Biometrics 64, 1–22 DOI: 10.1111/j.1541-0420.2005.00454.x
February 2015

Circular Piecewise Regression with an Application to Cell-cycle Biology

Cristina Rueda[1,†,*], Miguel A. Fernández[1,†,], Sandra Barragán[1,***],**

Kanti V. Mardia[2,**] and Shyamal D. Peddada[3,*****]**

[1] Departamento de Estadística e I.O., Universidad de Valladolid, 47011 Valladolid, Spain

[2] Department of Statistics, University of Oxford, Oxford, UK,

and Department of Statistics, University of Leeds, Leeds, UK

[3] Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

[†] These authors contributed equally to the paper

email: crueda@eio.uva.es

**email:* miguelaf@eio.uva.es

***email:* sandraba@eio.uva.es

****email:* K.V.Mardia@leeds.ac.uk

*****email:* peddada@niehs.nih.gov

Summary: Applications of circular regression models appear in many different fields such as evolutionary psychology, motor behavior, biology, and ,in particular, in the analysis of gene expressions in oscillatory systems. Specifically, for the gene expression problem, we need to model the relation among peak expressions of cell-cycle genes in two species with different cell phase lengths. This challenging problem reduces to the problem of constructing a piecewise circular regression model and, with this objective in mind, we propose a flexible circular regression model which allows different parameter values depending on sectors along the circle. We give a detailed interpretation of the parameters in the model and provide maximum likelihood estimators. We also provide a model selection procedure based on the concept of generalized degrees of freedom. The model is then applied to the analysis of two different cell-cycle data sets and, through these examples we highlight the power of our new methodology.

Key words: Circular data; Circular–circular regression; Change points; Gene expression; Generalized AIC; Von Mises distribution.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Analysis of circular data has a long history with well-developed theory and methodology documented in several books (see, for example, Fisher (1993); Mardia and Jupp (2000)). Until recently much of the literature was developed for describing circular models and drawing inferences on individual angular parameters, such as comparing the mean directions of two or more populations. In recent years, the circular-circular regression problem (i.e. when both regressor and response are circular variables) has received some attention (Kato and Jones (2010), Di Marzio et al. (2013)). Although there are some previous models for this problem (Sarma and Jammalamadaka (1993), Lund (1999)), the first tractable and interpretable model was proposed by Downs and Mardia (2002) under the assumption that the angular variables were distributed according to the von Mises distribution. More recently, several extensions to Downs and Mardia model were considered under different distributional assumptions. For example Kato et al. (2008) replaced von Mises by the wrapped Cauchy distribution and Kato and Jones (2010) considered extensions to von Mises and wrapped Cauchy distributions. More recently, a non-parametric approach to the problem was taken in Di Marzio et al. (2013), with Oliveira et al. (2014) providing the appropriate software for fitting the corresponding model. However, as we see below, we have found that none of these approaches gives plausible results for the application motivating this paper; the details of our application are as follow.

This paper is motivated by a problem encountered in cell biology where researchers are interested in correlating angular data from two or more sources (e.g. experiments or species). A cell-cycle among eukaryotic cells goes through four major phases of distinct biological functions, namely the G1, S, G2 and M phases. Genes participating in the cell-cycle tend to have a periodic pattern of expression over time. Consequently, the time to peak expression (known as the phase angle) of such genes can be mapped onto a unit circle.

Cell-biologists are often interested in drawing inferences regarding the phase angle of cell-cycle genes since they are considered to be associated with the gene's biological function (Jensen et al. (2006)). Often two types of inferences are of interest. For example, using the data obtained from a single experiment on a given species, one may be interested in estimating the phase angle of a set of cell cycle genes in that species when the relative order of peak expression is known (Rueda et al. (2009)). Another question of interest is to detect whether the order of the phase angles of a set of cell-cycle genes is consistent across multiple experiments on the same species (Liu et al. (2004)), or more broadly if the order of the phase angles of a set of cell-cycle genes is same across multiple species (Fernández et al. (2012)).

In this paper we develop a piecewise regression model able to relate the phase angles of two sets of data, such as data on two different experiments on the same species or two different species. While the regression model proposed in Downs and Mardia (2002) is likely to perform well when the duration of time spent by a cell in different phases of a cell-cycle is same across all species, it may be too rigid when the duration of time is not same across different species as the lengths of the four phases in which the cell-cycle is divided change from one species to another, so that the functional relationship between species may be different in each of the phases. For this reason, in Section 2 of this paper we introduce a flexible piecewise regression model that can be useful for drawing inferences when the duration of time spent in different phases by a cell varies across species.

Piecewise regression, although not defined for manifolds until now, has been well studied in the linear setting (see Seber and Wild (1989) or Motulsky and Christopoulos (2004)). To highlight some challenges in circular piecewise regression, we consider the simplest linear case. Namely, the case of a single change point with no error

$$y = a_1 + b_1 x, \quad x \leqslant c$$

$$y = a_2 + b_2 x, \quad x \geqslant c$$

with the continuity constraint

$$a_1 + b_1 c = a_2 + b_2 c. \tag{1}$$

We note that if x is a circular variable, the change point c has no meaning so there should be at least two change points. Further, in the linear case, this problem for computational purpose can be reparameterized as

$$y = A + Bx + C(x - D)SGN(x - D) \tag{2}$$

where now

$$c = D,\ a_1 = A + CD,\ a_2 = A - CD,\ b_1 = B - C,\ b_2 = B + C$$

so the constraint (1) is included in (2). We will see that such a simplification i s not available for the circular case. Of course, where the noise is added the inference problems become more intricate.

As we already mentioned, a possible non-parametric circular approach to tackle our problem might be that in Di Marzio et al. (2013). However, we note that this approach is not meaningful since in the cell-cycle the function relating the peak expressions has to be increasing as there is no return possible in the cycle and that the response run has to one cycle as the explicative variable runs through one cycle. These conditions are not always fulfilled when this non-parametric approach is considered.

There are many interesting biological questions related to cell cycle that can be answered from the circular piecewise model that we propose in this paper. Our proposed methodology can be useful for answering a variety of questions that are of potential interest to biologists. For example, we can test whether there is a significant difference in the duration of time spent by cells in different phases in two given species. Similar to Liu et al. (2004), we can test if the phase angles of cell-cycle genes from different labs/experiments on the same species are equal. We can also evaluate if the phase of peak expression has evolutionarily changed. To illustrate the methodology we use cell-cycle data available from the cyclebase data base

www.cyclebase.org (Santos et al. (2015)). This database contains data obtained from 20 different experiments conducted in different laboratories on budding yeast (*S. cerevisiae*), fission yeast (*S. pombe*).

The methodological contributions of this paper are provided in Section 2 where we develop the piecewise circular regression model (which is our first contribution), describe the interpretation of its parameters and their estimation. In this same Section, we also develop the other main contribution of this paper, namely, the construction the construction of a model selection procedure based on a recent criterion called Generalized Akaike Criterion.

In Section 3 we apply the proposed model to the analysis of cell-cycle gene expression data, which was the motivation for the methodology developed in this paper. We consider two different situations. In the first example, both data sets are from the same species but from different laboratories. In the second example, the data are obtained from two different species, namely, fission yeast and budding yeast. As noted in the literature (e.g. Fernández et al. (2012)), these two species of yeasts spend drastically different amounts of time in the different phases of cell cycle. Consequently, in this example we expect the proposed model to outperform the Downs-Mardia model. We illustrate how the model selection procedure distinguishes between these two cases and draw the appropriate conclusions for each case. Finally, in Section 4, we discuss variety of other biological applications, and indicate why the model is very general.

2. The circular piecewise regression model

In Section 2.1, we define the piecewise circular regression model with as many different functional relationships of the type of those in Downs and Mardia (2002) as sectors established in the circle. As usual with piecewise models we impose continuity on the global circular relationship. In this way, the model can cope with situations where the circular relationship among variables changes between those sectors. In the subsequent sections, Section 2.2–

Section 2.3, we describe the interpretation of the parameters and the maximum likelihood estimators; we show why the model is very flexible in the sense that conditions such as monotonicity can be imposed easily to allow for the restrictions on the parameters when needed. Section 2.4, contains the definition of a criterion for model selection based on the concept of generalized degrees of freedom by Ye (1998).

2.1 *The Model*

Consider a circular response variable ψ and a circular independent variable θ. We will denote as k the number of different pieces or sectors in the unit circle and as θ_i^*, $i = 1, 2, ..., k$ the sector borders (or change/break points in the linear piecewise regression model in the line) and we assume here that these borders are known. We denote as $\mathbf{\Theta}$ the vector of values for the independent variable with components θ_{ij} with $i = 1, 2, ..., k$ and $j = 1, 2, ..., n_i$, where the first index is the sector the observation belongs to, so that $\theta_i^* < \theta_{ij} \leqslant \theta_{i+1}^*$ with $\theta_{k+1}^* = \theta_1^*$, the second index j is the number of the observation in the corresponding sector, n_i is the number of observations in sector i and $N = \sum_{i=1}^{k} n_i$ is the total number of observations. Accordingly we denote as $\mathbf{\Psi}$ the vector of observed values and as ψ_{ij} the corresponding components of this vector. We further assume that ψ_{ij} given θ_{ij} comes from independent von Mises distributions $M(\mu_{ij}, \kappa)$ with density function $f(\theta_{ij}, \mu_{ij}, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta_{ij} - \mu_{ij})}$, where I_0 denotes the modified Bessel function of the first kind and order 0.

We now describe the circular–circular (c–c) regression model of Downs and Mardia (2002) where we take the circular response variable ψ, and the circular independent variable θ but in this case, there are no borders for θ. Further, ψ given θ comes from a von Mises distribution with $M(\mu, \kappa)$ and

$$\tan \frac{1}{2}(\mu - \beta) = \omega \tan \frac{1}{2}(\theta - \alpha), \tag{3}$$

where α and β are angular location parameters and ω is a slope parameter which is restricted to the closed interval $[-1, 1]$ by adjusting α and β appropriately. Next, we propose our

piecewise circular regression model with

$$\tan\frac{1}{2}(\mu_{ij}-\mu)=\omega_i\tan\frac{1}{2}(\theta_{ij}-\nu_i)\ \text{ for } j=1,\ldots,n_i \text{ and } i=1,\ldots,k, \tag{4}$$

where, to ensure continuity, we take

$$\omega_i\tan\frac{1}{2}(\theta_i^*-\nu_i)=\omega_{i-1}\tan\frac{1}{2}(\theta_i^*-\nu_{i-1})\ \text{ for } i=1,\ldots,k, \tag{5}$$

and $\nu_0=\nu_k$, $\omega_0=\omega_k$. This model maintains the functional relationship of the Downs and Mardia model but allows a different parameters in each of the sectors while imposing continuity on the global function. Equivalently, our model (4) can be rewritten as

$$\mu_{ij}=\mu+2\arctan\left(\omega_i\tan\frac{1}{2}(\theta_{ij}-\nu_i)\right), \tag{6}$$

or, simply $\mu_{ij}=f\left(\theta_{ij},\mu,\omega_1,\nu_1,\ldots,\omega_k,\nu_k\right)$ for $j=1,\ldots,n_i$ and $i=1,\ldots,k$.

Note that obviously $k>1$ and that for the simplest case of $k=2$, the model reduces to the Downs and Mardia model since from (5) $\nu_2=\nu_1$, $\omega_2=\omega_1$. Further, in view of the continuity condition, at least three sectors of the circle are required in order to fit a piecewise regression model on the circle.

2.2 *Interpretation of the model parameters*

Since the meaning of the parameters in our model is not straightforward (as for example in the normal linear regression model), we will give detail interpretation of each of them. Parameter μ can be interpreted as a global location parameter quantifying the rotation of the response that allows effective alignment with the independent variable. Since there are different sectors, different rotation parameters in each of the sectors are also needed for an appropriate alignment between the independent variable and the response; this is accounted for by the ν_i parameters. The ω_i are slope parameters for the θ_{ij} observations in the sector $(\theta_i^*,\theta_{i+1}^*]$. It is important to note that (i) there is dependence among the ν_i and the ω_i parameters as the continuity condition (5) has to be fulfilled and (ii) the model is piecewise non-linear so a higher ω_i does not always mean a steeper curve.

To illustrate these points we consider the following example coming from the data analyzed in Section 3.2. That is, we assume a model with the known parameters (estimated in Section 3.2) as given in Table 2 for which the sector boundaries considered are $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.1, 2.8, 4, 5.75)$. From these parameters it may appear that the slope on the first sector $(2.1, 2.8]$ ($\omega_1 = 1.658$) should be lower than in the fourth one $(5.75, 2.1]$ ($\omega_4 = 6.517$) while it is clear from the model graph at Figure 1 (bottom) that it is not so. From the same Figure, when compared with the third sector $(4, 5.75]$, it is also apparent that the slopes in the sectors are not constant. The ω_3 parameter of this third sector (65.711) is the highest of the four but there are some parts of the sector where the curve is not steep.

To give further detail on the relationship between the ν_i and the ω_i parameters, consider Figure 1 (top left); it gives the four curves from which the model is built. The blue curve corresponds to $\omega_1 = 1.658$, the red one to $\omega_2 = 0.066$, the grey one to $\omega_3 = 65.711$ and the green one to $\omega_4 = 6.517$. The ν_i parameters determine which part of each curve is used in the model. Depending on ν_i, the same ω value can translate into a more or less steep curve.

[Figure 1 about here.]

Figure 1 (top right) shows how these curves are combined in Figure 1 (top left) using the ν_i parameters, finally leading to the model graph as shown in the bottom of the Figure 1 (bottom). The solution appearing in Figure 1 (bottom) follows the thin blue line from 2.1 (where it intersects the green line for the second time) to 2.8 (where the thin blue line intersects the red line), then the red line from 2.8 to 4 (where the red line intersects the grey line), next the grey line from 4 to 5.75 (where the grey line intersect the green line) and finally the green line from 5.75 to 2.1.

It is also importance to note that, this behavior also happens with the circular-circular model of Downs and Mardia (2002) as the parameters are not uniquely determined. For example, if in the circular-circular model we consider $\mu' = \mu - \pi$, $\omega' = 1/\omega$ and $\nu' = \nu - \pi$

the model does not change. However, in the piecewise model it is not possible to make a transformation to ensure that $\omega_i \in [-1, 1]$ simultaneously for all $i = 1, \ldots, k$ as that would require more than one value for the single μ parameter.

2.3 *Estimation*

We now describe how to obtain the maximum likelihood estimates. Under our assumption that the response components are independent $M(\mu_{ij}, \kappa)$, the log-likelihood is given by

$$\max_{\kappa,\mu,\omega,\nu} \left[-N \ln I_0(\kappa) + \kappa \sum_{i=1}^{k} \sum_{j=1}^{n_i} \cos\left(\psi_{ij} - \mu - 2 \arctan\left(\omega_i \tan \frac{1}{2}(\theta_{ij} - \nu_i) \right) \right) \right] \tag{7}$$

This expression is to be maximized under the continuity condition (5) which leads to the following constrains on the parameters for the existence of a non-trivial continuous solution.

$$\begin{aligned} \tan\left(\frac{\theta_1^* - \nu_k}{2}\right) \prod_{i=1}^{k-1} \tan\left(\frac{\theta_{i+1}^* - \nu_i}{2}\right) &= \prod_{i=1}^{k} \tan\left(\frac{\theta_i^* - \nu_i}{2}\right) \\ \omega_i = \frac{\tan\left(\frac{\theta_{i+1}^* - \nu_{i+1}}{2}\right)}{\tan\left(\frac{\theta_{i+1}^* - \nu_i}{2}\right)} \omega_{i+1} &\text{ for } i = 1, \ldots, k-1. \end{aligned} \tag{8}$$

These constraints (8) are general for our piecewise circular regression model.

However two additional conditions have to be imposed for the cell-cycle application. One is the condition for monotonicity and another is of "synchronicity" which we now describe. The monotonicity condition ensures that the solution is an increasing function which leads to the following constrains:

$$\omega_i \geqslant 0 \text{ for } i = 1, \ldots, k. \tag{9}$$

By "synchronicity" condition, we mean that the response runs only one cycle as the explicative variable runs one cycle. For this purpose, we impose the constrains that the solution only crosses the 0 barrier once. Let

$$z_i = \nu_i + 2 \arctan\left(\frac{1}{\omega_i} \tan\left(\frac{-\mu}{2} \right) \right) \text{ for } i = 1, \ldots, k.$$

The z_i value is the possible zero of i^{th} piece of the function. It will be a zero of the global function if this value belongs to the appropriate interval. Thus, the constrains under the synchronicity condition can be written as

$$\sharp\left\{z_i : z_i \in \left(\theta_i^*, \theta_{i+1}^*\right]\right\} = 1, \tag{10}$$

with $\theta_{k+1}^* = \theta_1^*$. Hence, we need to optimize the log-likelihood (7) with the additional constrains given by (9) and (10).

Implementation. In order to compute the maximum likelihood estimates of our model, we rewrite the model as the circular-linear model

$$\tan\frac{1}{2}(\mu_{ij} - \mu) = \omega_1 X_1 + \ldots \omega_k X_k,$$

where the explanatory variables X_i are defined as $X_i = \tan\left(\frac{\theta_{ij}-\nu_i}{2}\right) I_{(\theta_i^* < \theta_{ij} \leqslant \theta_{i+1}^*]}$. Now, we can compute the maximum likelihood estimates of $\{\kappa, \mu, \omega_1, \nu_1, \ldots, \omega_k, \nu_k\}$ using the theory developed by Fisher and Lee (1992) for circular-linear models, and the R package of Agostinelli and Lund (2011); the optimization has to be performed under the constrains given by (8), (9) and (10).

2.4 *Model selection*

In applications of our piecewise model, we need to have a procedure to assess our model ($k > 2$) versus the Down and Mardia model ($k = 2$). The question of model selection has received a lot of attention in the literature in the past (starting with the well-known paper by Akaike (1973)), and also in recent years due, among other reasons, to the increasing complexity of modeling approaches. To our best knowledge, the question has not been specifically considered in the context of circular models. We introduce briefly the question in the next paragraphs, including some important references. A more complete discussion on the problem is out the aim of this paper.

A simple strategy for the selection of models is to measure how well each model fits the

data. A simple approach would be to use the model residuals that can be defined using the circular variance as

$$e_{ij} = 1 - \left(\cos\left(\psi_{ij} - \left(\hat{\mu} + 2\arctan\left(\hat{\omega}_i \tan\frac{1}{2}(\theta_{ij} - \hat{\nu}_i)\right)\right)\right)\right), \tag{11}$$

and define a circular distance criterion (CDC) as $CDC = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$.

Although the difference between the two CDC values may give some insight, it is not a good idea to decide between the models on this basis as a complex model (the piecewise model) will always have a lower CDC than a simpler one (the circular-circular model). If we try to correct this as it is usually done in model selection (for example, in the normal linear regression model), the number of parameters in the model, and the related concept of degrees of freedom (DF) and the AIC (Akaike Information Criterion) play a major role.

The value of AIC for a model M is defined as $AIC(M) = 2\ln(l(M)) - 2D$ where $l(M)$ is the model likelihood and D is a penalization factor, originally (Akaike (1973)) equal to p, the number of parameters in the model. The model with highest value of AIC is then selected. It also usual to use DF instead of p, DF coincides with p for simple models as the normal linear regression model and with the number of *free parameters* or the parameters of the final model, in other cases. However, this is not always so simple for more complex models like the lasso or shrinkage estimation (Kato (2009), Tibshirani and Taylor (2012)), or in the presence of mixed effects (Muller et al. (2013)). This is also the case in our context. On one hand, we have 3 parameters in the original circular-circular model (not including κ), also 3 are the free parameters or the DF of the model. On the other hand, in the piecewise model, we have $2k + 1$ parameters and k restrictions (8) which reduce the number of *free parameters*, or DF, initially, to $1 + k$. However, this number does not take into account the effect of the restrictions (9) and (10).

A lot of work has been done during last years in deriving measures of the complexity of models in complex cases, to be used, particularly, as a penalization factor in the AIC. One

of the proposed solutions is to use the concept of divergence, which is an estimate of the *effective degrees of freedom*. See, for example, Rueda (2013) or Hansen and Sokol (2014). In fact, for piecewise regression on the line, some authors have also used modified AIC or Bayesian Information Criteria (BIC) criteria for selecting the best model (Muggeo and Adelfio (2011); Malash and El-Khaiary (2011); Painting and Holwell (2013)), although these proposals obviously do not take into account the restrictions (10) or the manifold we are considering in this paper.

Other works dealing with model selection have used a related but different concept, the Generalized Degrees of Freedom (GDF), originally defined in Ye (1998) for normal models and also considered in different models by Gao and Fang (2011), Zhang et al. (2012) among others. The most interesting advantages of GDF are its applicability to complex modeling procedures and also its ease of calculation using a data perturbation approach. We propose to use the GDF approach to measure the model-complexity in the same way as defined in Ye (1998). The GDF are based on the sum of sensitivity of each fitted value to perturbation in the corresponding observed value. In fact, in a normal linear regression model with p parameters, if we denote as $\mathbf{Y}$ the N dimensional vector of observed response values and as $\hat{\mu}$ the corresponding vector of predicted values, Ye (1998) shows that $GDF = p = \sum_{i=1}^{N} \frac{\partial \hat{\mu}_i}{\partial y_i}$. where $\mathbf{Y} = (y_1, ..., y_N)$ and $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_N)$. The data perturbation and the GDF estimation procedure we use in this paper are similar to those one proposed in Ye (1998), taking into account that we are denoting as Ψ the response vector. These procedures are detailed in Web Appendix A. Once we finally compute the estimated generalized degrees of freedom $\widehat{GDF}$, we use the Generalized Akaike Criterion given by,

$$GAIC = AIC(M) = 2\ln(l(M)) - 2\widehat{GDF}$$

to choose among different models.

3. Applications

When a study or experiment is conducted/repeated by multiple labs then it is common to ask how reproducible the data and results are. With the advent of microarray technology, during the past decade multiple labs conducted cell-cycle experiments to compare phases of cell-cycle genes in the genome of various species. However, researchers have been concerned about the reproducibility of results across labs even within the same species. If the phase angles, within the same species, obtained from different labs or experiments poorly correlate with each other then it will be difficult to compare phase angles of cell-cycle genes across multiple species. There is a need for a methodology to assess relations among phase angles between a pair of experiments. However, with the exception of the geometric approach of Liu et al. (2004), to the best of our knowledge, there does not seem to exist a formal statistical procedure to make such diagnostics for angular data. Often biologists use visual displays such as the heatmaps (e.g. Peng et al. (2005)) when assessing similarities between experiments or studies. Such graphical tools ignore variability in the data and hence are not very satisfactory.

Using some recently published cell-cycle data, in the following we demonstrate that the methodology developed in this paper can be used to assess correlations between a pair of experiments. We consider two examples. In the first example we apply our methodology on a pair of experiments conducted in two different labs on the same species of yeast, namely, *S. cerevisiae*. In the second example we apply the methodology on data from two different species of yeast, namely, *S. cerevisiae* and *S. pombe*, conducted in two different labs.

3.1 *Within species between labs correlation of phase angles of cell-cycle genes*

In this example we considered phase angle estimates of 32 *S. cerevisiae* cell-cycle genes obtained from Spellman cdc experiment (Spellman et al., 1998) and from Pramilla38 experiment (Pramila et al., 2006). The phase angle data obtained from Spellman cdc experiment were taken to be regressors (θ) and those from Pramilla38 experiment were taken to be the

response variables (ψ). Using the information available in the cyclebase database, we placed the change points at $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (1.5, 2.8, 4, 5.75)$. The estimated phase angles, for the 32 cell-cycle genes, the sector they are placed in and a scatterplot of these data can be found in Web Table 1 and in Web Figure 1 respectively.

Using the proposed piecewise circular regression model we obtain the results summarized in Table 1. The fitted model is plotted in Figure 2 while the residuals (as in (11)) are in Web Figure 2. For comparison purposes, in each table and graph we also provide results using the circular-circular model (c-c model throughout this Section) by Downs and Mardia (2002).

[Table 1 about here.]

[Figure 2 about here.]

The circular correlation coefficient of Jammalamadaka and Sarma (1988) for these variables takes a value of 0.829 indicating good association among the two experiments. The value of the CDC measure for the piecewise model is 0.138 while for the c-c model is 0.148. The CDC has been reduced by 6.76%. To evaluate better if this improvement is big enough we consider the selection diagnostics we defined in subsection 2.4. The GDF values appearing in Table 2 are the mean values obtained averaging the results of several runs of the GDF algorithm with different values of the tuning parameter τ. The value of the generalized Akaike information criterion $GAIC$ for the piecewise model is 55.474 while for the c-c model is 55.977 so that there is no evidence of significant improvement due to the piecewise model.

If the role of the variables is reversed, i.e. the Pramilla38 experiment is taken as regressor and the Spellman cdc experiment is taken as response, and the appropriate change points are considered the same result is obtained. The $GAIC$ of the piecewise model is lower than that of the c-c model.

As the piecewise model does not yield a significant improvement in this case over the c-c model, we can infer that there is no need for different functional relationships in the

different cell phases. In this sense we may also say that there is congruence between these two experiments performed by different laboratories.

3.2 *Between species and between labs correlation of phase angles of cell-cycle genes*

In this example we considered phase angle estimates of 32 cell-cycle genes obtained from two different species *S.cerevisiae* and *S. pombe*. Furthermore, the data were obtained from two different labs. We used the phase angle estimates from Spellman cdc (Spellman et al. (1998)) on *S. cerevisiae* as the regressors (θ) and those from Oliva elut2 experiment (Oliva et al. (2005)) on *S. pombe* as the response variables (ψ). For this case, we placed the change points at $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.1, 2.8, 4, 5.75)$. There is only a small change in the first change point with respect to the previous example that is also congruent with the information available in the cyclebase database. The estimated phase angles for the 32 cell-cycle genes, the sector they are placed in and a scatterplot of the data can be found in Web Table 2 and in Web Figure 3 respectively.

Using the proposed piecewise circular regression model we obtain the results summarized in Table 2. The fitted model is plotted in Figure 3 and the residuals are plotted in Web Figure 4. As in the previous example, for comparison purposes, in each table and graph we also provide results using the c-c model by Downs and Mardia (2002).

[Table 2 about here.]

[Figure 3 about here.]

Unlike in the previous example, where the data were obtained from the same species, in this present example, consisting of data from different species, the circular correlation coefficient of Jammalamadaka and Sarma (1988) for these variables takes a value of -0.391 which indicates a poor circular-circular association between the regressors and the response variables so that we are not expected to find as good fits as in the previous example. The value

of the CDC measure for the piecewise model is 0.332 while for the c-c model is 0.394. The CDC has been reduced by 9.91%. As in the previous example the values of the GDF appearing in Table 2 are obtained averaging the results from several runs of the GDF algorithm with different values of the tuning parameter τ. The value of the generalized Akaike information criterion $GAIC$ for the piecewise model is 19.809 while for the c-c model is 19.412 so that there is some evidence of improvement due to the piecewise model in spite of its higher complexity. The difference between these two values is not as big as might be expected from the log-likelihood values (16.448 for the piecewise model and 13.553 for the c-c model, see Table 2) as the GDF are 6.544 and 3.847 for the for the piecewise and the c-c model respectively. Notice that our model selection proposal is somehow conservative as considering the usual DF values (5 for the piecewise model and 3 for c-c would give a bigger difference (22.896 vs 21.106) between the models. As this conservative approach still yields a difference in favor of the piecewise model we are reinforced in our conclusions. Moreover, reversing the role of the variables as we did in the previous example also yields the same result as the $GAIC$ for the piecewise model is higher than that of the c-c model.

Another application of our method is to estimate the duration of time the cells spend in various phases of cell-cycle. It is well-known among cell biologists that during the cell-division cycle, *S. cerevisiae* spends equal time in all phases (nearly 25% in each phase) whereas *S. pombe* spends a large proportion of time (according to some estimates nearly 70%) in the G2 phase. Interestingly, our method allows us to estimate these phase durations. More precisely, the images of the change points $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.1, 2.8, 4, 5.75)$ by the piecewise model are $(5.613, 6.248, 1.476, 4.762)$. Thus the lengths of the phases in the regressor species (*S. cerevisiae*) are $(0.7, 1.2, 1.75, 2.633)$, while using the piecewise model the estimated phase lengths for the response (*S. pombe*) are $(0.635, 1.511, 3.286, 0.851)$ so that there are differences in the 3rd and 4th sectors which correspond to cell phases G2 and

M respectively. Thus according to our estimates, *S. cerevisiae* spends 27.85% time spent in G2 phase whereas *S. pombe* spends 52.30% of its time in the G2 phase during the cell-cycle.

4. Discussion

Piecewise regression models have proved very useful in the case of data on the line to describe multi-linear relationships representing the different effects of the explicative variable on the response before and after some change points values on the explicative. Something similar is likely to occur in circular models. This paper is a first contribution to the study of these models in the circular setting. The model where the break points are assumed to be known have been described, estimated and applied to cell-cycle gene expression data.

It is to be noted the model we propose is not only interesting for cell biology applications. As seen from recent literature, related problems arise naturally in a wide range of fields. Some examples include circadian biology (Cermakian et al. (2011), Kondratova and Kondratov (2012)), metabolic cycle (Slavov et al. (2012)), evolutionary psychology (De Quadros-Wander and Stokes (2007)) or motor behavior (Baayen et al. (2012)).

Another application appears when dealing with woman's menstrual cycle. These cycle consists of three phases, namely, follicular, ovulation (single point) and luteal. Although the blood or urine concentration of hormones such as estrogen and progesterone are periodic during the approximate 28 day cycle, they have distinct patterns according to the phase of the cycle. For example, during the follicular phase, beginning with the menses, the concentration of estrogen sharply rises (almost like an exponential curve) and the blood estrogen levels sharply drop to the baseline at ovulation and then start to rise during the luteal phase and attain its peak in the middle of the luteal phase and begin to drop slowly towards the end of the monthly period. Humans are exposed to thousands of compounds and some of which are known to be estrogen like compounds such as, for example, bisphenol A (BPA), dichlorodiphenyltrichloroethane (DDT) or polychlorinated biphenyls (PCBs). Many

of these chemicals are often used in the manufacture of pesticides, flame retardants, etc. Effects of these compounds on the reproductive health (especially during puberty) is of great interest to researchers (see Kanno et al. (2003)). Among various concerns is the changes in the menstrual cycle among girls/women who are exposed to such endocrine disrupting estrogen like compounds. The piecewise regression model developed in this paper will be useful for characterizing patterns of various hormones during the menstrual cycle and use these characterizations to compare women in different chemical exposure groups.

From the methodological point of view, there are some extensions of the proposed model that can be dealt with not only in the model formulation but also in the estimation of the parameters and in the model selection issues.

One of these extensions is the inclusion of other explanatory variables in the model. In the linear piecewise regression literature this is done assuming that the breakpoints depend on only one of the explanatory variables or on the time the observations are taken (see, for example, Liu et al. (1997); Qu and Perron (2007)). In this way, it is easy to include other circular variables $\theta_2, \ldots, \theta_s$ and/or linear ones $Z_1, \ldots, Z_t$ in the model, replacing formulation (6) by

$$\mu_{ij} = \mu + 2 \arctan \left(\sum_{l=1}^{s} \omega_{il} \tan \frac{1}{2}(\theta_{ijl} - \nu_{il}) + \sum_{m=1}^{t} \beta_{im} z_{ijm} \right),$$

where β_{im} is the slope of the linear variable Z_m in sector i and replacing also the continuity conditions (5) by the corresponding ones. Notice that, since our estimation scheme relies on the circular-linear model from Fisher and Lee (1992) the estimation of the parameters can be performed when only one circular explanatory variable is present although restrictions on the parameters may be more involved. The same holds for model selection. Our $GAIC$ criterion and its computation does not depend on how many variables are in the model, or if they are circular or linear, thus making easy the task of variable selection.

Another extension that can be solved easily is that of dropping the known θ_i^* assumption

and estimating the phase borders as unknown parameters, or even the assumption of an unknown number of sectors. As with the previous extension no change on the estimation or model selection procedure is needed. However, it is obvious that the computational burden of both parameter estimation and model evaluation will be highly increased.

Acknowledgements

This work was supported by Spanish Ministerio de Ciencia e Innovación grant (MTM2012-37129 to S.B., C.R. and M.A.F.) and Junta de Castilla y León, Consejería de Educación and the European Social Fund within the (Programa Operativo Castilla y León 2007-2013 to S.B.) and the Intramural Research Program of the National Institute of Environmental Health Sciences (Z01 ES101744-04 to S.D.P.).

Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2.4, 3.1 and 3.2 are available with this paper at the Biometrics website on Wiley Online Library.

References

Agostinelli, C. and Lund, U. (2011). *circular: Circular Statistics*. R package version 0.4-3.

Akaike, H. (1973). Information theory and the maximum likelihood principle. In *International Symposium on Information Theory*. Akademiai Kiádo.

Baayen, C., Klugkist, I., and Mechsner, F. (2012). A test for the analysis of order constrained hypotheses for circular data. *Journal of Motor Behavior* **44,** 351–363.

Cermakian, N., Lamont, E., Bourdeau, P., and Boivin, D. (2011). Circadian clock gene expression in brain regions of alzheimer's disease patients and control subjects. *Journal of Biological Rhythms* **26,** 160–170.

De Quadros-Wander, S. and Stokes, M. (2007). The effect of mood on opposite-sex judgments of males' commitment and females' sexual content. *Evolutionary Psychology* **4,** 453–475.

Di Marzio, M., Panzera, A., and Taylor, C. (2013). Non-parametric regression for circular responses. *Scandinavian Journal of Statistics* **40,** 238 – 255.

Downs, T. and Mardia, K. (2002). Circular regression. *Biometrika* **89,** 683–697.

Fernández, M., Rueda, C., and Peddada, S. (2012). Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research* **40,** 2823–2832.

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

Fisher, N. and Lee, A. (1992). Regression models for an angular response. *Biometrics* **48,** 665–677.

Gao, X. and Fang, Y. (2011). A note on the generalized degrees of freedom under L1 loss function. *Journal of Statistical Planning and Inference* **141,** 677–686.

Hansen, N. and Sokol, A. (2014). Degrees of freedom for nonlinear least squares estimation. *Preprint* .

Jammalamadaka, S. and Sarma, Y. (1988). A correlation coefficient for angular variables. *Statistical Theory and Data Analysis II* pages 349–364.

Jensen, J., Jensen, T., Lichtenberg, U., Brunak, S., and Bork, P. (2006). Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443,** 594–597.

Kanno, J., Onyon, L., Peddada, S., Ashby, J., Jacob, E., and Owens, W. (2003). The OECD program to validate the rat uterotrophic bioassay. Phase 2: dose response studies. *Journal of Environmental Health Perspectives* **111,** 1530–1549.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis* **100,** 1138–1352.

Kato, S. and Jones, M. (2010). A family of distributions on the circle with links to, and applications arising from, möbius transformation. *Journal of the American Statistical Association* **105,** 249–262.

Kato, S., Shimizu, K., and Shieh, G. (2008). A circular-circular regression model. *Statistica Sinica* **18,** 633–645.

Kondratova, A. and Kondratov, R. (2012). The circadian clock and pathology of the ageing brain. *Nature Reviews Neuroscience* **13,** 325–335.

Liu, D., Weinberg, C., and Peddada, S. (2004). A geometric approach to determine association and coherence of the activation times of cell-cycling genes under differing experimental conditions. *Bioinformatics* **20,** 2521–2528.

Liu, J., Wu, S., and Zidek, J. (1997). On segmented multivariate regression. *Statistica Sinica* **7,** 497–525.

Lund, U. (1999). Least circular distance regression for directional data. *Journal of Applied Statistics* **26,** 723–733.

Malash, G. and El-Khaiary, M. (2011). Piecewise linear regression: A statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models. *Chemical Engineering Journal* **163,** 256–263.

Mardia, K. and Jupp, P. (2000). *Directional Statistics*. John Wiley & Sons.

Motulsky, H. and Christopoulos, A. (2004). *Fitting Models to Biological Data Using Linear and Nonlinear Regression*. Oxford University Press.

Muggeo, V. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* **27,** 161–166.

Muller, S., Scealy, J., and Welsh, A. (2013). Model selection in linear mixed models. *Statistical Science* **28,** 135–167.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B., and Leatherwood, J. (2005). The cell-cycle-regulated genes of schizosaccharomyces pombe. *PLoS Biology* **3,** 1239–1260.

Oliveira, M., Crujeiras, R., and Rodríguez-Casal, A. (2014). Npcirc: An R package for

nonparametric circular methods. *Journal of Statistical Software* **61,** 1 – 26.

Painting, C. and Holwell, G. (2013). Exaggerated trait allometry, compensation and trade-offs in the New Zealand giraffe weevil (Lasiorhynchus barbicornis). *PLoS ONE* **8,** e82467.

Peng, X., Karuturi, R., Miller, L., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L., Liu, E., Balasubramanian, M., and Liu, J. (2005). Identification of cell cycle-regulated genes in fission yeast. *The American Society for Cell Biology* **16,** 1026–1042.

Pramila, T., Wu, W., Miles, S., Noble, W., and Breeden, L. L. (2006). The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development* **22,** 2266–2278.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica* **75,** 459–502.

Rueda, C. (2013). Degrees of freedom and model selection in semiparametric additive monotone regression. *Journal of Multivariate Analysis* **117,** 88–99.

Rueda, C., Fernández, M., and Peddada, S. (2009). Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *Journal of the American Statistical Association* **104,** 338–347.

Santos, A., Wernersson, R., and Jensen, L. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research* **43,** D1140–D1144.

Sarma, Y. and Jammalamadaka, S. (1993). Circular regression. In *Statistical Science and Data Analysis. Proceedings of the Third Pacific Area Statistical Conference*, pages 109–128, Utrecht, Netherlands. VPS.

Seber, G. and Wild, C. (1989). *Nonlinear Regression.* John Wiley and Sons.

Slavov, N., Airoldi, E., van Oudenaarden, A., and Botstein, D. (2012). A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes.

Molecular Biology of the Cell **23,** 1986–1997.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9,** 3273–3297.

Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40,** 1198–1232.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93,** 120 – 131.

Zhang, B., Shen, X., and Mumford, S. (2012). Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis* **56,** 574 – 586.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

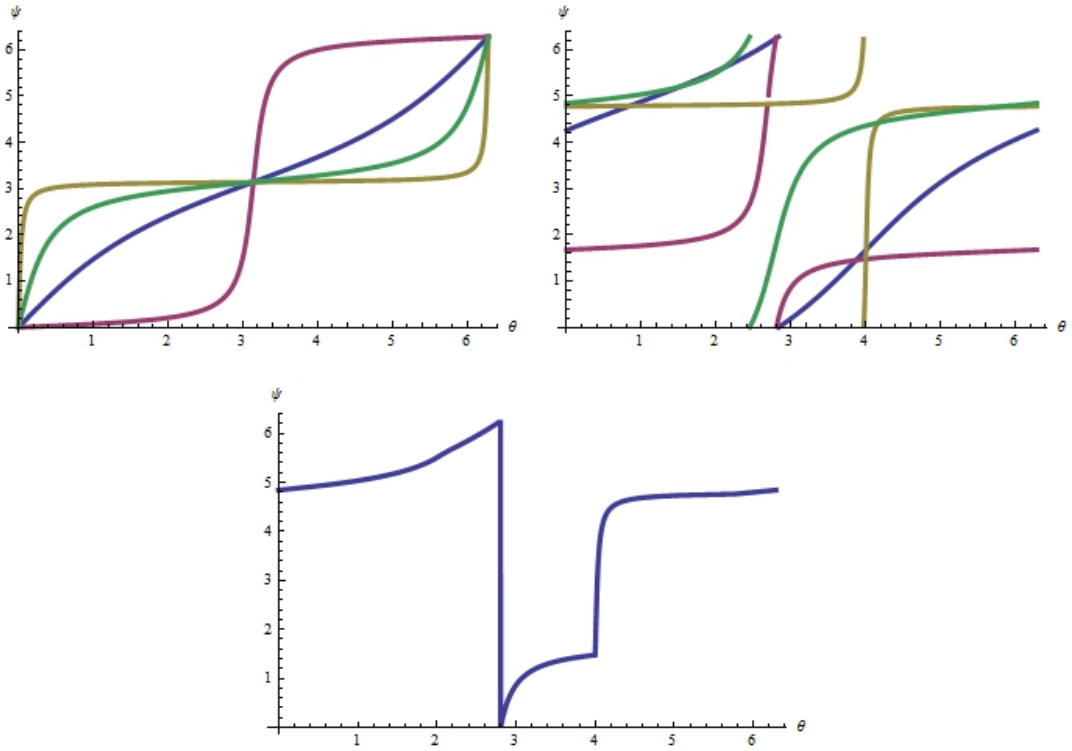

Figure 1. Top left: curves used in the piecewise model; Top right: curves used in the piecewise model shifted to their actual location; Bottom: final regression curve.

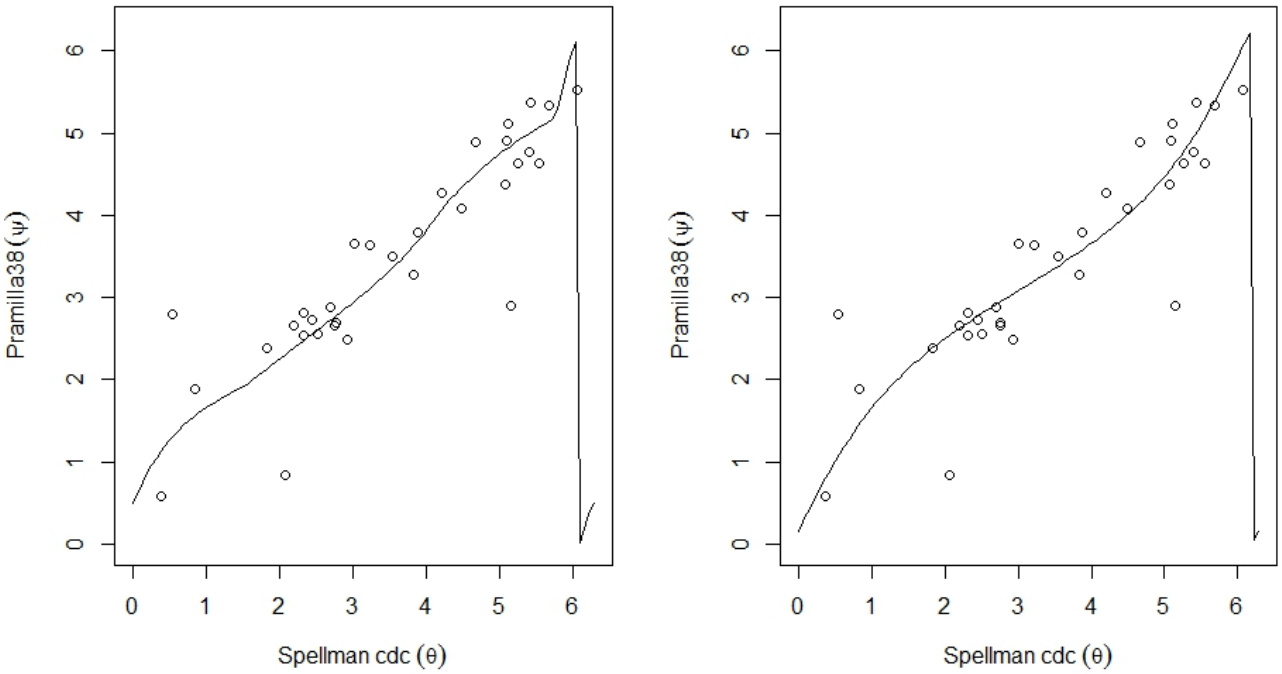

Figure 2. Fit of the estimated models for the *S. Cerevisiae* data (piecewise at left and c-c at right).

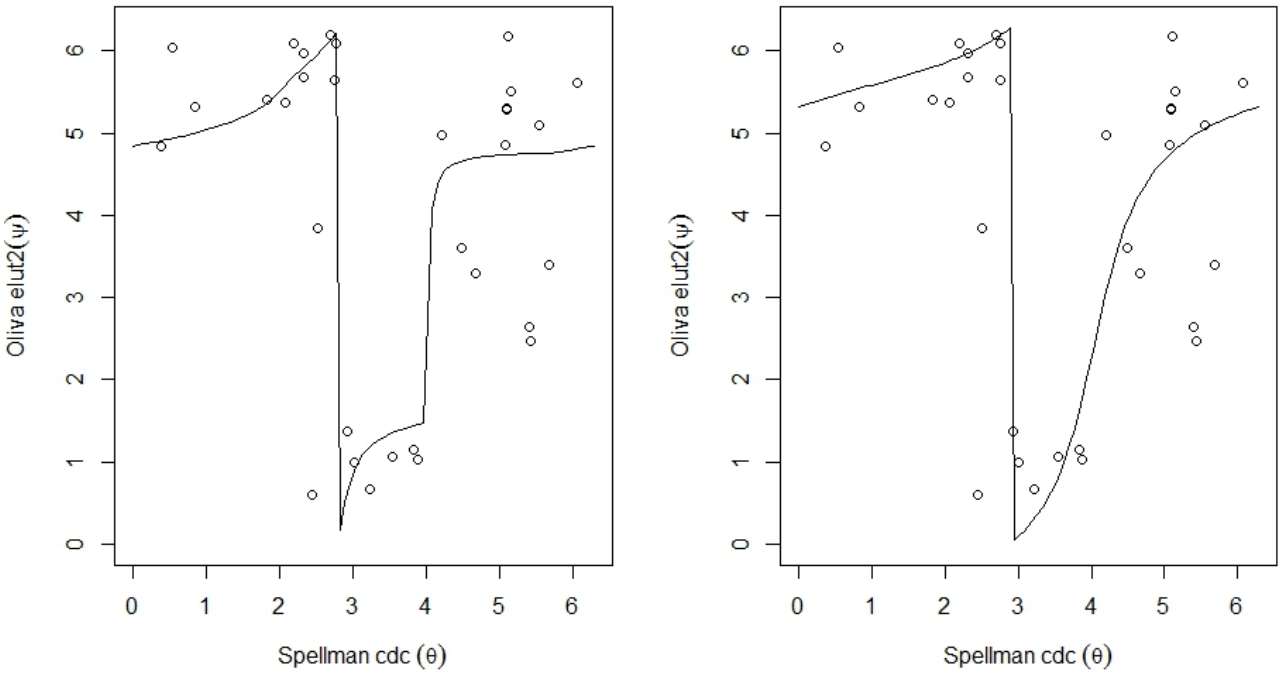

Figure 3. Fit of the estimated models for the two species data (piecewise at left and c-c at right)

Table 1
Maximum likelihood estimates and diagnostics under the piecewise regression model and the c-c model for the S. Cerevisiae data

Parameters	Estimated values (piecewise)	Estimated values (c-c)
μ	-0.840	3.099
ω	(1.491, 1.664, 0.485, 3.416)	0.546
ν	(5.214, 5.138, 0.003, 5.828)	3.032
κ	3.734	3.511
CDC	0.138	0.148
$\ln(l(m))$	32.723	31.421
GDF	4.986	3.432
$GAIC$	55.474	55.977

Table 2
Parameter estimations and diagnostics obtained using the piecewise regression model and the c-c model for the two species data

Parameters	Estimated values (piecewise)	Estimated values (c-c)
μ	1.646	-0.725
ω	(1.658, 0.066, 65.711, 6.517)	0.244
ν	(3.986, 5.824, 4.003, 2.774)	0.897
κ	1.822	1.592
CDC	0.332	0.394
$\ln(l(M))$	16.448	13.553
GDF	6.544	3.847
$GAIC$	19.809	19.412