Universidad de Valladolid

Facultad de Ciencias

Departamento de Estadística e Investigación Operativa

TESIS DOCTORAL:

Statistical methodology and software to analyse oscillatory signals with applications to biology

Presentada por Yolanda Larriba González para optar al grado de Doctora por la Universidad de Valladolid

Dirigida por:

Dra. Cristina Rueda Sabater
Dr. Miguel Alejandro Fernández Temprano

Cristina Rueda Sabater, Catedrática de Universidad, y Miguel Alejandro Fernández Temprano, Catedrático de Universidad, certifican que la presente memoria ha sido realizada, bajo su dirección, por Yolanda Larriba González, en el Departamento de Estadística e Investigación Operativa de la Universidad de Valladolid.

Valladolid, 12 de Marzo de 2020

Agradecimientos:

En primer lugar me gustaría dar las gracias a mi directora, Cristina, por su tiempo, su dedicación y tesón en este trabajo, y también por haber sido capaz de contagiarme su inquietud y entusiasmo por la investigación. Gracias por todo lo que me has enseñado.

También me gustaría dar las gracias a mi codirector, Miguel, por su esfuerzo para que este trabajo saliera adelante, por su guía, sus consejos y por haber estado siempre dispuesto a ayudarme.

Gracias a los dos por haber confiado en mí desde el primer momento para trabajar en este proyecto y por poner a mi alcance todos los medios posibles para dar visibilidad a nuestro trabajo. Por extensión, me gustaría agradecer al resto de miembros del Departamento de Estadística e Investigación Operativa su acogida durante estos años. En especial, a aquellos que un día también fueron mis profesores y que me 'engancharon' a la Estadística.

Cabe agradecer, la implicación y orientación de Shyamal Peddada en esta tesis y a título personal la atención recibida durante mis meses de estancia en Pittsburgh. También me gustaría dar las gracias al profesor Frank Scheer, por haberme brindado la oportunidad de trabajar en su equipo, ofreciéndonos una visión más cercana de los problemas biológicos.

Es de recibo agradecer a todas las instituciones que me han permitido disfrutar de ayudas para realizar esta tesis, en especial me gustaría dar las gracias a la fundación IMFAHE, por su beca Women in STEM, con la que pude realizar una segunda estancia de investigación en la prestigiosa Universidad de Harvard.

De manera general, gracias a todos aquellos que de un modo u otro me habéis enseñado a aprender y me habéis acompañado durante estos años. Comenzando por mis padres, a quienes les quiero agradecer la educación que he recibido, su esfuerzo y el haber despertado en mí la curiosidad por todo lo que me rodeaba; y por supuesto gracias a mis abuelos, por su cariño, por haberme cuidado siempre y por todo lo que he aprendido de ellos. Gracias también a mis amigas, esas que han permanecido a mi lado desde que éramos pequeñas, y a las nuevas amistades que he encontrado en mis estancias y en los congresos durante estos años. Gracias a mi 'incondicional', por todas esas tardes de viernes que me has aguantado sin reprocharme nunca nada. Gracias a todos aquellos que no os habéis cansado de preguntarme –*¿Qué tal vas Yol?* Y gracias sobre todo a ti, por no soltarte, por tu ilusión, por creer en mí y por hacer que el final de este trayecto haya sido tan fácil.

Por último, me gustaría dedicar esta tesis a Lara, para que nunca pierda esa curiosidad que ahora reflejan sus ojos.

Resumen

Numerosos procesos biológicos, como por ejemplo el ciclo menstrual, la actividad reproductora, el ciclo celular o el ciclo circadiano exhiben un comportamiento rítmico que se manifiesta en señales con patrones oscilatorios que se repiten de forma periódica, véanse Zhang et al. (2014), Liu et al. (2017), Caba et al. (2018), Draper et al. (2018) y Seney et al. (2019). El estudio de estas señales temporales rítmicas y su variación a lo largo del tiempo se conoce como cronobiología.

Desde un punto de vista estadístico, el análisis de datos en cronobiología presenta ciertos retos: (a) la señal puede adoptar una gran variedad de patrones rítmicos (Korenčič et al. (2012), Zhang et al. (2014), Rueda et al. (2019)); (b) con mucha frecuencia se dispone de información de pocos periodos y el número de observaciones por periodo es bajo (Panda et al. (2002), Hughes et al. (2007, 2009), Yang and Su (2010)); (c) en este tipo de datos subyace una estructura circular; (d) diversas fuentes de variabilidad afectan a las observaciones; (e) en ocasiones se desconocen los instantes de tiempo en los que se tomaron las muestras. Por todo ello, los modelos estándar de series de tiempo, o los modelos de Fourier no son adecuados para el análisis de este tipo de datos. Es habitual encontrar en la literatura modelos paramétricos clásicos, basados en funciones sinusoidales, como Cosinor (Tong (1976), Cornelissen (2014)). Sin embargo, estas funciones sinusoidales en ocasiones son demasiado rígidas, ya que en la mayoría de procesos biológicos se observan patrones temporales rítmicos asimétricos.

Varios son los problemas estadísticos que se plantean en cronobiología. Quizás el problema más clásico sea el de la identificación de señales con patrones temporales rítmicos; ya que es habitual encontrar múltiples señales asociadas a estos procesos, algunas de las cuales no son rítmicas. Este problema ha sido muy estudiado en la literatura y existen diferentes procedimientos para la detección de señales rítmicas, véanse Levine et al. (2002), Liu et al. (2004) o Straume (2004) entre otros. Por su uso extendido entre los biólogos destacamos los algoritmos JTK_Cycle (JTK) (Hughes et al. (2010)) y RAIN (Thaben and Westermark (2014)), basados en el test de Jonckheere-Terpstra y la correlación Tau de Kendall. Pese a ser dos de los procedimientos de detección de ritmicidad más utilizados en la práctica, ambos presentan una alta tasa de clasificación errónea en patrones rítmicos asimétricos.

En la mayoría de aplicaciones, se asume que se conocen los instantes de tiempo en los que se toman las muestras. Sin embargo, hay casos, como por ejemplo en el análisis de datos de biopsias humanas o de datos de expresiones de genes en cadáveres, en los que se desconocen esos instantes de tiempo y deben estimarse previamente a cualquier análisis de ritmicidad (Bossé et al. (2012), Li et al. (2013), Seney et al. (2019)). Dicho problema de estimación se conoce en cronobiología como estimación del orden temporal. Algunos de los procedimientos propuestos recientemente en la literatura para abordar este problema son Oscope (Leng et al. (2015)) y CYCLOPS (Anafi et al. (2017)). Oscope se diseñó específicamente para la reconstrucción de la dinámica del ciclo celular y su aplicabilidad se reduce únicamente a experimentos scRNA-Seq. CYCLOPS,

lejos de dar una formulación matemática del problema, propone una solución basada en redes neuronales, lo que dificulta la evaluación e interpretación de los resultados, y requiere de información adicional que no siempre está disponible.

Además de los dos problemas principales, mencionados anteriormente, existen otras muchas cuestiones interesantes planteadas relacionadas con el análisis de señales rítmicas, como la estimación del momento de máxima expresión y/o la comparación y clasificación de señales rítmicas.

La motivación fundamental de esta tesis es resolver problemas asociados al análisis de expresiones de genes, cuya actividad está gobernada por el ciclo circadiano. A estos genes se los conoce como genes circadianos porque presentan patrones de expresión sincronizados con dicho ciclo. En concreto, el problema que supuso el inicio de esta tesis, fue el de la identificación, entre las miles de señales que encontramos en un estudio genético, aquellos genes con patrones de expresión rítmicos.

En esta tesis se propone un marco estadístico teórico, basado principalmente en metodología de Inferencia con Restricciones de Orden (ORI), que supone un nuevo paradigma donde formular una gran variedad de problemas en cronobiología. Todos los procedimientos desarrollados han sido implementados en el software estadístico R, para hacerlos accesibles en la práctica.

La metodología ORI se caracteriza por la incorporación de información que se conoce a priori en términos de restricciones sobre los parámetros del modelo. Los procedimientos de inferencia resultan más eficaces y las soluciones son biológicamente interpretables. Esta metodología surgió en la primera mitad de los años 50 con los trabajos de Brunk (1955), Van Eeden (1956) y Bartholomew (1959). En Robertson and Wright (1980), Robertson et al. (1988) y Menéndez and Salvador (1991) se ahondó en estas primeras teorías y se desarrollaron nuevos algoritmos para hacerlas accesibles en la práctica. Los trabajos de Rueda et al. (2009), Fernández et al. (2012) y Barragán et al. (2015) extendieron algunos de los procedimientos clásicos de la ORI en el espacio circular. En la actualidad, la metodología ORI aparece en numerosas investigaciones en estadística aplicada, véanse, entre otros, Rueda et al. (2016), Wang and Zhong (2017) o Rueda et al. (2019).

La clave fundamental de la que arranca el desarrollo metodológico ORI presentado en esta tesis es la representación matemática de señal oscilatoria mediante restricciones *up-down-up*, entre los paramentos en el espacio Euclídeo, y la formulación equivalente de orden circular, entre los parámetros en el espacio circular (Rueda et al. (2009)). Las referencias básicas de la Estadística Direccional son Fisher (1993) y Mardia and Jupp (2000). Algunos de los avances más recientes en este campo pueden encontrarse en Ley and Verdebout (2017) y Ley and Verdebout (2018).

Esta tesis, que se presenta como compendio de publicaciones, consta de cuatro contribuciones científicas; tres artículos y un capítulo de libro, existiendo entre todos ellos una cohesión temática. La exposición sigue el orden cronológico en el que se desarrollaron.

En Larriba et al. (2016) se aborda el problema específico de la detección de ritmicidad para datos de expresión de genes derivados de la tecnología de microarrays (Irizarry et al. (2003a)). Se establece, por primera vez, una definición de señal rítmica en el espacio Euclídeo usando restricciones de orden y se diseña ORIOS, un algoritmo basado en tests de hipótesis anidados que involucran restricciones, para detectar y clasificar señales rítmicas. Para resolver estos contrastes de hipótesis los autores proponen el uso de test condicionales (Bartholomew (1961), Barlow et al. (1972), Robertson et al. (1988)). Los resultados derivados de este trabajo muestran que ORIOS presenta mayor potencia en la detección de genes rítmicos que sus principales competidores (JTK y RAIN), controlando la tasa de falsos positivos. Además, ORIOS identifica posibles nuevos genes circadianos que pueden ser relevantes para los biólogos.

Larriba et al. (2018) surgió a raíz de la constatación de que los datos de expresión de genes derivados de microarrays están sujetos a distintas fuentes de variabilidad y que la elección del método de normalización (preprocesado para eliminar/reducir el ruido sistemático de los datos) podía afectar sustancialmente en la detección de ritmicidad (Tu et al. (2002), Gautier et al. (2004)). Con el objeto de cuantificar y eliminar dicha dependencia, en este trabajo se introduce una medida de ritmicidad, basada en metodología bootstrap, que identifica genes rítmicos de forma robusta frente a la elección de la estrategia de normalización. Además, la metodología bootstrap desarrollada se presenta como una herramienta útil para simular datos de expresiones de genes. Se demuestra que la nueva medida de ritmicidad es eficaz para la detección de genes rítmicos independientemente de la normalización utilizada. En particular, se obtienen correlaciones muy altas entre los 'rankings' de ritmicidad de los genes obtenidos a partir de esta nueva medida para todas las normalizaciones consideradas en el trabajo. En Larriba et al. (2019) se propone una extensión de este trabajo. En concreto, dicha metodología bootstrap se aplica para el análisis de ritmicidad de las líneas celulares humanas U2OS, incluyendo detalles computacionales y extendiendo los resultados iniciales obtenidos en Larriba et al. (2018).

Finalmente, Larriba et al. (2020) presenta el marco teórico general, basado en metodología ORI, para el análisis de datos en cronobiología, y puede considerarse la contribución más importante de la tesis. En concreto, se establece la definición rigurosa de *señal circular*, utilizando restricciones de orden tanto en el espacio Euclídeo como en el espacio circular y se propone un modelo estadístico de *señal circular* más ruido para el análisis de señales rítmicas. Esta formulación equivalente de ritmicidad entre ambos espacios sustenta la metodología desarrollada en este trabajo.

En primer lugar, en el marco de este modelo con restricciones, se resuelve el problema de la estimación de *señal circular* como un problema de Regresión Isotónica (IR) (Robertson et al. (1988)). En segundo lugar, se formula el problema de la detección de ritmicidad como un problema de contraste de hipótesis con restricciones. En tercer lugar, se plantea el problema de la estimación del orden temporal como un problema de optimización que busca el orden entre los instantes de tiempo que minimiza la distancia entre el estimador de IR bajo ese orden y los datos. Este problema de minimización que incorpora restricciones inicialmente no tiene solución directa (NP-hard) (Bartholdi III et al. (1989)). Sin embargo, se puede abordar el problema considerando su representación en un grafo, donde los nodos son las observaciones y el objetivo es buscar la ruta más corta que recorra todas las observaciones exactamente una vez, empezando y acabando en la misma observación, lo que se corresponde con un orden circular. La formulación equivalente de ritmicidad entre los espacios Euclídeo y circular, traduce dicho orden circular en un orden *up-down-up* entre las observaciones euclídeas. Así, el problema de optimización puede resolverse como un problema del viajante (TSP), muy estudiado en investigación operativa (Flood (1956), Lawler et al. (1985), Reinelt (1994)). Aunque el problema TSP no tiene solución exacta, existen numerosas heurísticas, algunas de ellas implementadas en R, que llevan a soluciones aproximadas. Por último, en este trabajo, se definen varias medidas que son útiles para validar la bondad de este procedimiento frente a otras alternativas en la literatura y de uso muy extendido entre los biólogos. Los resultados arrojados de este trabajo se concretan en que la nueva metodología para la detección de ritmicidad mejora las tasas de error de JTK; y en que la reconstrucción del orden temporal con la nueva propuesta proporciona resultados más verosímiles que CYCLOPS para las medidas de validación mencionadas anteriormente.

Esta tesis supone una sólida aportación metodológica en el campo de la Estadística con Restricciones y un avance importante en la resolución de diversos problemas reales. La metodología ORI recogida en este trabajo permite una mejor interpretación de muchos de los problemas clásicos de cronobiología y mejora los resultados de sus competidores en los problemas estudiados. La versatilidad, simplicidad e interpretabilidad biológica de los elementos y procedimientos estadísticos desarrollados en esta tesis, y el éxito de su aplicación en la solución de muy diversos problemas han supuesto el nacimiento de una nueva metodología para el análisis de señales oscilatorias que está teniendo un avance vertiginoso y que cuenta con aplicaciones no solo en el campo de la cronobiología, sino también en otras áreas tan dispares como la astrofísica (Rueda et al. (2019)). Una aplicación particularmente interesante, en la que estamos trabajando arduamente en este momento, está relacionada con la electrofisiología del corazón. Los resultados de este trabajo pueden ser de gran repercusión para la salud, ya que esperamos avances significativos en el diagnóstico automático de enfermedades cardiovasculares, lo que podría significar una posible reducción en la tasa de mortalidad por esas causas.

Contents

List of Figures

List of Tables

1. Introduction

Motivation

Many physiological and biological phenomena, such as menstrual cycles (Draper et al. (2018)), reproductive activity (Simonneaux and Bahougne (2015), Caba et al. (2018)), cell cycle (Liu et al. (2017)) or circadian biology (Hughes et al. (2009), Zhang et al. (2014), Andreani et al. (2015), Seney et al. (2019)), are governed by oscillatory systems consisting of numerous signals that exhibit rhythmic patterns over time. For example, the circadian clock is a molecular pacemaker that orchestrates daily functional activity including metabolic state, endocrine activity or neural excitability. Genes involved in those processes that exhibit rhythmic expression patterns along ∼24-hour periods are called circadian genes. The study of such signals with temporal rhythmic patterns, and how these patterns change under different conditions, is called chronobiology.

Chronobiology has been an active area of research during the past two decades, with major impact on treating cardiovascular disorders like hypertension (Halberg et al. (2013)), detecting genes associated with neurodegenerative disorders (Li et al. (2013)) or depression (Chauhan et al. (2017)), and improving the effectiveness of cancer treatments (Chan et al. (2017)). For instance, Haus (2009) demonstrated that the timing of radiation according to host and/or tumour rhythms improves the toxic/therapeutic ratio of the treatment. These and other findings in biomedical sciences have increased interest in chronobiological experiments.

From a statistical point of view, the analysis of rhythmic signals ($\boldsymbol{\mu}$) in chronobiology has several challenges because of: (a) $\boldsymbol{\mu}$ displays a wide variety of rhythmic patterns over time, which are not exactly sinusoidal or even symmetric (Korenčič et al. (2012), Zhang et al. (2014), Rueda et al. (2019)); (b) the density of the time points and the number of periods of data is usually very small (Panda et al. (2002), Hughes et al. (2007, 2009), Yang and Su (2010)); (c) the intrinsic circular nature of data from oscillatory systems; (d) the variability in time course expression data due to noisy nature of the data; (e) in some applications, the temporal order among samples may be unknown. For these reasons, standard time series or Fourier models are not convenient for the analysis of chronobiological rhythms (Elkum and Myles (2006), Wijnen et al. (2006), Leise (2013)). Models based on parametric functions of time, such as Cosinor, have been proposed in chronobiology to model these patterns (Tong (1976), Cornelissen (2014)). The main drawback of these approaches is that such parametric functions are too rigid, as signals in oscillatory systems very often exhibit asymmetric patterns.

There are several commonly encountered problems in chronobiology. The main problem to solve in this context is rhythmicity detection as not all patterns observed in an oscillatory system display rhythmic patterns. For a given signal $\boldsymbol{\mu}$, rhythmicity detection can be formulated as the following hypothesis test:

$$\begin{aligned} H_0 &: \boldsymbol{\mu} \textit{ is a flat signal} \\ H_1 &: \boldsymbol{\mu} \textit{ is rhythmic signal.} \end{aligned} \tag{1.1}$$

This problem has been studied extensively in literature, existing a wide variety of procedures to address it including, among others, those based on sinusoidal curve fitting (Liu et al. (2004), Straume (2004), Cornelissen (2014)), autocorrelation (Levine et al. (2002)) or Fourier analyses (Wichert et al. (2004)). Some non-parametric approaches, such as JTK_Cycle (JTK) (Hughes et al. (2010)) and RAIN (Thaben and Westermark (2014)), based on Jonckheere-Terpstra test and Kendall's tau correlation, are widely employed by biologists. However, these two latter approaches do not detect asymmetric rhythmic patterns properly.

A fundamental assumption made in the above discussion is that the time corresponding to each biological sample is known. However, in many instances, such as when dealing with samples obtained from human cadavers (Li et al. (2013), Seney et al. (2019)) or human organ biopsies, (Lamb et al. (2011), Bossé et al. (2012)), the exact time corresponding to each biological sample may be unknown. In such cases, one needs to first estimate or determine the time associated with each sample before investigating rhythmicity. This problem, known as temporal order estimation, is other crucial issue in chronobiology. Some recent procedures to cope with this problem are Oscope (Leng et al. (2015)) and CYCLOPS (Anafi et al. (2017)). Oscope was specifically designed to recover cell cycle dynamic, and it is only applicable in single cell RNA-Seq experiments. CYCLOPS is far from a mathematical close-fitting formulation. It is based on a neural network framework (which is like a black box) and uses additional rhythmicity information which is not always available.

In addition to the major rhythmicity issues mentioned above, other interesting questions related to the analysis of oscillatory signals, such as peak time estimation or rhythm-pattern comparisons, deserve consideration. For instance, when dealing with circadian genes, time peak estimation reveals crucial information for biologists about the timings at which genes' biological function is carried out.

The main motivation of this thesis was to solve appealing rhythmicity questions specifically related to the analysis of circadian gene expression. In particular, the starting problem of this thesis was to identify among the several thousand of genes registered in a genetic study, those that display rhythmic expression patterns.

Methodology

This thesis proposes a general statistical framework, based on Order Restricted Inference (ORI) and R-based software, which states a new paradigm to formulate and solve a wide range of problems in chronobiology.

ORI methodology is characterized by incorporating prior information to the statistical model in terms of order restrictions on the model parameters to increase the efficiency, flexibility and interpretability of statistical inference procedures. ORI is an active research field in statistics which arose in the mid-1950s (Brunk (1955), Van Eeden (1956), Bartholomew (1959)). Robertson and Wright (1980), Robertson et al. (1988) and Menéndez and Salvador (1991) extended theories proposed in the early years providing new algorithms for making them accessible in practice. Moreover, Rueda et al. (2009), Fernández et al. (2012) and Barragán et al. (2017) developed ORI procedures to the circular manifold. These days, ORI research usually arises in applied statistics, see Rueda et al. (2016), Wang and Zhong (2017) or Rueda et al. (2019) among others.

The core of the ORI methodology proposed in this thesis is the formulation of rhythmic signal ($\boldsymbol{\mu}$) using mathematical inequalities, which we call *up-down-up* restrictions, among the parameters in the Euclidean space and the equivalent formulation of circular order among the parameters in the cicular space (Rueda et al. (2009)). Fisher (1993) and Mardia and Jupp (2000) describe the basics of Directional Statistics. Recent advances in this field can be found in Ley and Verdebout (2017) and Ley and Verdebout (2018).

Publications in the compendium

This thesis is presented as a compendium of publications. Specifically, it consists of four scientific contributions, three articles and one book chapter dealing with interrelated problems in chronobiology. Methodological details, contributions and results relative to each of them are shown later in the text. An overall description of each of these four contributions is outlined in the next paragraph.

Larriba et al. (2016) deals with the rhythmicity detection problem for gene expression data obtained from microarray technology. ORIOS algorithm, based on several nested ordered rhythmicity tests (see (1.1)), proposed in this paper identifies circadian genes in those data bases, but also classifies them according to their expression patterns. Larriba et al. (2018) emerged from observing that microarray gene expression data are subject to several sources of noise, and that the normalization (data preprocessing) choice may substantially impact on the determination of a gene to be rhythmic. In this paper, a bootstrap-based methodology was proposed to robustly identify circadian genes. Moreover, this bootstrap-based methodology can be used as a useful tool to simulate microarray gene expression data. An extension of this work can be found in Larriba et al. (2019), where the bootstrap-based methodology is applied in the U2OS human cell lines. The analyses conducted support the initial findings appearing in Larriba et al. (2018). Larriba et al. (2019) also provides step-by-step instructions to implement the R-based code developed by Larriba et al. (2018). Finally, Larriba et al. (2020) presents a general statistical framework, based on ORI methodology, Gaussian models and likelihood, to analyse rhythmicity. The ORI methodology developed relies on the definition of what we call *circular signal*, i.e. rhythmic signals, which are equivalently formulated both in the Euclidean and in the circular space. A standard *circular signal* plus error model is stated to address, among others, signal and peak time estimation, rhythmicity detection or temporal order estimation. The methodology is very flexible, provides a better insight of many problems and is broadly applicable in the context of oscillatory systems.

Contributions and results

Next, we describe in detail the contributions and results of each of the publications considered in this thesis. Full versions of the papers can be found in Section 2.

To improve the readability of this section, the mathematical notation from the original contributions has been slightly modified, so that an homogeneous notation is considered in the rest of the section. For simplicity in the exposition, we shall use the term “gene" to describe the response variable of interest and “gene expression" for the outcome.

For each gene g, $g = 1, \ldots, G$, assume that its expression is obtained at n time points $(t_i, i = 1, \ldots, n)$ in each of J periods of data, with T being the length of each period. Let X_{ij}^g be the observed expression data collected from gene g at time point t_i, within the jth period, and $\boldsymbol{X}_j^g = (X_{1j}^g, \ldots, X_{nj}^g)'$ denote the vector of data from gene g at the n time points in the jth period, for

$j = 1, \dots, J$ and $g = 1, \dots, G$. Let us further denote $\boldsymbol{Y}^g = (\overline{X}^g_{1.}, \dots, \overline{X}^g_{n.})'$, where $\overline{X}^g_{i.}$ denotes the average of data from gene g collected at time point t_i across the J periods, for $i = 1, \dots, n$ and $g = 1, \dots, G$. In our set-up both J and T are known. To improve readability, we shall delete the superscript g whenever possible.

Order restricted inference for oscillatory systems for detecting rhythmic signals. Larriba et al. (2016)

In Larriba et al. (2016), it is assumed that $J = 2$ and $T = 24$; that for each given time point t_i, the data collected across periods has the same expected value ($\mathbb{E}(X_{ij}) = \mu_i$); that the covariance matrix of $\boldsymbol{X}_j$ is a diagonal matrix, and that the period samples are independent from one another, for $i = 1, \dots, n$, and $j = 1, 2$. In other words, the observed data are modelled by the signal plus error model:

$$\boldsymbol{X}_j = \boldsymbol{\mu} + \boldsymbol{\epsilon}_j,$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ is assumed to be a *cyclic signal* and $\boldsymbol{\epsilon}_j = (\epsilon_{1j}, \dots, \epsilon_{nj})'$ with $\boldsymbol{\epsilon}_j \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$, for $j = 1, 2$.

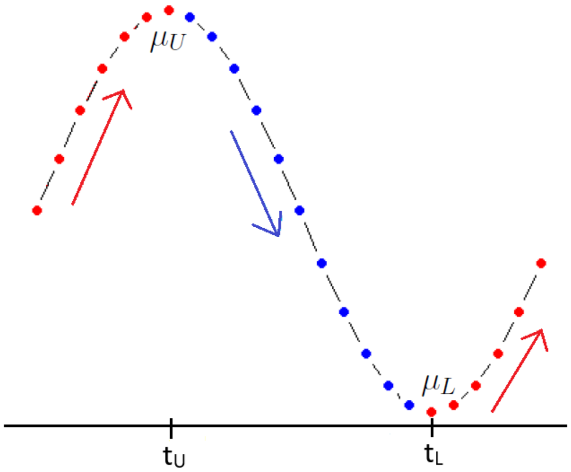


Figure 1.1: *Cyclic signal* $\boldsymbol{\mu}$

Cyclic signals proposed in Larriba et al. (2016) are the first attempt to formulate *up-down-up* signals, that underlie in many oscillatory systems, using order restrictions. Several authors model $\boldsymbol{\mu}$ as a parametric function of time (e.g. sinusoidal), see Straume (2004), Mockler et al. (2007) or Cornelissen (2014) among others. However, these functions are too rigid as they do not cope with asymmetric patterns that frequently appear in rhythmic processes. Mathematically, a signal $\boldsymbol{\mu}$ is said to be *cyclic* iff $\boldsymbol{\mu} \in C = \bigcup_{LU} C_{LU}$, where $L, U \in \{1, \dots, n\}$ and $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \leq \cdots \leq \mu_U \geq \cdots \geq \mu_L \leq \cdots \leq \mu_n \leq \mu_1\}$, where without loss of generality we assume $L > U$. Graphically, a *cyclic signal* $\boldsymbol{\mu}$ displays an *up-down-up* pattern with a unique peak (μ_U) and a unique trough (μ_L). It monotonically increases up to time point t_U and then decreases up to a time point t_L before increasing again, see Figure 1.1. They are entirely described by mathematical inequalities which provides flexibility, so that this class of *cyclic* functions allows to capture very heterogeneous patterns (not only sinusoidal) as shown in Figure 1.2.

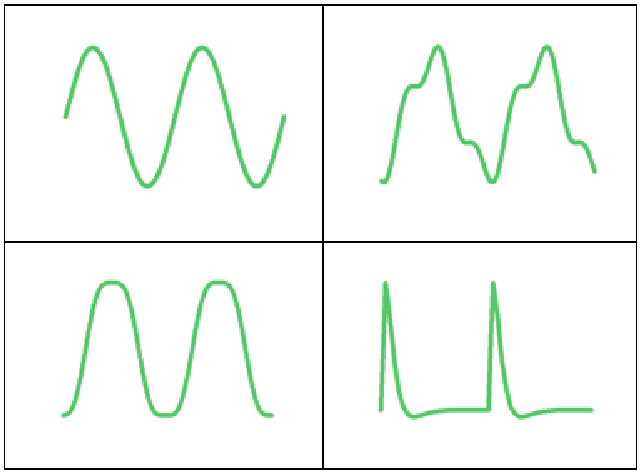


Figure 1.2: Four examples of rhythmic signal across two periods

The main goal in Larriba et al. (2016) is to discover circadian genes exhibiting rhythmic gene expression patterns. Yet, we observed that for gene data bases, in addition to genes displaying *cyclic* expression pattern, there were genes displaying multiple peaks and/or troughs within each period that are interesting for biologists. We refer to these patterns as *quasi-cyclic*. Thus, both genes with *cyclic* and *quasi-cyclic* patterns are considered as rhythmic. On the other hand, these data bases contain non-rhythmic genes. Both genes with noisy pattern across the two periods (called *flat*), and those with *flat* pattern in one of the periods and *cyclic* pattern in the other one (called *non-flat* & *non-periodic*), are considered as non-rhythmic genes.

ORIOS (Order Restricted Inference for Oscillatory Systems), a three step algorithm, is designed in Larriba et al. (2016) to deal with rhythmicity detection problem. Unlike widespread rhytmicity detection procedures, such as JTK or RAIN, ORIOS not only detects circadian (rhythmic) genes, but also classifies them into the four classes described above considering tests of ordered hypotheses. The statistical tests performed in ORIOS are conditional tests (CT) (Bartholomew (1961), Barlow et al. (1972), Robertson et al. (1988)). CTs are conditional versions of the likelihood ratio test commonly used in ORI as they are computationally simple and often more powerful for interesting alternatives (Robertson et al. (1988), Fernández et al. (2012)). Since there are a large number of genes, a large number of tests are performed. Consequently, resulting P-values are adjusted by Benjamini-Hochberg (BH) to control the false discovery rate (Benjamini and Hochberg (1995)).

In the first step of the procedure, ORIOS estimates the peak and trough time points indexes for each gene as follows: $\hat{L} = \arg \min_{i=1,\ldots,n} Y_i$ and $\hat{U} = \arg \max_{i=1,\ldots,n} Y_i$. Next, in the *filtering step*, ordered rhythmicity test (1.1) is conducted separately for each period. For each gene, the higher P-value of each period is considered. Genes rejecting the null in (1.1), i.e. H_0 : $\boldsymbol{\mu}$ *is a flat signal*, are declared as potentially rhythmic. Otherwise, they are non-rhythmic genes. Then, genes pass through the *classification step*. Test (1.1) is conducted for mean expression values of those genes declared as non-rhythmic in the *filtering step*, so that, those rejecting H_0 are declared to be *non-flat* & *non-periodic*. Otherwise, they are declared to be *flat*.
For mean expression values of the genes declared as potentially rhythmic in the *filtering step*, two nested ordered tests are conducted as follows. If we denote $H_2 : \boldsymbol{\mu} \in \mathbb{R}$, H_1 is tested against $H_2 \setminus H_1$ so that, genes rejecting H_1 are periodic but *non-cyclic* or *flat*, and we declare them as *quasi-cyclic*. For those genes which do not reject H_1, the ordered testing problem (1.1) is conducted. If H_0 is rejected, genes are declared as *cyclic*, otherwise they are *flat*. Table 1.1 summarizes the gene classification according to the results obtained from the testing procedure.

Table 1.1: Gene classification according to ORIOS algorithm

Filtering Stage	Classification Stage		Result
	H_1 vs $H_2 - H_1$	H_0 vs $H_1 - H_0$	
Reject?	Reject?	Reject?	
Yes	Yes	-	*Quasi-cyclic*
Yes	No	Yes	*Cyclic*
Yes	No	No	*Flat*
No	-	Yes	*Non-flat* & *non-periodic*
No	-	No	*Flat*

ORIOS performance was studied using simulated as well real data in terms of false positive/negative rates and compared to those of JTK and RAIN.
We analysed four well-known microarray gene expression data sets (Hughes et al. (2010), Thaben and Westermark (2014) and Larriba et al. (2016)) which are online available at NCBI GEO (`http://www.ncbi.nlm.nih.gov/geo/`). The mouse liver and pituitary gland as well as the NIH3T3 cell lines data consisted of 45101 genes each, whereas the U2OS human cell lines data consisted

of 32321 genes. Gene expression data are given along two periods of 24 hours length with a sampling frequency of $1hour/2days$. This same four data bases are used throughout this thesis to illustrate and validate the methods.

Results show that ORIOS has substantially higher power to detect a wide variety of rhythmic patterns than its competitors, controlling missclassification rates. In real settings, the detection of *quasi-cyclic* expression patterns presents a distinct advantage of ORIOS against its competitors. Specifically, ORIOS detects 2618 possible new circadian genes in mouse liver, 1188 in mouse pituitary, 781 in the NIH3T3 mouse cell lines and 452 in the U2OS human cell lines. It is important to note that, until now, those genes had not been considered as such. Full detailed results are given in the Results Section of Larriba et al. (2016). The R-based software for ORIOS algorithm developed in this paper is fully functional on `yolandalarriba.wixsite.com/myresearchsite/software-1`.

A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. Larriba et al. (2018)

Methods proposed in Larriba et al. (2016) addressed rhythmicity detection, the key problem in chronobiology, overcoming challenges (a), (b) and (c). Yet, as mentioned in challenge (d), chronobiological rhythms are intrinsically noisy in nature (Tu et al. (2002), Klebanov and Yakovlev (2007)), specifically, those obtained from high-throughput technologies, such as the four gene expression data sets analysed in Larriba et al. (2016) which are derived from microarray technologies (Irizarry et al. (2003a)). Accordingly, these raw microarray data need to be preprocessed before being formally analysed.

Preprocessing methods in literature (MBEI (Li and Wong (2001)), MAS5.0 (Hubbell et al. (2002), Liu et al. (2003)), RMA (Irizarry et al. (2003b))) usually consists of three steps, namely background correction, normalization and summarization. Normalization is an important component of preprocessing since it removes technical variation from expression data (Bolstad et al. (2003), Cheng et al. (2016)). There is no universally accepted normalization strategy and each of them is based on certain model assumptions. Larriba et al. (2018) focuses on seven popular normalization methods called *Quantile, (Cyclic) Loess, Contrast, Constant, Invariant Set, Qspline* and *Variance Stabilization Normalization (VSN)*, see Gautier et al. (2004) for details. In Larriba et al. (2018), it was observed that the choice of normalization may substantially impact on the determination of a gene as rhythmic. To overcome this problem, Larriba et al. (2018) develops a bootstrap-based rhythmicity measure that is robust to the normalization choice. A byproduct of the methodology is that it can be used as a tool for simulating realistic microarray experiments from a reference data set.

Microarray gene expression data are intensity measurements. Genes in the genome are represented in a microarray chip by different DNA probes. RNA samples are tagged with fluorescence molecules which hybridize with the DNA probes in the chip. The intensity of these bindings is measured on the tagged fluorescence molecules using a laser which provides gene expression values. This complex process involves several sources of noise. Data preprocessing is required to remove non-biological noise (normalization step) and to take data from probe level to single expression value (summarization step).

Preprocessing can be outlined as follows $[\boldsymbol{R}] \rightarrow [\boldsymbol{Z}] \rightarrow [\boldsymbol{W}] \rightarrow [\boldsymbol{X}]$, where $[\boldsymbol{R}]$ is the tri-dimensional array, expressed at probe level, of raw intensities; $[\boldsymbol{Z}]$ is that of corrected intensities; $[\boldsymbol{W}]$ denotes that of normalized intensities and $[\boldsymbol{X}]$ is the output gene expression matrix. In particular, background correction is a preliminary filter step, that it is usually conducted in microarray technology, with no impact on subsequent analyses. Normalization reduces systematic variation across arrays, so that biological differences are more easily detected. Summarization provides single gene expression values from probe level intensities. To do so, a two-way linear model on $[\boldsymbol{W}]$, involving probe and array (time points) effects, is assumed. Estimates from array effects correspond to gene expression data $[\boldsymbol{X}]$. Note that the expression matrices $[\boldsymbol{X}]$ are those used in Larriba et al. (2016) to detect circadian genes. However, the performance of rhythmicity detection algorithms (ORIOS, JTK, RAIN) on these data sets potentially depends upon, among other factors, the normalization choice. For instance, the mouse liver gene *Serpina3k* is declared as rhythmic (from BH P-values) by ORIOS when *Quantile* normalization is conducted, while it is non-rhythmic if the *Constant* strategy is followed, see Figure 1.3. Similar results are given by JTK or RAIN.

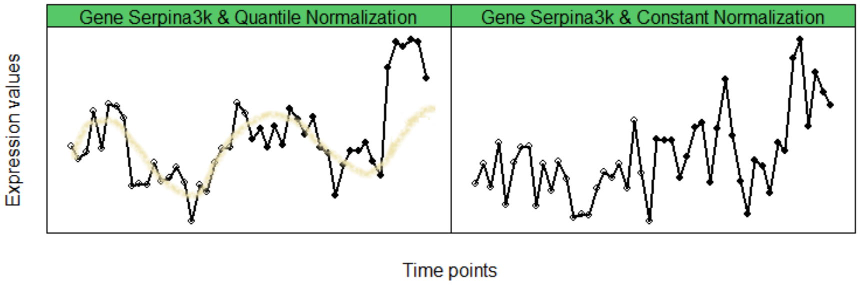


Figure 1.3: *Serpina3k* time course gene expression data regarding *Quantile* (left) and *Constant* (right) normalization methods

Given a normalization method m and a rhythmicity detection algorithm a, as a first attempt to assess normalization choice effects, Larriba et al. (2018)

defines the standard rhythmicity measure as follows:

$$M^g(m,a) = 1 - \text{P-value}^g(m,a),$$

where $\text{P-value}^g(m,a)$ denotes BH P-value of gene g, for $g = 1,\ldots,G$. In vector notation, we write $\boldsymbol{M}(m,a) = [M^1(m,a),\ldots,M^G(m,a)]$, whose components take values from 0 to 1. Values closer to 0 indicate potentially nonrhythmic gene and values closer 1 indicate potentially rhythmic gene. For instance, in the above example $M^{Serpina3k}(\mathit{Quantile}, ORIOS) = 0.996$ and $M^{Serpina3k}(\mathit{Constant}, ORIOS) = 0.639$, thus implying *Serpina3k* is potentially rhythmic ($M^g(m,a) > 0.99$) under *Quantile* normalization but not under *Constant*. To overcome this sort of discrepancy, Larriba et al. (2018) proposes a bootstrap-based methodology, based on a linear model, to define a robust rhythmicity measure ($\boldsymbol{M}_{Robust}$) as follows.

Consider B bootstrap replications. For each replication b ($b = 1,\ldots,B$), background corrected intensities $[\boldsymbol{Z}^{(b)*}]$ are generated according to parametric bootstrap considering the model:

$$\log_2(Z_{pi}^{(b)g*}) = \hat{\alpha}_p^g + \hat{\beta}_i^g + \epsilon_{pi}^{(b)g*},$$

where $\{\hat{\alpha}_p^g\}_{g=1}^G$ and $\{\hat{\beta}_i^g\}_{g=1}^G$ are the original estimates of probe and array effects obtained from $[\boldsymbol{Z}]$, $\epsilon_{pi}^{(b)g*} \overset{i.i.d.}{\sim} N(0, \hat{\sigma}^{2g})$ and $\hat{\sigma}^{2g}$ is the usual MSE (mean squared error) under the original 2-way model, for $i = 1\ldots,n$, $g = 1,\ldots,G$, and $p = 1,\ldots,P$, where P denotes the number of probes. Note that bootstrap generates data at probe level and mimics realistic characteristics from reference set as well as non-specific noise from probe distribution. Hence, given a reference data set, this procedure can be used as a reasonable tool to simulate microarray experiments.

Given $[\boldsymbol{Z}^{(b)*}]$, normalization and summarization steps are performed to obtain replicated gene expression matrices $[\boldsymbol{X}^{(b)*}]$, for $b = 1,\ldots,B$. Rhythmicity statistics are computed on these gene expression matrices to derive the robust rhythmicity measure $\mathbf{M}_{Robust}$.
For a rhythmicity detection algorithm a, a normalization strategy m and a random realization of data, consider the rhythmicity statistic $\mathbf{M}(m,a)$. Let $\boldsymbol{\theta}(m,a) = \mathbb{E}(\mathbf{M}(m,a))$ be the parameter of interest and $\hat{\boldsymbol{\theta}}(m,a) = \mathbf{M}(m,a)$ be its estimator.
For the b^{th} bootstrap sample, let $\hat{\boldsymbol{\theta}}^{(b)*}(m,a) = (\hat{\theta}^{1(b)*}(m,a),\ldots,\hat{\theta}^{G(b)*}(m,a))$ denote the bootstrap estimate of $\boldsymbol{\theta}(m,a)$ computed from $[\boldsymbol{X}^{(b)*}]$, $b = 1,\ldots,B$. Let $\widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(m,a)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(m,a))$ and
$\widehat{RMSE}(\hat{\boldsymbol{\theta}}(m,a)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(m,a) - \hat{\boldsymbol{\theta}}(m,a))^2}$. Then, $\mathbf{M}_{\text{Robust}}$ is defined as

$$\mathbf{M}_{\text{Robust}}(m,a) = \widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(m,a)) - \widehat{RMSE}(\hat{\boldsymbol{\theta}}(m,a)).$$

This value is said to be a “robust" measure of gene rhythmicity because by correcting for sample to sample variation in the rhythmicity measure, i.e. $RMSE$, the effect of the normalization method used is reduced.

In order to show that $\mathbf{M}_{\mathrm{Robust}}(m, a)$ is generally robust with respect to the normalization methods, we computed the Spearman and Pearson correlation coefficients between $\mathbf{M}_{\mathrm{Robust}}(m_r, a)$ and $\mathbf{M}_{\mathrm{Robust}}(m_s, a)$, for all pairs of normalization methods m_r, m_s, $r \neq s$ and compared the correlations with those corresponding to the standard measure $\mathbf{M}(m, a)$.

In Larriba et al. (2018), we analysed the mouse liver, pituitary and NIH3T3 cell lines data sets that were described in Larriba et al. (2016). The analyses are limited to only those genes that were considered to be rhythmic by the criterion $M^g(m, a) \geq 0.99$ for at least one normalization method m, $B = 100$ and $a = ORIOS$ where $m = \{$*Quantile, (Cyclic) Loess, Contrast, Constant, Invariant Set, Qspline, VSN*$\}$. Focusing on mouse liver data set, the percentage of common rhythmic genes for all normalizations increases from 24.39% (w.r.t $\boldsymbol{M}(m, a) \geq 0.99$)) to 48.61% (w.r.t $\boldsymbol{M}_{\mathrm{Robust}}(m, a) \geq 0.99$) and pairwise correlation among the normalization methods improves dramatically using the proposed methodology. For instance, the Spearman correlation coefficient between *Qspline* and *VSN* increases from 0.65 to 0.95. The increase is even more dramatic regarding Pearson correlation coefficient which nearly triples from 0.31 to 0.91. Consequently, using the robust measure defined here, if a gene has a high rank of rhythmicity under one normalization method, then it is also expected to have a similarly high rhythmicity rank under other normalization method. In the motivation example, *Serpina3k* is finally declared as non-rhythmic regarding the values from $\mathbf{M}_{Robust}$ for the normalization methods considered.

Microarray data normalization and robust detection of rhythmic features. Larriba et al. (2019)

In addition to these findings, Larriba et al. (2019) broadens the above results to U2OS human cell lines. Again, it is demonstrated that rhythmicity rankings are correlated across normalization methods considered after bootstrapping. In particular, Larriba et al. (2019) considers ten genes appearing in the first positions of the rhythmicity ranks under several normalization methods. Unlike what happens for the rank of these 10 genes regarding the standard rhythmicity measure, after bootstrapping all of them appear on very good rank positions, see Table 1.2. Full detailed results are given in the Results Sections of Larriba et al. (2018) and Larriba et al. (2019). The R-based microarray simulator software, which relies on the bootstrap methodology developed in this paper, is publicly available on `yolandalarriba.wixsite.com/myresearchsite/software-1`.

Table 1.2: Standard (left side) and robust (right side) rhythmicity measures for 10 genes in the first rhythmicity rank positions

	$\mathbf{M}(m, ORIOS)$						$\mathbf{M}_{Robust}(m, ORIOS)$					
Genes	*Quant.*	*Loess*	*Contr.*	*Inv. Set*	*Qspl.*	*VSN*	*Quant.*	*Loess*	*Contr.*	*Inv. Set*	*Qspl.*	*VSN*
7894590	133	106	283	98	35	15	7	6	6	9	6	6
Per3	9	4	6	11	14	107	3	3	4	3	2	11
Arntl	2	1	1	1	2	3	1	1	1	1	1	1
7973867	105	139	348	101	102	57	27	24	18	25	29	24
Pdpk1	1175	1382	2373	1019	1181	1146	15	18	17	11	15	15
8013519	749	360	357	637	781	381	6	8	8	23	7	5
Rnu2-1	1153	1311	1689	998	1199	967	33	25	11	22	32	17
Ccdc74a	1460	818	360	1236	1713	156	9	10	9	8	12	7
Nr1d2	1	2	3	2	1	2	2	2	2	2	3	3
8180318	276	299	98	236	526	53	4	5	3	5	5	4

Order restricted inference in chronobiology. Larriba et al. (2020)

Larriba et al. (2020) is the most relevant contribution of this thesis. Going an step further, it states a general statistical framework, based on ORI methodology, to formulate, interpret and solve crucial rhythmicity related issues in chronobiology. This ORI-based approach is broadly applicable to different oscillatory systems such as cell cycle, postmortem RNA-Seq gene expression data or metabolic cycle, among others.

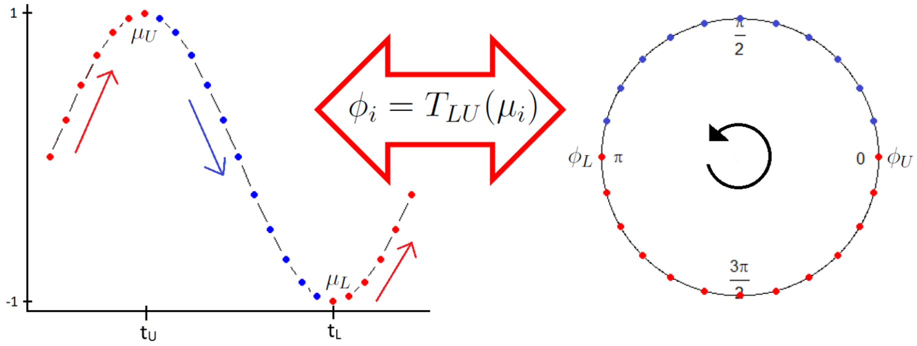
The ORI methodology developed here is underpinned by what we call *circular signal*. *Circular signals* ($\boldsymbol{\mu}$) are rhythmic signals equivalently formulated both in the Euclidean and in the circular space using order restrictions. In particular, *cyclic signals*, proposed in Larriba et al. (2016), are the Euclidean space representation of *circular signals*. Larriba et al. (2020) states its equivalent circular ordered representation ($\boldsymbol{\phi}$) in the circular space. A signal $\boldsymbol{\phi}$ in the circular space is said to be circular ordered iff $\boldsymbol{\phi} \in C_o = \{\boldsymbol{\phi} \in [0, 2\pi)^n : \phi_1 \preceq \cdots \preceq \phi_n \preceq \phi_1\}$, where $\preceq$ can be read as "is followed by". If $\boldsymbol{\phi} \in C_o$, then $\boldsymbol{\phi}$ is said to follow the circular order o, see Fernández et al. (2012) or Barragán et al. (2017) for details. *Circular signals* are equivalently formulated both in the Euclidean and in the circular space throughout the mapping T_{LU}, see Figure 1.4. This equivalence between the order in both spaces allows us to obtain better insights to the problems, as in the case of the temporal order estimation.

Consequently, the model proposed in Larriba et al. (2016) is generalized to the following *circular signal* plus error model:

$$\boldsymbol{X}_j = \boldsymbol{\mu} + \boldsymbol{\epsilon}_j,$$

where $\boldsymbol{\mu}$ is a *circular signal*. No distributional assumptions are needed at this moment.

Circular signal estimation problem is solved as the following mean squares



$$\phi_i = T_{LU}(\mu_i) = \begin{cases} \arcsin(\mu_i) - \frac{\pi}{2} & \text{, if } i \in \{1,\dots,U\} \cup \{L,\dots,n\} \\ \frac{\pi}{2} - \arcsin(\mu_i) & \text{, otherwise} \end{cases}$$

Figure 1.4: Equivalent formulation of *circular signal*. Left: Euclidean space. Right: circular space

optimization problem

$$\boldsymbol{Y}^\star = \arg\min_{\boldsymbol{Z}\in C} \sum_{i=1}^{n} (Y_i - Z_i)^2,$$

where $C = \bigcup_{LU} C_{LU}$ with $L, U \in \{1,\dots,n\}$ and $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \leq \cdots \leq \mu_U \geq \cdots \geq \mu_L \leq \cdots \leq \mu_n \leq \mu_1\}$ where without loss of generality we assume $L > U$. $\boldsymbol{Y}^\star$ is called the Isotonic Regression (IR) estimator of $\boldsymbol{Y}$ with respect to C. Given a model that incorporates restrictions on its parameters, as in this case, IR provides the best model-fitting by solving a least squared optimization problem. Since the order defined by C is not a closed convex cone, the derivation of $\boldsymbol{Y}^\star$ is non-trivial (Robertson et al. (1988)). In Larriba et al. (2020), a computationally efficient IR-based algorithm is designed to derive $\boldsymbol{Y}^\star$. Point estimators for t_U and t_L can be immediately derived from that algorithm as the point estimation of the indexes U and L.

The estimation of the *circular signal* provided by IR is the key in the solution of many of the rhythmicity related problems. It is well-known that under normality assumption, i.e. $\boldsymbol{\epsilon}_j \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$, $\boldsymbol{Y}^\star$ is the maximum likelihood estimator (MLE) of the *circular signal*, $\boldsymbol{\mu}$ (Robertson et al. (1988), Silvapulle and Sen (2005)). Moreover, under this assumption, the rhythmicity detection problem (1.1) is efficiently addressed as the following *circular signal* testing problem:

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_n \\ H_1 &: \boldsymbol{\mu} \in C. \end{aligned} \tag{1.2}$$

The standard ORI theory is not applicable directly because $C = \bigcup_{LU} C_{LU}$ in (1.2) is not a convex cone. Hence, this testing problem is conducted in a two-step approach. First, we estimate L and U by $L^\star$ and $U^\star$ using the IR algorithm. Next, the testing problem is solved assuming L and U to be known and using CTs and $\boldsymbol{Y}^\star$, the MLE of *circular signals*, which notably increases the efficiency of this procedure.

As mentioned in challenge (e), in some applications the temporal order among biological samples may be unknown, and one needs to estimate these times before analysing rhythmicity, see Lamb et al. (2011), Bossé et al. (2012), Li et al. (2013), Mure et al. (2018) or Seney et al. (2019) among others. The equivalent formulation of rhythmicity between the Euclidean and circular spaces provides a better insight of this problem. Temporal order estimation relies on the IR estimator of *circular signal* $\boldsymbol{Y}^\star$. Let $\mathcal{D} = \{\boldsymbol{Y}^g\}_{g=1}^G$ be the observed gene expression data set and $\boldsymbol{Y}_o^{\star g} = (Y_{o,i}^{\star g})_{i=1}^n$ denote the IR of $\boldsymbol{Y}_o^g = (Y_{o,1}^g, \ldots, Y_{o,n}^g)'$ under the *circular signal* model that generates the order o. In the temporal order estimation problem, we look for the order among time points that better reconstructs the underlying rhythmic pattern in the data base. As *circular signals* are equivalently formulated both in the Euclidean and in the circular space, this problem is reduced to that of looking for the circular order (permutation) among the indexes of the time points that is closest to the data. This problem is mathematically formulated as follows:

$$\arg\min_{o \in \Pi} d(\mathcal{D}, o), \tag{1.3}$$

where Π is the set of all possible circular orderings of all indexes $S = \{1, \ldots, n\}$ around the unit circle and $d(\mathcal{D}, o) = \sum_{g=1}^G \sum_{i=1}^n \nu_g (Y_i^g - Y_{o,i}^{\star g})^2$, is a measure of distance between o and $\mathcal{D}$ with ν_g being a positive weight. Note that since $\#\Pi = (n-1)!$, (1.3) is a NP-hard problem (Bartholdi III et al. (1989)). ORI methodology provides an approximate solution to (1.3) by formulating it as a traveling salesman problem (TSP) (Flood (1956), Lawler et al. (1985), Reinelt (1994)). For each gene, data are represented by a weighted directed graph where the nodes represent the indexes of the time points to be ordered. Each pair of nodes is connected by an edge whose length represents the intensity of the relationship, i.e. the distance among each pair of nodes in the graph. Therefore, the problem is reduced to find the tour that goes exactly once through all nodes in the graph, starting and ending at the same node. The tour that minimizes the total length is the solution of the well-known TSP. Altough TSP has not an exact solution, there are several heuristic procedures, many of them implemented in R, which provide accurate approximate solutions to this problem (Hahsler and Hornik (2007)). Moreover, relative efficiency rates of agreement (RRE), between a circular order and a data set, and concordance (CRE), between two circular orders are defined in Larriba et al. (2020) to validate the procedure. These measures are defined in terms of IR and MSE and they resemble those used in linear regression. RRE and CRE are positive and easily interpretable, the lower the values the more reliable the order reconstruction.

The ORI methodology proposed outperforms classical methods in chronobiology for signal estimation, rhythmicity detection and temporal order estimation. In simulation settings, IR peak and signal estimates dramatically improve Cosinor ones, with lower MSE values, when the signal is not exactly sinusoidal. ORI outperforms JTK detecting asymmetric rhythmic patterns and controls missclassification rates. Finally, the reliability of the temporal order reconstruction is assessed in terms of RRE and CRE. The values of these measures are about two or three times higher for CYCLOPS than for ORI-based solutions.

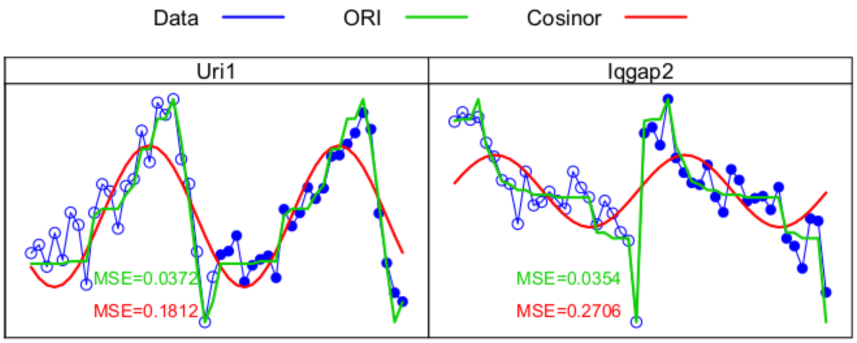


Figure 1.5: ORI (green) and Cosinor (red) model-fittings and MSE values for *Uri1* (left) and *Iqgap2* (right) gene expression data (blue)

When considering real data bases, impressive findings are obtained. In Figure 1.5, we can see that ORI procedure correctly identifies circadian genes with very asymmetric patterns, has lower MSE than Cosinor and provides accurate peak estimates. In addition to this, in Figure 1.6, we illustrate gene expression patterns for three well-known core clock circadian genes, called *Per2*, *Per3* and *Tspan8,* plotted according to three orders: the REAL, at which data were obtained; and the two estimated orders derived from ORI and CYCLOPS procedures. Unlike what happens for CYCLOPS, ORI reordered expressions sustain the rhithmicity nature of such core clock genes. Even more, CYCLOPS reordered expressions might be identified as non-rhythmic ones which may lead biologists to wrong conclusions. Full detailed results are given in the Result Section of Larriba et al. (2020). The R-based software developed to perform the analyses in this paper is available upon request since a full R-package including, among others the ORI-based methodology described here, is being developed by the authors.

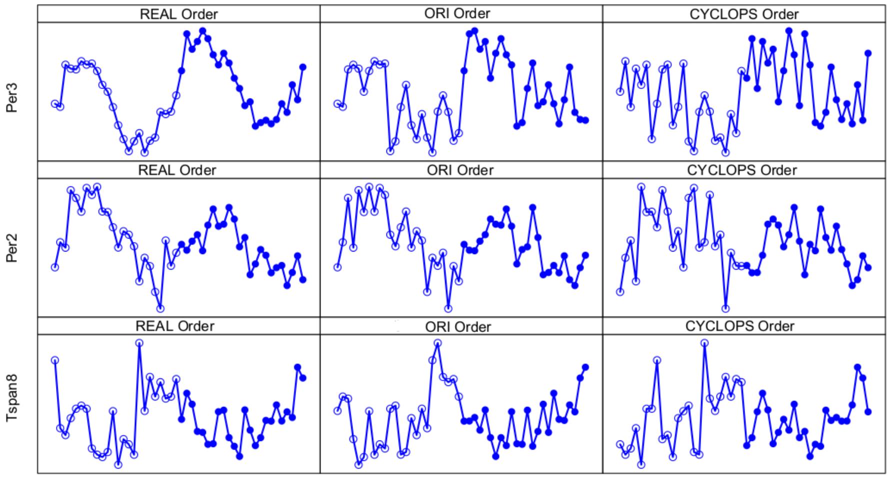


Figure 1.6: Core clock genes *Per3*, *Per2* and *Tspan8* from NIH3T3 plotted under REAL (left), ORI (middle) and CYCLOPS (right) orders

In conclussion, Larriba et al. (2020) provides a general ORI-based framework for analysing rhythmic patterns from oscillatory systems. The proposed methodology provides solutions to a wide range of problems associated with the analysis of rhythmic data such as rhythmicity detection, order reconstruction and peak time estimation. There are several advantages in using the proposed methodology. First of all, the methodology is simple to describe and use. Secondly, the equivalence between the order in the Euclidean and circular spaces makes easy to translate between both spaces and obtain a better insights to the problems. Moreover, the methodology is very flexible. The formulation does not require a rigid mathematical function to describe a rhythmic pattern. It is all done through mathematical inequalities. Rigorous mathematical formulation, which allows a deeper study of the methodology and its properties is another advantage of the methodology. Finally, ORI methodology is computationally efficient solving all the rhythmicity problems described in this work and it is broadly applicable to other different oscillatory systems.

Conclusions

This section summarizes the major contributions of this thesis and provides a brief remark about the present and future research to be developed from this thesis.

From a statistical point of view, this thesis lays the foundations of a novel and broadly applicable ORI methodology to formulate, interpret and accurately solve a wide range of problems in chronobiology outperforming recent methods in literature. As an added value, the bootstrap methodology developed here can be used as a tool to simulate microarray gene expression data that mimic realistic features.

From a biological point of view, it is very remarkable that the rhythmicity analysis methods proposed in this thesis allow us to identify new circadian genes and provide accurate peak and signal estimates. This is crucial in, among others, early cancer diagnosis or neurodegenerative disorders. Focusing on mouse liver data set, there are 2618 new circadian genes which have not been considered as such until now. Many of them display *quasi-cyclic* or markedly asymmetric gene expression patterns, such as *Iqgap2*, (see right panel in Figure 1.5). In particular, *Iqgap2* is a circadian gene involved in ovarian cancer detection and reveals a difference of ~ 4 hours in peak time estimation, compared to standard methods employed in practice, which may be crucial for biologists. Even more, these findings are consistent in the case of human experiments (U2OS cell lines) where rhythmicity rate is much lower ($\sim 3\%$) than it is for other data sets like mouse liver ($\sim 20\%$) identifying well-know rhythmic genes such as *Arntl, Per3* or *Nr1d2* among the 10 circadian genes with the highest robust rhythmicity measures.

The methodology developed in this thesis is the starting point to propose new models and to design answers to specific problems posed in different fields that share the fundamental characteristic of dealing with outcomes generated by oscillatory processes and the challenge raised by modelling in the presence of noise. Statistical methods are necessary to understand the systems, to predict their evolution, to find out how these systems interact with others and to generate new experiments, hypotheses and specific statistical procedures, required by the diverse problems that arise in those different fields. Specifically, the methodology has already been extended to solve problems in astrophysics and biology (Rueda et al. (2019)). Moreover, in addition to more interesting chronobiological problems, other very interesting applications are emerging. A particularly interesting application on which we are hard working at this moment is related to electrophysiology. We think that the results will have an immediate application in human health, as we expect to provide advances in the automatic diagnosis of cardiovascular diseases that should lead to a reduction in mortality from those causes.

2. Publications

This section provides the main bibliographic data of the three articles and the book chapter that are part of this thesis. The three articles of this compendium have been published in JCR-indexed peer-reviewed journals of the first quartile of their category. The book chapter is included in a book series by Springer Nature that appears in the SJR Scopus index list. The contributions are listed acording to publication date.

Larriba et al. (2016)

Manuscript type:	Article
Title:	Order restricted inference for oscillatory systems for detecting rhythmic signals
Authors:	Yolanda Larriba, Cristina Rueda, Miguel A. Fernández and Shyamal D. Peddada
DOI:	`10.1093/nar/gkw771`
Abstract:	Many biological processes, such as cell cycle, circadian clock, menstrual cycles, are governed by oscillatory systems consisting of numerous components that exhibit rhythmic patterns over time. It is not always easy to identify such rhythmic components. For example, it is a challenging problem to identify circadian genes in a given tissue using time-course gene expression data. There is a great potential for misclassifying non-rhythmic as rhythmic genes and vice versa. This has been a problem of considerable interest in recent years. In this article we develop a constrained inference based methodology called Order Restricted Inference for Oscillatory Systems (ORIOS) to detect rhythmic signals. Instead of using mathematical functions (e.g. sinusoidal) to describe shape of rhythmic signals, ORIOS uses mathematical inequalities. Consequently, it is robust and not limited by the biologist's choice of the mathematical model. We studied the performance of ORIOS using simulated as well as real data obtained from mouse liver, pituitary gland and data from NIH3T3, U2OS cell lines. Our results suggest that, for a broad collection of patterns of gene expression, ORIOS has substantially higher power to detect true rhythmic genes in comparison to some popular methods, while also declaring substantially fewer non-rhythmic genes as rhythmic. A user friendly code implemented in R language can be downloaded from `yolandalarriba.wixsite.com/myresearchsite/software-1`.
Reference:	Larriba, Y., Rueda, C., Fernández, M. A., & Peddada, S. D. (2016). Order restricted inference for oscillatory systems for detecting rhythmic signals. Nucleic Acids Research, 44(22), e163.

Larriba et al. (2018)

Manuscript type:	Article
Title:	A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data
Authors:	Yolanda Larriba, Cristina Rueda, Miguel A. Fernández and Shyamal D. Peddada
DOI:	`10.3389/fgene.2018.00024`
Abstract:	Gene-expression data obtained from high throughput technologies are subject to various sources of noise and accordingly the raw data are pre-processed before formally analyzed. Normalization of the data is a key pre-processing step, since it removes systematic variations across arrays. There are numerous normalization methods available in the literature. Based on our experience, in the context of oscillatory systems, such as cell-cycle, circadian clock, etc., the choice of the normalization method may substantially impact the determination of a gene to be rhythmic. Thus rhythmicity of a gene can purely be an artifact of how the data were normalized. Since the determination of rhythmic genes is an important component of modern toxicological and pharmacological studies, it is important to determine truly rhythmic genes that are robust to the choice of a normalization method. In this paper we introduce a rhythmicity measure and a bootstrap methodology to detect rhythmic genes in an oscillatory system. Although the proposed methodology can be used for any high-throughput gene expression data, in this paper we illustrate the proposed methodology using several publicly available circadian clock microarray gene-expression datasets. We demonstrate that the choice of normalization method has very little effect on the proposed methodology. Specifically, for any pair of normalization methods considered in this paper, the resulting values of the rhythmicity measure are highly correlated. Thus it suggests that the proposed measure is robust to the choice of a normalization method. Consequently, the rhythmicity of a gene is potentially not a mere artifact of the normalization method used. Lastly, as demonstrated in the paper, the proposed bootstrap methodology can also be used for simulating data for genes participating in an oscillatory system using a reference dataset. A user friendly code implemented in R language can be downloaded from `yolandalarriba.wixsite.com/myresearchsite/software-1`
Reference:	Larriba, Y., Rueda, C., Fernández, M. A., & Peddada, S. D. (2018). A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. Frontiers in Genetics, 9, 24.

Larriba et al. (2019)

Manuscript type:	Book chapter
Title:	Microarray data normalization and robust detection of rhythmic features
Authors:	Yolanda Larriba, Cristina Rueda, Miguel A. Fernández and Shyamal D. Peddada
DOI:	10.1007/978-1-4939-9442-7_9
Abstract:	Data derived from microarray technologies are generally subject to various sources of noise and accordingly the raw data are pre-processed before formally analysed. Data normalization is a key pre-processing step when dealing with microarray experiments, such as circadian gene-expressions, since it removes systematic variations across arrays. A wide variety of normalization methods are available in the literature. However, from our experience in the study of rhythmic expression patterns in oscillatory systems (e.g. cell-cycle, circadian clock), the choice of the normalization method may substantially impair the identification of rhythmic genes. Hence, the identification of a gene as rhythmic could be just as an artefact of how the data were normalized. Yet, gene rhythmicity detection is crucial in modern toxicological and pharmacological studies, thus a procedure to truly identify rhythmic genes that are robust to the choice of a normalization method is required. To perform the task of detecting rhythmic features, we propose a rhythmicity measure based on bootstrap methodology to robustly identify rhythmic genes in oscillatory systems. Although our methodology can be extended to any high-throughput experiment, in this chapter, we illustrate how to apply it to a publicly available circadian clock microarray gene-expression data and give full details (both statistical and computational) so that the methodology can be used in an easy way. We will show that the choice of normalization method has very little effect on the proposed methodology since the results derived from the bootstrap-based rhythmicity measure are highly rank correlated for any pair of normalization methods considered. This suggests, on the one hand, that the rhythmicity measure proposed is robust to the choice of the normalization method, and on the other hand, that gene rhythmicity detected using this measure is potentially not a mere artefact of the normalization method used. In this way the researcher using this methodology will be protected against the possible effect of different normalizations, as the conclusions obtained will not depend so strongly on them. Additionally, the described bootstrap methodology can also be employed as a tool to simulate gene-expression participating in an oscillatory system from a reference data set.
Reference:	Larriba, Y., Rueda, C., Fernández, M. A., & Peddada, S. D. (2019). Microarray data normalization and robust detection of rhythmic features. In: Bolón-Canedo V., Alonso-Betanzos A. (eds) Microarray Bioinformatics. Methods in Molecular Biology. Humana, New York, NY.

Larriba et al. (2020)

Manuscript type:	Article
Title:	Order restricted inference in chronobiology
Authors:	Yolanda Larriba, Cristina Rueda, Miguel A. Fernández and Shyamal D. Peddada
DOI:	10.1002/sim.8397
Abstract:	This paper is motivated by applications in oscillatory systems where researchers are typically interested in discovering components of those systems that display rhythmic temporal patterns. The contributions of this paper are twofold. First, a methodology is developed based on a circular signal plus error model that is defined using order restrictions. This mathematical formulation of rhythmicity is simple, easily interpretable and very flexible, with the latter property derived from the nonparametric formulation of the signal. Second, we address various commonly encountered problems in the analysis of oscillatory systems data. Specifically, we propose a methodology for (a) detecting rhythmic signals in an oscillatory system and (b) estimating the unknown sampling time that occurs when tissues are obtained from subjects whose time of death is unknown. The proposed methodology is computationally efficient, outperforms the existing methods, and is broadly applicable to address a wide range of questions related to oscillatory systems.
Reference:	Larriba, Y., Rueda, C., Fernández, M. A., & Peddada, S. D. (2020). Order restricted inference in chronobiology. Statistics in Medicine. 39, 265–278.

Bibliography

R. Anafi, L. Francey, J. Hogenesch, and J. Kim. Cyclops reveals human transcriptional rhythms in health and disease. *Proceedings of the National Academy of Sciences of the United States of America*, 114(20):5312–5317, 2017. doi: 10.1073/pnas.1619320114.

T. Andreani, T. Itoh, E. Yildirim, D. Hwangbo, and R. Allada. Genetics of circadian rhythms. *Sleep medicine clinics*, 10(4), 2015. doi: 10.1016/j.jsmc.2015.08.007.

R. Barlow, D. Bartholomew, J. Bremner, and H. D. Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York, 1972.

S. Barragán, C. Rueda, M. Fernández, and S. Peddada. Determination of temporal order among the components of an oscillatory system. *PLoS ONE*, 10 (7):1–14, 2015. doi: 10.1371/journal.pone.0124842.

S. Barragán, C. Rueda, and M. Fernández. Circular order aggregation and its application to cell-cycle genes expressions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(4):819–829, 2017. doi: 10.1109/TCBB.2016.2565469.

J. Bartholdi III, C. Tovey, and M. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241, 1989. doi: 10.1007/BF00295861.

D. Bartholomew. A test of homogeneity for ordered alternatives. ii. *Biometrika*, 46(3-4):328–335, 1959.

D. Bartholomew. A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 23(2):239–281, 1961.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

B. Bolstad, R. Irizarry, M. Åstrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. doi: 10.1093/bioinformatics/19.2.185.

Y. Bossé, D. Postma, D. Sin, M. Lamontagne, C. Couture, N. Gaudreault, P. Joubert, V. Wong, M. Elliott, M. Van Den Berge, C. Brandsma, C. Tribouley, V. Malkov, J. Tsou, G. Opiteck, J. Hogg, A. Sandford, W. Timens, P. Paré, and M. Laviolette. Molecular signature of smoking in human lung tissues. *Cancer Research*, 72(15):3753–3763, 2012. doi: 10.1158/0008-5472.CAN-12-1160.

H. Brunk. Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):607–616, 1955. doi: 10.1214/aoms/1177728420.

M. Caba, G. González-Mariscal, and E. Meza. Circadian rhythms and clock genes in reproduction: Insights from behavior and the female rabbit's brain. *Frontiers in Endocrinology*, 9:106, 2018. doi: 10.3389/fendo.2018.00106.

S. Chan, L. Zhang, L. Rowbottom, R. McDonald, G. Bjarnason, M. Tsao, E. Barnes, C. Danjoux, M. Popovic, H. Lam, C. DeAngelis, and E. Chow. Effects of circadian rhythms and treatment times on the response of radiotherapy for painful bone metastases. *Annals of Palliative Medicine*, 6(1):14–25, 2017. doi: 10.21037/apm.2016.09.07.

R. Chauhan, K. Chen, B. Kent, and D. Crowther. Central and peripheral circadian clocks and their role in alzheimer's disease. *Disease Models & Mechanisms*, 10(10):1187–1199, 2017. doi: 10.1242/dmm.030627.

L. Cheng, L.-Y. Lo, N. Tang, D. Wang, and K.-S. Leung. Crossnorm: a novel normalization strategy for microarray data in cancers. *Scientific Reports*, 6:18898, 2016. doi: 10.1038/srep18898.

G. Cornelissen. Cosinor-based rhythmometry. *Theoretical biology & medical modelling*, 11:16, 2014. doi: 10.1186/1742-4682-11-16.

C. Draper, K. Duisters, B. Weger, A. Chakrabarti, A. Harms, L. Brennan, L. Goulet, T. Konz, F. Martin, S. Moco, and J. van der Greef. Menstrual cycle rhythmicity: metabolic patterns in healthy women. *Scientific Reports*, 8:14568, 2018. doi: 10.1038/s41598-018-32647-0.

N. Elkum and J. Myles. Modeling biological rhythms in failure time data. *Journal of Circadian Rhythms*, 4:14, 2006. doi: 10.1186/1740-3391-4-14.

M. Fernández, C. Rueda, and S. 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.

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

M. Flood. The traveling-salesman problem. *Operations Research*, 4(1):61–75, 1956.

L. Gautier, L. Cope, B. Bolstad, and R. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. doi: 10.1093/bioinformatics/btg405.

M. Hahsler and K. Hornik. Tsp – Infrastructure for the traveling salesperson problem. *Journal of Statistical Software*, 23(2):1–21, 2007. doi: 10.18637/jss.v023.i02.

F. Halberg, D. Powell, K. Otsuka, Y. Watanabe, L. Beaty, P. Rosch, J. Czaplicki, D. Hillman, O. Schwartzkopff, and G. Cornelissen. Diagnosing vascular variability anomalies, not only mesor-hypertension. *American Journal of Physiology-Heart and Circulatory Physiology*, 305(3):H279–H294, 2013. doi: 10.1152/ajpheart.00212.2013.

E. Haus. Chronobiology in oncology. *International Journal of Radiation Oncology Biology Physics*, 73(1):3–5, 2009. doi: 10.1016/j.ijrobp.2008.08.045.

E. Hubbell, W.-M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, 2002. doi: 10.1093/bioinformatics/18.12.1585.

M. Hughes, L. Deharo, S. Pulivarthy, J. Gu, K. Hayes, S. Panda, and J. Hogenesch. High-resolution time course analysis of gene expression from pituitary. *Cold Spring Harbor Symposia on Quantitative Biology*, 72:381–386, 2007. doi: 10.1101/sqb.2007.72.047.

M. Hughes, L. DiTacchio, K. Hayes, C. Vollmers, S. Pulivarthy, J. Baggs, S. Panda, and J. Hogenesch. Harmonics of circadian gene transcription in mammals. *PLoS Genetics*, 5(4):1–12, 2009. doi: 10.1371/journal.pgen.1000442.

M. Hughes, J. 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.

R. Irizarry, B. Bolstad, F. Collin, L. Cope, B. Hobbs, and T. Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003a. doi: 10.1093/nar/gng015.

R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003b. doi: 10.1093/biostatistics/4.2.249.

L. Klebanov and A. Yakovlev. How high is the level of technical noise in microarray data? *Biology Direct*, 2:9, 2007. doi: 10.1186/1745-6150-2-9.

A. Korenčič, G. Bordyugov, R. Košir, D. Rozman, M. Goličnik, and H. Herzel. The interplay of cis-regulatory elements rules circadian rhythms in mouse liver. *PLoS ONE*, 7(11):1–13, 2012. doi: 10.1371/journal.pone.0046835.

J. Lamb, C. Zhang, T. Xie, K. Wang, B. Zhang, K. Hao, E. Chudin, H. Fraser, J. Millstein, M. Ferguson, C. Suver, I. Ivanovska, M. Scott, U. Philippar, D. Bansal, Z. Zhang, J. Burchard, R. Smith, D. Greenawalt, M. Cleary, J. Derry, A. Loboda, J. Watters, R. Poon, S. Fan, C. Yeung, N. Lee, J. Guinney, C. Molony, V. Emilsson, C. Buser-Doepner, J. Zhu, S. Friend, M. Mao, P. Shaw, H. Dai, J. Luk, and E. Schadt. Predictive genes in adjacent normal tissue are preferentially altered by scnv during tumorigenesis in liver cancer and may rate limiting. *PLoS ONE*, 6(7):1–17, 2011. doi: 10.1371/journal.pone.0020090.

Y. Larriba, C. Rueda, M. Fernández, and S. Peddada. Order restricted inference for oscillatory systems for detecting rhythmic signals. *Nucleic Acids Research*, 44(22):e163, 2016. doi: 10.1093/nar/gkw771.

Y. Larriba, C. Rueda, M. Fernández, and S. Peddada. A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. *Frontiers in Genetics*, 9:24, 2018. doi: 10.3389/fgene.2018.00024.

Y. Larriba, C. Rueda, M. Fernández, and S. Peddada. Microarray data normalization and robust detection of rhythmic features. In V. Bolón-Canedo and A. Alonso-Betanzos, editors, *Microarray Bioinformatics*, pages 207–225. Springer New York, 2019. ISBN 978-1-4939-9442-7.

Y. Larriba, C. Rueda, M. Fernández, and S. Peddada. Order restricted inference in chronobiology. *Statistics in Medicine*, 39(3):265–278, 2020. doi: 10.1002/sim.8397.

E. Lawler, J. Lenstra, A. Rinnooy Kan, and D. Shmoys. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, 1985.

T. Leise. Wavelet analysis of circadian and ultradian behavioral rhythms. *Journal of Circadian Rhythms*, 11(1):5, 2013. doi: 10.1186/1740-3391-11-5.

N. Leng, L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. Stewart, J. Thomson, and C. Kendziorski. Oscope identifies oscillatory genes in unsynchronized single-cell rna-seq experiments. *Nature Methods*, 12(10):947–950, 2015. doi: 10.1038/nmeth.3549.

J. Levine, P. Funes, H. Dowse, and J. Hall. Signal analysis of behavioral and molecular cycles. *BMC Neuroscience*, 3:1, 2002. doi: 10.1186/1471-2202-3-1.

C. Ley and T. Verdebout. *Modern directional statistics.* Chapman and Hall/CRC, 2017.

C. Ley and T. Verdebout. *Applied Directional Statistics: Modern Methods and Case Studies.* Chapman and Hall/CRC, 2018.

C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):31–36, 2001. doi: 10.1073/pnas.011404098.

J. Li, B. Bunney, F. Meng, M. Hagenauer, D. Walsh, M. Vawter, S. Evans, P. Choudary, P. Cartagena, J. Barchas, A. Schatzberg, E. Jones, R. Myers, S. Watson Jr., H. Akil, and W. Bunney. Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *Proceedings of the National Academy of Sciences of the United States of America*, 110 (24):9950–9955, 2013. doi: 10.1073/pnas.1305814110.

D. Liu, D. Umbach, S. Peddada, L. Li, P. Crockett, and C. Weinberg. A random-periods model for expression of cell-cycle genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7240–7245, 2004. doi: 10.1073/pnas.0402285101.

G. Liu, A. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. Siani-Rose. Netaffx: Affymetrix probesets and annotations. *Nucleic Acids Research*, 31(1):82–86, 2003. doi: 10.1093/nar/gkg121.

Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O. Aparicio, M. Zhang, R. Jianh, and T. Chen. Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nature Communications*, 8(22), 2017. doi: 10.1038/s41467-017-00039-z.

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

J. Menéndez and B. Salvador. Anomalies of the likelihood ratio tests for testing restricted hypothesis. *The Annals of Statistics*, 19(2):889–898, 1991.

T. Mockler, T. Michael, H. Priest, R. Shen, C. Sullivan, S. Givan, C. McEntee, S. Kay, and J. Chory. The diurnal project: Diurnal and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harbor Symposia on Quantitative Biology*, 72:353–63, 2007. doi: 10.1101/sqb.2007.72.006.

L. Mure, H. Le, G. Benegiamo, M. Chang, L. Rios, N. Jillani, M. Ngotho, T. Kariuki, O. Dkhissi-Benyahya, H. Cooper, and S. Panda. Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science*, 359 (6381):307–315, 2018. doi: 10.1126/science.aao0318.

S. Panda, M. Antoch, B. Miller, A. Su, A. Schook, M. Straume, P. Schultz, S. Kay, J. Takahashi, and J. Hogenesch. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, 109(3):307–320, 2002. doi: 10.1016/S0092-8674(02)00722-5.

G. Reinelt. *The Traveling Salesman. Computational Solutions for TSP Applications.* Springer-Verlag, 1994.

T. Robertson and F. Wright. Algorithms in order restricted statistical inference and the Cauchy mean value property. *The Annals of Statistics*, 8(3):645–651, 1980.

T. Robertson, F. Wright, and R. Dykstra. *Order Restricted Statistical Inference.* John Wiley & Sons, 1988.

C. Rueda, M. Fernández, and S. 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.

C. Rueda, M. Ugarte, and A. Militino. Checking unimodality using isotonic regression: an application to breast cancer mortality rates. *Stochastic Environmental Research and Risk Assessment*, 30(4):1277–1288, 2016. doi: 10.1007/s00477-015-1111-8.

C. Rueda, Y. Larriba, and S. Peddada. Frequency modulated möbius model accurately predicts rhythmic signals in biological and physical sciences. *Scientific Reports*, 9:18701, 2019. doi: 10.1038/s41598-019-54569-1.

M. Seney, K. Cahill, J. Enwright III, R. Logan, Z. Huo, W. Zong, G. Tseng, and C. McClung. Diurnal rhythms in gene expression in the prefrontal cortex in schizophrenia. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-019-11335-1.

M. Silvapulle and P. Sen. *Constrained Statistical Inference: Inequality, Order and Shape Restrictions.* Wiley Series in Probability and Statistics. John Wiley & Sons, 2005.

V. Simonneaux and T. Bahougne. A multi-oscillatory circadian system times female reproduction. *Frontiers in Endocrinology*, 6:157, 2015. doi: 10.3389/fendo.2015.00157.

M. Straume. Dna microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. *Methods Enzymol*, 383:149–166, 2004. doi: 10.1016/S0076-6879(04)83007-6.

P. Thaben and P. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014. doi: 10.1177/0748730414553029.

Y. Tong. Parameter estimation in studying circadian rhythms. *Biometrics*, 32 (1):85–94, 1976.

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22):14031–14036, 2002. doi: 10.1073/pnas.222164199.

C. Van Eeden. Maximum likelihood estimation of ordered probabilities. *Indagationes Mathematicae (Proceedings)*, 59:444 – 455, 1956. ISSN 1385-7258. doi: 10.1016/S1385-7258(56)50060-1.

H. Wang and P.-S. Zhong. Order-restricted inference for means with missing values. *Biometrics*, 73(3):972–980, 2017. doi: 10.1111/biom.12658.

S. Wichert, K. Fonkianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 2004. doi: 10.1093/bioinformatics/btg364.

H. Wijnen, F. Naef, C. Boothroyd, A. Claridge-Chang, and M. Young. Control of daily transcript oscillations in drosophila by light and the circadian clock. *PLoS Genetics*, 2(3):0326–0343, 2006. doi: 10.1371/journal.pgen.0020039.

R. Yang and Z. Su. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics*, 26(12):i168–i174, 2010. doi: 10.1093/bioinformatics/btq189.

R. Zhang, N. Lahens, H. Ballance, M. Hughes, and J. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):16219–16224, 2014. doi: 10.1073/pnas.1408886111.