UNIVERSIDAD DE VALLADOLID


TRABAJO FIN DE MÁSTER

# A new method for detection of cyclic circadian genes using order restricted inference

Yolanda Larriba González

2015

**Acknowledgements:**

I owe my deepest gratitude to my supervisors Cristina Rueda Sabater and Miguel Alejandro Fernández Temprano whose guidance and help have made this TFM possible. In the same line, I would like to give special thanks to the Department of Statistic and Operational Research of the University of Valladolid. Additionally, I gratefully acknowledge the support of my family and friends for their patience and encouragement.

*Valladolid, July 1, 2015*

**Abstract**

Identification of periodic patterns in gene expression data is important for studying the regulation mechanism of the circadian system. The information available is often given only by one or two cycles. Consequently, the number of observations is not enough to fit certain models, such as Fourier's models, properly. Some authors have already developed procedures or algorithms among which the JTK CYCLE algorithm is the most popular one.

We propose a new method to identify cyclic gene expressions based on euclidean and circular order restricted inference. Validation of the method is made through real data sets and simulations. Moreover, we compare the results obtained by the method with other detecting methods developed in the literature.

**Key words:** Circadian Cycle, Circular Data, Order Restricted Inference.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The study of biological rhythms is receiving a lot of attention in the Biological literature in recent years. At the core of research on biological rhythms lies the methodological problem of how to detect periodicities in measured data. This is reflected in the richness of the literature on this subject as well as in the wealth of methods and algorithms devoted to this task. The main purpose of this work is the development of new methodology and algorithms for the detection of periodic signals based on order restricted statistical methods.

Night and day, or dark and light patterns impact on human health in many different ways. It is well documented in the literature that in the US people are having, on average, less sleep during the night and this disruption or reduction in sleep is associated with numerous health outcomes including obesity. Among teenagers this may affect the production of their growth hormones and result in abnormal growth patterns.

For these reasons, researchers are in studying the effect of sleep on the circadian clock in human body during various stages of life. Important component of this clock are the circadian genes which have periodic expression overtime with phases suitably matching the night and day. It is important to recognize that the expression of circadian clock genes is tissue specific. Thus there may be differences in the phases as well as periods of these clock genes depending upon the tissue. There could be a

potential lag in peak expression, i.e. moment where the gene expression reaches its maximum, of the same gene in two different tissues depending upon its function.

Consequently, the identification of circadian clock genes in various tissues, such as heart, liver, etc, and the estimation of the lag in peak expression between tissues, so that the sequence of events can be correctly understood, are problems of considerable interest for biologists. Note that since circadian clock genes follow the night and day cycle, the genes involved in circadian clock have a periodic expression. Therefore, the problem of interest to a biologist 'reduces' to identifying genes that have a periodic expression and to comparing the phase angles (and other parameters) of circadian genes in two or more tissues.

Notice that these problems are not exclusive of circadian clock as similar ones also appear in other biological areas as, for example, cell-cycle research, where the role of all genes that express periodically in the cycle of cells is studied. The methodology developed here will obviously be useful for a wide kind of problems in different areas of research.

In this work, we address the problem of identification of periodic genes proposing novelty approaches. This relevant biological problem will be tackled in a parallel way within the Euclidean and the Circular spaces (i.e., this work involves dealing with circular data). The biological problem will be translated by formulations of several hypotheses testing problem and will be solved, in each of the spaces, using Order Restricted Inference (from now on ORI) techniques. ORI is a specific statistical methodology that allows us to incorporate *a priori* information in the model.

Due to the cyclic nature of the problems considered, one of the main features of the data to be analyzed is that they, by using an appropriate transformation, can be represented as points in the circle. Circular data have to be treated with care as the most simple statistics as, for example, the circular mean have to be adapted to the geometry of the space. In Chapter 2 we collect the basic concepts and definitions of circular data that will be needed to understand the document. A full account can

be found in, for example Mardia and Jupp (2000).

Other important aspect in this work is the order within the circadian gene expressions. To address this question we use techniques of ORI, see (Robertson et al. (1988), Silvapulle and Sen (2005)). The order of the gene expressions is represented by restrictions on the parameters of the models.

It might be said that the statistical methodology developed for the analysis of gene expression in the cell cycle has leaded to a new field in Statistics, which may be called, inference with constraints in circular models. Pioneering works in this field are Rueda et al. (2009) or Fernández et al. (2012). The first one defines and proposes an estimator for the circular isotonic regression (CIRE), and the second one deals with an hypothesis testing problem to test a given circular order using a conditional test. Some notes about these results appear in this work, see Section 2.2.1. Following this line, the previous works were used in Barragán (2014) to determine the order of activation of a set of genes in a species, and if this order changes with the evolution of the species or not. Recently, Barragán et al. (2015) deals with the problem of analysing gene expressions in two correlated oscillatory systems (peak expressions of periodic genes from different dose levels, species, organs,...). In Militino et al. (2015), the aim is the development of statistical tools to check if the unimodality pattern persists in some diseases like breast cancer, in different regions of developed countries using order restricted inference.

Identification of periodic genes among several thousands of genes using microarray data is not a simple problem due to the large variability in the data and the high number of data (several thousands of genes) to be processed. There are numerous ad-hoc methods available in the literature that are not entirely satisfactory. The most common ones are JTK Cycle algorithm (Hughes et al. (2010)) and RAIN, (Thaben and Westermark (2014)). Both these two methods are non-parametric methods. This work proposes new parametric statistical methodology based on circular data and aimed to identify periodic genes within the circadian cycle.

The techniques developed in the previously mentioned works can also help to

know, in the circadian cycle, about the biological clock if a tissue and of this biological clock (or order of activation of the genes) is conserved among tissues. These moments of activations can be identified as a part of a cyclic process, and as a consequence it is possible to understand them as circular data, establishing a connection between the euclidean and the circular data spaces.

However, as we have already mentioned, the circular data geometry does not allow to employ the usual statistical methods or developments. In addition to this, in many of the available data sets the number of available data is not high enough to properly fit mathematical models such as Fourier's models.

Therefore, the aim of this work consists in developing a new parametric statistical methodology and software to adapt the analysis of the circular data considering the fact that the number of observations available to determine if a gene expression is or not cyclic is low, using the previous methods and results obtained in constrained statistical inference applied to circular data, (e.g. Rueda et al. (2009) or Fernández et al. (2012)) as well as the software developed up to date Barragán et al. (2013). The layout of this work is the following:

1. The main ORI definitions and results within the Circular space, and a brief outline about one of the most employed algorithms in the literature to identify periodic genes, the JTK algorithm, can be found in Chapter 2.

2. A brief introduction to explain the notation employed and the basic definitions to formulate mathematically the problem of detecting cyclic genes is given in Section 3.1 of Chapter 3.

3. Next, we formulate, in a parallel way, Euclidean and Circular models using nested hypotheses testing problems, and we also design an algorithm to identify different periodic genes patterns for both spaces, see Section 3.2 of Chapter 3.

4. The last part of this work contains simulations and a real example to validate new methodology proposed, see Chapter 4

5. Finally, in the Chapter 5, we expose the main conclusions and the future work.

# Chapter 2

# Background

The theoretical bases of the algorithm proposed in Chapter 3 to detect cyclic genes, belong to the fields of order restricted inference and circular data. The literature related with both ORI and circular data is extremely widespread. In this brief review we start by giving the basic concepts for circular data and see how the ORI techniques have been incorporated so far for the analysis of these data, since ORI, has mainly been developed in the Euclidean space, being the works of Rueda et al. (2009) and Fernández et al. (2012) the first ones to incorporate constrains in the circular data analysis. Both researches arise to solve problems related to the analysis of gene expressions in the cell cycle.

To understand the ORI methodology for circular data, we must first define the main statistics and distributions (see Section 2.1) as well as circular orders (see Section 2.2) within the Circular space. Moreover, we need to study concepts such as CIRE (see Section 2.2.1) which allow making restricted inference on the von Mises model, (see Section 2.2.2). The last part of the Chapter describes one of the most widespread methods of detecting cyclic genes in the literature, the JTK CYCLE algorithm.

## 2.1 Basics on Circular Data Analysis

This Subsection introduces the basic concepts needed to understand this work. Circular data is a complex field, the key references being Mardia and Jupp (2000) and Fisher (1993). These are the basic concepts.

**Definition 2.1.1.** *Circular Data*

*A circular data is a point in the unit circle. In an equivalent way, it is a direction vector in the plane. When an initial direction and an orientation are fixed, it can be represented by the angle between the initial direction and the observed point.*

We consider the counter clockwise direction, that is also the most widely used in the literature. Circular data can be classified with respect to its precedence: compass data or clock data. In the last case, circular data depend on the time, where the circle represents the cycles which are going to be repeated again and again. Whatever was the origin of the data the features and tools are the same in both analyses.

We continue describing the most relevant measures of location, concentration, distance and association, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ a vector of $n$ observations in the circle.

### 2.1.1 Measures of Location and Concentration

**Definition 2.1.2.** *Circular or Directional Mean*

$$\overline{\theta} = Ave(\boldsymbol{\theta}) = \begin{cases} \arctan(\frac{\overline{S}}{\overline{C}}) & \text{if} \quad \overline{S} \geq 0, \overline{C} \geq 0 \\ \frac{\pi}{2} & \text{if} \quad \overline{S} > 0, \overline{C} = 0 \\ \arctan(\frac{\overline{S}}{\overline{C}}) + \pi & \text{if} \quad \overline{C} < 0 \\ \arctan(\frac{\overline{S}}{\overline{C}}) + 2\pi & \text{if} \quad \overline{S} < 0, \overline{C} \geq 0 \end{cases}$$

*where,*

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} \sin(\theta_i), \qquad \overline{C} = \frac{1}{n}\sum_{i=1}^{n} \cos(\theta_i).$$

*Note that if $\overline{S} = \overline{C} = 0$, then $\overline{\theta}$ is not defined.*

**Observation 2.1.1.** *The circular mean $\overline{\theta}$ does not verify the Cauchy mean value property, which is a key property in some results in the Euclidean space, see Example 2.1.1.*

**Example 2.1.1.** *Cauchy Mean Value Property*

*Suppose two birds are flying east at angles $\theta_1 = 0.52$ radians $(30^o)$ and $\theta_2 = 5.76$ radians $(330^o)$, respectively. Then, the arithmetic mean is $\overline{\theta} = \pi$ radians (i.e. $180^o$), suggesting that the birds on the average are actually flying westward, which refutes common sense. Instead angular mean is 0 radians (i.e., $0^o$). Moreover, note that 0 radians does not lie between 0.52 and 5.76, and hence, the angular mean does not satisfy the Cauchy mean value property.*

**Definition 2.1.3.** *Mean Resultant Length, MRL*

*The mean resultant length, MRL, is the most common measure of concentration for circular data, it is defined as:*

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i - \overline{\theta}) = \sqrt{\overline{S}^2 + \overline{C}^2}.$$

*It is a measure of the length of the mean direction defined before for $\boldsymbol{\theta}$.*

**Observation 2.1.2.** *Note that $0 \leq \overline{R} \leq 1$, therefore if $\theta_1, \ldots, \theta_n$ are tightly clustered, then $\overline{R}$ will be almost 1. On the other hand, if $\theta_1, \ldots, \theta_n$ are widely dispersed then $\overline{R}$ will be almost 0.*

**Observation 2.1.3.** *From the MRL, we can derive the definition of the resultant length $R$, of the vector as $R = n\overline{R}$, which can be understood as a dispersion measure too.*

**Example 2.1.2.** *Circular Mean and MRL for roulette wheel*

*A roulette wheel was spun and the positions at which it stopped were measured. The stopping positions in 9 trials were 0.75 $(43^o)$, 0.79 $(45^o)$, 0.91 $(52^o)$, 1.06 $(61^o)$, 1.31 $(75^o)$, 1.54 $(88^o)$, 1.54 $(88^o)$, 4.87 $(279^o)$, 6.23 $(357^o)$ radians. The circular raw data plot in Figure 2.1 suggest that there is a preferred direction.*

*Moreover,* $\overline{C} = 0.447$ *and* $\overline{S} = 0.553$*, so the mean direction* $\overline{\theta} = 0.89$ *radian* $(51^o)$ *and* $\overline{R} = 0.711$*. Figure 2.1 shows* $\overline{\theta}$ *and* $\overline{R}$ *for this data set and indicates the preferred direction of* $51^o$*.*

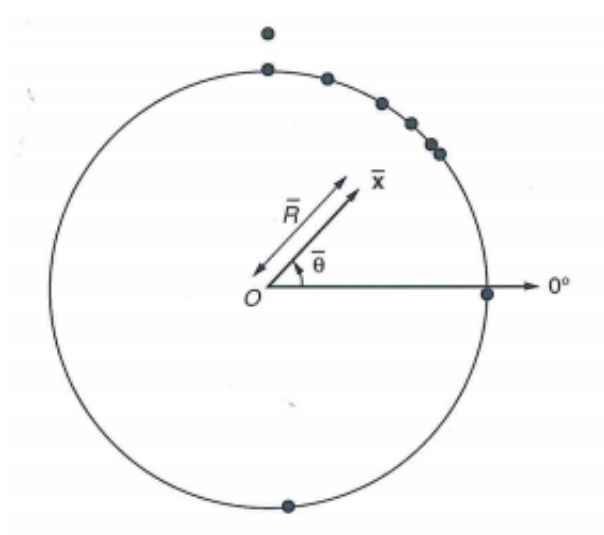



Figure 2.1: The mean direction $\overline{\theta}$, and the mean resultant $\overline{R}$, for the Example 2.1.2

### 2.1.2 Measures of Distance and Association

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ two vectors of circular observations. We define the distance between two vectors, and between a vector and a set of observations.

**Definition 2.1.4.** *Angular Distance between* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$

*The angular distance between* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *is given by:*

$$d(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} [1 - \cos(\alpha_i - \beta_i)]$$

**Definition 2.1.5.** *Sum of Circular Errors, SCE*

*The sum of circular errors between a vector* $\boldsymbol{\alpha}$ *and a vector of circular mean* $\overline{\boldsymbol{\theta}} = (\overline{\theta}_1, \ldots, \overline{\theta}_n)$*, where* $\overline{\theta}_i$ *is the circular mean of the vector* $(\theta_{i1}, \ldots, \theta_{ic})'$ *is defined*

*by:*

$$SCE(\overline{\boldsymbol{\theta}}, \boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{c}[1-\cos(\theta_{ij}-\alpha_i)]$$
$$= \sum_{i=1}^{n} R_i[1-cos(\overline{\theta}_i-\alpha_i)],$$

*where $R_i$ is the resultant length of the vector $(\theta_{i1},\ldots,\theta_{ic})$*

## 2.2 Basics on Order Restricted Inference

An intrinsic feature of circular data is the fact that, the smallest number of elements in a set of circular data to establish a circular association or order between them is three, while in the Euclidean space two elements have an order among them.

To be able to define an order in the circle is necessary to consider a third one element, so that the initial point in the circle has not influence into the order. So the circular order is defined at least for three elements, namely, $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_1$. The notation $\leq$ for the circular order is inherited of circular data can be understood as a cyclical process.

**Definition 2.2.1.** *Circular Order*

*We will say that a vector $\boldsymbol{\theta}$ follows a circular order $O$ where $O = \{o_1,\ldots,o_n\}$ is a permutation of $\{1,\ldots,n\}$ if $\theta_{o_1}$ precedes $\theta_{o_2}$ which precedes $\theta_{o_3}$ and so on with $\theta_{o_n}$ preceding $\theta_{o_1}$. We will denote this precedence relation as $\theta_{o_1} \leq \cdots \leq \theta_{o_n} \leq \theta_{o_1}$ and we will also write that $\boldsymbol{\theta} \in C_O$. With this notation, $C_O = \{\theta \in [0,2\pi)^{2n} : \theta_{o_1} \leq \cdots \leq \theta_{o_n} \leq \theta_{o_1}\}$ is said to be an order cone.*

These orders are invariants with respect to rotations and they are independents of initial direction choice, (see Mardia and Jupp (2000)).

### 2.2.1 Circular Isotonic Regression Estimator. CIRE

The Circular Isotonic Regression Estimator, from now on, CIRE, is the natural extension of Isotonic Regression in the Euclidean space into the Circular space. An outline about Isotonic Regression within the Euclidean space can be found in Appendix A. The CIRE and other relevant statistical results within the Circular space were studied in Rueda et al. (2009). Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ the CIRE of $\boldsymbol{\theta}$ with respect an order $O$ can be defined as the closest vector to $\boldsymbol{\theta}$ verifying the order $O$.

**Definition 2.2.2.** *Circular Isotonic Regression Estimator, CIRE*

*The Circular Isotonic Regression Estimator, CIRE, of* $\boldsymbol{\theta}$ *with respect to* $\boldsymbol{C_O}$ *is:*

$$\tilde{\boldsymbol{\theta}}^{(\boldsymbol{O})} = \arg\min_{\boldsymbol{\eta} \in C_O} SCE(\boldsymbol{\theta}, \boldsymbol{\eta}) \tag{2.1}$$

Moreover, $\tilde{\boldsymbol{\theta}}$ determines a partition $\mathcal{P} = \{1, \ldots, m\}$ into sets of coordinates on which $\tilde{\theta_j}$ is constant. These sets are called level sets. Both this property and the existence, (almost surely) uniqueness and others properties are proved in Rueda et al. (2009). It is not possible to obtain the CIRE by any well-known algorithm for constrained estimators in the Euclidean space, or by adapting them to the Circular space. When it is clear which cone we are reference to we will drop the super-index $(O)$.

As it is shown in Appendix A and according to Rueda et al. (2009), $\tilde{\boldsymbol{\theta}}^{(O)}$ is achieved through a specific algorithm for circular data based on the PAVA (Pool Adjacent Violator Algorithm) proposed in Robertson and Wright (1980) which solves the problem of isotonic regression for euclidean data with distinct order constrains. In order to that, adjacent observations which violate the order constrains are averaged in sets for which the restricted estimator takes the same value. These sets are also called, level sets. See Appendix A for details.

The CIRE implementation is available both for SAS and for R code. A freely downloadable SAS based user-friendly software can be obtained in `http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/peddada/index.cfm`.

The R package is included in the R-package **isocir** (**iso**tonic inference for **cir**cular data), Barragán et al. (2013).

### 2.2.2 Inferences on the von Mises Model. Conditional Test

In the same way that the Normal model is the most widespread within the Euclidean space, the von Mises model plays the central role in the Circular space. The model proposed in this work will assume, whenever was necessary, von Mises distribution. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ be a circular random vector of von Mises distribution, we denote $\theta_i \sim VM(\phi_i, \kappa), \quad i = 1, \ldots, n$ where $\phi_i$ and $\kappa$ are parameters of location and dispersion, respectively. A brief outline about von Mises distribution can be found in Appendix D.

The assumption of von Mises distribution is needed in most previous results obtained in circular ORI, for instance in Rueda et al. (2009) where under the assumption of von Mises distribution it was proved that if $\boldsymbol{\phi} \in C_O$ then, the CIRE provides its restricted maximum likelihood estimator (RMLE).In Appendix B we include a brief outline for RMLE on Normal models.

Moreover, to solve the hypotheses testing problems proposed in this work, we act in the same way that in Fernández et al. (2012) where a conditional procedure is used to test a fixed circular order. The use of conditional test is not new in ORI, see for example Bartholomew (1961), or Iverson and A. (1987). Compared with the standard LRT where weights depend on unknown parameters that are difficult to compute, the conditional test is computationally much simpler and it also benefits from an increase in power in interesting alternatives.

This issue has been discussed within both the Euclidean space (Wollan and Dykstra (1986), Robertson et al. (1988), Menéndez et al. (1991), or recently Militino et al. (2015)) and within the Circular space (Fernández et al. (2012),Barragán (2014)). In Appendix C we include the main guidelines to conduct simple conditional tests in both spaces. In particular, one of the conditional test proposed in this work within the Circular space is the first time appears in the literature.

## 2.3 JTK Algorithm

Most available methods for periodicity detection can be traced back to Fourier methods in some form (Halberg et al. (1967), Straume (2004), Wichert et al. (2004), Wijnen et al. (2005)). These methods generally assume an underlying rhythm in the form of one or more sine waves and the general assumption that the noise variance is both Gaussian distributed and independent of measurement magnitude. However sometimes it is not even close to reality for biological data. For these cases the literature offers non parametric statistical methods, being JTK CYCLE algorithm the non parametric method which has had the largest impact and has been widely adopted in the field, see Wu et al. (2014), Deckard et al. (2013), Li et al. (2015).

JTK CYCLE builds on the non parametric Jonckheere-Terpstra test (Jonckheere (1954), Terpstra (1952)), which detects monotonous trends in data consisting of a dependent variable (e.g., mRNA expression levels) and an independent variable (e.g., time). JTK CYCLE acts designing tests measurements in both rising and falling parts of the underlying rhythm pattern against each other, i.e., by default it assumes a perfectly symmetric wave form, where the falling part has the mirror-image shape of the rising part. In practice, JTK CYCLE algorithm allows the user to choose a set of periods , as well as, possible phases, in this way, the pvalues returned are false discovery rate, FDR adjusted pvalues. The JTK CYCLE's weakness is that it is not able to detect asymmetric shapes, for instance, the case of an initial increasing and a subsequent sharper decrease in the dependent variable.

# Chapter 3

# Methodology

This Chapter contains some previous definitions and notation. We also describe the models and methods developed to solve the problem of identifying cyclic signals.

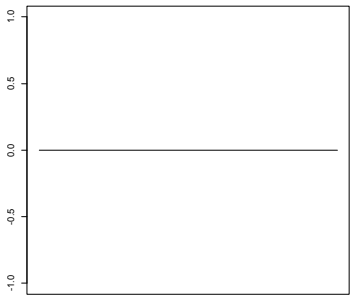
## 3.1 Basic Concepts and Notation

Let $T$ a fixed period, and let $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_c')$ be the vector of $n \times c$ expressions of a gene, where $c$ is the number of cycles of period $T$, $n$ is the number of observations per cycle and $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{ni})'$ is the vector containing the $n$ observations of the gene expression in the cycle $i$. The sequence of times $t_1, ..., t_n$ where the observations are measured is called *timepoint set* . Thus, $X_{ij}$ denotes the $i$th observation of the $j$th cycle of the gen, which is measured at instant $t_i$.

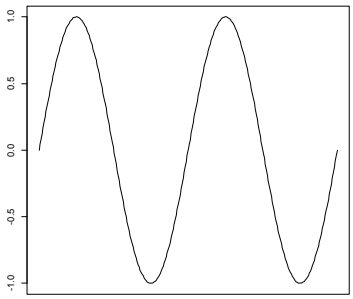
In many practical situations, the researcher is interested in the detection of periodic genes. For us, periodic genes will be those for which $E(\boldsymbol{X}_1) = \cdots = E(\boldsymbol{X}_c) = \boldsymbol{\mu}$ is a *cyclic signal*, (to be defined below).

The models and methods we propose in chapter 3.2 are designed to detect *cyclic signals*, $\boldsymbol{\mu}$, and to distinguish between the following 4 different patterns: Non cyclic & Non periodic; Non cyclic & Periodic; Cyclic & Periodic; Cyclic & Periodic & Constant. See Figure 3.1.

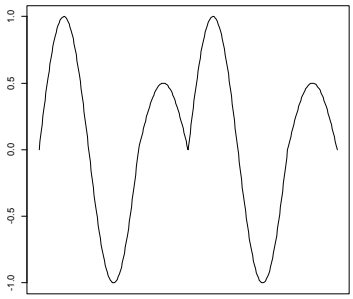
Giving a rigorous definition of *cyclic signal*, is a difficult task. According to



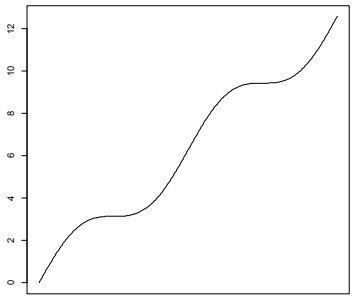

(a) Constant pattern




(b) Cyclic & Periodic pattern




(c) Non cyclic & Periodic pattern




(d) Non cyclic & Non periodic pattern

Figure 3.1: Four different cyclic signal patterns

biological interest, we assume that a *cyclic signal* is a periodic signal with one local maximum, called *phase* $P$, and one local minimum per cycle. A possible periodic gene with an underlying cyclic signal is shown in Figure 3.2.

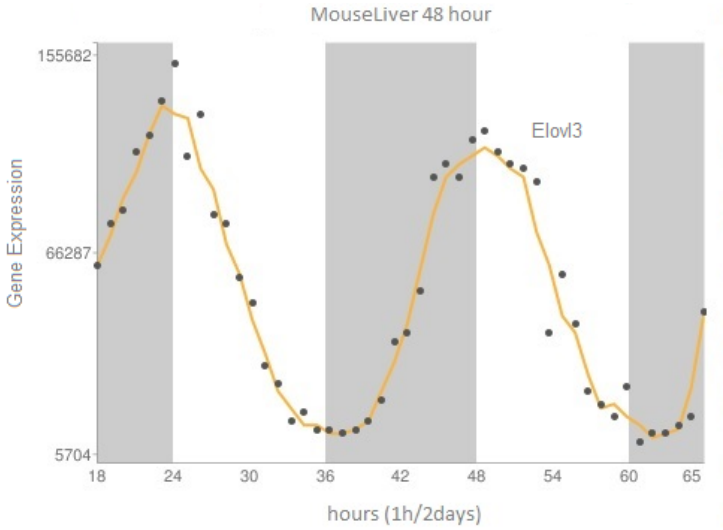



Figure 3.2: Periodic gene Elovl3 from Mouse Liver with a possible underlying cyclic signal.

According to this, a *cyclic signal*, $\boldsymbol{\mu}$, can be mathematically defined as follows:

**Definition 3.1.1.** *Cyclic Signal (I)*

$\boldsymbol{\mu}$ *cyclic signal* $\iff \exists\ \boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ *with* $\mu_i = \sin(\phi_i)$ *and* $\boldsymbol{\phi} \in C_O$ *where* $C_O = \{\boldsymbol{\phi} \in [0, 2\pi)^{2n} : \phi_1 \leq \cdots \leq \phi_n \leq \phi_1\}$ *is a circular order, (see Subsection 2.2.1).*

Taking into account this definition, a natural space to state the problem is the Circular Space, (see Figure 3.3). Therefore, we assume that $\exists\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ a vector of circular data, (see Section 2.1) such that $X = \sin(\boldsymbol{\theta})$ and $\boldsymbol{\phi} = E(\boldsymbol{\theta})$ (at least asymptotically).

**Definition 3.1.2.** *Cyclic Signal (II)*

$\boldsymbol{\mu}$ *cyclic signal* $\iff \boldsymbol{\mu} \in \mathcal{C} = \bigcup_{L,U} C_L^U$, *where* $L, U \in \{1, \ldots, n\}$ *and* $C_L^U = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_{Lj} \leq \mu_{L+1j} \leq \cdots \leq \mu_{Uj} \geq \mu_{U+1j} \geq \cdots \geq \mu_{L-1j} \geq \mu_{Lj}\}$.

**Observation 3.1.1.** *Note that the previous definitions involve both euclidean and circular parameters.*

**Proposition 3.1.1.** *The definitions (3.1.1) and (3.1.2) for a cyclic signal* $\boldsymbol{\mu}$ *are equivalent.*

***Proof:***

- *(3.1.1)* $\Rightarrow$ *(3.1.2)*

  *Given* $\boldsymbol{\phi} \in OC$*, define* $\boldsymbol{\mu} = \sin(\boldsymbol{\phi})$ *and* $L$ *and* $U$ *as the indexes such that* $\phi_L = arg\min_{i=1,\dots,n} \sin(\phi_i)$ *and* $U = arg\max_{i=1,\dots,n} \sin(\phi_i)$*. Since* $\boldsymbol{\phi} \in C_O$ *we have that* $\boldsymbol{\mu} \in C_{LU} \subset \mathcal{C}$*.*

- *(3.1.2)* $\Rightarrow$ *(3.1.1)*

  *Consider the indexes* $L$ *and* $U$ *such that* $\boldsymbol{\mu} \in C_{LU}$*, and denote*

  $$LU = \{i \in \{1, \dots, n\} : i \in [\min\{L, U\}, \max\{L, U\}]\}.$$

  *Then, for* $i = 1, \dots, n$*, define* $\phi_i$ *as*

  $$\phi_i = \begin{cases} \frac{\pi}{2} + \arcsin(\mu_i) & \text{if } i \in LU \\ \frac{3\pi}{2} - \arcsin(\mu_i) & \text{otherwise} \end{cases}$$

  *It is straightforward to check that* $\boldsymbol{\phi} \in C_O$*.*

In these terms, the goal of this work is identifying and distinguishing among periodic patterns from genes expressions, i.e., a gene expression will be periodic if the underlying function of this gene $\boldsymbol{\mu}$ is a *cyclic signal*. This problem turns into solving hypotheses testing problems within the Euclidean and the Circular spaces, as we explain in the following Sections.

## 3.2 Models and Methods

The equivalence between (3.1.1) and (3.1.2) allows to consider in a parallel way an Euclidean and a Circular model. The problem of identifying and distinguishing

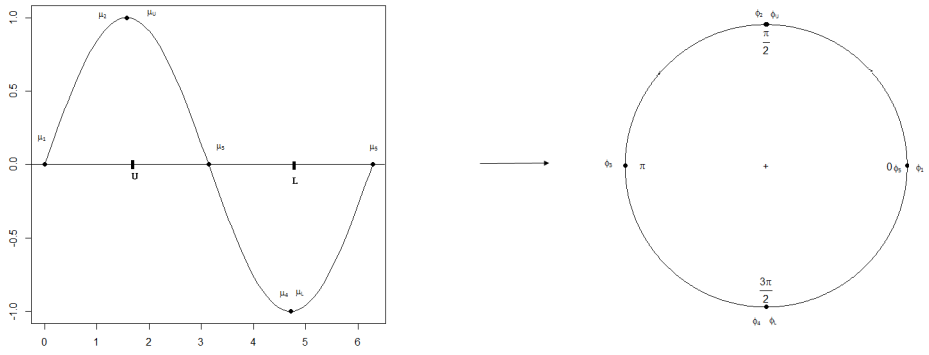


Figure 3.3: Relation between the Euclidean and Circular spaces

between the four different patterns cited before (see Figure 3.1) generates a nested testing hypothesis problem on $\boldsymbol{\mu}$ to be tested. To be more precise, for a cyclic signal $\boldsymbol{\mu}$, we will test the following four nested hypotheses:

$$
\begin{aligned}
H_0 &: \boldsymbol{\mu} \text{ Constant} \\
H_1 &: \boldsymbol{\mu} \text{ Cyclic \& Periodic} \\
H_2 &: \boldsymbol{\mu} \text{ Cyclic \& Non periodic} \\
H_3 &: \boldsymbol{\mu} \text{ Non cyclic \& Non periodic}
\end{aligned}
$$

A general procedure in both spaces is to conduct the following testing problems sequentially: $H_2$ against $H_3 - H_2$, $H_1$ against $H_2 - H_1$ and $H_0$ against $H_1 - H_0$. In this way, non periodic (therefore non cyclic) patterns are detected with the first test; periodic but non cyclic patterns are detected by accepting $H_2$ but not $H_1$; cyclic periodic patterns appear when $H_2$ and $H_1$ are not rejected but $H_0$ is rejected against $H_1 - H_0$; and finally, (periodic) constant patterns are concluded if none of the three null hypotheses are rejected.

The method we present for solving the hypotheses testing problems $H_1$ against $H_2 - H_1$ and $H_0$ against $H_1 - H_0$ is based on likelihood ratio test (LRT). However, the problem of testing $H_2$ against $H_3 - H_2$ which detects patterns such as (d) in

Figure 3.1, is harder to derive than in the other ones. Although we have studied some alternatives to the problem, it is only solved in a partial way. Therefore, this task will be dealt in the future work, see 5.3.

A direct use of conditional test based on LRT is an intractable problem in both spaces, therefore we propose a method to solve the problem following a 2-stage algorithm in both spaces:

- In a **first stage** we estimate $L$ and $U$ as follows:

$$\begin{aligned} \hat{L} &= arg \min_{i=1,\dots,n} \overline{X}_{i.} \\ \hat{U} &= arg \max_{i=1,\dots,n} \overline{X}_{i.}, \end{aligned} \tag{3.1}$$

  where $\overline{X}_{i.} = \frac{\sum_{j=1}^{2} X_{ij}}{2}$.

- In a **second stage** we consider the hypotheses testing problems under the assumption that $L$ and $U$ are known. In both spaces, we have to consider the three following steps:

  1. In a **first step** we reformulate the hypotheses of the testing problem to the estimations of $L$ and $U$.
  2. The **second step** is dedicated to the computation of the MLE of the model's parameters.
  3. In a **third step** the testing problems are solved using conditional test based on LRT in order to get a pvalue.

Finally, we use a pvalue adjustment to take into account the multiple testing and get the final results.

The first stage is direct and common in both spaces. The second stage needs to be specified in each space. Subsections 3.2.1 and 3.2.2 describe the model and the three mentioned steps to consider the testing problems within the Euclidean and the Circular spaces respectively. Throughout this Chapter, and without loss of generality we assume $c = 2$.

### 3.2.1 Euclidean Space Approach

From the first stage we assume that $L$ and $U$ are known, (see 3.1). Let us further assume that $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ independent, where $\mu_{ij}$ and $\sigma^2$ are parameters of location and dispersion respectively. The following algorithm describes the three steps of the second stage to consider the testing problems within the Euclidean space.

- **First Step:**

  Under these assumptions, the hypothesis testing problem is written as follows:

$$\begin{aligned} H_0^E: &\quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}^* \\ H_1^E: &\quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \in C_{LU}^E \\ H_2^E: &\quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \end{aligned}$$

  where $\boldsymbol{\mu}_j = (\mu_{1j}, \ldots, \mu_{nj})'$, $\boldsymbol{\mu}^* = \mu \cdot (1, \ldots, 1)$ and $C_{LU}^E$ is the order cone $C_{LU}^E = \{\mu_{iL} \leq \ldots \mu_{iU-1} \leq \mu_{iU} \geq \mu_{iU+1} \geq \cdots \geq \mu_{uL-1} \geq \mu_{iL}\}$. The order cone $C_{LU}^E$ is usually named *unimodality cone*. In this way, the hypotheses testing problem above involves order constraints.

- **Second Step**

  The log-likelihood for a general model is:

$$l\left((\boldsymbol{\mu_1}, \boldsymbol{\mu}_2), \sigma_i^2; (\boldsymbol{X}_1, \boldsymbol{X}_2)\right) \propto n \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} \omega_i \sum_{j=1}^{2} (X_{ij} - \mu_{ij}). \tag{3.2}$$

  where $\omega_i = \frac{2}{\sigma_i^2}$ and $\Sigma = diag(\frac{\sigma_1^2}{2}, \frac{\sigma_2^2}{2})$.

  From (3.2) we can derive the MLE of the parameters in the model. For $s = 0, 1, 2$, we denote by $\hat{\boldsymbol{\mu}}_s$ the MLE under $H_s^E$ for $\boldsymbol{\mu}$. Differentiating (3.2) with respect to $\sigma^2$ we obtain the MLE for parameter $\sigma^2$. Let $\hat{\sigma}_s^2$ denote the the MLE under hypothesis $s$, with $s = 0, 1, 2$.

Then, under $H_0^E$, $H_1^E$ and $H_2^E$, the MLEs of $\boldsymbol{\mu}$ and $\sigma^2$ can be written as follows:

**MLE under $H_0^E$:**

- According to Robertson et al. (1988), $\hat{\boldsymbol{\mu}}_0 = \hat{\mu}\cdot(1,\ldots,1)$, where $\hat{\mu}$ is written as:
$$\hat{\mu} = \frac{\sum_{i=1}^{n}\omega_i\sum_{j=1}^{2}X_{ij}}{\sum_{i=1}^{n}\omega_i}. \tag{3.3}$$
- According to Robertson et al. (1988), the estimator $\sigma^2$ can be written as:
$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ij}-\hat{\boldsymbol{\mu}}_0)^2}{2n}, \tag{3.4}$$
with $\hat{\boldsymbol{\mu}}_0 = \hat{\mu}\cdot(1,\ldots,1)$, and $\hat{\mu}$ given in (3.3).

**MLE under $H_1^E$:**

- The fact that $L$ and $U$ are fixed, let us calculate the MLE of $\boldsymbol{\mu}$ under $H_1^E$ using isotonic regression. An outline of isotonic regression can be found in Appendix A. Then, we can write:
$$\hat{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}^* \tag{3.5}$$
where $\boldsymbol{\mu}^*$ is the isotonic regression of $\overline{\boldsymbol{X}} = (\overline{X}_1,\ldots,\overline{X}_n)'$, being $\overline{X}_i = \overline{X}_{i\cdot}$ and with weight vector $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_n)'$ under the order which determines $C_{LU}^E$.
- According to Robertson et al. (1988), the corresponding estimator of $\sigma^2$ is:
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ij}-\hat{\boldsymbol{\mu}}_1)^2}{2n}, \tag{3.6}$$
with $\boldsymbol{\mu}_1$ given in (3.5).

**MLE under $H_2^E$:**

- Under $H_2$ the MLE of $\boldsymbol{\mu}$ can be written as:
$$\hat{\boldsymbol{\mu}}_2 = \overline{\boldsymbol{X}}, \tag{3.7}$$

see Robertson et al. (1988).

– Under $H_2^E$, the expression to estimate $\sigma^2$ is the most well-known:

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2}(X_{ij} - \hat{\boldsymbol{\mu}}_2)^2}{2n}, \tag{3.8}$$

with $\boldsymbol{\mu}_2$ given in (3.7), (see Robertson et al. (1988)).

- **Third Step:**

We assume $\sigma^2$ is known and the MLE of $\sigma^2$ obtained in the second step as the real value. The conditional tests are based on likelihood ratio test statistics (LRT) that are defined below. We denote by $LRT_{12}^E$ and $LRT_{01}^E$ the likelihood ratio test statistics for testing $H_1^E$ against $H_2^E - H_1^E$ and $H_0^E$ against $H_1^E - H_0^E$ respectively, they can be written as:

$$\begin{aligned} T_{12}^E = LRT_{12}^E &= -2l_{H_1^E}\left((\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \sigma^2; (\boldsymbol{X}_1, \boldsymbol{X}_2)\right) + 2l_{H_2^E}\left((\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \sigma^2; (\boldsymbol{X}_1, \boldsymbol{X}_2)\right) = \\ &= \sum_{i=1}^{n}\omega_i(\overline{X}_i - \hat{\boldsymbol{\mu}}_1)^2 - \sum_{i=1}^{n}\omega_i(\overline{X}_i - \hat{\boldsymbol{\mu}}_2)^2 \\ &= \sum_{i=1}^{n}\omega_i(\overline{X}_i - \boldsymbol{\mu}^*)^2, \end{aligned} \tag{3.9}$$

where $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)'$.

$$\begin{aligned} T_{01}^E = LRT_{01}^E &= -2l_{H_0^E}\left((\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \sigma^2; (\boldsymbol{X}_1, \boldsymbol{X}_2)\right) + 2l_{H_1^E}\left((\boldsymbol{\mu}_1, \boldsymbol{\mu}_2), \sigma^2; (\boldsymbol{X}_1, \boldsymbol{X}_2)\right) = \\ &= \sum_{i=1}^{n}\omega_i(\overline{X}_i - \hat{\boldsymbol{\mu}}_0)^2 - \sum_{i=1}^{n}\omega_i(\overline{X}_i - \hat{\boldsymbol{\mu}}_1)^2 \\ &= \sum_{i=1}^{n}\omega_i(\boldsymbol{\mu}^* - \hat{\mu})^2 + 2\sum_{i=1}^{n}\omega_i(\overline{X}_i - \boldsymbol{\mu}^*)(\boldsymbol{\mu}^* - \hat{\mu}). \end{aligned} \tag{3.10}$$

Applying Theorem 1.3.6 of Robertson et al. (1988) the last term in (3.10) is

seen to be zero. Hence the LRT under $H_0^E$ is:

$$T_{01}^E = LRT_{01}^E = \sum_{i=1}^{n} \omega_i (\boldsymbol{\mu}^* - \hat{\mu})^2. \tag{3.11}$$

where $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)'$.

The conditional test is described in detail in Appendix C. Let us denote as $t_{12}^E$ and $t_{01}^E$ the observed values of the statistics $T_{12}^E$ and $T_{01}^E$ respectively. Let us further denote by $m^E$ the computing level sets of $\boldsymbol{\mu}^*$ under $H_1^E$. Then the corresponding pvalues can be obtained as follows:

$$p12E = P[\chi^2_{n-m^E} \geq t_{12}^E] \tag{3.12}$$

$$p01E = P[\chi^2_{m^E-1} \geq t_{01}^E], \tag{3.13}$$

where p12E and p01E denotes the pvalue of testing $H_1^E$ against $H_2^E - H_1^E$ and $H_0^E$ against $H_1^E - H_0^E$ respectively within the Euclidean space.

### 3.2.2 Circular Space Approach

From the first stage we assume that $L$ and $U$ are known, (see 3.1). Let us further assume that $\theta_{ij} \sim VM(\phi_{ij}, \kappa)$ independent, where $\phi_{ij}$ and $\kappa$ are parameters of location and dispersion respectively.

To define $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{nj})$ we need a previous normalization of the data, i.e. of $\boldsymbol{X}$, into $[-1, 1]$. Next, as $L$ and $U$ are fixed from first stage, let us define $\boldsymbol{\theta}_j$ as follows:

$$\boldsymbol{\theta}_j = \begin{cases} \frac{\pi}{2} + \arcsin(\boldsymbol{X}_j) & \text{if } i \in LU \\ \frac{3\pi}{2} - \arcsin(\boldsymbol{X}_j) & \text{otherwise} \end{cases} \tag{3.14}$$

where $LU = \{i \in \{1, \ldots, n\} : i \in [\min\{L, U\}, \max\{L, U\}]\}$.

The following algorithm describes the three steps of the second stage to consider the testing problems within the Circular space.

- **First Step:**

  Under these assumptions, the hypothesis testing problem is written as:

$$\begin{aligned} H_0^C: &\quad \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 = \boldsymbol{\phi}^* \\ H_1^C: &\quad \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 \in C_{LU} \\ H_2^C: &\quad \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 \end{aligned}$$

  where $\boldsymbol{\phi}^* = \phi \cdot (1, \ldots, 1)$ if $i \in LU$ or $\boldsymbol{\phi}^* = \phi \cdot (1, \ldots, 1) + \pi$ in case $i \notin LU$, and $C_{LU}$ is the order cone $C_{LU} = \{0 \leq \phi_{Lj} \leq \cdots \leq \phi_{Uj} \leq \pi < \phi_{U+1} \leq \cdots \leq \phi_{L-1} < 2\pi \}$. In this way, the hypothesis testing problem above involves circular constraints, i.e., prefixed circular orders.

- **Second Step:**

  The log-likelihood for a general model is:

$$l\left((\boldsymbol{\phi_1}, \boldsymbol{\phi}_2), \kappa; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right) \propto \sum_{i=1}^{n} \left( 2\log 2\pi + \kappa \sum_{j=1}^{2} \cos(\theta_{ij} - \phi_{ij}) - 2\log I_0(\kappa) \right). \tag{3.15}$$

  For $s = 0, 1, 2$, we denote by $\hat{\boldsymbol{\phi}}_s$ and by $\hat{\kappa}_s$ the MLEs under $H_s$ for $\boldsymbol{\phi}$ and for $\kappa$. respectively. The MLE of $\boldsymbol{\phi}$ and $\kappa$ are derived as follow:

  **MLE under $H_0^C$:**

  - Taking into account that the maximum of $\cos x$ occurs at $x = 0$, we obtain:

$$\hat{\boldsymbol{\phi}}_0 = \begin{cases} \theta^* & \text{if } i \in LU \\ \theta^* + \pi & \text{otherwise} \end{cases}$$

    where $\theta^* = Ave(\theta_{ij}) \quad \forall i \in LU$ and for $j = 1, 2$.

  - We have to solve the following equation:

$$\frac{dl}{d\kappa} = R - 2nA(\kappa) = 0, \tag{3.16}$$

    where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, being $I_0(\kappa)$ and $I_1(\kappa)$ the modified Bessel

functions of the first kind and order 0 and 1 respectively. Then:

$$\hat{\kappa}_0 = A^{-1}(\overline{R}), \tag{3.17}$$

note that $R$ is the MRL of $(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$.

**MLE under $H_1^C$:**

- In Rueda et al. (2009) were proved that the CIRE of $\boldsymbol{\theta}$, is the maximum likelihood estimator (MLE) of $\boldsymbol{\phi}$, when $\boldsymbol{\phi} \in C_O$, see Definition (2.2.2). Then:

$$\hat{\boldsymbol{\phi}}_1 = \tilde{\boldsymbol{\theta}}^{(LU)}. \tag{3.18}$$

- The equation to solve is:

$$\frac{dl}{d\kappa} = \sum_{i=1}^{n} \overline{R}_i \cos(\overline{\theta}_i - \phi_i) - 24\frac{I_1(\kappa)}{I_0(\kappa)} = 0, \tag{3.19}$$

Thus, $\hat{\kappa}_1 = A^{-1}\left(\frac{\sum_{i=1}^{n} \overline{R}_i \cos(\theta_i - \phi_i)}{24}\right)$.

**MLE under $H_2^C$:**

- The log-likelihood of the model can be written as:

$$l\left((\boldsymbol{\phi}_1', \boldsymbol{\phi}_2'), \kappa; (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')\right) \propto \sum_{i=1}^{n} \left(2\log 2\pi + \kappa R_i \cos(\overline{\theta_i} - \phi_i) - 2\log I_0(\kappa)\right). \tag{3.20}$$

Then:

$$\hat{\boldsymbol{\phi}}_2 = \overline{\boldsymbol{\theta}} = (\overline{\theta}_1, \ldots, \overline{\theta}_n)', \tag{3.21}$$

where $\overline{\theta}_i = Ave(\theta_{i1}, \theta_{i2})$.

- The estimator of $\kappa$ under $H_2^C$ is solution of the equation:

$$\frac{dl}{d\kappa} = \sum_{i=1}^{n} \overline{R}_i - 24\frac{I_1(\kappa)}{I_0(\kappa)} = 0. \tag{3.22}$$

Thus, $\hat{\kappa}_1 = A^{-1}\left(\frac{\sum_{i=1}^{n} \overline{R}_i}{24}\right)$.

- **Third Step:** We assume $\kappa$ is known and the MLE of $\kappa$ obtained in the second step as the real value. The conditional tests are based on LRT that are defined below. We denote by $LRT_{12}^C$ and $LRT_{01}^C$ the likelihood ratio test statistics for testing $H_1^C$ against $H_2^C - H_1^C$ and $H_0^C$ against $H_1^C - H_0^C$ respectively. They can be written as:

$$\begin{aligned} T_{12}^C = LRT_{12}^C &= -2l_{H_1^C}((\boldsymbol{\phi}_1, \boldsymbol{\phi}_2), \kappa; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) + 2l_{H_2^C}((\boldsymbol{\phi}_1, \boldsymbol{\phi}_2), \kappa; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) = \\ &= 2\kappa \left( \sum_{i=1}^{n} R_i - \sum_{i=1}^{n} R_i \cos(\overline{\theta}_i - \tilde{\theta}_i) \right) \\ &= 2\kappa \left( \sum_{i=1}^{n} R_i [1 - \cos(\overline{\theta}_i - \tilde{\theta}_i)] \right) \\ &= 2\kappa SCE(\overline{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}) \end{aligned} \tag{3.23}$$

$$\begin{aligned} T_{01}^C = LRT_{01}^C &= -2l_{H_0^C}((\boldsymbol{\phi}_1, \boldsymbol{\phi}_2), \kappa; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) + 2l_{H_1^C}((\boldsymbol{\phi}_1, \boldsymbol{\phi}_2), \kappa; (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) = \\ &= 2\kappa \left( \sum_{i=}^{n} R_i \cos(\overline{\theta}_i - \tilde{\theta}_i) - R \right) \end{aligned} \tag{3.24}$$

where $R_i$ is the resultant length of $i$th components of each cycle and $R$ is the resultant length of $(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$. Moreover, note that R is the MLE of $(\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')$ under $H_0^C$.

The conditional test to conduct $H_1^C$ against $H_2^C - H_1^C$ is described in detail in Appendix C. Let us denote as $t_{12}^C$ the observed values of the statistics $T_{12}^C$. Let us further denote by $m^C$ the computing level sets of $\tilde{\boldsymbol{\theta}}^{(LU)}$ under $H_1^C$. Then, the corresponding pvalue can be obtained as follows:

$$p12C = P[\chi^2_{n-m^C} \geq t_{12}^C] \tag{3.25}$$

where p12C denotes the pvalue of testing $H_1^C$ against $H_2^C - H_1^C$ within the Circular space.

In a similar way, we also propose to use an $\alpha$ level conditional test to conduct the testing problem $H_0^C$ against $H_1^C - H_0^C$. We assume that asymptotically, the distribution of $T_{01}^C$ is a $\chi^2$ distribution when $\kappa$ is known. Therefore, the $\alpha$ level conditional test involved to solve $H_0^C$ against $H_1^C - H_0^C$ rejects $H_0^C$ when $T_{01}^C \geq c(m)$, where $c(m)$ is defined as de $1-\alpha'$ percentil of the $\chi^2_{m-1}$ such that:

$$\alpha' = P(\chi^2_{m-1} \geq c(m)) = \frac{\alpha}{1 - P_{\phi^0}(T_{01}^C = 0)} \tag{3.26}$$

where $P_{\phi^0}(T_{01}^C = 0)$ is the probability under $H_0$ that $T_{01}^C = 0$, $\phi^0$ verifying $\phi_1^0 = \cdots = \phi_n^0$, is assumed to be the least favourable configuration, under the hypothesis $H_1^C$ for the $LRT$ in regular testing problems and $m$ is the number of level sets of CIRE of $\boldsymbol{\phi}$ under $H_1^C$, see Fernández et al. (2012). Thus, we assume that the conditional test is asymptotically an $\alpha$ level test, and it allows to obtain pvalues from a $\chi^2_{m-1}$ distribution.

Let us denote as $t_{01}^C$ the observed values of the statistics $T_{01}^C$. Then, the corresponding pvalue can be obtained as follows:

$$p01C = P[\chi^2_{m^C-1} \geq t_{01}^C] \tag{3.27}$$

where p01C denotes the pvalue of testing $H_0^C$ against $H_1^C - H_0^C$ within the Circular space.

## 3.3 False Discovery Correction

In this section we explain how we obtain the final values from which we decide if a signal can be considered cyclic or not. Notice that the pvalues obtained from the test described in previous section must not be directly used as, among other issues, we are performing multiple testing.

In many practical situations, multiple hypotheses testing problems are usually controled by FDR (False Discovery Rate) procedures, see Dudoit et al. (2003). Ac-

curately estimating the rate of false discoveries is a well-recognized problem in any high-throughput analysis, see MacArthur (2012).

FDR is one way of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR controlling procedures are designed to control the expected proportion of rejected null hypotheses that were incorrect rejections, i.e. false discoveries. One of these procedures is the Benjamini–Hochberg procedure, (see Hochberg and Tamhane (1987)), which controls the false discovery rate at level of signification $\alpha$.

The procedure is easy, given a pvalue output and a level of signification $\alpha$, the pvalues must be sorted increasingly $p_{(1)} \leq \ldots p_{(i)} \leq \ldots p_{(n)}$, and only will be considered the pvalues verifying $p_{(i)} < (\frac{i}{n})\alpha$, see Hochberg and Benjamini (1990). The pvalue adjustment is done in two steps:

1. **First Step:** In a similar way to JTK CYCLE, for each gene we define the following final pvalues for the test $H_1$ against $H_2 - H_1$:

$$pp01E = \begin{cases} 1 & \text{if } p12E + p12C < 0.1 \\ p01E & \text{otherwise} \end{cases} \tag{3.28}$$

$$pp01C = \begin{cases} 1 & \text{if } p12E + p12C < 0.1 \\ p01C & \text{otherwise} \end{cases} \tag{3.29}$$

2. **Second Step:**

   - Euclidean Space:

     Let $pp01E_{(1)}, \ldots, pp01E_{(i)}, \ldots, pp01E_{(n)}$ the sorted pvalue output from testing $H_0^E$ against $H_1^E - H_0^E$, once considered (3.28). A gene corresponding to the position $(i)$ will be selected as cyclic with the FDR adjustment procedure if $pp01E_{(i)} \leq (\frac{i}{n})\alpha$. Otherwise, the gene will be identifying as non cyclic.

   - Circular Space:

Let $pp01C_{(1)}, \ldots, pp01C_{(i)}, \ldots, pp01C_{(n)}$ the sorted pvalue output from testing $H_0^C$ against $H_1^C - H_0^C$, once considered (3.29). A gene corresponding to the position $(i)$ will be selected as cyclic with the FDR adjustment procedure if $pp01C_{(i)} \leq (\frac{i}{n})\alpha$. Otherwise, the gene will be identifying as non cyclic.

# Chapter 4

# Numerical Studies

This Chapter presents the results of different numerical studies to validate the new methodology for detecting cyclic genes and to compare it with the JTK CYCLE algorithm 2.3.

This Chapter is divided in three Sections. The first one shows the simulation results for fixed cyclic and non cyclic patterns. In the second one, we present the simulation results from a data base which has been generated imitating those that appear in practical studies. Finally, in the third Section, we analyze a real data base of 250 circadian genes.

## 4.1 Simulation Results

According to literature and behaviour of circadian genes, see Wu et al. (2014), the simulation design has been carried out using observations from two cycles (48 hours, 1h/2days) with period 24 hours. The simulated data are generated from $N_{48}(\boldsymbol{\mu}, \sigma^2 I)$ where $\boldsymbol{\mu}$ follows eight different patterns (as we see below, see Tables 4.1 and 4.2) and $\sigma^2$ is fixed to be 1. We have generated 100 repetitions for each scenario, i.e. for each pattern.

Patterns used to simulate data were generated according to what it usual in the literature, see Deckard et al. (2013). To simulate cyclic genes, we have considered

six patterns, called: *cosine, cosine2, cosinePeak, sineSquare, asymmetric and cosinePeakExtreme*, see Table 4.1. And to simulate non cyclic genes the patters we have considered are named: *flat* and *nonCyclic*, see Table 4.2. All patterns in Tables 4.1 and 4.2 have been generated for two cycles.

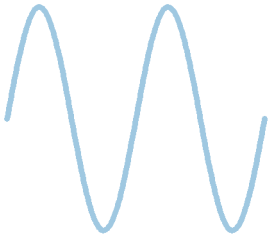
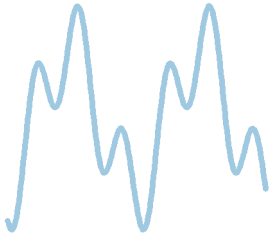
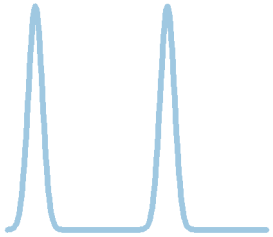
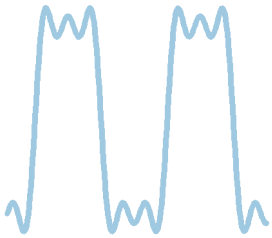
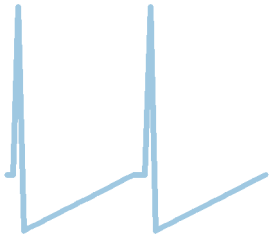
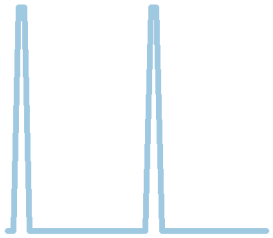
| | | |
|---|---|---|
| Cosine | Cosine2 | CosinePeak |
|  |  |  |
| SineSquare | Asymmetric | CosPeakExt |
|  |  |  |

Table 4.1: Cyclic Patterns

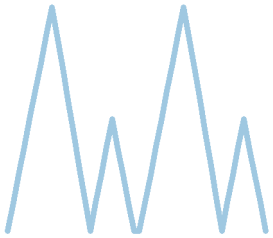
| | |
|---|---|
| Flat | NonCyclic |
|  |  |

Table 4.2: Non Cyclic Patterns

Table 4.3 contains summarizing statistics of pvalues for the testing problem $H_1$ vs $H_2 - H_1$, i.e., we offer summarizing statistics for p12E and p12C in the eight different scenarios (six from cyclic patterns and two from non cyclic patterns) simulated.

Figure 4.1 and 4.2 show the p12E and p12C distributions for cyclic and non cyclic patterns respectively.

| Pattern | Mean p12E | Sd p12E | Mean p12C | Sd p12C |
|---|---|---|---|---|
| Cosine | 0.743 | 0.245 | 0.601 | 0.274 |
| Cosine2 | 0.605 | 0.296 | 0.484 | 0.293 |
| Cosine Peak | 0.648 | 0.265 | 0.500 | 0.277 |
| Sine Square | 0.680 | 0.247 | 0.527 | 0.261 |
| Asymmetric | 0.647 | 0.284 | 0.588 | 0.287 |
| Cosine Peak Extreme | 0.594 | 0.286 | 0.518 | 0.259 |
| Flat | 0.611 | 0.277 | 0.502 | 0.273 |
| Non Cyclic | 0.039 | 0.089 | 0.032 | 0.068 |

Table 4.3: Mean and sd of the pvalues from testing $H_1$ vs $H_2 - H_1$ in each pattern

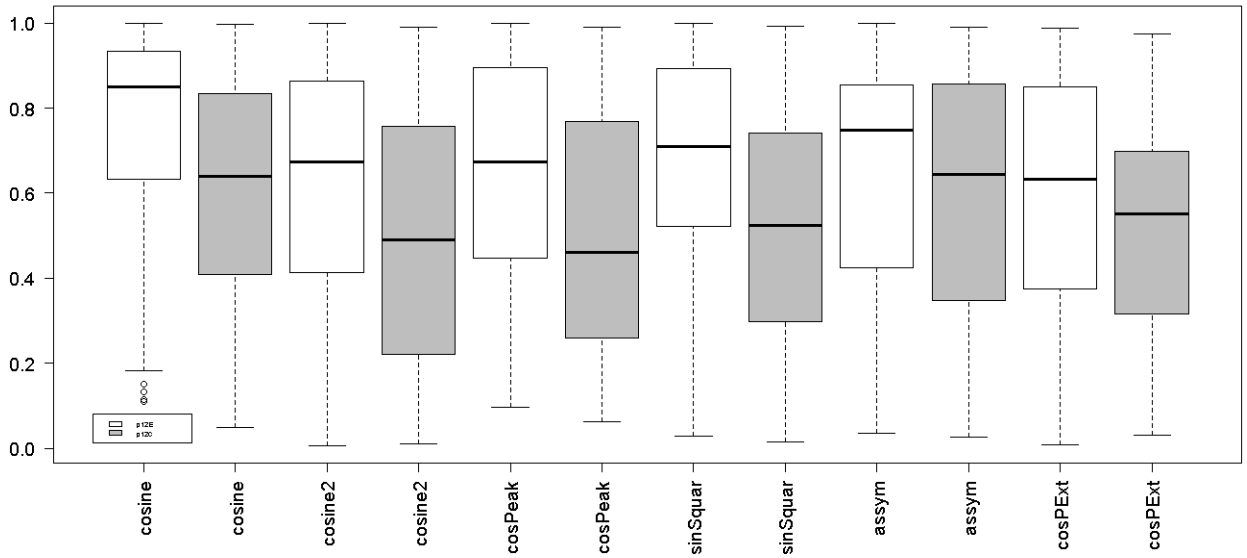



Figure 4.1: Pvalue distribution of cyclic patterns from testing $H_1$ vs $H_2 - H_1$

From Table 4.3 and Figures 4.1 and 4.2 we deduce that according to the cyclic signal definition given, in 3.1.1, only *nonCyclic* patterns present evidences against $H_1$.

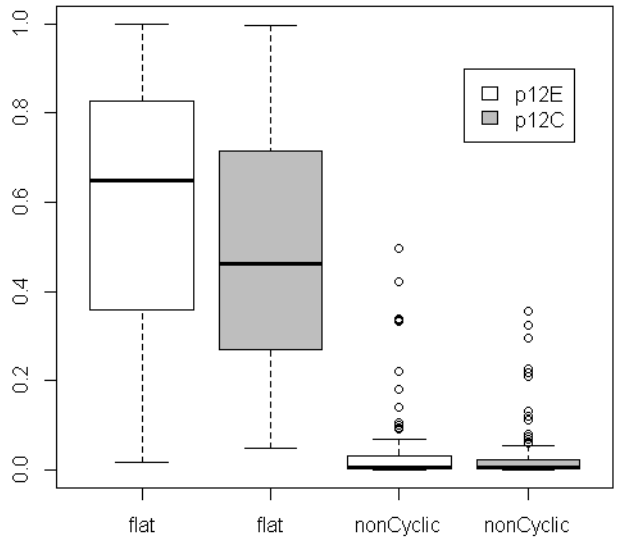



Figure 4.2: Pvalue distribution of non cyclic patterns from testing $H_1$ vs $H_2 - H_1$

Table 4.4 contains summarizing statistics of pvalues for the testing problem $H_0$ vs $H_1 - H_0$, i.e., we offer summarizing statistics for p01E, p01C and pJTK in the eight different scenarios simulated. Figure 4.3 and 4.4 show distributions of the p01E and p01C for cyclic and non cyclic patterns respectively. Finally, Figure 4.5 and 4.6 show for this same testing problem the pJTK, i.e. the pvalues for JTK CYCLE distributions for cyclic and non cyclic patterns respectively.

| Pattern | Mean p01E | Sd p01E | Mean p01C | Sd p01C | Mean pJTK | Sd pJTK |
|---|---|---|---|---|---|---|
| Cosine | $3.5{\cdot}10^{-4}$ | $1.9{\cdot}10^{-3}$ | $1.8{\cdot}10^{-3}$ | $2.8{\cdot}10^{-3}$ | $5.6{\cdot}10^{-6}$ | $1.9{\cdot}10^{-5}$ |
| Cosine2 | $3.7{\cdot}10^{-3}$ | $1.6{\cdot}10^{-2}$ | $6.6{\cdot}10^{-3}$ | $1.7{\cdot}10^{-2}$ | $4.4{\cdot}10^{-3}$ | $1.7{\cdot}10^{-2}$ |
| Cosine Peak | $2.6{\cdot}10^{-4}$ | $1.1{\cdot}10^{-3}$ | $1.6{\cdot}10^{-3}$ | $3.1{\cdot}10^{-3}$ | $4.1{\cdot}10^{-3}$ | $2.6{\cdot}10^{-2}$ |
| Sine Square | $5.8{\cdot}10^{-3}$ | $1.9{\cdot}10^{-2}$ | $1.1{\cdot}10^{-2}$ | $2.8{\cdot}10^{-2}$ | $2.2{\cdot}10^{-2}$ | $1.0{\cdot}10^{-1}$ |
| Asymmetric | $9.5{\cdot}10^{-8}$ | $6.2{\cdot}10^{-7}$ | $7.4{\cdot}10^{-5}$ | $1.4{\cdot}10^{-4}$ | $2.1{\cdot}10^{-1}$ | $3.0{\cdot}10^{-1}$ |
| Cos Peak Ext | $1.0{\cdot}10^{-4}$ | $5.0{\cdot}10^{-4}$ | $8.0{\cdot}10^{-4}$ | $3.0{\cdot}10^{-3}$ | $8.3{\cdot}10^{-1}$ | $3.1{\cdot}10^{-1}$ |
| Flat | $1.2{\cdot}10^{-1}$ | $1.3{\cdot}10^{-1}$ | $9.7{\cdot}10^{-2}7$ | $1.1{\cdot}10^{-1}$ | $9.3{\cdot}10^{-1}9$ | $2.1{\cdot}10^{-1}6$ |
| NonCyclic | $1.6{\cdot}10^{-6}$ | $8.7{\cdot}10^{-6}$ | $2.3{\cdot}10^{-4}$ | $4.7{\cdot}10^{-4}$ | $8.9{\cdot}10^{-5}$ | $2.5{\cdot}10^{-4}$ |

Table 4.4: Mean and sd of the pvalues from testing $H_0$ vs $H_1 - H_0$ in each pattern

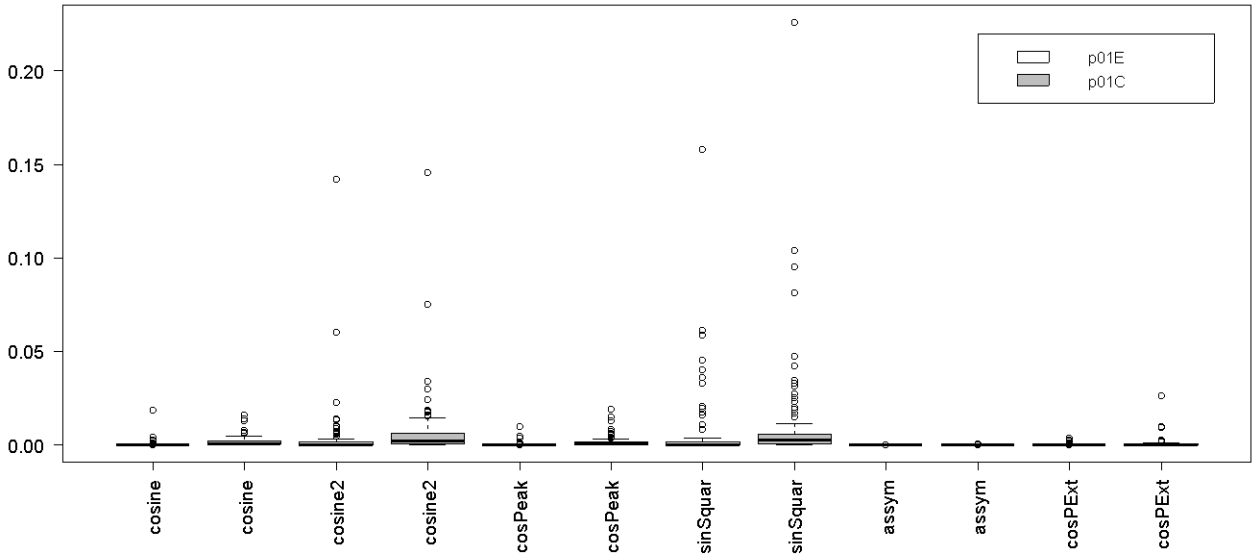



Figure 4.3: Pvalue distribution of cyclic patterns from testing $H_0$ vs $H_1 - H_0$

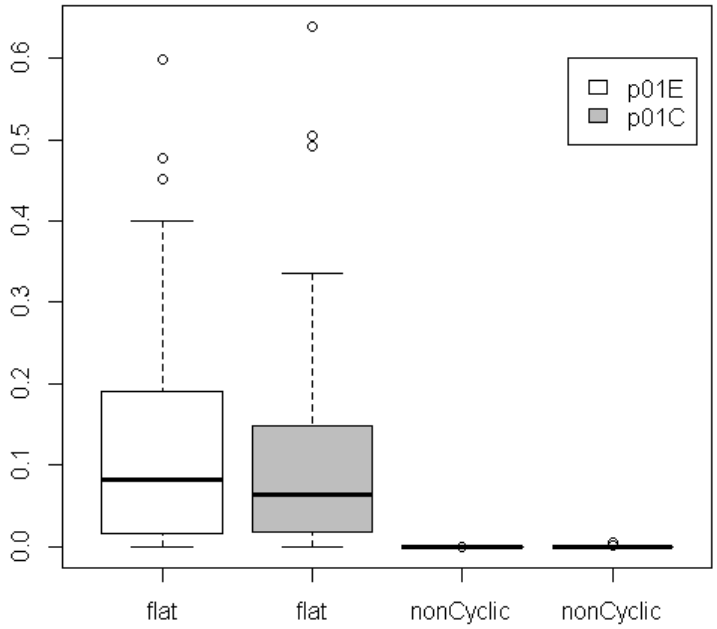



Figure 4.4: Pvalue distribution of non cyclic patterns from testing $H_0$ vs $H_1 - H_0$

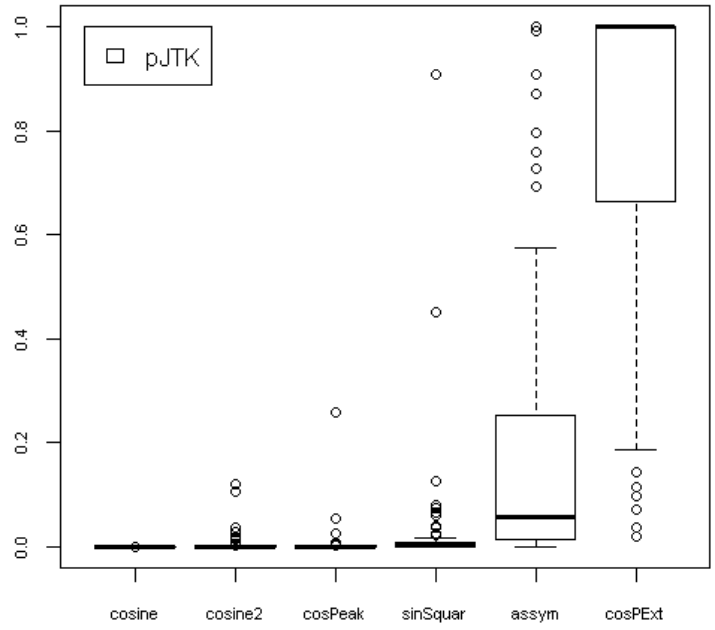



Figure 4.5: JTK Pvalue distribution of cyclic patterns from testing $H_0$ vs $H_1 - H_0$

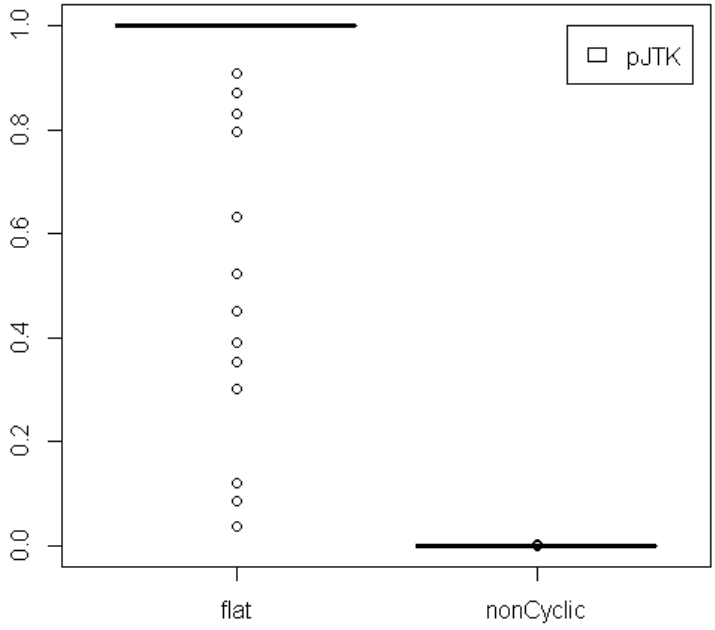



Figure 4.6: JTK Pvalue distribution of non cyclic patterns from testing $H_0$ vs $H_1 - H_0$

Some important comments are that JTK CYCLE classifies non cyclic but periodic patterns as cyclic, while ORI approaches consider them as non cyclic in the testing problem $H_1$ vs $H_2 - H_1$. Moreover, note that asymmetric patterns, (*asymmetric* or *cosinePeakExtreme*) are not detected as cyclic by JTK CYCLE algorithm while the ORI approaches does detect them.

## 4.2 Simulation of 250 'artificial' genes

In order to imitate the behaviour of real data bases we have generated an artificial data base containing a proportion of cyclic and non cyclic patterns close to those appearing in real data bases.

We have generated 250 data (genes) from $N_{48}(\boldsymbol{\mu}, \sigma^2 I)$, where 120 times $\boldsymbol{\mu}$ comes from a cyclic pattern and 130 from a non cyclic pattern and $\sigma^2$ is fixed to be 1. To simulate the 120 cyclic genes we have generated 20 repetitions for each one of the six cyclic patterns, see Table 4.1. The 130 non cyclic genes have been simulated from 110 *flat* patterns and 20 *nonCyclic* patterns, see Table 4.2. The purpose of this simulation is reproduce a real case, where non cyclic genes are larger than cyclic genes. We must underline the novelty of the simulation we present, getting away from the usual guidelines of simulations in this field.

We fix an error measure $\alpha = 0.05$ or $\alpha = 0.01$, each gene is defined as cyclic or not for each approach using the FDR controlling procedure described in Section 3.3. If the gene is classified as cyclic the corresponding indicator variable will take value 1, otherwise it will be 0. The label of indicator variables depends on the method M and on the level of signification L as follows: SEL+M+L, where SEL (from selection) is fixed, M is the method used:

$$M = \begin{cases} E & \text{if the Euclidean method is being used} \\ C & \text{if the Circular method is being used} \\ JTK & \text{if the JTK CYCLE method is being used} \end{cases} \tag{4.1}$$

and L the level of signification chosen:

$$L = \begin{cases} 01 & \text{if the alpha value is 0.01} \\ 05 & \text{if the alpha value is 0.05} \end{cases} \tag{4.2}$$

For example, SELE01=1 means that the gene is detected as cyclic with Euclidean ORI approach and $\alpha = 0.01$. Note that we will also use SELE01 to make reference to Euclidean approach with $\alpha = 0.01$.

Tables 4.5 to 4.10 show the number of cyclic and non cyclic genes identified by the different methods, for each $\alpha$ and for each pattern.

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELE05=0 | 0 | 0 | 0 | 1 | 0 | 0 | 80 | 18 | 99 |
| SELE05=1 | 20 | 20 | 20 | 19 | 20 | 20 | 30 | 2 | 151 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.5: Number of cyclic and non cyclic patterns identified with SELE05

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELE01=0 | 0 | 1 | 0 | 3 | 0 | 0 | 102 | 18 | 124 |
| SELE01=1 | 20 | 19 | 20 | 17 | 20 | 20 | 8 | 2 | 126 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.6: Number of cyclic and non cyclic patterns identified with SELE01

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELC05=0 | 0 | 1 | 0 | 1 | 0 | 0 | 69 | 18 | 89 |
| SELC05=1 | 20 | 19 | 20 | 19 | 20 | 20 | 41 | 2 | 161 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.7: Number of cyclic and non cyclic patterns identified with SELC05

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELC01=0 | 3 | 7 | 1 | 5 | 0 | 1 | 101 | 18 | 136 |
| SELC01=1 | 17 | 13 | 19 | 15 | 20 | 19 | 9 | 2 | 114 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.8: Number of cyclic and non cyclic patterns identified with SELC01

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELJTK05=0 | 0 | 1 | 1 | 2 | 14 | 20 | 110 | 0 | 148 |
| SELJTK05=1 | 20 | 19 | 19 | 18 | 6 | 0 | 0 | 20 | 102 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.9: Number of cyclic and non cyclic patterns identified with SELJTK05

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| SELJTK01=0 | 0 | 2 | 2 | 4 | 16 | 20 | 110 | 0 | 154 |
| SELJTK01=1 | 20 | 18 | 18 | 16 | 4 | 0 | 0 | 20 | 96 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.10: Number of cyclic and non cyclic patterns identified with SELJTK01

From Tables 4.5 to 4.10 we can conclude:

- Both symmetric and asymmetric cyclic patterns are well detected as cyclic genes using ORI approaches.
- JTK CYCLE does not work properly detecting asymmetric patterns such us *cosinepeakExtreme* or *asymetric*. In addition to this, if the pattern is periodic but non cyclic (*nonCyclic*) JTK CYCLE identifies it as cyclic too.
- With respect to the error measures $\alpha$, Euclidean approaches seem working better for $\alpha = 0.01$, and JTK work well for $\alpha = 0.05$.

For each method and $\alpha$ value, Table 4.11 contains two different error rates (ER) using two different weights ($\omega_1$ and $\omega_2$); the proportion of absent events that yield positive test outcomes, i.e. false positive rate (FPR); and the proportion of events that are being tested for which yield negative test outcomes with the test, i.e the false negative rate (FNR). To be more precise, the first column of Table 4.11 indicates the method and error considered, the second one shows the error rates obtained using equal weights for each gene. The third column shows the error rate that results of using weights which are equal in each one of the eights patterns. The fourth and fifth columns show the FPR and FNR respectively.

| Method | $ER\ \omega_1$=1/250 | $ER\ \omega_2$=1/8 | FPR | FNR |
|---|---|---|---|---|
| SELE05 | 13.2 | 5.3 | 24.6 | 0.8 |
| SELE01 | 5.6 | 4.7 | 7.7 | 3.3 |
| SELC05 | 18 | 7.2 | 33.1 | 1.7 |
| SELC01 | 11.2 | 12.8 | 8.5 | 14.2 |
| SELJTK05 | 23.2 | 31.3 | 15.4 | 31.7 |
| SELJTK01 | 25.6 | 32.5 | 15.4 | 36.7 |

Table 4.11: % of Weight Error rates, FPR and FNR for each method

An immediate future task will be repeating this simulation $N$ times to obtain mean error rates, see Section 5.3. Even so, according to different simulations made, the results seem to be quite stable with respect to what we show in Table 4.11.

The main conclusions from Table 4.11 are:

- SELE01 is the method with the smallest error rates. And it is the method where the false positive and false negative rates are more similar, which is a good property if we consider a ROC curve.
- Euclidean approaches seem to work better than Circular do (with both $\alpha$'s).Thus way be due to the fact the data are generated under normal models.
- JTK CYCLE approach exhibits the largest error rates.

- According to the literature, (see Sehgal (2004)) whatever was the tissue the non cyclic genes number is higher than the cyclic ones. In consonance with biologists, a possible weakness of the JTK CYCLE could be the fact that it detects more cyclic genes than expected, which means a high false positive rates for this approach. On the other hand, SELE01 and SELC01, have lower FPRs.

In order to achieve lower FPRs we have defined a mix approach which is simply defined as MIXTsel=SELJTK05*SELE01. Table 4.12 shows the number of cyclic genes identified in each pattern with MIXTsel, and Table 4.13 shows the error rates, FPR and FNR for this method.

| Indicator | Cos | Cos2 | CosPeak | SinSq | Asym | CosPExt | Flat | NonCyc | Sum |
|---|---|---|---|---|---|---|---|---|---|
| MIXTsel=0 | 0 | 2 | 1 | 5 | 14 | 20 | 110 | 18 | 170 |
| MIXTsel=1 | 20 | 18 | 19 | 15 | 6 | 0 | 0 | 2 | 80 |
| Sum | 20 | 20 | 20 | 20 | 20 | 20 | 110 | 20 | 250 |

Table 4.12: Number of cyclic and non cyclic patterns identified with MIXTsel

| Method | $ER\omega$=1/250 | $ER\omega$=1/8 | FPR | FNR |
|---|---|---|---|---|
| MIXTsel | 17.6 | 27.5 | 0.15 | 35 |

Table 4.13: Weight Error rates, FPR and FNR for MIXTsel

From Table 4.12 non cyclic patterns are well detected by MIXTsel, although asymmetric patterns are detected as non cyclic. From Table 4.13, MIXTsel's FPR is 0.15% in the line with biologist recommendations.

Table 4.14 contains the simultaneous detection of cyclic genes with the two methods involved in MIXTsel.

| SELJTK05/SELE01 | SELE01=0 | SELE01=1 | Sum |
|---|---|---|---|
| SELJTK05=0 | 102 | 46 | 148 |
| SELJTK05=1 | 22 | 80 | 102 |
| Sum | 124 | 126 | 250 |

Table 4.14: Number of simultaneous cyclic genes detecting with SELJTK05 and SELE01

From Table 4.14 we can conclude:

- 102 genes are not selected by any of the methods, all of them *flat* patterns.
- 80 genes are selecting using SELE01 and SELJTK05.
- Among theses 80 cyclic genes detected by both methods, we know that 82 are well detected by SELJTK05 with a false positive rate of 15.4%. Whereas MIXTsel detect 78 genes are cyclic with a false positive rate of 0.15%.

Therefore, according to biologists recommendation, the method MIXTsel is good as it gets a low false positive rate. MIXTsel is also in line with the fact that the number of non cyclic patterns in tissues is larger than cyclic ones.

**Observation 4.2.1.** *A note about FDR*

*According to the literature, FDR controlling procedures work when a set of pvalues are obtained from a given statistic test that is used repeatedly under the null hypothesis. In those cases, the expected error rates must be of the same order than the error measure $\alpha$ given, (see Hochberg and Benjamini (1990)). However, in this work we have to take into account two issues that could explain why we obtain higher rates than expected:*

- *The pvalues that we consider are $P(T_{01}^{L,U} \geq t_{obs})$, where $L,U$ are sample-dependent. Then, these pvalues are obtained from different statistic tests $T_{01}^{L,U}$, i.e. the statistic depends on $L$ and $U$, and the expected number of positive is not easy to calculate, although we simulate from normal distributions with equal means.*

- *Moreover, the FDR that we are interested in is not exactly related with $H_0$ because we consider nonCyclic patterns different from the equality (flat patterns).*

*This explain the proposal of using 0.01 instead of 0.05 for the ORI approach and also the proposal of using a mixture approach (MIXTSEL=SELE01*SELJTK05) which has a FDR around 0.025.*

## 4.3 Real Data Results

This Section shows the results obtained from the analysis of a real data base of 250 circadian genes with period 24 hours, from mouse liver. The observations were taken in two cycles of 24 hours, i.e., 1hour/2days. This data base can be found in *CircaDB* (`http://circadb.hogeneschlab.org/query`) an online data base of circadian gene expression which have implemented more useful algorithms to detect cyclic genes. In particular, this data base has been used to test JTK CYCLE algorithm, (see Hughes et al. (2010)) and `http://openwetware.org/wiki/HughesLab:JTK_Cycle`.

From results in Section 4.2 we have selected SELE01, SELJTK05 and MIXTsel approaches to deal with this circadian gene data base. In Figure 4.7 we illustrate the circadian cyclic genes detected by the three approaches.

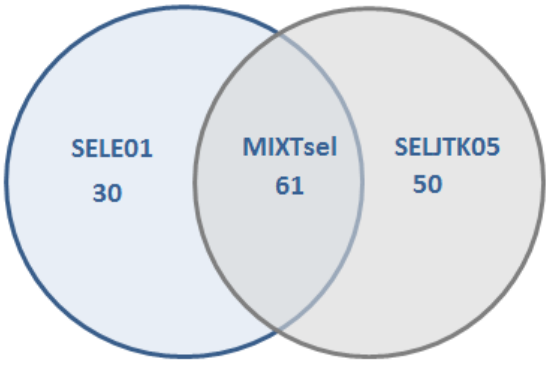



Figure 4.7: Number of circadian cyclic genes detected with SELE01, MIXTsel and SELJTK05

Bellow we analyse particular examples of genes. The genes 8, 146 and 217 showed in Table 4.15 are detected as cyclic by JTK CYCLE algorithm, while ORI approaches do not classify them as cyclic. None of them seems to be cyclic, the first could be cyclic but with other period. In the other two genes expressions the first and second cycles look like different.

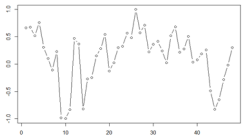
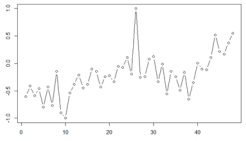
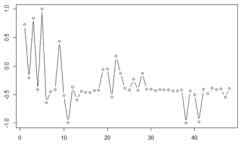
| GENE 8 | GENE 146 | GENE 127 |
|---|---|---|
|  |  |  |

Table 4.15: Profiles of genes 8, 146 and 217 in circadian data base

The mix-up in ORI approaches appears for instance in genes 25, 98 and 199, see Table 4.16. In all of them, the problem seems to be the same, they are not periodic genes, i.e., first and second cycles are clearly different, that should be eliminated testing $H_2$ against $H_3 - H_2$, (this will be considered in a future research, see Chapter 5.3). In contrast to JTK CYCLE which detects some as cyclic, for instance gene 199 with pJTK equals to 0.066; and other ones as non cyclic, such us genes 25 and 98 with pJTKs equal to 1 and 0.134 respectively.

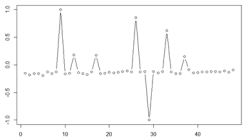
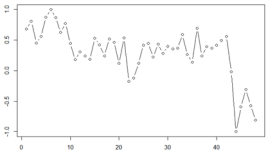
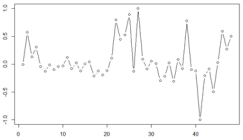
| GENE 25 | GENE 98 | GENE 199 |
|---|---|---|
|  |  |  |

Table 4.16: Profiles of genes 25, 98 and 199 in circadian data base

One of the most clear examples which shows the weakness of JTK to detect periodic but non cyclic genes as cyclic, (*nonCyclic* patterns), occurs with gene 144, (see Table 4.17) whose pJTK is 0.042, i.e., it is detected as cyclic. However, the

Euclidean ORI methodology detects it as periodic but non cyclic in the hypothesis testing problem $H_1$ against $H_2 - H_1$.

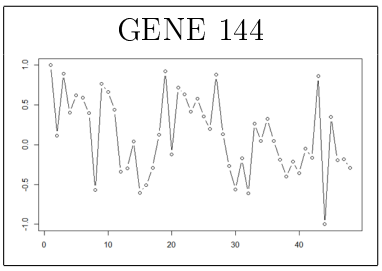



Table 4.17: Profile of gene 144 in circadian data base

Finally, we assess the behaviour of methodology described by rank analysis. We study the Spearman correlation between the pvalue outputs, (see Table 4.18), where according to theoretical bases, Euclidean and Circular approaches are very correlated being but not so much with JTK algorithm.

| Pvalues | p01E | p01C | pJTK05 |
|---|---|---|---|
| p01E | 1 | 0.913 | 0.565 |
| p01C | 0.913 | 1 | 0.494 |
| pJTK05 | 0.565 | 0.494 | 1 |

Table 4.18: Spearman correlation coefficients between pvalues from SELE01, SELC01, SELJTK05

The Spearman coefficients with the 61 circadian cyclic genes detected by MIXTsel are shown in Table 4.19

| Pvalues | p01E | p01C | pJTK05 |
|---|---|---|---|
| p01E | 1 | 0.796 | 0.396 |
| p01C | 0.796 | 1 | 0.373 |
| pJTK05 | 0.396 | 0.373 | 1 |

Table 4.19: Spearman correlation coefficients between pvalues from SELE01, SELC01, SELJTK05 for the 61 circadian genes detected with MIXTsel

The correlation coefficients decrease compared them with those on Table 4.18. As a consequence, the rank of the 61 circadian genes obtained with SELE01 and with SELJTK05 are significantly different. Table 4.21 shows these ranks.

Some interesting differences are remarked in bold in the ranking, (see Table 4.21). Their corresponding gene profiles are shown in Table 4.20. The gene in the first position, whose probeset is 1415673_$at$, coincides for the two approaches. The probeset 1415705_$at$ and 1415743_$at$ with rank positions $5^{th}$ and $8^{th}$ in SELJTK05 respectively, and $50^{th}$ and $46^{th}$ in SELE01 respectively, are apparently better positioned using the ORI approaches. Finally, the gene expression with probeset 1415817_$s$_$at$ which in SELJTK05 ranking appears in $14^{th}$, SELE01 approach considers it as a more relevant circadian gene expression. In fact, the profile of this last gene looks more like cyclic than the two cited before. The ORI methodology for the 61 circadian gene expressions selected as cyclic with MIXTsel approach describe a more consistent and homogeneous ranking than the JTK CYCLE ranking does.

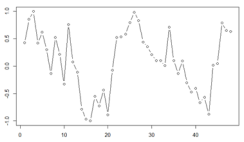
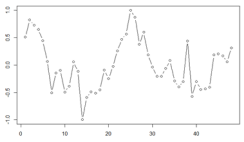
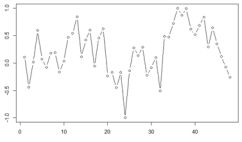
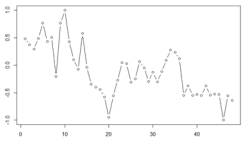
| 1415673_at | 1415705_at | 1415743_at | 1415817_s_at |
|---|---|---|---|
|  |  |  |  |

Table 4.20: Profiles of genes 4, 148, 36 and 74 in circadian data base with probesets 1415673_at, 1415705_at, 1415743_at and 1415817_s_at respectively

| Gene | Probeset | JTK Rank | SELE01 Rank | 224 | 1415893$_$at | 31 | 32 |
|---|---|---|---|---|---|---|---|
| **4** | **1415673_at** | **1** | **1** | 88 | 1415757_at | 32 | 42 |
| 9 | 1415678_at | 2 | 2 | 109 | 1415778$_$at | 33 | 21 |
| 20 | 1415689_s_at | 3 | 3 | 87 | 1415756_a_at | 34 | 57 |
| 22 | 1415691_at | 4 | 36 | 119 | 1415788_at | 35 | 22 |
| **36** | **1415705_at** | **5** | **50** | 150 | 1415819_a_at | 36 | 33 |
| 49 | 1415718_at | 6 | 4 | 16 | 1415685_at | 37 | 23 |
| 54 | 1415723_at | 7 | 5 | 170 | 1415839_a_at | 38 | 61 |
| **74** | **1415743_at** | **8** | **46** | 247 | 1415916_a_at | 39 | 54 |
| 82 | 1415751_at | 9 | 6 | 198 | 1415867_at | 40 | 37 |
| 102 | 1415771$_$at | 10 | 7 | 56 | 1415725_at | 41 | 52 |
| 127 | 1415796_at | 11 | 8 | 64 | 1415733_a_at | 42 | 24 |
| 133 | 1415802_at | 12 | 9 | 78 | 1415747_s_at | 43 | 40 |
| 147 | 1415816_at | 13 | 51 | 90 | 1415759_a_at | 44 | 35 |
| **148** | **1415817_s_at** | **14** | **10** | 110 | 1415779_s_at | 45 | 25 |
| 152 | 1415821_at | 15 | 41 | 67 | 1415736_at | 46 | 45 |
| 158 | 1415827_a_at | 16 | 11 | 190 | 1415859_at | 47 | 26 |
| 159 | 1415828_a_at | 17 | 48 | 68 | 1415737_at | 48 | 27 |
| 171 | 1415840_at | 18 | 12 | 76 | 1415745_a_at | 49 | 47 |
| 181 | 1415850_at | 19 | 13 | 43 | 1415712_at | 50 | 28 |
| 186 | 1415855_at | 20 | 34 | 100 | 1415769_at | 51 | 44 |
| 191 | 1415860_at | 21 | 31 | 225 | 1415894_at | 52 | 49 |
| 223 | 1415892_at | 22 | 14 | 27 | 1415696_at | 53 | 55 |
| 230 | 1415899_at | 23 | 15 | 107 | 1415776_at | 54 | 59 |
| 235 | 1415904_at | 24 | 16 | 47 | 1415716_a_at | 55 | 58 |
| 240 | 1415909_at | 25 | 17 | 91 | 1415760_s_at | 56 | 39 |
| 101 | 1415770_at | 26 | 18 | 17 | 1415686_at | 57 | 29 |
| 57 | 1415726_at | 27 | 56 | 42 | 1415711_at | 58 | 60 |
| 162 | 1415831_at | 28 | 19 | 38 | 1415707_at | 59 | 43 |
| 50 | 1415719_s_at | 29 | 38 | 206 | 1415875_at | 60 | 30 |
| 221 | 1415890_at | 30 | 20 | 37 | 1415706_at | 61 | 53 |

Table 4.21: Ranking sorted by JTK CYCLE rank according to the pvalues obtained. The first column is the position of the gene in the data base, the second one the probeset, and the third and fourth the rank in the pvalues from SELJTK05 and SELE01.

**Observation 4.3.1.** *Computational Note*

*The simulation studies have been carried out under version 3.2.1 of R, using the R-Packages 'Iso' and 'isocir'. Details of these packages can be found in* http://cran.r-project.org/web/packages/Iso/Iso.pdf *and* http://cran.r-project.org/web/packages/isocir/isocir.pdf*, respectively.*

# Chapter 5

# Conclusions

This Chapter summarizes the most relevant conclusions of this work. Section 5.1 recapitulates the main methodological contributions. The conclusions from numerical studies are included in Section 5.2. Finally, Section 5.3 describes an outline of the future work than will be developed from the results obtained here.

## 5.1 Methodological Contributions

In this work we solve the problem of detecting cyclic patterns by designed a new methodology. We start proposing a definition of cyclic signal which incorporates order restrictions. This definition involves both Euclidean and Circular parameters and let us establish a novel mathematical formulation of the model using nested hypotheses testing problems in both spaces. It is assumed that observations in an Euclidean space are available, but the inferences are performed in both Euclidean and a latent Circular space.

The algorithm designed to solve hypotheses testing problems is based on, euclidean and circular order restricted inference methodology in a two stage algorithm common to both spaces. In the first stage we estimate two parameters needed to redefine the hypothesis testing problem. In the second one, specific for each one of the two spaces, the hypotheses testing problems are reformulated and are conducted using conditional tests.

In relation with the validation of the approaches, we propose a new simulation study by generating an 'artificial' data base which imitates the real case. Both cyclic and non cyclic patterns similar to those of real gene expression profiles are used as underlying signals to simulate the data. Moreover, the selection of cyclic genes from the simulated data base has been done using a procedure to control the FDR in a similar way as the biologist suggest to do with real data sets.

Therefore, the methodological contribution can be summarized in four points:

- A new definition of cyclic signal using restrictions.
- Novel formulation of the problem.
- Design of an algorithm to solve the nested hypotheses testing problems.
- New design for the simulations which imitates real data base and controls cyclic patterns detection using FDR.

Moreover, the new methodology supposes an advantage with respect to other algorithms in literature to detect cyclic patters because it provides a general formulation to the problem. On the one hand, we have developed the methodology based on a general and non parametric definition of cyclic signal in contrast to other approaches in the literature which depends on an specific underlying periodic pattern (e.g. sinusoidal patterns). On the other hand, the nested hypotheses testing problems provide new methodology with flexibility enough not only to detect cyclic patters, but also to distinguish between them.

In addition to this, the rigour with we have formulated the model, designed the algorithm and validated the results is a guarantee of the potential of the procedure, in contrast with other approaches in literature whose methodological description turns into a 'black box'.

As a consequence of these intrinsic features of the methodology developed the range of applications is very broad as many periodic phenomena appear in biology and other fields. For instance, we find a direct application of new methodology to

model Hemodynamic Response Functions, (see 5.3), that has an underlying cyclic signal, or any kind of action potential with a similar response profile.

## 5.2 Numerical Studies Conclusions

In this Section we enumerate the main conclusions from numerical studies. However, we also want to recall that other biological conclusions could be obtained if experts carry out a complete analysis from the results in particular cases, such us in the ranking of Table 4.21.

First, the simulation results provide some interesting conclusions about the features of the JTK:

- The JTK CYCLE algorithm is not able to detect as cyclic asymmetric patterns, (see Figure 4.5), in addition to this, periodic but non cyclic patterns are considered as cyclic by JTK CYCLE algorithm, (see Figure 4.6).
- Flat patterns are well detected by JTK CYCLE algorithm. However, this methodology is not able to distinguish among them, because most times their pvalues are equal to 1, (see Figures 4.4 and 4.6).
- The error rates (ER) of JTK CYCLE are higher than any other ER for ORI approaches, (see Table 4.11).

Moreover, related to the ORI methodology, the main points we can conclude from simulation results are:

- Apart from *cosine* pattern, mean and sd pvalues from Euclidean approaches for cyclic patterns are lower than obtained with JTK CYCLE algorithm, (see Table 4.4).
- The testing problem $H_1$ against $H_2 - H_1$ let us identify periodic but non cyclic patterns in contrast to JTK CYCLE algorithm which considers them as cyclic, (see mean p12E or mean p12C for *nonCyclic* patterns in Table 4.3).

- The distribution of pvalues obtained with ORI approaches for cyclic and non cyclic patterns let us rank them to compare them or to validate the performance of the new methodology, (see Figures 4.3 and 4.4).

- According to biologist recommendation, ORI approaches with $\alpha = 0.01$ give the lowest false positive rates (FPRs). Moreover, SELE01 is the method where FPRs and FNR (i.e, false discovery rates) are more similar, (see Table 4.11).

- SELE01 and MIXTsel reduces to a half the false positive rate (FPR) obtained with JTK CYCLE algorithm, (see Table 4.11).

In general, from the simulation studies we conclude that JTK CYCLE algorithm presents failures in detection of asymmetric and periodic but non cyclic patterns. Moreover, their FPRs look like higher than expected in a real case. Since different methods detect different cyclic genes, a mixture of them seems to be appropriate, being our proposal MIXTsel the approach which obtains closer results to the real case with lower FPRs.

## 5.3 Future work

This work takes a further step as statistical inference on angular parameters when they are ordered around a unit circle. However, there are several problems that remain to be addressed and serve as topics for future research. The most relevant are:

- Design of approaches for solving the hypothesis testing problem $H_2$ against $H_3 - H_2$.

- Analysis and validation of the models under the assumptions that $\kappa$ and $\sigma^2$ are unknown.

- Design of a procedure takes into account the uncertainty of unknowing $L$ and $U$. Our initial proposal consists on developing a complete procedure on the EM algorithm.

- Exploring of theoretical properties of the proposed inference procedures. Including the determination of standard errors and confidence intervals to the estimation procedures explained.
- Developing of statistic software to incorporate the new methodology to detect cyclic patterns.
- Checking the performance of the new methodology in other scenarios, as the case of large series from various periods, attenuated patterns in time, detection of outliers,...
- Exploration the ability of new methodology with other real data sets and applications. For example, we find a direct application of new methodology to model Hemodynamic Response Functions, that has an underlying cyclic signal, or any kind of action potential with a similar response profile.

# Appendixes

# Appendix A

# Isotonic Regression

The concept of Isotonic Regression is the baseline within the ORI context. It tries to look for the order closest vector to the observations under certain constraints. It means restricting the least square problem.

In the framework of Euclidean space, we denote by $\boldsymbol{X} = (X_1, ..., X_n)'$ the vector of observed mean values from $n$ populations with sizes $(n_1, ..., n_n)$. We suppose that $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n.$ is the vector of means, and $\Sigma$ the the matrix of covariates with $\Sigma = diag(\frac{\sigma_1^2}{n_1}, ..., \frac{\sigma_n^2}{n_n})$. Let us further assume $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$ be a weight vector associated with the order in the cone $C$. The isotonic regression of $\boldsymbol{X}$ with weight vector $\boldsymbol{\omega}$ is:

$$\boldsymbol{X}^* = \arg\min_{\boldsymbol{Z} \in C} \sum_{i=1}^{n} \omega_i (Z_i - X_i) = \arg\min_{\boldsymbol{Z} \in C} (\boldsymbol{X} - \boldsymbol{Z})' W (\boldsymbol{X} - \boldsymbol{Z}), \tag{A.1}$$

where $W$ denote a positive defined matrix so that $W = diag(\omega_1, ..., \omega_n)$, and $\omega_i$, is a positive weight $\forall \quad i = 1, ..., n$.

If we consider the metric defined by the scalar product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}' W \boldsymbol{v}$, the isotonic regression is unique and is determined by the orthogonal projection of $\boldsymbol{X}$ on the cone $C$, (see Robertson et al. (1988)), it means:

$$\boldsymbol{X}^* = P_W(\boldsymbol{X}/C). \tag{A.2}$$

According to Robertson and Wright (1980) the isotonic regression is obtained as the average of components, as a consequence of the *Cauchy mean value* property; the mean of two values is bounded by them. Then, it exists a partition $\{(l)\}_{l=1}^{m}$ of the index $\{1, ..., n\}$ such as $X_i^* = Av(G_{(l)}) \forall i \in (l)$, where $G_{(l)} = \{X_i\}_{i \in (l)}$ and $Av(G_{(l)}) = \frac{\sum_{i \in (l)} \omega_i X_i}{\sum_{i \in (l)} \omega_i}$. These ones are the equivalent in the Euclidean space to the level sets defined in Rueda et al. (2009) in the Circular space, we call them level sets of $\boldsymbol{X}^*$ too.

The PAVA (*pool adjacent violator algorithm*) is the algorithm proposed in Robertson and Wright (1980) to solve the problem of isotonic regression in case that $\boldsymbol{\mu} \in C$, it is based on averaging adjacent observations which violate the order constrains. The Cauchy mean value property is essential to PAVA runs correctly. The complete algorithm appears in Barragán (2014)

When the order constrains are not the simple order there are other algorithms which work out the isotonic regression problem, Dykstra (1981), Lee (1983) or Pardalos and Xue (1999).

In the context of Circular data, Rueda et al. (2009) establish the beses for Circular isotonic regression. There is also proposed an algorithm based on PAVA to solve the circular isotonic regression.

# Appendix B

# Restricted Maximum Likelihood Estimator, RMLE

The great majority of works related to this field have developed methods for Normal models, i.e. we assume that $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is the vector of observations, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$ is the vector of means and $\Sigma$ the the matrix of covariates, $\Sigma = diag(\frac{\sigma_1^2}{n_1}, ..., \frac{\sigma_n^2}{n_n})$. In order to simplify we consider $\Sigma$ known and $\boldsymbol{\mu} \in C$, with $C$ order cone. Under these assumptions, the likelihood function is:

$$L(\boldsymbol{X}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|} \exp\{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\} \tag{B.1}$$

In this Appendix we show the restricted maximum likelihood estimator, (RMLE) associated to the Normal model.

The RMLE for the parameter $\boldsymbol{\mu}$ is exactly the isotonic regression for the order represented by the cone $C$, i.e.:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}^* = P_{\Sigma^{-1}}(\boldsymbol{X}/C), \tag{B.2}$$

where $P_{\Sigma^{-1}}$ denotes the corresponding orthogonal projection, (see AppendixA for details). Moreover, in Chapter 1 of Robertson et al. (1988) is included a detailed construction of the RMLE in exponential families.

We also conclude from Robertson et al. (1988), properties such as RMLE is biased and its mean squared error is less than the MLE one:

$$E[(\boldsymbol{X}^* - \boldsymbol{\mu})'(\boldsymbol{X}^* - \boldsymbol{\mu})] \geq E[\boldsymbol{X} - \boldsymbol{\mu})'(\boldsymbol{X} - \boldsymbol{\mu})]. \tag{B.3}$$

However this feature is not always true estimating parameter functions.

The first references with regard to the inference with constrains estimation are Lee (1981), Lee (1988) and Kelly (1989). The same issue has been studied in later works by Menéndez and Salvador (1991), Rueda et al. (1997a,b), Fernández (1995) and Fernández et al. (1997, 1998, 1999, 2000).

Note that the RMLE is the analogous to the CIRE within the Circular space, (see Rueda et al. (2009)), i.e. associated to von Mises models.

# Appendix C

# Conditional test

Conditional tests have been widely studied in the literature. Related to this work, Bartholomew (1961) discussed a conditional likelihood ratio test to testing homogeneity of means versus linear order constrains between them for the isotonic normal means problem. And Robertson and Wegman (1978) discussed the corresponding conditional test for testing if the means satisfy a linear order versus there were no restrictions under them. Both tests are the base to conduct hypotheses testing problems proposed in this work within the Euclidean space.

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)' \in \mathbb{R}^n$ denotes the vector of sample means of a random sample from $N_n(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ is the vector of populations means and $\Sigma$ the matrix of covariances $\Sigma = diag(\frac{\sigma_1^2}{n_1}, \ldots \frac{\sigma_1^2}{n_n})$. We assume, we have independent samples where the $i$-th sample is of size $n_i$, with $i = 1, \ldots, n$.

Under these assumptions, we propose consider a conditional test to settle the following hypotheses test:

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_n \\ H_1 &: \mu_1 \leq \cdots \leq \mu_n \\ H_2 &: \boldsymbol{\mu} \in \mathbb{R}^n \end{aligned} \tag{C.1}$$

The likelihood ratio test constructed in Bartholomew (1961) rejects $H_0$ for large

values of the statistic:

$$LRT_{01} = \sum_{i=1}^{n} (X_i^* - \overline{X}_i)^2 n_i, \tag{C.2}$$

where $\boldsymbol{X}^* = (X_1^*, \ldots, X_n^*)$ and $\overline{\boldsymbol{X}} = \frac{\sum_{i=1}^n n_i X_i}{\sum_{i=1}^n n_i}(1, \ldots, 1)$ are the maximum likelihood estimator of $\boldsymbol{\mu}$ and under $H_1$, and $H_0$, respectively, see Robertson et al. (1988).

The distribution of $LRT_{01}$ under $H_0$ is given by:

$$P(LRT_{01} \geq t) = \sum_{m=1}^{n} P_0(m, n) P(\chi^2_{m-1} \geq t), \tag{C.3}$$

where $\chi_i^2$ denotes a chi-square random variable with $i$ degrees of freedom and $P_0(m, n)$ are defined to be the probabilities under $H_0$ of obtaining $m$ distinct values among $\boldsymbol{X}^*$. A distribution of this form is know as a *Chi-Bar-Squared* (mixture of $\chi^2$ distributions), and critical values are easily obtained if the coefficients $P_0(m, n)$ are known. They are easy to calculate if $n_1 = n_2 = \cdots = n_n$, but very difficult for other situations.

Bartholomew discussed a related procedure which largely avoids this difficulty. The idea was to condition on $m$, the number of distinct values in $\boldsymbol{X}^*$, and to compare $LRT_{01}$ to a critical value for a chi-square with degrees of freedom determined by $m$. This is valid because under $H_0$ the conditional distribution of $LRT_{01}$ given $m$ is equal to the restricted distribution, and is a chi-square with $m-1$ degrees of freedom (this is implicit in the proof of Theorem 3.1 Barlow et al. (1972)). In order to obtain a size $\alpha$ test, one must allow for the fact that $LRT_{01} = 0$ with probability $P_0(1, n)$, and adjust the chi-square critical value accordingly, so the test rejects $H_0$ if $LRT_{01} > t_m$, where $t_m$ satisfies:

$$P(\chi^2_{m-1} \geq t_m) = \frac{\alpha}{1 - P_0(1, n)} \tag{C.4}$$

The likelihood ratio test of $H_1$ versus $H_2$ was constructed by Robertson and Wegman (1978), it rejects $H_1$ for large values of

$$LRT_{12} = \sum_{i=1}^{n} (X_i - X_i^*)^2 n_i. \tag{C.5}$$

The distribution of $LRT_{12}$ for arbitrary $\boldsymbol{\mu} \in H_1$ is intractable, but in Robertson and Wegman (1978) is also shown that the least favourable configuration (i.e., the $\boldsymbol{\mu}$ for which the probability of a type I error is maximized) is $\boldsymbol{\mu} \in H_0$. Moreover, the distribution of $LRT_{12}$ under $H_0$ is again a chi-bar-square, involving the same $P_0(m, n)$ coefficients:

$$P(LRT_{12} \geq t) = \sum_{m=1}^{n} P_0(m, n) P(\chi^2_{n-m} \geq t). \tag{C.6}$$

As with $LRT_{01}$, the conditional distribution under $H_0$ of $LRT_{12}$ given $m$ is a chi-square. Hence, we can construct a conditional test of $H_1$ versus $H_2$ which rejects $H_1$ if $LRT_{12} > t_m$, where $t_m$ satisfies:

$$P(\chi^2_{n-m} > t_m) = \frac{\alpha}{1 - P_0(n, n)}. \tag{C.7}$$

Here, the coefficients $P_0(n, n)$ is "in a sense least favourable," the probability under $H_0$ that $LRT_{12} = 0$ (or that there are $n$ distinct values in $\boldsymbol{X}^*$), and it is proved in Barlow et al. (1972) that this probability is smaller for $\boldsymbol{\mu} \in H_0$ than for another $\boldsymbol{\mu} \in H_1$. Finally, it is shown in Bartholomew (1961) that both tests are asymptotically of level $\alpha$.

The tests below, are simple cases where conditional likelihood ratio tests can be conducted to solve hypothesis testing problem. However, conditional tests go beyond and can be spread out to solve very general hypotheses testing problems, see Militino et al. (2015).

In particular, conditional tests have also been studied within the Circular space, see Fernández et al. (2012). This document proposes a conditional test within the Circular space, based on the $LRT^C_{12}$ defined in (3.23), to conduct the analogous test $H_1$ against $H_2 - H_1$ explained in this Appendix within the Circular space, it will be one of the test we solve in this work, see (3.15)). This document also concludes that, asymptotically the conditional test proposed to solve this testing problem is of level $\alpha$ and that $LRT_{12}$ under $H_1$ follows $\chi^2_{n-m}$ distribution when the dispersion parameter

$\kappa$ is known, where $m$ is the number of level sets of the maximum likelihood estimator CIRE (see Rueda et al. (2009)), of the parameter involved in the hypothesis testing problem under $H_1$.

# Appendix D

# The von Mises distribution

From the point of view of the statistical inference, the most useful distribution on the circle is the von Mises distribution which plays a similar role than Normal distribution on the Euclidean spaces.

A random variable $\theta$ has a von Mises distribution $\theta \sim VM(\phi, \kappa)$, with $\phi \in [0, 2\pi)$ and $\kappa \geq 0$, if the probability density function is:

$$g(\theta; \phi, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \phi)} \qquad x \in [0, 2\pi), \tag{D.1}$$

where $I_0$ denotes the modified Bessel function of the first kind and order 0. The modified Bessel function of first kind and order $q$ is given by:

$$I_q(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} cos(qx) e^{\kappa \cos(\theta)} d\theta. \tag{D.2}$$

The parameter $\phi$ is the mean direction and the parameter $\kappa$ is known as the concentration parameter. This distribution is unimodal and symmetric with respect to the mean direction $\phi$, as it is showed in the figure D.1.

**Maximum likelihood Estimation for von Mises distributions**

Let $\theta_1, \ldots, \theta_n$ a random sample from $VM(\phi, \kappa)$. The log-likelihood is:

$$\begin{aligned} l(\theta_1, \ldots, \theta_n; \phi, \kappa) &= n \log 2\pi + \kappa \sum_{i=1}^{n} \cos(\theta_i - \phi) - n \log I_0(\kappa) \\ &= n\{\log 2\pi + \kappa \overline{R} \cos(\overline{\theta} - \phi) - \log I_0(\kappa)\} \end{aligned} \tag{D.3}$$

Since the function $\cos x$ has its maximum at $x = 0$, the maximum likelihood estimate $\hat{\phi}$ is:

$$\hat{\phi} = \overline{\theta}. \tag{D.4}$$

Differentiating D.3 with respect to $\kappa$ and using the fact that $I_0'(\kappa) = I_1(\kappa)$ gives:

$$\frac{\partial l}{\partial \kappa} = n\{\overline{R} \cos(\overline{\theta} - \phi) - A(\kappa)\} \tag{D.5}$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$. Then the maximum likelihood estimation $\hat{\kappa}$ of $\kappa$ is the solution of :

$$A(\kappa) = \hat{\kappa}, \tag{D.6}$$

i.e.

$$\hat{\kappa} = A^{-1}(\overline{R}) \tag{D.7}$$

Different values of $\kappa$ produce different shapes in the distribution such as we can see in the Figure D.1.

The von Mises distribution is related with other distributions; if $\kappa = 0$, the we obtain the uniform distribution in the circle. When $\kappa \to \infty$, we conclude $\kappa^{-1/2}(\theta - \phi) \sim N(0, 1)$ (see pp.38 Mardia and Jupp (2000)).

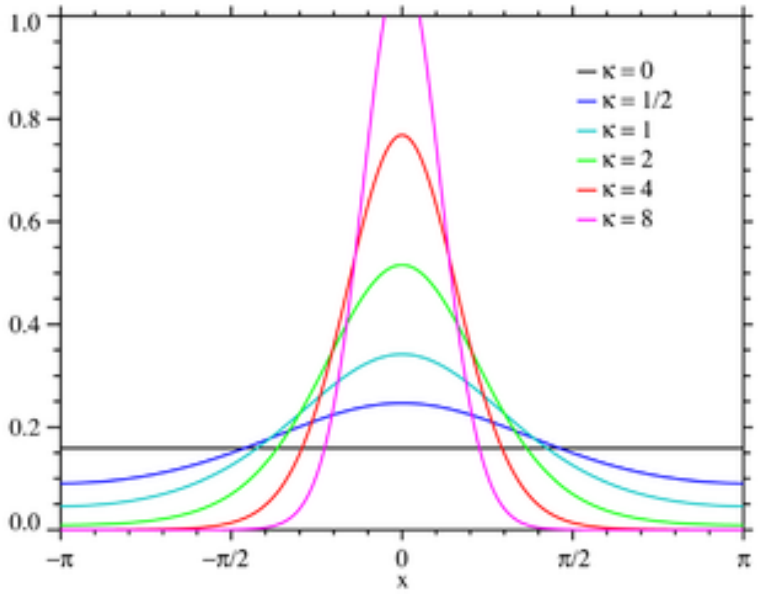



Figure D.1: Density of the von Mises distribution $VM(0, \kappa)$, with $\kappa = 0, 1/2, 1, 2, 4, 8$.

# Bibliography

R.E. Barlow, D.J. Bartholomew, J.M. Bremer, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Willey, 1972.

S. Barragán. *Procedimientos estadísticos para modelos circulares con restricciones de orden aplicados al análisis de expresiones de genes*. PhD thesis, Universidad de Valladolid, 2014.

S. Barragán, M.A. Fernández, C. Rueda, and S.D. Peddada. isocir: An R package for constrained inference using isotonic regression for circular data, with an application to cell biology. *Journal of Statistical Software*, 54(4):1–17, 2013. URL http://www.jstatsoft.org/v54/i04/.

S. Barragán, M.A. Fernández, K.V. Mardia, S.D. Peddada, and C. Rueda. Circular piesewise regression with an application to cell-cycle gene biology. *Plos One*, In Press, 2015.

D.J. Bartholomew. A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 239–281, 1961.

A. Deckard, R.C. Anafi, J.B. Hogenesch, S.B. Haase, and J. Harer. Design and analysis of large-scale biological rhythm studies: A comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics*, 2013. doi: 10.1093/bioinformatics/btt541. URL http://bioinformatics.oxfordjournals.org/content/early/2013/09/20/bioinformatics.btt541.abstract.

S. Dudoit, J.P. Shaffer, , and J.C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.

R.L. Dykstra. An isotonic regression algorithm. *Journal of Statistical Planning and Inference*, 5(4):355–363, 1981.

M.A. Fernández. *Comportamiento del estimador máximo verosímil para un parámetro k-dimensional en modelos con restricciones*. PhD thesis, Universidad de Valladolid, 1995.

M.A. Fernández, C. Rueda, and B. Salvador. On the maximum likelihood estimator under order restrictions in uniform probability models. *Communications in Statistics-Theory and Methods*, 26(8):1971–1980, 1997.

M.A. Fernández, C. Rueda, and B. Salvador. Simultaneous estimation by isotonic regression. *Journal of statistical planning and inference*, 70(1):111–119, 1998.

M.A. Fernández, C. Rueda, and B. Salvador. The loss of eficiency estimating contrast under restrictions. *Scandinavian Journal of Statistics*, (26):579–592, 1999.

M.A. Fernández, C. Rueda, and B. Salvador. Parameter estimation under orthant restrictions. *Canadian Journal of Statistics*, 28(1):171–181, 2000.

M.A. Fernández, C. Rueda, and S. D. Peddada. 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(7):2823–2832, 2012. doi: 10.1093/nar/gkr1077. URL http://nar.oxfordjournals.org/content/40/7/2823.abstract.

N.I. Fisher. *Statistical analysis of Circular Data*. Cambridge, 1993.

F. Halberg, Y.L. Tong, and E.A. Johnson. Circadian system phase — an aspect of temporal morphology; procedures and illustrative examples. In H. von Mayersbach, editor, *The Cellular Aspects of Biorhythms*, pages 20–48. Springer Berlin Heidelberg, 1967. ISBN 978-3-540-03744-6. doi: 10.1007/978-3-642-88394-1_2. URL http://dx.doi.org/10.1007/978-3-642-88394-1_2.

Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, 1990. ISSN 1097-0258. doi: 10.1002/sim.4780090710. URL http://dx.doi.org/10.1002/sim.4780090710.

Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. Willey, 1987.

M.E. Hughes, J.B. Hogenesch, and K. Kornacker. Jtk cycle: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380, 2010. doi: 10.1177/0748730410379711. URL http://jbr.sagepub.com/content/25/5/372.abstract.

G.J. Iverson and Harp S. A. A conditional likelihood ratio test for order restrictions in exponential families. *Mathematical Social Sciences*, 14(2):141 – 159, 1987. ISSN 0165-4896. doi: http://dx.doi.org/10.1016/0165-4896(87)90018-7. URL http://www.sciencedirect.com/science/article/pii/0165489687900187.

A.R. Jonckheere. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41:133–145, 1954. doi: http://dx.doi.org/10.2307/2333011.

R.E. Kelly. Stochastic reduction of loss in estimating normal means by isotonic regression. *The Annals of Statistics*, pages 937–940, 1989.

C.C. Lee. The quadratic loss of isotonic regression under normality. *The Annals of Statistics*, pages 686–688, 1981.

C.C. Lee. The min-max algorithm and isotonic regression. *The Annals of Statistics*, pages 467–477, 1983.

C.C. Lee. Quadratic loss of order restricted estimators for treatment means with a control. *The Annals of Statistics*, pages 751–758, 1988.

J. Li, G.R. Grant, J.B. Hogenesch, and M.E. Hughes. Chapter sixteen - considerations for rna-seq analysis of circadian rhythms. In Amita Sehgal, editor, *Circadian Rhythms and Biological Clocks, Part A*, volume 551 of *Methods in Enzymology*, pages 349 – 367. Academic Press, 2015. doi: http://dx.doi.org/10.1016/bs.mie.

2014.10.020. URL http://www.sciencedirect.com/science/article/pii/S0076687914000214.

D. MacArthur. Methods: Face up to false positives. *Nature*, 487:427–428, 2012. doi: doi:10.1038/487427a.

K.V. Mardia and P.E. Jupp. *Directional statistics*, volume 494. John Wiley and Sons, 2000.

J.A. Menéndez and B. Salvador. Anomalies of the likelihood ratio test for testing restricted hypotheses. *The Annals of Statistics*, pages 889–898, 1991.

J.A. Menéndez, C. Rueda, and B. Salvador. Conditional test for testing a face of the tree order cone. *Communications in Statistcs. Simulation and computation*, 20(1,2):751–762, 1991.

A.F. Militino, C. Rueda, and M.D. Ugarte. Checking unimodality and locating the break-point: an application to breast cancer mortality trend. *SERRA*, Submited for publication, 2015.

P.M. Pardalos and G. Xue. Algorithms for a class of isotonic regression problems. *Algorithmica*, 23(3):211–222, 1999.

T. Robertson and E.J. Wegman. Likelihood ratio tests for order restrictions in exponential families. *The Annals of Statistics*, pages 485–505, 1978.

T. Robertson and F.T. Wright. Algorithms in order restricted statistical inference and the cauchy mean value property. *Ann. Statist.*, 8(3):645–651, 05 1980. doi: 10.1214/aos/1176345014. URL http://dx.doi.org/10.1214/aos/1176345014.

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restristed Statistical Inference*. John Willey and Sons, 1988.

C. Rueda, B. Salvador, and M.A. Fernández. A good property of the maximum likelihood estimator in a restricted normal model. *Test*, 6(1):127–135, 1997a. ISSN 1133-0686. doi: 10.1007/BF02564430. URL http://dx.doi.org/10.1007/BF02564430.

C. Rueda, B. Salvador, and Fernández M.A. Simultaneous estimation in a restricted linear model. *Journal of Multivariate Analysis*, 61(1):61 – 66, 1997b. ISSN 0047-259X. doi: http://dx.doi.org/10.1006/jmva.1997.1657. URL http://www.sciencedirect.com/science/article/pii/S0047259X97916572.

C. Rueda, M.A. Fernández, and S.D. Peddada. 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(485):338–347, 2009. doi: 10.1198/jasa.2009.0120. URL http://dx.doi.org/10.1198/jasa.2009.0120.

A. Sehgal. *Molecular Biology of Circadian Rhythms*. Willey, 2004.

M.J. Silvapulle and P. K. Sen. *Constrained Statistical Inference. Inequality, Order, and Shape Restrictions*. John Willey and Sons, 2005.

M. Straume. Dna microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. In Ludwig Brand and Michael L. Johnson, editors, *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 149 – 166. Academic Press, 2004. doi: http://dx.doi.org/10.1016/S0076-6879(04)83007-6. URL http://www.sciencedirect.com/science/article/pii/S0076687904830076.

T.J. Terpstra. The asymptotic normality and consistency of kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14(1952): 327–333, 1952.

P.F. Thaben and P.O. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014. doi: 10.1177/0748730414553029. URL http://jbr.sagepub.com/content/29/6/391.abstract.

S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 2004. doi: 10.1093/bioinformatics/btg364. URL http://bioinformatics.oxfordjournals.org/content/20/1/5.abstract.

H. Wijnen, F. Naef, and M.W. Young. Molecular and statistical tools for circadian transcript profiling. *Methods in enzymology*, 393:341–365, 2005.

P.C. Wollan and R. L. Dykstra. Conditional tests with an order restriction as a null hypothesis. In *Advances in Order Restricted Statistical Inference*, pages 279–295. Springer, 1986.

G. Wu, J.g Zhu, J. Yu, L. Zhou, J.Z. Huang, and Z. Zhang. Evaluation of five methods for genome-wide circadian gene identification. *Journal of biological rhythms*, 29 (4):231–242, 2014.