Circular Rank Aggregation and its Application to Cell-cycle Genes Expressions

Sandra Barragán, Cristina Rueda and Miguel A. Fernández

Abstract—The aim of circular rank aggregation is to find a circular rank or order on a set of n items using angular values from p heterogeneous data sets. This problem is new in the literature and has been motivated by the biological question of finding the order among the peak expression of a group of cell cycle genes. In this paper, two very different approaches to solve the problem that use pairwise and triplewise information are proposed. Both approaches are analyzed and compared using theoretical developments and numerical studies, and applied to the cell cycle data that motivated the problem.

Index Terms—Cell-cycle Genes, Circular Data, Hodge Theory, R package **isocir**, Rank Aggregation, Traveling Salesman Problem

1 Introduction

In this paper we deal with the problem of obtaining a circular rank (order) on a set of n items by using angular values from p heterogeneous data sets that typically are observations from experiments conducted under different conditions.

The question of circular rank aggregation (equivalently circular order aggregation) has been motivated by an application in molecular biology related to the analysis of expression data from cell cycle genes that play an important role on the process of cell division.

The circular problem has a counterpart in the line, the classical problem of determining the *true* order or rank among n objects using the ranks assigned by p independent *judges*. There exists a huge literature in rank aggregation for Euclidean data [1, 2, 3, 4]. In fact, a broad list of techniques to tackle the problem has been developed and numerous settings have been considered. The problem can be presented in a general form as that of finding the rank that is "closest to" a given set of data according to an objective function or criteria. The techniques can be classified considering several aspects:

1) The type of objective function or criteria.
2) The type of input information (ordinal or cardinal).
3) The mathematical representation of the input information: As vectors (when ranks are given by the judges), or matrices (when the initial information are multiple preferences between pair of items, see [5]).
4) The statistical assumptions considered. These may range assuming fixed distributions for the observed data to no assumption at all passing by assuming a distribution on the permutations.
5) The available information, producing supervised (some information available) or unsupervised (no information available) aggregation methods.

• *S. Barragán, Cristina Rueda and Miguel A. Fernández are with the Departamento de Estadística e I.O., Universidad de Valladolid, 47011 Valladolid, Spain. E-mails: sandraba@eio.uva.es, crueda@eio.uva.es, miguelaf@eio.uva.es*

The spectrum of the problems where the methodology of rank aggregation is applied is wide, starting with applications in social sciences, where the subject initially appears under the name of social choice problem. In this field the most studied problem is vote aggregation [6, 7, 8]. Nowadays, one of the most relevant areas facing the problem of rank aggregation is information retrieval. In this area, rank aggregation methodology is being applied in web searching [9, 10, 11] and one of the most popular algorithms, the PageRank algorithm (used by Google) has been developed [12]. Other areas with interesting applications include biology [c.f. 13, 14, 15, 16], sport competitions [17], or quality assessment [18] among others.

Although rank aggregation is extensively studied in the line, the problem is practically unexplored in the circular setting. As we show along this paper, due to the underlying geometry of the circle, the Euclidean space based methods cannot be directly applied. Although the problem of circular order aggregation has been briefly introduced in [19] and [20], this is the first paper on the subject which is characterized by the use of circular ranks, or angular data sets, as input information and by formulating the problem as the search of a common circular order. For a revision of the basic elements to analyze circular data we refer to the books of [21] and [22]. Two papers dealing with statistical related problems are [23] and [24]. The first solves the problem of estimating points in a unit circle subject to an order restriction and the second provides a test for testing a given circular order. In both papers cell-cycle gene expression data have been analyzed.

To illustrate the difficulties of the problem at hand, see example in Figure 1 below with three items to be ordered: $1, 2, 3$ and two experiments (a),(b). The observed values in the experiments (a) and (b) verify the same circular order, that we will denote as $1 \leq 3 \leq 2 \leq 1$ to emphasize its circularity. However, the direct approach to aggregate angular information, the circular means of the observed values, do not verify the same order (Figure 1(c)). The problem is due to the non-convexity of the set of vectors verifying a circular order. If the data in the same example are rotated it can be checked that the arithmetic mean is

not a valid approach either. Then, the classical approach of Borda method is not appropriate.

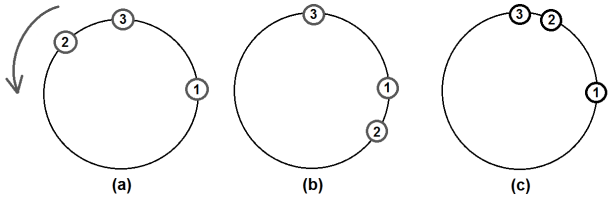

Fig. 1. Example: n=3, p=2. Observed scores in (a) and (b). Circular means in (c)

Many possible techniques can be developed by considering the different variants of the five aspects enumerated above about rank aggregation. We have restricted the set of candidate methods to those designed to solve problems related to the motivation of our research and those with a general purpose. In particular, we consider unsupervised algorithms which search the optimum of an objective function which is defined as a distance between a target order and the data sets. We consider, cardinal and ordinal input data, two different objective functions and we make no distributional assumptions on the data. We present two different techniques. The originality of these proposals are twofold. First, the idea of using the circular isotonic regression estimator (CIRE), see [23], which allows the definition of a new objective function for cardinal data. Second, the use of triplewise information in a novel technique.

This latter novel technique is based on Hodge theory [25]. The use of triplewise information allows the introduction of angular preferences between three items by using *triangular flows*. To our best knowledge, this is the first time that a triplewise data approach is proposed in the literature. In fact, the triplewise data seems to be the natural input information in the circular setting as three is the minimum number of items to be uniquely ordered in the circle. In this setting, we propose a squared-loss optimization problem in $\mathbb{R}^{n\times n\times n}$ to obtain the aggregated circular rank and we develop an algorithm to solve it. We also prove, using Hodge theory, good theoretical properties for the proposed algorithm.

Besides Hodge approach, we also consider another technique based on solving a Traveling Salesman Problem (TSP). The TSP is one of the most intensively studied problems in optimization. It can be formulated as the search of the shortest tour in a graph where the vertexes are the items to be ordered and the lengths of the edges (between each pair of vertexes) measure pairwise relationships. This technique has been previously explored in the works by [19] and [20].

Several interesting examples are included that illustrate the weaknesses and strengths of the methods, and a very extensive simulation study is conducted. The value of the objective functions as well as the computational time are the criteria used to compare the solutions from the different techniques.

Moreover, the different approaches are used to solve the problem of finding the order of activation of cell-cycle genes, which is the problem which initially motivated this research.

The algorithms developed in this paper have been implemented as part of an R package called **isocir** (isotonic inference for circular data), that is available on CRAN [26].

The outline of the paper is as follows. We address the basic estimation problem in Section 2 where the objective function and the related elements are defined. The TSP pairwise proposal, and the Hodge triplewise technique are presented in Sections 3 and 4 respectively. Section 5 is devoted to the analysis of the numerical results and Section 6 to the problem of ordering cell-cycle genes from heterogenous experiments. Finally, conclusions are given in Section 7.

2 THE ORDER AGGREGATION PROBLEM USING ANGULAR DATA

Let $V = \{1, 2, ..., n\}$ be the set of items to be ordered on a circle and let $j = 1, ..., p$ be the experiments. We assume that each experimenter j assigns circular scores (cardinal information), or gives a circular ordering, (ordinal information), to a fraction of the $i = 1, \ldots, n$ items. We will see that to find the aggregated order, one may use the individual observations, θ_{ij}, directly; pairwise information, Y^j_{ih} measuring the degree of preference of item i over item h ; or triplewise information, Ψ^j_{ihk}, measuring the degree of circular preference of the triplet i, then h then k. Let us also denote

$$\boldsymbol{\Theta}_j = (\theta_{1j}, \ldots, \theta_{ij}, \ldots, \theta_{nj})' \text{ for } j = 1, \ldots, p.$$

Gathering all such observations from the p experiments together, we have the matrix $\boldsymbol{\Theta} = (\Theta_1, ..., \Theta_p)$. We also denote as $\boldsymbol{T_j} = (\tau_{1j}, \ldots, \tau_{nj})'$ the vector of ordered positions for observations in experiment j. In this way, when cardinal scores are observed we have,

$$\tau_{ij} = k, i = 1, ..., n, \Leftrightarrow \theta_{(k)j} = \theta_{ij}.$$

On the other hand, when only ordinal information is provided, $\boldsymbol{T_j}$ gives the positions in the order starting with the item i such that $\tau_{ij} = 1$. Both $\boldsymbol{T_j}$ and the angular values derived from the positions, $\boldsymbol{T_j}$ (assigning $\theta_{ij} = 2\pi\frac{(k-1)}{n} \Leftrightarrow \tau_{ij} = k$) can be used as inputs depending on the technique.

Let $\mathfrak{O}$ denote the set of all possible orders among the n objects and let us denote as $\boldsymbol{\alpha} \sim \boldsymbol{O}$ when an angular vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)'$ verifies the order $\boldsymbol{O} \in \mathfrak{O}$. Also, for $j = 1, ..., p$, let $\widetilde{\boldsymbol{\Theta}}_j^{(\boldsymbol{O})} = (\widetilde{\theta}_{1j}^{(\boldsymbol{O})}, \widetilde{\theta}_{2j}^{(\boldsymbol{O})}, \ldots, \widetilde{\theta}_{n}^{(\boldsymbol{O})})'$ denote the CIRE of $\boldsymbol{\Theta}_j$ under the circular order $\boldsymbol{O}$, ie: the vector verifying the circular order $\boldsymbol{O}$, closest to $(\theta_{1j}, ..., \theta_{nj})'$ using the sum of circular errors (SCE) distance:

$$\begin{aligned} \widetilde{\boldsymbol{\Theta}}_j^{(\boldsymbol{O})} &= \underset{\boldsymbol{\alpha}\sim\boldsymbol{O}}{\arg\min}\, SCE(\theta, \alpha) \\ &= \underset{\boldsymbol{\alpha}\sim\boldsymbol{O}}{\arg\min} \textstyle\sum_{i=1}^n (1 - \cos(\theta_{ij} - \alpha_i)). \end{aligned} \quad (1)$$

The CIRE is defined in [23] where also interesting properties and an algorithm to obtain the CIRE are given.

The distance between $\boldsymbol{\Theta}_j$ and the order $\boldsymbol{O}$ is then defined using the mean sum of circular errors (MSCE) as follows:

$$\begin{aligned} d(\boldsymbol{\Theta}_j, \boldsymbol{O}) &= MSCE(\boldsymbol{\Theta}_j, \widetilde{\boldsymbol{\Theta}}_j^{(\boldsymbol{O})}) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n (1 - \cos(\theta_{ij} - \widetilde{\theta}_{ij}^{(\boldsymbol{O})})). \end{aligned} \quad (2)$$

Finally, using the CIRE and the MSCE from the p experiments, we define a distance between the full data set $\boldsymbol{\Theta}$ and

the order $\boldsymbol{O}$, that is denoted by $d^*(\boldsymbol{\Theta}, \boldsymbol{O})$ and is given by the weighted mean of the MSCEs as follows:

$$d^*(\boldsymbol{\Theta}, \boldsymbol{O}) = MSCE(\Theta, \widetilde{\boldsymbol{\Theta}}^{(\boldsymbol{O})}) = \sum_{j=1}^{p} \omega_j d(\boldsymbol{\Theta}_j, \boldsymbol{O}), \quad (3)$$

where ω_j is the weight associated with the j^{th} experiment, which is usually used to take into account differences in variability within experiments. For instance, assuming $\theta_{ij} \sim M(\phi_{ij}, \kappa_j)$ with κ_j known, the weights may be defined as $\omega_j = \frac{\kappa_j}{\sum_{j=1}^{p} \kappa_j}$.

With this notation, the problem of searching for a global circular order, $\boldsymbol{O}^* \in \mathfrak{O}$ from the information given by the p experiments, can be written as the following optimization problem:

$$\boldsymbol{O}^* = \arg\min_{\boldsymbol{O} \in \mathfrak{O}} d^*(\boldsymbol{\Theta}, \boldsymbol{O}) = \arg\min_{\boldsymbol{O} \in \mathfrak{O}} \sum_{j=1}^{p} \omega_j d(\boldsymbol{\Theta}_j, \boldsymbol{O}). \quad (4)$$

Moreover, there may exist applications where ordinal information is provided as input. In these cases, the objective function of interest is defined using the vectors of positions and the optimization problem is defined as follows,

$$\boldsymbol{O}^{**} = \arg\max_{\boldsymbol{O} \in \mathfrak{O}} \sum_{j=1}^{p} \omega_j \hat{\Delta}(\boldsymbol{T}_j, \boldsymbol{T}), \quad (5)$$

where $\boldsymbol{T}_j$ is the vector of positions for experiment j, $\boldsymbol{T}$ is the vector of positions for order $\boldsymbol{O}$ and $\widehat{\Delta}(\boldsymbol{T}_j, \mathbf{T})$ is the circular version of Kendall's Tau defined in [21] that we will denote as CKτ.

Unfortunately, even the Euclidean equivalent problems to (4) and (5) are NP-hard [see 9]. This means that there is no guarantee that the optimum can be attained in polynomial time. In this paper, we design several techniques that provide good approximations to problems (4) and (5) and have in common a general structure in two steps. In a first step an initial solution denoted as $\widehat{\boldsymbol{O}}^0$ is provided, which is refined in step 2. The final order coming from the whole procedure is denoted as $\widehat{\boldsymbol{O}}$. In sections 3 and 4 several alternative techniques are proposed to obtain $\widehat{\boldsymbol{O}}^0$ using information of individual scores, pairwise flows or triangular flows respectively. Step 2 is the same for all these techniques and it involves the implementation of a local search algorithm called CLMA (Circular Local Minimization Algorithm) whose objective is to make local improvements in $\widehat{\boldsymbol{O}}^0$. This algorithm is based on a well-known algorithm for rank aggregation called *Local Kemenization* and developed in [9]. CLMA considers each triple of consecutive elements and checks if a permutation of those items improves the objective function. Full details on how CLMA works are given in the Supplementary Information.

3 A PAIRWISE ORDERING TECHNIQUE BASED ON THE TSP

In this section we present a technique where each experiment j is represented by a directed graph where the nodes represent the items to be ordered. Each pair of nodes (h, k) is connected by an edge with length E^j_{hk} that measures the preference of h over k in experiment j. Different definitions for the lengths E^j_{hk} are proposed at the end of this section. The information given by each experiment is aggregated in a matrix $E = (E_{hk})_{n \times n}$ of aggregated edge lengths, where $E_{hk} = \sum_{j=1}^{p} \omega_j E^j_{hk}$.

The problem of finding a circular order using the representation of an aggregated directed graph, defined by E, is reduced to that of finding the shortest tour that passes exactly once by each of the nodes in the graph, starting and ending at the same node. Then, an approximate solution to (4) is given by the circular order associated with the tour that minimizes the total length. This latter problem is the well-known Traveling Salesman Problem (TSP) that is one of the most famous combinatorial problems, and perhaps the best studied one, in the field of computational combinatorial optimization and graph theory [27, 28, 29].

Let $\mathcal{X}$ be the set of $n \times n$ binary matrices. The mathematical formulation of our TSP is,

$$\widehat{X} = \arg\min_{X \in \mathcal{X}} \sum_{hk} X_{hk} E_{hk} \quad (6)$$

restricted to,

$$(i) \quad \sum_{h=1}^{n} X_{hk} = 1 \quad \forall k = 1, \ldots, n$$
$$(ii) \quad \sum_{k=1}^{n} X_{hk} = 1 \quad \forall h = 1, \ldots, n$$
$$(iii) \quad \sum_{h,k \in S} X_{hk} \leq |S| - 1 \;\; \forall S \subset V, |S| > 1.$$

A binary matrix X verifying restrictions (i), (ii) and (iii), represents a tour that goes exactly once by all nodes in the graph, starting and ending at the same node with $X_{hk} = 1$ iff the edge (h, k) is active in the tour. Therefore, there is an obvious one to one relationship between matrices verifying the three restrictions in (6) and circular orders.

The order $\widehat{\boldsymbol{O}}^0$ corresponding to the solution to (6), $\widehat{X}$, is the approximate solution to (4) given by this approach.

A major advantage of this formulation is computational as there are multiple heuristics to solve the TSP offering good approximations [29].

3.1 Definition of E^j_{hk}

In order to obtain more general results, we consider directed distances, that allow taking into account the rotation direction in the definition of the lengths of the edges. This is the general formulation we use to define the lengths of the edges.

$$E^j_{hk} = d_\alpha(\theta_{hj}, \theta_{kj}) = \min\left(d_R(\theta_{hj}, \theta_{kj}), \alpha \cdot d_C(\theta_{hj}, \theta_{kj})\right), \quad (7)$$

where d_R and d_C are distances, on the rotation direction and on the opposite direction, respectively, and $\alpha \geq 1$ is a penalization constant. The idea behind this type of penalty is from a problem presented by [30]. Notice that $\alpha = 1$ and $d_R = d_C$ would lead to an undirected distance while $\alpha = \infty$ would yield a distance that does only allow moving on the rotation direction. It is also interesting to note that in order to use the TSP algorithms, it is not necessary that the distances define a metric, it is enough that they verify some basic properties, namely they have to be bounded, positive,

continuous and verify a relaxed triangular inequality (see Lemma 3.1 below).

Many different choices for d_R, d_C and α have been considered in preliminary analysis. From the huge range of distances considered we have selected the simplest ones having the best performance in numerical studies. That selection is given in Table 1, with Table 3 containing a full description of the labels. Among the **TSP**α options with $\alpha \in [1, \infty)$ we selected for further analysis **TSP1** and **TSP3**, the first one because is a symmetric distance and the second, because of its good behavior in simulations (similar that of **TSP2** or **TSP4**) and also because it has an interesting geometric interpretation, which can be briefly explained as follows. Consider a traveler who has reached location k forgetting to stop at location h. To correct this error he/she has to go back to h and travel again from h to k to continue the route. In this way, he/she has traveled a total of three times the distance between h and k ($\alpha = 3$).

TABLE 1
Labels and definitions of the edge lengths

Label	Selected α	Lengths of the edges
TSPb	∞	$d_R(\theta_{hj}, \theta_{kj}) = \begin{cases} 0 & \text{if } k = h \\ 1 & \text{if } \tau_{kj} = \tau_{hj} + 1 \pmod n \\ 2 & \text{if } \tau_{kj} \neq \tau_{hj} + 1 \pmod n \end{cases}$ $d_C(\theta_{hj}, \theta_{kj}) = d_R(\theta_{hj}, \theta_{hj})$
TSPp	∞	$d_R(\theta_{hj}, \theta_{kj}) = \tau_k - \tau_h \pmod n$ $d_C(\theta_{hj}, \theta_{kj}) = d_R(\theta_{kj}, \theta_{hj})$
TSPα	1,3	$d_R(\theta_{hj}, \theta_{kj}) =$ $= \begin{cases} 1 - \cos(\theta_{kj} - \theta_{hj}) & \text{if } \theta_{kj} - \theta_{hj} \pmod{2\pi} \leq \pi \\ 3 - \cos(\theta_{kj} - \theta_{hj} - \pi) & \text{if } \theta_{kj} - \theta_{hj} \pmod{2\pi} > \pi \end{cases}$ $d_C(\theta_{hj}, \theta_{kj}) = d_R(\theta_{kj}, \theta_{hj})$

It is easy to see that the distances E^j_{hk} defined as in (7) using the definitions of $d_R(\theta_{hj}, \theta_{kj})$ and $d_C(\theta_{hj}, \theta_{kj})$ in Table 1 are positive, bounded and continuous. The required property of relaxed triangular inequality appearing in Lemma 3.1 is proved in the Supplementary Information.

Lemma 3.1. Let $\theta_i, \theta_h, \theta_k \in [0, 2\pi]$ and $\alpha \geq 1$, then,
$d_\alpha(\theta_h, \theta_k) \leq 2(d_\alpha(\theta_h, \theta_i) + d_\alpha(\theta_i, \theta_k))$

As we have noted before, there are multiple heuristics offering good approximations for the solution of the optimization problem. When comparing the heuristics have found that there is not an absolute winner and that a better approximation to the optimum is obtained by repeatedly running different heuristics and selecting the best solution in terms of the objective function.

In Section 5, where the different alternatives are compared numerically, we consider different TSP methods from Table 1 as well as different heuristics. Specifically, we use those implemented in the R package called **TSP** [31].

4 A TRIPLEWISE ORDERING TECHNIQUE BASED ON HODGE THEORY

The idea behind this proposal is to use triplewise information instead of using scores or pairwise values. In the same line that skew-symmetric matrices are used to define pairwise flows, skew-symmetric hypermatrices can be used to define triplewise flows as we show in this section.

Although from a formal point of view this technique requires an important theoretical basis the Hodge approach has the following advantages:

1) The triplewise format is a natural way of representing information on circular orders (as a set of three elements is the minimal set with an order relationship on the circle) that allows to combine information from different sources directly even if the starting points of the circle differ among sources.
2) It is flexible in, at least, two senses. On one hand, it allows alternative ways of introducing the information, i.e. by direct specification of relations among the elements of each triple, or by specification of individual (using vectors) or pairwise information (through matrices). On the other hand, it allows several ways to aggregate the information from the different sources.
3) The calculations are very simple and thus the execution time is very short so that the method is computationally efficient.
4) Hodge theory allows to define indexes of inconsistency to evaluate the results.

The intensity of the relationship among the elements of a triple for each experiment j is represented by an hypermatrix in $\mathbb{R}^{n\times n\times n}$ that we denote as Ψ^j. The elements of Ψ^j, ψ^j_{ihk} measure the degree of preference of the order $i \leq h \leq k \leq i$ over the order $h \leq i \leq k \leq h$ in the jth experiment and verify the basic property of being skew-symmetric, i.e. $\psi^j_{ihk} = \psi^j_{hki} = \psi^j_{kih} = -\psi^j_{ikh} = -\psi^j_{khi} = -\psi^j_{hki}$, for any $i, h, k = 1, \ldots, n$ and $j = 1, \ldots, p$.

There are many different ways of defining this "degree of preferences" depending on the problem at hand and on the objective function. At the end of this section we will see several ways of defining ψ^j_{ihk} that follow the general rule given by:

$$\psi^j_{ihk} = sign^j(i, h, k) \cdot \lambda^j_{ihk}, \qquad \begin{matrix} i, h, k = 1, \ldots, n, \\ j = 1, \ldots, p, \end{matrix} \tag{8}$$

where, $sign^j(i, h, k) = sign^j(\theta_{hj} - \theta_{ij}) + sign^j(\theta_{kj} - \theta_{hj}) + sign^j(\theta_{ij} - \theta_{kj})$ is the sign of the triple (i, h, k) in experiment j and λ^j_{ihk} is a nonnegative value that measures the degree of separation of the items i, h, k. Notice that as the ψ^j_{ihk} values are independent of the starting point of the circle we can directly aggregate the Ψ^j hypermatrices into the hypermatrix $\overline{\Psi}$ using directly the arithmetic or the circular mean.

The problem of finding the circular order under Hodge theory is similar to that of finding a linear rank from pairwise information given by [25]. For this reason, the terminology that we describe below has been borrowed from that paper which in turn comes from graph theory, linear algebra and topology.

Consider the inner product defined in $\mathbb{R}^{n\times n\times n}$ as

$$< \Psi^1, \Psi^2 >_2 = \sum_{i,h,k} w_{ihk} \psi^1_{ihk} \psi^2_{ihk}, \tag{9}$$

where w is a weight function that may account for missing information or for the weight of the items and experiments. In order to simplify the exposition we eliminate the weights

from now on. Let us also define operators $\delta_1^*, \delta_1, \delta_0^*$ and δ_0 as,

$$\begin{array}{ll}\delta_1^*(\Psi) = Y \in \mathbb{R}^{n\times n} & \text{with } Y_{ih} = \sum_k \psi_{ihk} \\ \delta_1(Y) = \Psi \in \mathbb{R}^{n\times n\times n} & \text{with } \psi_{ihk} = Y_{ih} + Y_{hk} + Y_{ki} \\ \delta_0^*(Y) = s \in \mathbb{R}^n & \text{with } s_i = \sum_h Y_{ih} \\ \delta_0(s) = Y \in \mathbb{R}^{n\times n} & \text{with } Y_{ih} = s_h - s_i.\end{array}$$

Notice that, δ_1^* takes a skew-symmetric hypermatrix Ψ into a skew-symmetric matrix Y, δ_1 takes a skew-symmetric matrix Y into a score function s, δ_0^* takes a skew-symmetric matrix into a vector and δ_0 takes a vector into a skew-symmetric matrix and they are defined as follows,

We also need to define the superindex (l) which, in a subspace or in a subset, indicates that the index l has been eliminated and the dimension has been reduced in one unit. This will be useful to pass from a set of circular scores $\phi_i, i = 1, ..., n$ which define circular order on V, to a set of scores $s_i = \phi_i - \phi_l, i \neq l$ which determine a unique rank on $V^{(l)}$ and also to define a set of circular scores from a set of Euclidean scores adding the missing element in the right place as we will se below.

Now, we can define $\mathcal{H}_C$ which, as we prove below, is the subset of skew-symmetric hypermatrices inducing a circular order.

$$\mathcal{H}_C = \{\Psi \in \mathbb{R}^{n\times n\times n} : \exists Y \in \mathcal{M}_C \text{ with } \Psi = \delta_1(Y)\},$$

where,

$$\mathcal{M}_C = \{Y \in \mathbb{R}^{n\times n} : Y_{il} = -\sum_{j\neq l} Y_{ij} \text{ and } \exists l \text{ with } Y^{(l)} \in \mathcal{M}_G^{(l)}\},$$

and,

$$\mathcal{M}_G^{(l)} = \{X \in \mathbb{R}^{(n-1)\times(n-1)} : \exists s : V^{(l)} \to \mathbb{R} \text{ with } X_{ih} = \delta_0(s)\}.$$

$\mathcal{M}_G^{(l)}$ is the set of skew-symmetric matrices that induces a rank on $V^{(l)} = \{i_1, ..., i_{n-1}\}$ (see [25]). Then, a matrix $X \in \mathcal{M}_G^{(l)}$ induces a circular order on V via the following rule. Consider the s function such that $X_{ih} = \delta_0(s)$ and assume that $s(i_1) \leq ... \leq s(i_{n-1})$ (i.e. s defines the order $i_1 \leq ... \leq i_{n-1}$ in $V^{(l)}$). Then the circular order induced in V is $l \leq i_1 \leq ... \leq i_{n-1} \leq l$.

Now any hypermatrix $\Psi \in \mathcal{H}_C$ is generated by a matrix in $\mathcal{M}_C$ which comes from a matrix in $\mathcal{M}_G^{(l)}$ which in turn is generated by a score function that induces a circular order. Therefore we have checked that any $\Psi \in \mathcal{H}_C$ induces a circular order.

Reciprocally, given a set of n circular scores $\{\phi_i, i = 1, ..., n\}$, a hypermatrix $\Psi \in \mathcal{H}_C$ can be easily defined as follows. Take $l \in V$. Then, for $i, h \in V, i, h \neq l$ define $s_i = \phi_i - \phi_l$ for $i = 1, ..., n$ and $Y_{ih}^{(l)} = s_h - s_i$, $Y_{ih} = Y_{ih}^{(l)}, Y_{il} = -\sum_{h\neq l} Y_{ih}$. By construction, $Y \in \mathbb{R}^{n\times n} \in \mathcal{M}_C$ and $\Psi = \delta_1(Y) \in \mathcal{H}_C$.

The expression of Ψ in terms of the initial scores is given by:

$$\begin{array}{l}\psi_{ihk} = 0 \text{ if } i, h, k \in V^{(l)} \\ \psi_{lih} = \phi_h - \phi_i \text{ if } h, i \in V^{(l)} \\ \psi_{ihk} = \psi_{hki} = \psi_{kih} = -\psi_{hik} = -\psi_{khi} = -\psi_{ikh} \\ \text{for any } i, h, k \in V.\end{array} \tag{10}$$

Now, as $\mathcal{H}_C$ is the subset of hypermatrices inducing a circular order, the problem of finding the closest circular order to the aggregated hypermatrix $\overline{\Psi}$ can be formulated as follows

$$\widehat{\Psi} = \arg\min_{\Psi\in\mathcal{H}_C} ||\overline{\Psi} - \Psi||^2. \tag{11}$$

From the definition of $\mathcal{H}_C$ it is straightforward that the latter problem is also equivalent to finding

$$\widehat{Y} = \arg\min_{Y\in\mathcal{M}_C} ||\overline{\Psi} - \delta_1(Y)||^2. \tag{12}$$

The solution to these problems is given in Theorem 4.1. The proof of the result, which is obtained solving in $\mathbb{R}^{n-1\times n-1}$ a well-known problem on Hodge theory (equation (7) in [25]), is given in the Supplementary Information.

Theorem 4.1. Let $\overline{\Psi}$ be a skew-symmetric hypermatrix and $\overline{Y} = \frac{1}{n}\delta_1^*(\overline{\Psi})$. Then $\widehat{\Psi} = \delta_1(\widehat{Y})$ where,

$\widehat{Y}_{il_0} = -\sum_j \widehat{Y}_{ij}^{(l_0)}, \widehat{Y}^{(l_0)} = \delta_0(\boldsymbol{s}),$ with,
$l_0 = \arg\max_l \sum_h \overline{Y}_{lh}^2$ and $s_i = -\frac{1}{n-1}\sum_{h\neq l_0} \overline{Y}_{ih}\ \forall i \neq l_0.$

The order defining $\widehat{Y}$ is our $\widehat{O}^0$ and is derived as follows. First, the s function in Theorem 4.1 defines a rank in $V^{(l_0)}$ ($i_1 \leq \cdots \leq i_{n-1} \Leftrightarrow s(i_1) \leq \cdots \leq s(i_{n-1})$). Then, the corresponding circular order in V is given by $l_0 \leq i_1 \leq \cdots \leq i_{n-1} \leq l_0$, which is, by construction, the order defined by $\widehat{Y}$.

As we have commented above, one of the main advantages of this approach is the flexibility in the definition of the triplewise flow, Ψ^j. Some alternatives, that share the general formulation (8) and have been selected from preliminary numerical studies, are given in Table 2, with Table 3 containing a full description of the labels, and compared in Section 5.

TABLE 2
Labels and descriptions of variants of the Hodge approach

Label	$\lambda_{ihk}^j\ \forall i, h, k \in V, j = 1, \ldots, p.$
HODb	$1\ \forall i, h, k, j$
HODr	$1 - \overline{R}_{ihk}^j$
HODp	0 iff $i, h, k \in V^{(l)}$; $\frac{2\pi}{n} \mid \tau_{hj} - \tau_{kj} \mid$ iff $h, k \in V^{(l)}, i = l$
HODs	0 iff $i, h, k \in V^{(l)}$; $\mid \theta_{hj} - \theta_{kj} \mid$ iff $h, k \in V^{(l)}, i = l$

where $\overline{R}_{ihk}^j$ is the mean resultant length of θ_{ij}, θ_{hj} and θ_{kj} (see [22] p. 17).

The selection of the element l in **HODp** and **HODs** can be done in different ways. Using a rule similar to the one used in the theorem, we propose to select $l_0 = \arg\max_l \gamma_l$ where $\gamma_l = \sum_{hk}(\overline{\psi}_{lhk})^2$.

In [25] many advantages of the Hodge approach are commented. We can emphasize that the Hodge approach is much better adapted to incomplete and unbalanced data sets. These advantages also appear in the circular case. In fact, we add as an additional advantage the use of γ_l to select l_0 which derives in a method that is computationally effective.

We also want to stress that many other alternatives can be also considered in the definition of Ψ^j such as those based on probabilistic arguments [32] or on the Bradley-Terry extension [33].

5 NUMERICAL STUDIES

In this section, examples and numerical studies are considered to compare the different techniques and their variants. The complete list of methods is given in Table 3 below.

TABLE 3
Labels for different Circular Order Aggregation Methods

Label	Description
TSPb	TSP binary length edges
TSPp	TSP positions length edges
TSP1	TSP Symmetric distance
TSP3	TSP Penalized distance with $\alpha = 3$
HODb	Hodge binary triplewise flows
HODr	Hodge resultant length triplewise flows.
HODp	Hodge position-based triplewise flows
HODs	Hodge score-based triplewise flows

Table 4 contains several simple examples showing how the performance of the methods depends on the scenario. The table includes the data and the MSCE (3) for the circular order given in Step 1 (as the second step only makes local changes in the order given by the first step, similar conclusions would have been obtained also with two steps approaches if the number of items to be ordered is increased). The non-optimal solutions obtained for each example appear in bold in Table 4. From these results we can conclude that there is not an universal winner and that several alternatives using the TSP and Hodge approaches should be considered. This strategy will be better illustrated in the application.

In the rest of the Section, we compare the results of different methods proposed using randomly generated data. We use von Mises distribution (see [22] for its definition) as it is the most widely used in circular data. We will assume $\theta_{ij} \sim^{independent} M(\phi_{ij}, \kappa_j)$ with $i = 1, \ldots, n$, $j = 1, \ldots, 6$. We have considered two artificial (EQGR with equally spaced values of ϕ_{ij} in sector circles and EXAM extending the first example in Table 4) and one "real" scenario. The real scenario use the estimated values of the κ values and phases angles from *S. cerevisiae* data analyzed in Section 6. The parameter values for the different scenarios are given in Table 5. Many other scenarios have been considered. As similar results to those appearing here were obtained, we only detail the most significant ones in order to simplify the exposition.

In Table 5 a total of 36 scenarios are described. For each escenario and each of the two objective functions (4) and (5) we performed 200 numerical simulations. In this section we show the most relevant results for these scenarios. Some more results appear in the Supplementary Material. The rest of the results would lengthen the paper unnecessarily as they would lead to the same conclusions that we expose below.

Figure 2 shows the boxplots for the 200 values from the first step of MSCE, CKτ and computational time in seconds for the different aggregation methods considered. It can be seen that **TSPp** is the method with poorest results in terms of the MSCE, while **TSP1** is the worst one when CKτ is the objetive function. **HODr** is the most computationally expensive while the rest of the methods are executed in less than 0.6 seconds being the two fastest **HODp** and **HODs**.

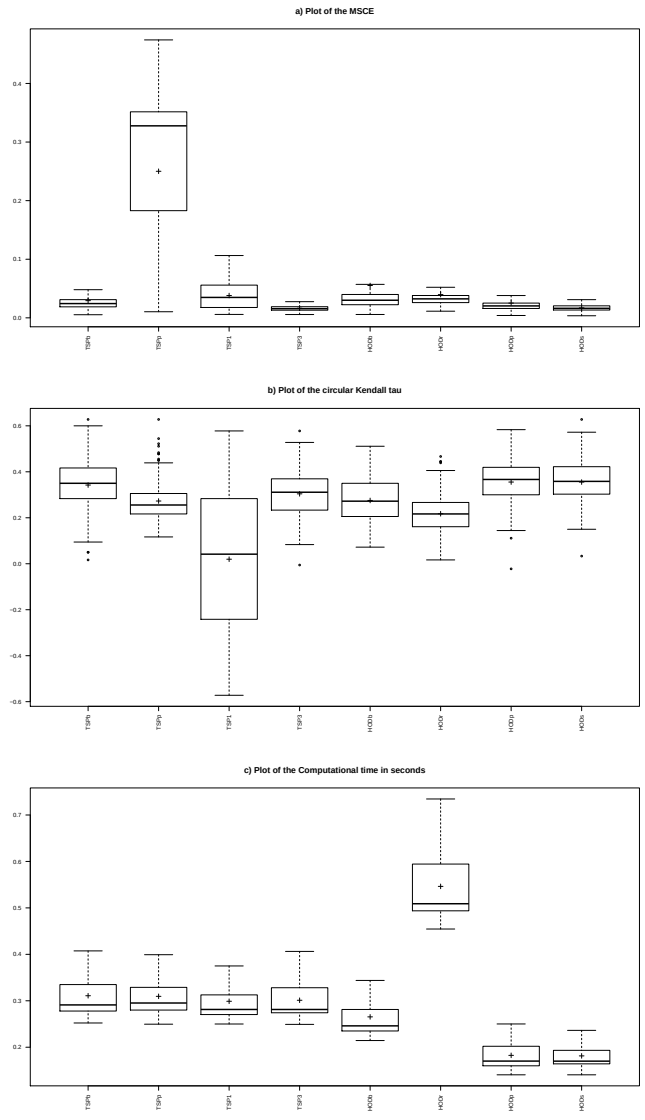

Fig. 2. Values of MSCE (a), CKτ (b) and joint computational time for MSCE and CKτ (c) from the first step for the scenario EQGR n=10 p=6 κ_j=8

Then, in Figure 3 an analysis of the behavior of the criteria when the number of elements n is increased can be seen. The figure shows the mean values of MSCE (a), CKτ (b) and computational time in seconds (c) for the best methods of each technique, namely **TSPb**, **TSP3**, **HODp** and **HODs**. **TSP1** has been also considered as it yieled better results than **TSPb** in other scenarios not detailed here. **HODp** and **HODs** were the best Hodge methods in all the explored scenarios.

We can observe the evolution of the mean values of the MSCE when n increases in Figure 3 (a). It is obvious that **TSP3** and **HODs** are the most stable methods and that they give the best approximation to the optimum in terms of MSCE. In Figure 3 (b) we can see that the two best methods when the objective function is CKτ are clearly those that use the Hodge approach, **HODp** and **HODs**. Figure 3 (c) shows that, although both approaches have a reasonable execution time, the TSP approach is moderately faster.

TABLE 4
Performance of Circular Order Aggregation methods in Simple Examples. MSCE values for step 1

DATA: p vectors in $\mathbb{R}^n$	TSPb	TSPp	TSP1	TSP3	HODb	HODr	HODp	HODs
$(0, 1/10, 1/9)\pi$ $(0, 1/4, 1/10)\pi$	**0.0041**	**0.0041**	0.00001	**0.0041**	**0.0041**	0.00001	**0.0041**	0.00001
$(0, 3/4, 1/2)\pi$ $(0, 11/6, 1/2)\pi$	0	0	0	0	0	0	0	**0.0367**
$(0, 1/4, 3/4, 5/4)\pi$ $(0, 1, 3/4, 5/4)\pi$ $(0, 3/4, 1/4, 7/4)\pi$	0.0488	0.0488	**0.0615**	0.0488	0.0488	0.0488	0.0488	0.0488
$(0, 7/11, 9/11, 8/5, 9/5)\pi$ $(0, 8/11, 1/2, 3/2, 1/20)\pi$	**0.0179**	**0.0179**	**0.1349**	0.0087	**0.0224**	**0.0179**	**0.0132**	0.0087
$(0, 3/4, 1/2, 1, 11/10, 10/9)\pi$ $(0, 11/6, 1/2, 1, 5/4, 11/10)\pi$	**0.0335**	**0.0381**	**0.0335**	**0.0335**	0.0230	**0.1202**	0.0230	**0.0335**

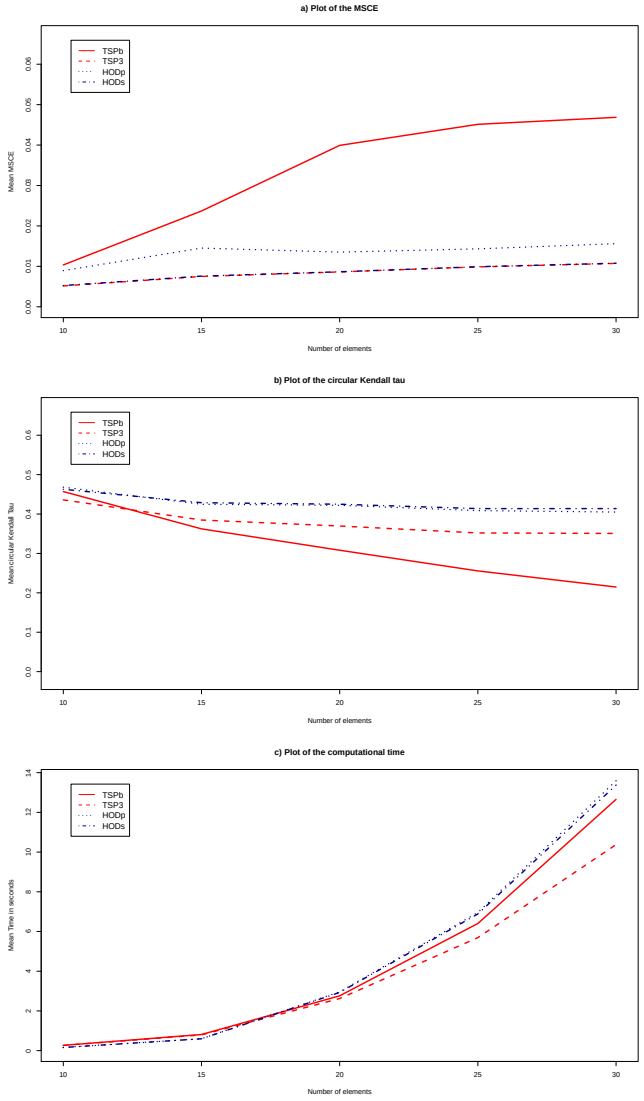

Fig. 3. Evolution of the mean values of MSCE, CKτ and joint time for MSCE and CKτ in the first step for the pattern EQGR p=6 κ_j=20 when the number of elements n to be ordered increases

TABLE 5
Parameter configuration for the simulation scenarios

Scenario	n	Φ	κ_j
EQGR	10,15,20	$\phi_{ij} = i\frac{\pi/4}{n/2}$ $1 \le i \le [n/2]$; $\phi_{ij} = i\frac{\pi/8}{n/2} + \pi$ $[n/2] < i \le n$; $1 \le j \le p$	8,20
EXAM	25,30,35	$\boldsymbol{\phi_j} = (0, \frac{1}{10}, \frac{1}{9})\pi$ $1 \le j \le [p/2]$; $\boldsymbol{\phi_j} = (0, \frac{1}{4}, \frac{1}{10})\pi$ $[p/2] < j \le p$	
REAL		Taken from *S. cerevisiae* data in Section 6	

The computational effort of all these methods obviously increases with both the values of n and p. We want to stress that we have observed that this effort increases much more when the MSCE is computed so that we recommend to consider CKτ when n or p is high.

Now, we study in Figures 4 and 5 the improvement obtained with the Circular Local Minimization Algorithm (CMLA) performed in step 2 and the increase in the computation time due to this second step. We considered the 5 scenarios in Table 5 with n=10. The mean values of the MSCE and computation time after each step for the different methods are represented in Figure 4. In this case we dropped **TSPp** as its inclusion would hide the differences among the rest of the methods due to scale problems (recall from Figure 2 (a) that it was by far the worst method under MSCE for the EQGR scenario). In Figure 5 the same type of analysis is done for CKτ.

We can observe that both the reduction of the mean values of the MSCE and the increase in the mean values of CKτ due to the CLMA are higher when the values in the first step are not too good. On the other hand, this second step does not improve too much the results for the methods whose step 1 already yielded good results (such as **TSP3** or **HODs** under MSCE). We can also see that this CLMA increases the computational effort in all cases and more significantly if the MSCE criterion is used. For these reasons, we can say that CLMA is suggested just for those situations where it is convenient to refine the initial solution.

As final conclusions from the numerical studies we can say that, although there is not an absolute winner method, in most of the situations the best methods are **TSP3** and **HODs** if the MSCE criterion is considered and **HODs** or **HODp** if the criterion used is CKτ. Moreover, the second step of the proposed methodology (CLMA), can be useful to refine the approximation but it is not worthy in most of the cases due

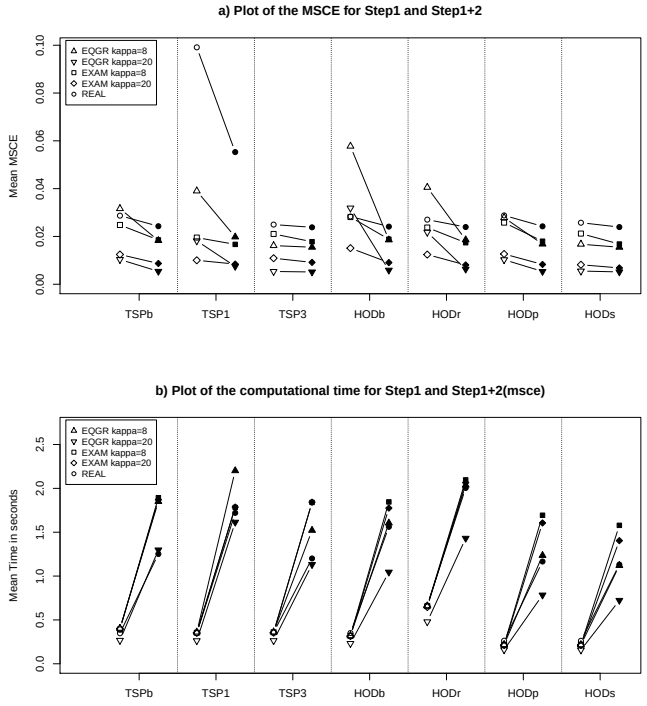

Fig. 4. Mean values of MSCE (a) and computational time in seconds (b) for step 1 (without filling) and after step 2 (filling in black) for the scenarios with n=10 p=6

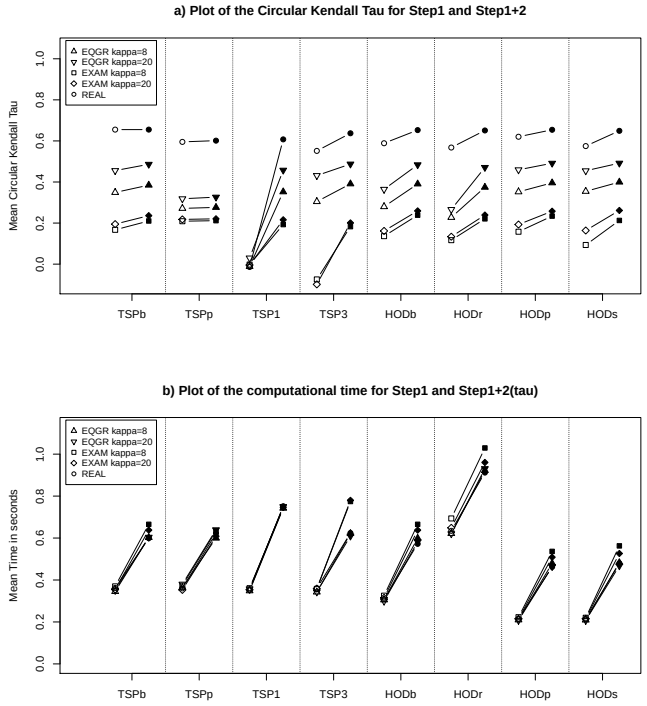

Fig. 5. Mean values of CKτ (a) and computational time in seconds (b) for step 1 (without filling) and after step 2 (filling in black) for the scenarios with n=10 p=6

to the increment in execution time.

6 Ordering cell-cycle genes from heterogenous experiments

According to [34], genes participating in a cell division cycle have a cyclical pattern of expression with peak (phase angle) attained just before their function. Therefore, these phase angles, that are circular parameters, are expected to be ordered on the circle according to the biological functions of the genes. For this reason, biologists are interested in estimating the order of the phase angles to know the order of activation of the genes. However, the task is not easy as the data available comes from experiments performed in different laboratories using different technologies (for example the cells may be arrested at different points of the cell cycle at the start of the experiment). Due to these heterogeneities significantly different estimates for the phase angle of a given gene are obtained from the experiments considered.

In this paper we use data from 10 experiments for *S. pombe* [35, 36, 37], 6 experiments for *S. cerevisiae* [38, 39, 40, 41] and $p = 4$ experiments for *humans* (HeLa cells) [42]. These data, which have also been considered in other papers [23, 24, 26], are publicly available in Cyclebase [43], which is an online database (www.cyclebase.org) that offers results from genome-wide cell-cycle-related experiments. As in [24], for *S. pombe* we consider 34 genes with a high periodicity level and their corresponding *S. cerevisiae* orthologs. For *humans*, as in [26], we consider 11 genes also with high periodicity level and with orthologs in both yeasts. The names of all these genes can be found in the Supplementary Material of the paper. The phase angle estimators for these genes were obtained from the Random Periods Model (RPM), a nonlinear regression model for estimating the peak expression of a cell-cycle gene from its cyclical pattern of expression [44]. The weight of each experiment is assigned depending on the data variability. For these weights, we refer the reader to [24] where the values used here were derived.

Tables 6, 7 and 8 show the circular order aggregation results using the methods **TSP3**, **HODp** and **HODs** recommended in Section 5. We also compare them with the results obtained with the orders given by Cyclebase for each of the species. If we globally compare our results with the cyclebase order we can see that there are significant improvements for each of the three species which means that the orders obtained with our methodology may lead to relevant biological hypotheses. These orders are also detailed in the Supplementary Material.

TABLE 6
S.pombe. 34 genes (orthologs with *S. cerevisiae*)

	TAU		MSCE	
	Step 1	Step 2	Step 1	Step 2
TSP3 Order	0.128	0.445	0.084	0.077
HODp Order	0.657	0.686	0.081	0.076
HODs Order	0.503	0.673	0.083	0.075
Cyclebase Order	0.103		0.092	

The performance of the methods is globally very good as for both criteria (CKτ and MSCE) the results improve

TABLE 7
S. cerevisiae. 34 genes (orthologs with *S. pombe*)

	TAU		MSCE	
	Step 1	Step 2	Step 1	Step 2
TSP3 Order	0.724	0.819	0.030	0.028
HODp Order	0.778	0.816	0.029	0.028
HODs Order	0.723	0.816	0.031	0.028
Cyclebase Order	0.467		0.088	

TABLE 8
Human. 11 genes orthologs with *S. pombe* and *S. cerevisiae*

	TAU		MSCE	
	Step 1	Step 2	Step 1	Step 2
TSP3 Order	0.820	0.890	0.007	0.007
HODp Order	0.890	0.890	0.009	0.007
HODs Order	0.820	0.890	0.007	0.007
Cyclebase Order	-0.075		0.099	

significantly those obtained from the cyclebase orders. In the *S. pombe* case The TSP and Hodge approaches yield very different orders (see Supplementary Material) which suggests that there may not be a clear order among the 34 genes. For *S. cerevisiae* slightly different orders, with only one of the 34 genes changing its position when the CKτ criterion is considered, are provided by the TSP and Hodge techniques so that we may be fairly confident on the results obtained. Finally, for *humans* the same order is attained by the four methods in the second step and the improvement in the MSCE and CKτ values compared with those of cyclebase orders is really impressive.

7 Final discussion and Future Research

The problem of rank aggregation is formulated in this paper as an optimization problem with the objective function defined depending on the input information, based on the MSCE for cardinal data and based on CKτ for ordinal data. With the aim of solving that optimization problem, we have developed methods by using two different levels of processing the initial information: in pairs and in triples.

The main innovation is the use of triples, which arises naturally to measure circular association. The information on triples is represented using hypermatrices. Hodge theory is used to find the closest circular order for a given hypermatrix measuring the triplewise flows between the items to be ordered. This approach based on Hodge theory is computationally very simple and efficient. It provides results that are good in terms of the objective function and in comparison with other alternative solutions, and to what we have seen, also biologically interpretable. Regarding the TSP approach, although at first it may look not so simple, it provides very good results in terms of the MSCE in most scenarios.

As for computational issues, all the methodology and algorithms developed are implemented in R code as part of an R package called **isocir** (isotonic inference for circular data), that is available on CRAN [26]. Details about the Circular Local Minimization Algorithm (CLMA) are given in Supplementary Information. This algorithm is used as a second step to improve the approximation to the optimum. However, as it has been shown in the numerical studies, the improvement is not remarkable and increases significantly the execution time.

Apart from the theoretical development, we have illustrated the use of the methodology in practice by determining the activation order of cell-cycle genes. There are other problems in computational biology where the methodology could be applied, such as those of ordering genes along the circadian cycle [45], or hormones cycles [46].

In general terms, this methodology is applicable in studies where the interest is the order of occurrence of cyclic events. For example, in meteorology, the aim could be to order wind directions from different atmospheric phenomena or to order the spread of fires [47, 48, 49].

Finally, the ideas that we have presented here, in particular the use of the Isotonic Regression and the objective function derived, can also be of interest in the rank aggregation problem in the Euclidean context. This will be part of our future research.

Acknowledgment

This work was supported by Spanish Ministerio de Ciencia e Innovación grant (MTM2012-37129 to S.B., C.R. and M.A.F.) and Junta de Castilla y León, Consejería de Educación and the European Social Fund within the Programa Operativo Castilla y León 2007-2013 (to S.B.).

References

[1] P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 2, pp. 262–268, 1977.

[2] J. Borda, *Memorie sur les elections au scrutin*. Historie de l´Academie Royal des Science., 1781.

[3] M. Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Facsimile reprint of original published in Paris, 1972, by the Imprimerie Royale, 1785.

[4] F. Schalekamp and A. Zuylen, "Rank aggregation: Together we are strong," in *Proceedings of 11th ALENEX*, 2009, pp. 38–51.

[5] M. N. Volkovs and R. S. Zemel, "A flexible generative model for preference aggregation," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 479–488.

[6] J. Bartholdi, C. Tovey, and M. Trick, "Voting schemes for which it can be difficult to tell who won the election." *Social Choice Welfare*, vol. 6, pp. 157–165, 1989.

[7] A. Caplin and B. Nalebuff, "Aggregation and social choice: A mean voter theorem," *Econometrica*, vol. 59, no. 1, pp. 1–23, 1991.

[8] F. Hassanzadeh, "Distances on rankings: From social choice to flash memories," Ph.D. dissertation, University of Illinois, 2013.

[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th International World Wide Web Conference*, 2001, pp. 613–622.

[10] A. Shishkin, P. Zhinalieva, and K. Nikolaev, "Quality-biased ranking for queries with commercial intent," in

Proceedings of the 22nd international conference on World Wide Web companion, ser. WWW '13 Companion, 2013, pp. 1145–1148.

[11] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz, "Pairwise ranking aggregation in a crowd-sourced setting," in *Proceedings of the sixth ACM international conference on Web search and data mining*, ser. WSDM '13, 2013, pp. 193–202.

[12] A. N. Langville and C. D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton, NJ, USA: Princeton University Press, 2006.

[13] R. DeConde, S. Hawley, and S. Falcon, "Combining results of microarray experiments: A rank aggregation approach," *Statistical Applications in Genetics and Molecular Biology*, vol. 5, no. 1, pp. 1–23, 2006.

[14] V. Pihur, S. Datta, and S. Datta, "Finding common genes in multiple cancer types through meta-analysis of microarray experiments: A rank aggregation approach," *Genomics*, vol. 92, no. 6, pp. 400–403, 2008.

[15] I. Simko and D. Pechenick, "Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods," *Communications in Biometry and Crop Science*, vol. 5, no. 1, pp. 41–55, 2010.

[16] K. Kadota and K. Shimizu, "Evaluating methods for ranking differentially expressed genes applied to microarray quality control data," *BMC Bioinformatics*, vol. 12, no. 1, p. 227, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/227

[17] R. Sizemore, "Hodgerank: Applying combinatorial hodge theory to sports ranking," Ph.D. dissertation, Wake Forest University, 2013.

[18] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao, "Hodgerank on random graphs for subjective video quality assessment," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 844–857, 2012.

[19] S. Barragán, C. Rueda, M. Fernández, and S. Peddada, "Statistical framework for determining the temporal order in an oscillatory system," *Preprint*, 2015.

[20] C. Rueda, M. Fernández, S. Barragán, and S. Peddada, "Some advances in constrained inference for ordered circular parameter in oscillatory systems," in *Geometry Driven Statistics*. Wiley, 2015.

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

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

[23] 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*, vol. 104, no. 485, pp. 338–347, 2009.

[24] 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*, vol. 40, no. 7, pp. 2823–2832, 2012.

[25] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial Hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, 2011.

[26] S. Barragán, M. Fernández, C. Rueda, and S. Peddada, "isocir: An R package for constrained inference using isotonic regression for circular data, with an application to cell biology," *Journal of Statistical Software*, vol. 54, no. 4, pp. 1–17, 2013. [Online]. Available: http://www.jstatsoft.org/v54/i04/

[27] M. M. Flood, "The traveling-salesman problem," *Operations Research*, vol. 4, no. 1, pp. 61–75, 1956.

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

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

[30] J. S. Naor and R. Schwartz, "The directed circular arrangement problem," *ACM Transactions on Algorithms (TALG)*, vol. 6, no. 3, p. 47, 2010.

[31] M. Hahsler and K. Hornik, *Traveling Salesperson Problem (TSP)*, 2011, r package version 1.0-6.

[32] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, ser. Information Science and Statistics. Springer, 2004.

[33] R. A. Bradley and M. E. Terry, "Rank analisis of incomplete block designs: The method of paired comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345, 1952.

[34] J. Jensen, T. Jensen, U. Lichtenberg, S. Brunak, and P. Bork, "Co-evolution of transcriptional and post-translational cell-cycle regulation," *Nature*, vol. 443, pp. 594–597, 2006.

[35] A. Oliva, A. Rosebrock, F. Ferrezuelo, S. Pyne, H. Chen, S. Skiena, B. Futcher, and J. Leatherwood, "The cell-cycle-regulated genes of Schizosaccharomyces pombe," *PLoS. Biology*, vol. 3, pp. 1239–1260, 2005.

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

[37] G. Rustici, J. Mata, K. Kivinen, P. Lio, C. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bahler, "Periodic gene expression program of the fission yeast cell cycle," *Nature Genetics*, vol. 36, pp. 809–817, 2004.

[38] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[39] U. de Lichtenberg, R. Wernersson, T. Jensen, H. Nielsen, A. Fausboll, P. Schmidt, F. Hansen, S. Knudsen, and S. Brunak, "New weakly expressed cell cycle-regulated genes in yeast," *Yeasts*, vol. 22, no. 5, pp. 1191–1201, 2005.

[40] T. Pramila, W. Wu, S. Miles, W. Noble, and L. L. Breeden, "The forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle," *Genes & Development*, vol. 20, no. 16, pp. 2266–2278, 2006.

[41] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray

hybridization," *Molecular biology of the cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[42] M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, and P. O. Brown, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular biology of the cell*, no. 6, pp. 1977–2000, 2002.

[43] A. Santos, R. Wernersson, and L. Jensen, "Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1140–D1144, 2015. [Online]. Available: http://www.cyclebase.org/

[44] 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*, vol. 101, no. 19, pp. 7240–7245, 2004.

[45] Y. Li, G. Li, H. Wang, J. Du, and J. Yan, "Analysis of a gene regulatory cascade mediating circadian rhythm in zebrafish," *PLoS computational biology*, vol. 9, no. 2, p. e1002940, 2013.

[46] G.-J. Hendriks, D. Gaidatzis, F. Aeschimann, and H. Grosshans, "Extensive oscillatory gene expression during C. elegans larval development," *Molecular Cell*, vol. 53, no. 3, pp. 380 – 392, 2014.

[47] J. Bowers, I. Morton, and G. Mould, "Directional statistics of the wind and waves," *Applied Ocean Research*, vol. 22, pp. 13–30, 2000.

[48] E. García-Portugués, A. M. Barros, R. M. Crujeiras, W. González-Manteiga, and J. Pereira, "A test for directional-linear independence, with applications to wildfire orientation and size," *Stochastic Environmental Research and Risk Assessment*, pp. 1–15, 2013.

[49] M. D. Muñoz, A. Mata, E. Corchado, and J. M. Corchado, "(obifs) isotropic image analysis for improving a predicting agent based systems," *Expert Systems with Applications*, vol. 40, no. 12, pp. 5011 – 5020, 2013.

Sandra Barragán Sandra Barragán earned a PhD in Statistical Science from the Universidad de Valladolid in 2014. She received the Degree in Statistics also from the Universidad de Valladolid. Her research interests include Statistical Inference methods under restrictions, Circular data, and Applied Algorithms.

Cristina Rueda Cristina Rueda earned a PhD in Statistical Science from the Universidad de Valladolid in 1991. She received the BS in Mathematics also from the Universidad de Valladolid. She is currently Catedrática de Universidad at the Department of Statistics and Operational Research of the Universidad de Valladolid. Her research interests include Statistical Inference methods under restrictions, Circular data, Computational Biology, and Small Area Estimation.

Miguel A. Fernández Miguel A. Fernandez earned a PhD in Statistical Science from the Universidad de Valladolid in 1995. He received the BS in Mathematics also from the Universidad de Valladolid. He is currently Profesor Titular de Universidad at the Department of Statistics and Operational Research of the Universidad de Valladolid. His research interests include Statistical Inference methods under restrictions, Circular data, Computational Biology, and Reliability and Maintenance of Complex Systems.