Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Order Restricted Inference in Chronobiology

Yolanda Larriba[1] | Cristina Rueda[1] | Miguel A. Fernández*[1] | Shyamal D. Peddada[2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain
[2]Department of Biostatistics, Public School of Health, University of Pittsburg, Pittsburgh, USA

Correspondence
*Miguel A. Fernández. Email: miguelaf@eio.uva.es

Summary

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 the 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 non-parametric 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, (b) estimating the unknown sampling time which 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.

KEYWORDS:
Constrained Inference, Circular Data, Rhythmicity Detection, Timing Estimation, Oscillatory Systems

1 | INTRODUCTION

Locomotor activity, blood pressure or body temperature are just a few of the physiological and biological phenomena exhibiting rhythmic or oscillatory processes in nature. Such oscillatory systems contain one or more components that display periodic or rhythmic patterns over each observed period. For example, genes (i.e. the components) participating in a circadian clock often display a rhythmic pattern of expression as shown in Figure 1. The study of such components with temporal rhythmic patterns, and how these patterns change under different conditions, is called chronobiology.[1,2,3,4] Chronobiology has been an active area of research during the past two decades, with major impact on treating cardiovascular disorders like hypertension,[5] detecting genes associated with neurodegenerative disorders[6] or depression,[7] and improving the effectiveness of cancer treatments.[8] For instance, Haus[9] 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, particularly in identifying and/or characterizing rhythmic processes.

Although this article is motivated by gene expression studies associated with circadian clock, the proposed methodology is very general and is broadly applicable to other oscillatory systems such as the cell-cycle,[10,11] the endocrinology,[12] vascular processes,[13] etc. In fact, there is considerable interest in pharmacology, psychiatry and other areas of medical sciences to discover genes belonging to oscillatory systems that have rhythmic temporal expression.

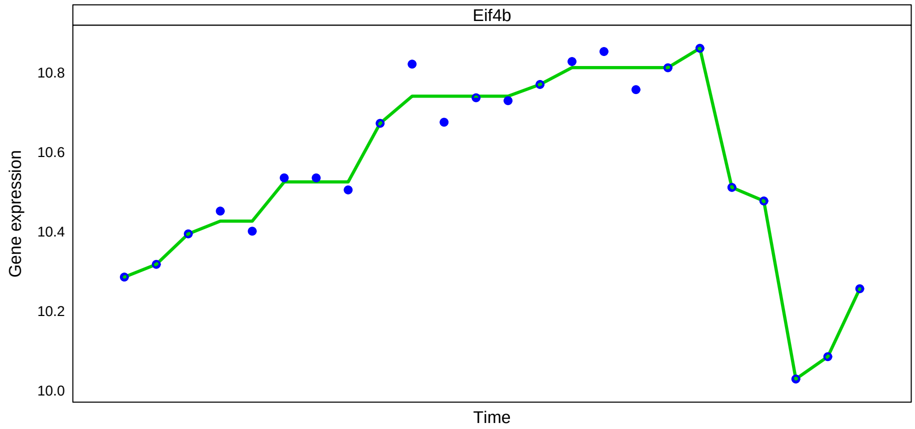


FIGURE 1 Observed gene expression (blue dots) along 24 hours for gene Eif4b. Green line represents the underlying estimated rhythmic signal

From a statistical point of view, the modelling of chronobiological rhythms in biomedical sciences is a challenge because, unlike the time course financial data in the stock market or heart rate data in intensive care units, the density of time points is generally low[14,15] and the number of periods of data is usually very small.[16,17] For these reasons, standard time series or Fourier models are not convenient.[18,19,20] Another challenge with these chronobiological data is that not all components display the same pattern of expression over time.[21,22,23] Despite this heterogeneity, rhythmic patterns of components of an oscillatory system usually display up-down-up patterns. Yet, it is important to note that in many cases the shape is not exactly sinusoidal or even symmetric. Figure 1 shows an example of these up-down-up asymmetric rhythmic patterns. Models based on parametric functions of time, such as Cosinor,[24,3], have been proposed in chronobiology to model these patterns.[25] However, these parametric functions are too rigid, as other patterns (e.g. asymmetric ones) frequently appear in biological systems. The first statistical problem to solve in this context is to determine if the observed pattern is rhythmic or not. There are a wide variety of procedures in the literature to detect rhythmicity including, among others, those based on autocorrelation[26]; cosine curve-fitting[10,27,3] or Fourier analysis.[28] Some non-parametric methods such as JTK_Cycle[29] (JTK) and RAIN,[21] that use Jonckheere-Terpstra test and the Kendall's tau correlation, have also been proposed in the literature. However, these approaches do not detect asymmetric rhythmic patterns properly. Recently, Larriba et al[22] designed an algorithm that successfully identify and classify circadian clock gene expression patterns. This latter approach can be considered as a precursor to the methods we present here.

A fundamental assumption made in the above discussion is that the time corresponding to each biological sample is known or can be ascertained. However, in many instances, such as when dealing with samples obtained from human cadavers[6] or human organ biopsies,[30,31] the timing, i.e. 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.

There are, to our knowledge, two main procedures in the literature to cope with this problem of timing estimation, namely Oscope[32] and CYCLOPS.[33] Both procedures present limitations and do not address the problem in a general context. Oscope, is specifically designed to recover cell cycle dynamic in single cell RNA-Seq experiments and works on the single transcript level, so that it is computationally intensive and highly sensitive to the inter-subject variability, inherent in human experiments. CYCLOPS is based on developing neural networks from the first eigenvectors (called eigengenes) from a singular value decomposition (SVD) analysis. Notice that eigengenes suggest the fundamental (gene) expression patterns across the samples, which in turn represent a biological theme if the data are well organized,[34,35] see Figure S1 in the Supporting Information. CYCLOPS overcomes the drawbacks in Oscope but suffers form its own weaknesses. First one is that the data should cover the entire periods, a downside in human studies, since target population must be substantially increased to fill in under-represented times of the day. Second, it uses additional information, such as gene rhythmicity evolutionary information, that it not always available. Third, the optimization problem solved is quite far, mathematically speaking, from a close-fitting formulation. And fourth, in a neural network framework (which is like a black box), it is difficult to assess the influence of outliers, due to biological or sample noise, or of other artifacts of the data.

In this paper, we develop a general methodology, based on Order Restricte Inference (ORI), that solves the two main questions of rhythmic pattern detection and timing estimation commented above besides addressing other scientific questions related to oscillatory systems. In Section 2 we introduce a *(circular) signal* plus error model which is the basis of the methodology, we develop estimation and testing procedures that can be used to solve the problem of rhythmic pattern detection and the interesting question arising in real practice of the estimation of peaks. Also in this Section, a solution to the problem of timing estimation under the framework of *circular signal* models is proposed reformulating the statistical problem as that of deriving the optimal *circular order*. Section 3 is devoted to simulation experiments that illustrates the good performance of the new solutions by comparing them to those provided with alternative approaches, while Section 4 shows the results obtained using several real data sets. Finally, concluding remarks are provided in Section 5.

2 | METHODS

The key of our methodology is the definition of what we call *circular signals*. A *circular signal* can be graphically mapped as a function displaying a temporal up-down-up pattern. Such patterns are commonly seen in biological rhythmic processes as in cell-cycle and the circadian clock (see Figure 1). This up-down-up pattern over a discrete number of values can be described using mathematical inequalities that establish order restrictions among those values. We refer to these signals as *circular*, since periodic events in the Euclidean space can be mapped as circular processes in the Circular space.[36] Moreover, *circular signals* can be equivalently formulated both in the Euclidean as well as in the Circular space (see Section 2 in the Supporting Information).

For simplicity of exposition, throughout this paper we shall use the term "gene" to describe the response variable of interest and the term "gene expression" for the outcome.

2.1 | Circular signal model

For each gene, suppose its expression is obtained at time points $t_i, i = 1, \ldots, n$ in each of the p periods of data, with T being the length of each period. In our set-up both p and T are known. Let X_{ij} be the observed data collected at time point $t_i, i = 1, \ldots, n$, within the jth period, $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})'$ denote the vector of data at the n time points in the jth period $j = 1, \ldots, p$ and $\boldsymbol{Y} = (\overline{X}_{1.}, \ldots, \overline{X}_{n.})'$, where $\overline{X}_{i.}$ denotes the average of data collected at time point t_i across the p periods, for $i = 1, \ldots, n$. We assume that, for each given time point t_i, the data collected across p periods has same expected value, that the covariance matrix of $\boldsymbol{X}_j$ is a diagonal matrix, and that the period samples are independent from one another. Independence, is a reasonable assumption in many applications if p is not too large. Thus, we assume that, for each $j = 1, \ldots, p$, the data satisfy the following signal plus error model:

$$\boldsymbol{X}_j = \boldsymbol{\mu} + \boldsymbol{\epsilon}_j, \tag{1}$$

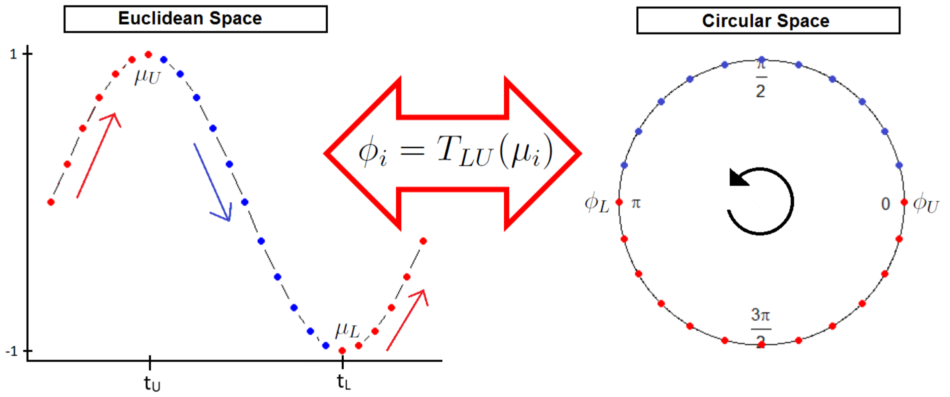
where the signal term $\boldsymbol{\mu}$ has an up-down-up shape which is defined more precisely below. For convenience, we shall refer to the signal as *circular signal* and it should not be confused that the components of $\boldsymbol{\mu}$ are angular parameters. These values are points in the Euclidean space that have a rhythmic pattern as shown in Figure 2. It is important to notice that no distributional assumption is needed for the results obtained below unless otherwise stated in the text.

Denote $L = \underset{i=1,\ldots,n}{\arg\min}\, \mu_i$ and $U = \underset{i=1,\ldots,n}{\arg\max}\, \mu_i$, the indices on the time point vector for which the signal reaches its minimum and maximum respectively. Common signal shapes in chronobiology describe up-down-up patterns, that is, the signal monotonically increases up to μ_U, and then decrease up to μ_L before increasing again so that these indices are usually unique. A typical pattern is provided in the left hand panel of Figure 2. However, as illustrated in this work, the real data exhibit more irregular patterns, such as those in Figure 1 and Figure S2 in the Supporting Information. Notice that $L > U$ or $L < U$ is not usually known to the analyst. Without loss of generality we assume that $L > U$ and we maintain in the rest of the paper, while in practise both options are tested.

Now we provide the Euclidean space representation of the *up-down-up signals* ($\boldsymbol{\mu}$) and their equivalent circular ordered representation (ϕ) in the Circular space. Proposition 1 (Supporting Information) checks that these two representations are equivalent using a transformation T_{LU} between the Euclidean and Circular spaces. One may refer to Section 2 in the Supporting Information for full details. Given this equivalence, for convenience, we refer to both representations as *circular signals*.

A signal $\boldsymbol{\mu}$ in the Euclidean space is said to be up-down-up iff $\boldsymbol{\mu} \in C = \bigcup_{LU} C_{LU}$, where $L, U \in \{1, \dots, n\}$, $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \leq \dots \leq \mu_U \geq \dots \geq \mu_L \leq \dots \leq \mu_n \leq \mu_1\}$.

A signal $\boldsymbol{\phi}$ in the Circular space is said to be circular ordered iff $\boldsymbol{\phi} \in C_o = \{\boldsymbol{\phi} \in [0, 2\pi)^n : \phi_1 \preceq \dots \preceq \phi_n \preceq \phi_1\}$ where $\preceq$ can be read as "is followed by". $\boldsymbol{\phi}$ it is said to follow the circular order o.



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

FIGURE 2 Equivalent formulation of circular signal with $L > U$. Left: Euclidean space. Right: Circular space.

The equivalence between circular signal in the Euclidean space and circular signal on the unit circle is illustrated in Figure 2. The utility of the equivalence of the two formulations to solve the problem of temporal order estimation is discussed in Section 2.4. Moreover, other methodological strategies will arise from the equivalence. Some of them are described in the Discussion Section.

2.2 | Circular signal estimation

We propose solution to the following mean squares optimization problem as estimator of the circular signal:

$$\boldsymbol{Y}^\star = \underset{\boldsymbol{Z} \in C}{\arg\min} \sum_{i=1}^{n} (Y_i - Z_i)^2. \tag{2}$$

The vector $\boldsymbol{Y}^\star$ is called the Isotonic Regression (IR) of $\boldsymbol{Y}$ with respect to C with equal weights. If variances are unequal then we use weighted least squares where the weights are inverse of variances at each time point. The IR estimator is a step function, with sets of consecutive components for which the estimator takes the same value, called level sets. See Robertson et al[37] for applications and algorithms to solve IR problems.

Note that the order defined by C is not a closed convex cone but a union of closed convex cones (recall that $C = \bigcup_{LU} C_{LU}$). Consequently the derivation of $\boldsymbol{Y}^\star$ is non-trivial for C, unlike when the cone of interest is a closed convex cone, see Robertson et al.[37]

In order to derive the IR estimator defined in (2), we have designed a computationally efficient algorithm based on theoretical results on the IR estimator. Both the algorithm and the theoretical results are given in Section 3 of the Supporting Information.

Implicit in the estimation of $\boldsymbol{\mu} \in C$, is the problem of estimating t_U and t_L, the times at which the peak and the trough occur. In the case of genes with periodic expression, these peak time points correspond to the biological/functional activity of the genes. Point estimators for t_U and t_L can be immediately derived from the point estimation of the indices U and L obtained in Algorithm 1, given in Subsection 3.2 in the Supporting Information.

2.3 | Inference for circular signal normal models

In chronobiology it is reasonable to assume that the random errors $\boldsymbol{\epsilon}_j$, $j = 1, 2, \ldots, p$, are independently and normally distributed and that the variability in the random errors are not dependent on the value of the signal. Thus, we assume that the variances are homoscedastic. It is well-known that, under this assumption, IR yields the maximum likelihood estimator (MLE) of the corresponding parameter.[37,38] Therefore, in this case $\boldsymbol{Y}^\star$ as defined in (2), is the MLE of the circular signal $\boldsymbol{\mu}$.

In particular, under the assumption of normality, it is straightforward to obtain confidence intervals for t_U and t_L using standard parametric bootstrap.[39] Subsections 3.1 and 4.1 below compare the performance of this approach with Cosinor, an extended methodology to model rhythms in chronobiology, in simulations and real data, respectively.

Testing circular signals

We formulate the problem of identifying circadian genes using the following testing problem:

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_n \ \textit{(flat signal)} \\ H_1 &: \boldsymbol{\mu} \in C \ \textit{(circular signal).} \end{aligned} \tag{3}$$

Detection of rhythmic signals has been considered in the literature by several authors[2,29,21] as a testing problem. In Larriba et al[22], the authors even classified signals into one of four different patterns.

As remarked in Section 2.2, it is important to recognize that (3) is not a standard testing problem. Again, the standard ORI theory is not applicable directly because $C = \bigcup_{LU} C_{LU}$ in (3) is not a convex cone.[37,38] In fact, the non-convexity issue arises even in the case of simpler alternative hypotheses such as the umbrella order where the location of the peak (or trough) is unknown, i.e. $\mu_1 \leq \cdots \leq \mu_r \geq \cdots \geq \mu_n$ with r unknown.[40,41,42] For this reason, as done in the case of umbrella alternative, we consider a two-step approach. First, we estimate L and U by $L^\star$ and $U^\star$ using the IR algorithm (see Subsection 3.2 in the Supporting Information) and we define the testing problem assuming L and U are known. This way, we do not propagate the uncertainty estimates associated with $L^\star$ and $U^\star$ when dealing with Type 1 errors associated with the proposed methodology. Thus, the hypothesis testing problem of interest is formulated as follows:

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_n \ \textit{(flat signal)} \\ H_1 &: \boldsymbol{\mu} \in C_{LU} \ \textit{(circular signal. } L\textit{, } U \textit{ known).} \end{aligned} \tag{4}$$

We test the above hypotheses using the conditional test based on the likelihood ratio (LR) statistic. The use of conditional tests is not new, see Bartholomew,[43] Menéndez and Salvador[44] and Fernández et al[45] among others. It is basically a conditional version of the classical LR test for the above hypotheses where the critical value depends upon the number of level sets m used to derive $Y^\star$. The conditional test is computationally simple and often more powerful for interesting alternatives than the LR test (see the above references).

The conditional $\alpha-$level test is given by:

$$\text{Reject } H_0 \text{ if } R \geq c(m),$$

where the LR statistic is given by $R = \sum_{i=1}^{n} \omega_i (Y_i^\star - \overline{Y})^2$, $\overline{Y} = (\sum_{i=1}^{n} \omega_i Y_i) / \sum_{i=1}^{n} \omega_i$ and $c(m)$ is chosen so that $Pr(\chi^2_{m-1} \geq c(m)) = \alpha$. Recall that in this work we are assuming that $\omega_i = \frac{p}{\sigma^2}$ and that σ^2 is assumed initially known.

In the case σ^2 unknown, the R statistic changes and includes an estimator of σ^2. The new statistic has a beta distribution with parameters $((m-1)/2, (2n-m)/2)$ so that percentiles coming from that distribution are used, instead of the ones from χ^2_{m-1}, in the determination of the critical value. See Robertson et al[37] pages 69-70.

The performance of the above conditional test is compared with JTK, a well known method used in chronobiology to detect rhytmic patterns.[46,47] In Section 3.1 both methods are compared using synthetic data and in Section 4.1 using real data.

2.4 | Temporal order estimation

As noted earlier, in some applications, time points at which samples were obtained may be unknown. For example, when dealing with autopsy data, the time of death of a set of patients is usually unknown. Yet, using data from those patients, chronobiologists are interested in determining the temporal order among samples (patients' time of death) regarding gene expression data. This section provides a mathematical solution to the problem of temporal order estimation by deriving the optimal circular order among the time point indices of an observed data set. Note that, in a second stage, this method can be combined with the one developed in Subsection 2.3 to identify rhythmic genes. The proposed methodology provides a biologically interpretable solution and overcomes some of the shortcomings of recently introduced CYCLOPS methodology. We shall evaluate the performance of the two methods using measures based on mean squared error (MSE).

Suppose $\boldsymbol{Y}^k = (Y_1^k, \ldots, Y_n^k)'$ represents the (gene expression) data corresponding to the k^{th} gene, $k = 1, \ldots, K$ and $\mathcal{D} = \{\boldsymbol{Y}^k\}_{k=1}^K$ denotes set of all data vectors from the K genes.

Let Π be the set of all possible circular orderings of all indices $S = \{1, \ldots, n\}$ around a unit circle. Notice that for each circular order $o \in \Pi$, there is a circular signal model so that $\boldsymbol{\mu} \in C_o$. For a given circular order $o \in \Pi$, we define a measure of the distance between o and $\mathcal{D}$ as follows:

$$d(o, \mathcal{D}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \nu_k \left(Y_i^k - Y_{o,i}^{\star k}\right)^2, \tag{5}$$

where $\boldsymbol{Y}_o^{\star k} = (Y_{o,i}^{\star k})_{i=1}^n$ denotes the IR of $\boldsymbol{Y}_o^k = (Y_{o,i}^k)_{i=1}^n$ under the circular signal model that generates o, and ν_k denotes the weight associated with the k^{th} element in the data set. For instance, when the experiments are subject to different variability, then $\nu_k = \frac{1}{\sigma_k^2}$.

The problem of determining the temporal order among the specimens using the gene expression data is to solve the following optimization problem:

$$\arg\min_{o \in \Pi} d(o, \mathcal{D}). \tag{6}$$

The above problem (6) is a NP-hard problem,[48] since the unknown order is one among $\#\Pi = (n-1)!$ possible orders.

An optimization problem that resembles (6) is formulated in Barragán et al,[49] although in this latter case the statistical problem is one defined in a Circular space.

We obtain an approximate solution to the optimization problem (6) by formulating it as a traveling salesman problem (TSP) as follows. The data on gene k are represented by a weighted directed graph where the nodes represent the items (or points) to be ordered. Each pair of nodes (i, j) is connected by an edge of length L_{ij}^k that represents the intensity of the relationship or the distance between i to j in gene k. The information is aggregated resulting in a matrix L of aggregated edge lengths, $L_{ij} = \sum_{k=1}^K \delta_k L_{ij}^k$. Some details on the choice of the weights are given below.

In the problem at hand, if L_{ij} measures the temporal distance between the data at time points t_i and t_j, then we propose to use the L^1 distance for the expression values, $L_{ij} = \sum\limits_{k=1}^{K} \delta_k |Y_i^k - Y_j^k|$ for $i, j \in S$. Let the binary matrix Γ, satisfying restrictions (i) and (ii) below, represent a tour that goes exactly once through all nodes in the graph, starting and ending at the same node with $\Gamma_{ij} = 1$ iff the edge (i, j) is active in the tour.

There is an obvious one to one relationship between Γ and orders o among the set of indices S. Therefore, the problem of finding a circular order using the representation of an aggregated directed graph, defined by L, is reduced to finding a 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 which, in our case, is mathematically formulated as follows,

$$\widehat{\Gamma} = \arg\min_{\Gamma} \sum_{ij} \Gamma_{ij} L_{ij} \tag{7}$$

restricted to

$$(i)\ \Gamma_{ij}, \text{ is a doubly stochastic matrix,}$$
$$(ii)\ \sum_{i,j\in V}\Gamma_{ij} \le |V|-1\ \ \forall V\subset S, |V|>1.$$

We conducted several heuristic procedures to provide a set of approximate solutions for the tour $\widehat{\Gamma}$ defined in (7). Among those tours, we choose the one ($\hat{o}$) that minimizes (5). One may refer to Subsection 4.1 in the Supporting Information for a detailed description of the temporal order estimation methodology including a flowchart.

The above optimization algorithm is flexible as different weights can be chosen. For example, in the case of gene expression data, information regarding rhythmicity of a set of genes is often available, for instance because they are known to be rhythmic in other organs or species so that the weights can be assigned accordingly. Moreover, the above procedure can be combined with a previous SVD on the initial data matrix $\mathcal{D}$, as in CYCLOPS, using only the first eigengenes proposed in Anafi et al.[33]

In Subsections 3.2 and 4.2, we compare the performances of ORI and CYCLOPS approaches when determining the temporal orders among time points for simulations and real data, respectively.

Validation measures

We define several measures of agreement between a circular order and a data set based on IR. Relative efficiency rates (RRE) are similar to measures used in linear regression and also resemble those proposed in Anafi et al.[33] The measure compares the total sum of squared errors of a given order (o), that may correspond to a linear or a circular (up-down-up) relationship, in reconstructing characteristic expression patterns, relative to the total sum of errors as follows:

$$RRE_T(o) = \frac{\sum_{k=1}^{K} SRE_t(o,\boldsymbol{Y}^k)}{\sum_{k=1}^{K} SRE_t(\cdot,\boldsymbol{Y}^k)}, \tag{8}$$

where, $SRE_t(\cdot,\boldsymbol{Y}^k) = \sum\limits_{i=1}^{n}\left(\frac{Y_i^k-\overline{Y^k}}{Y_i^k}\right)^2$; $\overline{Y^k} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$, $SRE_t(o,\boldsymbol{Y}^k) = \sum\limits_{i=1}^{n}\left(\frac{Y_i^k-Y_{o,i}^{\star k}}{Y_i^k}\right)^2$ and $\boldsymbol{Y}_o^{\star k}$ is the IR of $\boldsymbol{Y^k}$ under the circular signal model that generates the order o for $k=1,\ldots,K$.

RRE is a positive measure of the percentage of variability not explained by order o. Smaller values of the RRE measure indicate that the order generates estimators that are closer to the observed values suggesting a more reliable order reconstruction.

On the other hand, for experiments where the real timing is known, a measure of concordance (CRE) between the real order and the circular order o is defined as follows:

$$CRE_T(o,REAL) = \frac{\sum_{k=1}^{n} SRE_t(o,\boldsymbol{Y}_{REAL}^{\star k})}{\sum_{k=1}^{n} SRE_t(\cdot,\boldsymbol{Y}_{REAL}^{\star k})}. \tag{9}$$

Again, notice that smaller values of the CRE measure indicates a higher concordance among the real timing and the circular orders considered.

3 | SIMULATIONS

We generate a data set combining four signal shapes that represent real gene expression patterns, called *Cosine, Cosine Two, Asymmetric* and *Flat* (see Section 5 in the Supporting Information for full pattern definitions). The first three ones represent patterns from rhythmic genes and the *Flat* pattern represents non-rhythmic genes. A data set with $15,000$ genes is generated, 20% corresponding to rhythmic patterns ($1,000$ genes from each of the three rhythmic signals) and the rest to non-rhythmic patterns, imitating real scenarios. Corresponding to each pattern, we simulate data $\boldsymbol{X}_j = (X_{1j},\ldots,X_{nj})'$ for $n = 24$ time points and $j = 1,2$ periods using the simulated data set equation $\boldsymbol{X}_j \sim N_{24}(\boldsymbol{\mu},\sigma^2\boldsymbol{I})$ where σ^2 is fixed to be 1, so that $\boldsymbol{Y} \sim N_{24}(\boldsymbol{\mu},\frac{\sigma^2}{2}\boldsymbol{I})$. The values of $\boldsymbol{\mu}$ are chosen so as to represent the four different signal shapes considered. Moreover, following Wu et al[50] and Larriba et al[22], in this simulation study we sampled every hour for 2 full days (denoted by 1h/2 days), phase shift (t_U) is chosen from an uniform distribution in $[0,24)$, and a median amplitude level of 2.5 is also fixed. We performed exhaustive simulations using a wide range of values of error variance, phase angles and amplitudes. Since the conclusions were similar to the ones summarized here, we do not present those results in the paper. More details on the simulation design are given in Section 5 in the Supporting Information.

3.1 | Results for circular signal inference

First, we consider the estimation of $\boldsymbol{\mu}$, t_U, and t_L. In this case, we compare ORI with Cosinor, in terms of the MSE of the estimators. We drive confidence intervals (CI) using parametric bootstrap.[39] For these simulations we have computed 500 percentile confidence intervals. Each of these intervals is based on 200 bootstrap replications. Since data periodicity is 24 hours, we computed MSE taking into account the equivalence $0 \equiv 24$ hours.

Table 1 illustrates the average MSE for ORI and Cosinor estimators for the signal ($\boldsymbol{\mu}$) and for peak and troughs times (t_U, t_L) for each of the three simulated rhythmic patterns. For each of these three patterns, the average MSE is computed over the corresponding genes in the $15,000$ genes data set. Apart from the case when the data are generated according to *Cosine* function, which is the underlying assumption in the Cosinor model, in all other cases, ORI outperforms Cosinor by having smaller MSE (see Table 1). Notice that, although Cosinor gives very good fits for the *Cosine* function, this methodology is unable to give reasonable fits for data coming from models, as the *Cosine Two* pattern, that deviate slightly from that function.[50] This fact is quite important in practice where the *Cosine* model will not hold perfectly in almost any case.

In Table 2, we also compared the procedures in terms of estimated coverage probabilities of 95% confidence intervals. As expected, neither procedure does a good job of achieving the nominal 95% level for all patterns. However, between the two procedures, ORI performs substantially better by getting coverage probability closer to the true level of 95%. On the other hand Cosinor performs disastrously when the the underlying model is not *Cosine* shaped function. In fact, in some cases, the coverage probability of Cosinor can be as low as zero.

TABLE 1 Mean MSE for $\boldsymbol{\mu}$, t_U and t_L

	Cosine		*Cosine Two*		*Asymmetric*	
	ORI	Cosinor	ORI	Cosinor	ORI	Cosinor
$\boldsymbol{\mu}$	0.32	0.06	0.32	0.84	0.23	1.77
t_U	2.10	0.08	1.30	4.15	0.07	1.80
t_L	2.25	0.08	1.26	4.11	12.12	25.07

TABLE 2 95 % CI bootstrap coverage percentages (average lengths)

	Cosine		*Cosine Two*		*Asymmetric*	
	ORI	Cosinor	ORI	Cosinor	ORI	Cosinor
t_U	91 (3.16)	100 (0.68)	92 (2.46)	0 (0.75)	100 (0.44)	71 (2.01)
t_L	93 (3.20)	100 (0.69)	94 (2.36)	0 (0.76)	88 (4.19)	0 (2.00)

We compared the performance of ORI with the commonly used JTK procedure for testing hypotheses regarding rhythmicity of a gene. Since hypotheses regarding a large number of genes is being performed, to control for false discovery rate (FDR) we applied the Benjamini-Hochberg (BH) procedure. In Table 3 we provide the FDR as well as the false negative rate (FNR), i.e. a gene with a rhytmic pattern is declared to be non-rhytmic. We performed simulations at nominal FDR $\alpha = 0.01$.

ORI controlled both the false discovery rate (FDR) and the false negative rate (FNR) for different patterns of true signal (Table 3). JTK algorithm fails to detect *Asymmetric* signal patterns with FNR =0.956, while ORI works well in that, although it has a slightly higher FDR value for *Flat* pattern than expected (0.025 instead 0.01).

TABLE 3 FNR and FDR comparisons at nominal level of $\alpha = 0.01$.

False Negative Rate						False Discovery Rate	
Cosine		*Cosine Two*		*Asymmetric*		*Flat*	
ORI	JTK	ORI	JTK	ORI	JTK	ORI	JTK
0.000	0.000	0.000	0.000	0.000	0.956	0.025	0.000

3.2 | Results for temporal order estimation

In this section we consider the problem of temporal order estimation. We compare the performance of 4 different methods for estimating the true temporal order. In addition to ORI methodology described earlier, we consider the recently developed neural network based methodology called CYCLOPS.[33] This methodology performs a reduction of dimensionality of the full gene expression data using those eigengenes[34] that contribute to 85% of the total variability. The third methodology, that we denote as $\text{SVD}_{85\%}$+ORI, is a variation to ORI methodology where we apply ORI on the same 85% CYCLOPS eigengenes. The fourth method we consider is LINEAR which temporally orders the data using the first eigengene.

The above four methods are compared in terms of *RRE* and *CRE* described in Section 2.4. As for the data sets considered the true temporal order among samples is known, as a measure of comparison, we computed these two metrics for the true real order (REAL). Notice that for these cases, *CRE* is obviously zero. In (8) and (9) we defined these agreement (*RRE*) and concordance measures (*CRE*) for the orders derived from the full data set. Since CYCLOPS and $\text{SVD}_{85\%}$+ORI are based on the first eigengenes gathering for 85% of the variability of the full data set, we also considered, for comparison purposes, similar measures based on these eigengenes, denoted as $RRE_{85\%}$ and $CRE_{85\%}$. These measures are fully detailed in Subsection 4.2 of the Supporting Information.

Figure 3 shows that the first eigengene from the simulated data set captures well the rhythm, i.e. the fundamental expression pattern across samples. This fact is revealed by the high *RRE* values for LINEAR ordering in comparison with circular counterparts (see Table 4) which suggests the existence of a temporal order in this simulated data set. The reconstruction of the temporal order is much better using ORI or $\text{SVD}_{85\%}$+ORI than using CYCLOPS as the latter gives the worst *RRE* and *CRE* values in all sets considered, as it is shown in Table 4. In particular, the *CRE* value is around three times higher for CYCLOPS than for the ORI based solutions. The good performance of ORI and $\text{SVD}_{85\%}$+ORI approaches is also illustrated graphically in Figure S3 in the Supporting Information, where the pattern of different simulated rhythmic genes are plotted under four different orders.

TABLE 4 *RRE* and *CRE* values for the entire simulated data set and for the set of the first eigengenes accounting for 85% of data variability

Measure	REAL	LINEAR	ORI	$\text{SVD}_{85\%}$+ORI	CYCLOPS
RRE_T	0.1971	0.6653	**0.1965**	0.1986	0.2520
$RRE_{85\%}$	0.1104	0.5197	0.1048	**0.0970**	0.1626
CRE_T	0.0000		**0.0453**	0.0491	0.1471
$CRE_{85\%}$	0.0000		**0.0242**	0.0406	0.1309

4 | REAL DATA APPLICATION

We apply the ORI methodology to four well-known data sets in chronobiology, see Hughes et al,[29] Thaben and Westermark[21] and Larriba et al,[22] which are available online 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 45,101 genes each, whereas the U2OS human cell lines data consisted of 32,321 genes. Each data had 48 time points representing two periods of data, i.e. a sampling frequency of 1 h/2 days. As in simulation study, results for circular signal estimation and detection, and for temporal order estimation are compared with Cosinor, JTK and CYCLOPS, respectively.

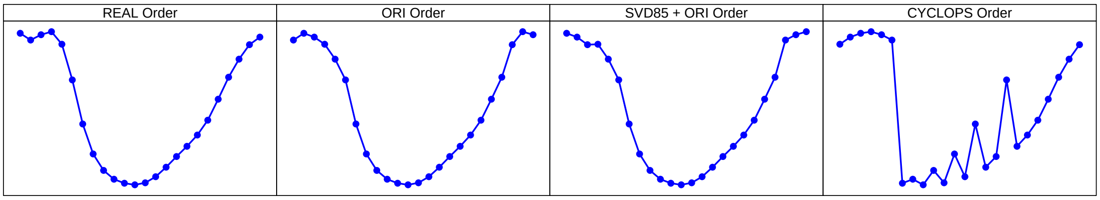


FIGURE 3 First eigengene from the simulated data set, plotted under REAL, ORI, $\text{SVD}_{85\%}$+ORI and CYCLOPS orders.

4.1 | Results for circular signal inference

First, we illustrate that the IR estimator of circular signals proposed in Section 2.2 is flexible enough to capture the pattern heterogeneity usually exhibited by circadian data bases. As an example, three rhythmic genes from mouse liver,[51] namely *Gys2, Uri1* and *Iqgap2* with different rhythmic patterns are shown in Figure 4. This figure compares ORI and Cosinor methodology and, besides the good fit provided by IR, it also exposes how differences in signal and peak time estimators increase as patterns become more asymmetric.

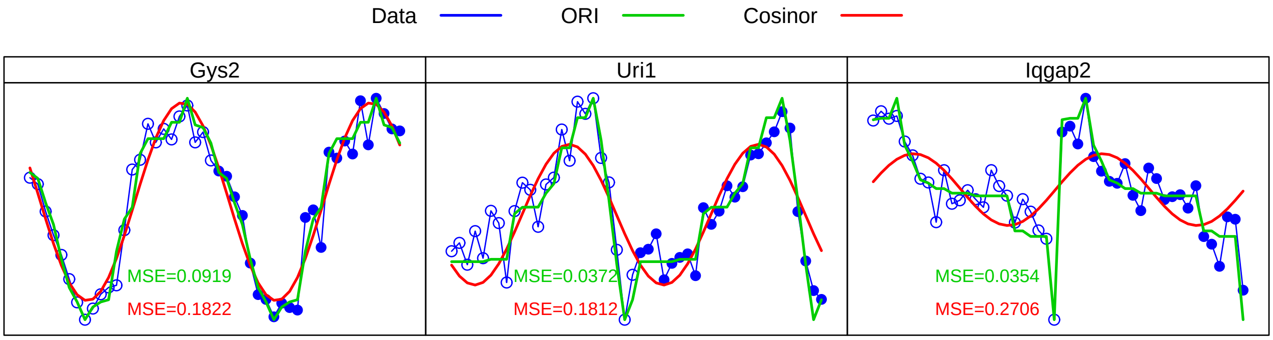


FIGURE 4 ORI (green) and Cosinor (red) model-fittings and MSE values for three different gene expression data (blue) called Gys2 (left), Uri1 (middle) and Iqgap2(right).

Now, we compare the results of ORI and JTK in the four mentioned data sets. In Table 5 we can see that for each of the data sets, there is a significant number of genes which are identified as rhythmic by ORI, but declared as non rhythmic by JTK, e.g. 5095 in mouse liver. Yet, according to the simulation study JTK, tends to have a higher FNR than our procedure (see results for the *Asymmetric* pattern in Table 3). To illustrate this fact, Figure 5 displays specific gene expression examples, among those 5095 genes in mouse liver, that are declared as non rhythmic by JTK despite that they present a clear rhythmic pattern.

4.2 | Results for temporal order estimation

Here, we investigate the performance of the temporal order estimation in these four real data sets, assuming that the order is unknown. As in simulations, we consider the REAL, LINEAR, ORI, $\text{SVD}_{85\%}$+ORI and CYCLOPS orders. For each data set, *RRE* and *CRE* measures are computed following the lines described in the simulation study.

Table 6 shows the results for the different methods considered in the four data sets. In all cases the existence of a temporal order is supported by the low values of *RRE* for the circular orders compared with the linear counterparts.

TABLE 5 Rhythmic and non-rhythmic joint gene detection for ORI vs JTK in the four data sets considered ($\alpha = 0.01$)

	ORI	JTK	
		Rhythmic	Non-rhythmic
Liver	Rhythmic	3952	5095
	Non-rhythmic	1046	35008
Pituitary	Rhythmic	602	2589
	Non-rhythmic	115	41795
NIH3T3	Rhythmic	35	1318
	Non-rhythmic	12	43736
U2OS	Rhythmic	30	823
	Non-rhythmic	3	31465

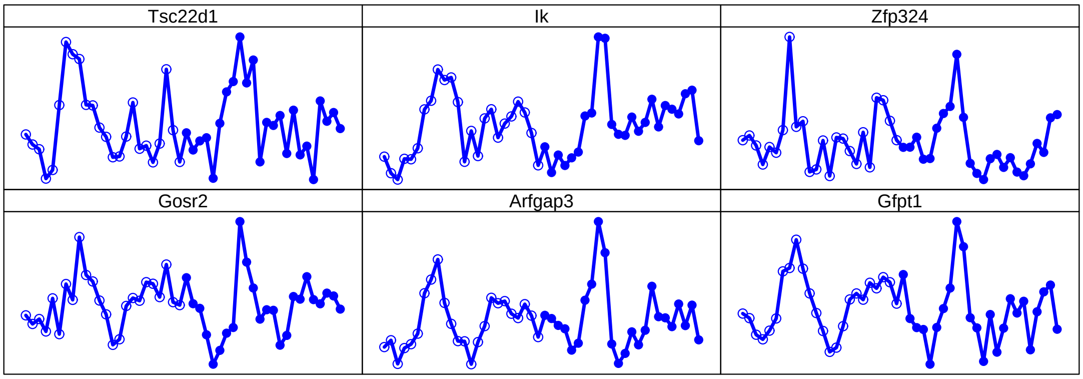


FIGURE 5 Some examples of rhythmic circadian genes in mouse liver according to ORI, which are detected as non-rhythmic by JTK ($\alpha = 0.01$).

However, there are differences among the data sets due to the different noise levels in each of them. In mouse liver the *RRE* values are much smaller than in the other data sets, as it is known that liver is a tissue with a marked presence of circadian genes.[52] As a result, the first eigengene in mouse liver (see row 1 of Figure 6) captures the rhythmicity in the data very well. The reconstruction of the temporal order is quite good under ORI and CYCLOPS, being $\text{SVD}_{85\%}$+ORI the one giving the best concordance with the REAL order. On the other hand, pituitary exhibits a high level of noise. Consequently, unlike the other data sets, the first eigengene does not display a rhythmic pattern and no strong conclusions on order estimation can be obtained. Finally, the other two data sets (NIH3T3 and U2OS cell lines) exhibit a moderate-high noise level. The first eigengene, in spite of not having a clear periodic pattern exhibits rhythmic characteristics (see rows 3 and 4 in Figure 6). Again, ORI methods outperform CYCLOPS in terms of *RRE* and *CRE* measures. Additionally, Figures S4, S5, S6 and S7 in the Supporting Information show examples of specific genes in these four data sets plotted under the REAL, LINEAR, $\text{SVD}_{85\%}$+ORI and CYCLOPS orders. For all these examples, except for row 3 in Figure S5, where noise impairs diagnosis, $\text{SVD}_{85\%}$+ORI provides a more reliable order reconstruction than CYCLOPS does.

5 | DISCUSSION

In this paper we introduce a general framework for describing and discovering rhythmic patterns of expression for oscillatory data. The proposed methodology provides solutions to a wide range of problems associated with the analysis

TABLE 6 *RRE* and *CRE* (RRE_T and CRE_T) values for the entire data set and for the set of the first eigengenes accounting for 85% ($RRE_{85\%}$ and $CRE_{85\%}$) of data variability computed across five and three different orders respectively

	Measure	REAL	LINEAR	ORI	$SVD_{85\%}$+ORI	CYCLOPS
Liver	RRE_T	0.2758	0.7636	0.2640	0.2678	**0.2633**
	$RRE_{85\%}$	0.1869	0.6254	0.1894	**0.1851**	0.1982
	CRE_T	0.0000		0.1021	**0.0896**	0.1489
	$CRE_{85\%}$	0.0000		0.0733	**0.0528**	0.1039
Pituitary	RRE_T	0.4507	0.7284	**0.2804**	0.2990	0.4433
	$RRE_{85\%}$	0.4456	0.6148	0.2674	**0.2523**	0.4336
	CRE_T	0.0000		0.6557	0.6272	**0.4273**
	$CRE_{85\%}$	0.0000		0.7564	**0.7303**	0.7372
NIH3T3	RRE_T	0.3911	0.8030	**0.3323**	0.3367	0.4070
	$RRE_{85\%}$	0.3405	0.6451	0.2763	**0.2597**	0.3895
	CRE_T	0.0000		**0.3289**	0.3438	0.4334
	$CRE_{85\%}$	0.0000		**0.4127**	0.4495	0.4723
U2OS	RRE_T	0.4698	0.8231	**0.4468**	0.4525	0.4670
	$RRE_{85\%}$	0.4127	0.7768	0.3794	**0.3646**	0.4259
	CRE_T	0.0000		**0.4892**	0.5168	0.6009
	$CRE_{85\%}$	0.0000		0.4615	**0.4535**	0.5884

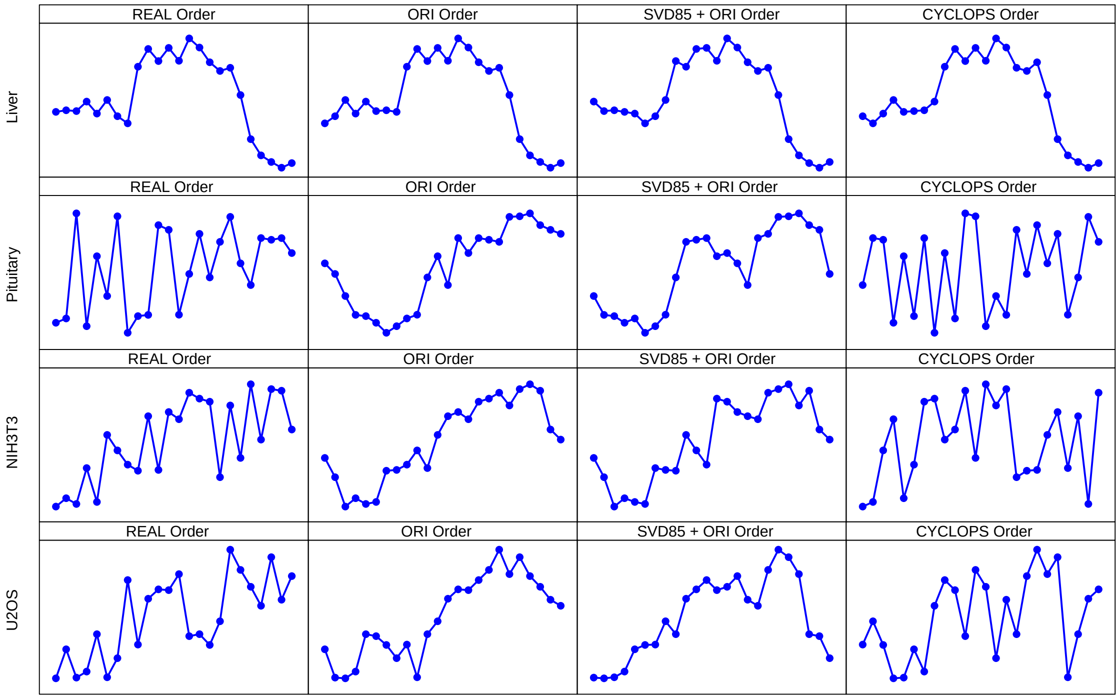


FIGURE 6 First eigengene from each of the four data sets plotted using four different orders. First row is for the mouse liver, second for mouse pituitary, third for cell lines NIH3T3 and fourth for cell lines U2OS

of rhythmic data such as rhythmicity detection, order reconstruction and peak time estimation, outperforming the available methods in literature. In particular, for signal estimation, ORI outperforms one component Fourier methods (*Cosinor*) and it does not suffer from drawbacks that appear in multicomponent Fourier methods that may yield estimated signals with multiple local maxima. As a result, ORI provides accurate peak and signal estimators which could be crucial for a more reliable solution of the problem, especially when the associated signal is asymmetric. For instance, *Iqgap2* (right panel in Figure 4) is a rhythmic gene with a markedly asymmetric gene expression pattern involved in ovarian cancer detection.[53]

There are several advantages in using the proposed methodology. First of all, the methodology is simple to describe and use. Thus applied researchers will not be intimidated by complicated theory or formulas. Secondly, we demonstrate the equivalence between the order in the Euclidean and Circular space. As a consequence, it is easy to translate between the two spaces and obtain 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. We also want to stress that, as shown in simulations and real data cases, the developed methodology outperforms other recently developed ones such as JTK for rhythmicity detection and CYCLOPS for temporal order reconstruction. 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.

The methodology developed in this work does not consider other shapes or patterns that may appear in different applications. Furthermore, in the present paper we have not considered any covariates and adjacent time points are assumed to be independent since serial correlation is embedded by the (up-down-up) signal shape. In the future, we plan to extend our ORI based methodology to deal with other patterns, covariates, as well as possible serial time correlation using ideas from Follmann and Proschan.[54]

There are several challenges with chronobiological data that require further development of methodology. For example, for the timing estimation problem, in some instances the investigator may know a priori about the time of sampling for some subset of points. In such cases, the route of the traveling salesman may be constrained by those fixed time points. Moreover, in many real cases, as in human biopsies, samples are almost exclusively obtained during the day, see Anafi et al,[33] so that data are not recorded on the entire period. Our methodology would work perfectly for those cases while CYCLOPS (and other proposals in the literature) do not. Another interesting aspect is that our methods are not affected if the data points are not equispaced.

Finally, we have developed an R code to perform all the analysis exposed here that can be obtained from the authors upon request.

For all of these reasons and chances of future developments, we have promising expectations about the ORI methodology for being favorably received by biologists.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support received by the Spanish Ministerio de Ciencia e Innovación and European Regional Development Fund; Ministerio de Economía y Competitividad grant [MTM2015-71217-R to CR and MF] and Spanish Ministerio de Educación, Cultura y Deporte [FPU14/04534 to YL].

Author contributions

CR conceived and supervised the study. YL processed original data, generated simulations, performed statistical analyses and implemented the model . CR, YL and MF developed theoretical results. CR, YL, MF and SP interpreted the results and wrote the manuscript.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

Data accessibility

The data that support the findings of this study are openly available at NCBI GEO, (http://www.ncbi.nlm.nih.gov/geo/).

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:
Figure S1. Biological theme from mouse liver (GSE11923) data set.
Figure S2. Examples of circadian gene expressions from mouse liver (GSE11923) data set describing up-down up patterns.
Figure S3. Simulated rhythmic genes with Cosine (top) and Asymmetric (bottom) patterns plotted under REAL, ORI, SVD85%+ORI and CYCLOPS orders.
Figure S4. Rhythmic circadian genes Psph, Eif5 and Errfi1 from mouse liver plotted under REAL, LINEAR, SVD85%+ORI and CYCLOPS orders.
Figure S5. Rhythmic circadian genes Marcks, Usp2 and Fkbp and from pituitary plotted under REAL, LINEAR, SVD85%+ORI and CYCLOPS orders.
Figure S6. Rhythmic circadian genes Per3, Per2 and Tspan8 from NIH3T3 plotted under REAL, LINEAR, SVD85%+ORI and CYCLOPS orders.
Figure S7. Rhythmic circadian genes 7893966, Itga5 and Atp6v0c from U2OS plotted under REAL, LINEAR, SVD85%+ORI and CYCLOPS orders.
Figure S8. Temporal order estimation flowchart.
Figure S9. Four signal shapes considered in the simulated data set along two periods.
Table S1. Functions (yt) of time (t) used to generate signal shape profiles in the simulation study.

References

1. Halberg F. Chronobiology. *Annu Rev Physiol* 1969; 31: 675-726.

2. Refinetti R, Cornelissen G, Halberg F. Procedures for numerical analysis of circadian rhythms. *Biol Rhythm Res* 2007; 38(4): 275-325.

3. Cornelissen G. Cosinor-based rhythmometry. *Theor Biol Med Model* 2014; 11: 16. doi: 10.1186/1742-4682-11-16

4. Cornelissen G, Otsuka K. Chronobiology of Aging: A Mini-Review. *Gerontology* 2017; 63(2): 118-128.

5. Halberg F, Powell D, Otsuka K, et al . Diagnosing vascular variability anomalies, not only MESOR-hypertension. *Am J Physiol Heart Circ Physiol* 2013; 305(3): H279-H294. doi: 10.1152/ajpheart.00212.2013

6. Li J, Bunney B, Meng F, et al . Circadian patterns of gene expression in the human brain and disruption in major depressive disorder. *P Natl Acad Sci USA* 2013; 110(24): 9950-9955.

7. Chauhan R, Chen K, Kent B, Crowther D. Central and peripheral circadian clocks and their role in Alzheimer's disease. *Dis Model Mech* 2017; 10(10): 1187-1199.

8. Chan S, Zhang L, Rowbottom L, et al . Effects of circadian rhythms and treatment times on the response of radiotherapy for painful bone metastases. *Ann Palliat Med* 2017; 6(1): 14-25. doi: 10.21037/apm.2016.09.07

9. Haus E. Chronobiology in Oncology. *Int J Radiat Oncol Biol PhysInt J Radiat Oncol Biol Phys* 2009; 73(1): 3-5. doi: 10.1016/j.ijrobp.2008.08.045

10. Liu D, Umbach D, Peddada S, Li L, Crockett P, Weinberg C. A random-periods model for expression of cell-cycle genes. *Proc Natl Acad Sci USA* 2004; 101(19): 7240-7245.

11. Oliva A, Rosebrock A, Ferrezuelo F, et al . The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol* 2005; 3(7): 1239-1260.

12. Xiao E, Xia-Zhang L, Barth A, Zhu J, Ferin M. Stress and the menstrual cycle: Relevance of cycle quality in the short- and long-term response to a 5-day endotoxin challenge during the follicular phase in the rhesus monkey. *J Clin Endocrinol Metab* 1998; 83(7): 2454-2460.

13. Halberg F, Cornelissen G, Wang Z, et al . Chronomics: Circadian and circaseptan timing of radiotherapy, drugs, calories, perhaps nutriceuticals and beyond. *J Exp Ther Oncol* 2003; 3(5): 223-260.

14. Hughes M, DiTacchio L, Hayes K, et al . Harmonics of circadian gene transcription in mammals. *PLoS Genet* 2009; 5(4): 1-12. doi: 10.1371/journal.pgen.1000442

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

16. Panda S, Antoch M, Miller B, et al . Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* 2002; 109(3): 307-320.

17. Hughes M, Deharo L, Pulivarthy S, et al . High-resolution time course analysis of gene expression from pituitary. *Cold Spring Harb Symp Quant Biol* 2007; 72: 381-386. doi: 10.1101/sqb.2007.72.047

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

19. Wijnen H, Naef F, Boothroyd C, Claridge-Chang A, Young M. Control of daily transcript oscillations in Drosophila by light and the circadian clock. *PLoS Genet* 2006; 2(3): 0326-0343.

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

21. Thaben P, Westermark P. Detecting Rhythms in Time Series with RAIN. *J Biol Rhythms* 2014; 29(6): 391-400.

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

23. Hughes M, Abruzzi K, Allada R, et al . Guidelines for Genome-Scale Analysis of Biological Rhythms. *J Biol Rhythms* 2017; 32(5): 380-393.

24. Tong Y. Parameter estimation in studying circadian rhythms. *Biometrics* 1976; 32(1): 85-94.

25. Jang T, Kim H, Kang S, Choo S, Lee IS, Choi K. Circadian rhythm of wrist temperature among shift workers in South Korea: A prospective observational study. *Int J Environ Res Public Health* 2017; 14(10): 1109. doi: 10.3390/ijerph14101109

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

27. Straume M. DNA Microarray Time Series Analysis: Automated Statistical Assessment of Circadian Rhythms in Gene Expression Patterning. *Methods Enzymol* 2004; 383: 149-166.

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

29. Hughes M, Hogenesch J, Kornacker K. JTK CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *J Biol Rhythms* 2010; 25(5): 372-380.

30. Lamb J, Zhang C, Xie T, et al . Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PLoS ONE* 2011; 6(7): 1-17. doi: 10.1371/journal.pone.0020090

31. Bossé Y, Postma D, Sin D, et al . Molecular signature of smoking in human lung tissues. *Cancer Res* 2012; 72(15): 3753-3763.

32. Leng N, Chu L, Barry C, et al . Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat Methods* 2015; 12(10): 947-950.

33. Anafi R, Francey L, Hogenesch J, Kim J. CYCLOPS reveals human transcriptional rhythms in health and disease. *Proc Natl Acad Sci USA* 2017; 114(20): 5312-5317.

34. Alter O, Brown P, Botstein D. Singular value decomposition for genome-Wide expression data processing and modeling. *Proc Natl Acad Sci USA* 2000; 97(18): 10101-10106.

35. Zhang W, Edwards A, Fan W, Zhu D, Zhang K. SvdPPCS: An effective singular value decomposition-based method for conserved and divergent co-expression gene module identification. *BMC Bioinformatics* 2010; 11. doi: 10.1186/1471-2105-11-338

36. Winfree A. *The Geometry of Biological Time.* Springer Science Business Media . 2001.

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

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

39. Efron B, Tibshirani R. *An introduction to the bootstrap.* New York : Chapman & Hall . 1993.

40. Peddada S, Harris S, Zajd J, Harvey E. ORIOGEN: Order restricted inference for ordered gene expression data. *Bioinformatics* 2005; 21(20): 3933-3934.

41. Rueda C, Ugarte M, Militino A. Checking unimodality using isotonic regression: an application to breast cancer mortality rates. *Stoch Env Res Risk A* 2016; 30(4): 1277-1288.

42. Wylupek G. An Automatic Test for the Umbrella Alternatives. *Scand J Stat* 2016; 43(4): 1103-1123.

43. Bartholomew D. A test of homogeneity of means under restricted alternatives. *J Roy Stat Soc B Met* 1961; 23(2): 239-281.

44. Menéndez J, Salvador B. Anomalies of the likelihood ratio tests for testing restricted hypothesis. *Ann Statist* 1991; 19(2): 889-898.

45. Fernández M, Rueda C, Peddada S. Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Res* 2012; 40(7): 2823-2832.

46. Bekker dC, Will I, Hughes D, Brachmann A, Merrow M. Daily rhythms and enrichment patterns in the transcriptome of the behavior-manipulating parasite Ophiocordyceps kimflemingiae. *PLoS ONE* 2017; 12(11): 1-20. doi: 10.1371/journal.pone.0187170

47. Ferrari C, Proost S, Janowski M, et al . Kingdom-wide comparison reveals the evolution of diurnal gene expression in Archaeplastida. *Nat Commun* 2019; 10(1): 387316. doi: 10.1038/s41467-019-08703-2

48. Bartholdi III J, Tovey C, Trick M. The computational difficulty of manipulating an election. *Soc Choice Welfare* 1989; 6(3): 227-241.

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

50. Wu G, Zhu J, Yu J, Zhou L, Huang J, Zhang Z. Evaluation of five methods for genome-wide circadian gene identification. *J Biol Rhythms* 2014; 29(4): 231-242.

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

52. Zhang R, Lahens N, Ballance H, Hughes M, Hogenesch J. A circadian gene expression atlas in mammals: Implications for biology and medicine. *P Natl Acad Sci USA* 2014; 111(45): 16219-16224.

53. Deng Z, Wang L, Hou H, Zhou J, Li X. Epigenetic regulation of IQGAP2 promotes ovarian cancer progression via activating Wnt/Îš-catenin signaling. *Int J Oncol* 2016; 48(1): 153-160.

54. Follmann D, Proschan M. A Simple Permutation-Type Method for Testing Circular Uniformity with Correlated Angular Measurements. *Biometrics* 1999; 55(3): 782âĂŞ791.

How to cite this article: Larriba Y, Rueda C, Fernández MA, and Peddada SD (), Order Restricted Inference in Chronobiology, *Stat Med*, *XX*.