbol{\gamma}$ was known, an estimate of $\boldsymbol{m}$ could be constructed of the form

$$\hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{D\gamma}),$$

where $\boldsymbol{S}$ is a smoothing matrix. Substituting this expression into (1.12) and after some readjustment, the next equation is derived:

$$(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y} = (\boldsymbol{I_n} - \boldsymbol{S})\boldsymbol{D\gamma} + \boldsymbol{\varepsilon},$$

with $\boldsymbol{I}_n$ being the identity matrix of order n. Using the least squares method, the following estimator is obtained:

$$\hat{\boldsymbol{\gamma}} = [\boldsymbol{D'}(\boldsymbol{I}_n - \boldsymbol{S_1})'(\boldsymbol{I}_n - \boldsymbol{S_1})\boldsymbol{D}]^{-1}\boldsymbol{D'}(\boldsymbol{I}_n - \boldsymbol{S_1})'(\boldsymbol{I}_n - \boldsymbol{S_1})\boldsymbol{Y} = \boldsymbol{AY}, \tag{1.13}$$

where $\boldsymbol{S_1}$ is a preliminary smoothing matrix, different from the one used for the estimation of $\hat{\boldsymbol{m}}$. Speckman (1988) noted that estimator (1.13) has asymptotically normal distribution with negligible bias. Since the estimator involves the smoothing matrix $\boldsymbol{S_1}$, it therefore involves a smoothing parameter h_1. Bowman and Azzalini (1997, Section 6.5) recommend exploring several bandwidths, although a small smoothing parameter should be selected to minimize the bias in the estimation of $\hat{\boldsymbol{\gamma}}$. As a simple guideline they use the smoothing parameter $2R/n$, where R is the range of the design points. This bandwidth restricts the smoothing to approximately eight neighboring observations when the data are equally spaced (if a normal kernel is used).

Once an estimation of $\boldsymbol{\gamma}$ is obtained, it can be replaced in (1.12) and, thus, the vector of fitted values under the null hypothesis is derived as

$$\hat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{D}\hat{\boldsymbol{\gamma}}),$$

with $\boldsymbol{S}$ being the smoothing matrix using a given bandwidth for estimation (for example the one given by the cross-validation criterion). Under the alternative, the vector of fitted values is estimated as

$$\hat{\boldsymbol{m}} = \boldsymbol{S_d}\boldsymbol{Y},$$

where $\boldsymbol{S_d}$ is a block matrix as described in the equality test. The next statistic is then used for the parallelism test:

$$L_3 = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\hat{\gamma}_i + \hat{m}(X_{ij}) - \hat{m}_i(X_{ij})]^2, \tag{1.14}$$

with $\hat{\sigma}^2$ being either estimator (1.10) or (1.11). The numerator of this statistic can be expressed as a quadratic form in the data:

$$\boldsymbol{Y'}[\boldsymbol{DA} + \boldsymbol{S}(\boldsymbol{I}_n - \boldsymbol{DA}) - \boldsymbol{S_d}]'[\boldsymbol{DA} + \boldsymbol{S}(\boldsymbol{I}_n - \boldsymbol{DA}) - \boldsymbol{S_d}]\boldsymbol{Y},$$

therefore, its structure is of the type $\boldsymbol{Y'QY}$ and the results used for calculating the p-value in the previous tests can be applied again to derive an approximate distribution for the statistic under H_0, which is a shifted and scaled χ^2 distribution with the parameters obtained as in (1.8). Note that, as in the equality test, the estimators $\hat{m}$ and $\hat{m}_i,\ i \in \{1, ..., I\}$ must be obtained using the same smoothing parameter, so that the bias is canceled out.

The test is now applied to the onion data employed in the previous section. It could be natural that the shape of the regression curve was the same for the two locations but the yield in one of the locations was always larger than in the other one. Therefore it is of interest to investigate if the curves are parallel, for which the parallelism test is used. When selecting the smoothing parameter with the cross-validation criterion, a p-value of .0277 is obtained for the test, rejecting the null hypothesis for $\alpha = .05$. Since the outcome of the test depends on the bandwidth, Figure 1.8 shows the significance trace of the test over a sequence of smoothing parameters. It is shown that for all values of the smoothing parameter except for the lowest ones (which are very unrealistic given the sample size), the p-value is below .05, leading to the rejection of the parallelism hypothesis for this value of α.

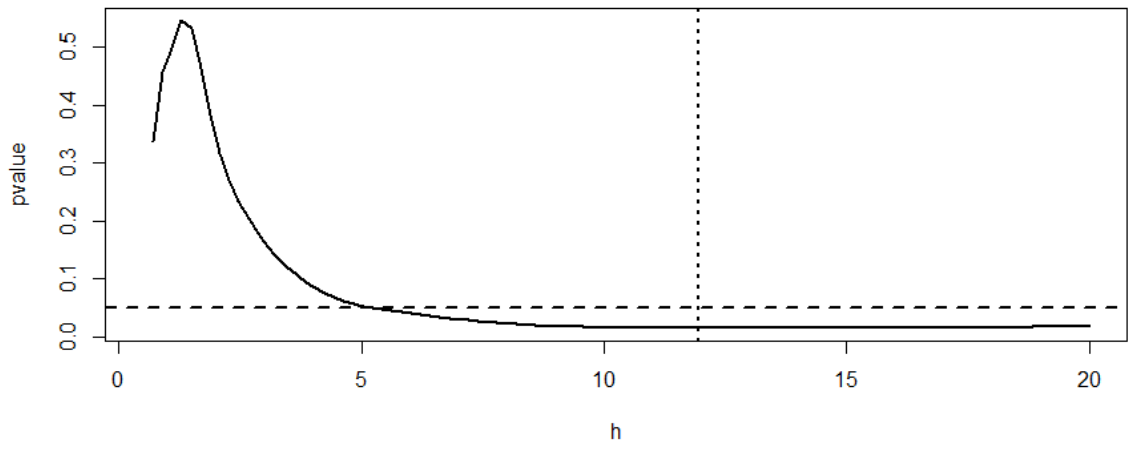


Figure 1.8: Trace of the parallelism test applied to the onion data. Dotted vertical line representing the bandwidth selected with the cross-validation criterion using all the data. Dashed horizontal line represents $\alpha = .05$.

1.2.4 Simulation study

Up to the author's knowledge, there is not a simulation study of the significance test presented in Section 1.2.1 in the statistical literature. For that reason a brief simulation study of the no-effect test will be carried out in this section. The performance of the test will be analyzed using several smoothing parameters in order to study the dependence of the test on the bandwidth. Regarding the two revisited ANCOVA tests (equality and parallelism), a complete simulation study can be found in Young and Bowman (1995). Part of this simulation study is redone in the present section.

Signification test

In the following, the power and calibration of the no-effect test will be studied through a brief simulation study. The computation was done with `R`'s package `sm` (Bowman and Azzalini, 2018), where the test

is implemented. Two models will be considered:

A. $Y = 2 - \beta X^2 + \varepsilon, \quad \beta = 0,\ .05,\ .1$

B. $Y = \beta \sin(10X) + \varepsilon, \quad \beta = 0,\ .1,\ .15$

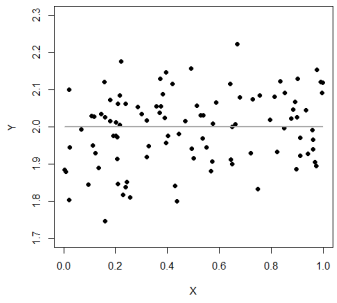
where $\varepsilon \sim N(0, \sigma)$, and σ takes different values for each model. The first value of β corresponds to the null hypothesis being true, and when the other values of β are used, the alternative hypothesis holds. The sample size takes values in $\{50, 100, 250, 400\}$. Realizations of data drawn from the models and the true regression functions are represented in Figure 1.9 for all the values of β.

The test was applied to 500 replications of the simulated data and the percentages of rejection for the nominal level $\alpha = .05$ were recorded. The dependence of the test on the smoothing parameter was investigated by considering different values of the bandwidth: cv, $\frac{1}{2}cv$, $2cv$, $3cv$ and $4cv$, where cv is the parameter selected by cross-validation.

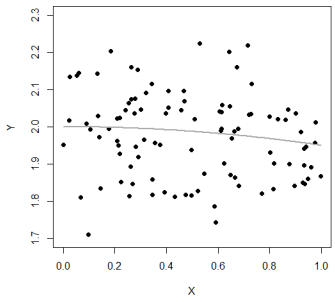
Model A Table 1.1 contains results for model A. Under the null hypothesis the test with bandwidth selected by cross-validation (cv) obtains percentages of rejection much larger than the nominal level $\alpha = .05$ (around 9% or 10%, for instance). When considering $\frac{1}{2}cv$, the results are close to α, and when using multiples of cv, the larger the smoothing parameter, the closer the percentages of rejection are to α. On the other hand, under the alternative hypothesis, the percentages of rejection tend to 1 as the data size increases, at least when $\beta = .1$. However, when using $\frac{1}{2}cv$ as the smoothing parameter, the percentages of rejection are not so large as with the other parameters. Note that when using large bandwidths the estimation of the curve will be oversmoothed, and therefore it will be close to a straight line, but since this line will not have zero slope, the test can still detect the significance of the predictor variable. Therefore, even for large bandwidths as $4cv$, the percentages of rejection under H_1 are still high.

Model B Results for model B are displayed in Table 1.2. Under H_0 the percentages of rejection obtained with the cv bandwidth are much larger than α and, with the largest bandwidths, the percentages of rejection tend to α. Nevertheless, as before, using $\frac{1}{2}cv$ results close to .05 are obtained. Under the alternative hypothesis, the largest percentages of rejection are obtained with the cv bandwidth, and increasing the smoothing parameter has the effect of obtaining smaller percentages of rejection. This behavior is due to the shape of the regression function under H_1, since now when using very large bandwidths the estimation of the curve will tend to a straight line with zero slope, so with large bandwidths as $4cv$ the test is not able to detect the effect of the predictor variable as often as with other parameters. On the other hand, with a small bandwidth like $\frac{1}{2}cv$, the percentages of rejection are quite high in comparison.

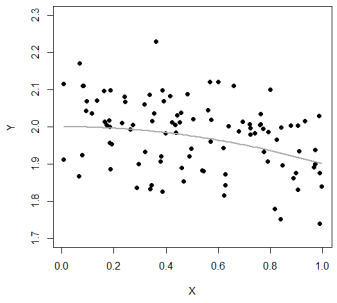
As a general conclusion on the bandwidth, given the simulation's results, a good choice in practice would be to use $\frac{1}{2}cv$ as a bandwidth. With this smoothing parameter the test is well calibrated and the power of the test is relatively high compared to the results obtained for other bandwidths.



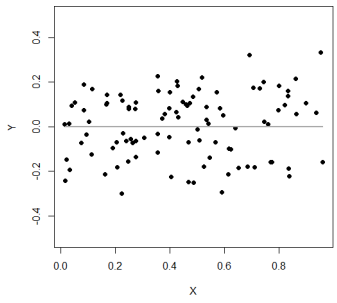
(a) Model A. $\beta = 0$



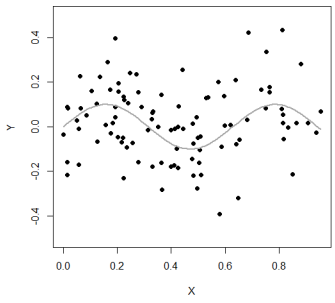
(b) Model A. $\beta = .05$



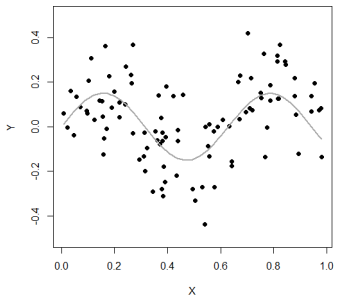
(c) Model A. $\beta = .1$



(d) Model B. $\beta = 0$



(e) Model B. $\beta = .1$



(f) Model B. $\beta = .15$

Figure 1.9: Scatter plots of simulated data drawn from models A (first row) and B (second row) with $n = 100$, along with the true regression functions. The standard deviation is $\sigma = .1$ in Model A and $\sigma = .15$ in Model B.

Test of no effect. Model A															
σ	n	$\frac{1}{2}cv$			cv			$2cv$			$3cv$			$4cv$	
		$\beta = 0$	$\beta = .05$	$\beta = .1$	$\beta = 0$	$\beta = .05$	$\beta = .1$	$\beta = 0$	$\beta = .05$	$\beta = .1$	$\beta = 0$	$\beta = .05$	$\beta = .1$	$\beta = 0$	$\beta = .05$	$\beta = .1$
.05	50	.068	.374	.924	.130	.524	.968	.076	.502	.966	.056	.480	.964	.052	.474	.968
	100	.050	.738	1	.080	.834	1	.050	.818	1	.038	.812	1	.032	.810	1
	250	.042	.970	1	.084	.992	1	.050	.818	1	.038	.812	1	.032	.810	1
	400	.064	1	1	.092	1	1	.064	1	1	.044	1	1	.038	1	1
.1	50	.056	.130	.416	.092	.186	.560	.066	.168	.548	.054	.158	.526	.050	.156	.530
	100	.056	.204	.742	.102	.316	.816	.062	.290	.808	.040	.270	.802	.036	.264	.800
	250	.042	.512	.992	.070	.628	1	.040	.634	1	.030	.628	.998	.030	.628	.998
	400	.058	.724	1	.086	.820	1	.062	.822	1	.048	.822	1	.046	.824	1

Table 1.1: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test in Model A based on 500 simulations.

Test of no effect. Model B															
σ	n	$\frac{1}{2}cv$			cv			$2cv$			$3cv$			$4cv$	
		$\beta=0$	$\beta=.1$	$\beta=.15$	$\beta=0$	$\beta=.1$	$\beta=.15$	$\beta=0$	$\beta=.1$	$\beta=.15$	$\beta=0$	$\beta=.1$	$\beta=.15$	$\beta=0$	$\beta=.1$	$\beta=.15$
.1	50	.054	.840	.982	.106	.932	.998	.074	.846	.988	.058	.644	.930	.056	.440	.824
	100	.050	.998	1	.092	1	1	.074	.998	1	.056	.962	1	.052	.876	1
	250	.052	1	1	.104	1	1	.070	1	1	.060	1	1	.060	1	1
	400	.062	1	1	.106	1	1	.088	1	1	.074	1	1	.072	1	1
.15	50	.044	.390	.834	.086	.598	.936	.060	.426	.826	.054	.244	.628	.044	.166	.428
	100	.056	.862	1	.094	.920	1	.058	.838	.996	.050	.628	.970	.040	.468	.876
	250	.042	1	1	.086	1	1	.084	1	1	.060	.990	1	.052	.954	1
	400	.060	1	1	.096	1	1	.076	1	1	.058	1	1	.054	.998	1

Table 1.2: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test in Model B based on 500 simulations.

Replication of original simulation study for ANCOVA

In order to study the performance of the nonparametric ANCOVA tests described in this chapter, the simulation study in Young and Bowman (1995) was replicated using the function `sm.ancova` from the `sm`[5] package (Bowman and Azzalini, 2018). Three sets of design points were used for obtaining the values in the independent variable:

- Both groups equally spaced on the interval $(0, 1)$.
- Both groups identical, as determined by a single random sample from a $U(0, 1)$ distribution.
- Each group determined by a different sample from a $U(0, 1)$ distribution.

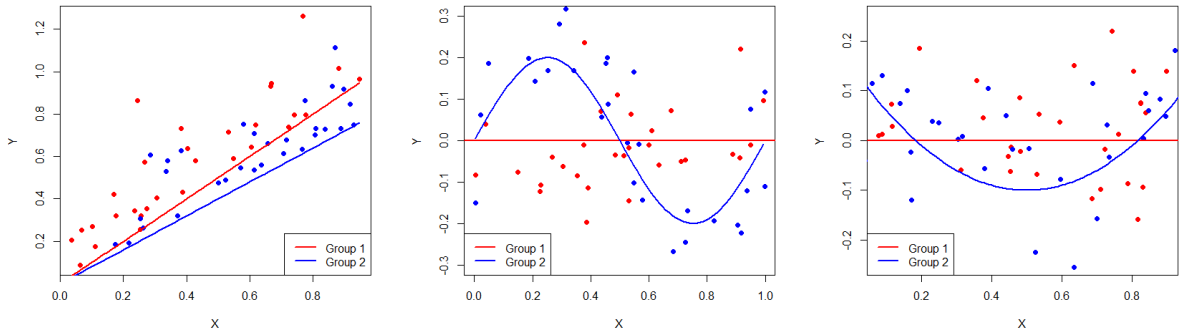
The different regression relationships used were

A. Group 1: $Y = X$. Group 2: $Y = \beta X, \quad \beta = 1,\ .9,\ .8$

B. Group 1: $Y = 0$. Group 2: $Y = \beta \sin(2\pi X), \quad \beta = 0,\ .1,\ .2$

C. Group 1: $Y = 0$. Group 2: $Y = \beta(X^2 - X + 0.15), \quad \beta = 0,\ .5,\ 1$

When the first value of β is used, the null hypothesis holds, while when considering the other two values of β, the alternative hypothesis is true.

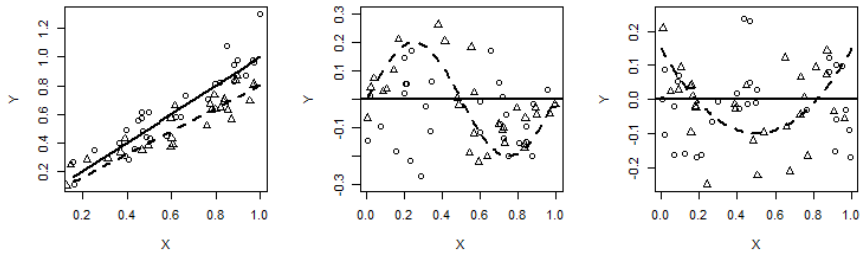


(a) Model A (b) Model B (c) Model C

Figure 1.10: Representations of simulated data from models A, B and C under the alternative hypothesis of different regression curves ($\beta = .8$ in A, $\beta = .2$ in B and $\beta = 1$ in C), along with the true regression curves. Number of observations is 30 for each group and $\sigma = .1$.

For the test of parallelism another parameter $\gamma = .05$ was added to the responses for the second group, describing the shift between the curves for each group. The number of simulated observations was 60, with both groups having 30 observations each, as in the original simulation study. The errors follow a $N(0, \sigma)$ distribution, with two different values for the standard deviation: .05 and .1. Simulated data from all three models and the true regression curves under H_1 are displayed in Figure 1.10. The smoothing parameter was obtained through cross-validation and the nominal level considered was $\alpha = .05$. Along with the percentages of rejection for the nonparametric tests (NP), the percentages of rejection for the linear ANCOVA tests (L) were obtained. Results were determined using 500 simulations for each combination. Results for the test of equality are displayed in Table 1.3

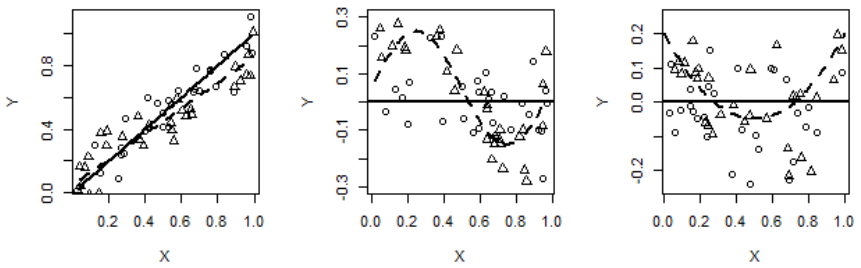
Test of equality



σ		A			B			C		
		$\beta = 1$	$\beta = .9$	$\beta = .8$	$\beta = 0$	$\beta = .1$	$\beta = .2$	$\beta = 0$	$\beta = .5$	$\beta = 1$
		Equally spaced design								
.05	NP	.056	.928	1	.052	.956	1	.048	.568	1
	L	.062	.986	1	.050	.838	1	.038	.020	.004
.1	NP	.032	.324	.950	.062	.338	.968	.056	.128	.592
	L	.052	.484	.972	.058	.328	.840	.080	.044	.038
		$U(0,1)$ design (same for both groups)								
.05	NP	.050	.912	1	.036	.948	1	.046	.458	.964
	L	.048	.986	1	.044	.856	1	.044	.074	0.12
.1	NP	.052	.366	.910	.042	.368	.978	.038	.182	.506
	L	.072	.491	.978	.050	.410	.808	.048	.032	.042
		$U(0,1)$ design (different for each group)								
.05	NP	.048	.970	1	.050	.994	1	.060	.434	.984
	L	.066	.998	1	.046	.966	1	.048	.246	.038
.1	NP	.042	.346	.852	.048	.464	.970	.052	.138	.492
	L	.056	.540	.976	.052	.528	.982	.050	.030	.036

Table 1.3: Percentages of rejections (for $\alpha = .05$) for the nonparametric test of equality (NP) and the linear model test (L) in Models A, B and C based on 500 simulations and with $n_i = 30$, for $i = 1, 2$.

Test of parallelism



		A			B			C		
σ		$\beta = 1$	$\beta = .9$	$\beta = .8$	$\beta = 0$	$\beta = .1$	$\beta = .2$	$\beta = 0$	$\beta = .5$	$\beta = 1$
		Equally spaced design								
.05	NP	.044	.364	.956	.054	.974	1	.048	.628	.994
	L	.064	.642	.992	.046	.948	1	.050	.020	.002
.1	NP	.056	.104	.412	.060	.436	.990	.052	.222	.644
	L	.058	.202	.608	.054	.460	.938	.058	.054	.016
		$U(0,1)$ design (same for both groups)								
.05	NP	.042	.360	.932	.048	.972	1	.038	.412	.998
	L	.044	.590	.984	.068	.960	1	.038	.022	.054
.1	NP	.042	.106	.412	.042	.554	.990	.050	.120	.574
	L	.046	.188	.640	.038	.562	.988	.042	.050	.058
		$U(0,1)$ design (different for each group)								
.05	NP	.048	.342	.932	.050	.974	1	.048	.526	.994
	L	.048	.574	.996	.038	.920	1	.044	.024	.012
.1	NP	.038	.118	.434	.046	.432	.990	.050	.160	.512
	L	.044	.192	.646	.050	.572	.988	.064	.064	.034

Table 1.4: Percentages of rejections (for $\alpha = .05$) for the nonparametric test of parallelism (NP) and the linear model test (L) in Models A, B and C based on 500 simulations and with $n_i = 30$, for $i = 1, 2$.

and results for the test of parallelism are presented in Table 1.4. A discussion on the results for each model is given next.

Model A Since the model considered has a linear shape, it is expected that the linear ANCOVA tests perform better than the nonparametric ones. When testing equality, both the linear and the nonparametric tests present percentages of rejection close to the nominal level $\alpha = .05$ under the null hypothesis. Under H_1, percentages of rejection are close to 1 for $\sigma = .05$, and the percentages are lower when the error variance increases. Although the results obtained for the linear and the nonparametric tests are similar, the linear test always obtains percentages of rejection slightly higher than the nonparametric test. As noted before, this is an expected behavior of the tests given that the underlying shape of the curves is linear.

As for the tests of parallelism, both of them obtain percentages of rejection close to .05 under H_0. On the other hand, under both variations of the alternative hypothesis, the percentages of rejection are lower than in the equality case. Again, the performance of the tests is worse when the error variance increases, since it is more difficult to capture the differences between the groups. In addition, in this scenario the linear test also obtains slightly higher percentages of rejection under H_1, as in the equality setting.

Model B Although the parametric assumption is not correct for this model, the parametric tests (both equality and parallelism) still perform well under H_0 (given that under the null hypothesis the linear assumption is correct), obtaining percentages of rejection oscillating around $\alpha = .05$. Results for the nonparametric equality and parallelism tests under the null hypothesis are also close to the nominal level.

Under the alternative hypotheses ($\beta = .1$ and $\beta = .2$) results for equality and parallelism are similar: when $\sigma = .05$ the percentages of rejection are close to 1 for the parametric and the nonparametric tests, and for $\sigma = .1$, as expected, the percentages of rejection are lower. However, although in Model A the parametric tests obtained rather higher results than their nonparametric counterparts under H_1, now both tests obtain similar results, which is at first surprising given that under the alternative hypothesis one of the regression functions is not linear. The reason why the test for linear still works well here is that it fits the model with two different lines, so although the shape of the fitted regression functions is not correct, the test is able to reject H_0 when it is false.

Model C Under H_0 the nonparametric tests perform well, obtaining results close to α both in the equality and the parallelism settings. The linear tests also obtain percentages of rejection close to the nominal level (note that under H_0 the regression function is a horizontal line, so the linear assumption is correct).

However, under H_1 one of the underlying regression curves is a quadratic function, far away from the linear assumption in the parametric tests, and when fitting a linear model, the two fitted regression lines are similar. Therefore, the parametric tests obtain percentages of rejection close to zero, which shows that these linear tests are not capable of detecting the differences between the two groups. On the other hand, the results for the nonparametric tests are close to one when $\sigma = .05$. Again, when increasing the value of σ, the percentages of rejection are lower.

In conclusion, the parametric test is well calibrated but it fails when trying to capture the differences between the groups. Even when the test is able to detect such differences, the shape of the estimated regression curve is far from the real model, unsuccessfully modeling the regression function. On the other hand, the nonparametric test seems to be well calibrated and it is able to reject H_0 when it is false, whichever the shape of the regression function.

[5]The `sm` library uses, by default, the variance estimator proposed by Gasser *et al.* (1986).

In this section it was assumed that the errors follow a normal distribution with constant variance. Now, the performance of the equality and parallelism tests when the normality assumption does not hold will be studied. For this aim, the same models considered in the original simulation study by Young and Bowman (1995) will be used, as well as the data size. Instead of the normal distribution, the errors will be drawn from the following distributions:

- Exponential distribution: $\varepsilon \sim \text{Exp}(\lambda), \quad \lambda = 7, 10$
- Student's T distribution: $\varepsilon \sim T_k, \quad k = 15, 30$

The first distribution is skewed, whereas the second one is heavy tailed. Percentages of rejection are displayed in Table 1.5 for the test of equality and in Table 1.6 for the test of parallelism. A brief discussion on them follows.

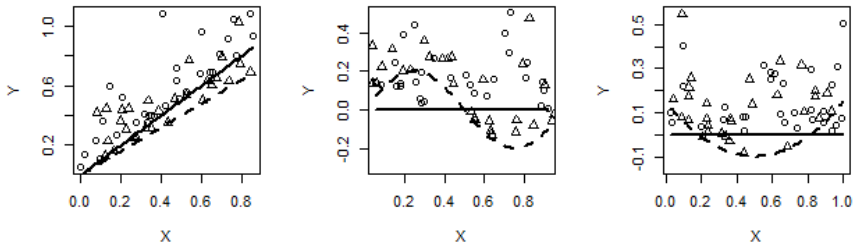
Exponentially distributed errors Under the null hypothesis both the test of equality and the test of parallelism obtain some percentages of rejection quite higher than the nominal level $\alpha = .05$, even reaching or surpassing .09. On the other hand, under the alternative hypothesis both tests are capable of detecting the differences between the groups, although the power is not certainly high.

***T* distributed erros** When the errors are drawn from a T distribution, under the null hypothesis the percentages of rejection are closer to the nominal level than when considering exponential errors, although some of the results are moderately high compared to α. On the other hand, under the alternative hypothesis the tests of equality and parallelism are unable to capture the differences between the groups, leading to very low percentages of rejection (lower than 9% in all cases).

The problem of studying the equality of I regression curves with non-normal or heteroscedastic errors was approached by Dette and Neumeyer (2001), who also worked with statistic L_2 defined in (1.9). These authors used statistic L_2 to construct a bootstrap version of the test, affirming that the test obtained a good performance under heteroscedasticity and non-normality even for small sample sizes ($n_i \in \{10, 20, 30, 50\},\ i = 1, 2$).

Tests of equality & parallelism

Exponentially distributed errors

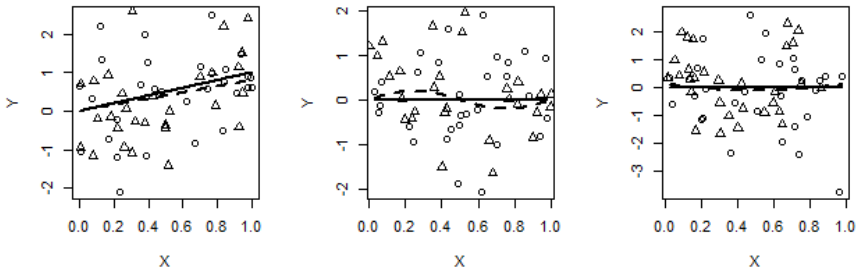


	λ	A			B			C		
		$\beta = 1$	$\beta = .9$	$\beta = .8$	$\beta = 0$	$\beta = .1$	$\beta = .2$	$\beta = 0$	$\beta = .5$	$\beta = 1$
					Equally spaced design					
Eq	7	.040	.236	.736	.096	.228	.662	.062	.084	.240
	10	.090	.474	.943	.062	.414	.906	.068	.132	.430
Par	7	.064	.106	.376	.046	.282	.756	.064	.082	.282
	10	.078	.198	.584	.048	.484	.948	.060	.160	.510
					$U(0,1)$ design (same for both groups)					
Eq	7	.068	.254	.734	.034	.280	.696	.052	.108	.194
	10	.054	.454	.936	.044	.462	.950	.064	.126	.422
Par	7	.066	.118	.324	.056	.274	.754	.044	.094	.246
	10	.062	.194	.574	.046	.516	.970	.068	.152	.464
					$U(0,1)$ design (different for each group)					
Eq	7	.076	.258	.688	.070	.244	.706	.050	.090	.194
	10	.074	.448	.928	.056	.436	.920	.064	.100	.352
Par	7	.042	.114	.284	.056	.312	.798	.050	.082	.222
	10	.064	.184	.544	.068	.502	.950	.080	.144	.480

Table 1.5: Percentage of rejections (for $\alpha = .05$) for the nonparametric tests of equality (Eq) and parallelism (Par) for Models A, B and C with exponential erros, based on 500 simulations and with $n_i = 30$, for $i = 1, 2$.

Tests of equality & parallelism

T distributed errors



	k	A			B			C		
		$\beta = 1$	$\beta = .9$	$\beta = .8$	$\beta = 0$	$\beta = .1$	$\beta = .2$	$\beta = 0$	$\beta = .5$	$\beta = 1$
					Equally spaced design					
Eq	$k = 15$	.080	.058	.064	.044	.062	.074	.058	.052	.058
	$k = 30$	.054	.068	.076	.052	.048	.060	.048	.050	.066
Par	$k = 15$	.060	.070	.060	.062	.056	.086	.042	.038	.042
	$k = 30$	.066	.060	.066	.062	.060	.076	.044	.066	.060
					$U(0,1)$ design (same for both groups)					
Eq	$k = 15$	.046	.070	.058	.068	.062	.090	.060	.046	.070
	$k = 30$	.050	.050	.072	.034	.052	.056	.058	.062	.064
Par	$k = 15$	.058	.068	.066	.064	.056	.084	.064	.066	.066
	$k = 30$	.064	.068	.048	.070	.060	.066	.028	.076	.054
					$U(0,1)$ design (different for each group)					
Eq	$k = 15$	.052	.062	.074	.054	.050	.046	.062	.060	.056
	$k = 30$	.036	.070	.068	.052	.064	.068	.062	.072	.068
Par	$k = 15$	.044	.048	.056	.046	.048	.054	.040	.062	.070
	$k = 30$	.056	.046	.066	.050	.064	.070	.062	.062	.058

Table 1.6: Percentage of rejections (for $\alpha = .05$) for the nonparametric tests of equality and parallelism for Models A, B and C with T-distributed errors, based on 500 simulations and with $n_i = 30$, for $i = 1, 2$.

Chapter 2

Some background on circular data

Circular data refers to observations that can be represented as points on the circumference of the unit circle. Once a zero direction and a sense of rotation have been chosen, these observations can be expressed as angles. Most examples of this type of data come from compass or clock measurements, such as directions or observations with time or calendar effects. Circular data can be found in many different fields: biology (orientation of the red wood ants in reaction to different stimuli; Jander (1957)), geology (cross-beds; SenGupta and Rao (1966) and Mardia (1972)), environmetrics and oceanography (wind and waves directions; Jona-Lasinio *et al.* (2012) and Oliveira *et al.* (2014a)), medicine (sudden infant death syndrome; Mooney *et al.* (2003)) or ecology (wildfire occurrences in the Iberian Peninsula; Ameijeiras-Alonso *et al.* (2019)).

Due to the periodicity of the data, classical statistical techniques might not be suitable to handle circular observations, so different methods must be applied to this kind of data. Jammalamadaka and SenGupta (2001), Pewsey *et al.* (2013) and Ley and Verdebout (2017) present a detailed review of circular statistics. Also, Crujeiras (2017) presents a review of different statistical methods for this kind of data, including nonparametric techniques. Circular data can also be viewed as a particular case of directional data, defined in a hypersphere of arbitrary dimension (see Mardia and Jupp, 2000). Hence, methods for spherical data can be adapted to the circle.

This chapter aims also to give some background on circular data: Section 2.1 gives an introduction to the main population and sample measures for circular data. Section 2.2 reviews the most important parametric circular distributions. In Section 2.3, parametric regression models for this kind of data will be presented. Finally, in Section 2.4 parametric ANOVA and ANCOVA models for circular data are revisited.

2.1 Population and sample measures

The present section will be devoted to the study of different measures of a circular population, such as location, concentration and dispersion measures, as well as their sample analogues. For that aim it is necessary to define, first, the concept of a circular density function.

When considering a circular random variable Θ with support in $[0, 2\pi)$, a function f will be a circular density if it verifies:

Condition 1. $f(\theta) \geq 0$ for $\theta \in [0, 2\pi)$,

Condition 2. $f(\theta + 2\pi k) = f(\theta)$ for $\theta \in [0, 2\pi)$ and $k \in \mathbb{Z}$,

Condition 3. $\int_0^{2\pi} f(\theta) d\theta = 1$.

Therefore, the main difference with a linear density function is the addition of Condition 2.

Population measures

Consider the circular variable Θ with density function f. Its first cosine and sine moments are defined, respectively, as

$$C = \mathbb{E}[\cos\Theta] = \int_0^{2\pi} \cos(\theta) f(\theta) d\theta \quad \text{and} \quad S = \mathbb{E}[\sin\Theta] = \int_0^{2\pi} \sin(\theta) f(\theta) d\theta.$$

The basic circular measure of location is the population mean direction, which is defined as

$$\mu = \text{atan2}(S, C),$$

where $\text{atan2}(S, C)$ is an operator returning the angle between the x-axis and the vector from the origin to (C, S). It is defined as

$$\text{atan2}(a, b) = \begin{cases} \arctan(b/a) & \text{if} \quad a > 0, \\ \arctan(b/a) + \pi & \text{if} \quad b \geq 0,\ a < 0, \\ \arctan(b/a) - \pi & \text{if} \quad b < 0,\ a < 0, \\ \pi/2 & \text{if} \quad b > 0,\ a = 0, \\ -\pi/2 & \text{if} \quad b < 0,\ a = 0, \\ \text{undefined} & \text{if} \quad b = 0,\ a = 0. \end{cases} \tag{2.1}$$

Note that the mean direction does not always exist, since it is undefined when both S and C are zero, which means null concentration. An alternative measure of location is the median direction, defined as any angle Ψ that minimizes

$$\mathbb{E}[\pi - |\pi - |\Theta - \Psi||],$$

where $|\cdot|$ denotes the absolute value function. This measure is not necessarily unique, although it is unique for unimodal distributions.

In the circle, given the nature of the variable's support, namely $[0, 2\pi)$, measures of concentration are more frequently used than measures of dispersion. The fundamental measure of concentration is the mean resultant length, which is defined as

$$\rho = (S^2 + C^2)^{1/2} \in [0, 1].$$

The mean resultant length always exists and the case when $\rho = 0$ corresponds to the situation where μ does not exist.

Although measures of concentration are far more used in the circular context, it is also possible to define measures of dispersion in the circle. For instance, the population circular variance is defined in Mardia and Jupp (2000, Chapter 3) as

$$V = 1 - \rho \in [0, 1]. \tag{2.2}$$

Other measures of dispersion can be considered, such as the circular standard deviation

$$\sigma = (-2\log(1 - V))^{1/2} \in [0, \infty].$$

Sample measures

Given a set of circular observations $\{\theta_1, ..., \theta_n\}$, it is possible to apply some statistics to summarize the data. For unimodal data, it is of interest to calculate the sample mean direction, which can be obtained as

$$\bar{\theta} = \text{atan2}(\bar{S}, \bar{C}),$$

where

$$\bar{S} = \frac{1}{n}\sum_{j=1}^{n}\sin(\theta_j), \quad \bar{C} = \frac{1}{n}\sum_{j=1}^{n}\cos(\theta_j).$$

Therefore, $\bar{C}$ and $\bar{S}$ are the sample analogues of C and S, respectively. Note that the sample mean direction is not defined when $\bar{S}$ and $\bar{C}$ are zero. As for the median direction, its sample counterpart can be obtained minimizing the dispersion measure

$$\frac{1}{n}\sum_{j=1}^{n}(\pi - |\pi - |\theta_j - \Psi||).$$

It is also possible to define the sample mean resultant length in terms of $\bar{S}$ and $\bar{C}$, as

$$\bar{R} = (\bar{S}^2 + \bar{C}^2)^{1/2}.$$

The sample mean resultant length $\bar{R}$ equals 1 only when all the data points are located at the same point on the unit circle. Then, a value close to 1 means that the data are closely congregated around the mean direction. On the other hand, a value close to 0 does not necessarily indicate that the data are evenly distributed around the circle. For example, the set of observations $\{\theta_1, ..., \theta_{n/2}, \theta_1 + \pi, ..., \theta_{n/2} + \pi\}$, for even n, has a mean resultant length of $\bar{R} = 0$.

Furthermore, the sample measures of dispersion on the circle are easily obtained. The sample circular variance is defined as

$$\bar{V} = 1 - \bar{R},$$

which is also defined in $[0, 1]$. In addition, the sample circular standard deviation is obtained as

$$\hat{\sigma} = (-2\log(1 - \bar{V}))^{1/2} \in [0, \infty].$$

The distance between two points in the circle ω and ϕ is usually measured by the circular distance

$$d(\omega, \phi) = 1 - \cos(\omega - \phi). \tag{2.3}$$

Thus, the measure of dispersion of the sample $\{\theta_1, ..., \theta_n\}$ about a direction Ψ is given by

$$d(\Psi) = \frac{1}{n}\sum_{j=1}^{n}[1 - \cos(\theta_j - \Psi)],$$

which is minimized when $\Psi = \bar{\theta}$. Note that then $d(\bar{\theta}) = 1 - \bar{R} = \bar{V}$.

2.2 Circular models

Many parametric models for circular distributions have been proposed in literature. In this section the most important circular models will be described, presenting some unimodal (both symmetric[1] and asymmetric) and multimodal models.

[1]Several kinds of symmetry exist in the circle. When referring to symmetry in the circle in this manuscript, the reflective symmetry is being considered. A circular density function f is considered to be reflectively symmetric about Ψ if it satisfies $f(\theta + \Psi) = f(-\theta + \Psi)$.

2.2.1 Unimodal models

One of the most important and useful circular distributions is the von Mises, introduced by von Mises (1918). The von Mises distribution $vM(\mu, \kappa)$, is a symmetric and unimodal distribution determined by a mean direction $\mu \in [0, 2\pi)$ and a concentration parameter $\kappa \geq 0$ with density function

$$f(\theta, \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu)), \quad \theta \in [0, 2\pi), \tag{2.4}$$

where $I_j(\cdot)$ denotes the modified Bessel function of the first kind and order j and in this case it is used as a normalizing constant. The von Mises distribution is the unique continuous circular distribution for which the sample mean direction, $\bar{\theta}$, is the maximum likelihood estimate of the population mean direction, μ. For this reason, the von Mises distribution is sometimes known as the circular Normal, as noted by Gumbel *et al.* (1953), who studied its properties and remarked its similarities with the Normal distribution. The spread of the Gaussian distribution is controlled by the standard deviation σ, while in the case of the von Mises this is controlled by the concentration parameter κ. The larger κ is, the larger the concentration, hence the lower the spread of the distribution. When the $\kappa \to 0$, the von Mises density function tends to

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

which corresponds to the density function of the circular uniform distribution. First row in Figure 2.1 shows representations of the von Mises density functions with mean direction $\mu = \pi/2$ and different concentration parameters. The density in Figure 2.1a corresponds to the circular uniform distribution.

A simple way of obtaining circular densities is to "wrap" a linear density around the unit circle. If X is a random variable in $\mathbb{R}$ with density function g, then $\Theta = X (\text{mod } 2\pi)$ and has density function

$$f(\theta) = \sum_{k=-\infty}^{\infty} g(\theta + 2\pi k).$$

The most widely used examples of wrapped densities are the wrapped normal and the wrapped Cauchy. For the wrapped normal, when considering a random variable $X \sim N(\xi, \sigma)$ with mean $\xi \in (-\infty, \infty)$ and variance $\sigma^2 > 0$, the "wrapping" procedure produces a random circular variable Θ which has density

$$f(\theta) = \frac{1}{(-4\pi \log \rho)^{-1/2}} \sum_{k=-\infty}^{\infty} \exp\left[\frac{(\theta - \mu + 2\pi k)^2}{4 \log \rho}\right],$$

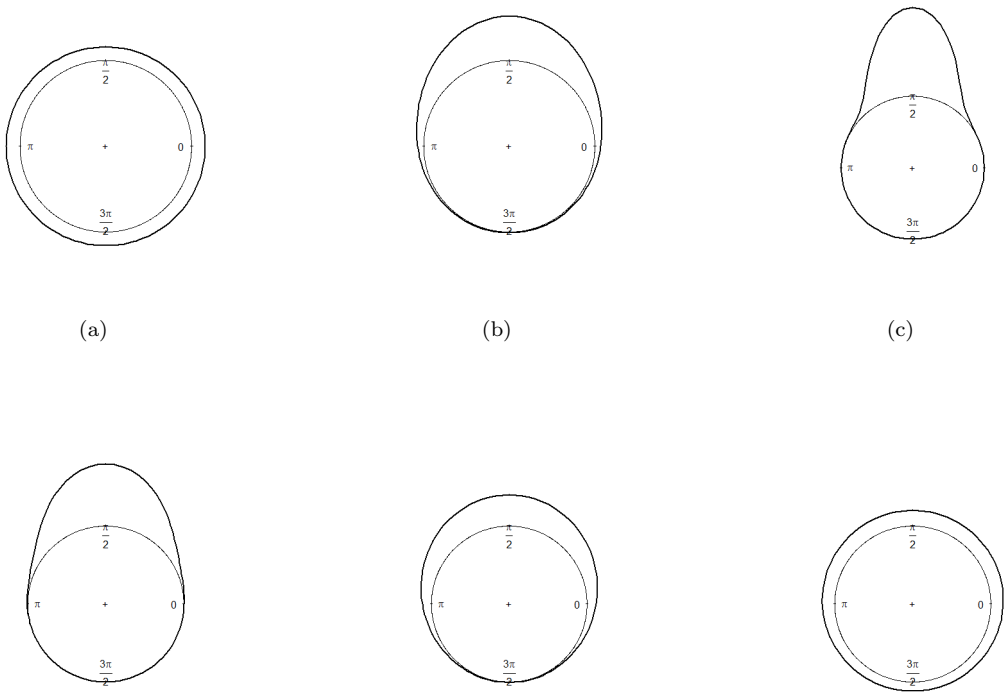
where $\rho = e^{-\sigma^2/2}$ is the concentration and $\mu = \xi (\text{mod } 2\pi)$ is the mean direction. Since the density expression involves an infinite sum, values of $f(\theta)$, $\theta \in [0, 2\pi)$, must be obtained using numerical methods. Second row in Figure 2.1 displays the representations of three wrapped normal densities with common mean direction, $\mu = \pi/2$, and different values of σ.

The von Mises distribution, the uniform distribution, the wrapped Cauchy and many other important distributions are just a particular case of a three parameter family of symmetric circular distributions known as the Jones-Pewsey family (Jones and Pewsey, 2005). The density function of this family is

$$f(\theta) = \frac{[\cosh(\kappa\Psi) + \sinh(\kappa\Psi)\cos(\theta - \mu)]^{1/\Psi}}{2\pi P_{1/\Psi}(\cosh(\kappa\Psi))},$$

where μ is a location parameter, $\kappa \geq 0$ is a concentration parameter, $\Psi \in (-\infty, \infty)$ is a kurtosis parameter and $2\pi P_{1/\Psi}$ is a normalizing constant[2].

[2] $P_{1/\Psi}(\cdot)$ is the associated Legendre function of the first kind of degree $1/\Psi$ and order zero (Gradshteyn and Ryzhik, 1994).



(a) (b) (c)

(d) (e) (f)

Figure 2.1: Polar representations of von Mises densities (first row) and wrapped normal densities (second row). Von Mises densities have mean direction $\mu = \pi/2$ and concentration parameters $\kappa = 0$ in (a), $\kappa = 2$ in (b) and $\kappa = 10$ in (c). Wrapped normal densities have mean direction $\mu = \pi/2$ and standard deviation $\sigma = 0.5$ in (d), $\sigma = 1$ in (e) and $\sigma = 2$ in (f).

Circular distributions can also be obtained by radial projection of bivariate distributions on the plane. Let the random variable $\boldsymbol{X} = (X_1, X_2)'$ follow a bivariate normal idstribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and covariance matrix $\boldsymbol{\Sigma}$. The vector $\boldsymbol{U} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$, with $\boldsymbol{U} = (U_1, U_2)'$, is the projection of $\boldsymbol{X}$ onto the unit circle. Then, the variable Θ is defined by taking $U_1 = \cos\Theta$ and $U_2 = \sin\Theta$, and it is said to follow a projected normal distribution $PN_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Small, 1996; Mardia and Jupp, 2000). Since the general structure of the covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

the density function of Θ depends on five parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, and its explicit form is complicated. When $\boldsymbol{\Sigma} = \boldsymbol{I}_2$, with $\boldsymbol{I}_2$ being the identity matrix of order 2, the density function of Θ is unimodal and symmetric. In addition, if $\mu_1 = \mu_2 = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_2$, the distribution is uniform on the circle.

All the distributions presented so far are symmetric (except for some particular cases of the projected normal distribution with $\boldsymbol{\Sigma} \neq \boldsymbol{I}_2$), but there are also asymmetric circular models, such as the Kato-Jones density (Kato and Jones, 2015). Pewsey *et al.* (2013, Chapter 4) and Ley and Verdebout (2017, Chapter 2) can be consulted for other distributions.

2.2.2 Multimodal models

In many situations, circular data will have more than one preferred direction, as it happens in Figure 2.2, where a polar representation of wind direction measurements[3] in Santiago de Compostela, Spain during December, 2017 is displayed (data avaliable in MeteoGalicia[4]). The representation includes a rose diagram (a circular alternative to the histrogram). It seems clear that there are two modes in the data considered, hence, it is not possible to model this example with any of the models described above, since all of them are unimodal. Some multimodal models are considered in literature, as the generalized von Mises distribution (Gatto and Jammalamadaka, 2007. In addition, when considering the projected normal distribution with $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2$, the resulting density is symmetric and with antipodal modes, *i.e.*, it has two equal modes in opposite directions.

However, with the objective of having a better parameter interpretation, the general approach for modeling multimodal data is to use finite mixtures of unimodal densities. The mixture density function is constructed as

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

with $\sum_m p_m = 1$ and f_m being circular densities. Finite mixtures of von Mises distributions are widely used and have been studied by many authors (Mardia and Sutton, 1975, Spurr, 1981 or Bartels, 1984).

2.3 Regression for circular data

Circular observations might be accompanied by other measurements, either circular or linear. It may be of interest to study how both variables interfere with each other from a regression perspective. For example, exploring the relationship between temperature and wind direction in Vinciguerra Glacier (Oliveira *et al.*, 2013) or predicting the angles moved by small blue periwinkles after relocation given the distance moved (Fisher and Lee, 1992; Di Marzio *et al.*, 2012). Depending on the nature of each variable, three different regression scenarios are possible: circular covariates and linear responses

[3]The wind direction data are time dependent. Although the models considered in this section are meant for independent data, the example is just meant as an illustration.

[4]MeteoGalicia website: http://www.meteogalicia.gal/observacion/rede/redeIndex.action

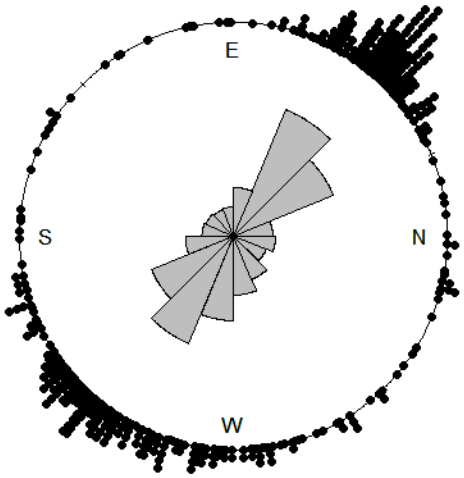


Figure 2.2: Representation in the circle of wind direction data recorded on December 2017 in Santiago de Compostela, Spain, along with a rose diagram. Data were recorded every two hours.

(circular-linear regression), linear covariates and circular responses (linear-circular regression) and circular covariates and circular responses (circular-circular regression).

2.3.1 Circular-Linear regression

In the case where the explanatory variable (Θ) has a circular nature and the response variable (Y) is linear, the following parametric model[5] is usually considered:

$$Y = \beta_0 + \beta_1 \cos\Theta + \beta_2 \sin\Theta + \varepsilon, \tag{2.5}$$

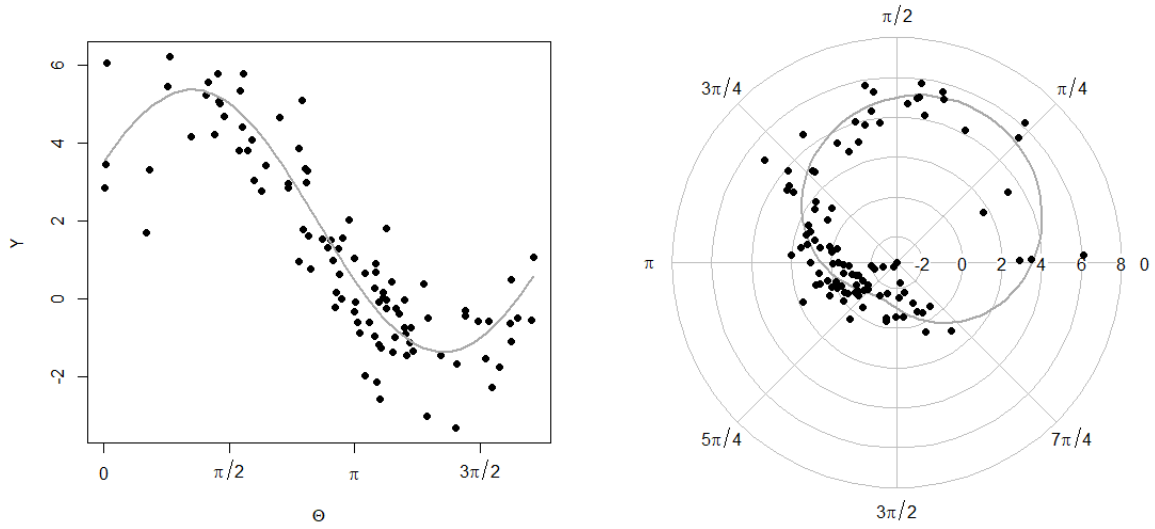
where β_0, β_1 and β_2 are parameters that must be estimated and the ε are supposed to be independent and identically distributed errors from a normal distribution with mean 0 and constant variance σ^2. This can be considered as a multiple linear model and the parameters can be estimated via least squares (see Sheather, 2009). Figure 2.3 presents simulated data from this model along with the true regression curve. The left panel shows a linear representation, where the periodicity can be seen by joining the extremes of the lines. On the other hand, the right panel displays a circular representation. It is important to note that model (2.5) is equivalent to the following cosine model

$$Y = \gamma_0 + \gamma_1 \cos(\Theta - \omega) + \varepsilon,$$

taking $\gamma_0 = \beta_0$, $\gamma_1 = \sqrt{\beta_1^2 + \beta_2^2}$ and $\omega = \text{atan2}(\beta_2, \beta_1)$. As for the polynomial regression linear case, model (2.5) can be extended to include more sine and cosine terms. Its general form is given by

$$Y = \beta_0 + \beta_1 \cos(\Theta) + \beta_2 \sin(\Theta) + \beta_3 \cos(2\Theta) + \beta_4 \sin(2\Theta) + ... + \beta_{2k-1} \cos(k\Theta) + \beta_{2k} \sin(k\Theta) + \varepsilon.$$

[5]Although only one predictor variable will be considered in this work, the model can be extended to include more covariates.



(a) (b)

Figure 2.3: Scatter plots of $n = 100$ sample points generated from model (2.5) with parameters $\beta_0 = 2$, $\beta_1 = 3/2$ and $\beta_2 = 3$, with the true regression function. (a) Linear representation. (b) Circular representation.

Other models can be contemplated, for example, Batschelet (1981) proposes a skew-cosine model to account for regression functions exhibiting skewness by including a fourth parameter and another cosine term inside the trigonometric functions in model (2.5). The same author discusses an alternative model capable of describing some departures from sinusoidal oscillations.

2.3.2 Linear-Circular regression

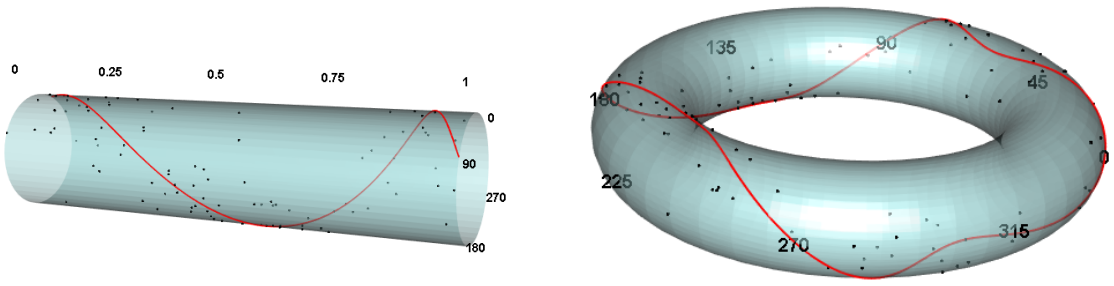
Consider now the case where the response variable Φ is circular, which is quite different from the previous setting. If the predictor variable is linear, the regression function can be regarded as a curve on the surface of an infinite cylinder, as it can be seen in Figure 2.4a. The image displays simulated data from a linear-circular model jointly with the true regression function. The dependence of Φ on the explanatory variable X can be modeled by ensuring that its mean direction depends on X. Classical linear-circular regression models are based on the von Mises distribution. Let $\Phi_1, ..., \Phi_n$ be angular observations such that each observation follows a von Mises distribution where the mean direction depends on the predictor, *i.e.,*

$$\Phi_j \sim vM(\mu(X_j), \kappa), \quad j \in \{1, ..., n\},$$

where κ is constant. One of the earliest linear-circular regression models was proposed by Gould (1969). This model assumes that the mean direction depends linearly on X:

$$\mu(X_j) = \gamma + \beta X_j, \quad j \in \{1, ..., n\}, \ \gamma \in [-\pi, \pi), \ \beta \in \mathbb{R},$$

where the parameters γ and β are estimated through maximum likelihood. This model is considered to be implausible in many situations and, in addition, it has been pointed out that problems arise when fitting the model, since the maximum likelihood estimators for the parameters are not unique. Johnson



(a) (b)

Figure 2.4: Representation of simulated data in (a) the cylinder and (b) the torus, along with the true regression functions in red. Data size is 100 in both cases. Circular units are in degrees.

and Wehrly (1978) proposed a regression model which was generalized by Fisher and Lee (1992). The location model proposed by Fisher and Lee assumes that the μ's are related to the covariate by means of a link function g so that

$$\mu(X_j) = \gamma + g(\beta X_j), \quad j \in \{1, ..., n\}, \ \gamma \in [-\pi, \pi), \ \beta \in \mathbb{R}, \tag{2.6}$$

where γ and β are parameters that must be estimated. The link function must be chosen in order to map the real line to the circle, so g should be a monotone function with support $(-\infty, \infty)$ and range $[-\pi, \pi)$. Since the parameter γ is considered as the origin, the link function should verify $g(0) = 0$. One possibility is to choose the link function from a parametric family of links, such as

$$g(u) = 2\tan^{-1}(\text{sgn}(u)|u|^{\lambda}),$$

where $\text{sgn}(\cdot)$ is the sign function. with $\lambda = 0$ corresponding to the logarithmic transformation. Although λ could be estimated from the data as in the estimation of Box-Cox transformations, usually $\lambda = 1$ is taken obtaining:

$$g(u) = 2\tan^{-1}(u).$$

Note that the previous models assume that the circular variable takes values in $[-\pi, \pi)$ and, if the support of the variable was $[0, 2\pi)$, the g functions described before would not verify $g(0) = 0$, so the parameter γ would lose its interpretation as the origin. The estimation of the parameters is done through maximum likelihood using a iteratively reweighted least squares algorithm, although Presnell *et al.* (1998) point out some computational problems that arise from this procedure. These authors present a more flexible model which does not rely on the von Mises assumption, the Projected Multivariate Linear Model (PMLM). This is a parametric model which assumes that the directions conditioned on the values of the predictor variables follow a Projected Normal distribution, i.e., the directions are projections onto the unit circle of a bivariate Normal distribution (see Section 2.2).

Regression models where the parameter dependent on the predictor variable is the concentration are also plausible. Fisher and Lee (1992) generalized a concentration model originally proposed by Johnson and Wehrly (1978), in which the concentration parameters are modeled by

$$\kappa_j = h(X_j), \quad j \in \{1, ..., n\},$$

where the function h maps $\mathbb{R}$ to $[0, \infty)$. The most common option is to choose

$$h(u) = a \exp(\omega u), \quad a \in [0, \infty), \ \omega \in \mathbb{R}. \tag{2.7}$$

Again, the parameters a and ω must be estimated from the data. Fisher and Lee (1992) also added a hybrid model, in which both the location and the concentration parameters depend on the predictors. Such model is a combination of models (2.6) and (2.7), and it is obtained straightforward.

2.3.3 Circular-Circular regression

In the case where both variables are circular, the interpretation of the regression function could be a curve on the surface of a torus, as shown in Figure 2.4b, where simulated data from a circular-circular model is displayed. Consider the explanatory variable Θ and the response variable Φ. Jammalamadaka and SenGupta (2001) use the fact that the functions $\sin(\cdot)$ and $\cos(\cdot)$ are periodic with period 2π, and they can be expressed in terms of their Fourier series expansions. Thus, the authors propose fitting the following general model:

$$\cos(\Phi_j) = \gamma^c + \sum_{k=1}^{p} (\beta_k^c \cos(k\Theta_j) + \omega_k^c \sin(k\Theta_j)) + \varepsilon_j^c,$$

$$\sin(\Phi_j) = \gamma^s + \sum_{k=1}^{p} (\beta_k^s \cos(k\Theta_j) + \omega_k^s \sin(k\Theta_j)) + \varepsilon_j^s,$$

where $(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2)$ is an error vector with mean $(\mathbf{0}, \mathbf{0})$ and whose covariance matrix must be estimated from the data. This model does not assume any specific distribution for the errors. The estimation of the parameters is done trough the least squares method. When selecting the degree p of the polynomial, it is recommended to start with a small value of p and then compute the reduction in the error sum of squares when using a degree of $(p + 1)$. If the reduction is significantly large, then $(p + 1)$ is considered and the process is repeated until there is not a significantly reduction (see Jammalamadaka and SenGupta, 2001, for details).

2.4 ANOVA and ANCOVA for circular regression

In the previous section, parametric regression models with circular variables were reviewed. However, circular measurements can also be accompanied by discrete covariates, leading to the appearance of different groups in the data. In such setting, parametric ANOVA and ANCOVA models can also be considered. This section will briefly review a parametric ANOVA model for circular data and a parametric ANCOVA model for circular predictors and linear responses. In the linear-circular and circular-circular regression contexts there are not, up to the author's knowledge, any parametric ANCOVA models.

2.4.1 ANOVA for circular variables

In order to compare the mean values of a circular variable across several groups, different techniques have been studied. The parametric approach introduced by Watson and Williams (1956) and subsequently modified by Stephens (1972) will be now revised.

Let $\{\Phi_{ij}\}$, $i \in \{1, ..., I\}$, $j \in \{1, ..., n_i\}$ be a sample of a circular variable Φ where each of the observations belongs to one out of I groups. Assume

$$\Phi_{ij} = (\mu_i + \varepsilon_{ij}) (\text{mod } 2\pi), \quad \mu_i \in [0, 2\pi) \ \forall \ i \in \{1, ..., I\}, \ j \in \{1, ..., n_i\},$$

where the errors ε_{ij} are generated from a $vM(0,\kappa)$, $\kappa > 1$. The goal is to test the equality of the circular mean directions, $\mu_1, ..., \mu_I$. Hence, the considered hypotheses are

$$H_0 : \mu_1 = ... = \mu_I,$$
$$H_1 : \mu_i \neq \mu_k, \quad \text{for some } i, k \in \{1, ..., I\}.$$

The statistic proposed by Watson and Williams (1953) takes the from

$$\frac{(\sum_{i=1}^{I} \bar{R}_i - \bar{R})/(I-1)}{\sum_{i=1}^{I}(n_i - \bar{R}_i)/(n-I)},$$

where $\bar{R}$ is the mean resultant length of the whole sample and $\bar{R}_i$ is the mean resultant length of the observations belonging to the ith group. The above statistic follows an approximate $F_{I-1,n-I}$ distribution as long as κ is large. The approximation is said to be good when $\kappa \geq 2$. For $1 < \kappa < 2$, Stephens (1972) improves the approximation by using the statistic

$$\left(1 + \frac{3}{8\hat{\kappa}}\right) \frac{(\sum_{i=1}^{I} \bar{R}_i - \bar{R})/(I-1)}{\sum_{i=1}^{I}(n_i - \bar{R}_i)/(n-I)}$$

and using the same F-distribution approximation. Here, $\hat{\kappa}$ is the overall maximum likelihood estimate of κ.

The above test is only valid when the concentration of the errors is larger than 1. In cases where such condition cannot be satisfied, a large-sample nonparametric test proposed by Watson (1983) may be employed. For small sample sizes, a bootstrap version of the test was obtained by Fisher and Hall (1991). These two versions of the test, as well as the test proposed by Watson and Williams (1953) are implemented in R's package `circular` (Agostinelli and Lund, 2017).

2.4.2 ANCOVA for circular-linear regression

In the circular regression context, a discrete predictor may also be added to a regression model. Regarding the circular-linear setting, Anderson-Cook (1999) considers the full parametric ANCOVA model derived from (2.5):

$$Y_{ij} = \gamma_i + \beta_i \cos\Theta_{ij} + \omega_i \sin\Theta_{ij} + \varepsilon_{ij}, \quad i \in \{1, ..., I\},\ j \in \{1, ..., n_i\}. \tag{2.8}$$

Since model (2.5) can be regarded as a multiple linear model, a parametric linear ANCOVA may be used to test the differences between the groups, considering $\cos\Theta$ and $\sin\Theta$ as two continue explanatory variables. The parameters of the full model can be estimated applying the least squares method for each group. Anderson-Cook (1999) introduces several tests for the ANCOVA model: one of them can be regarded as the equality test for the interaction model and another one can be thought as the parallelism test. Both tests will be reviewed and, in order to establish an analogy with the tests studied in Section 1.1, an equality test for the non-interaction model will be proposed.

Test for the non-interaction model

Recall that in a non interaction model, the effects of the continuous and the discrete predictors sum up to each other, but do not interact between them. Thus, in a model without interaction the regression function will have a different intercept for each group, while the other parameters remain equal:

$$Y_{ij} = \gamma_i + \beta \cos\Theta_{ij} + \omega \sin\Theta_{ij} + \varepsilon_{ij}, \quad i \in \{1, ..., I\},\ j \in \{1, ..., n_i\}.$$

Therefore, the interest lies on determining if, effectively, each group has a different intercept or if the regression function is the same for all groups. The hypotheses considered to test the equality of the curves are

$$H_0 : Y_{ij} = \gamma + \beta \cos\Theta_{ij} + \omega \sin\Theta_{ij} + \varepsilon_{ij}, \quad \forall\ i \in \{1, ..., I\},$$
$$H_1 : Y_{ij} = \gamma_i + \beta \cos\Theta_{ij} + \omega \sin\Theta_{ij} + \varepsilon_{ij}, \quad \gamma_i \neq \gamma_k \text{ for some } i, k \in \{1, ..., I\}.$$

Following the theory in linear parametric models, it can be derived that the corresponding test statistic is

$$\frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-I-2)} \in F_{I-1,n-I-2}, \tag{2.9}$$

where

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma} - \hat{\beta}\cos\Theta_{ij} - \hat{\omega}\sin\Theta_{ij})^2, \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}\cos\Theta_{ij} - \hat{\omega}\sin\Theta_{ij})^2.$$

In RSS_0, the parameters are estimated through the least squares method. On the other hand, the parameters in RSS are estimated using partitioned regression. The number of degrees of freedom under the null hypothesis is $(n-3)$ and under the alternative $(n-I-2)$. Then, $(RSS_0 - RSS)$ has $(I-1)$ degrees of freedom.

Test for the interaction model

Analogously to the linear case, in an interaction model the effects of the continuous and discrete variables interfere with each other resulting in not only different intercepts but different sine and cosine parameters as well, as in the full model (2.8). As before, it is of interest to test the equality of the regression functions. For that aim, the hypotheses of the test proposed by Anderson-Cook (1999) are

$$H_0 : Y_{ij} = \gamma + \beta\cos\Theta_{ij} + \omega\sin\Theta_{ij} + \varepsilon_{ij}, \quad \forall\ i \in \{1,...,I\},$$
$$H_1 : Y_{ij} = \gamma_i + \beta_i\cos\Theta_{ij} + \omega_i\sin\Theta_{ij} + \varepsilon_{ij}, \quad \gamma_i \neq \gamma_k,\ \beta_i \neq \beta_k \text{ or } \omega_i \neq \omega_k \text{ for some } i,k \in \{1,...,I\}.$$

Now, the test statistic is

$$\frac{(RSS_0 - RSS)/(3I-3)}{RSS/(n-3I)} \in F_{3I-3,n-3I}, \tag{2.10}$$

where

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma} - \hat{\beta}\cos\Theta_{ij} - \hat{\omega}\sin\Theta_{ij})^2, \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}_i\cos\Theta_{ij} - \hat{\omega}_i\sin\Theta_{ij})^2.$$

Here, under H_1 there are $(n-3I)$ degrees of freedom, since $3I$ parameters are being estimated. Thus, the number of degrees of freedom in $(RSS_0 - RSS)$ is $(3I-3)$.

Test of parallelism

Anderson-Cook (1999) proposes a test to determine if the regression curves are parallel, against the alternative hypothesis of different regression curves, in the full model (2.8). As in the linear case, this can be regarded as a test to ascertain the existence of interaction. Consequently, the next hypotheses statement is used:

$$H_0 : Y_{ij} = \gamma_i + \beta\cos\Theta_{ij} + \omega\sin\Theta_{ij} + \varepsilon_{ij}, \quad \forall\ i \in \{1,...,I\},$$
$$H_1 : Y_{ij} = \gamma_i + \beta_i\cos\Theta_{ij} + \omega_i\sin\Theta_{ij} + \varepsilon_{ij}, \quad \beta_i \neq \beta_k \text{ or } \omega_i \neq \omega_k \text{ for some } i,k \in \{1,...,I\}.$$

A reasonable statistic here is

$$\frac{(RSS_0 - RSS)/(2I-2)}{RSS/(n-3I)} \in F_{2I-2,n-3I}, \tag{2.11}$$

where

$$RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}\cos\Theta_{ij} - \hat{\omega}\sin\Theta_{ij})^2, \quad RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\gamma}_i - \hat{\beta}_i\cos\Theta_{ij} - \hat{\omega}_i\sin\Theta_{ij})^2.$$

The numbers of degrees of freedom on the F statistic are determined by the number of parameters estimated: $(n - 3I)$ in RSS and $(n - I - 2)$ in RSS_0, which makes $(2I - 2)$ degrees of freedom in $(RSS_0 - RSS)$.

Illustration with simulated data

There is not a simulation study in Anderson-Cook analyzing the performance of the tests previously described. For that reason the three tests (equality for non-interaction and interaction models and parallelism) will be now analyzed through a brief simulation study, where different scenarios are considered. In each setting, the percentages of rejection after 1000 replications of the simulated data were recorded, for a nominal level $\alpha = .05$. For the three tests, the predictor variables were drawn from the following distributions:

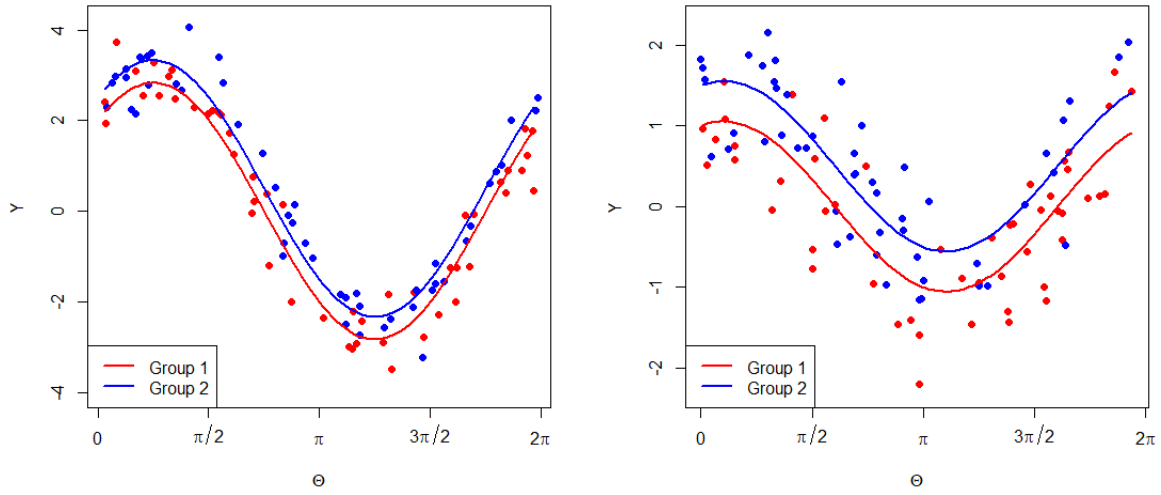
- Both groups equally spaced on the interval $[0, 2\pi)$.
- Both groups identical, as determined by a single random sample from a circular uniform distribution.
- Each group determined by a different sample from a circular uniform distribution.

For the non interaction model, consider the following regression relationships:

A. Group 1: $Y = 2\cos\theta + 2\sin\theta + \varepsilon$. Group 2: $Y = \gamma + 2\cos\theta + 2\sin\theta + \varepsilon, \quad \gamma = 0, .25, .5$

B. Group 1: $Y = \cos\theta + \frac{1}{3}\sin\theta + \varepsilon$. Group 2: $Y = \gamma + \cos\theta + \frac{1}{3}\sin\theta + \varepsilon, \quad \gamma = 0, .25, .5$

Here, $\gamma = 0$ corresponds to H_0 being true and H_1 holds for the other values of γ. In all cases, the error ε was drawn from a $N(0, \sigma)$, where $\sigma = .25, .5$. For each group 50 observations were simulated, making a total of 100 observations. Figure 2.5 displays a linear representation of simulated data from both models, where $\sigma = .5$ and $\gamma = .5$. The image includes the true regression curves for each model.



(a) Model A

(b) Model B

Figure 2.5: Representations of simulated data from the non-interaction models A and B under the alternative hypothesis $\gamma = .5$ with $\varepsilon \sim N(0, .5)$, along with the true regression curves for each group. Number of observations is 50 for each group.

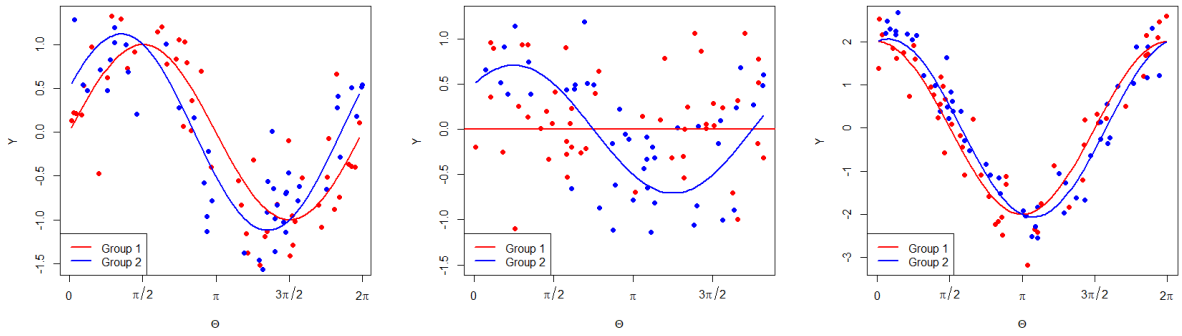
Regarding the test for the interaction model, the contemplated regression relationships were

A. Group 1: $Y = \sin\theta + \varepsilon$. Group 2: $Y = \beta\cos\theta + \sin\theta + \varepsilon, \quad \beta = 0, .25, .5$

B. Group 1: $Y = \varepsilon$. Group 2: $Y = \beta\cos\theta + \beta\sin\theta + \varepsilon, \quad \beta = 0, .25, .5$

C. Group 1: $Y = 2\cos\theta + \varepsilon$. Group 2: $Y = 2\cos\theta + \beta\sin\theta + \varepsilon, \quad \beta = 0, .25, .5$

Therefore, when $\beta = 0$ is considered, the null hypothesis holds. As in the previous case, $\varepsilon \sim N(0, \sigma)$, where $\sigma = .25, .5$, and the number of simulated observations in each group is 50. A representation of the true regression models under the alternative hypotheses ($\beta = .5$) with simulated data originated from such models can be found in Figure 2.6. The same models were considered for the parallelism test, but in this case a shift parameter was added to the responses of the second group: .2 in Model A, .05 in Model B and .3 in Model C. Results on the three tests are summarized next.



(a) Model A (b) Model B (c) Model C

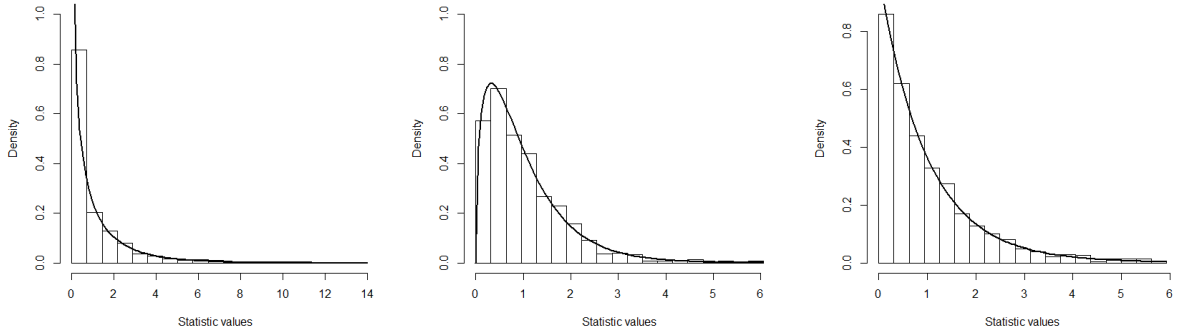
Figure 2.6: Representations of simulated data from interaction models A, B and C under the alternative hypothesis $\beta = .5$ with ε $N(0, .5)$, along with the true regression curves for each group. Number of observations is 50 for each group.

Equality test for the non-interaction model Table 2.1 collects the percentages of rejection for the equality test in the model without interaction. Results for both models, A and B, are quite similar. Under the null hypothesis, percentages of rejections vary around the nominal level .05, indicating the test is well calibrated. When the alternative hypothesis is true, percentages of rejections are large, specially when $\gamma = .5$ is used, with percentages of rejection close to 1. As expected, percentages of rejections are lower when σ increases.

Equality test for the interaction model Results for the equality test in the interaction model are displayed on Table 2.2. Under the null hypothesis all percentages of rejection (for models A, B and C) are close to $\alpha = .05$. On the other hand, under H_1 the percentages of rejection grow when increasing the differences between the curves ($\beta = .5$) and diminish when increasing the standard deviation of the errors ($\sigma = .5$).

Parallelism test Lastly, Table 2.3 summarizes the results of the parallelism test. As before, the test seems to be well calibrated, since percentages of rejection under H_0 are always close to the nominal level α. As for the alternative hypothesis, the more different the curves are, the larger percentages of rejection obtained. Naturally, increasing the error variance leads to smaller percentages of rejection.

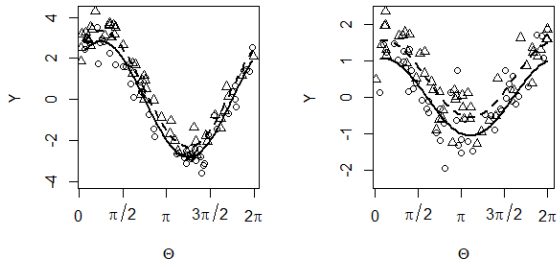
Figure 2.7a displays three histograms of the statistic values for the different tests considering Model A ($\sigma = .25$) under the null hypothesis. The theoretical F distributions are also represented, and it is easy to observe that the histograms match the true functions. The Kolmogorov-Smirnoff test was used to determine if the values of the statistics follow a F distribution under H_0, with the parameters specified before. In all cases, the corresponding p-values were much larger than the usual significance levels, therefore, there are no evidences to reject the hypothesis of the statistic values following a F distribution (under H_0).



(a) Test for the non-interaction model (b) Test for the interaction model (c) Test of parallelism

Figure 2.7: Histograms for the values of the statistics in the different tests, for the corresponding Model A ($\sigma = .25$) under H_0. The density function for the distribution of the corresponding statistics under the null hypothesis is represented as a black curve.

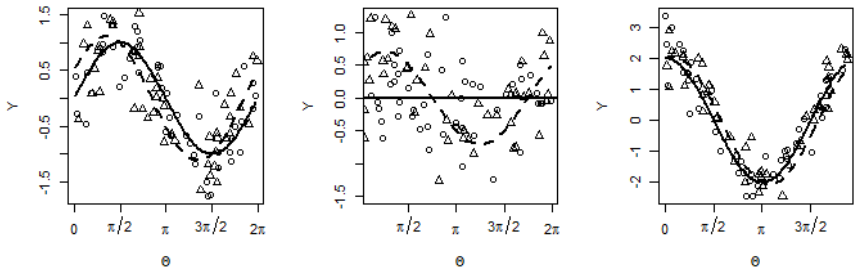
Test for the non-interaction model



σ	A			B		
	$\gamma = 0$	$\gamma = .25$	$\gamma = .5$	$\gamma = 0$	$\gamma = .25$	$\gamma = .5$
	Equally spaced design					
.25	.051	.1	1	.056	.999	1
.5	.047	.694	1	.043	.687	.997
	Circular uniform design (same for both groups)					
.25	.049	1	1	.057	.998	1
.5	.046	.711	.999	.053	.725	.997
	Circular uniform design (different for each group)					
.25	.057	1	1	.053	1	1
.5	.052	.698	.998	.042	.700	.999

Table 2.1: Percentages of rejections (for $\alpha = .05$) for the non-interaction test based on 1000 simulations and with $n_i = 50$ for $i = 1, 2$.

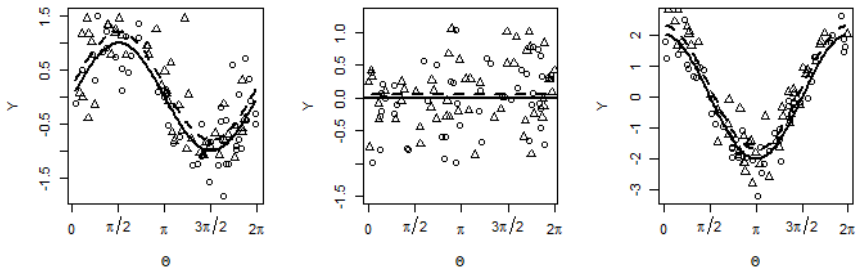
Test for the interaction model



σ	A			B			C		
	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
				Equally spaced design					
.25	.050	.842	1	.047	.995	1	.047	.851	1
.5	.053	.283	.843	.047	.502	.993	.045	.281	.857
				Circular uniform design (same for both groups)					
.25	.046	.824	1	.045	.993	1	.049	.843	1
.5	.053	.277	.850	.050	.551	.986	.039	.277	.838
				Circular uniform design (different for each group)					
.25	.053	.837	1	.051	.989	1	.050	.837	1
.5	.052	.264	.805	.053	.473	.990	.049	.302	.818

Table 2.2: Percentages of rejections (for $\alpha = .05$) for the interaction test based on 1000 simulations and with $n_i = 50$ for $i = 1, 2$.

Parallelism test



σ	A			B			C		
	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
				Equally spaced design					
.25	.050	.895	1	.041	.992	1	.044	.882	1
.5	.050	.344	.897	.043	.595	1	.051	.334	.887
				Circular uniform design (same for both groups)					
.25	0.49	.875	1	.047	.997	1	.049	.876	1
.5	.047	.304	.854	.059	.563	.992	.053	.324	.876
				Circular uniform design (different for each group)					
.25	.047	.871	1	.053	.996	1	.049	.879	1
.5	.060	.327	.887	.049	.567	.994	.051	.313	.863

Table 2.3: Percentages of rejections (for $\alpha = .05$) for the parallelism test based on 1000 simulations and with $n_i = 50$ for $i = 1, 2$.

Chapter 3

Nonparametric ANCOVA for circular regression

The regression models for circular predictors and/or circular responses presented in Chapter 2 have been approached in the statistical literature by a parametric perspective: it is usually assumed that the data are drawn from a parametric model and its parameters are estimated. However, this approach may not be the most suitable one in complex settings. Nonparametric models avoid the assumption of a specific parametric shape for the regression function and provide a flexible alternative. In this chapter, nonparametric regression models for circular data (both for circular predictors and circular responses) are reviewed in Section 3.1. In the following sections, the main contributions of this manuscript are introduced: proposals for no-effect tests for nonparametric circular models are given in Section 3.2 and Section 3.3 contains proposals for nonparametric ANCOVA models with circular variables.

3.1 Nonparametric regression for circular data

Nonparametric methods in the circular setting were first introduced in the different context of density estimation by Hall *et al.* (1987) and Bai *et al.* (1988) in the spherical case, and by Fisher (1989) in the circular context. These works adapt the classical kernel density estimator (Parzen, 1962; Rosenblatt, 1956) to the circular (or spherical) case. Hall *et al.* (1987) give two proposals to estimate a density function in the sphere making use of the fact that distance in the sphere may be measured by an angle. These authors also use different kernel functions to those in the linear case. The same approach is used by Bai *et al.* (1988) to estimate a density function in a k-dimensional sphere. Furthermore, Fisher (1989) adapts the classical density estimator for linear data to the circular case by using a quartic kernel function[1].

However, in the regression context most efforts have been concentrated on parametric models, and the nonparametric proposals are relatively recent. Di Marzio *et al.* (2009) were the first to extend local polynomial regression to the circular context, contemplating circular predictors and linear responses. The authors give the definition of circular kernels and use them to adapt the local polynomial estimator to the directional case. The asymptotic properties of the kernels and the estimator are also explored. This subject was also studied by Oliveira *et al.* (2013), who explore the performance of the estimator proposed by Di Marzio *et al.* (2009) through a simulation study. Another alternative for nonparametric circular-linear regression is to use periodic splines, introduced by Cogburn and Davis (1974) and studied by Wahba (1990) and Wood (2006). These methods are not thought specifically for circular predictors, but for any periodic covariate.

The proposal in Di Marzio *et al.* (2009) for circular covariates adapts the standard theory for local

[1] A quartic kernel function is of the form $K(u) = \frac{15}{16}(1-u^2)^2$, for $u \in [-1, 1]$.

polynomial regression by using circular kernels and a local trigonometric fit. This approach is not feasible for circular responses, given that it is not possible to (locally) average over angular variables. Thus, a different approach is used by Di Marzio *et al.* (2012), who introduce kernel regression methods for circular responses, with the support of the predictor variable being either the real line or the circle. The authors propose estimating the regression function trough the arc-tangent of a ratio between the locally weighted components of the first sample trigonometric moment of the response variable. This method avoids the selection of an arbitrary link function, as it happened in the parametric context.

In this section, a review of kernel regression methods both for circular predictors and circular responses will be given. The regression models and estimators will be presented, and methods for choosing the smoothing parameter will also be revisited.

3.1.1 Circular-linear regression

Consider a random sample $\{(\Theta_j, Y_j)\}_{j=1}^n$ from (Θ, Y), a circular and a linear variable respectively. The relationship between both variables may be described as

$$Y_j = m(\Theta_j) + \varepsilon_j, \quad j \in \{1, ..., n\}, \tag{3.1}$$

where ε_j has zero mean and standard deviation σ and Θ has a circular density function f. In the same way a local linear fit is used in nonparametric estimation with real-valued variables, under these circumstances Di Marzio *et al.* (2009) consider a local trigonometric polynomial fit

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

The parameters β_0 and β_1 are estimated via weighted local least squares, where the weights are given by a circular kernel K_κ. In practice, this kernel is usually taken as a von Mises density (2.4) with mean direction 0 and concentration parameter κ. For each $\theta \in [0, 2\pi)$, the weights given to each observation $\Theta_j,\ j = 1, ..., n$, will depend on the distance to the fixed point θ. Thus, the parameters β_0 and β_1 are estimated as

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{j=1}^{n} K_\kappa(\theta - \Theta_j)[Y_j - (a + b\sin(\theta - \Theta_j))]^2.$$

The circular-linear estimator is given by $\hat{\beta}_0 = \hat{m}(\theta)$. The vector of fitted values $\hat{\boldsymbol{m}}$ can be obtained as $\hat{\boldsymbol{m}} = \boldsymbol{SY}$, where $\boldsymbol{S}$ is a smoothing matrix. Here, κ plays a role analogous to h when considering linear data, since it controls the degree of smoothing. Large values of κ lead to undersmoothed estimations of m, tending to the interpolation of the data. In contrast, small values of κ result in a global averaging, and in a oversmoothed estimation. Under some regularity conditions it can be proven (see Di Marzio *et al.*, 2009) that for $\theta \in [0, 2\pi)$ the conditional asymptotic mean squared error (AMSE) of $\hat{m}(\theta)$ equipped with von Mises kernels is

$$\text{AMSE}[\hat{m}(\theta)|\Theta_1, ..., \Theta_n] = \frac{1}{4}\left[\frac{I_1(\kappa)}{\kappa I_0(\kappa)} m''(\theta)\right]^2 + \frac{I_0(2\kappa)}{2\pi I_0^2(\kappa)} \frac{\sigma^2}{n f(\theta)},$$

where f is the density function of the predictor variable Θ and, as before, $I_j(\cdot)$ denotes the modified Bessel function of the first kind and order j. After some calculations it can be shown that the concentration parameter which minimizes the AMISE is

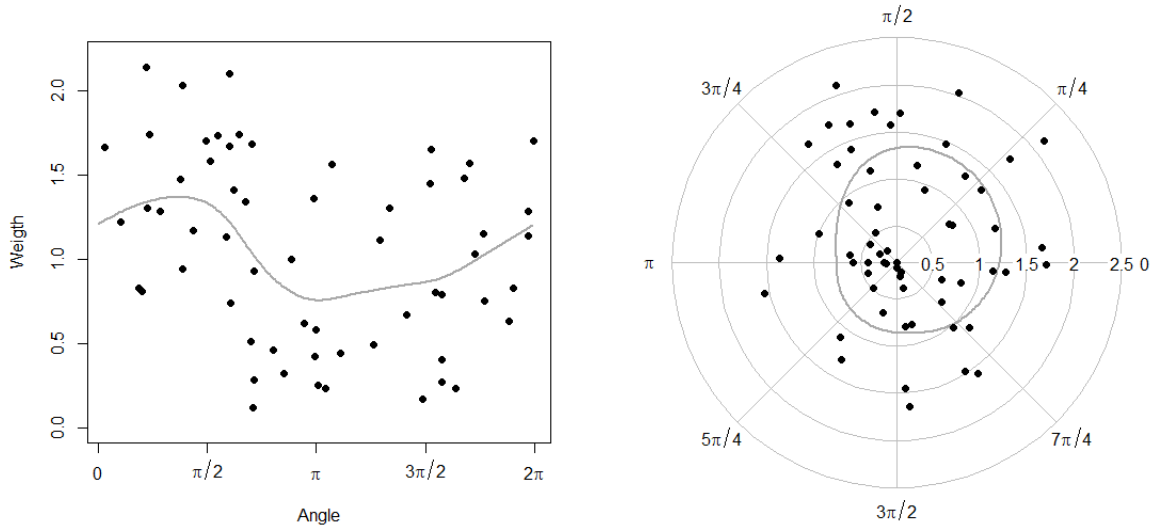
$$\kappa_{op} = \left[\frac{4n^2\pi \int_0^{2\pi} [m''(\theta)]^4 d\theta \int_0^{2\pi} [f(\theta)]^2 d\theta}{\sigma^4}\right]^{1/5}.$$

However, as it happened with (1.6) in the linear case, this optimal global parameter will depend on the unknown quantities $\int_0^{2\pi} [m''(\theta)]^4 d\theta$ and $\int_0^{2\pi} [f(\theta)]^2 d\theta$. It is then necessary to make use of other

methods to find an optimal value of κ in practice. The cross-validation method is a simple and well known alternative. It selects κ as the value that minimizes

$$CV(\kappa) = \frac{1}{n}\sum_{j=1}^{n}[Y_j - \hat{m}^{-j}(\Theta_j)]^2, \tag{3.2}$$

where $\hat{m}^{-j}$ is the leave-one-out estimator of the regression function, for $j = 1, ..., n$. As an example of the performance of the circular-linear estimator, Figure 3.1 represents the flywheel data, which will be presented and analyzed in detail in Chapter 5, with the regression function estimated with the method described above and the smoothing parameter selected by cross-validation.



(a) (b)

Figure 3.1: Scatter plots of the flywheel data with the regression curve estimated with the nonparametric circular estimator. Smoothing parameter selected by cross-validation. (a) Linear representation. (b) Circular representation.

For exploratory tasks, the selection of the smoothing parameter can be avoided by the construction of a CircSiZer (Oliveira *et al.,* 2014a). CircSiZer is a visualization method used both for regression and density estimation which allows to represent which features of the estimated curve are really present and are not sampling noise. This tool is a modification of SiZer (Chaudhuri and Marron, 1999) for circular data. For the construction of the CircSiZer map, the estimator $\hat{m}$ is evaluated also on the same grid of values of Θ (for the different values of κ). The map (see Figure 3.2) represents the regions where the slope of the curve is significantly increasing (blue), decreasing (red) or not significantly different from zero (purple). In this specific case, one can see that the regression function is significantly increasing from approximately $5\pi/4$ to $7\pi/4$ (for all the bandwidths considered) and significantly decreasing from $\pi/4$ to $5\pi/6$ (approximately).

Di Marzio *et al.* (2014) consider a generalization of the model (3.1) where the predictor variable is defined on a hypersphere of arbitrary dimension. The authors present a local linear estimator for the regression function which, however, does not match the proposal of Di Marzio *et al.* (2009) in the particular case of a circular predictor variable. García-Portugués *et al.* (2016) use a different approach

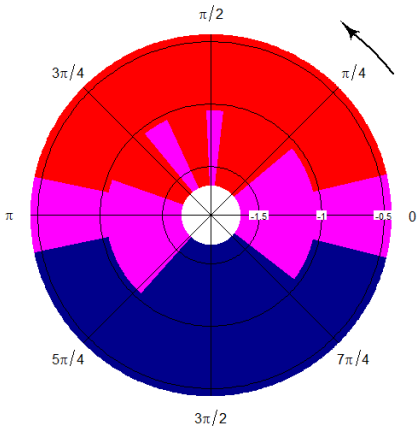


Figure 3.2: CircSiZer map for circular-linear regression applied to 100 realizations of simulated data from model (2.5) with parameters $\beta_0 = 2$, $\beta_1 = 3/2$ and $\beta_2 = 3$.

for this problem, obtaining a projected local estimator, where the estimator for the regression function is a weighted average of the response values involving directional kernels. In the particular case of a circular covariate, the estimator corresponds to the proposal given by Di Marzio *et al.* (2009).

3.1.2 Regression for circular responses

Consider now the case where the response variable, Φ, is circular and depends on a predictor variable Δ, with Δ being either circular or linear. Given the sample $\{(\Delta_1, \Phi_1), ..., (\Delta_n, \Phi_n)\}$ drawn from (Δ, Φ), the relation between these two variables can be modeled by

$$\Phi_j = [m(\Delta_j) + \varepsilon_j](\text{mod} 2\pi), \quad j \in \{1, ..., n\}, \tag{3.3}$$

where ε_j are random angles with zero mean direction, finite concentration and are independent of the Δ_js. In order to obtain an estimator for the regression function m it is necessary to recall the circular distance $d(\cdot, \cdot)$ defined in (2.3). Given two circular variables, Θ and Ψ, associated risk of $d(\Theta, \Psi)$ is

$$\mathbb{E}[1 - \cos(\Theta - \Psi)].$$

Then, the risk associated to the distance between the response variable Φ and the function of the predictor variable $m(\Delta)$ in (3.3) is given by

$$\mathbb{E}[1 - \cos(\Theta - m(\Delta))]. \tag{3.4}$$

Given an angular or scalar value δ, let $m_1(\delta) = \mathbb{E}[\sin(\Phi)|\Delta = \delta]$ and $m_2(\delta) = \mathbb{E}[\cos(\Phi)|\Delta = \delta]$, and let $g_k(\delta) = m_k(\delta)f(\delta)$, $k = 1, 2$, with f being the density function of Δ. The function that minimizes the risk (3.4) is

$$m(\delta) = \text{atan2}[g_1(\delta), g_2(\delta)],$$

with atan2 defined as in (2.1). Therefore, Di Marzio *et al.* (2012) propose estimating m as

$$\hat{m}(\delta) = \text{atan2}[\hat{g}_1(\delta), \hat{g}_2(\delta)], \tag{3.5}$$

where

$$\hat{g}_1(\delta) = \frac{1}{n}\sum_{i=j}^{n} \sin(\Phi_j)W(\Delta_j - \delta), \quad \hat{g}_2(\delta) = \frac{1}{n}\sum_{j=1}^{n} \cos(\Phi_j)W(\Delta_j - \delta),$$

and with W being a local weight function, which is defined on the real line or on the circle, depending on the nature of Δ.

Di Marzio *et al* (2012) consider different alternatives for the local weights. For linear covariates, the first option is to use a linear kernel as a weight function:

$$W(X_j - x) = K_h(X_j - x) = \frac{1}{h} K\left(\frac{X_j - x}{h}\right),$$

where K is a linear and symmetric density function and h is the smoothing parameter. This method corresponds to the circular analogue of the Nadaraya-Watson estimator. The other alternative is to use local linear weights, given by

$$W(X_j - x) = \frac{1}{n} K_h(X_j - x) \left[\sum_{k=1}^{n} [K_h(X_k - x)(X_k - x)] - (X_j - x) \sum_{k=1}^{n} [K_h(X_k - x)(X_k - x)]\right],$$

where Δ is now replaced by X (scalar variable). This local linear weights function gives higher weights to those points nearer to x. Then, the method is considered as a circular analogue for the local linear estimator.

On the other hand, for circular covariates, the first alternative is to use circular kernels as local weights (as a Nadaraya-Watson type estimator)

$$W(\Theta_j - \theta) = K_\kappa(\Theta_j - \theta),$$

and the second option is to use the following local linear weights:

$$W(\Theta_j - \theta) = \frac{1}{n} K_\kappa(\Theta_j - \theta) \left[\sum_{k=1}^{n} K_\kappa(\Theta_k - \theta) \sin^2(\Theta_k - \theta) - \sin(\Theta_j - \theta) \sum_{k=1}^{n} K_\kappa(\Theta_k - \theta) \sin(\Theta_k - \theta)\right],$$

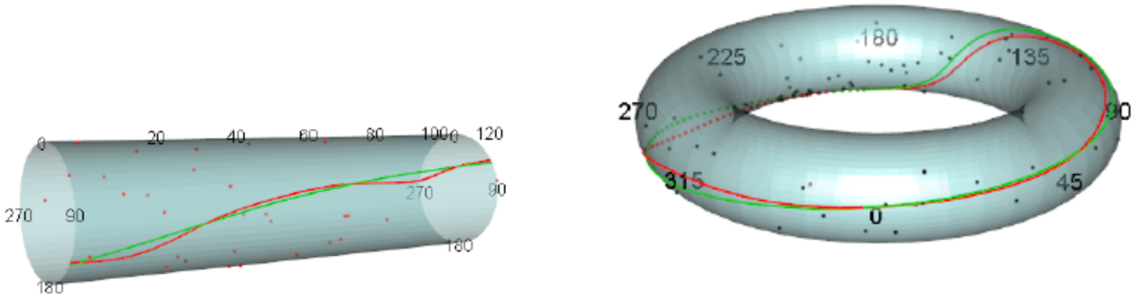
where Δ is now replaced by Θ (circular variable). As in the linear case, this function assigns larger weights to the points closer to θ. In both cases a smoothing parameter (h or κ) must be chosen. For each estimator, it is possible to calculate the AMISE and then minimize it with respect to h or κ (for details see Di Marzio *et al.*, 2012), but the optimal parameter will depend on unknown quantities. Thus, the smoothing parameter selection is usually done by cross-validation. This method selects the parameter that minimizes

$$\sum_{j=1}^{n} [-(\cos(\Phi_j - \hat{m}^{-j}(\Delta_j))],$$

where $\hat{m}^{-j}$ denotes the nonparametric estimator (and hence depending on h/κ) computed with all the observations except (Δ_j, Φ_j). The minimization of the above expression is equivalent to minimizing the sum of angular distances:

$$\sum_{j=1}^{n} d(\hat{m}^{-j}(\Delta_j), \Phi_j).$$

Figure 3.3 shows the periwinkle data given in Fisher and Lee (1992) and Di Marzio *et al.* (2012) (and mentioned in Section 3.1) represented on the cylinder and the wind direction data examined in Oliveira *et al.* (2014b) represented on the torus. In both cases the regression function was estimated with the method described above, using both the Nadaraya-Watson type weights and the local linear weights. The smoothing parameter was selected by cross-validation.



(a) (b)

Figure 3.3: (a) Representation on the cylinder of the periwinkle data with the Nadaraya-Watson estimator (red) and the local linear estimator (green). (b) Representation on the torus of the wind directions data with the Nadaraya-Watson estimator (red) and the local linear estimator (green).

3.2 Nonparametric significance test for circular data

In the previous section, regression models involving circular variables (response and/or explanatory) were reviewed. As it happened with linear data, these kernel-type estimators are constructed so that they depend heavily on the data and, sometimes, it is not clear if the features of the estimated function are actually real or if they are a consequence of the variability in the data. Specifically, it seems of great importance to asses if the response actually depends on the predictor variable. This section provides novel proposals to formally investigate the significance of the predictor variable. First, a proposal is given for the circular-linear regression setting. Afterwards, a different method is introduced for the circular response case.

3.2.1 Test for circular-linear regression

Consider the circular-linear regression scenario, with Θ being a circular predictor variable and Y a linear response variable. Let $\{(\Theta_1, Y_1), ..., (\Theta_n, Y_n)\}$ be a sample from (Θ, Y). The goal is to construct a no-effect test for regression model (3.1). Therefore, the following hypotheses are considered:

$$H_0 : Y_j = \gamma + \varepsilon_j, \quad \gamma \in \mathbb{R},$$
$$H_1 : Y_j = m(\Theta_j) + \varepsilon_j, \quad m(X_j) \neq \gamma \text{ for some } j \in \{1, ..., n\}.$$

Recall that the errors ε_j are independent and identically distributed and independent from Θ and, in addition, follow a normal distribution with zero mean and constant standard deviation σ. A test statistic can be constructed by adapting the ideas by Bowman and Azzalini (1997) to the circular context, using the nonparametric estimator presented in Section 3.1.1. Therefore, the residual sums of squares are used to quantify how much the models explain the data under each of the two hypotheses:

$$RSS_0 = \sum_{j=1}^{n} (Y_j - \hat{\gamma})^2, \quad \text{and} \quad RSS = \sum_{j=1}^{n} (Y_j - \hat{m}(X_j))^2,$$

where $\hat{m}$ is the nonparametric estimator for circular predictors introduced by Di Marzio *et al.* (2009) (see Section 3.1.1 of this manuscript) and $\hat{\gamma}$ is the sample mean of the responses. The test statistic is then constructed as a ratio:

$$C_1 = \frac{RSS_0 - RSS}{RSS}.$$

The derivation of the distribution of C_1 under H_0 is based on the normality assumption on the errors. The nonparametric estimator for circular predictors is a linear form in the data, i.e.,

$$\hat{\boldsymbol{m}} = \boldsymbol{S}\boldsymbol{Y},$$

where $\hat{\boldsymbol{m}}$ is the vector with the fitted values and $\boldsymbol{S}$ is the smoothing matrix. Consequently, the residual sums of squares can be expressed in vector-matrix notation

$$RSS_0 = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{L})'(\boldsymbol{I}_n - \boldsymbol{L})\boldsymbol{Y} \quad \text{and} \quad RSS = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{S})'(\boldsymbol{I}_n - \boldsymbol{S})\boldsymbol{Y},$$

where $\boldsymbol{L}$ is a $n \times n$ matrix with n^{-1} in all its components. Thus, the test statistic can be rewritten as

$$C_1 = \frac{\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}},$$

with $\boldsymbol{A} = (\boldsymbol{I}_n - \boldsymbol{S})'(\boldsymbol{I}_n - \boldsymbol{S})$ and $\boldsymbol{B} = \boldsymbol{I}_n - \boldsymbol{L} - \boldsymbol{A}$. Now, a p-value for the test is obtained as

$$p = \mathbb{P}\left(\frac{\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}} > Obs\right) = \mathbb{P}(\boldsymbol{Y}'(\boldsymbol{B} - \boldsymbol{A} \cdot Obs)\boldsymbol{Y} > 0),$$

with Obs being the observed value of the statistic. As discussed in Section 1.2.1, under the null hypothesis $\mathbb{E}(Y_j) = \gamma$, and in order to apply the results about quadratic forms in normal variables it is necessary that these normal variables have zero mean. However, because of the construction of C_1, it is easy to see that γ disappears due to the differences involved. Then, the p-value calculation is equivalent to

$$p = \mathbb{P}(\boldsymbol{\varepsilon}'(\boldsymbol{B} - \boldsymbol{A} \cdot Obs)\boldsymbol{\varepsilon} > 0).$$

Now, given that matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, $\boldsymbol{B} - Obs \cdot \boldsymbol{A}$ is also symmetric and the results used in Section 1.2.1 can be directly applied. Therefore, the first three cumulants of $\boldsymbol{\varepsilon}'(\boldsymbol{B} - \boldsymbol{A} \cdot Obs)\boldsymbol{\varepsilon}$ are obtained as

$$\nu_s = 2^{s-1}(s-1)!\mathrm{tr}(\boldsymbol{V}\boldsymbol{C})^s, \quad s = 1, 2, 3.$$

Then, the distribution of $\boldsymbol{\varepsilon}'(\boldsymbol{B} - \boldsymbol{A} \cdot Obs)\boldsymbol{\varepsilon}$ is approximated to a shifted and scaled χ^2, with parameters calculated as

$$a = |\nu_3|/(4\nu_2), \quad b = (8\nu_2^3)/\nu_3^2, \quad c = \nu_1 - ab, \tag{3.6}$$

with a being the scale parameter, c being the location parameter and b the number of degrees of freedom. Now, the p-value for the test is calculated as $p = 1 - q$, where

$$q = \mathbb{P}[\chi_b^2 \leq -c/a].$$

It is important to note that, as in the linear case described in Section 1.2.1, the test is very influenced by the smoothing parameter, and because of the bias present in the estimation of m, the smoothing parameter obtained by cross-validation will not be the most suitable one in many settings. In the simulation study carried out in Chapter 4 this matter will be studied and different smoothing parameters will be considered.

3.2.2 Test for circular responses

A significance test in the regression setting where the response is circular will be presented in this section. Consider a circular response variable Φ and a predictor variable Δ which can be either circular or real-valued. Let $\{(\Delta_1, \Phi_1), ..., (\Delta_n, \Phi_n)\}$ be a sample from (Δ, Φ). The hypotheses needed for a nonparametric significance test are

$$H_0 : Y_j = [\gamma + \varepsilon_j](\text{mod}2\pi), \quad \gamma \in [0, 2\pi),$$
$$H_1 : Y_j = [m(\Theta_j) + \varepsilon_j](\text{mod}2\pi), \quad m(\Theta_j) \neq \gamma + 2l\pi \text{ for some } j \in \{1, ..., n\}, \ \forall\ l \in \mathbb{Z}.$$

Recall that the errors ε_j are angles with zero mean direction, finite concentration and are independent from Δ. The model under H_0 is fitted by obtaining the sample mean direction of the responses, namely $\hat{\gamma}$. Under the alternative hypotheses the model is fitted using estimator (3.5), denoted by $\hat{m}$.

When trying to construct a suitable test statistic, one would consider adapting statistic L_1 defined in (1.7), but the approach should be discarded since its construction as a ratio of the residual sums of squares is not adequate for the circular response case. The residual sums of squares measures the quadratic distance between the observed data and the fitted data (under the null and the alternative). When the responses are of a circular nature, the residual sums of squares cannot be used because the quadratic distance is not well defined on the circle. As a consequence, the newly proposed approach will also be constructed as a ratio, but in order to measure how well the fitted model explains the data, the circular distance (2.3) will be employed. Consequently, the proposed test statistic takes the form

$$C_2 = \frac{RSD_0 - RSD}{RSD},$$

where RSD_0 and RSD are, respectively, the residual sums of distances under H_0 and H_1, defined as

$$RSD_0 = \sum_{j=1}^{n}[1 - \cos(\Phi_j - \hat{\gamma})] \quad \text{and} \quad RSD = \sum_{j=1}^{n}[1 - \cos(\Phi_j - \hat{m}(\Delta_j))].$$

Unlike in the previous section, the test statistic cannot be written in vector-matrix notation, and the arguments used for obtaining the distribution of the statistic under H_0 are not valid in this setting. Yet, the distribution of statistic C_2 under the null hypothesis can be obtained through bootstrap methods. The resampling strategy is specified hereafter.

1. Given a smoothing parameter h or κ (depending on the nature of the predictor variable), compute the nonparametric estimator of the regression function, namely $\hat{m}$. In addition, obtain the sample mean direction of the responses, $\hat{\gamma}$, and compute the observed value of the statistic C_2, denoted by Obs.

2. Obtain the residuals under the null hypothesis: $\hat{\varepsilon}_j = \Phi_j - \hat{\gamma}, \quad j \in \{1, ..., n\}$.

3. Construct the resampled responses as

$$\Phi_j^* = [\hat{\gamma} + \hat{\varepsilon}_j^*](\text{mod } 2\pi),$$

   where $\hat{\varepsilon}_j^*$ are obtained from sampling the residuals randomly with replacement.

4. Compute the bootstrap versions of $\hat{m}$ and $\hat{\gamma}$, denoted by $\hat{m}^*$ and $\hat{\gamma}^*$. In order to obtain $\hat{m}^*$, the smoothing parameter used in step 1 is again employed.

5. Evaluate the bootstrap version of the statistic

$$C_2^* = \frac{\sum_{j=1}^{n}[1 - \cos(\Phi_j^* - \hat{\gamma}^*)] - \sum_{j=1}^{n}[1 - \cos(\Phi_j^* - \hat{m}^*(\Delta_j))]}{\sum_{j=1}^{n}[1 - \cos(\Phi_j^* - \hat{m}^*(\Delta_j))]}$$

6. Repeat steps 3-5 B times to obtain the bootstrap versions of the statistic, $C_2^{*(1)}, ..., C_2^{*(B)}$ and approximate the critical value as

$$p - \text{value}^* = \frac{\#\{C_2^{*(b)} \geq Obs\}}{B}.$$

As in the no-effect test for circular-linear regression, the previously proposed test depends heavily on the smoothing parameter, and again, given the bias present in the estimation of m, the parameters selected by cross-validation are usually not adequate. This matter will be studied in detail in Chapter 4.

3.3 Nonparametric ANCOVA for circular regression

A discrete covariate can be added to the regression models studied in Section 3.1, allowing the observations to belong to different groups. Proposals of nonparametric approaches for ANCOVA models with circular variables will be detailed in this section, presenting testing methods for equality and parallelism, for regression models with circular response and/or covariate.

3.3.1 Tests for circular-linear regression

Let $\{(\Theta_{ij}, Y_{ij})\}_{ij}$ $i \in \{1, ..., I\}$, $j \in \{1..., n_i\}$ be a sample drawn from (Θ, Y), where Θ is a circular covariate and Y a linear response variable, and suppose each observation belongs to one out of I groups. Under this scenario, an ANCOVA regression model is formulated as

$$Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}, \quad i \in \{1, ..., n\}, \quad j \in \{1, ..., n_i\},$$

where the ε_{ij} are independent and $N(0, \sigma)$. This model assumes that there is a different regression curve for each group. In what follows, two different tests will be proposed: one for testing equality and another one for testing parallelism.

Test of equality

The goal for the test of equality is to determine if the regression curves are the same for all groups or, on the contrary, if there are two or more groups with different regression functions. Therefore, the hypothesis testing problem is formulated as

$$\begin{aligned} H_0 &: Y_{ij} = m(\Theta_{ij}) + \varepsilon_{ij}, \quad \forall\, i \in \{1, ..., I\}, \\ H_1 &: Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}, \quad m_i(\cdot) \neq m_k(\cdot) \text{ for some } i, k \in \{1, ..., I\}. \end{aligned}$$

Now, the regression functions are estimated using the nonparametric circular-linear estimator presented in Section 3.1.1. The estimator of m, namely $\hat{m}$, is obtained fitting the regression model with all the data, while the estimator of m_i, $\hat{m}_i$, is constructed using only the data belonging to group i, with $i \in \{1, ..., I\}$.

In the circular-linear scenario, although the predictor variable is circular, the response of the regression function is still linear. Thus, the quadratic distance is adequate to measure the differences between the global estimator and the estimated regression curves for each group. Therefore, for the equality test, the next statistic is proposed:

$$C_3 = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{m}_i(\Theta_{ij}) - \hat{m}(\Theta_{ij})]^2, \tag{3.7}$$

where $\hat{\sigma}^2$ is an estimator of the variance. Despite the simple form of the test which resembles the corresponding one for real-valued covariates, the circular behavior of the predictor variable plays an

important role in the variance estimation. In the linear case, two estimators were presented, (1.10) and (1.11). When the nature of the explanatory variable is circular, because of its periodic behavior, these variance estimators should be adjusted. Modifications of both estimators for the circular predictors case are proposed.

In order to estimate the variance in each group, the modification of estimator (1.10) takes the form

$$\hat{\sigma}_i^2 = \frac{1}{2n_i}\sum_{j=1}^{n_i}[Y_{i[j+1]} - Y_{i[j]}]^2, \quad i \in \{1, ..., I\}.$$

In the above expression, $Y_{i[j]}$, with $j \in \{1, ..., n_i\}$, denotes the value of Y corresponding to $\Theta_{i[j]}$, where $\Theta_{i[j]}$ represents the jth smallest value on the real line of the sample from Θ in group i (given that an origin has been chosen) and $Y_{i[n_i+1]} = Y_{i[1]}$. Consequently, this estimator also measures the differences between the observations corresponding to the smallest and the largest values of the sample from Θ (in group i), given an origin. Hence, the global variance would be estimated as

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I} n_i\hat{\sigma}_i^2.$$

The previous modified estimator can also be expressed in vector-matrix notation as $\boldsymbol{Y'BY}$, where $\boldsymbol{B}$ is a $n \times n$ block matrix composed of i blocks, where the ith block is a $n_i \times n_i$ matrix of the form

$$(2n-2I)^{-1}\begin{bmatrix} 2 & -1 & & & & & -1 \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \dots & \dots & \dots & & \\ & & & \dots & \dots & \dots & \\ & & & & -1 & 2 & -1 \\ -1 & & & & & -1 & 2 \end{bmatrix}.$$

Note that the $\boldsymbol{B}$ matrix is different from the one used in the linear context, since it adjusts for the circular nature of the covariate. Specifically, in the linear context the first and last terms of the diagonal where 1 and all the terms outside the three main diagonals where zero.

In addition, estimator (1.11) is also adapted to the circular setting. In this context, the modified pseudo-residuals are defined as

$$\tilde{\varepsilon}_{i[j]} = \frac{\Theta_{i[j+1]} - \Theta_{i[j]}}{\Theta_{i[j+1]} - \Theta_{i[j-1]}}Y_{i[j-1]} + \frac{\Theta_{i[j]} - \Theta_{i[j-1]}}{\Theta_{i[j+1]} - \Theta_{i[j-1]}}Y_{i[j+1]} - Y_{i[j]}, \quad i \in \{1, ..., I\},\ j \in \{1, ..., n_i\},$$

where $Y_{i[n_i+1]} = Y_{i[1]}$, $Y_{i[n_i+2]} = Y_{i[2]}$ and $Y_{i[0]} = Y_{i[n_i]}$. Note that in the linear case, the pseudo-residuals introduced by Gasser *et al.* (1986) are only defined for $j \in \{2, ..., n_i - 1\}$. The new pseudo-residuals can then be expressed as $\tilde{\varepsilon}_{i[j]} = a_{i[j]}Y_{i[j-1]} + b_{i[j+1]}Y_{i[j+1]} - Y_{i[j]}$, and thus, the variance in each group is estimated as

$$\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{c_{i[j]}^2}\tilde{\varepsilon}_{i[j]}^2, \tag{3.8}$$

where $c^2_{i[j]} = a^2_{i[j]} + b^2_{i[j]} + 1$, $i \in \{1, ..., I\}$, $j \in \{1, ..., n_i\}$. Therefore, the global variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{n-I} \sum_{i=1}^{I} n_i \hat{\sigma}_i^2.$$

This estimator can also be written as $\boldsymbol{Y'BY} = \boldsymbol{Y'A'AY}$, where $\boldsymbol{A}$ is a $n \times n$ block matrix, with the ith block taking the form

$$(n-I)^{-1/2} \begin{bmatrix} -1/c_{i[1]} & b_{i[1]}/c_{i[1]} & & & & & & a_{i[1]}/c_{i[1]} \\ a_{i[2]}/c_{i[2]} & -1/c_{i[2]} & b_{i[2]}/c_{i[2]} & & & & & \\ & a_{i[3]}/c_{i[3]} & -1/c_{i[3]} & b_{i[3]}/c_{i[3]} & & & & \\ & & \dots & \dots & \dots & & & \\ & & & \dots & \dots & \dots & & \\ & & & & a_{i[n_i-1]}/c_{i[n_i-1]} & -1/c_{i[n_i-1]} & b_{i[n_i-1]}/c_{i[n_i-1]} & \\ b_{i[n]}/c_{i[n]} & & & & & a_{i[n]}/c_{i[n]} & -1/c_{i[n]} & \end{bmatrix}.$$

Note that the difference with the matrix used in the linear context is that, in that case, the first and last rows of each block were zeros, while that does not happen here.

Now, C_3 can be written in vector-matrix notation as $\boldsymbol{Y'QY}/\boldsymbol{Y'BY}$. The matrix $\boldsymbol{Q}$ is of the form $[\boldsymbol{S_d} - \boldsymbol{S}]'[\boldsymbol{S_d} - \boldsymbol{S}]$, with $\boldsymbol{S}$ being the smoothing matrix under the null hypothesis and $\boldsymbol{S_d}$ being the block matrix made out of the smoothing matrices of each group. In order to obtain a p-value for the test it is necessary to calculate the distribution of C_3 under the null hypothesis. As in the linear case exposed in Section 1.2.2, it will be seen that such distribution is almost equivalent to the distribution of $\boldsymbol{\varepsilon'B\varepsilon}$. Given that under the null hypothesis $\boldsymbol{Y} = \boldsymbol{m} + \boldsymbol{\varepsilon}$, the statistic C_3 can be expressed as

$$\frac{\boldsymbol{m'Qm} + \boldsymbol{\varepsilon'Q\varepsilon}}{\boldsymbol{m'Bm} + \boldsymbol{\varepsilon'B\varepsilon}}. \tag{3.9}$$

If a common concentration parameter κ is used to estimate $\hat{m}$ and $\hat{m}_i$, $i \in \{1, ..., I\}$, the first term in the numerator of the previous expression disappears asymptotically, because, as it happened with the nonparametric linear estimator, the means of the circular-linear estimators of m and m_i are approximately equal. This is due to their bias expressions, which are, respectively

$$\text{bias}[\hat{m}(\theta)|\Theta_1, ..., \Theta_n] \approx \frac{1}{4} \left[\frac{I_1(\kappa)}{\kappa I_0(\kappa)} \right]^2 m''(\theta) + \left| \frac{I_0(2\kappa)}{2\pi I_0^2(\kappa)} \right| \frac{\sigma^2}{nf(\theta)},$$

and

$$\text{bias}[\hat{m}_i(\theta)|\Theta_1, ..., \Theta_n] \approx \frac{1}{4} \left[\frac{I_1(\kappa)}{\kappa I_0(\kappa)} \right]^2 m_i''(\theta) + \left| \frac{I_0(2\kappa)}{2\pi I_0^2(\kappa)} \right| \frac{\sigma^2}{n_i f(\theta)}.$$

In addition, the first term in the denominator of (3.9) is very small in comparison with the term $\boldsymbol{Y} = \boldsymbol{m} + \boldsymbol{\varepsilon}$, because of how the $\boldsymbol{B}$ matrices are constructed, so such term can be ignored. Therefore, the distribution of the statistic is obtained by calculating the distribution of $\boldsymbol{\varepsilon'Q\varepsilon}/\boldsymbol{\varepsilon'B\varepsilon}$. Hence, the correspondent p-value of the test is obtained as

$$p = \mathbb{P}\left(\frac{\boldsymbol{\varepsilon'Q\varepsilon}}{\boldsymbol{\varepsilon'B\varepsilon}} > Obs \right) = \mathbb{P}\left(\boldsymbol{\varepsilon'}(\boldsymbol{Q} - \boldsymbol{B} \cdot Obs)\boldsymbol{\varepsilon} > 0 \right).$$

The variable $\boldsymbol{\varepsilon}'(\boldsymbol{Q}-\boldsymbol{B}\cdot Obs)\boldsymbol{\varepsilon}$ is again a quadratic form in normal variables of the type $\boldsymbol{z}'\boldsymbol{C}\boldsymbol{z}$, with $\mathbb{E}(\boldsymbol{z})=\boldsymbol{0}$ and $\boldsymbol{C}$ being a symmetric matrix. Consequently, the shifted and scaled χ^2 approximation also works in this scenario: the first three moments of $\boldsymbol{\varepsilon}'(\boldsymbol{Q}-\boldsymbol{B}\cdot Obs)\boldsymbol{\varepsilon}$ are calculated and used to obtain the parameters of a $a\chi^2_b+c$ distribution as in (3.6).

Test of parallelism

Once it is known that the regression curves are different, it is of interest to test parallelism, i.e., to test whether switching from one group to another just increases or decreases the value of the response variable by a constant. The hypotheses considered in this case are

$$\begin{aligned} &H_0: Y_{ij}=\gamma_i+m(\Theta_{ij})+\varepsilon_{ij}, \quad \gamma_1=0,\ \forall\ i\in\{1,...,I\},\\ &H_1: Y_{ij}=m_i(\Theta_{ij})+\varepsilon_{ij}, \quad m_i(\cdot)\neq m_k(\cdot)+\gamma \text{ for some } i,k\in\{1,...,I\},\ \forall\ \gamma\in\mathbb{R}. \end{aligned}$$

Under the null hypothesis, the model can be written in vector-matrix notation as

$$\boldsymbol{Y}=\boldsymbol{D}\boldsymbol{\gamma}+\boldsymbol{m}+\boldsymbol{\varepsilon}, \tag{3.10}$$

where $\boldsymbol{D}$ is a known matrix consisting of 0s and 1s. Given a vector $\boldsymbol{\gamma}$, an estimate of the regression function can be constructed:

$$\hat{\boldsymbol{m}}=\boldsymbol{S}(\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{\gamma}),$$

with $\boldsymbol{S}$ being a smoothing matrix constructed with the circular-linear regression method. Substituting this estimator in equation (3.10) and applying the least squares method, an estimate of $\boldsymbol{\gamma}$ is derived:

$$\hat{\boldsymbol{\gamma}}=[\boldsymbol{D}'(\boldsymbol{I}_n-\boldsymbol{S_1})'(\boldsymbol{I}_n-\boldsymbol{S_1})\boldsymbol{D}]^{-1}\boldsymbol{D}'(\boldsymbol{I}_n-\boldsymbol{S_1})'(\boldsymbol{I}_n-\boldsymbol{S_1})\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{Y},$$

where $\boldsymbol{S_1}$ is a preliminary smoothing matrix. After $\hat{\boldsymbol{\gamma}}$ is estimated, the regression function m is estimated as

$$\hat{\boldsymbol{m}}=\boldsymbol{S}(\boldsymbol{Y}-\boldsymbol{D}\hat{\boldsymbol{\gamma}}).$$

However, for estimating the vector of parameters $\hat{\boldsymbol{\gamma}}$ it is necessary to choose a first smoothing parameter κ_1, independent of the one used to estimate the actual curves. Although in practice it is recommended to explore several smoothing parameters, a new automatic rule was derived in order to be able to obtain a p-value. For obtaining the rule, the recommendation of Bowman and Azzalini (1997) in the linear case was followed, that is to restrict the smoothing to approximatively eight neighboring observations. Let $d_2(\cdot,\cdot)$ be defined as

$$d_2(\Phi,\Theta)=\min\{|\Phi-\Theta|,2\pi-|\Phi-\Theta|\}, \quad \Phi,\Theta\in[0,2\pi).$$

Then, d_2 is a distance in the circle, different from distance (2.3). The automatic rule consists of finding a preliminary vector of smoothing parameters, $\boldsymbol{h_1}$, containing one parameter for each observation, in which the parameter associated to observation Θ_{ij} will be the distance to its 8th[2] nearest neighbor (considering distance d_2). Then, $\boldsymbol{h_1}$ is used to obtain a vector of smoothing parameters valid for the circular case using the results in Gumbel *et al.* (1953), which show that for large values of κ the von Mises $vM(\mu,\kappa)$ converges in distribution to a $N(\mu,1/\sqrt{\kappa})$. Thus, if h_1 is the preliminary smoothing parameter corresponding to Θ_{ij}, the concentration parameter for this observation will be $\kappa_1=1/h_1^2$.

Now, the proposed statistic for the test of parallelism is of the form

$$C_4=\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[\hat{\gamma}_i+\hat{m}(\Theta_{ij})-\hat{m}_i(\Theta_{ij})]^2,$$

[2]The election of 8 as the number of neighbors used was motivated by the recommendations in Bowman and Azzallini (1997). However, simulations not showed in this manuscript revealed that when considering a number of neighbors close to 8 (e.g. 6,10,12) the bias of the estimator of $\boldsymbol{\gamma}$ remained practically unchanged.

with $\hat{\sigma}^2$ being one of the two variance estimators proposed in the equality test, accounting for the periodic behavior of Θ, and $\hat{m}$ and $\hat{m}_i$ are, respectively, the circular-linear estimators of m and m_i, with $i \in \{1, ..., I\}$. Again, C_4 can be written in matrix notation as $\boldsymbol{Y'QY}/\boldsymbol{Y'BY}$ where $\boldsymbol{B}$ is one of the matrices described above and the matrix $\boldsymbol{Q}$ is of the form

$$[\boldsymbol{DA} + \boldsymbol{S}(\boldsymbol{I}_n - \boldsymbol{DA}) - \boldsymbol{S_d}]'[\boldsymbol{DA} + \boldsymbol{S}(\boldsymbol{I}_n - \boldsymbol{DA}) - \boldsymbol{S_d}].$$

In order to derive the distribution of the statistic under the null, the same reasoning as in the equality test is carried out. Therefore, the desired distribution will be a shifted and scaled χ^2 distribution, where the shift and scale parameters and the number of degrees of freedom depend on the first three cumulants of $\boldsymbol{\varepsilon'}(\boldsymbol{Q} - \boldsymbol{B} \cdot Obs)\boldsymbol{\varepsilon}$, and are obtained as in the previous section.

3.3.2 Tests for circular responses

In this section, proposals for nonparametric ANCOVA models for circular responses will be given, both for linear or angular predictors. Two tests are proposed: equality and parallelism. The goal is to formulate a nonparametric ANCOVA model and to construct the hypotheses for testing equality and parallelism, as well as obtaining the corresponding test statistics and their distribution under the null hypothesis.

Consider the notation in Section 3.1.2, where Δ represents the predictor variable, which can be either real-valued or circular, and Φ represents the response variable, which is angular. Consider, in addition, a discrete predictor variable with I different groups as attributes. Let $\{(\Delta_{ij}, \Phi_{ij})\}_{ij},\ i \in \{1, ..., I\}$,
$j \in \{1, ..., n_I\}$, be a random sample from (Δ, Φ), where the index i indicates that the observation belongs to the ith group. Now, the ANCOVA model can be written as

$$\Phi_{ij} = [m_i(\Delta_{ij}) + \varepsilon_{ij}](\text{mod}2\pi), \quad i \in \{1, ..., I\},\ j \in \{1, ..., n_I\}.$$

The methods that will be described assume that the errors ε_{ij} have zero mean and constant concentration, but they do not need to follow any specific distribution. The equality and parallelism test will be introduced next.

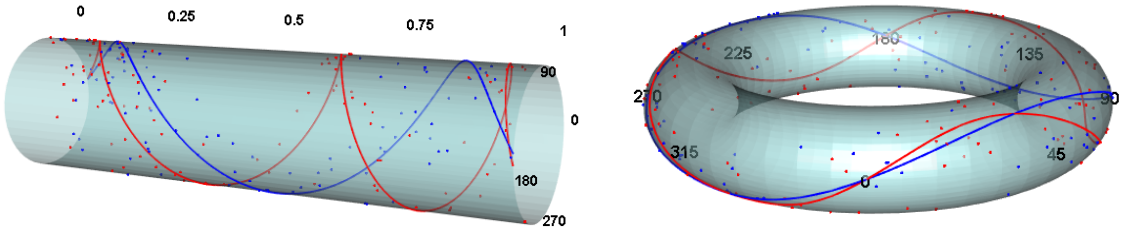
Test of equality

For testing the equality of the regression curves the following hypotheses are needed:

$$H_0 : \Phi_{ij} = [m(\Delta_{ij}) + \varepsilon_{ij}](\text{mod}2\pi), \quad \forall\ i \in \{1, ..., I\},$$
$$H_1 : \Phi_{ij} = [m_i(\Delta_{ij}) + \varepsilon_{ij}](\text{mod}2\pi), \quad m_i(\cdot) \neq m_k(\cdot) + 2l\pi \text{ for some } i, k \in \{1, ..., I\},\ \forall\ l \in \mathbb{Z}.$$

Then, the null hypotheses assumes that there is only one regression curve, independently from the group, while the alternative presumes that there are at least two different regression functions, as in the situation displayed in Figure 3.4, which shows simulated data from two distinct regression curves, both in the cylinder and the torus. The statistics used in Section 3.3.1 are a modification of the ones proposed by Young and Bowman (1995), and reviewed in Section 1.2, to account for the circular nature of the predictor variable. The idea under those statistics is to measure the quadratic distance between the estimations of the regression curves for each group and the estimation of the regression function for all the data. Such approach is not feasible in this scenario, since the response variable is now circular and the quadratic distance is not well defined on the circle. It is necessary, then, to consider the circular distance given in (2.3). Therefore, a proposal of test statistic in this context is given by

$$C_5 = \frac{1}{\bar{D}} \sum_{i=1}^{I} \sum_{j=1}^{n_I} d(\hat{m}_i(\Delta_{ij}), \hat{m}(\Delta_{ij})) = \frac{1}{\bar{D}} \sum_{i=1}^{I} \sum_{j=1}^{n_I} \left[1 - \cos(\hat{m}_i(\Delta_{ij}) - \hat{m}(\Delta_{ij}))\right], \tag{3.11}$$



(a) (b)

Figure 3.4: Simulated data from different regression functions in (a) the cylinder and (b) the torus, with the true regression functions. Data size is 100 for each group in both cases. Units in the circle given in degrees.

where $\bar{D}$ is an estimator of the circular variance of the errors defined in (2.2) and it is given by given by

$$\bar{D} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}[1-\cos(Y_{ij}-\hat{m}_i(\Delta_{ij}))]. \tag{3.12}$$

The distribution of C_5 under the null hypothesis is obtained using bootstrap methods. The resampling strategy is described next.

1. Compute the dispersion estimator $\bar{D}$ and the nonparametric estimates $\hat{m}$ (with all the data) and $\hat{m}_1, ..., \hat{m}_I$ (for each group) using the same smoothing parameter, for example the one selected by cross-validation. Obtain the observed value of statistic C_5, namely *Obs*.

2. Obtain the residuals under the null hypothesis: $\hat{\varepsilon}_{ij} = \Phi_{ij} - \hat{m}(\Delta_{ij})$.

3. Construct the resampled responses as

$$\Phi^*_{ij} = [\hat{m}(\Delta_{ij}) + \hat{\varepsilon}^*_{ij}](\text{mod } 2\pi),$$

where $\hat{\varepsilon}^*_{ij}$ are obtained from sampling the residuals randomly with replacement.

4. Compute the bootstrap versions of $\hat{m}, \hat{m}_1, ..., \hat{m}_I$ and $\bar{D}$, denoted by $\hat{m}^*, \hat{m}^*_1, ..., \hat{m}^*_I$, and $\bar{D}^*$, using the same smoothing parameter employed in step 1 for the estimation of the regression functions.

5. Evaluate the bootstrap version of the statistic

$$C^*_5 = \frac{1}{\bar{D}^*}\sum_{i=1}^{I}\sum_{j=1}^{n_I}\left[1-\cos(\hat{m}^*_i(\Delta_{ij})-\hat{m}^*(\Delta_{ij}))\right].$$

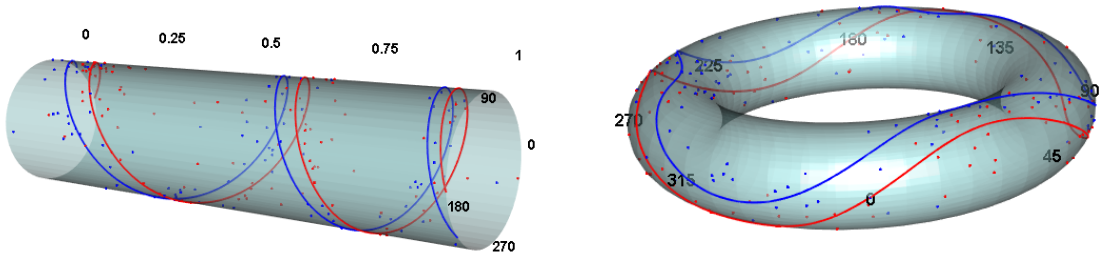
6. Repeat steps 3-5 B times to obtain the bootstrap versions of the statistic, $C_5^{*(1)}, ..., C_5^{*(B)}$ and approximate the significant value as

$$p - \text{value}^* = \frac{\#\{C_5^{*(b)} \geq Obs\}}{B}.$$

Note that in step 4 the bootstrap versions of the estimators are obtained using the smoothing parameter applied initially, instead of selecting a new bandwidth for each bootstrap replication. This is the usual procedure when dealing with bootstrap methods in nonparametric regression (Hardle and Bowman, 1988; Politis, 2014), because the goal here is to obtain the distribution of the statistic under the null hypothesis, and changing the smoothing parameter would highly increase the variability of the estimator under H_0. In addition, in most literature concerning this matter, the bootstrap versions of the responses are constructed with an oversmoothed estimator, in order to correct the bias (Hardle and Marron, 1991; Cao and González-Manteiga, 1993). However, in this case such correction is unnecessary, given that the statistic in (3.11) involves the differences between $\hat{m}_i$ and $\hat{m}$, and since the same smoothing parameter is used for both estimations, the bias is canceled out.

Test of parallelism

On top of the test of equality, it can be useful to determine if the regression curves are "parallel", meaning that the regression functions have the same shape, except for an angular shift between the curves. Figure 3.5 shows this "parallel" behavior of the functions with simulated data belonging to two different groups. The left plot corresponds to the case with a linear covariate (represented on the cylinder), while the right plot displays data coming from a model with a circular predictor (represented on the torus).



(a) (b)

Figure 3.5: Simulated data from parallel regression functions in (a) the cylinder and (b) the torus, with the true regression functions. Data size is 100 for each group in both cases. Units in the circle given in degrees.

For testing parallelism, the considered hypotheses are:

$$H_0 : \Phi_{ij} = [\gamma_i + m(\Delta_{ij}) + \varepsilon_{ij}](\text{mod}2\pi), \quad \forall\ i \in \{1, ..., I\},$$
$$H_1 : \Phi_{ij} = [m_i(\Delta_{ij}) + \varepsilon_{ij}](\text{mod}2\pi), \quad m_i(\cdot) \neq [m_k(\cdot) + \gamma](\text{mod}2\pi) \text{ for some } i, k \in \{1, ..., I\},\ \forall\ \gamma \in [0, 2\pi).$$

In order to fit the model under H_1, the regression estimator (3.5) is applied to the data of each group, leading to the estimations $\hat{m}_1, ..., \hat{m}_I$. The estimation under the null hypothesis is more complicated. The approach followed will be similar to the one used in the circular-linear case: if $\gamma_1, ..., \gamma_I$ were known, the global regression function could be estimated applying estimator (3.5) to the data $\{(\Delta_{ij}, \Phi_{ij} - \gamma_i)\}$, with $i \in \{1, ..., I\}$ and $j \in \{1, ..., n_i\}$. Hence, estimations $\hat{\gamma}_1, ..., \hat{\gamma}_I$ of the shift parameters will be obtained, and then they will be used to estimate the global m. In the circular-linear case, the model is written in vector-matrix notation and after some rearrangement the vector containing the shift parameters is estimated through the least squares method, but several problems arise when trying to use the same approach in this scenario. In the first place, it is difficult to write the fitted model in vector-matrix notation, since the regression estimator proposed by Di Marzio *et al.* (2012) is not linear on the responses, *i.e.*, it can not be written as a smoothing matrix multiplied by the vector of responses. In the second place, applying least squares would not be adequate here, given that the quadratic distance is not well defined on the circle. In order to overcome these problems, the circular distance (2.3) is again considered. Therefore, $\hat{\gamma}_1, ..., \hat{\gamma}_I$ will be the parameters that solve the next minimization problem:

$$\begin{aligned} &\underset{\gamma_1,...,\gamma_I}{\arg\min} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [1 - \cos(\Phi_{ij} - \gamma_i - \hat{m}^1(\Delta_{ij}))] \\ &\text{s.t.} \qquad \gamma_i \in [0, 2\pi), \ \forall \ i \in \{1, ..., I\}, \end{aligned}$$

where $\hat{m}^1$ is a preliminary estimation of m. Given the difficulty of the optimization problem, it is solved with numerical methods. Specifically, the optimization method used is a limited memory BFGS (L-BFGS) proposed by Byrd *et al.,* (1995), which is meant for bound constraint optimization. The estimations $\hat{\gamma}_1, ..., \hat{\gamma}_I$ obtained will not be unbiased, due to the bias of the preliminary estimator $\hat{m}^1$ (simulations showed that when using the true values of m, the estimators of the shift parameters were unbiased). However, the bias is smaller as the sample size increases.

At the same time, a first smoothing parameter needs to be chosen to obtain $\hat{m}^1$, and it should be selected so that it minimizes the bias in the preliminary estimation of m and therefore in the estimation of $\gamma_1, ..., \gamma_I$. For this aim, when the predictor variable is linear, the smoothing parameter should be large, and for circular covariates the smoothing (concentration) parameter should be small. Although it is recommended to explore several parameters, an automatic rule was derived.

When the predictors are linear ($\Delta = X$), the rule consists of using a vector of smoothing parameters in which each of them corresponds to one observation. Each parameter will be the distance to the 8th nearest observation. On the other hand, in the case where the predictor is of a circular nature ($\Delta = \Theta$), the rule is the same as in the test of parallelism for circular-linear regression (Section 3.3.1).

Now, using the same ideas behind the statistic for the equality test, the next statistic is proposed for the test of parallelism:

$$C_6 = \frac{1}{\bar{D}} \sum_{i=1}^{I} \sum_{j=1}^{n_I} d(\hat{\gamma}_i + \hat{m}(\Delta_{ij}), \hat{m}_i(\Delta_{ij})) = \frac{1}{\bar{D}} \sum_{i=1}^{I} \sum_{j=1}^{n_I} \left[1 - \cos(\hat{\gamma}_i + \hat{m}(\Delta_{ij}) - \hat{m}_i(\Delta_{ij}))\right],$$

with $\bar{D}$ being the dispersion estimator in (3.12). As in the test of equality, the distribution of C_6 under the null hypothesis is calculated through bootstrap methods. The resampling plan is as follows:

1. Choose a preliminary smoothing parameter h_1 or κ_1 (depending on the nature of the explanatory variable) and obtain the nonparametric estimator $\hat{m}^1$ using all the data. Calculate $\hat{\gamma}_1, ..., \hat{\gamma}_I$.

2. Compute the nonparametric estimate $\hat{m}$ using all the data and the shift parameters estimators. Compute $\hat{m}_1, ..., \hat{m}_I$ using the data belonging to each group and calculate $\bar{D}$. Evaluate statistic C_6 obtaining the observed value Obs.

3. Obtain the residuals under the null hypothesis: $\hat{\varepsilon}_{ij} = \Phi_{ij} - \hat{\gamma}_i - \hat{m}(\Delta_{ij})$.

4. Construct the resampled responses as

$$\Phi_{ij}^* = [\hat{\gamma}_i + \hat{m}(\Delta_{ij}) + \hat{\varepsilon}_{ij}^*](\text{mod } 2\pi),$$

where $\hat{\varepsilon}_{ij}^*$ are obtained from sampling the residuals randomly with replacement.

5. Using h_1 (or κ_1) as a smoothing parameter, compute the bootstrap version of $\hat{m}^1$, $\hat{m}^{1*}$, with the resampled data, and use it to obtain the bootstrap estimates of the shift parameters, $\hat{\gamma}_1^*, ..., \hat{\gamma}_I^*$.

6. Compute the bootstrap versions of $\hat{m}, \hat{m}_1, ..., \hat{m}_I$ and $\bar{D}$ denoted by $\hat{m}^*, \hat{m}_1^*, ..., \hat{m}_I^*$, $\bar{D}^*$, using the smoothing parameter employed in step 2 for the estimation of the regression functions.

7. Evaluate the bootstrap version of the statistic

$$C_6^* = \frac{1}{\bar{D}^*} \sum_{i=1}^{I} \sum_{j=1}^{n_I} \left[1 - \cos(\hat{m}_i^*(\Delta_{ij}) - \hat{\gamma}_i^* - \hat{m}^*(\Delta_{ij}))\right].$$

8. Repeat steps 4-7 B times to obtain the bootstrap versions of the statistic, $C_6^{*(1)}, ..., C_6^{*(B)}$ and approximate the significant value as

$$p - \text{value}^* = \frac{\#\{C_6^{*(b)} \geq Obs\}}{B}.$$

3.4 Contributions of this chapter

This chapter has been focused on different hypotheses testing problems for regression involving circular variables. In addition to surveying the existing regression models for this kind of data, the present chapter included several new proposals for significance tests and ANCOVA tests. To sum up, this section collects the principal contributions of the chapter.

To begin with, two significance tests in the circular regression context were presented in Section 3.2: one for circular predictors and linear responses and one for circular responses, with the predictors being either real-valued or circular. In the circular-linear regression context, the significance test is an adaptation of the one for real-valued variables proposed by Bowman and Azzalini (1997), where the regression function is estimated with the circular-linear estimator proposed by Di Marzio *et al.* (2009). The second test, designed for the linear-circular and the circular-circular regression settings, is based on a new statistic which makes use of the circular distance defined in (2.3).

Secondly, the main contribution of the manuscript (and primary goal of this MSc Thesis) was presented in Section 3.3, which contains proposals for ANCOVA tests in the three circular regression contexts. In each case, a test of equality and a test of parallelism were provided. In the circular-linear regression setting, the test statistics are based on their linear counterparts, but they were modified, not only by using the circular-linear regression estimator, but also by providing two different variance estimators for the circular-linear regression context. Regarding the linear-circular and circular-circular regression scenarios, new test statistics for testing equality and parallelism were proposed, again making use of the circular distance. In addition, a novel estimator of the circular variance for nonparametric regression models with circular response variables was introduced.

Chapter 4

Simulation study

This chapter contains an extensive simulation study which analyzes the performance of the nonparametric tests proposed in Chapter 3. The tests will be applied to simulated data generated under both the null and the alternative hypotheses, therefore analyzing the calibration and power of the tests. It must be highlighted that the code for the proposed tests as well as the code for the simulation study was self programmed. Section 4.1 is devoted to the study of the significance tests, whereas Section 4.2 contains the analysis of the ANCOVA tests.

4.1 Significance tests

In Section 3.2, different significance tests for circular variables were proposed, providing tools for assessing the effect of the predictor variable in each of the three regression scenarios (circular-linear, linear-circular and circular-circular). The objective of this section is to analyze the performance of the three tests, which will be applied to simulated data drawn from three different regression models (in each setting). As explained in Section 3.2, the no-effect tests relies heavily on the smoothing parameter, and because of the bias present in the estimation of the regression function, using the smoothing parameter selected by cross-validation is not adequate for the correct calibration of the tests. For that reason in this section the simulations will be carried out considering different values of the smoothing parameter. In addition, various settings will be contemplated in order to analyze the performance of the tests under different situations: several values of the sample size and the variance/concentration and different design points for the predictor variables will be used.

After B replications of the simulated data are drawn, the percentages of rejection for $\alpha = .05$ will be computed. As a way to determine if a certain percentage of rejection $\hat{p}$ obtained under the null hypothesis is large, one can easily construct a 95% confidence interval for the proportion:

$$\left(\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{B}}, \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{B}}\right), \tag{4.1}$$

where $z_{0.025}$ is the quantile of the Normal distribution function leaving a probability of .025 to the right. If the nominal level $\alpha = .05$ does not fall inside the interval, then the percentage of rejections is significantly different from 5%, indicating a poor calibration of the test (if the null hypothesis is true).

The next subsections contain, for each of the three regression settings, descriptions of the simulation scenarios and models, as well as the percentages of rejection obtained and comments on the results.

4.1.1 Circular-linear regression

The no-effect test for circular predictors and real-valued responses, presented in Section 3.2.1, is analyzed in this section. The following models will be studied:

A. $Y = \beta \sin\Theta + \varepsilon, \quad \beta = 0, .2, .3$

B. $Y = 3 + \beta \sin\Theta \cos\Theta + \varepsilon \quad \beta = 0, .25, .5$

C. $Y = 1 + \beta \sin(\Theta + 2\sin\Theta) + \varepsilon, \quad \beta = 0, .4, .6$

The errors follow a $N(0, \sigma)$ distribution where the standard deviation σ takes different values depending on the model (which will be specified in the corresponding tables with the results). The sample size takes the values 50, 100, 250 and 400. If the first value of β is used, the data are drawn from the null hypothesis. When the other values of β are considered the alternative hypothesis holds, and the effect of the predictor variable is more noticeable when employing the last value of β. Regarding the design points, three scenarios were contemplated:

- Design 1: Circular uniform distribution.
- Design 2: $vM(\pi, 1.5)$.
- Design 3: $vM(0, 1.5)$.

Figure 4.1 displays realizations of the models under H_1, with $\beta = .3$ in Model A, $\beta = .5$ in Model B and $\beta = .6$ in Model C. In addition, the curves representing the regression models under the null hypothesis are also shown. As mentioned before, the tests are applied using different values of the smoothing parameters, namely the one obtained by cross-validation (cv) and modifications of it ($\frac{1}{8}cv$, $\frac{1}{4}cv$, $2cv$ and $4cv$). The number of replications is $B = 1000$. For this value, using the confidence interval in (4.1) it can be determined that a percentage of rejection under H_0 is significantly different from α if it is higher than .065. The results are summarized next:

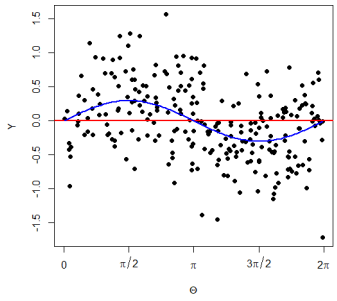
Model A The percentages of rejection obtained in Model A are collected in Table 4.1 (Design 1), Table 4.2 (Design 2) and Table 4.3 (Design 3). When the null hypothesis is true ($\beta = 0$) percentages of rejection when using cv are very large compared to the nominal level $\alpha = .05$ (around 9% or 10%). When increasing the smoothing parameter, for example when using $4cv$, percentages are slightly smaller than α for the smallest sample size ($n = 50$), but for $n = 400$ some results are significantly larger than .05. On the other hand, when $\frac{1}{8}cv$ is used, percentages of rejection under H_0 are close to the nominal level α.

If the alternative hypothesis is true, the largest percentages of rejection are obtained for cv, and increasing the smoothing parameter makes the percentages diminish. On the other hand, when reducing the smoothing parameter ($\frac{1}{4}cv$ or $\frac{1}{8}cv$) these results are lower than with cv, but still quite high. As it was expected, when augmenting the sample size and reducing the variance, the percentages of rejection increase.

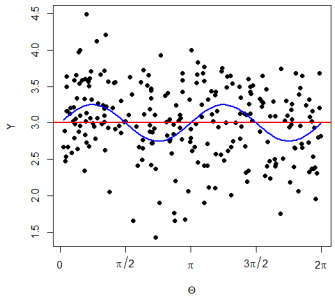
As for the differences between the three designs, under H_0 results are similar for the three scenarios. However, under the alternative hypothesis the power of the test is larger for the circular uniform design.

Model B The results obtained for Model B are shown in Table 4.4 for Design 1, Table 4.5 for Design 2 and Table 4.6 for Design 3. When the null hypothesis is true, large percentages of rejection (compared to α) are obtained with cv and the percentages decrease when either increasing or decreasing the smoothing parameter. If $\frac{1}{8}cv$ is used, percentages of rejection are close to .05. Also with $4cv$ results are close to the nominal level, although some percentages are significantly different from .05.

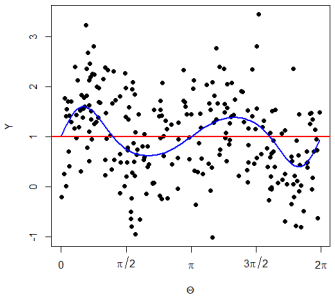
When the alternative hypothesis holds, the largest percentages of rejection are gotten with cv. Under Design 1, $\frac{1}{8}cv$ obtains not certainly large percentages, since the low concentration makes it difficult to ascertain the effect of the predictor variable. However, if Designs 2 or 3 are considered, the power of the test with $\frac{1}{8}cv$ is almost as high as with cv. On contrast, with $4cv$ results are high when using a uniform design, but they are lower if Designs 2 and 3 are considered. In general, percentages of rejection are higher with the von Mises designs.



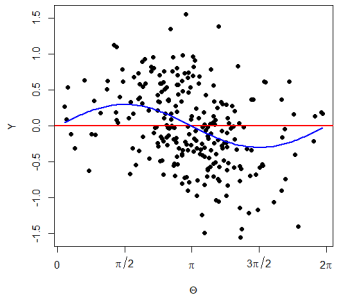
(a) Model A. Design 1



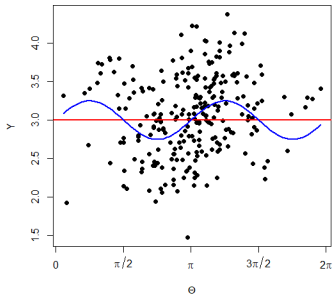
(b) Model B. Design 1



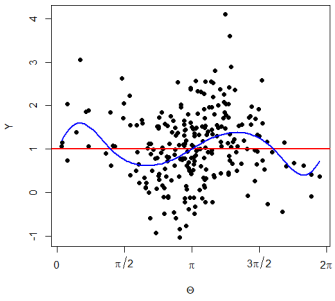
(c) Model C. Design 1



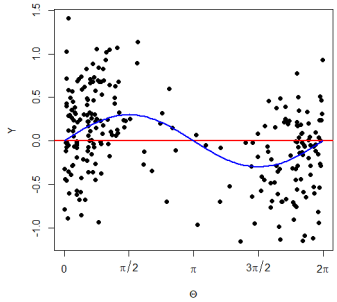
(d) Model A. Design 2



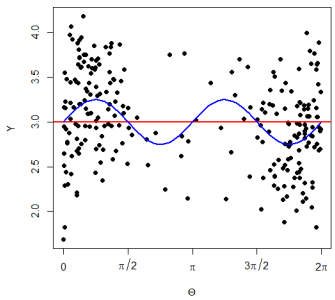
(e) Model B. Design 2



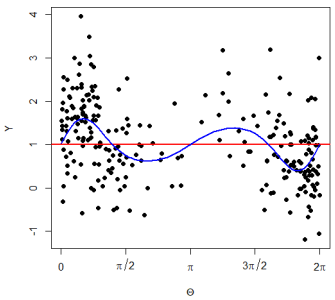
(f) Model C. Design 2



(g) Model A. Design 3



(h) Model B. Design 3



(i) Model C. Design 3

Figure 4.1: Representations of simulated data from models A (first column), B (second column) and C (third column) under the alternative hypothesis ($\beta = .3$ in A, $\beta = .5$ in B and $\beta = .6$ in C) under the three different designs, along with the true regression curves for each group (blue) and the regression lines assumed by the null hypothesis (red). Number of observations is 250 in all cases.

Model C Results for Model C are displayed in Tables 4.7, 4.8 and 4.9 respectively for Designs 1, 2 and 3. On the one hand, if the null hypothesis is true, the test obtains percentages of rejection close to the nominal level $\alpha = .5$ when using $\frac{1}{8}cv$. Results for $4cv$ are also close to α but several percentages obtained were significantly different from .05. Again, when using the smoothing parameter selected by cross-validation (cv) percentages of rejection lie around 9% and 10%.

On the other hand, if H_1 holds, the power of the test changes fairly with the different designs. As it happened in Model B, if Design 1 is used the percentages of rejection are quite low with $\frac{1}{8}cv$ but they are larger in the von Mises designs. This behavior is turned around when using $4cv$, which obtains higher percentages of rejection with the circular normal design. Predominantly, results under H_1 are higher under Design 3. This is due to the shape of the regression function, since the region where the data are concentrated in Design 3 is the region where the regression function deviates more from a horizontal line.

As a general conclusion on the smoothing parameter, with a small bandwidth ($\frac{1}{8}$ of the parameter obtained by cross-validation) the test seems to be well calibrated. Also when using $4cv$ results under H_0 are close to α, but sometimes percentages of rejection with this selection of the smoothing parameter are either too high or too low. Therefore, the best results for calibration are obtained with $\frac{1}{8}$, although in practice it is recommended to use a sequence of smoothing parameters and to obtain the significance trace of the test.

Model A. Design 1: circular uniform distribution

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.2$	$\beta=.3$	$\beta=0$	$\beta=.2$	$\beta=.3$	$\beta=0$	$\beta=.2$	$\beta=.3$	$\beta=0$	$\beta=.2$	$\beta=.3$	$\beta=0$	$\beta=.2$	$\beta=.3$
.25	50	.047	.917	.998	.052	.928	.999	.096	.922	.999	.076	.878	.997	.029	.740	.972
	100	.050	.999	1	.050	.999	1	.086	1	1	.068	.996	1	.038	.990	1
	250	.055	.999	1	.065	1	1	.096	1	1	.092	1	1	.065	1	1
	400	.052	1	1	.058	1	1	.105	1	1	.097	1	1	.071	1	1
.5	50	.055	.352	.691	.061	.369	.715	.088	.390	.714	.061	.294	.626	.031	.184	.434
	100	.037	.653	.954	.053	.675	.961	.095	.687	.957	.086	.607	.939	.053	.476	.872
	250	.051	.981	1	.066	.980	1	.096	.976	1	.087	.965	1	.054	.935	1
	400	.049	.998	1	.067	.998	1	.100	.998	1	.087	.996	1	.059	.993	1

Table 4.1: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model A (Design 1) based on 1000 simulations.

Model A. Design 2: $vM(\pi, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$
.25	50	.051	.804	.991	.062	.839	.995	.075	.831	.994	.057	.757	.980	.035	.646	.943
	100	.049	.984	1	.053	.991	1	.094	.994	1	.072	.983	1	.053	.963	1
	250	.058	1	1	.066	1	1	.096	1	1	.086	1	1	.064	1	1
	400	.051	1	1	.067	1	1	.096	1	1	.092	1	1	.069	1	1
.5	50	.060	.295	.534	.072	.323	.575	.096	.336	.589	.073	.254	.501	.046	.168	.377
	100	.052	.534	.864	.063	.576	.892	.085	.596	.905	.064	.499	.857	.037	.373	.770
	250	.053	.920	1	.064	.939	1	.088	.950	1	.082	.925	1	.055	.867	.999
	400	.059	.990	1	.067	.997	1	.104	.997	1	.091	.994	1	.061	.979	1

Table 4.2: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model A (Design 2) based on 1000 simulations.

Model A. Design 3: $vM(0, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$	$\beta = 0$	$\beta = .2$	$\beta = .3$
.25	50	.039	.808	.987	.046	.830	.991	.063	.829	.990	.046	.759	.984	.033	.650	.959
	100	.054	.987	1	.062	.992	1	.075	.991	1	.057	.985	1	.044	.960	1
	250	.059	1	1	.072	1	1	.106	1	1	.096	1	1	.063	1	1
	400	.051	1	1	.065	1	1	.099	1	1	.074	1	1	.053	1	1
.5	50	.050	.308	.555	.058	.348	.592	.077	.338	.594	.057	.253	.509	.038	.175	.387
	100	.060	.542	.899	.069	.576	.913	.094	.592	.918	.076	.512	.874	.054	.401	.800
	250	.049	.907	1	.061	.940	1	.103	.954	1	.089	.923	.999	.055	.845	.999
	400	.052	.985	1	.062	.993	1	.105	.994	1	.093	.990	1	.051	.979	1

Table 4.3: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model A (Design 3) based on 1000 simulations.

Model B. Design 1: circular uniform distribution

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
.25	50	.047	.167	.577	.050	.245	.830	.081	.429	.960	.048	.362	.929	.021	.218	.822
	100	.053	.290	.977	.056	.503	.997	.091	.740	1	.079	.699	1	.052	.571	.999
	250	.055	.841	1	.061	.959	1	.097	.994	1	.095	.994	1	.065	.984	1
	400	.057	.986	1	.062	1	1	.091	1	1	.079	1	1	.056	1	1
.5	50	.056	.069	.163	.068	.094	.231	.107	.173	.388	.083	.127	.321	.028	.059	.201
	100	.045	.094	.290	.053	.137	.502	.090	.262	.733	.080	.237	.690	.056	.147	.581
	250	.036	.194	.855	.042	.333	.967	.089	.562	.993	.078	.527	.991	.043	.406	.980
	400	.049	.337	.993	.062	.566	.999	.107	.790	.999	.092	.759	.999	.068	.668	.999

Table 4.4: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model B (Design 1) based on 1000 simulations.

Model B. Design 2: $vM(\pi, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
.25	50	.061	.464	.964	.064	.466	.972	.072	.472	.965	.061	.396	.946	.036	.302	.894
	100	.053	.801	1	.059	.802	1	.080	.805	1	.063	.745	.999	.050	.633	.999
	250	.045	.995	1	.056	.996	1	.092	.997	1	.076	.997	1	.050	.990	1
	400	.048	1	1	.057	1	1	.089	1	1	.075	1	1	.045	1	1
.5	50	.048	.154	.464	.056	.166	.477	.078	.172	.469	.063	.126	.398	.051	.078	.315
	100	.048	.249	.770	.055	.259	.779	.078	.277	.796	.062	.227	.736	.044	.166	.631
	250	.047	.623	.997	.060	.644	.997	.105	.676	.998	.093	.616	.997	.052	.494	.992
	400	.063	.807	1	.078	.820	1	.118	.849	1	.108	.822	1	.068	.732	1

Table 4.5: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model B (Design 2) based on 1000 simulations.

Model B. Design 3: $vM(0, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
.25	50	.045	.463	.969	.053	.485	.967	.081	.471	.960	.058	.400	.937	.041	.308	.882
	100	.053	.798	1	.065	.820	1	.099	.805	1	.079	.761	1	.051	.660	.999
	250	.058	.995	1	.067	.995	1	.105	.998	1	.080	.994	1	.059	.985	1
	400	.053	1	1	.067	1	1	.108	1	1	.092	1	1	.063	1	1
.5	50	.058	.138	.465	.069	.158	.471	.087	.164	.469	.074	.141	.407	.056	.089	.315
	100	.058	.293	.783	.074	.300	.803	.102	.320	.826	.074	.263	.765	.063	.194	.660
	250	.056	.611	.991	.058	.625	.997	.089	.673	.998	.066	.616	.996	.052	.495	.990
	400	.060	.819	.999	.076	.835	.999	.115	.865	1	.104	.838	1	.078	.740	1

Table 4.6: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model B (Design 3) based on 1000 simulations.

Model C. Design 1: circular uniform distribution

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.4$	$\beta=.6$	$\beta=0$	$\beta=.4$	$\beta=.6$	$\beta=0$	$\beta=.4$	$\beta=.6$	$\beta=0$	$\beta=.4$	$\beta=.6$	$\beta=0$	$\beta=.4$	$\beta=.6$
.5	50	.067	.224	.524	.080	.364	.762	.117	.584	.914	.080	.512	.866	.042	.315	.722
	100	.057	.531	.955	.069	.766	.995	.115	.926	1	.084	.904	1	.050	.822	.996
	250	.053	.988	1	.062	1	1	.090	1	1	.083	1	1	.059	1	1
	400	.056	1	1	.069	1	1	.096	1	1	.092	1	1	.061	1	1
.75	50	.066	.127	.218	.075	.173	.349	.093	.062	.228	.470	.289	.542	.022	.116	.319
	100	.060	.215	.544	.069	.360	.773	.106	.590	.911	.099	.547	.890	.066	.414	.804
	250	.051	.656	.988	.053	.848	1	.092	.962	1	.088	.948	1	.059	.901	1
	400	.053	.932	1	.056	.982	1	.096	.997	1	.090	.996	1	.068	.991	1

Table 4.7: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model C (Design 1) based on 1000 simulations.

Model C. Design 2: $vM(\pi, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$
.5	50	.046	.554	.836	.054	.565	.836	.071	.539	.839	.058	.456	.778	.036	.354	.687
	100	.047	.845	.990	.063	.854	.990	.102	.866	.993	.081	.820	.984	.058	.730	.969
	250	.048	.999	1	.061	.998	1	.103	.999	1	.085	.997	1	.063	.988	1
	400	.055	1	1	.075	1	1	.122	1	1	.098	1	1	.070	1	1
.75	50	.058	.280	.522	.064	.294	.527	.069	.317	.526	.055	.244	.459	.037	.173	.342
	100	.052	.489	.855	.063	.512	.858	.090	.521	.868	.079	.449	.817	.056	.510	.719
	250	.058	.901	.998	.061	.909	.998	.083	.934	.999	.063	.907	.998	.039	.833	.993
	400	.056	.985	1	.064	.985	1	.113	.988	1	.088	.983	1	.068	.969	1

Table 4.8: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model C (Design 2) based on 1000 simulations.

Model C. Design 3: $vM(0, 1.5)$

Test of no effect. Circular-linear regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$	$\beta = 0$	$\beta = .4$	$\beta = .6$
.5	50	.048	.651	.947	.053	.695	.972	.062	.749	.978	.041	.681	.968	.029	.576	.933
	100	.039	.942	.999	.044	.958	1	.081	.969	1	.065	.962	1	.053	.927	1
	250	.051	1	1	.063	1	1	.104	1	1	.086	1	1	.055	1	1
	400	.048	1	1	.059	1	1	.094	1	1	.080	1	1	.060	1	1
.75	50	.051	.336	.636	.063	.364	.685	.083	.392	.731	.065	.343	.661	.047	.279	.555
	100	.054	.658	.953	.068	.686	.967	.088	.716	.974	.070	.675	.968	.048	.584	.939
	250	.059	.982	1	.071	.990	1	.107	.991	1	.086	.990	1	.059	.979	1
	400	.056	1	1	.061	1	1	.093	1	1	.082	1	1	.043	1	1

Table 4.9: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-linear regression in Model C (Design 3) based on 1000 simulations.

4.1.2 Linear-circular regression

This section is devoted to the analysis of the performance of the test proposed in Section 3.2.2 in the specific case of a linear predictor. The models studied are the following:

A. $\Phi = [\beta \tan(X) + \varepsilon](\text{mod} 2\pi), \quad \beta = 0, .25, .5,$

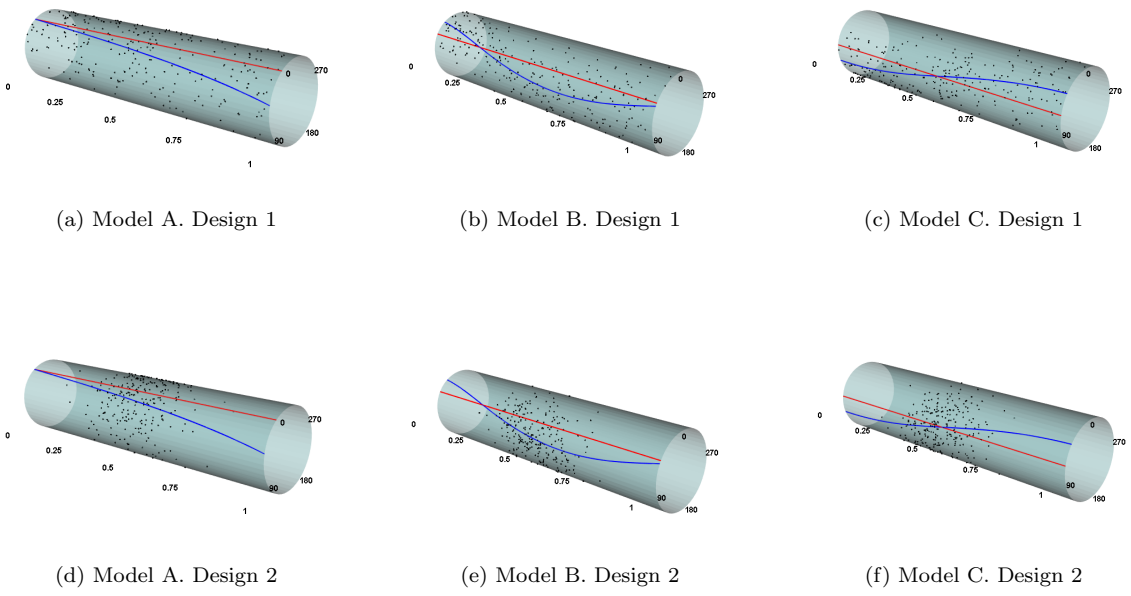
B. $\Phi = [\frac{\pi}{4} + \beta \sin(4X - 1) + \varepsilon](\text{mod} 2\pi), \quad \beta = 0, .5, 1,$

C. $\Phi = [\frac{3\pi}{8} + \beta \cos(3X) + \varepsilon](\text{mod} 2\pi), \quad \beta = 0, .5, 1,$

where the errors ε follow a von Mises distribution with mean direction 0 and concentration κ. The values of κ are different for each model and are specified in the corresponding tables with the results. Again, the first value of β corresponds to the null hypothesis being true, and with the other two values the performance of the test under H_1 is being studied. Two designs of the predictor variable are considered:

- Design 1: $U(0, 1)$.
- Design 2: $N(.5, .1)$.

Figure 4.2 shows realizations of simulated data under the alternative hypothesis ($\beta = .5$) with the true regression functions. The regression lines under the null hypothesis are also represented. It can be seen that when Design 2 is considered it is difficult to discern if the data are drawn from the null or the alternative hypothesis.



(a) Model A. Design 1 (b) Model B. Design 1 (c) Model C. Design 1

(d) Model A. Design 2 (e) Model B. Design 2 (f) Model C. Design 2

Figure 4.2: Representations on the cylinder of simulated data from models A, B and C under the alternative hypothesis ($\beta = .5$) under Design 1 (top row) and Design 2 (bottom row). True regression curves (blue) and regression lines under H_0. Number of observations is 250 and the value of κ is 2 in Model A and 4 in Models B and C. Circular units are in degrees.

As in the previous section, different smoothing parameters will be contemplated: cv (the bandwidth selected by cross-validation), $\frac{1}{4}cv$, $\frac{1}{2}cv$, $2cv$ and $4cv$. In this case, since the test includes a bootstrap

procedure, the number of bootstrap resamples is 500. Percentages of rejection are computed after applying the test to $B = 500$ replications of the simulated data, therefore, considering the confidence interval for the proportion in (4.1), a percentage of rejection will be considered large (compared to the nominal level .05) when it is .074 or larger. A summary of the results is given next.

Model A Percentages of rejection corresponding to Model A are displayed on Table 4.10 (Design 1) and Table 4.11 (Design 2). When $\beta = 0$ (i.e. when H_0 is true), percentages of rejection when using cv are many times significantly greater than $\alpha = .05$ (around 8% of rejections). On the other hand, when augmenting or diminishing the smoothing parameter, percentages of rejection turn closer to α. In fact, when using $\frac{1}{4}cv$ or $4cv$, the nominal level falls in all the confidence intervals of the form (4.1) for the correspondent percentages.

If the alternative hypothesis is true, percentages of rejection for cv are the highest ones, however, with larger smoothing parameters such as $4cv$ the percentages are almost as high. On the contrary, when using smaller bandwidths, like $\frac{1}{4}cv$, the power of the test is lower. Under Design 2 the percentages of rejection are much smaller than under Design 1, which is not surprising given that, as mentioned before, when the data are concentrated in the middle of the cylinder the regression functions under H_0 and H_1 are not very different. Regarding the concentration of the errors, it can be seen that the percentages of rejection under H_1 are larger when the concentration is higher.

Model B The results for Model B can be found in Table 4.12 for Design 1 and Table 4.12 for Design 2. Under the null hypothesis, as in Model A, the test seems to be well calibrated for $\frac{1}{4}cv$ and $4cv$, while when using the smoothing parameter selected by cross-validation large percentages of rejection are obtained (even reaching 10% of rejections).

On the other side, if H_1 is true, percentages of rejection are larger with $4cv$ than with $\frac{1}{4}cv$. This means that when the smoothing parameter is small it is more difficult for the test to reject H_0 when it is false. As before, percentages obtained for Design 2 and under H_1 are lower than for Design 1.

Model C Tables 4.14 and 4.15 contain percentages of rejection for Model C, under Designs 1 and 2 respectively. When the null hypothesis holds, using $\frac{1}{4}cv$ and $4cv$ as bandwidths makes percentages of rejection close to the level .05. When the alternative hypothesis is true, out of the two parameters that make the test well calibrated under the null, $4cv$ obtains higher percentages of rejection.

Summing up, the parameters $\frac{1}{4}cv$ and $4cv$ make the significance test well calibrated under the null hypothesis but, in general, with $4cv$ the test has more power under the alternative, so it should be considered as a usable bandwidth for the test. However, as for the test for circular-linear data, the test in practice should be carried out using several smoothing parameters and obtaining the significance trace of the test.

Model A. Design 1: $U(0,1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.25$	$\beta=.5$	$\beta=0$	$\beta=.25$	$\beta=.5$	$\beta=0$	$\beta=.25$	$\beta=.5$	$\beta=0$	$\beta=.25$	$\beta=.5$	$\beta=0$	$\beta=.25$	$\beta=.5$
4	50	.046	.216	.652	.066	.242	.722	.074	.266	.760	.068	.254	.756	.044	.240	.754
	100	.050	.390	.936	.064	.444	.968	.096	.472	.972	.084	.468	.972	.064	.450	.972
	250	.058	.822	1	.074	.860	1	.084	.876	1	.078	.868	1	.064	.864	1
	400	.048	.952	1	.068	.970	1	.080	.974	1	.068	.974	1	.054	.974	1
2	50	.042	.106	.338	.050	.124	.376	.072	.140	.424	.060	.134	.424	.048	.120	.418
	100	.064	.190	.558	.062	.222	.626	.092	.246	.640	.076	.230	.642	.056	.214	.628
	250	.062	.440	.938	.068	.506	.958	.084	.538	.962	.078	.524	.962	.066	.512	.960
	400	.054	.660	.996	.068	.706	1	.082	.746	1	.068	.742	1	.062	.734	1

Table 4.10: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model A (Design 1) based on 500 simulations.

Model A. Design 2: $N(.5, .1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$	$\beta = 0$	$\beta = .25$	$\beta = .5$
4	50	.040	.090	.154	.046	.100	.166	.062	.118	.196	.056	.112	.192	.042	.092	.178
	100	048	.070	.180	.058	.094	.206	.072	.110	.224	.062	.108	.230	.050	.090	.220
	250	.052	.102	.396	.048	.118	.454	.062	.150	.498	.060	.148	.496	.060	.132	.484
	400	.044	.222	.604	.052	.250	.648	.068	.288	.694	.062	.276	.688	.048	.266	.678
2	50	.050	.040	.064	.054	.044	.068	.064	.050	.074	.066	.050	.074	.058	.044	.070
	100	.060	.056	.108	.068	.066	.118	.080	.084	.142	.078	.072	.124	.064	.068	.122
	250	.064	.072	.190	.074	.080	.212	.086	.096	.252	.076	.086	.234	.064	.072	.236
	400	.040	.092	.286	.058	.110	.330	.084	.136	.364	.072	.134	.360	.056	.124	.340

Table 4.11: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model A (Design 2) based on 500 simulations.

Model B. Design 1: $U(0,1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$
4	50	.048	.604	.978	.058	.800	.996	.080	.878	1	.066	.856	1	.058	.800	.996
	100	.058	.948	1	.066	.986	1	.090	.996	1	.074	.992	1	.066	.986	1
	250	.052	1	1	.060	1	1	.086	1	1	.064	1	1	.060	1	1
	400	.060	1	1	.078	1	1	.100	1	1	.082	1	1	.072	1	1
2	50	.048	.250	.730	.054	.412	.908	.068	.524	.950	.070	.494	.950	.054	.412	.908
	100	.052	.532	.984	.066	.726	1	.100	.818	1	.086	.796	1	.066	.726	1
	250	.050	.956	1	.062	.992	1	.076	.994	1	.070	.994	1	.062	.992	1
	400	.042	.998	1	.052	1	1	.066	1	1	.056	1	1	.052	1	1

Table 4.12: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model B (Design 1) based on 500 simulations.

Model B. Design 2: $N(.5, .1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$
4	50	.032	.248	.626	.048	.272	.700	.074	.314	.752	.068	.308	.764	.058	.286	.734
	100	.048	.354	.862	.046	.408	.926	.054	.468	.954	.052	.462	.952	.040	.438	.940
	250	.052	.708	1	.056	.800	1	.078	.856	1	.072	.850	1	.060	.832	1
	400	.044	.880	1	.062	.936	1	.078	.968	1	.056	.962	1	.048	.956	1
2	50	.046	.108	.332	.056	.106	.356	.070	.124	.402	.066	.130	.420	.052	.128	.404
	100	.044	.170	.494	.052	.186	.582	.066	.214	.642	.062	.212	.644	.040	.192	.618
	250	.046	.384	.876	.056	.446	.940	.076	.502	.968	.072	.500	.964	.060	.472	.958
	400	.040	.506	.976	.050	.618	.992	.072	.690	.996	.056	.686	.994	.040	.662	.994

Table 4.13: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model B (Design 2) based on 500 simulations.

Model C. Design 1: $U(0,1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$
4	50	.056	.796	.984	.064	.846	.986	.080	.860	.990	.078	.858	.990	.056	.854	.990
	100	.040	.980	1	.042	.990	1	.064	.992	1	.054	.992	1	.040	.992	1
	250	.042	1	1	.064	1	1	.074	1	1	.060	1	1	.048	1	1
	400	.038	1	1	.044	1	1	.068	1	1	.048	1	1	.042	1	1
2	50	.042	.410	.980	.054	.446	.718	.070	.494	.772	.066	.486	.772	.056	.474	.756
	100	.066	.734	1	.080	.792	.966	.100	.808	.974	.086	.806	.974	.068	.802	.974
	250	.044	.978	1	.062	.990	1	.066	.990	1	.056	.990	1	.046	.990	1
	400	.044	1	1	.062	1	1	.082	1	1	.078	1	1	.054	1	1

Table 4.14: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model C (Design 1) based on 500 simulations.

Model C. Design 2: $N(.5, .1)$

Test of no effect. Linear-circular regression

κ	n	$\frac{1}{4}cv$			$\frac{1}{2}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$	$\beta=0$	$\beta=.5$	$\beta=1$
4	50	.058	.178	.360	.064	.196	.394	.072	.242	.426	.066	.232	.424	.052	.218	.412
	100	.040	.360	.694	.054	.410	.730	.084	.452	.772	.080	.450	.766	.066	.440	.760
	250	.034	.788	.900	.050	.834	.986	.068	.862	.990	.058	.866	.992	.046	.860	.992
	400	.050	.918	1	.060	.956	1	.080	.968	1	.070	.964	1	.060	.964	1
2	50	.046	.118	.198	.056	.128	.202	.060	.148	.232	.058	.154	.240	.052	.148	.232
	100	.036	.210	.352	.040	.226	.376	.054	.238	.408	.044	.244	.420	.036	.234	.410
	250	.044	.386	.704	.046	.426	.716	.056	.456	.754	.052	.450	.754	.042	.446	.744
	400	.050	.578	.900	.060	.632	.912	.072	.676	.936	.064	.676	.944	.050	.664	.938

Table 4.15: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for linear-circular regression in Model C (Design 2) based on 500 simulations.

4.1.3 Circular-circular regression

The last one of the nonparametric significance tests, for circular covariates and circular responses, will be analyzed in this section. The models studied are

A. $\Phi = [\frac{\pi}{4} + \beta \exp(\sin(\Theta + \pi)) + \varepsilon] (\mathrm{mod} 2\pi), \quad \beta = 0, .3, .5$

B. $\Phi = [\frac{3\pi}{4} + \beta \sin(2\Theta + 2\sin(\Theta + \frac{\pi}{2})) + \varepsilon] (\mathrm{mod} 2\pi), \quad \beta = 0, .35, .5$

C. $\Phi = [\frac{3\pi}{2} - \beta \cos(2\Theta + 4\cos\Theta + 3\sin\Theta) + \varepsilon] (\mathrm{mod} 2\pi), \quad \beta = 0, .3, .5$

When the value of β is 0, the predictor variable Θ has no effect on the response, therefore the null hypothesis is true. On the contrary, with the other values of β the alternative hypothesis is being considered. The errors ε follow a von Mises distribution with zero mean and variable concentration (depending on the model). The values of the concentration for each model are specified in the tables containing the results, which will be presented later. The data corresponding to the predictor variable are drawn from three different distributions:

- Design 1: Circular uniform distribution.
- Design 2: $vM(\pi, 2)$.
- Design 3: $vM(0, 2)$.

Simulated data from all the models and designs drawn under the alternative hypothesis ($\beta = .3$ in Models A and C, $\beta = .35$ in Model B) are represented in Figure 4.3. The true regression functions are also included, as well as the regression curves assumed under the null hypothesis.

The number of bootstrap resamples is 500 in all cases. As in the previous section, the percentages of rejection for the nominal level $\alpha = .05$ are computed after applying the test to 500 realizations of the simulated data. Consequently, as before, a percentage under H_0 will be significantly larger than α when it is .074 or more. In this case, the considered smoothing parameters are cv, $\frac{1}{8}cv$, $\frac{1}{4}cv$, $2cv$ and $4cv$. A recapitulation of the results follows.

Model A Results obtained for Model A are collected in Table 4.16 (Design 1), Table 4.17 (Design 2) and Table 4.18 (Design 3). Under the null hypothesis, where the data are generated from a constant regression function, percentages of rejection are close to $\alpha = .05$ when using $\frac{1}{8}cv$ and $4cv$, although a couple percentages of rejection are slightly high. On contrast, when using the other smoothing parameters results under H_0 are usually large, specially for cv (surpassing 11% of rejections several times).

On the other hand, under the alternative hypothesis, with $4cv$ percentages of rejection are low compared to the ones corresponding to other bandwidths. Diminishing the smoothing parameter makes the percentages of rejection slightly lower. In addition, the power of the test is higher when considering Design 1, although percentages of rejection for the von Mises designs are not very low in comparison with the uniform design.

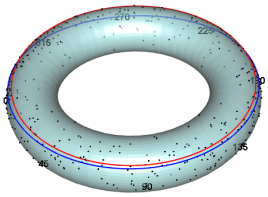
Model B Tables 4.19, 4.20 and 4.21 show the percentages of rejection obtained for Model B under Designs 1, 2 and 3, respectively. Results are similar to the ones obtained in Model A. Under H_0, with cv the percentages of rejection are large compared to the nominal level .05. When incrementing or diminishing the bandwidth, results are closer to α. For instance, for $\frac{1}{8}cv$ and $4cv$ only a couple percentages are significantly larger than .05.

When the alternative hypothesis is true, and the regression function changes with the values of Θ, the percentages of rejection obtained for $\frac{1}{8}cv$ are larger than the ones gotten with $4cv$. In this case, percentages for the von Mises designs under H_1 are larger than the ones obtained in the uniform setting. In fact, with Design 3 the power of the test is generally higher. This is due to the shape of the regression function under the alternative hypothesis: it departs more from a constant in the region where the data are concentrated in Design 3.

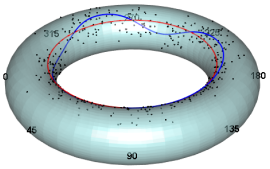
Model C The results for this model are displayed in Table 4.22 for Design 1, Table 4.23 for Design 2 and Table 4.24 for Design 3. Under H_0, as in the previous models, if the smoothing parameter selected by cross-validation (cv) is used, large percentages of rejection are obtained, which lie between 8% and 12% of rejections. However, when incrementing the smoothing parameter, percentages are closer to α. Using a smaller parameter than cv also gives results closer to the nominal level. Specifically, percentages of rejection for $\frac{1}{8}cv$ and $4cv$ are the closest to .05.

Regarding the performance of the test under the alternative hypothesis, the power of the test when using $\frac{1}{8}cv$ is generally higher than the power for $4cv$. Comparing results between the different designs, the lowest percentages of rejection under H_1 are obtained with Design 2. However, in all cases percentages of rejection are close to one at least when the sample size is large ($n = 250, 400$).

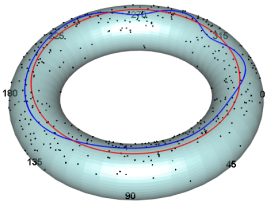
To sum up, if the smoothing parameter is selected by cross-validation the test will not be well calibrated. However, using $\frac{1}{8}cv$ as the bandwidth seems the best option since, for this parameter, the test obtains percentages of rejection close to α under H_0 and the power of the test under H_1 is still high compared to the results obtained for other smoothing parameters. Still, in practice the test must be applied using several bandwidths and obtaining the significance trace of the test, rejecting the null hypothesis if it is rejected for the majority of the smoothing parameters.



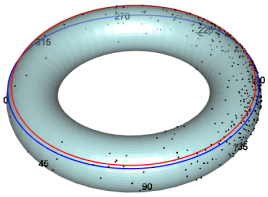
(a) Model A. Design 1



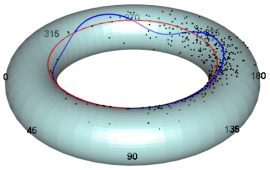
(b) Model B. Design 1



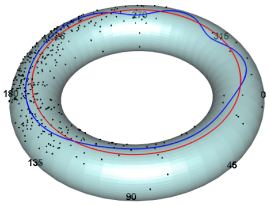
(c) Model C. Design 1



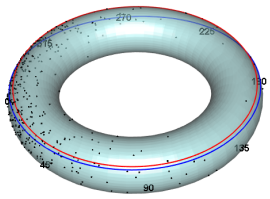
(d) Model A. Design 2



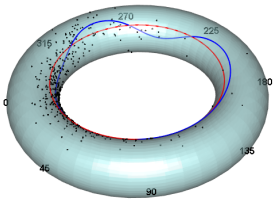
(e) Model B. Design 2



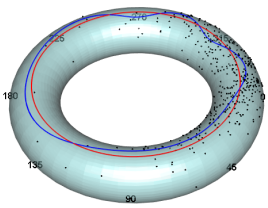
(f) Model C. Design 2



(g) Model A. Design 3



(h) Model B. Design 3



(i) Model C. Design 3

Figure 4.3: Representations on the torus of simulated data from models A, B and C under the alternative hypothesis (second value of β) under Design 1 (top row), Design 2 (middle row) and Design 3 (bottom row). The true regression curves (blue) and regression lines under H_0 (red) are included. Number of observations is 400 and the value of κ is 3 in Model A, 5 in Model B and 4 in Model C. Circular units are in degrees.

Model A. Design 1: circular uniform distribution

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$
4	50	.064	.606	.948	.068	.620	.956	.088	.632	.950	.060	.524	.912	.030	.324	.834
	100	.050	.914	1	.054	.920	1	.086	.928	1	.066	.880	1	.032	.782	.998
	250	.046	1	1	.062	1	1	.086	1	1	.072	1	1	.048	1	1
	400	.064	1	1	.082	1	1	.126	1	1	.104	1	1	.082	1	1
2	50	.042	.394	.768	.048	.418	.808	.078	.418	.800	.050	.280	.692	.032	.154	.514
	100	.040	.670	.988	.046	.682	.990	.074	.712	.992	.058	.626	.982	.028	.434	.924
	250	.046	.978	1	.052	.974	1	.084	.974	1	.068	.962	1	.054	.934	1
	400	.048	.996	1	.062	.996	1	.106	.996	1	.100	.996	1	.074	.996	1

Table 4.16: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model A (Design 1) based on 500 simulations.

Model A. Design 2: $vM(\pi, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$
4	50	.052	.408	.824	.056	.414	.842	.066	.418	.832	.050	.350	.762	.044	.278	.662
	100	.072	.728	.990	.042	.758	.990	.064	.760	.990	.050	.684	.986	.036	.568	.970
	250	.054	.984	1	.062	.992	1	.098	.994	1	.088	.982	1	.050	.964	1
	400	.054	1	1	.074	1	1	.108	1	1	.102	1	1	.078	1	1
2	50	.056	.238	.548	.072	.254	.582	.072	.236	.546	.066	.188	.480	.062	.152	.382
	100	.074	.498	.912	.072	.524	.916	.090	.534	.894	.062	.464	.842	.052	.356	.758
	250	.056	.908	1	.064	.912	1	.096	.930	1	.076	.890	1	.064	.806	.998
	400	.060	.972	1	.060	.974	1	.098	.982	1	.090	.970	1	.070	.944	1

Table 4.17: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model A (Design 2) based on 500 simulations.

Model A. Design 3: $vM(0, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$
4	50	.048	.454	.866	.064	.470	.872	.082	.454	.846	.072	.388	.800	.054	.298	.704
	100	.068	.734	.990	.078	.744	.994	.094	.734	.992	.080	.674	.986	.058	.562	.976
	250	.054	.998	1	.060	.994	1	.094	.998	1	.076	.986	1	.054	.972	1
	400	.052	1	1	.078	1	1	.118	1	1	.090	.998	1	.064	.998	1
2	50	.058	.256	.578	.064	.268	.596	.076	.256	.544	.066	.202	.446	.044	.140	.362
	100	.068	.506	.884	.090	.510	.908	.106	.506	.902	.084	.428	.852	.072	.316	.758
	250	.058	.914	1	.070	.902	1	.112	.914	1	.088	.874	1	.064	.780	1
	400	.060	.992	1	.064	.992	1	.098	.992	1	.082	.984	1	.056	.958	1

Table 4.18: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model A (Design 3) based on 500 simulations.

Model B. Design 1: circular uniform distribution

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$
5	50	.042	.266	.518	.058	.338	.650	.070	.498	.806	.050	.394	.738	.032	.258	.570
	100	.074	.606	.908	.076	.738	.976	.108	.878	.992	.076	.842	.990	.046	.734	.978
	250	.050	.996	1	.058	1	1	.098	1	1	.082	1	1	.056	1	1
	400	.060	1	1	.068	1	1	.122	1	1	.102	1	1	.060	1	1
4	50	.068	.298	.480	.078	.336	.576	.118	.436	.712	.094	.366	.632	.048	.244	.462
	100	.038	.488	.834	.056	.604	.932	.088	.768	.980	.054	.714	.976	.038	.598	.928
	250	.044	.934	1	.046	.982	1	.088	.998	1	.070	.998	1	.048	.986	1
	400	.058	.998	1	.060	1	1	.102	1	1	.092	1	1	.058	1	1

Table 4.19: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model B (Design 1) based on 500 simulations.

Model B. Design 2: $vM(\pi, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$
5	50	.054	.330	.638	.082	.378	.700	.090	.428	.724	.068	.374	.692	.056	.304	.616
	100	.058	.716	.968	.082	.782	.982	.106	.818	.982	.074	.714	.984	.058	.698	.970
	250	.052	.994	1	.064	.996	1	.098	.998	1	.084	.998	1	.072	.992	1
	400	.072	1	1	.082	1	1	.126	1	1	.106	1	1	.080	1	1
4	50	.064	.328	.582	.070	.364	.628	.047	.388	.674	.082	.374	.638	.054	.304	.556
	100	.054	.572	.904	.058	.620	.932	.064	.666	.944	.070	.774	.922	.038	.698	.878
	250	.040	.966	1	.064	.978	1	.108	.984	1	.060	.996	1	.054	.992	1
	400	.052	1	1	.076	1	1	.108	1	1	.094	1	1	.068	1	1

Table 4.20: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model B (Design 2) based on 500 simulations.

Model B. Design 3: $vM(0, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$	$\beta = 0$	$\beta = .35$	$\beta = .5$
5	50	.070	.414	.696	.068	.456	.762	.086	.472	.806	.068	.422	.758	.054	.350	.668
	100	.076	.704	.936	.076	.740	.952	.094	.760	.974	.092	.732	.974	.070	.678	.954
	250	.058	.992	1	.070	.998	1	.110	1	1	.080	1	1	.062	.998	1
	400	.070	1	1	.084	1	1	.118	1	1	.112	1	1	.068	1	1
4	50	.056	.272	.538	.080	.318	.594	.092	.334	.636	.062	.298	.578	.060	.254	.500
	100	.056	.580	.904	.062	.636	.934	.078	.682	.950	.056	.624	.940	.054	.538	.896
	250	.058	.972	.998	.052	.976	1	.078	.982	1	.082	.978	1	.054	.968	1
	400	.056	1	1	.062	1	1	.108	1	1	.092	1	1	.066	1	1

Table 4.21: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model B (Design 3) based on 500 simulations.

Model C. Design 1: circular uniform distribution

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$	$\beta=0$	$\beta=.5$	$\beta=.5$	$\beta=0$	$\beta=.3$	$\beta=.5$
4	50	.056	.334	.750	.058	.360	.786	.092	.452	.866	.068	.376	.794	.040	.272	.668
	100	.060	.678	.996	.062	.754	.998	.092	.816	1	.068	.778	1	.048	.716	.982
	250	.064	.992	1	.074	.998	1	.116	1	1	.104	1	1	.068	.998	1
	400	.058	1	1	.062	1	1	.094	1	1	.088	1	1	.058	1	1
3	50	.058	.264	.612	.072	.312	.674	.100	.364	.760	.064	.286	.698	.036	.186	.518
	100	.056	.534	.938	.060	.588	.954	.078	.670	.976	.062	.604	.968	.034	.486	.924
	250	.050	.954	1	.058	.962	1	.108	.986	1	.104	.982	1	.064	.968	1
	400	.056	.998	1	.052	1	1	.076	1	1	.068	1	1	.044	1	1

Table 4.22: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model C (Design 1) based on 500 simulations.

Model C. Design 2: $vM(\pi, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .5$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$
4	50	.060	.252	.608	.080	.282	.668	.102	.304	.670	.080	.252	.614	.062	.186	.550
	100	.046	.534	.932	.052	.564	.944	.076	.570	.952	.052	.504	.926	.036	.410	.896
	250	.052	.926	1	.062	.944	1	.078	.950	1	.068	.938	1	.042	.894	1
	400	.066	.994	1	.074	.994	1	.100	.998	1	.082	.998	1	.066	.994	1
3	50	.048	.176	.470	.050	.222	.508	.064	.228	.514	.062	.186	.450	.056	.142	.362
	100	.062	.420	.782	.076	.436	.804	.112	.490	.826	.102	.436	.786	.078	.356	.730
	250	.042	.788	.998	.054	.812	1	.088	.856	1	.076	.796	1	.054	.742	.998
	400	.074	.946	1	.080	.950	1	.122	.964	1	.098	.952	1	.072	.906	1

Table 4.23: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model C (Design 2) based on 500 simulations.

Model C. Design 3: $vM(0, 2)$

Test of no effect. Circular-circular regression

σ	n	$\frac{1}{8}cv$			$\frac{1}{4}cv$			cv			$2cv$			$4cv$		
		$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$	$\beta = 0$	$\beta = .5$	$\beta = .5$	$\beta = 0$	$\beta = .3$	$\beta = .5$
4	50	.052	.298	.626	.056	.336	.718	.088	.398	.816	.068	.372	.796	.056	.320	.734
	100	.044	.550	.938	.056	.600	.980	.064	.702	.990	.050	.670	.990	.040	.614	.984
	250	.056	.968	1	.072	.988	1	.100	.998	1	.084	.996	1	.066	.992	1
	400	.064	.998	1	.068	.998	1	.106	1	1	.092	1	1	.066	1	1
3	50	.050	.198	.488	.068	.222	.564	.096	.272	.686	.092	.230	.648	.058	.172	.566
	100	.058	.410	.840	.066	.458	.896	.098	.542	.956	.068	.520	.958	.050	.450	.930
	250	.044	.848	1	.050	.900	1	.088	.950	1	.082	.940	1	.054	.916	1
	400	.068	.990	1	.078	.996	1	.128	1	1	.114	1	1	.074	.994	1

Table 4.24: Percentages of rejections (for $\alpha = .05$) for the nonparametric significance test for circular-circular regression in Model C (Design 3) based on 500 simulations.

4.2 ANCOVA tests

In Section 3.3 nonparametric equality and parallelism tests were proposed for the three different circular regression settings. The performance of those tests will be analyzed in this section. First, the simulation scenarios will be presented. Afterwards, the obtained results will be shown and examined.

4.2.1 Circular-linear regression

In this section the equality and parallelism tests for circular predictors and linear responses described in Section 3.3.1 will be considered. For both tests, data will be drawn from three different models under both the null and the alternative hypothesis. In all cases 1000 replications of the data will be simulated and the percentages of rejection for $\alpha = .05$ will be recorded, so a null hypothesis is rejected when the associated p-value is smaller than α. Taking into account the number of replications and using the 95% confidence interval in (4.1), a percentage of rejection will be considered significantly larger that α when it is greater than .065.

For the equality test, the three following models will be studied:

A. Group 1: $Y = \sin\Theta + \varepsilon$.
   Group 2: $Y = \beta\cos\Theta + \sin\Theta + \varepsilon, \quad \beta = 0, .2, .3$

B. Group 1: $Y = \exp(\sin\Theta) + 2 + \varepsilon$.
   Group 2: $Y = \beta\exp(\sin\Theta) + 2 + \varepsilon, \quad \beta = 1, 1.15, 1.3$

C. Group 1: $Y = \cos\Theta\sin\Theta + \varepsilon$.
   Group 2: $Y = \beta\cos\Theta\sin\Theta + \varepsilon, \quad \beta = 1, 1.5, 1.75$

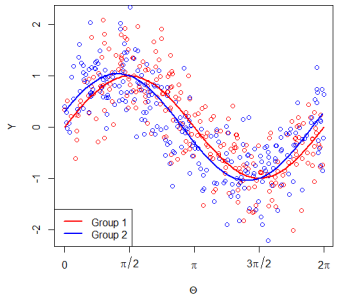
The number of observations in each group takes the values $50, 100,$ and 250. The errors follow a $N(0, \sigma)$ distribution, where σ takes different values depending on the model. When using the first value of β, the null hypothesis is being considered, while the other values correspond to H_1 being true.

At the same time, for the test of parallelism the same models are contemplated, but adding a shift of .2 to the responses in the second group. Thus, for the first value of β the regression curves are parallel (as assumed in the null hypothesis), while for the other two values of β the regression curves are different, so the data is drawn under H_1. As for the design points, three scenarios were examined:

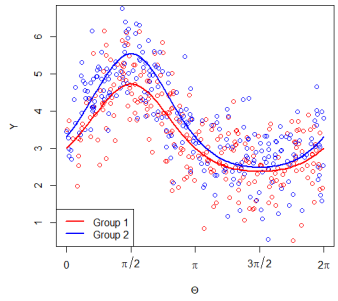
- Design 1: Both groups determined by a different sample from a circular uniform distribution.
- Design 2: Both groups determined by a different sample from a von Mises distribution, with mean $\mu = \pi$ and concentration $\kappa = 1.5$.
- Design 3: Both groups determined by a different sample from a von Mises distribution, with mean $\mu = 0$ and concentration $\kappa = 1.5$.

As an example of the models considered in the test of equality, Figure 4.4 shows simulated data from the three models under H_1 ($\beta = 0.3$ in A, $\beta = 1.3$ in B and $\beta = 1.75$ in C) and the true regression functions. On the other hand, Figure 4.5 displays representations of simulated data from the three models for the test of parallelism under the null hypothesis, where the parallel curves can be observed.

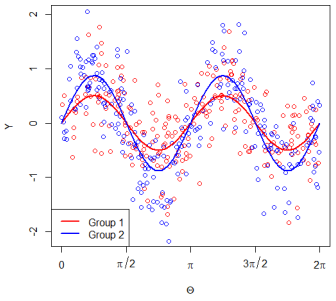
In addition to the nonparametric tests for circular predictors presented in Section 3.3 (which will be denoted by NPC), the nonparametric tests for linear data (NPL) presented in Section 1.2 and the parametric interaction and parallelism tests for circular predictors (PCP) introduced in Section 2.4.2, will be also studied in order to compare their performances. Results for the NPL test were obtained using the `sm` library, while the code for the PCP and NPC tests was self-programmed. Regarding the estimation of the curves, the cross-validation criterion (3.2) was used to select the smoothing parameter in the NPC test. For the NPL test, cross-validation was also the criterion employed for



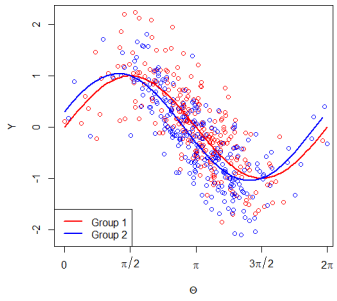
(a) Model A. Design 1



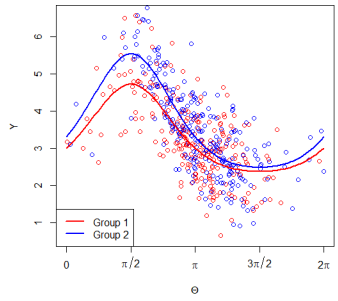
(b) Model B. Design 1



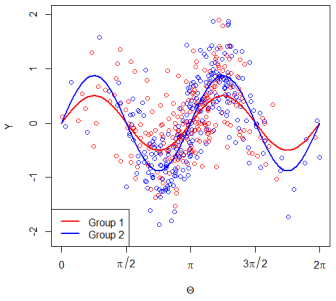
(c) Model C. Design 1



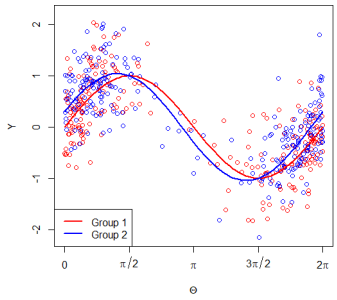
(d) Model A. Design 2



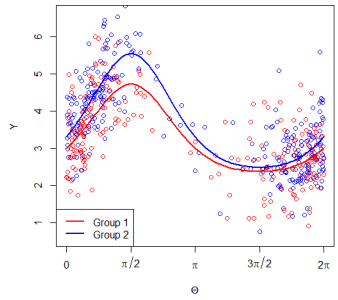
(e) Model B. Design 2



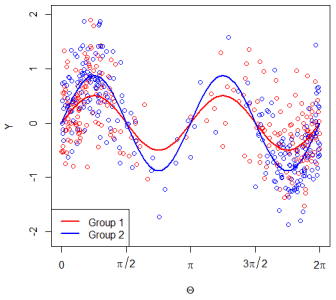
(f) Model C. Design 2



(g) Model A. Design 3

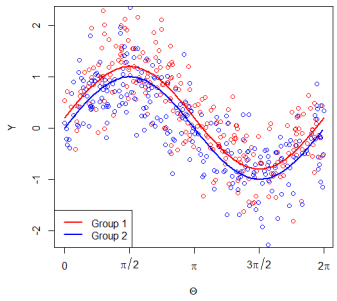


(h) Model B. Design 3

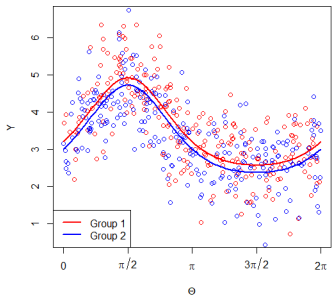


(i) Model C. Design 3

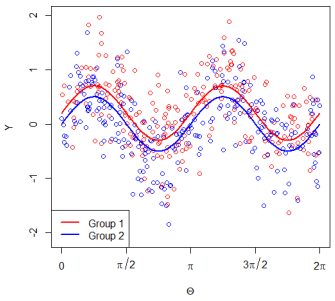
Figure 4.4: Representations of simulated data from models A (first column), B (second column) and C (third column) under the alternative hypothesis ($\beta = 0.3$ in A, $\beta = 1.3$ in B and $\beta = 1.75$ in C) under the three different designs, along with the true regression curves for each group. Number of observations is 250 for both groups in all cases.



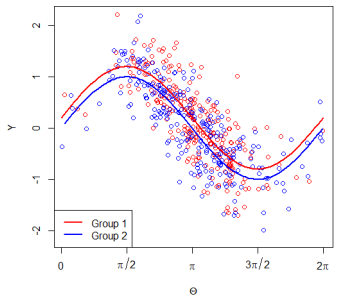
(a) Model A. Design 1



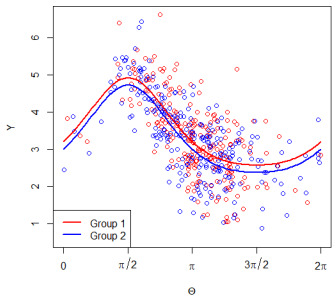
(b) Model B. Design 1



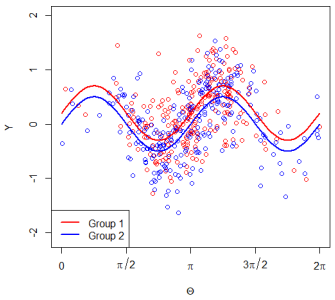
(c) Model C. Design 1



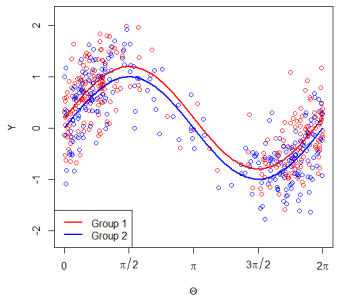
(d) Model A. Design 2



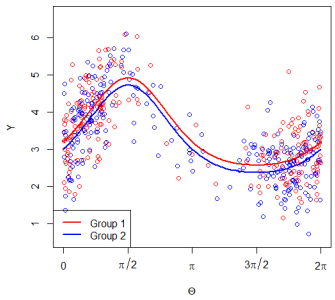
(e) Model B. Design 2



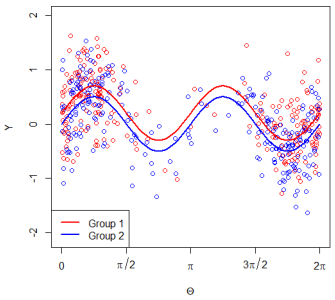
(f) Model C. Design 2



(g) Model A. Design 3



(h) Model B. Design 3



(i) Model C. Design 3

Figure 4.5: Representations of simulated data from models A (first column), B (second column) and C (third column) under the null hypothesis under the three different designs, along with the true regression curves for each group. Number of observations is 250 for both groups in all cases.

the selection of the smoothing parameter. As for the tests of parallelism, recall that it is necessary to choose a preliminary smoothing parameter. For the NPL test, the rule implemented in the `sm` library and proposed by Young and Bowman (1995) is to use $2R/n$ as a first smoothing parameter, where R is the range of the design points. For the NPC test, the preliminary concentration parameter is selected with the rule proposed in Section 3.3.1. A discussion on the results is presented next.

Model A Results for the test of equality for Model A are displayed on Table 4.25 (Design 1), Table 4.26 (Design 2) and Table 4.27 (Design 3). The parametric model is correct in this case and, as expected, the PCP test outperforms the rest in this scenario. Under the null hypothesis, the PCP and the NPC tests obtain percentages of rejection close to the nominal level .05 under the three designs. On the other hand, results for the NPL test under H_0 are also close to α when the uniform design is used (Design 1), but for Designs 2 and 3 the percentages of rejection are slightly higher, surpassing the 7% of rejections many times and even reaching 8%.

Under the alternative hypotheses, the nonparametric tests obtain similar results, although the power of the parametric test is higher since the parametric shape of the regression function assumed by the PCP test is correct. As expected, for the three tests, percentages of rejection under H_1 are lower when the value of σ is increased.

As for the test of parallelism, Table 4.28 displays the results obtained for Model A under Design 1, while Tables 4.29 and 4.30 contain results for Designs 2 and 3, respectively. As before, Model A follows the parametric shape assumed by the parametric test. Consequently it is expected that this test will outperform the rest in this setting. Under H_0, percentages of rejection of the PCP test, under the three designs are close to the nominal level α. The results for the NPC test are also around α, although it does reject 7.4% of the times in one scenario, under Design 3 with the lowest sample size. As for the nonparametric test for linear data, with the uniform design percentages of rejection vary around the nominal level, but under Design 2 it rejects more than 7% of the times in many occasions, although these numbers decrease by increasing the sample size. When using Design 3 the results are slightly higher than α (around 6.5%).

Furthermore, under the alternative hypothesis the PCP test obtains the largest percentages of rejection, as expected, although with large sample sizes the results obtained by the two nonparametric tests are quite close to the parametric results. The percentages of rejection for the NPL and NPC tests are similar, but under Design 1 the results for the circular test are slightly larger than the ones corresponding to its linear counterpart, and this behavior is turned around with Designs 2 and 3.

Model B Tables 4.31, 4.32 and 4.33 present results for the test of equality applied to Model B under Designs 1, 2 and 3, respectively. In this case the model is not of the parametric form (2.5), but it is an exponential transformation of it and the parametric test is still able to detect the differences between the groups.

Under H_0 ($\beta = 1$), as it happened in Model A, results for the PCP and the NPC tests are close to the value of α. This also happens for the NPL test when Design 1 is being considered but not with the von Mises designs. In the cases where Design 2 is used, 7% of rejections are obtained several times, even surpassing 8% of rejections once. However, with the largest sample size (250 for each group) the results get closer to α. Under Design 3 results are not as high as in the previous case, but many percentages of rejection lie above .065, reaching 7.5% of rejections once, which is significantly higher than the nominal level .05. In this case it does not seem that the results are closer to α when increasing the sample size. This behavior shows that the NPL test is not invariant to changes in the mean of the distribution of Θ.

Under the alternative hypothesis the three tests present similar results, obtaining higher percentages of rejection as the values of n_1 and n_2 increase and as the value of σ decreases.

As for the test of parallelism, percentages of rejection obtained for Model B are collected in Tables 4.34, 4.35 and 4.36 for Designs 1, 2 and 3, respectively. In this case, under H_0, the parametric test obtains percentages of rejection close to $\alpha = .05$ when Design 1 is used, but for Designs 2 and 3

the test gets quite high results in some scenarios, even surpassing 8% of rejections a couple of times. This behavior is explained by the fact that Model B does not follow the parametric shape assumed by the PCP test. Concerning the nonparametric test for linear data, with the uniform design results are pretty close to α, but with Design 2 it obtains some high percentages (around 8%). Under Design 3 results for the NPL test are not excessively high but some percentages are still significantly higher then α in all cases.

Under H_1, although the parametric assumption of test PCP is not right, this test obtains slightly larger percentages of rejection than the nonparametric ones. However, with large sample sizes the differences between the results of the three tests are small. As expected, percentages of rejection when $\beta = 1.3$ are larger, because the differences between the two groups are more pronounced. At the same time, reducing the variance of the errors and increasing the sample size leads to larger percentages of rejection.

Model C Results for the test of equality in Model C are shown in Tables 4.37, 4.38 and 4.39, containing the percentages of rejection under Designs 1, 2 and 3, respectively.

Under the null hypothesis, the NPC test obtains results close to the nominal level, $\alpha = .05$, for the three considered designs. Results for the PCP test are also close to α under Design 1, but when the von Mises designs are used the test obtains results which double or triplicate α. Thus, since now the model is far from the parametric shape assumed in (2.5), the test rejects the null hypothesis many more times than it should. As for the NPL test, results are close to the nominal level when Design 1 is used and just slightly higher than α under Design 3, but under Design 2, 7% and 8% of rejections are reached several times.

Under the alternative hypothesis, if Design 1 is considered the PCP test is completely unable to detect the differences between the two groups, obtaining percentages of rejection below .1. When the other two designs are considered, the parametric test is able to reject the null hypothesis more often, but the power of the test is considerably low. The nonparametric tests obtain again similar results under H_1, with percentages close to 1 as the sample size increases. Again, a large value of σ leads to lower percentages of rejection, but they are still high if the sample is sufficiently large.

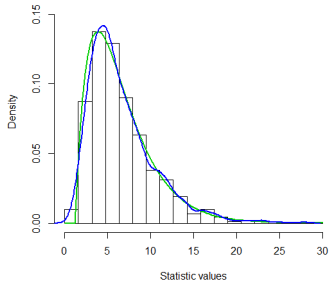
Results for the test of parallelism applied to Model C are found in Table 4.40 for Design 1, Table 4.41 for Design 2 and Table 4.42 for Design 3. When the null hypothesis holds, the parametric test performs well under Design 1, rejecting around 5% of the times in all cases. Nevertheless, the performance of the test is very unsatisfactory when Designs 2 and 3 are considered, obtaining between 10% and 20% of rejections. Results for the NPL test under H_0 are not too far from α when using the uniform design, but under Designs 2 some large percentages of rejection are recorded (around 7% or 8% of rejections), although results are closer to α when considering large sample sizes. Under Design 3 the results of the NPL test are not too high, but it was shown that the proportion of rejections was significantly larger that .05. Lastly, the nonparametric test for circular predictors performs well under the three designs, with percentages of rejection around $\alpha = .05$.

On another note, when the alternative hypothesis is true, the PCP test is unable to determine the differences between the groups under Design 1, with results lower than 6% of rejections in all cases. Under the other two designs the test is actually able to detect the two different curves, but the power is much lower than the obtained for the other two tests. Anyhow, given that the test is not well calibrated, it should not be used in practice. Results for the NPL and NPC tests are similar under H_0, with percentages of rejection close to 1 as the sample size and the differences between the curves increase.

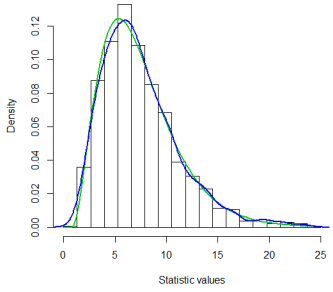
As a general summary of the results, it was found that the parametric test does not provide a correct calibration when the true model is not correctly imposed (being sometimes anticonservative). Thus, it should only be employed under this assumption. Regarding the nonparametric tests, focusing on the calibration, the linear test is sometimes anticonservative, while the same is not true for the new proposal in the circular setting. In the scenarios where the level under the null is close to α,

the percentage of rejections under the alternative is similar for both or slightly better for the circular proposal.

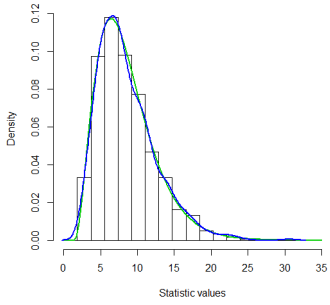
In Section 3.3 it was shown that the distribution of the test statistic C_3 (3.7) under H_0 was a shifted and scaled χ^2 distribution. In order to ascertain it, the values of the statistic obtained in the simulations were recorded. As an example, Figure 4.6 shows histograms of the statistic values obtained after applying the test of equality to simulated data from Model A with $\sigma = .25$ and different values for the sample size. The density function of the statistic values was estimated using the kernel density estimator, with the bandwidth selected by the *plug-in* method proposed by Sheather and Jones (1991). The $a\chi^2_b + c$ distribution was also represented in Figure 4.6 although the scale parameter a, the location parameter c and the number of degrees of freedom b were estimated by replacing the first three cumulants ν_1, ν_2 and ν_3 in (3.6) by the first three sample central moments of the statistic values.



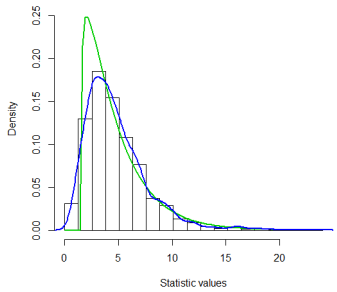
(a) Design 1. $n_1 = n_2 = 50$



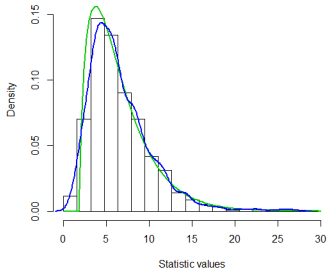
(b) Design 1. $n_1 = n_2 = 100$



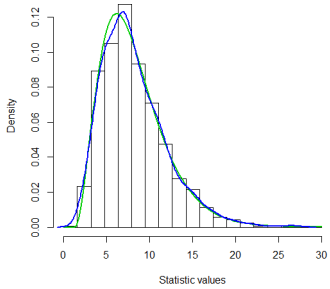
(c) Design 1. $n_1 = n_2 = 250$



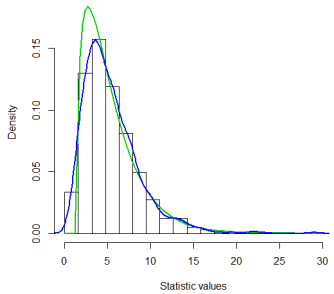
(d) Design 2. $n_1 = n_2 = 50$



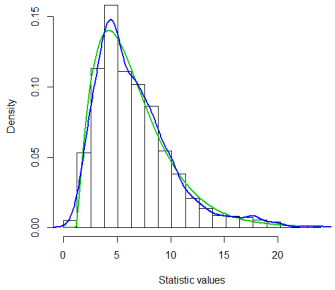
(e) Design 2. $n_1 = n_2 = 100$



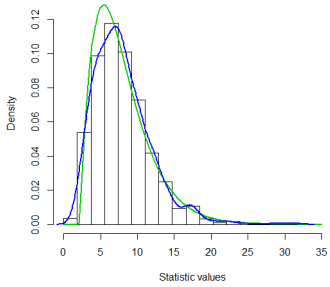
(f) Design 2. $n_1 = n_2 = 250$



(g) Design 3. $n_1 = n_2 = 50$



(h) Design 3. $n_1 = n_2 = 100$

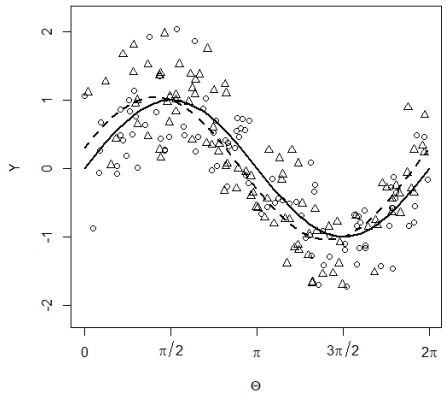


(i) Design 3. $n_1 = n_2 = 250$

Figure 4.6: Histograms of 1000 realizations of the statistic C_3 computed with data drawn from Model A with different designs and different sample sizes. The blue lines are the nonparametric density estimators of the statistic values and the green lines are $a\chi^2_b + c$ distributions with the parameters estimated from the data.

Model A. Design 1: circular uniform

Test of equality. Circular-linear regression

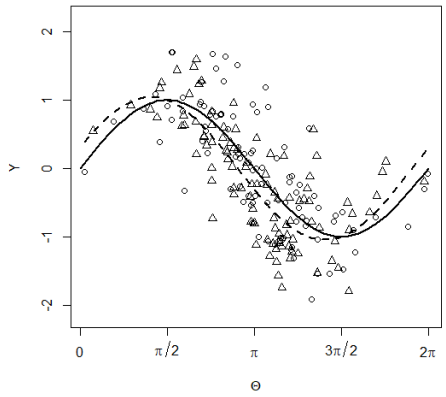


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.043	.067	.050	.621	.435	.428	.935	.827	.803
	50	100	.037	.042	.033	.752	.619	.613	.983	.954	.946
	100	100	.052	.059	.055	.928	.837	.838	.998	.994	.995
	100	250	.057	.048	.048	.984	.931	.933	1	1	1
	250	250	.041	.048	.048	1	.999	.999	1	1	1
.5	50	50	.047	.047	.043	.175	.143	.124	.355	.273	.270
	50	100	.045	.055	.055	.243	.176	.171	.505	.381	.381
	100	100	.047	.058	.047	.322	.250	.236	.694	.555	.578
	100	250	.051	.053	.043	.469	.367	.366	.867	.736	.752
	250	250	.057	.060	.057	.758	.615	.648	.988	.951	.962

Table 4.25: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model A and Design 1 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model A. Design 2: $vM(\pi, 1.5)$

Test of equality. Circular-linear regression

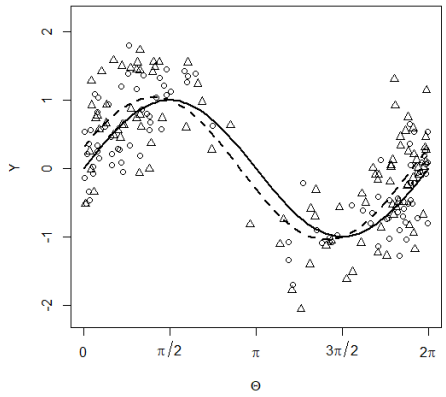


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.053	.078	.038	.723	.630	.529	.979	.934	.893
	50	100	.046	.077	.046	.850	.761	.708	.997	.985	.976
	100	100	.065	.080	.062	.967	.929	.909	1	.998	.998
	100	250	.062	.071	.053	.996	.983	.978	1	1	1
	250	250	.050	.062	.053	1	1	1	1	1	1
.5	50	50	.057	.071	.051	.228	.205	.149	.460	.398	.334
	50	100	.049	.071	.042	.268	.241	.199	.593	.498	.470
	100	100	.063	.076	.055	.421	.347	.330	.786	.683	.676
	100	250	.052	.060	.046	.615	.477	.468	.925	.858	.849
	250	250	.043	.058	.040	.849	.737	.740	.995	.983	.981

Table 4.26: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model A and Design 2 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model A. Design 3: $vM(0, 1.5)$

Test of equality. Circular-linear regression

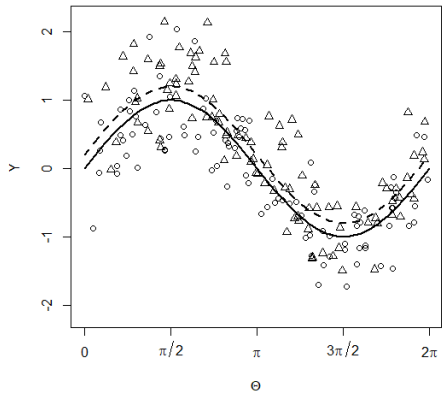


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.059	.069	.066	.717	.578	.570	.975	.921	.927
	50	100	.059	.074	.049	.868	.770	.755	.994	.979	.981
	100	100	.052	.064	.052	.958	.890	.896	1	1	.999
	100	250	.047	.054	.043	.998	.990	.987	1	1	1
	250	250	.042	.060	.048	1	1	1	1	1	1
.5	50	50	.046	.059	.048	.195	.191	.149	.449	.368	.334
	50	100	.054	.068	.048	.258	.222	.183	.594	.517	.475
	100	100	.059	.076	.064	.419	.374	.344	.771	.688	.672
	100	250	.049	.053	.041	.567	.474	.443	.921	.863	.850
	250	250	.044	.065	.050	.830	.752	.745	.998	.987	.988

Table 4.27: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model A and Design 3 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model A. Design 1: circular uniform

Test of parallelism. Circular-linear regression

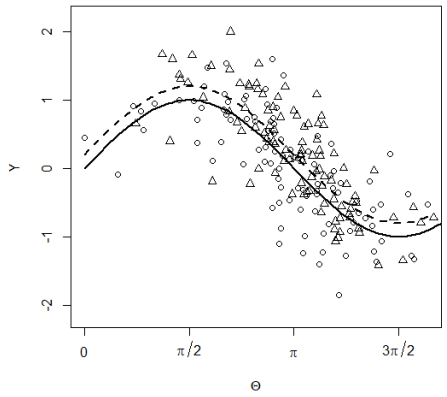


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.043	.056	.052	.675	.497	.489	.961	.851	.852
	50	100	.048	.052	.038	.815	.647	.650	.992	.956	.961
	100	100	.059	.070	.059	.943	.852	.860	.998	.993	.993
	100	250	.055	.040	.039	.987	.949	.947	1	1	1
	250	250	.050	.057	.048	1	1	1	1	1	1
.5	50	50	.043	.056	.044	.206	.158	.147	.404	.298	.320
	50	100	.054	.061	.053	.258	.187	.190	.577	.422	.446
	100	100	.047	.058	.049	.373	.272	.286	.758	.601	.626
	100	250	.045	.050	.042	.546	.396	.425	.904	.775	.798
	250	250	.049	.057	.057	.817	.653	.703	.991	.960	.969

Table 4.28: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model A and Design 1 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model A. Design 2: $vM(\pi, 1.5)$

Test of parallelism. Circular-linear regression

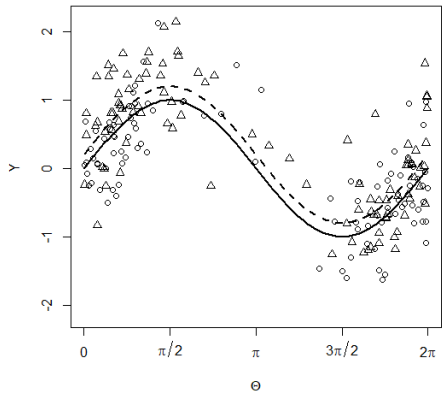


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.051	.074	.054	.378	.287	.227	.697	.536	.433
	50	100	.041	.074	.040	.489	.363	.298	.843	.712	.624
	100	100	.059	.075	.062	.674	.536	.459	.963	.875	.832
	100	250	.054	.067	.050	.858	.714	.645	.991	.973	.957
	250	250	.053	.058	.051	.982	.929	.908	1	1	1
.5	50	50	.053	.076	.050	.127	.130	.091	.249	.223	.163
	50	100	.047	.072	.045	.148	.142	.101	.292	.255	.202
	100	100	.054	.068	.048	.236	.202	.155	.442	.372	.307
	100	250	.048	.061	.049	.307	.245	.198	.577	.486	.416
	250	250	.046	.053	.046	.507	.378	.368	.857	.721	.723

Table 4.29: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model A and Design 2 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model A. Design 3: $vM(0, 1.5)$

Test of parallelism. Circular-linear regression

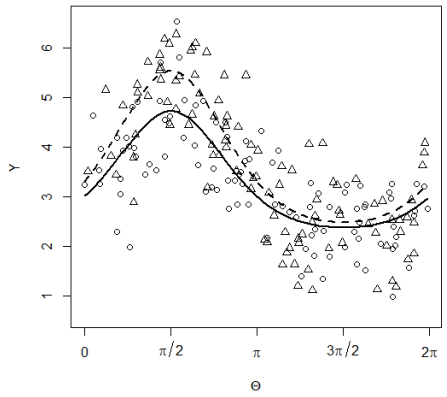


			$\beta = 0$			$\beta = .2$			$\beta = .3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.045	.077	.074	.345	.247	.250	.693	.522	.516
	50	100	.044	.063	.049	.507	.363	.339	.846	.689	.664
	100	100	.059	.071	.054	.710	.535	.511	.954	.872	.852
	100	250	.057	.052	.048	.862	.699	.664	.992	.956	.945
	250	250	.043	.056	.046	.980	.918	.913	1	1	.998
.5	50	50	.045	.059	.046	.105	.101	.087	.199	.165	.138
	50	100	.060	.067	.049	.148	.131	.106	.301	.222	.187
	100	100	.058	.064	.059	.224	.168	.166	.438	.327	.309
	100	250	.048	.065	.047	.312	.253	.220	.605	.463	.448
	250	250	.045	.067	.046	.500	.392	.380	.840	.744	.726

Table 4.30: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model A and Design 3 based on 1000 simulations. Results for $\beta = 0$ show empirical size, whereas $\beta = .2$ and $\beta = .3$ show empirical power.

Model B. Design 1: circular uniform

Test of equality. Circular-linear regression



			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.052	.068	.054	.345	.331	.314	.935	.918	.906
	50	100	.042	.044	.035	.454	.419	.416	.980	.975	.970
	100	100	.053	.059	.052	.676	.630	.630	.999	.998	.997
	100	250	.059	.048	.049	.845	.799	.810	1	1	1
	250	250	.042	.048	.045	.980	.970	.975	1	1	1
.75	50	50	.049	.049	.045	.187	.162	.161	.634	.558	.568
	50	100	.048	.055	.054	.245	.201	.208	.774	.707	.718
	100	100	.045	.053	.043	.348	.282	.298	.934	.891	.899
	100	250	.059	.053	.043	.488	.426	.439	.979	.970	.977
	250	250	.051	.060	.058	.799	.727	.744	1	.999	1

Table 4.31: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model B and Design 1 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

4.2.2 Linear-circular regression

In the present section the performance of the nonparametric tests for circular responses and linear predictors, proposed in Section 3.3.2 will be analyzed. For the test of equality and the test of parallelism, data will be drawn from three different models. The number of resamples in the bootstrap procedure will be 500. Percentages of rejection for $\alpha = .05$ will be computed after 500 realizations of the data. Consequently, in order to determine if a particular percentage of rejection is large, one can calculate the 95% confidence interval and check if it contains the nominal level .05. For 500 replications of the data, a percentage will be considered high when it is .074 or larger. The simulated models for the test of equality are presented next:

A. Group 1: $\Phi = [10\tan(\frac{2}{3}X - \frac{1}{4}) + \varepsilon](\text{mod } 2\pi)$.
   Group 2: $\Phi = [\beta\tan(\frac{3}{2}X - \frac{1}{4}) + \varepsilon](\text{mod } 2\pi), \quad \beta = 10, 8.5, 7.$

B. Group 1: $\Phi = [2\sin(4X - 1) + \varepsilon](\text{mod } 2\pi)$.
   Group 2: $\Phi = [\beta\sin(4X - 1) + \varepsilon](\text{mod } 2\pi), \quad \beta = 2, 1.75, 1.5.$

C. Group 1: $\Phi = [3\pi X^2 + \varepsilon](\text{mod } 2\pi)$.
   Group 2: $\Phi = [\beta\pi X^2 + \varepsilon](\text{mod } 2\pi), \quad \beta = 3, 3.2, 3.4.$

The sample size for each group takes values in the set $\{50, 100, 250\}$. The errors ε follow a von Mises distribution $vM(0, \kappa)$, where κ takes different values depending on the model. The null hypothesis holds when the first value of β is being used. The other two values match different situations where H_1 is true. The performance of the test of parallelism is studied using the same models as in the test of equality, but adding a shift of $\pi/8$ radians to the responses in the second group. The two tests are analyzed under two different scenarios for the design points:

- Design 1: Both group determined by a different sample from a $U(0, 1)$ distribution.
- Design 2: Both groups determined by a different sample from a $N(0.5, 0.1)$ distribution.

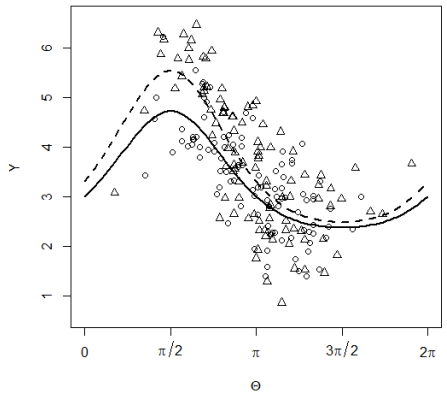
Figure 4.7 shows, as an example, a representation on the cylinder of data simulated from the three models used for the test of equality under H_1 with the last values of β, along with the true regression curves. Regarding the test of parallelism, the models under the null hypothesis are represented on the cylinder with realizations of the simulated data in Figure 4.8.

In order to estimate the regression curves, the tests were applied to the simulated data using the cross-validation criterion in order to select the smoothing parameter. For the test of parallelism it was necessary to select a preliminary smoothing parameter to estimate the shift, and it was done with the rule presented in Section 3.3.2. In this section the results of the test will not be compared with other methods as in the previous section. The reasons are, in the first place, that it does not exist a parametric ANCOVA test for circular responses. Secondly, the nonparametric equality and parallelism tests for linear data were applied to the simulated data and the percentages of rejection obtained under H_0 were undoubtedly high (e.g. 20% of rejections with $\alpha = .05$). Because it was clear that when applying the tests to linear-circular data they were not well calibrated, results were not included in the manuscript. The obtained percentages of rejection for the newly proposed tests for linear-circular data are shown and analyzed next.

Model A Results for the test of equality applied to Model A are displayed on Table 4.43. Under the null hypothesis the test rejects close to 5% of the times, although in some scenarios the percentage is even lower (2.4% or 3.2%, which are significantly lower than $\alpha = .05$). When the alternative hypothesis holds, under Design 1 (uniform distributed predictors) the percentages of rejection grow closer to 1 as the sample size increases, even when $\beta = 8.5$ is used. At the same time, results under the normal design are quite smaller than under the uniform design, which is understandable since under Design 2 the data

Model B. Design 2: $vM(\pi, 1.5)$

Test of equality. Circular-linear regression

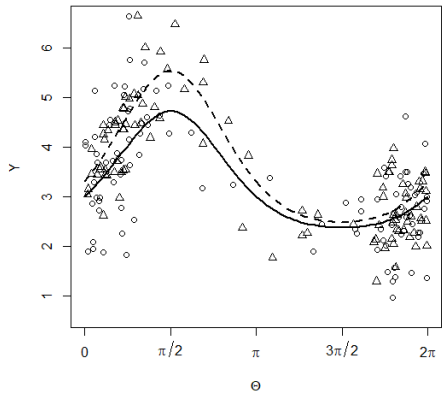


			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.065	.076	.044	.387	.376	.305	.917	.907	.856
	50	100	.068	.076	.046	.451	.446	.387	.969	.974	.963
	100	100	.069	.081	.063	.634	.595	.554	.998	.999	.998
	100	250	.076	.067	.053	.822	.773	.735	1	1	1
	250	250	.070	.055	.053	.962	.958	.956	1	1	1
.75	50	50	.065	.066	.048	.174	.182	.131	.623	.622	.554
	50	100	.061	.068	.041	.228	.222	.183	.718	.693	.646
	100	100	.076	.076	.051	.350	.336	.295	.916	.896	.873
	100	250	.070	.060	.047	.486	.444	.416	.981	.977	.965
	250	250	.069	.055	.042	.733	.667	.654	1	1	1

Table 4.32: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model B and Design 2 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

Model B. Design 3: $vM(0, 1.5)$

Test of equality. Circular-linear regression

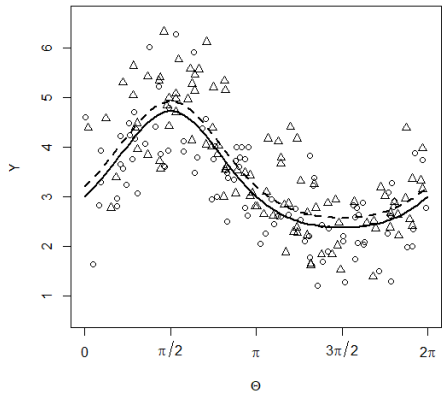


			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.069	.069	.060	.341	.290	.270	.905	.848	.861
	50	100	.061	.069	.047	.492	.396	.395	.969	.939	.949
	100	100	.076	.063	.053	.663	.586	.594	.998	.997	.997
	100	250	.070	.056	.043	.825	.752	.763	1	1	1
	250	250	.069	.059	.050	.967	.955	.959	1	1	1
.75	50	50	.059	.055	.042	.174	.175	.168	.611	.529	.540
	50	100	.067	.068	.047	.234	.192	.179	.761	.675	.674
	100	100	.060	.075	.060	.362	.323	.307	.904	.859	.872
	100	250	.049	.049	.043	.464	.389	.375	.976	.952	.962
	250	250	.049	.069	.050	.730	.650	.649	.999	.999	.999

Table 4.33: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model B and Design 3 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

Model B. Design 1: circular uniform

Test of parallelism. Circular-linear regression

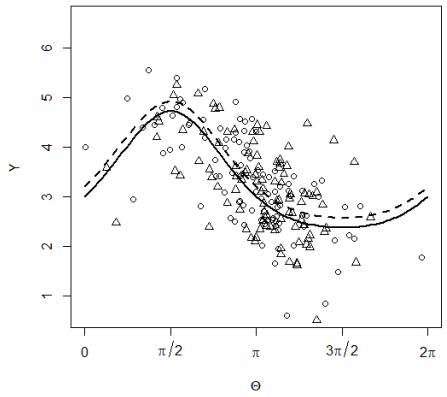


			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.050	.055	.039	.141	.118	.130	.487	.374	.381
	50	100	.043	.052	.040	.174	.146	.145	.589	.500	.496
	100	100	.054	.066	.045	.274	.218	.235	.771	.723	.737
	100	250	.051	.046	.052	.363	.278	.288	.907	.872	.892
	250	250	.043	.055	.059	.603	.495	.508	.996	.984	.993
.75	50	50	.046	.054	.042	.082	.082	.081	.251	.203	.170
	50	100	.048	.063	.047	.117	.084	.094	.336	.259	.273
	100	100	.052	.060	.041	.135	.127	.113	.470	.359	.384
	100	250	.053	.047	.056	.212	.168	.149	.609	.504	.534
	250	250	.045	.060	.051	.320	.225	.248	.855	.790	.826

Table 4.34: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model B and Design 1 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

Model B. Design 2: $vM(\pi, 1.5)$

Test of parallelism. Circular-linear regression

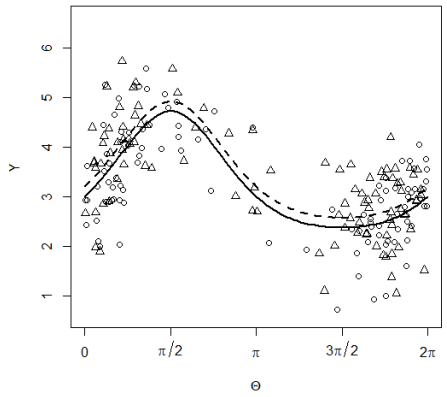


			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.068	.078	.051	.178	.152	.107	.422	.382	.306
	50	100	.073	.068	.048	.191	.147	.136	.549	.482	.409
	100	100	.068	.080	.044	.246	.215	.180	.724	.642	.661
	100	250	.084	.064	.052	.320	.245	.213	.879	.798	.796
	250	250	.065	.057	.049	.542	.426	.414	.979	.970	.979
.75	50	50	.067	.068	.053	.101	.098	.085	.249	.214	.174
	50	100	.070	.066	.043	.070	.112	.081	.269	.233	.199
	100	100	.053	.068	.040	.160	.133	.118	.418	.357	.355
	100	250	.064	.060	.051	.187	.152	.139	.546	.433	.403
	250	250	.059	.051	.049	.286	.205	.208	.774	.670	.713

Table 4.35: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model B and Design 2 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

Model B. Design 3: $vM(0, 1.5)$

Test of parallelism. Circular-linear regression

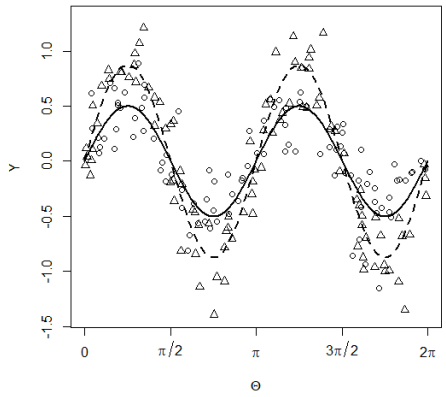


			$\beta = 1$			$\beta = 1.15$			$\beta = 1.3$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.5	50	50	.065	.067	.062	.126	.111	.123	.439	.333	.360
	50	100	.064	.056	.048	.202	.157	.126	.547	.420	.423
	100	100	.079	.068	.051	.273	.217	.162	.723	.622	.633
	100	250	.067	.055	.048	.309	.219	.232.	855	.772	.790
	250	250	.075	.060	.056	.535	.417	.429	.981	.969	.965
.75	50	50	.062	.062	.045	.095	.094	.099	.234	.160	.170
	50	100	.078	.066	.048	.107	.095	.086	.284	.209	.189
	100	100	.065	.065	.060	.142	.118	.121	.396	.299	.318
	100	250	.058	.065	.036	.179	.131	.116	.544	.406	.447
	250	250	.054	.068	.056	.292	.231	.222	.799	.669	.724

Table 4.36: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model B and Design 3 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.15$ and $\beta = 1.3$ show empirical power.

Model C. Design 1: circular uniform

Test of equality. Circular-linear regression

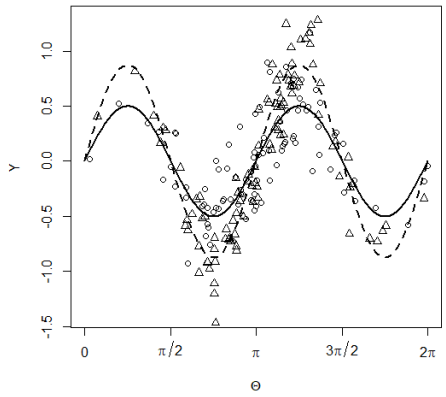


			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.059	.071	.055	.048	.541	.519	.055	.925	.917
	50	100	.055	.053	.042	.028	.680	.679	.019	.983	.987
	100	100	.045	.065	.058	.048	.918	.915	.042	.999	1
	100	250	.056	.046	.043	.018	.985	.987	.007	1	1
	250	250	.050	.048	.046	.055	1	1	.054	1	1
.5	50	50	.047	.059	.042	.047	.153	.152	.055	.302	.297
	50	100	.057	.058	.062	.029	.166	.174	.032	.383	.418
	100	100	.041	.053	.041	.044	.296	.333	.057	.649	.683
	100	250	.052	.054	.047	.030	.407	.425	.024	.834	.866
	250	250	.042	.063	.056	.054	.715	.755	.043	.987	.990

Table 4.37: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model C and Design 1 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.

Model C. Design 2: $vM(\pi, 1.5)$

Test of equality. Circular-linear regression

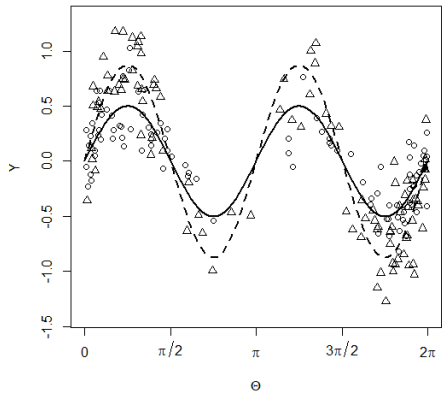


			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.192	.083	.046	.312	.617	.547	.403	.941	.917
	50	100	.152	.070	.053	.273	.741	.731	.380	.987	.984
	100	100	.177	.076	.068	.414	.939	.923	.577	.999	.999
	100	250	.170	.067	.049	.395	.988	.986	.636	1	1
	250	250	.153	.058	.055	.696	1	1	.909	1	1
.5	50	50	.106	.080	.052	.179	.171	.150	.255	.364	.336
	50	100	.094	.071	.050	.174	.192	.181	.257	.428	.432
	100	100	.103	.067	.051	.231	.294	.286	.354	.671	.681
	100	250	.093	.061	.046	.208	.402	.406	.397	.844	.845
	250	250	.090	.063	.051	.415	.736	.751	.709	.988	.999

Table 4.38: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model C and Design 2 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.

Model C. Design 3: $vM(0, 1.5)$

Test of equality. Circular-linear regression

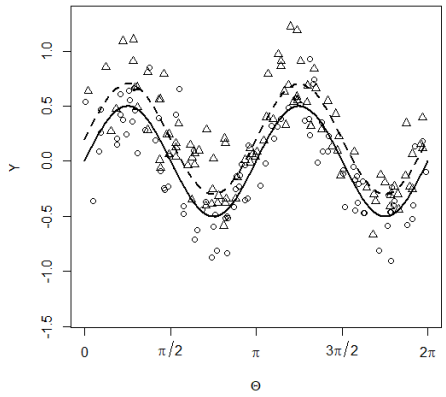


			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.161	.063	.048	.302	.522	.551	.423	.904	,917
	50	100	.178	.062	.044	.288	.672	.713	.393	.979	.986
	100	100	.160	.057	.056	.418	.909	.926	.576	1	1
	100	250	.161	.052	.045	.421	.978	.984	.645	1	1
	250	250	.164	.062	.052	.668	1	1	.901	1	1
.5	50	50	.098	.066	.040	.181	.191	.149	.254	.331	.333
	50	100	.097	.064	.050	.164	.214	.203	.242	.444	.454
	100	100	.103	.074	.068	.237	.313	.317	.397	.651	.701
	100	250	.079	.067	.046	.267	.406	.443	.392	.819	.859
	250	250	.098	.062	.049	.412	.723	.741	.671	.983	.988

Table 4.39: Percentages of rejection (for $\alpha = .05$) for the parametric interaction test for circular predictors (PCP), the nonparametric equality test for linear data (NPL) and the nonparametric equality test for circular predictors (NPC) for Model C and Design 3 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.

Model C. Design 1: circular uniform

Test of parallelism. Circular-linear regression

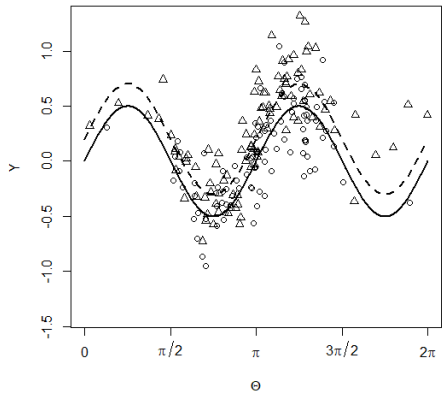


			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.060	.061	.048	.044	.584	.579	.056	.937	.915
	50	100	.049	.060	.052	.027	.707	.730	.026	.987	.974
	100	100	.046	.066	.053	.048	.930	.932	.044	1	1
	100	250	.047	.042	.054	.025	.987	.985	.017	1	1
	250	250	.058	.057	.064	.045	1	1	.049	1	1
.5	50	50	.053	.055	.043	.046	.154	.156	.051	.332	.327
	50	100	.059	.066	.044	.030	.187	.226	.036	.419	.476
	100	100	.046	.059	.042	.057	.334	.341	.056	.680	.697
	100	250	.049	.045	.059	.033	.442	.505	.023	.872	.880
	250	250	.041	.052	.047	.051	.754	.791	.047	.991	.995

Table 4.40: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model C and Design 1 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.

Model C. Design 2: $vM(\pi, 1.5)$

Test of parallelism. Circular-linear regression

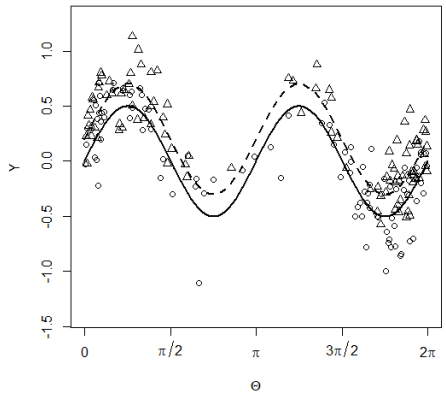


			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.218	.079	.058	.355	.661	.634	.446	.957	.957
	50	100	.178	.072	.050	.328	.779	.750	.450	.989	.990
	100	100	.182	.081	.045	.452	.950	.938	.659	.999	.998
	100	250	.172	.067	.058	.476	.988	.992	.712	1	1
	250	250	.169	.056	.048	.754	1	1	.932	1	1
.5	50	50	.120	.073	.051	.190	.186	.184	.284	.425	.377
	50	100	.111	.074	.040	.203	.226	.224	.298	.491	.517
	100	100	.105	.068	.047	.275	.334	.356	.417	.716	.744
	100	250	.117	.058	.049	.252	.438	.505	.469	.871	.887
	250	250	.096	.056	.039	.463	.779	.793	.762	.992	.990

Table 4.41: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model C and Design 2 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.

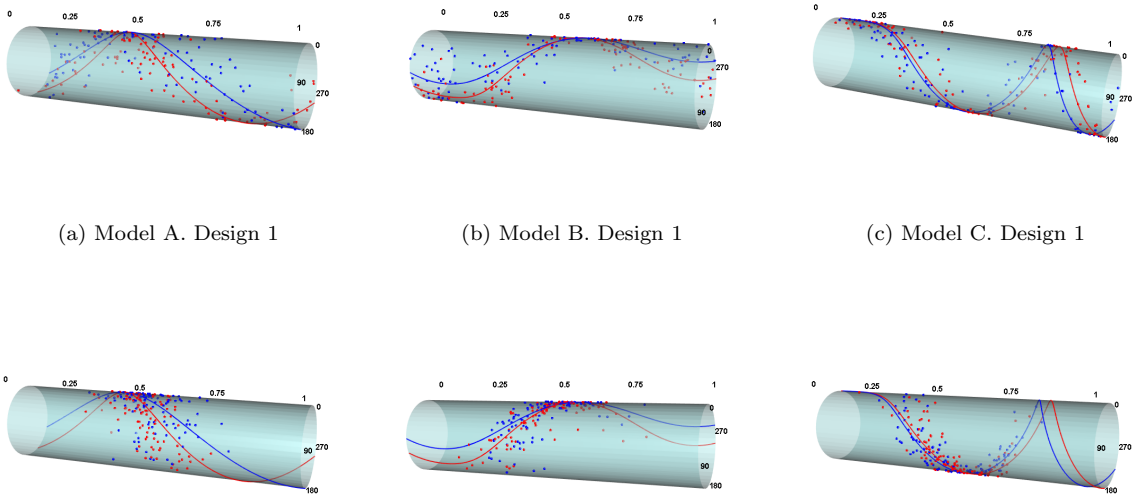
Model C. Design 3: $vM(0, 1.5)$

Test of parallelism. Circular-linear regression



			$\beta = 1$			$\beta = 1.5$			$\beta = 1.75$		
σ	n_1	n_2	PCP	NPL	NPC	PCP	NPL	NPC	PCP	NPL	NPC
.25	50	50	.171	.064	.060	.348	.568	.605	.473	.930	.955
	50	100	.189	.060	.045	.326	.703	.787	.455	.983	.988
	100	100	.180	.065	.045	.446	.916	.935	.644	1	1
	100	250	.182	.053	.052	.484	.986	.986	.709	1	1
	250	250	.177	.061	.055	.732	1	1	.925	1	1
.5	50	50	.106	.063	.047	.208	.206	.168	.291	.368	.368
	50	100	.113	.069	.039	.192	.240	.218	.289	.489	.498
	100	100	.104	.068	.059	.277	.339	.364	.428	.683	.721
	100	250	.090	.066	.037	.314	.445	.519	.458	.848	.885
	250	250	.113	.056	.058	.479	.759	.813	.733	.990	.995

Table 4.42: Percentages of rejection (for $\alpha = .05$) for the parametric parallelism test for circular predictors (PCP), the nonparametric parallelism test for linear data (NPL) and the nonparametric parallelism test for circular predictors (NPC) for Model C and Design 3 based on 1000 simulations. Results for $\beta = 1$ show empirical size, whereas $\beta = 1.5$ and $\beta = 1.75$ show empirical power.



(a) Model A. Design 1 (b) Model B. Design 1 (c) Model C. Design 1

(d) Model A. Design 2 (e) Model B. Design 2 (f) Model C. Design 2

Figure 4.7: Representations in the cylinder of simulated data from models A, B and C under the alternative hypothesis ($\beta = 7$ in A, $\beta = 1.5$ in B and $\beta = 3.4$ in C) under Design 1 (top row) and Design 2 (bottom row), along with the true regression curves for each group. Number of observations is 100 for both groups and the value of κ is 3 in Model A, 5 in Model B and 4 in Model C. Circular units are in degrees.

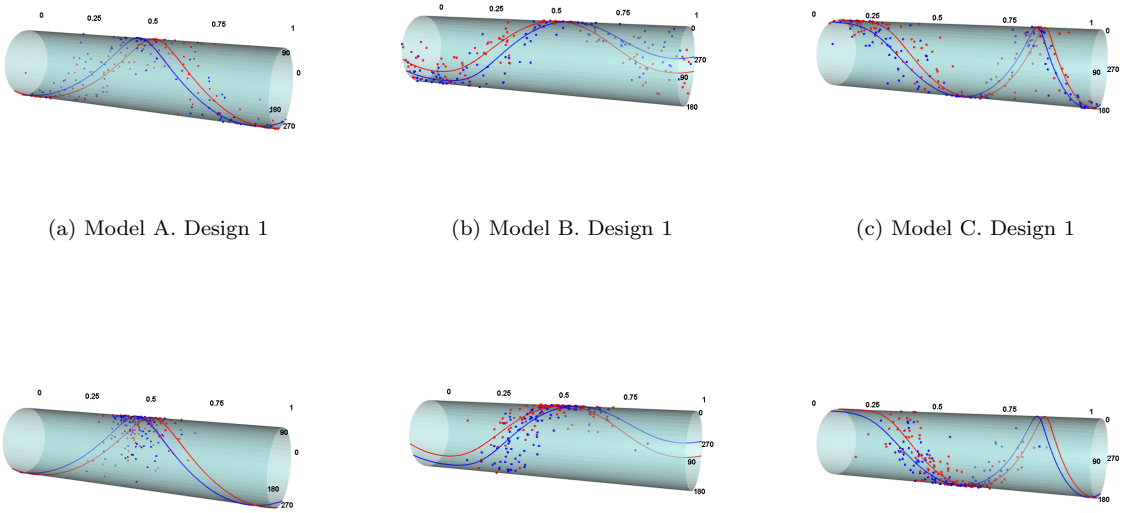
are concentrated near the point where the two curves intersect (see Figure 4.7d). Even so, when the differences between the curves are more noticeable ($\beta = 7$) and the sample size is large, the percentages of rejection are also close to 1. In addition, as expected, when diminishing the concentration of the errors the percentages of rejection under H_1 decrease.

For the test of parallelism, results are collected in Table 4.44. Under the null hypothesis the percentages of rejection are close to α, showing that the test is well calibrated. If the data are drawn under the alternative hypothesis, when considering Design 1 the power of the test is high, specially for the second value of α. If the data are concentrated as in Design 2 the percentages of rejection are lower, since it is more difficult to discern the different curves. However, when the sample size is large percentages of rejection are high.

Model B Table 4.45 shows the percentages of rejection obtained after applying the equality test to Model B. Under H_0, and when Design 1 is being considered, results are close to the nominal level $\alpha = .05$, while under the normal design the percentages of rejection are moderately smaller than α (around 3% our 4% of rejections, and sometimes being significantly smaller than α).

Under the alternative hypothesis, with Design 1 results converge to 1 as the sample size increases. As it happened with Model A, when using Design 2 the data are concentrated in the point where the regression curves are closer (see Figure 4.7e), so it is more difficult to detect the differences between both curves. Consequently, the percentages of rejection are quite lower than in the Design 1 scenario, but they are still close to 1 when the second value of β is used and the sample size is large. Again, a reduction in the concentration of the errors translates in lower percentages of rejection.

Table 4.46 contains percentages of rejection for the test of parallelism applied to Model B. Under



(a) Model A. Design 1 (b) Model B. Design 1 (c) Model C. Design 1

(d) Model A. Design 2 (e) Model B. Design 2 (f) Model C. Design 2

Figure 4.8: Representations in the cylinder of simulated data from models A, B and C under the alternative hypothesis ($\beta = 7$ in A, $\beta = 1.5$ in B and $\beta = 3.4$ in C) under Design 1 (top row) and Design 2 (bottom row), along with the true regression curves for each group. Number of observations is 100 for both groups and the value of κ is 3 in Model A, 5 in Model B and 4 in Model C. Circular units are in degrees.

H_0 all the percentages of rejection are close to the level .05. Regarding the performance of the test under H_1, results are close to the ones obtained for the test of equality. When Design 2 is considered, percentages of rejection are lower, but it is comprehensible given the shape of the regression functions where the data are concentrated.

Model C Table 4.47 contains the percentages of rejection corresponding to the test of equality applied to Model C. Under the null hypothesis the test rejects close to 5% of the times, although when the normal design is used the percentages are slightly lower than α. On the other hand, when the alternative hypothesis holds, results are similar under both designs, unlike in the other two models. In fact, percentages of rejection are sometimes larger under Design 2. The cause of this behavior is that under the normal design the points are concentrated in a zone where the curves are not close, as it can be seen in Figure 4.7f.

On the other hand, results for the test of parallelism are displayed on Table 4.48. In this case, all the percentages are close to .05 when the null hypothesis is true. However, when H_1 holds, percentages of rejection are not very high, specially under Design 2, because of the shape of the regression functions. When $\beta \neq 3$ the two curves are not parallel, but they are close to being parallel. Furthermore, when the data are concentrated as in Design 2 it is more difficult to discern the different curves, which explains the low power of the test.

As a general conclusion, it was found that both the equality and parallelism tests are well calibrated and are the only available method to test equality and parallelism when working with linear-circular

data, since, as mentioned before, the nonparametric test for linear data are far from providing a correct calibration. It shall be noticed that, given that a bootstrap procedure is used to calculate the distribution of the statistics under H_0, the tests are somewhat computationally costly, specially the test of parallelism, which also uses numerical optimization to estimate the shift parameter.

Model A

Test of equality. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5, 0.1)$		
κ	n_1	n_2	$\beta = 10$	$\beta = 8.5$	$\beta = 7$	$\beta = 10$	$\beta = 8.5$	$\beta = 7$
4	50	50	.038	.604	.994	.058	.230	.682
	50	100	.048	.776	1	.038	.250	.798
	100	100	.045	.932	1	.050	.384	.950
	100	250	.050	.992	1	.032	.538	.992
	250	250	.040	1	1	.046	.810	1
3	50	50	.036	.428	.956	.052	.148	.492
	50	100	.024	.592	.996	.046	.218	.670
	100	100	.062	.794	1	.052	.270	.842
	100	250	.034	.930	1	.048	.416	.954
	250	250	.052	.996	1	.038	.630	1

Table 4.43: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for linear predictors and circular responses for Model A based on 500 simulations. Results for $\beta = 10$ show empirical size, whereas $\beta = 8.5$ and $\beta = 7$ show empirical power.

Model A

Test of parallelism. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5,0.1)$		
κ	n_1	n_2	$\beta=10$	$\beta=8.5$	$\beta=7$	$\beta=10$	$\beta=8.5$	$\beta=7$
4	50	50	.056	.586	.990	.044	.098	.338
	50	100	.070	.738	1	.038	.132	.466
	100	100	.060	.896	1	.048	.196	.628
	100	250	.054	.982	1	.050	.254	.822
	250	250	.044	1	1	.034	.460	.974
3	50	50	.042	.420	.952	.054	.100	.282
	50	100	.038	.540	.990	.042	.100	.332
	100	100	.050	.750	1	.048	.154	.494
	100	250	.056	.908	1	.052	.224	.692
	250	250	.064	.998	1	.060	.330	.898

Table 4.44: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for linear predictors and circular responses for Model A based on 500 simulations. Results for $\beta = 10$ show empirical size, whereas $\beta = 8.5$ and $\beta = 7$ show empirical power.

Model B

Test of equality. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5, 0.1)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$
6	50	50	.066	.450	.976	.038	.154	.514
	50	100	.050	.650	.998	.028	.212	.718
	100	100	.046	.784	1	.036	.320	.882
	100	250	.046	.748	1	.036	.320	.882
	250	250	.046	.976	1	.040	.506	.994
5	50	50	.054	.286	.856	.030	.086	.268
	50	100	.056	.408	.956	.048	.132	.436
	100	100	.042	.584	.996	.052	.202	.620
	100	250	.048	.748	1	.042	.260	.846
	250	250	.054	.918	1	.038	.462	.968

Table 4.45: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for linear predictors and circular responses for Model B based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 1.75$ and $\beta = 1.5$ show empirical power.

Model B

Test of parallelism. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5, 0.1)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$
6	50	50	.072	.364	.930	.044	.142	.474
	50	100	.044	.488	.974	.046	.182	.590
	100	100	.052	.686	1	.042	.258	.740
	100	250	.046	.836	1	.056	.320	.894
	250	250	.058	.986	1	.056	.556	.986
5	50	50	.046	.340	.892	.040	.110	.382
	50	100	.056	.460	.960	.038	.134	.498
	100	100	.056	.582	.996	.056	.204	.682
	100	250	.052	.740	1	.056	.308	.844
	250	250	.066	.962	1	.048	.476	.966

Table 4.46: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for linear predictors and circular responses for Model B based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 1.75$ and $\beta = 1.5$ show empirical power.

Model C

Test of equality. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5, 0.1)$		
κ	n_1	n_2	$\beta = 3$	$\beta = 3.2$	$\beta = 3.4$	$\beta = 3$	$\beta = 3.2$	$\beta = 3.4$
5	50	50	.044	.326	.780	.038	.278	.838
	50	100	.058	.520	.868	.038	.346	.952
	100	100	.048	.730	.854	.034	.528	.988
	100	250	.048	.888	.926	.054	.710	1
	250	250	.060	.968	.966	.036	.924	1
4	50	50	.062	.274	.744	.040	.222	.742
	50	100	.054	.382	.852	.040	.320	.858
	100	100	.040	.598	.862	.032	.438	.970
	100	250	.052	.836	.922	.046	.570	.992
	250	250	.056	.950	.946	.062	.810	1

Table 4.47: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for linear predictors and circular responses for Model C based on 500 simulations. Results for $\beta = 3$ show empirical size, whereas $\beta = 3.2$ and $\beta = 3.4$ show empirical power.

Model C

Test of parallelism. Linear-circular regression

			Design 1: $U(0,1)$			Design 2: $N(0.5, 0.1)$		
κ	n_1	n_2	$\beta = 3$	$\beta = 3.2$	$\beta = 3.4$	$\beta = 3$	$\beta = 3.2$	$\beta = 3.4$
5	50	50	.044	.134	.514	.048	.060	.162
	50	100	.052	.232	.718	.054	.098	.194
	100	100	.066	.440	.844	.040	.100	.308
	100	250	.052	.602	.924	.068	.156	.410
	250	250	.056	.888	.934	.052	.206	.728
4	50	50	.048	.138	.412	.038	.058	.112
	50	100	.046	.190	.620	.048	.070	.170
	100	100	.042	.314	.806	.046	.106	.242
	100	250	.044	.478	.932	.048	.106	.322
	250	250	.068	.772	.906	.062	.170	.572

Table 4.48: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for linear predictors and circular responses for Model C based on 500 simulations. Results for $\beta = 3$ show empirical size, whereas $\beta = 3.2$ and $\beta = 3.4$ show empirical power.

4.2.3 Circular-circular regression

To finalize this simulation study, the equality and parallelism tests for circular predictors and circular covariates will be analyzed. As in the previous section, 500 replications of data will be drawn from simulated models and the percentages of rejection for the nominal level $\alpha = .05$ will be computed. Again, for this number of replications it will be considered that a percentage is significantly larger than α if it is .074 or higher. The test of equality will be applied to the next models:

A. Group 1: $\Phi = [2\sin(2\Theta) + \varepsilon](\text{mod } 2\pi)$.
Group 2: $\Phi = [\beta\sin(2\Theta) + \varepsilon](\text{mod } 2\pi), \quad \beta = 2, 2.5, 3.$

B. Group 1: $\Phi = [\cos(2\Theta + 3\cos(2\Theta)) + \varepsilon](\text{mod } 2\pi)$.
Group 2: $\Phi = [\beta[\cos(2\Theta + 3\cos(2\Theta)) + \varepsilon](\text{mod } 2\pi), \quad \beta = 1, 1.5, 1.75.$

C. Group 1: $\Phi = [\sin(3\Theta)(\Theta/2)^{1/2} + \varepsilon](\text{mod } 2\pi)$.
Group 2: $\Phi = [\beta\sin(3\Theta)(\Theta/2)^{1/2} + \varepsilon](\text{mod } 2\pi), \quad \beta = 1, 1.35, 1.5.$

The sample size for each group is 50, 100 or 250. The errors ε follow a von Mises distribution with mean direction zero and concentration κ, which variates in each model. As in the other sections, with the first value of β the data is drawn under H_0, whereas for the other two values the data is drawn from different regression functions. For the test of parallelism the previous models are modified by adding a shift of $\pi/8$ radians to the responses in each group. Thus, if the first value of β is used, the regression curves are parallel. A random design is used for the predictor variable, although three designs are used:

- Design 1: Both groups determined by a different sample from a circular uniform distribution.
- Design 2: Both groups determined by a different sample from a von Mises distribution, with mean $\mu = \pi$ and concentration $\kappa = 2$.
- Design 3: Both groups determined by a different sample from a von Mises distribution, with mean $\mu = 0$ and concentration $\kappa = 2$.

Realizations of data drawn from models for the test of equality under the alternative hypothesis (first value of β) can be seen in Figure 4.9. For the test of parallelism, representations on the torus of data drawn under the null hypothesis are displayed in Figure 4.10, where the parallel regression curves are shown.

Regarding the estimation of the regression functions, the cross-validation criterion was used to select the smoothing parameter. In addition, for the test of parallelism the rule proposed in Section 3.3.2 was employed to estimate the shift parameter. For the bootstrap procedure 500 bootstrap replicates were obtained. Now, a summary of the results is given.

Model A Table 4.49 contains the results of the test of equality applied to Model A. Under the null hypothesis percentages of rejection are mostly close to $\alpha = .05$, although one of the results (.074) is significantly higher than α. Under H_1, as the data size increases the percentages of rejection grow closer to 1. In fact, for $\beta = 3$ results are close to 1 even for $n_1 = 50$ and $n_2 = 50$.

With respect to the test of parallelism, results are collected in Table 4.50. When the null hypothesis hold, the corresponding percentages of rejection are close to the nominal level, indicating a correct calibration of the test. If the data are drawn from the alternative hypothesis, again the power of the test is high specially for large sample sizes.

Model B Table 4.51 shows percentages of rejection for the test of equality applied to Model B. Under the null hypothesis there are not percentages of rejection significantly different than α. On the other hand, if the alternative hypothesis is true, as in the previous model percentages of rejection are high, converging to 1 as n_1 and n_2 increase. In general, when H_1 hold results obtained under Designs 2 and 3 are higher than the ones gotten under Design 1.

Table 4.52 presents percentages of rejection for the test of parallelism, when it is applied to Model B. Although in one occasion a percentage of rejection under H_0 turned out to be significantly larger than .05 (7.4% of rejections) most percentages are close to α. Under the alternative hypothesis results are very similar to the ones obtained for the test of equality: percentages grow closer to 1 as the sample size increases, and the power of the test is slightly higher under the von Mises designs.

Model C Results obtained for the last model are displayed in Table 4.53 (test of equality) and Table 4.54 (test of parallelism). Under H_0 percentages of rejection in both tests are close to α except for a couple of exceptions (7.4% of rejections in the test of equality, Design 1 and 7.8% of rejections in the test of parallelism, Design 2). On the other hand, when the alternative hypothesis holds, percentages of rejection are similar for the two tests, which are able to detect the different curves, specially for large sample sizes.

As a general summary, the tests of equality and parallelism for circular-circular regression seem to be well calibrated, although sometimes the percentages under the null are slightly higher than the considered level. These vaguely high percentages under H_0 may be corrected when using a higher number of bootstrap replicates. In the present simulation study the number of replicates was held to 500 for computational reasons, but it is recommended to use a larger number in practice. It should be noted that because of the bootstrap procedure used to calculate the distribution of the statistic, the test are computationally slow, specially the parallelism case, since it also involves numerical optimization methods.

Model A

Test of equality. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 2.5$	$\beta = 3$	$\beta = 2$	$\beta = 2.5$	$\beta = 3$	$\beta = 2$	$\beta = 2.5$	$\beta = 3$
4	50	50	.060	.408	.966	.048	.508	.992	.070	.516	.988
	50	100	.062	.498	.994	.040	.646	.998	.036	.640	.996
	100	100	.052	.834	1	.054	.878	1	.040	.872	1
	100	250	.064	.942	1	.048	.974	1	.052	.980	1
	250	250	.064	1	1	.068	1	1	.058	1	1
2	50	50	.072	.144	.514	.062	.222	.708	.062	.204	.704
	50	100	.048	.166	.656	.042	.208	.848	.048	.214	.836
	100	100	.074	.326	.928	.048	.430	.970	.064	.436	.968
	100	250	.044	.410	.984	.060	.532	.998	.060	.528	.998
	250	250	.044	.878	1	.044	.808	1	.038	.816	1

Table 4.49: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for circular-circular regression for Model A based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.

Model A

Test of parallelism. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 2.5$	$\beta = 3$	$\beta = 2$	$\beta = 2.5$	$\beta = 3$	$\beta = 2$	$\beta = 2.5$	$\beta = 3$
4	50	50	.058	.426	.956	.050	.570	.992	.048	.532	.986
	50	100	.054	.524	.992	.040	.694	.998	.060	.692	.998
	100	100	.052	.858	1	.062	.878	1	.038	.880	1
	100	250	.066	.966	1	.050	.974	1	.050	.970	1
	250	250	.054	1	1	.040	1	1	.066	1	1
2	50	50	.062	.142	.528	.056	.250	.738	.054	.254	.730
	50	100	.070	.164	.650	.058	.262	.832	.056	.262	.844
	100	100	.054	.320	.942	.046	.414	.976	.054	.410	.974
	100	250	.058	.444	.992	.048	.520	.998	.046	.600	.998
	250	250	.066	.884	1	.028	.864	1	.040	.868	1

Table 4.50: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for circular-circular regression for Model A based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.

Model B

Test of equality. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$
4	50	50	.066	.474	.852	.066	.578	.936	.072	.586	.942
	50	100	.068	.530	.924	.070	.654	.974	.046	.662	.970
	100	100	.072	.862	.996	.052	.916	1	.046	.912	1
	100	250	.070	.948	1	.054	.978	1	.056	.982	1
	250	250	.044	1	1	.058	1	1	.038	1	1
3	50	50	.060	.300	.622	.060	.400	.792	.066	.380	.800
	50	100	.062	.380	.780	.044	.496	.872	.060	.496	.874
	100	100	.058	.722	.990	.066	.788	.996	.068	.792	.996
	100	250	.070	.810	1	.052	.878	1	.058	.862	1
	250	250	.062	.994	1	.066	.998	1	.062	.998	1

Table 4.51: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for circular-circular regression for Model B based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.

Model B

Test of parallelism. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$	$\beta = 2$	$\beta = 1.75$	$\beta = 1.5$
4	50	50	.044	.436	.814	.058	.514	.918	.062	.554	.868
	50	100	.054	.508	.920	.060	.644	.960	.054	.642	.962
	100	100	.062	.892	.998	.072	.894	.998	.058	.886	.996
	100	250	.056	.974	1	.056	.974	1	.074	.980	1
	250	250	.048	1	1	.048	1	1	.058	1	1
3	50	50	.070	.338	.656	.052	.366	.758	.054	.396	.744
	50	100	.072	.376	.786	.054	.454	.862	.062	.446	.834
	100	100	.072	.726	.976	.062	.730	.986	.050	.734	.976
	100	250	.058	.834	.992	.052	.856	1	.072	.856	.996
	250	250	.074	.994	1	.072	.998	1	.064	.992	1

Table 4.52: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for circular-circular regression for Model B based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.

Model C

Test of equality. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$
5	50	50	.070	.346	.692	.066	.520	.866	.056	.508	.862
	50	100	.068	.492	.840	.052	.624	.942	.040	.612	.938
	100	100	.074	.782	.990	.052	.868	.996	.042	.884	1
	100	250	.048	.944	1	.046	.944	1	.050	.964	1
	250	250	.058	1	1	.052	.996	1	.050	1	1
3	50	50	.060	.214	.418	.070	.264	.524	.040	.268	.560
	50	100	.052	.220	.472	.046	.306	.632	.064	.326	.654
	100	100	.066	.494	.838	.052	.558	.916	.040	.524	.908
	100	250	.064	.650	.938	.054	.702	.964	.060	.712	.960
	250	250	.042	.960	1	.070	.964	1	.054	.942	1

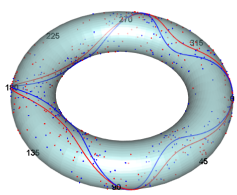
Table 4.53: Percentages of rejection (for $\alpha = .05$) for the nonparametric equality test for circular-circular regression for Model C based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.

Model C

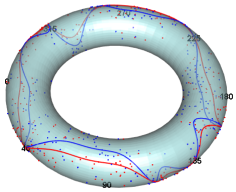
Test of parallelism. Circular-circular regression

			Design 1: CU			Design 2: $vM(\pi, 2)$			Design 3: $vM(0, 2)$		
κ	n_1	n_2	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$
5	50	50	.070	.412	.728	.054	.546	.856	.044	.490	.804
	50	100	.058	.470	.854	.054	.654	.948	.054	.608	.914
	100	100	.064	.814	.992	.038	.894	.998	.068	.846	.994
	100	250	.054	.926	.998	.070	.964	1	.050	.948	.998
	250	250	.056	1	1	.046	1	1	.030	1	1
3	50	50	.066	.222	.456	.068	.362	.616	.048	.228	.454
	50	100	.060	.220	.470	.078	.384	.694	.048	.270	.586
	100	100	.052	.430	.830	.060	.546	.910	.056	.522	.870
	100	250	.064	.636	.958	.052	.710	.970	.048	.626	.952
	250	250	.052	.974	1	.066	.952	1	.040	.944	1

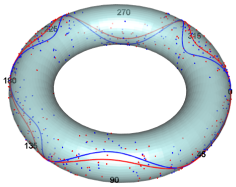
Table 4.54: Percentages of rejection (for $\alpha = .05$) for the nonparametric parallelism test for circular-circular regression for Model C based on 500 simulations. Results for $\beta = 2$ show empirical size, whereas $\beta = 2.5$ and $\beta = 3$ show empirical power.



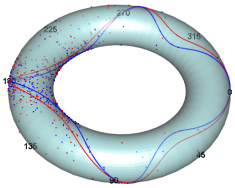
(a) Model A. Design 1



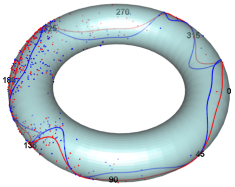
(b) Model B. Design 1



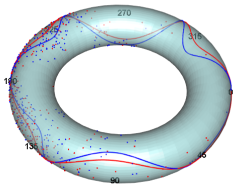
(c) Model C. Design 1



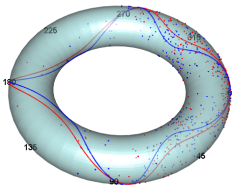
(d) Model A. Design 2



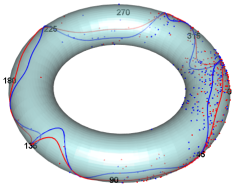
(e) Model B. Design 2



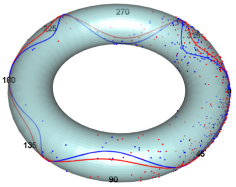
(f) Model C. Design 2



(g) Model A. Design 3

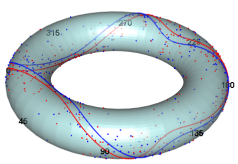


(h) Model B. Design 3

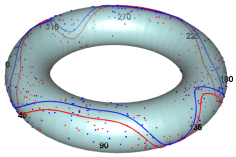


(i) Model C. Design 3

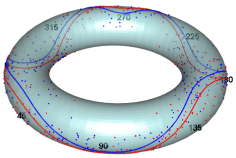
Figure 4.9: Representations in the torus of simulated data from models A, B and C of the equality test under the alternative hypothesis ($\beta = 2.5$ in A, $\beta = 1.75$ in B and $\beta = 1.5$ in C) under Design 1 (top row), Design 2 (middle row) and Design 3 (bottom row), along with the true regression curves for each group. Number of observations is 250 for both groups and the value of κ is 2 in Model A and 3 in Models B and C. Circular units are in degrees.



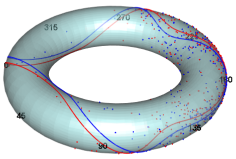
(a) Model A. Design 1



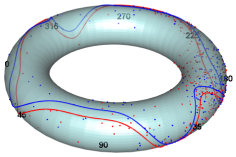
(b) Model B. Design 1



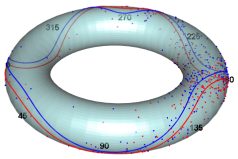
(c) Model C. Design 1



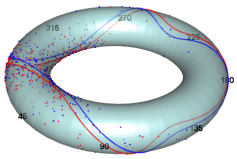
(d) Model A. Design 2



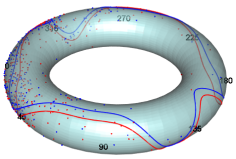
(e) Model B. Design 2



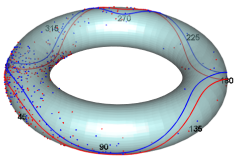
(f) Model C. Design 2



(g) Model A. Design 3



(h) Model B. Design 3



(i) Model C. Design 3

Figure 4.10: Representations in the torus of simulated data from models A, B and C of the parallelism test under the null hypothesis. Data drawn under Design 1 (top row), Design 2 (middle row) and Design 3 (bottom row), along with the true regression curves for each group. Number of observations is 250 for both groups and the value of κ is 2 in Model A and 3 in Models B and C. Circular units are in degrees.

Chapter 5

Application to real data

The aim of this chapter is to apply the tests proposed in Chapter 3 to real data, for which two different datasets will be used. The tests involving circular covariates and linear responses will be applied to a dataset which belongs to the automotive industry field, and can be found in Anderson-Cook (1999). On contrast, the dataset used to illustrate the tests for circular responses belongs to the animal orientation field, and it was provided by Professor Felicita Scapini from the Department of Biology of the University of Florence.

5.1 Flywheel data

In this section, the flywheel data given in Anderson-Cook (1999) will be analyzed with the nonparametric tests presented in Chapter 3. This dataset gives two measures on flywheels (a mechanical device used in the automotive industry), where one of them is circular, and it is used to predict the second measure, which is linear. In addition, the data belong to four different groups.

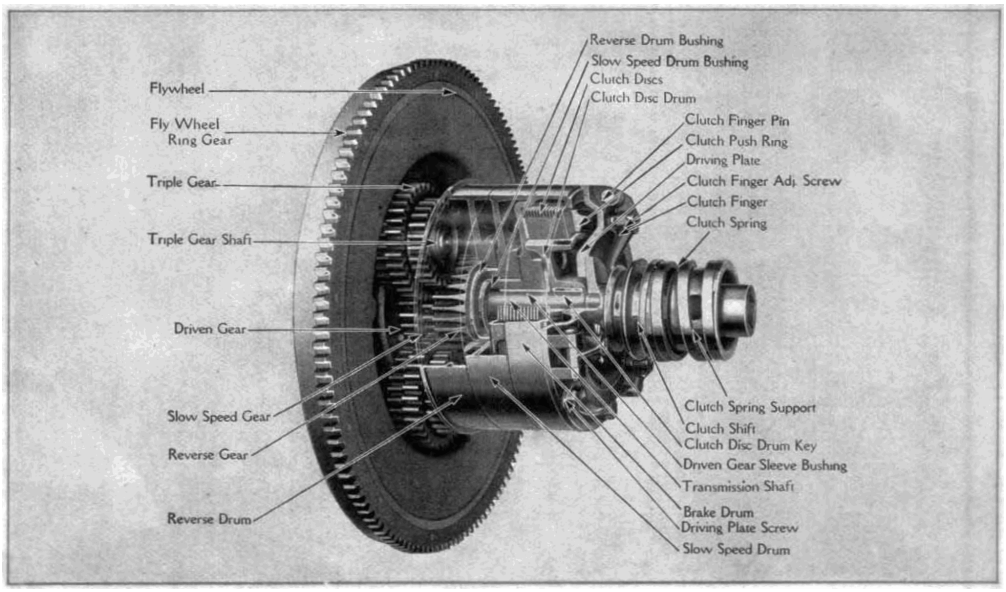
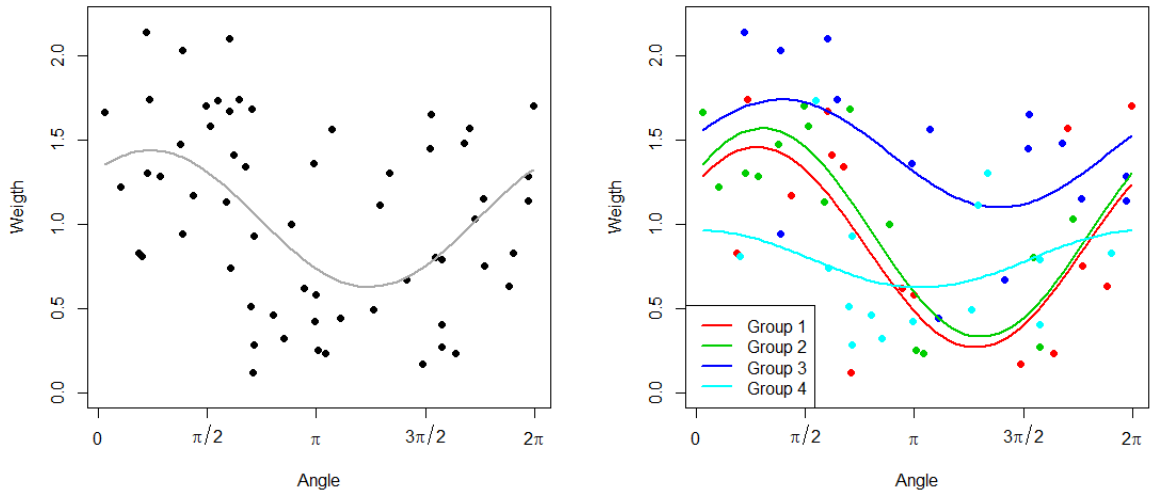


Figure 5.1: Transmission system of a car, including a flywheel on the left, from Wikimedia Commons (2013).

A flywheel is a device used in different machines designed to regulate an engine's rotation. It is, essentially, a heavy wheel attached to a rotating shaft (see Figure 5.1) and it is used to increase the machine's momentum, in order to produce stability and to store rotational energy in an efficient way. In vehicles production it is necessary to balance flywheels to ensure that the rotation transmits minimal vibration, since such vibration can lead to the malfunctioning of the system. When correcting the balance, the response obtained is cylindrical: an angular component measuring the angle of imbalance and a linear component evaluating the magnitude of the correction required to balance the flywheel. Modeling the relationship between the angle and the magnitude of correction can be helpful for a better understanding of the process, leading to the minimization of the costs by creating more efficient designs. In addition, a metallic molding is used in the production process. The metal's purity and density can influence both the angle and the correction magnitude.

The data given in Anderson-Cook (1999) contains measurements of the angles of imbalance of 60 flywheels, as well as the measurements of the corrections required (in inch-ounces). Four different kinds of metals were employed, with 15 flywheels corresponding to each type of metal. The goal of this chapter is twofold: first, to analyze the dependence of the two variables through a significance test. The second objective is to test whether the regression curves for the four types of metal are the same or, in case of concluding that they are different, to test if they are parallel. However, one must note that the data size in this example is small, since when obtaining the regression function estimators in each group only 15 observations are used.

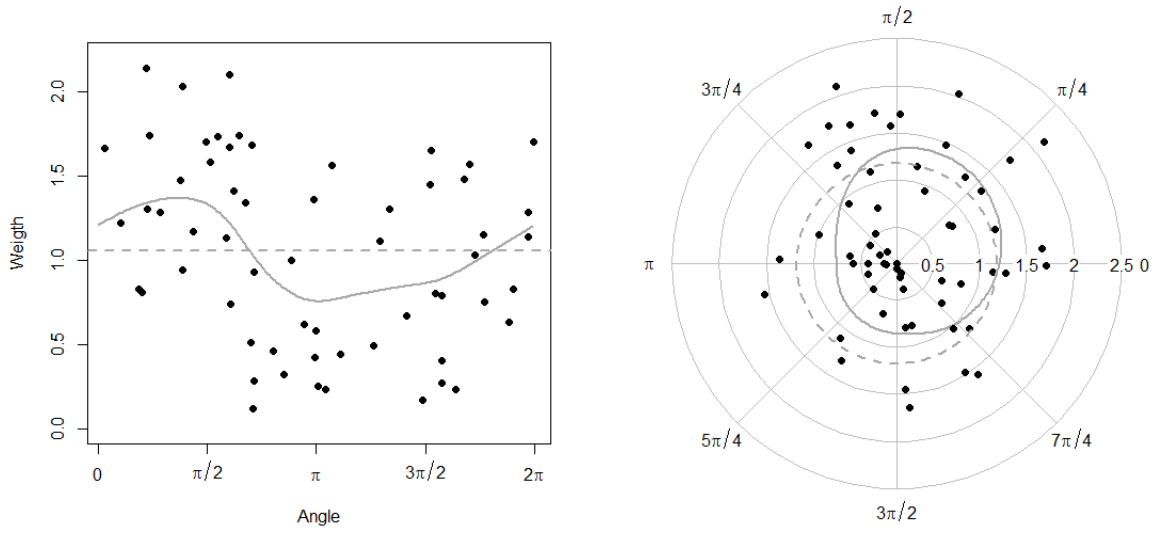


(a) (b)

Figure 5.2: Scatter plots of the angle of imbalance (in radians) against the measurements of correction (in inch-ounces). (a) Regression curve estimated parametrically with all the data. (b) Regression curves estimated parametrically for each group.

Anderson-Cook (1999) focuses on the detection of different regression functions, estimating the curves parametrically, using model (2.5). Figure 5.2 presents two scatter plots of the data with the parametric estimations of the regression curves. The left image shows the regression curve obtained using all the data while the image in the right displays four parametric estimations, one for each metal. In order to test if the regression functions are the same for all the groups, the approach taken by Anderson-Cook (1999) is the one described in Section 2.4.2. The author uses the test for the

interaction model, obtaining a p-value of .012, smaller than .05, and concluding, for this significance level, that a single function is inadequate to describe the behavior of the data. Afterwards, the author uses the test of parallelism, obtaining a large p-value, .32, and coming up to the conclusion of no evidence against parallel regression functions. Therefore, since the equality hypothesis was rejected and the parallelism hypothesis was not, the final conclusion is that the four curves are parallel.



(a) (b)

Figure 5.3: Scatter plots of the angle of imbalance (in radians) against the measurements of correction (in inch-ounces) for the four groups. The continuous line is the nonparametrically estimated regression curve for all the data. The dashed line represents the mean of the responses. (a) Linear representation. (b) Circular representation.

There is no justification of a particular parametric model in the original paper. Thus, the flywheel example will be analyzed with nonparametric methods, using the proposal described in Section 3.2.1 to test the significance of the predictor variable, and the proposal in Section 3.3.1 to test the existence of different regression functions according to the groups.

A single nonparametric model can be constructed, without considering the different groups, as in Figure 5.3 (continuous line), where the regression function was estimated with the circular-linear nonparametric estimator (see Section 3.1.1) using all the data. The dashed line represents the mean of the responses, corresponding to the estimated model under the hypotheses of no effect of the covariate. The nonparametric estimation of the regression function changes for the different values of the predictor variable, but it could be possible that the responses did not depend on the angle of imbalance and that the features of the curve were due to sample noise. To ascertain this, in the first place the CircSiZer map was constructed, obtaining the image shown in Figure 5.4. The map shows that the regression function is significantly increasing in the fourth quadrant for many values of the concentration parameter, and it is significantly decreasing in the second quadrant for most values of κ.

Given the CircSiZer map, it seems clear that the predictor variable affects the response. Even so, the no-effect test for circular predictors (presented in Section 3.2.1) is applied to the data. A range of smoothing parameters between 0 and 15 was considered, obtaining a p-value for each bandwidth. The

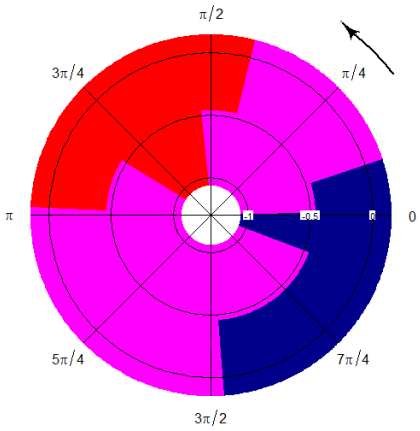


Figure 5.4: CircSiZer map for circular-linear regression for the flywheel data.

results are displayed in Figure 5.5. For all the smoothing parameters, the obtained critical value lie below the nominal level $\alpha = .05$, therefore the null hypothesis of no effect of the predictor is rejected for all the considered parameters.

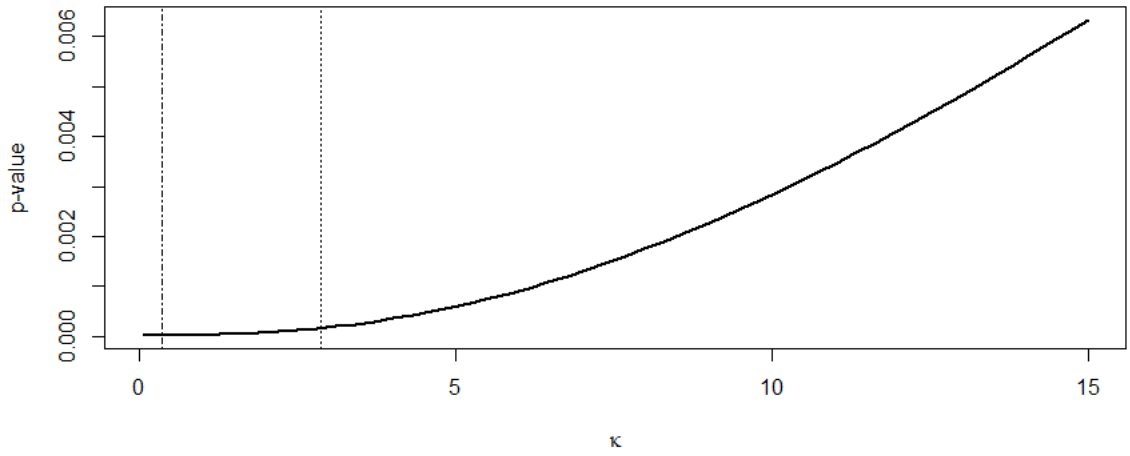
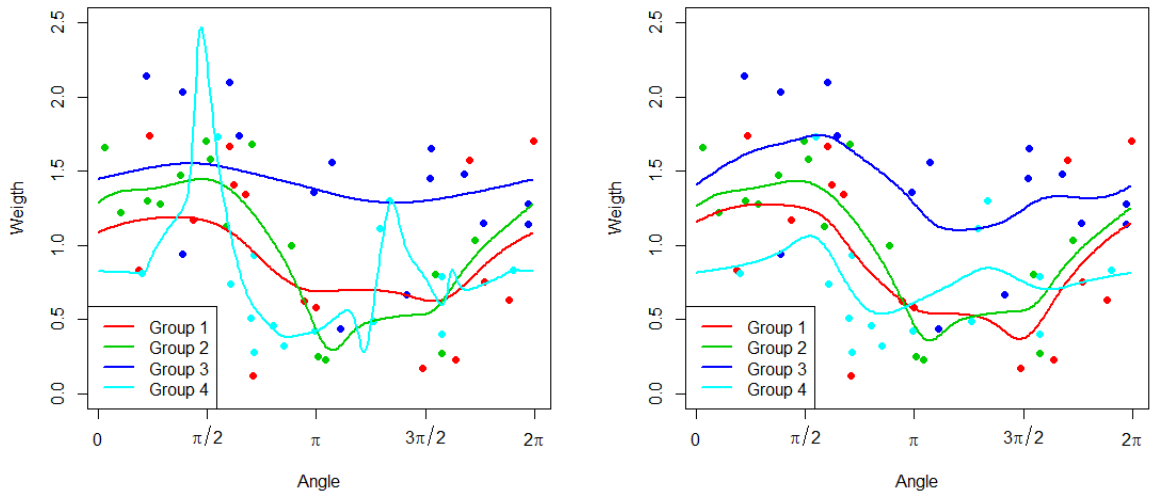


Figure 5.5: Trace of the no effect test applied to the flywheel data. Dotted vertical line representing the bandwidth selected by cross-validation. Dotted-dashed vertical line representing $\frac{1}{8}cv$.



(a) (b)

Figure 5.6: Scatter plots of the angle of imbalance (in radians) against the measurements of correction (in inch-ounces) for each of the groups with the nonparametric estimators for circular-linear regression. (a) Smoothing parameters selected by cross-validation in each group. (b) Same smoothing parameter is used in all groups.

However, since the metal used in the molding is different, it could be possible to have different regression curves for the groups. Figure 5.6 shows the data with different colors for each of the groups, with their corresponding estimated regression curves. In the left panel the smoothing parameter was selected by cross-validation in each of the groups. However, the equality and parallelism tests will be applied to the data using only one smoothing parameter for all the groups, in order to avoid the bias. The right panel of the figure shows the estimation of the regression functions in each group with the concentration parameter selected by applying cross-validation to the whole dataset.

The test of equality is first applied to the data, to test if all the regression curves are the same. The value of the statistic obtained is 20.96, while the p-value is .0263, lower than the nominal level $\alpha = .05$. Thus, for that significance level, the hypothesis of equal regression curves is rejected. This result is obtained using the smoothing parameter selected by cross-validation for all the data and the variance estimator (3.8). For a better application of the test, a sequence of concentration values is used. Figure 5.7 shows the trace of the test for the sequence of smoothing parameters. It is shown that the equality assumption is not rejected for concentration values approximately larger than 5, although given the sample size, large smoothing parameters are quite unrealistic in practice. Then, it can be concluded that there is evidence for saying that the four regression curves are not equal for a significance level of .05.

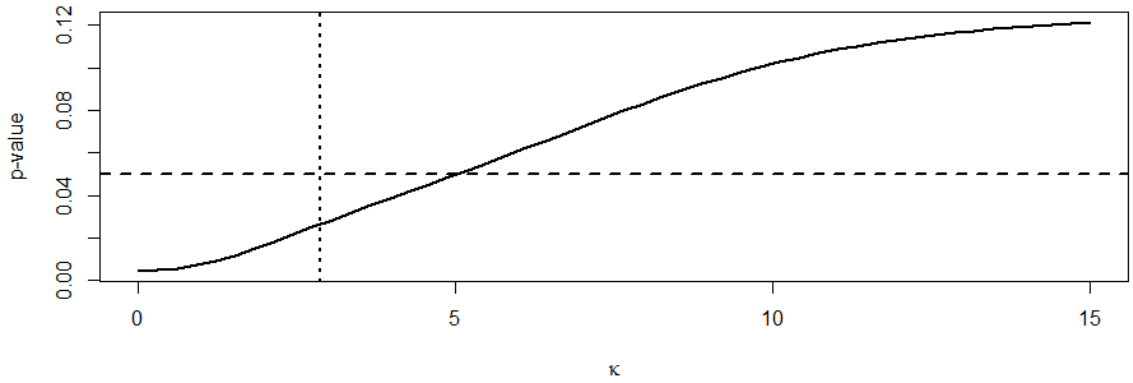


Figure 5.7: Trace of the equality test applied to the flywheel data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data. Horizontal dashed line representing the nominal level $\alpha = .05$.

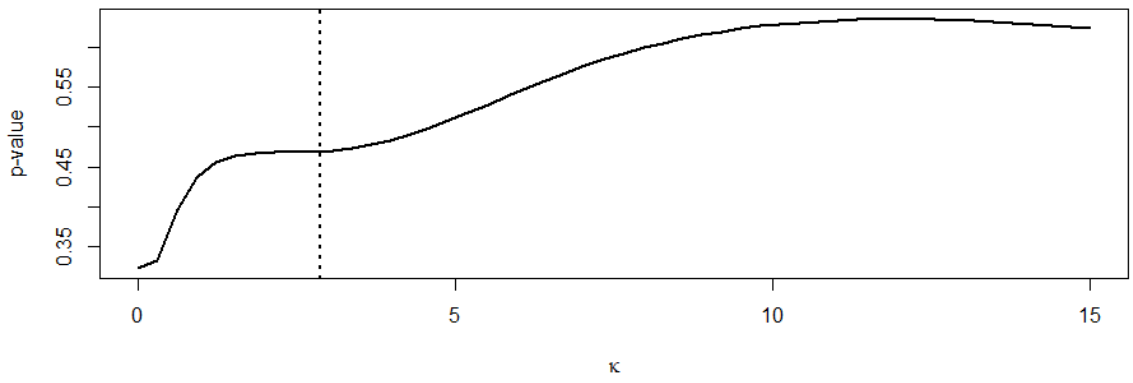


Figure 5.8: Trace of the parallelism test applied to the flywheel data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data.

Once the equality hypothesis is rejected, it could be checked if the regression curves are parallel. The parallelism test is applied with the smoothing parameter selected by cross-validation, and the obtained value for the test statistic is 5.44, while the p-value is 0.4695, much greater than $\alpha = .05$. Thus, there is no evidence for rejecting the null hypothesis of parallel regression curves. Again, to avoid difficulties derived from the selection of the smoothing parameter, the test is applied using a range of smoothing values. The trace of the test is displayed in Figure 5.8, which shows that the null hypothesis is not rejected for $\alpha = .05$ for any of the smoothing parameters considered (κ lying between .05 and 15).

5.2 Sandhoppers data

Biologists and zoologists have focused their research on analyzing the sources of variation in animal orientation under natural conditions. For some animals it is easy to observe their behavior and movements, but for some others the process requires more effort. A particular case is the study of sandhoppers (talitridae family) which are amphipod crustaceans (see Figure 5.9) known for their typical "hopping" movement, from which their common name is derived. The behavior of sandhoppers has been studied for years. They usually remain burrowed in the wet sand during the day and go out to the surface from sunset to sunrise. When they are displaced during the day due to the drying up of the sand layers, they return to the wet sand near the water. Several authors have linked the direction of sandhoppers movements to different factors, such as the slope of the beach or landscape features. Pardi and Ercolini (1986) stated that in Mediterranean shores the movement on sandhoppers relies on sun orientation, while this does not happen in Atlantic shores. Differences in sandhoppers orientation also arise between those living on tidal and non-tidal beaches, because of the risk of being swept away.

Figure 5.9: Sandhopper of the species *Talitrus saltator*, from Wikimedia Commons (2017).

In this section, data containing orientation directions of sandhoppers will be analyzed. The dataset was first studied in Scapini *et al.* (2002). It contains observations involving two sympatric sandhopper species (*Talitrus saltator* and *Talorchestia brito*). The study was conducted in an exposed and non-tidal beach in Tunisia, the Zouara beach. The main goal of the experiment was to analyze the direction of movement given different conditions. In order to record the data, the experimenters used two different circular arenas with 72 cross pit-fall traps each, placed at the circumference. The animals were released in the arenas, and once they made an orientation choice they were caught in one of the traps, which were separated from each other by an angle of 5°. One of the arenas allowed the view of both the sky and the landscape, while in the other the landscape was screened off, so that only the sky was visible. In addition to the direction of movement, other variables were recorded, such as sun azimuth, the species of sandhopper, the type of arena, the sex of the animal and several climatic variables (air temperature, air relative humidity and atmospheric preasure). The experiments were conducted in two different seasons (April and October).

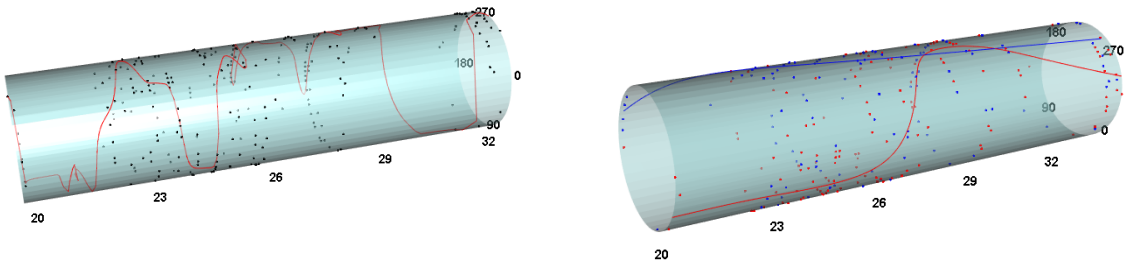
With the objective of studying the variation of directions taken by the sandhoppers in the presence of several factors, Scapini *et al.* (2002) use the Projected Multivariate Linear Model (PMLM) proposed by Presnell *et al.* (1998) (see Section 2.3.2). However, when using this model the authors consider all the predictor variables as discrete variables, factorizing the continuous ones such as the sun azimuth or the climatic variables.

Marchetti *et al.* (2003) also analyzed this dataset, focusing only on the data obtained for the

sandhoppers of the species *Talitrus saltator*. In this case the authors consider the variable sun azimuth as continuous, and fit regression models to explain the movement direction of the animals. As an exploratory tool, they fit a nonparametric regression model with the movement directions as a response variable and the sun azimuth as a predictor variable, and try to graphically asses the differences between the screened and the unscreened groups. However, the regression smoothers used are meant for both linear responses and linear predictors. The authors also apply the PMLM as in Scapini *et al.* (2002), but this time considering the sun azimuth as a continuous linear variable.

The objective in this section is to apply the nonparametric significance tests proposed in Section 3.2.2 and the ANCOVA tests proposed in Section 3.3.2 to the sandhoppers data. As in Marchetti *et al.* (2003), the animals considered will be those of the species *Talitrus saltator*, but only the data recorded on the October season will be employed. An analogous study could be conducted with the data recorded in the spring season, but it will not be shown here because the range of the variables which will be used as predictors is very small in this season, and the study lacks of interest. Two different regression models will be used: the first one treating the angle of direction as the dependent variable and the temperature as the predictor (therefore leading to a linear-circular regression), while the second model considers the angle of direction as a function of the sun azimuth, thus obtaining a circular-circular regression model. For the ANCOVA tests, the groups considered will be the sandhoppers placed in a screened arena and the ones situated in unscreened ones. The total number of observations is 278, with 146 belonging to the unscreened group and 132 in the screened group.

To begin with, the relationship between the angle of direction of the sandhoppers and the temperature will be analyzed. Figure 5.10a shows a representation on the cylinder of the angle of direction against temperature, with the estimated regression curve obtained with the cross-validation method for selecting the bandwidth. The first goal here is to ascertain if the temperature actually has an effect on the responses, for which the nonparametric significance test for linear-circular regression is used. The significance test introduced in Section 3.2.2 was applied using 1000 bootstrap replicates and over a sequence of smoothing parameters between .05 and 3. The resultant p-value was zero in all cases, concluding that there are evidences to reject the null hypothesis of no effect of the temperature over the direction of movement.



(a) (b)

Figure 5.10: Representations on the cylinder of the direction of movement against temperature with the estimated regression curve with the smoothing parameter selected by cross-validation. (a) Without distinguishing by group. (b) Distinguishing by group: unscreened group in blue and screened group in red.

Once it is known that the direction of movement is actually influenced by the temperature, the question relies on whether the regression functions for the screened and the unscreened animals are the same. Figure 5.10b shows representations of the data distinguishing between the screened and the unscreened groups, with the estimated regression functions. The smoothing parameter was selected by cross-validation in each group.

The plots suggest that the behavior of the screened animals was quite different from the behavior of the unscreened sandhoppers. This issue can be assessed by using the nonparametric test of equality for linear-circular regression. The test was applied to the data with the smoothing parameter selected by cross-validation and using 1000 bootstrap replicates, obtaining a critical value of .001. Then, there are evidences for a significance value $\alpha = .05$ to conclude that the two regression curves are different. Because the test depends on the smoothing parameter, its significance trace over a sequence of bandwidths is displayed in Figure 5.11. The corresponding p-values were much smaller than the significance level .05 in all cases, supporting the previous conclusion.

The test of parallelism is also applied to the data, obtaining a p-value of 0 when using the smoothing parameter selected by cross-validation and 1000 bootstrap replicates. Figure 5.12 shows the trace of the parallelism test, with all the critical values below .05 and rejecting, for this significance level, that the two curves are parallel, in favor of different regression curves for each group.

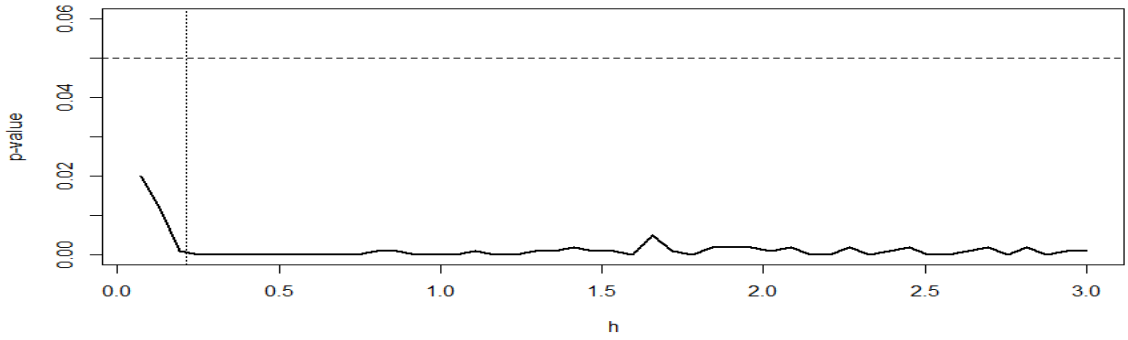


Figure 5.11: Trace of the equality test for linear-circular regression applied to the sandhoppers data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data. Horizontal dashed line representing the nominal level $\alpha = .05$.

Now the regression relationship between the direction of movement and the sun azimuth will be studied. As mentioned before, some studies have linked the direction of movement of sandhoppers in Mediterranean beaches to the position of the sun. Figure 5.13a displays a representation of the direction of movement against the sun azimuth on the torus, with the estimation of the regression function.

Now the objective lies on determining if the sun azimuth affects the direction of the animals. For such purpose it is necessary to consider several concentration parameters in order to apply the no-effect test for circular-circular regression. The significance trace of the test (using 1000 bootstrap replicates) is displayed in Figure 5.14, which shows that the p-values are much lower than .05 for all the values of the smoothing parameter considered. Therefore, for that significance level it is rejected that the sun azimuth has no effect on the direction of movement of the sandhoppers.

The next objective consists on studying if the relationship between the direction of movement and the sun azimuth is different for the two groups of sandhoppers. The estimated regression curves for

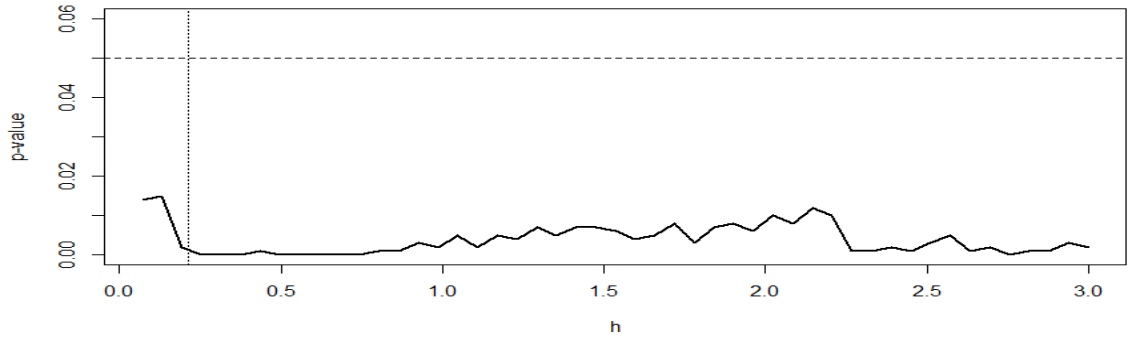
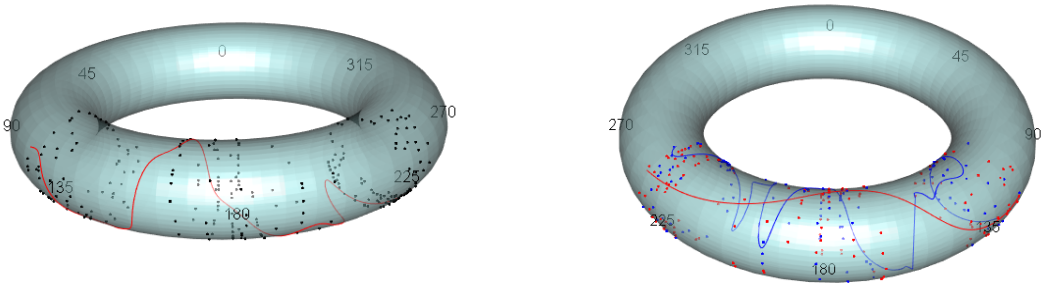


Figure 5.12: Trace of the parallelism test for linear-circular regression applied to the sandhoppers data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data. Horizontal dashed line representing the nominal level $\alpha = .05$.

each group are represented in Figure 5.13b, where the smoothing parameter was selected by applying the cross-validation method in each group. The obtained smoothing parameter in the screened group turned out to be quite large, resulting in a undersmoothed estimator of the regression function. However, recall that the test of equality is applied using the same concentration parameter to the global estimator and to the group estimators, in order to cancel out the bias. When this parameter is the one obtained by cross-validation (for all the data), the p-value of test is .011, lower than the significance level $\alpha = .05$. In order to avoid wrong conclusions because of the selection of the smoothing parameter, the test is also applied using different values for the concentration. Figure 5.15 displays the trace of the test, showing that the critical value lies below .05 for all the considered concentration parameters. Consequently, for that value of α, there are evidences to reject the hypothesis of equal regression curves for the sandhoppers placed in the unscreened and screened arenas.

Lastly, one can wonder if the two groups of sandhoppers change their movements in an equal manner depending on the sun azimuth, but one group just shifts their direction in an specific angle because of the presence or absence of the screen in the arena. The test of parallelism will be applied in this case, using 1000 bootstrap replicates. If the smoothing parameter selected by cross-validation is employed, the corresponding critical value for the test is .011, lower than $\alpha = .05$. Figure 5.16 present the significance trace of the test. The p-value lies below α for concentration values approximately larger than 20, while for lower concentrations the critical value is larger than the nominal level. However, taking into account the sample size, it is expected that the optimal concentration parameter is large, so the parameters for which the p-value is larger than α are quite unrealistic in practice. It can be concluded, then, that there are evidences to reject the parallelism of the two regression function.

To conclude, it has been shown that the effect of both the temperature and the sun azimuth on the direction of movement of the sandhoppers is significant. In addition, it has been concluded that the relationship between the angle of direction and the temperature and the angle of direction and the sun azimuth are different for the screened and the unscreened groups of sandhoppers.



(a) (b)

Figure 5.13: Representations on the torus of the direction of movement against sun azimuth with the estimated regression curve and the smoothing parameter selected by cross-validation. (a) Without distinguishing by group. (b) Distinguishing by group: unscreened data in blue and screened data in red.

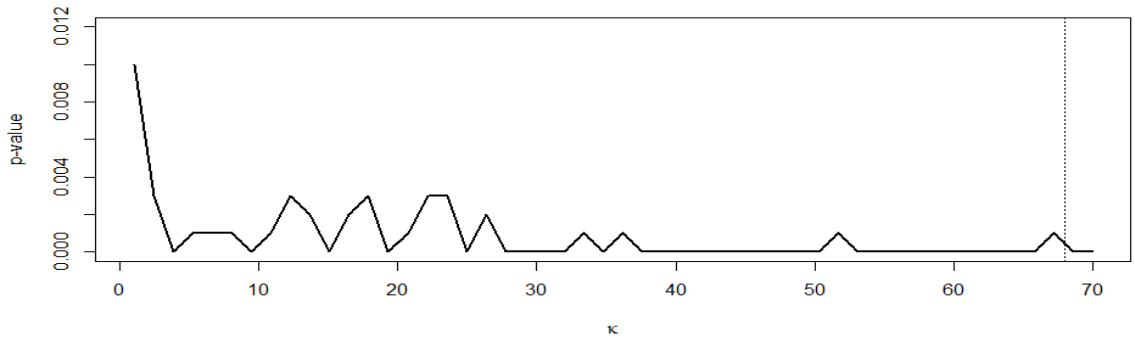


Figure 5.14: Trace of the significance test for circular-circular regression applied to the sandhoppers data. Dotted vertical line representing the bandwidth selected by cross-validation.

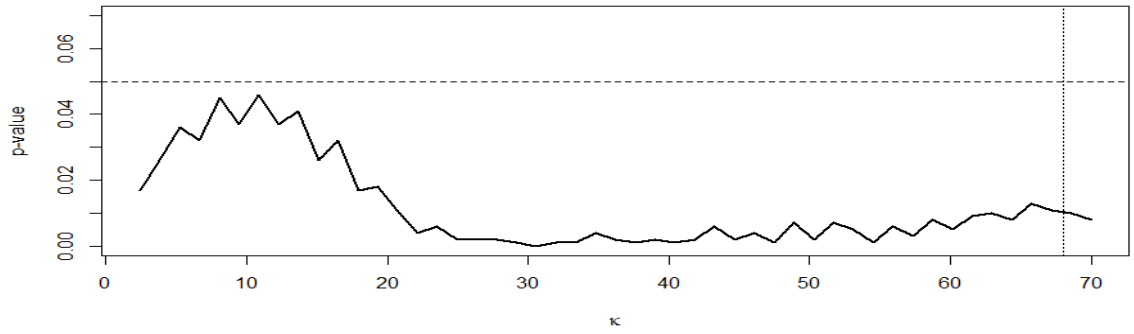


Figure 5.15: Trace of the equality test for circular-circular regression applied to the sandhoppers data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data. Horizontal dashed line representing the nominal level $\alpha = .05$.

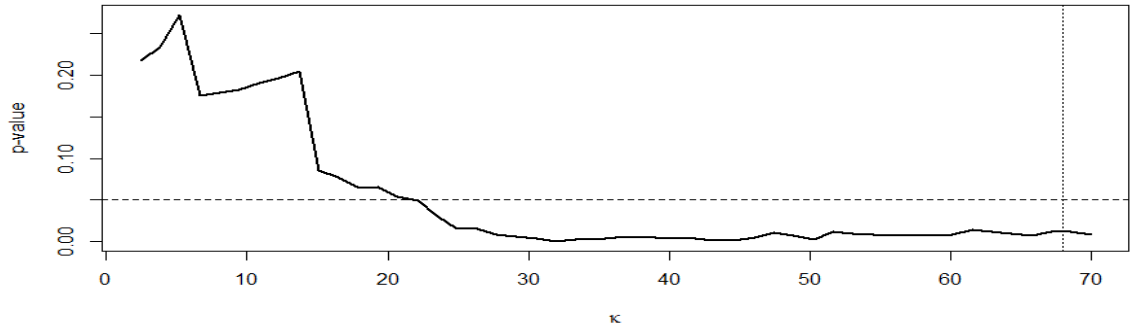


Figure 5.16: Trace of the parallelism test for circular-circular regression applied to the sandhoppers data. Dotted vertical line representing the bandwidth selected by cross-validation using all the data. Horizontal dashed line representing the nominal level $\alpha = .05$.

Chapter 6

Conclusions and discussion

Two goals have been accomplished in this MSc Thesis. In the first place, new nonparametric significance tests for regression involving circular variables have been proposed. The three tests have shown a high dependence on the smoothing parameter and simulations showed that with an adequate selection of the bandwidth, the tests are well calibrated. However, the tests must be applied in practice by constructing the significance trace of the tests. The second and main accomplished objective present in this manuscript is the construction of new nonparametric analysis of covariance models for circular regression. In the circular-linear context, the proposed tools for testing equality and parallelism were found to be well calibrated in all the scenarios considered, while their counterparts for linear data were found to be anticonservative in some settings. As for the power of the tests, it was ascertained that the new circular proposals were able to compete with the linear tests, even outperforming them in some cases. At the same time, in the regression scenarios where the response variable is circular (linear-circular and circular-circular regression) the novel methods proposed are the only tools available to correctly test equality and parallelism with this type of data, since methods for real-valued data provide a deficient calibration in these cases.

Regarding some possible extensions of the proposed methods, the ANCOVA models could be adapted to the multivariate case. Recently, the nonparametric regression estimators for circular data have been extended to the multivariate context, where the response variable depends on two or more predictors. Such estimators could be used to construct nonparametric ANCOVA models for circular data with two covariates. Moreover, the problem of considering more than one factor variable is a possible extension that could be studied in the future. In addition, there might be cases where at the time of representing the data, the existence of different groups seems clear, but where the factor variables are not known. In such scenarios, modal regression is a useful alternative. Therefore, a future line of work could be the study of modal regression models for circular variables.

To conclude, the computational developments carried out in this project must be highlighted. The different tests proposed in this work (no-effect and ANCOVA) have been programmed in the statistical software `R`. The new tools are to be included in the library `NPCirc` (Oliveira *et al.*, 2014b), which contains nonparametric methods for density and regression estimation involving circular variables. Again, the Supercomputing Center of Galicia (CESGA) must be recognized for supplying the resources necessary to perform most of the simulations included in this project.

Bibliography

[1] Agostinelli C and Lund U (2017). R package 'circular': Circular statistics (version 0.4-93). `https://r-forge.r-project.org/projects/circular/`.

[2] Ameijeiras-Alonso J, Lagona F, Ranalli M and Crujeiras RM (2019) A circular nonhomogeneous hidden Markov field for the spatial segmentation of wildfire occurrences. *Environmetrics*, 30(2).

[3] Anderson-Cook CM (1999) A tutorial on one-way analysis of circular-linear data. *Journal of Quality Technology*, 31(1), 109-119.

[4] Anderson E (1935) The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59(1), 2-5.

[5] Bai ZD, Rao CR and Zhao LC (1988) Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27(1), 24-39.

[6] Bartels R (1984) Estimation in a bidirectional mixture of von Mises distributions. *Biometrics*, 40(1), 777-784.

[7] Batschelet E (1981) *Circular Statistics in Biology.* Academic Press, London.

[8] Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(1), 353-360.

[9] Bowman AW and Azzalini A (1997) *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus illustrations* (Vol. 18). OUP Oxford.

[10] Bowman AW and Azzalini A (2018) R package 'sm': nonparametric smoothing methods (version 2.2-5.6). `https://cran.r-project.org/web/packages/sm`.

[11] Buckley MJ and Eagleson GK (1988) An approximation to the distribution of quadratic forms in normal random variables. *Australian Journal of Statistics*, 30(1), 150-159.

[12] Byrd RH, Peihuang LU, Nocedal J and Zhu C (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(1), 1190-1208.

[13] Cao R and González-Manteiga W (1993) Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics*, 2(1), 379-388.

[14] Chaudhuri P and Marron JS (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447), 807-823.

[15] Cogburn R and Davis HT (1974) Periodic splines and spectral estimation. *The Annals of Statistics*, 2(1), 1108-1126.

[16] Crujeiras RM (2017) An introduction to statistical methods for circular data. *Boletín de Estadística e Investigación Operativa*, 33(2), 85-107.

[17] Dette H and Neumeyer N (2001) Nonparametric analysis of covariance. *The Annals of Statistics*, 29(5), 1361-1400.

[18] Di Marzio M, Panzera A, and Taylor CC (2009) Local polynomial regression for circular predictors. *Statistics & Probability Letters*, 798(1), 2066-2075.

[19] Di Marzio M, Panzera A, and Taylor CC (2012) Non-parametric regression for circular responses. *Scandinavian Journal of Statistics*, 40(2), 238-255.

[20] Di Marzio M, Panzera A, and Taylor CC (2014) Nonparametric regression for spherical data. *Journal of the American Statistical Association*, 109(506), 748-763.

[21] Fan J (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420), 998-1004.

[22] Fan J and Gijbels I (1996) *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

[23] Faraway JJ (2004) *Linear Models with R*. Chapman and Hall, Boca Raton.

[24] Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.

[25] Fisher NI (1989) Smoothing a sample of circular data. *Journal of Structural Geology*, 11(1), 775-778.

[26] Fisher NI and Hall PG (1991) Bootstrap algorithms for small sample. *Journal of Statistical Planning and Inference*, 27(1), 157-169.

[27] Fisher NI and Lee AJ (1992) Regression models for an angular response. *Biometrics*, 48(1), 665-77.

[28] García-Portgués E, Van Keilegom I, Crujeiras RM and González-Manteiga W (2016) Testing parametric models in linear-directional regression. *Scandinavian Journal of Statistics*, 43(4) 1178-1191.

[29] Gasser T, Sroka L, and Jennen-Steinmetz C (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(1), 625-633.

[30] Gatto R and Jammalamadaka SR (2007) The generalized von Mises distribution. *Statistical Methodology*, 4(3), 341-353.

[31] Gould AL (1969) A regression technique for angular variates. *Biometrics*, 25(1), 683-700.

[32] Gradshteyn IS and Ryzhik IM (1994) *Table of Integrals, Series, and Products (5th Ed)*. Academic Press, San Diego.

[33] Gumbel EJ, Greenwood JA and Durand D (1953) The circular normal distribution: theory and tables. *Journal of the American Statistical Association*, 48(261), 131-152.

[34] Hall P, Watson GP and Cabrera J (1987) Kernel density estimation for spherical data. *Biometrika*, 74(4), 751-762.

[35] Hardle W and Bowman AW (1988) Bootstrapping in nonparametric regression: local adaptive smoothinf and confidence bands. *Journal of the America Statistical Association*, 83(1), 102-110.

[36] Hardle W and Marron JS (1991) Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19(1), 778-796.

[37] Hurvich CM, Simonoff JS and Tsai C (1988) Smoothing parameter selection in nonparametric regression using an improved Akaike Information Criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(2), 271-293.

[38] Jammalamadaka SR and SenGupta A (2001) *Topics in Circular Statistics.* World Scientific, Singapore.

[39] Jander R (1957) Die optische richtungsorientierung der roten waldameise (formicaruea l.) *Zeitschrift für vergleichende Physiologie*, 40(1), 162-238.

[40] Johnson NL and Kotz S (1972) *Distributions in Statistics: Continuous Univariate Distributions, Vol. II.* Wiley, New York.

[41] Johnson RA and Wehrly TE (1978) Some angular-linear distributions and related regression models, *Journal of the American Statistical Association*, 73(1), 602-606.

[42] Jona-Lasinio G, Gelfand A and Jona-Lasinio M (2012) Spatial analysis of wave direction data using wrapped gaussian processes. *Annals of Applied Statistics*, 6(4), 1478-1498.

[43] Jones MC and Pewsey A (2005) A family of symmetric distributions on the circle. *Journal of the American Statistical Association*, 100(472), 1422-1428.

[44] Kato S and Jones MC (2015) A tractable and interpretable four-parameter family of unimodal distributions on the circle, *Biometrika*, 102(1), 181-190.

[45] Ley C and Verdebout T (2017) *Modern Directional Statistics.* Chapman & Hall, Boca Raton.

[46] Marchetti G and Scapini F (2003) Use of multiple regression models in the study of sandhopper orientation under natural conditions. *Estuarine, Coastal and Shelf Science.* 58. 207-215.

[47] Mardia KV (1972) *Statistics of Directional Data.* Academic Press, New York.

[48] Mardia KV and Jupp PE (2000) *Directional Statistics.* John Wiley, Chichester.

[49] Mardia KV and Sutton TW (1975) On the modes of a mixture of two von Mises distributions. *Biometrika*, 62(1), 699-701.

[50] Maxwell SE and Delaney HD (2003) *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* Mahwah, New Jersey: Lawrence Erlbaum.

[51] Mooney JA, Helms PJ and Jollife IT (2003) Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, 41(1), 505-513.

[52] Nadaraya EA (1964) On estimating regression. *Theory of Probability and its Applications*, 9(1), 141-142.

[53] Oliveira M, Crujeiras RM and Rodríguez-Casal A (2013) Nonparametric circular methods for exploring environmental data. *Journal of Environmental and Ecological Statistics*, 20(1), 1-17.

[54] Oliveira M, Crujeiras RM and Rodríguez-Casal A (2014a) CircSiZer: an exploratory tool for circular data, *Journal of Environmental and Ecological Statistics*, 21(1), 143-159.

[55] Oliveira M, Crujeiras RM and Rodríguez-Casal A (2014b) NPCirc : An R package for nonparametric circular methods. *Journal of Statistical Software*, 61(1), 1-26.

[56] Pardi L and Ercolini A (1986) Zonal recovery mechanisms in talitrid crustaceans. *Bollettino di Zoologia Italiana*, 53(1), 139-160.

[57] Parzen E (1962) On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(1), 1065-1076.

[58] Pewsey A, Neuhuser M, and Ruxton GD (2013) *Circular Statistics in R.* Oxford University Press, Oxford.

[59] Politis DN (2014) Bootstrap confidence intervals in nonparametric regression without an additive model.*Springer Proceedings in Mathematics and Statistics*, 74(1), 271-282.

[60] Presnell B, Morrison SP and Littel RC (1998) Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, 93(443), 1068-1077.

[61] R Core Team (2018) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

[62] Ratkowsky DA (1983) *Nonlinear Regression Modelling* (Statistics: Textbooks and Monographs, Volume 48), New York: Marcel Dekker.

[63] Rice J (1984) Bandwidth choice for nonparametric kernel regression. *The Annals of Statistics*, 12(1), 1215-1230.

[64] Rosenblatt M (1956) Remarks on some nonparametric estimate of a density function. *The Annals of Mathematical Statistics*, 27(1), 832-837.

[65] Ruppert D, Sheather SJ and Wand MP (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(1), 1257-1270.

[66] Scapini F, Aloia A, Bouslama MF, Chelazzi L, Colombini I, El Gtari M, Fallaci M, and Marchetti, GM (2002) Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, Talitrus saltator and Talorchestia brito, from an exposed Mediterranean beach. *Behavioural Ecology and Sociobiology*, 51(1), 403-414.

[67] Sheather SJ (2009) *A Modern Approach to Regression with R.* Springer, New York.

[68] Sheather SJ and Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53(1), 683-690.

[69] SenGupta S and Rao JS (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(1), 165-174.

[70] Small CG (1996) *The Statistical Theory of Shape*, Springer, New York.

[71] Solomon H and Stephens MA (1977) Distribution of a sum of weighted chi-square variables. *Journal of the American Statistical Association*, 72(360), 881-885.

[72] Speckman P (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3), 413-436.

[73] Spurr BD (1981) On estimating the parameters in mixtures of circular normal distributions. *Mathematical Geology*, 13(1), 163-173.

[74] Stephens MA (1972) Multisample tests for the von Mises distribution. *Journal of the American Statistical Society*, 67(1), 456-461.

[75] Von Mises R (1918) Uber die 'ganzzahligkeit' der atomgewichte und verwandte fragen. *Physikalische Zeitschrift*, 19(1), 490-500.

[76] Wahba G (1990) *Spline Models for Observational Data*. Society for Industrial Mathematics, Philadelphia.

[77] Wand MP and Jones MC (1995) *Kernel Smoothing*. Chapman & Hall, New York.

[78] Watson GS (1964) Smooth regression analysis. *Sankhya*, series A, 26(1), 359-372.

[79] Watson GS (1983) *Statistics on Spheres*. John Wiley, New York.

[80] Watson GS and Williams E (1956) On the construction of significance tests on the circle and the sphere. *Biometrika*, 43(1), 344-352.

[81] Wikimedia Commons (2013) Transmission, from author: Ford Motor Company. `https://commons.wikimedia.org/wiki/File:Ford_model_t_1919_d029_transmission.png`. (Online; accessed June 7, 2019).

[82] Wikimedia Commons (2017) Creepy crawlies beach closeups, from author: Magnus Hagdorn. `https://commons.wikimedia.org/wiki/File:Going_-_sandhopper_(Talitrus_saltator)_(8666909448).jpg`. (Online; accessed June 7, 2019).

[83] Wood S (2006) *Generalized Additive Models: An Introduction with R.* Chapman and Hall, Boca Raton.

[84] Wright K (2018) R package 'agridat': Agricultural Datasets. (version 1.16). `https://CRAN.R-project.org/package=agridat`.

[85] Young S and Bowman AW (1995) Nonparametric analysis of covariance. *Biometrics*, 51(1), 920-931.