Universidad de Valladolid

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA

TESIS DOCTORAL:

PROCEDIMIENTOS ESTADÍSTICOS PARA MODELOS CIRCULARES CON RESTRICCIONES DE ORDEN APLICADOS AL ANÁLISIS DE EXPRESIONES DE GENES

Presentada por Sandra Barragán Andrés para optar al grado de doctora por la Universidad de Valladolid

Dirigida por:
Dra. Cristina Rueda Sabater
Dr. Miguel A. Fernández Temprano

Cristina Rueda Sabater, Catedrática de Universidad, y Miguel A. Fernández Temprano, Profesor Titular de Universidad, certifican que la presente memoria ha sido realizada, bajo su dirección, por Sandra Barragán Andrés en el Departamento de Estadística e Investigación Operativa de la Universidad de Valladolid.

Valladolid, 27 de Junio de 2014

Agradecimientos

En primer lugar, quiero agradecer a mis directores Cristina y Miguel el haberme dado la oportunidad de trabajar junto a ellos en un proyecto como este y también el haber sido la combinación perfecta para que esta tesis llegue al final. Gracias por lo mucho que me habéis enseñado y todo el esfuerzo y tiempo dedicados. Agradecer también al resto de miembros del departamento de Estadística e Investigación Operativa su acogida durante estos años.

I want to express my sincere gratitute to Shyamal D. Peddada, I can't forget all his invaluable help and guidance regarding the application problem but also his nice words of support during the last years.

Cabe agradecer el haber tenido la oportunidad de disfrutar de las ayudas para personal investigador que me han permitido realizar esta tesis.

De manera general, gracias a todos aquellos que me han enseñado a aprender y que han formado parte de mi vida en estos últimos años. Aunque no me resulta posible nombraros a todos, bien os merecéis como mínimo una mención en muestra de mi gratitud. Comenzando por ese formidable (más en calidad que en tamaño) grupo de doctorandos, gracias por echarme una mano (a veces el brazo entero) en esas ocasiones que es mejor dejar reposar el trabajo; sin olvidar a las que habéis sabido seguir ahí desde sitios remotos. Gracias, como no, a ese grupo de salseros que le ha dado sabor (ritmo y allegria) a mis horas libres. Gracias a esas compis de piso por hacer del mismo un hogar antiangustias y repleto de historietas inolvidables. Gracias a la hinchada de los juegos por (entre otras cosas) esos órdagos tan emocionantes. Gracias a mis colegialas de honor santa crucenses porque a pesar del paso del tiempo seguís haciendo que el número que más me emocione sea el tres-cinco. Gracias a mi gente de León por esa amistad eterna que sirve de ancla y de multiplicador de alegrías. Y gracias muy especialmente al responsable de una gran parte de mis momentos de felicidad, por saber estar siempre a mi lado tanto para disfrutar de las buenas ocasiones como para darme fuerza en las no tan buenas.

Por último, quiero dedicar esta tesis y mi más sincero agradecimiento a mis padres por su paciencia y apoyo incondicional en todas mis decisiones y a mis abuelos por todo lo que me transmitieron.

La gratitud es la memoria del corazón.

-Lao Tsé-

Resumen

En los últimos años ha surgido un interés creciente por el estudio de problemas estadísticos en modelos circulares con restricciones. Esta tesis supone una contribución a esta línea de investigación. En concreto, aborda por un lado, resolver el problema teórico de agregación de órdenes circulares y por otro lado el contraste de igualdad de órdenes circulares en dos o más poblaciones. El principal reto es realizar un tratamiento correcto de las características del espacio circular en el que se encuentran los datos. En este trabajo se presentan y evalúan diferentes métodos para agregación de órdenes circulares, algunos que utilizan técnicas novedosas como la que hace uso del problema del viajante (TSP). En relación al segundo problema, se propone un test no paramétrico para realizar el contraste de igualdad de órdenes circulares en varias poblaciones. Este procedimiento se basa un proceso de selección aleatoria sobre los experimentos y hace uso de un estadístico estandarizado. De esta manera se tiene en cuenta posibles diferencias en la variabilidad entre y dentro de las poblaciones y el hecho de disponer de muestras no balanceadas.

El estudio de estos problemas surge de la necesidad de resolver varias cuestiones en el ámbito de la biología molecular a partir del análisis de datos de expresiones de genes asociados al ciclo celular. En concreto, se trata de identificar aquellos genes cuyas funciones biológicas se han conservado en el proceso evolutivo.
Como resultado de aplicar los procedimientos desarrollados en esta memoria se ha conseguido determinar un conjunto de 7 genes cuyos momentos de máxima expresión siguen el mismo orden a lo largo del ciclo celular de las tres especies comparadas (*S.pombe*, *S.cerevisiae* y Humanos), lo que significa que conservan sus funciones a lo largo del proceso evolutivo.

Finalmente, con objeto de facilitar al usuario la utilización de los procedimientos diseñados se ha implementado en el lenguaje R toda la metodología desarrollada, dando lugar a un paquete denominado isocir (*isotonic inference for circular data*) que se encuentra disponible en el CRAN. Este paquete incluye también la implementación de procedimientos básicos de la inferencia con restricciones en modelos circulares que no habían sido implementados anteriormente en R.

Abstract

In recent years there has been a considerable interest in drawing inferences regarding order relationships among circular parameters. In particular, the aim in this work is to solve the theoretical problems regarding aggregation of circular orders and testing equality of circular orders in two or more populations. Due to the circular geometry, the suitable treatment of the data is challenging. In this thesis, we present and compare various methods to aggregate the observed information resulting in the aggregate circular order. Some of these methods used original techniques such as the one based on the traveling salesman problem. In relation to the second problem, we propose a non-parametric test for testing the equality of circular orders in various populations. This procedure is based on random selection of the experiments and a standardized statistic. This procedure takes into consideration diferences within and among populations and unbalanced samples.

The study of these statistical problems arises due to the lack of methodology in the literature to solve a problem encountered in the molecular biology field. Specifically, the biological problem is the identification of the maximum set of genes whose biological functions are evolutionarily conserved. As a result of the analysis by using the methodology developed in this work, we discover a set of 7 cell-cycle genes whose order among the peak expressions is conserved across three species (*S.pombe*, *S.cerevisiae* and Humans), then, they conserved their functions evolutionarily.

Finally, we have implemented in R all the methodology proposed here resulting in a user-friendly software which is part of the package called isocir (isotonic inference for circular data). This package contains also the basic methods for circular models with order restrictions and is publicly available in the CRAN.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

The secret to getting ahead is getting started.
Mark Twain

1.1. Motivación y objetivos

El interés científico en los sistemas biológicos y la interpretación de su funcionamiento ha crecido de manera notable en las últimas décadas. La secuenciación del ADN - conocer la secuencia que corresponde a cada gen - ha jugado un papel fundamental en esta cuestión. En la actualidad uno de los grandes desafíos es descubrir la función biológica de cada gen y establecer las relaciones existentes entre genes. Los estudios genómicos encaminados a avanzar en este problema se caracterizan por su interdisciplinariedad, ya que requieren utilizar tanto conocimientos biológicos como estadísticos e informáticos. La aportación que realizamos en este trabajo es metodológica, desarrollando procedimientos estadísticos que se enmarcan en el contexto del análisis de datos circulares con restricciones de orden y que resuelven problemas en este ámbito de la biología molecular, en concreto del análisis de expresiones de genes.

En términos generales, el análisis de la expresión de los genes sirve para alcanzar diferentes objetivos dependiendo del estudio biológico. Por ejemplo, un

oncólogo puede estar interesado en las variaciones de las expresiones de los genes relacionadas con un cierto tipo de cáncer, para su detección y tratamiento. En otro estudio, un endocrino puede interesarse en cómo responde la expresión ante un aumento del estrés en diferentes momentos del ciclo menstrual.
En nuestro caso, el objetivo biológico que nos ocupa es identificar aquellos genes cuyas funciones biológicas se han conservado en el proceso evolutivo (Jensen et al. (2006)). Con ese fin, se consideran datos de expresiones de un conjunto de genes asociados al ciclo celular y que son comunes a diferentes especies. La función biológica de un gen está directamente relacionada con el momento máximo de su nivel de expresión, el momento en que dicho gen se activa. Si cierto conjunto de genes conserva el mismo orden de activación en diferentes especies, se entiende que la función biológica de dichos genes se mantiene evolutivamente. Consideraremos únicamente aquellos genes cuyo patrón de expresion es claramente periódico y por tanto se puede hablar de un único momento de máxima expresión asociado a dicho gen.

La principal característica de los datos a analizar es que son puntos en el círculo. Esto se debe a que los momentos de máxima expresión son medidas angulares en el ciclo celular. Para analizar adecuadamente este tipo de datos, tanto la definición de los estadísticos más simples como los procedimientos más complejos se deben adaptar a la estructura específica de la geometría circular. Los elementos básicos del análisis de datos circulares (Mardia y Jupp (2000), Fisher (1993)) se introducen en los preliminares de esta memoria.
La otra cuestión relevante a tener en cuenta es el orden en el que participan los genes en el ciclo celular. Esto implica la necesidad de utilizar técnicas propias de la Inferencia Con Restricciones (Robertson et al. (1988), Silvapulle y Sen (2005)). El orden entre los genes se representa matemáticamente mediante restricciones sobre los parámetros (momentos de máxima expresión) del modelo. Este área de la estadística es poco conocida por lo que también se introducen las bases de la metodología en los preliminares.

Podríamos decir que los procedimientos estadísticos desarrollados a raíz de los problemas que han surgido en el análisis de la expresión de genes asociados al ciclo celular han dado lugar a una nueva rama de la estadística que denominamos Inferencia Con Restricciones en modelos circulares. Las diferencias geométricas existentes entre el círculo y la línea son el motivo por el que la adaptación de los métodos ya existentes en ICR en modelos euclídeos no es ni mucho menos inmediata. Los trabajos pioneros en este ámbito son: Rueda et al. (2009) y Fernández et al. (2012). En el primero se resuelve el problema de la definición y obtención del estimador de regresión isotónica circular (CIRE) y en el segundo se resuelve el problema del contraste de un orden circular dado mediante un test condicional. Un resumen de los resultados más relevantes en estos dos artículos pioneros puede encontrarse también en los preliminares de esta memoria.

El problema biológico que nos ocupa de comparar el orden de activación de un conjunto de genes en diferentes especies, se traduce estadísticamente, en un problema de contraste de hipótesis. No existe ningún procedimiento, hasta lo que nosotros conocemos, que resuelva este problema convenientemente dadas las características de los datos y del problema. En el presente trabajo se desarrolla la metodología adecuada, enmarcada en el análisis de datos circulares con restricciones. Como parte de dicha metodología surge el problema de estimar el orden circular a partir de la información de experimentos heterogéneos. Este problema que se puede formular como el de buscar el orden circular que mejor represente toda la información disponible, se encuadra dentro de la problemática de la agregación de órdenes y será el problema que se aborden en primer lugar en este trabajo. Al igual que el problema de contraste de órdenes circulares, la agregación de órdenes circulares es un problema inédito en la literatura. El antecedente más cercano es el estudio del problema de agregación de rankings en la línea. Una revisión de los métodos más populares para resolver este último problema se incluye en los preliminares de la tesis. De hecho, la adaptación de los procedimientos en la línea a la geometría del círculo será la

primera dirección que tomemos para resolver el problema.

El Capítulo 3 de este trabajo se dedica al estudio de los métodos de agregación de órdenes circulares, dónde se introducen los métodos adaptados de la línea y se diseñan otros métodos, mucho más novedosos, que hacen uso de ideas originales, propias de otras áreas de investigación, y que se adaptan a las características de los datos circulares. Se pueden destacar como propuestas más novedosas las dos siguientes: el enfoque que considera el problema del viajante (*Traveling Salesman Problem*, TSP Lawler et al. (1985)), un problema de optimización planteado habitualmente en áreas como la Investigación Operativa y la Logística cuya idea principal es la búsqueda de la ruta más corta a través de un conjunto de localizaciones; y otro enfoque de resolución que hace uso de la teoría de Hodge que se encuadra en la topología algebraica y que sólo recientemente ha comenzado a ser usada en el análisis de datos. Se presenta brevemente esta teoría en los preliminares y en el Apéndice A.
Las alternativas dentro de cada enfoque generan una variedad importante de métodos. Las simulaciones y los ejemplos numéricos ponen de manifiesto las ventajas y desventajas de cada perspectiva y señalan al TSP como candidato ganador.

El otro problema que se resuelve en esta tesis, como ya se menciona anteriormente, es el contraste de igualdad de órdenes circulares. En concreto, se plantea la hipótesis nula de la igualdad de los órdenes circulares de S poblaciones. Para resolverlo se propone un enfoque no paramétrico que tiene en cuenta la posibilidad de muestras no balanceadas y las diferencias en las características, además de la localización, entre las poblaciones. Debido a estas circunstancias un test estándar de permutaciones no funciona correctamente por lo que se diseña un procedimiento específico de remuestreo tipo bootstrap y se proponen diferentes estadísticos test. La validación y comparación de las alternativas se realiza mediante simulaciones tanto bajo la hipótesis nula como bajo la alternativa. El procedimiento final propuesto es el que tiene un mejor comportamiento para la potencia. Además, del procedimiento de remuestreo,

se obtienen otras salidas interesantes como son: un estimador del orden circular común a todas las poblaciones y un índice de confianza de dicho orden.

Por otro lado hay que destacar el trabajo realizado para la implementación de los métodos. A lo largo de nuestra investigación hemos desarrollado un paquete de programación en el lenguaje R, (R Core Team (2014)), llamado **isocir** (**iso**tonic inference for **cir**cular data) que presentamos en el Capítulo 6. La elección del lenguaje R se justifica fácilmente por la gran difusión y transparencia que este tiene al tratarse de software libre. Esto permite que sea accesible no sólo a los profesionales de la estadística sino a investigadores de cualquier campo. Algunos algoritmos se han programado en C++ para conseguir un tiempo más corto de ejecución. En ese caso, se ha hecho uso de los conocidos objetos `SEXP` de R y la interacción mediante la función `.Call` para mantener coherencia y que la interfaz con el usuario sea R en todo momento. El trabajo desarrollado en este aspecto computacional incluye también la implementación de los métodos desarrollados en Rueda et al. (2009) y Fernández et al. (2012) y de los nuevos métodos desarrollados en este trabajo. El material del paquete **isocir** se encuentra actualizado en el CRAN: `http://cran.r-project.org/web/packages/isocir/` y ha dado lugar a la publicación de un artículo, ver Barragán et al. (2013).

La metodología presentada en esta memoria está motivada por el análisis de expresión de genes asociados al ciclo celular. En este contexto se plantean diferentes cuestiones a cuya solución se ha dedicado una importante parte del trabajo. El problema biológico y las soluciones al mismo se presentan en el Capítulo 5. El objetivo principal es identificar aquellos genes asociados al ciclo celular que mantengan el mismo orden de activación en tres especies esenciales en la evolución: dos tipos de levaduras y los humanos. Se dispone para ello de las medidas de los niveles de expresión en un total de 20 experimentos de los cuales: 10 experimentos tienen datos de la levadura *S.pombe* (Rustici et al. (2004), Oliva et al. (2005), Peng et al. (2005)); 6 experimentos de la levadura

S.cerevisiae (Cho et al. (1998), de Lichtenberg et al. (2005), Pramila et al. (2006), Spellman et al. (1998)); y 4 experimentos de células humanas HeLa (Whitfield et al. (2002)). Entre experimentos hay heterogeneidad debido a diferencias tanto en los momentos del ciclo en los que se inician las mediciones como en la variabilidad por haber sido realizados en diferentes laboratorios e incluso con técnicas distintas. Se seleccionan 11 genes comunes a estas tres especies que comparten las secuencias de ADN y un comportamiento periódico a lo largo del ciclo celular. Esta información se obtiene de la base de datos Cyclebase, Gauthier et al. (2008). Los momentos del ciclo celular en los que ocurre la máxima expresión se estiman mediante el llamado *Random Periods Model* (Liu et al. (2004)) y componen el conjunto de datos circulares con el que trabajamos. A partir del análisis de las expresiones de los genes que son periódicos en estas 3 especies y utilizando la metodología desarrollada, se descubre un conjunto de 7 genes que comparten el orden en las 3 especies lo que significa un avance en el conocimiento de la evolución de las especies, de la función biológica de los genes y de las relaciones entre los mismos.

1.2. Contribuciones de esta tesis

Esta tesis supone un avance en el análisis de modelos circulares con restricciones, principalmente en los siguientes puntos:

- Definición y análisis de métodos de agregación y estimación de órdenes circulares.
- Desarrollo del contraste de la igualdad entre órdenes circulares de diferentes poblaciones.
- Resolución de problemas de la biología molecular enmarcados en el análisis de expresiones de genes asociados al ciclo celular.
- Implementación en R de los métodos desarrollados para el análisis de modelos circulares con restricciones para su ejecución mediante el paquete **isocir**.

1.3. Estructura de esta memoria

En este primer capítulo se ofrece una visión general de la motivación de los problemas planteados; los objetivos y cómo se alcanzan. En el Capítulo 2, se encuentran los preliminares de cada área utilizada: el análisis de datos circulares, la inferencia con restricciones en el espacio euclídeo y circular y la agregación de rankings. El problema de la agregación de un orden circular se presenta en el Capítulo 3. Se exponen diferentes métodos para su resolución y se comparan mediante estudios de simulación y ejemplos. En el Capítulo 4, se plantea y resuelve el contraste de igualdad de órdenes circulares mediante un procedimiento de remuestreo. En el Capítulo 5, se resuelve la aplicación que motivó el desarrollo de estos métodos, es decir el análisis de la expresión de genes asociados al ciclo celular comparando diferentes especies. Dedicamos el Capítulo 6 a mostrar el uso de la implementación en el paquete de R **isocir** de toda la metodología de los capítulos tercero y cuarto así como métodos anteriores a esta tesis. Por último, el Capítulo 7 se dedica a conclusiones.

Capítulo 2

Preliminares

Success depends upon previous preparation,
and without such preparation
there is sure to be failure.
Confucio

En este capítulo se introducen conceptos de diferentes ramas de la Estadística que, como ya se comenta en la introducción, son necesarias para el desarrollo de la metodología que se elabora posteriormente en este trabajo. En primer lugar, en la Sección 2.1 se presenta la notación y terminología que aparece a lo largo de la memoria. A continuación, en la Sección 2.2 se hace una revisión de los conceptos básicos en el análisis de datos circulares, la Sección 2.3 se dedica a la revisión de los conceptos y procedimientos básicos de la Inferencia con Restricciones y en la Sección 2.4 se resumen los avances realizados hasta la fecha en el análisis de modelos circulares con restricciones. Finalmente, en la Sección 2.5 se presentan los métodos de agregación de rankings en la línea.

2.1. Notación y terminología

Los conjuntos de datos con los que trabajamos están compuestos por observaciones circulares asociadas a n variables circulares cuyo parámetro de interés son las direcciones medias. Sea $\mathbf{\Phi}_s = (\phi_{1s}, \ldots, \phi_{is}, \ldots, \phi_{ns})'$, $s = 1, \ldots, S$,

$i = 1, \ldots, n$ el vector de direcciones medias poblacionales donde ϕ_{is} es la dirección media de la variable i en la población s.
Sea $\boldsymbol{\Theta_s} = (\boldsymbol{\Theta}_{1s}, \ldots, \boldsymbol{\Theta}_{js}, \ldots, \boldsymbol{\Theta}_{p_s s})'$, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$ el conjunto de datos observados en los p_s experimentos relativos a la población s, donde el vector $\boldsymbol{\Theta_{js}} = (\theta_{1js}, \ldots, \theta_{ijs}, \ldots, \theta_{njs})'$ contiene las observaciones obtenidas para dichas n variables en el experimento j de la población s. Denotaremos como p al número total de experimentos, $p = \sum_{s=1}^{S} p_s$.

En nuestra aplicación, cada población corresponde a una especie y las variables de interés son los momentos de máxima expresión de los n genes considerados. Las observaciones son datos circulares puesto que dichos momentos de máxima expresión se registran como ángulos en la circunferencia a partir del momento de arranque del experimento. Dado que uno de los objetivos de esta memoria es la construcción de un orden circular entre los genes a partir de la información de varios experimentos, y que en la literatura de agregación de órdenes en la línea se suele hablar de orden entre elementos, con frecuencia nos referiremos a los genes como los elementos a ordenar.
Por otro lado, con objeto de simplificar las expresiones matemáticas, y siempre que no haya posibles confusiones, porque tratemos una única población ó un único experimento, eliminaremos los subíndices s y j correspondientes.

Denotaremos por $\mathcal{O}$ al conjunto de todos los órdenes circulares posibles entre n elementos. Así, si $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ es un vector de observaciones angulares, $\eta_i \in [0, 2\pi)$, $i = 1, \ldots, n$. El orden que siguen los elementos de este vector se denotará por $\boldsymbol{O_\eta} = (o_1, \ldots, o_l, \ldots, o_n)' \in \mathcal{O}$, donde $o_l = i$ si el elemento i-ésimo se encuentra en la posición l-ésima del orden. En este caso diremos que $\boldsymbol{\eta}$ verifica el orden $\boldsymbol{O_\eta}$ y lo formularemos como $\boldsymbol{\eta} \circlearrowleft \boldsymbol{O_\eta}$. En relación al problema de agregación de órdenes, también son de interés las posiciones que ocupan cada una de las componentes de un vector en el orden. Para este mismo vector $\boldsymbol{\eta}$ denotaremos las posiciones que ocupan sus elementos en el orden $\boldsymbol{O_\eta}$ mediante el vector $\boldsymbol{T_\eta} = (\tau_1, \ldots, \tau_n)'$ donde $\tau_i = k$, si la componente η_i se encuentra en la posición k del orden circular $\boldsymbol{O_\eta}$ (es decir si $o_k = i$). Así, por

ejemplo, si $\boldsymbol{\eta} = (2.47, 0.56, 4.92, 1.23, 6.1)$ tendremos que $\boldsymbol{O_\eta} = (2, 4, 1, 3, 5)'$ y $\boldsymbol{T_\eta} = (3, 1, 4, 2, 5)'$.

Hay que tener en cuenta que, debido al carácter cíclico de la aplicación para la que se desarrolla la metodología, y muy probablemente de otras aplicaciones, los órdenes que se definen serán invariantes frente a la rotación (es decir independientes del origen del círculo) con lo que $\boldsymbol{O}_\eta = (o_1, \ldots, o_l, \ldots, o_n)' \equiv \boldsymbol{O^l_\eta} = (o_l, \ldots, o_n, o_1, \ldots, o_{l-1})', \forall l \in \{1, \ldots, n\}$. Por este motivo, cuando nos refiramos al orden entre las componentes de $\boldsymbol{\eta}$ frecuentemente utilizaremos también la notación $\eta_1 \leq \cdots \leq \eta_n \leq \eta_1$. Notar también que como consecuencia de esta relación de equivalencia $\#\mathcal{O} = (n-1)!$.

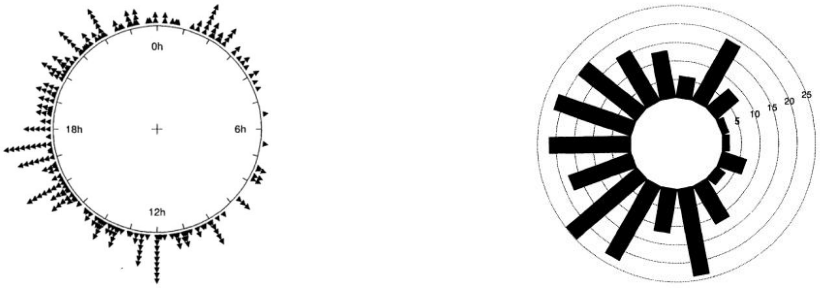
2.2. Revisión del Análisis de Datos Circulares

Los datos circulares tienen características singulares debidas al espacio en el que se encuentran, por lo que son necesarios modelos y métodos específicos para su análisis. Un dato circular es un punto en la circunferencia unidad o equivalentemente un vector de dirección en el plano. Se representa mediante el ángulo entre la dirección inicial y el punto observado una vez fijada dicha dirección inicial y el sentido de rotación. Habitualmente, en la literatura, el sentido de rotación en el círculo unidad es el contrario a las agujas del reloj, antihorario, y por tanto es el que usamos a lo largo de esta memoria. Los datos circulares se pueden clasificar en dos tipos: los procedentes de la brújula - el propio espacio físico - y los procedentes del reloj. En este último se tienen observaciones en el tiempo donde el círculo representa los ciclos que se repiten periódicamente, como ocurre en un reloj analógico o en nuestra aplicación en el ciclo celular. Sea cual sea su procedencia las características son comunes así como las herramientas para su análisis.

Además de la aplicación que motiva esta memoria hay otros muchos campos en los que aparecen datos circulares. Por ejemplo, en meteorología, midiendo

las direcciones del viento, Bowers et al. (2000); en ecología, al medir la dirección del movimiento de diferentes animales, como puede ser el caso del vuelo de los pájaros desde su nido, Cochran et al. (2004); en las ciencias sociales y políticas (Haskey (1988)), criminología (Brunsdon y Corcoran (2005)), psicología (Kibiak y Jonas (2007)), análisis de imágenes (Boles y Lohmann (2003)), medicina (Simpson y Edwards (2013); Maldonado et al. (1997)).

Un ejemplo concreto de datos circulares extraído de Fisher (1993) son las llegadas de 254 pacientes a una unidad de cuidados intensivos de un cierto servicio de urgencias. Su representación mediante un diagrama circular se encuentra en la Figura 2.1a y el correspondiente histograma circular en la Figura 2.1b. En este ejemplo el espacio representado por el círculo es un supuesto reloj de 24 horas.



(a) Diagrama de las llegadas (b) Histograma circular

Figura 2.1: Representaciones de un ejemplo de datos circulares con ciclo de 24h.

En las siguientes dos subsecciones se hace una revisión de conceptos, medidas y resultados básicos para el problema que nos ocupa; así como de la distribución de von Mises dada su relevancia en el análisis de datos circulares - se trata de la equivalente a la Normal en este contexto - y además, aparece en numerosas ocasiones a lo largo de nuestro trabajo. Las referencias básicas en este campo son los libros de Mardia y Jupp (2000), Fisher (1993) y Jammalamadaka y SenGupta (2001).

2.2.1. Medidas básicas en el análisis de datos circulares

Sea $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_i, \ldots, \theta_n)'$ un vector de direcciones observadas localizadas en el círculo unidad.

Medidas de localización y concentración

Introducimos a continuación la media y mediana circulares y la longitud media resultante comentando alguna característica de las mismas.

Definición 2.1. *La media circular del vector de direcciones* $\boldsymbol{\Theta}$ *viene dada por la siguiente expresión,*

$$\overline{\theta} = Ave(\boldsymbol{\Theta}) = \begin{cases} \arctan\left(\frac{\overline{S}}{\overline{C}}\right) & si\ \overline{S} \geq 0, \overline{C} > 0 \\ \frac{\pi}{2} & si\ \overline{S} > 0, \overline{C} = 0 \\ \arctan\left(\frac{\overline{S}}{\overline{C}}\right) + \pi & si\ \overline{C} < 0 \\ \arctan\left(\frac{\overline{S}}{\overline{C}}\right) + 2\pi & si\ \overline{S} < 0, \overline{C} \geq 0 \end{cases} \tag{2.1}$$

donde,

$$\overline{S} = \sum_{i=1}^{n} \sin\theta_i; \qquad \overline{C} = \sum_{i=1}^{n} \cos\theta_i$$

Notar que si $S = C = 0$, $\overline{\theta}$ *no está definida.*

La media circular, al contrario que la media aritmética, no verifica la propiedad del valor medio de Cauchy que dice que la media de dos valores se encuentra acotada por dichos valores y que es una propiedad fundamental para la obtención de algunos resultados con datos en la línea que, como consecuencia, no se obtienen para el caso circular como veremos más adelante (Sección 2.4).

Por su parte, la mediana circular también es usada como alternativa a la media en algunos métodos de los que proponemos en el Capítulo 3. Su definición es la siguiente.

Definición 2.2. *La mediana circular del vector de direcciones* $\boldsymbol{\Theta}$ *es,*

$$\breve{\theta} = Med(\boldsymbol{\Theta}) = \arg\min_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{i=1}^{n} \{\pi - |\pi - |\theta_i - \alpha_i||\}.$$

La medida usual de concentración con respecto a la media circular es la longitud media resultante definida a continuación.

Definición 2.3. *La longitud media resultante es,*

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

Esta es una medida de la longitud de la dirección media definida en (2.1) para un vector de direcciones $\boldsymbol{\Theta}$. Toma valores cercanos a 1 en caso de que las componentes de dicho vector estén muy concentradas y cercanos a 0 en caso de que se encuentren dispersas. Notar que si $\overline{R} = 0$ no implica que los valores se encuentren dispersos. Por ejemplo, si en un vector de tamaño 20 se consideran 10 valores en $\pi/2$ y otros 10 en $3\pi/2$, se tiene $\overline{R} = 0$ y sin embargo los valores se encuentran concentrados en dos puntos opuestos en el círculo. De esta manera, ocurre para cualquier conjunto de la forma: $(\theta_1, \ldots, \theta_n, \theta_1+\pi, \ldots, \theta_n+\pi)'$.

Para finalizar las medidas de localización y concentración vemos un ejemplo muy simple extraído de Mardia y Jupp (2000). Tenemos los siguientes datos circulares: 0.75, 0.79, 0.91, 1.06, 1.31, 1.54, 1.54, 4.87, 6.23. Entonces, la media circular es $\overline{\theta}$ =0.89 y la longitud media resultante es $\overline{R}$ =0.71. En la Figura 2.2 se encuentran representadas ambas medidas así como los datos.

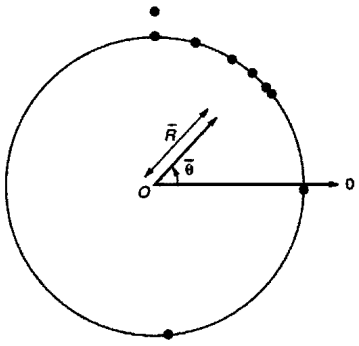


Figura 2.2: Representación de un ejemplo para la media circular y la longitud media resultante.

Medidas de distancia y asociación

Sean $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)'$ y $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)'$ dos vectores circulares. Introducimos en las siguientes definiciones las medidas de distancia y asociación entre dos vectores y respecto a un conjunto de datos. Estas medidas son de especial interés ya que la solución al problema principal que nos ocupa está relacionada con la cercanía a los datos circulares observados.

Definición 2.4. *La distancia angular entre los vectores $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ es:*

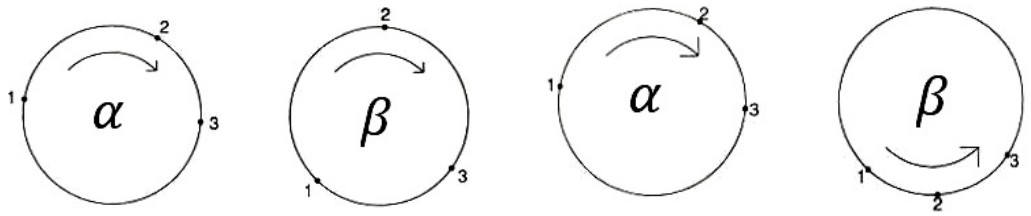
$$d(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} 1 - \cos(\alpha_i - \beta_i)$$

A continuación, se define la medida de la distancia entre un vector y un conjunto de observaciones, mediante la suma de errores circulares.

Definición 2.5. *La suma de errores circulares (SCE, Sum of Circular Errors) entre un vector $\boldsymbol{\alpha}$ y un conjunto de p vectores circulares n dimensionales $\{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_p\}$, donde $\boldsymbol{\Theta}_j = (\theta_{11}, \ldots, \theta_{nj})'$ se define como,*

$$SCE(\boldsymbol{\Theta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{p} d(\theta_{ij}, \alpha_i) = \sum_{j=1}^{p} \overline{R}_j (1 - \cos(\overline{\theta}_j - \alpha_j)).$$

Respecto a la asociación entre vectores angulares cabe decir que se mide a través de la relación entre tripletas. Esto se debe a que una tripleta es el conjunto mínimo de elementos que mantiene una relación de asociación circular. En la línea, dos elementos tienen un orden entre sí pero en el círculo es necesario considerar un tercer elemento para que el punto de inicio del círculo no influya en dicha asociación, (ver Figura 2.3, extraída de Fisher (1993)). Entre dos tripletas existe asociación positiva o de concordancia si ambas siguen el mismo sentido de rotación, bien en el sentido de las agujas del reloj o bien en el contrario, Figura 2.3a. Serán tripletas discordantes o con asociación negativa



(a) Tripletas concordantes (b) Tripletas discordantes

Figura 2.3: Tipos de asociación circular entre dos tripletas

si tienen diferentes sentidos de rotación, Figura 2.3b.

Definición 2.6. *El grado de asociación entre las dos tripletas* $(\alpha_i, \alpha_h, \alpha_k)$ *y* $(\beta_i, \beta_h, \beta_k)$ *se define como,*

$$\begin{aligned}\delta_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(i,h,k) = \; & signo(\alpha_i - \alpha_h) \cdot signo(\alpha_h - \alpha_k) \cdot signo(\alpha_k - \alpha_i) \\ & \cdot signo(\beta_i - \beta_h) \cdot signo(\beta_h - \beta_k) \cdot signo(\beta_k - \beta_i),\end{aligned}$$

En caso de asociación positiva $\delta_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = 1$ y si la asociación es negativa $\delta_{(\boldsymbol{\alpha},\boldsymbol{\beta})} = -1$.

A continuación, en base a la medida del grado de asociación entre todas las tripletas, definimos el coeficiente de correlación Tau circular de Kendall que es una medida no paramétrica de la asociación entre dos vectores circulares. La elección de este coeficiente se debe principalmente a que es el correspondiente en el círculo a la Tau de Kendall, medida esta ampliamente usada en los métodos de agregación en la línea (ver Sección 2.5).

Definición 2.7. *La Tau Circular de Kendall (Fisher (1993, p.146-147)) entre los vectores* $\boldsymbol{\alpha}$ *y* $\boldsymbol{\beta}$ *es,*

$$\hat{\Delta}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \binom{n}{3}^{-1} \sum_{1 \le i \le h \le k \le n} \delta_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(i,h,k).$$

La Tau circular de Kendall varía en el intervalo $[-1, 1]$, tomando el valor 1 cuando hay concordancia (asociación positiva) y -1 en caso de discordancia

(asociación negativa).

Existen otros muchos coeficientes de correlación definidos en la literatura que miden otros tipos de asociación, aunque no resultan de interés para nuestro caso. Algunos ejemplos son el basado en correlaciones canónicas (Jupp y Mardia (1980)), en la dependencia rotacional (Fisher y Lee (1982)), en rangos (Mardia (1975)), en rangos con signos (Fisher y Lee (1982)), para n pares (Jammalamadaka y Sarma (1988) y Jammalamadaka y SenGupta (2001)), para dependencia cluster y el coeficiente basado en la matriz de productos cruzados (Rivest (1982)).

2.2.2. Distribución de von Mises

En los modelos circulares la distribución von Mises (von Mises (1918)) juega un papel similar al modelo Normal en el espacio euclídeo. En los problemas que aquí se plantean, no siempre será necesario suponer una distribución subyacente, sin embargo, en el caso de hacerlo, asumiremos que dicha distribución es de von Mises.

Se dice que una variable θ tiene distribución de von Mises, $\theta \sim M(\phi, \kappa)$, con $\phi \in [0, 2\pi)$ y $\kappa \geq 0$, si la función de densidad viene dada por la siguiente expresión,

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

donde I_0 es la función modificada de Bessel de primera clase y de orden cero. La función de Bessel de primera clase de orden q se define como,

$$I_q(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(qx) e^{\kappa \cos(x)} dx.$$

El parámetro ϕ es la dirección media y κ es el parámetro de concentración de la distribución. Esta distribución es unimodal y simétrica respecto a la dirección media ϕ, como puede verse en la Figura 2.4. Los estimadores máximo

verosímiles (EMV) de los dos parámetros de interés son,

$$\widehat{\phi} = \overline{\theta}, \qquad \widehat{\kappa} = A^{-1}(\overline{R}), \tag{2.2}$$

donde $A^{-1}(\overline{R}) = I_0(\overline{R})/I_1(\overline{R})$.

Las variaciones en la forma de la distribución según los valores de κ se pueden ver en la Figura 2.4. Las relaciones más conocidas de la von Mises con otras distribuciones son:

- Si $\kappa = 0$, nos encontramos ante la distribución uniforme en el círculo.
- Si $\kappa \to \infty$ entonces se tiene que,

$$\kappa^{-1/2}(\theta - \phi) \sim N(0, 1)$$

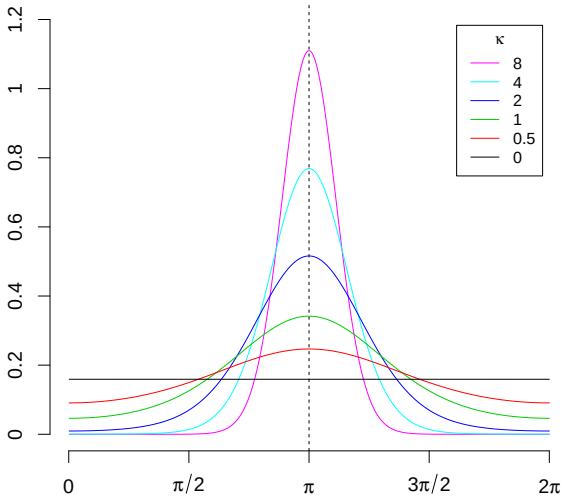


Figura 2.4: Función de densidad de la distribución de von Mises $M(\pi, \kappa)$ según las variaciones del parámetro de concentración κ

2.3. Inferencia con restricciones

La Inferencia con Restricciones (ICR) es la rama de la Estadística que estudia la definición y análisis de procedimientos que incorporan restricciones en los parámetros del modelo. Dichas restricciones se definen a partir de información adicional de la que en muchas situaciones se dispone o que se puede suponer por ser razonable, o que se obtiene indagando en el problema. El objetivo es conseguir métodos que sean más eficaces, más potentes, más robustos, más flexibles y poco costosos computacionalmente. En muchos casos, una ventaja añadida es la obtención de soluciones fácilmente interpretables en la práctica.

En esta sección, vemos brevemente los conceptos básicos de esta área por ser fundamentales en los procedimientos para modelos circulares con restricciones que se desarrollan en este trabajo. En primer lugar, se presenta la representación matemática de las restricciones de orden más comunes en el espacio euclídeo. A continuación el problema de mínimos cuadrados restringido conocido como regresión isotónica y el algoritmo que se usa para su resolución en el caso del orden simple. Finalmente, presentamos algunos métodos para realizar inferencias bajo la suposición del modelo Normal. Las referencias básicas en este ámbito son Barlow et al. (1972), Robertson et al. (1988), Silvapulle y Sen (2005).

La notación es diferente a la que encontramos en el resto de esta memoria ya que en esta sección se expone una revisión de la ICR en el espacio euclídeo. Denotaremos por $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ el vector de los valores medios de muestras de tamaños n_i, $i = 1, \ldots, n$, observadas para n poblaciones cuyas medias son $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$.

2.3.1. Representación de las restricciones de orden

Las restricciones sobre los parámetros que tratamos en este trabajo son de orden. En otras palabras la información adicional nos dice que las n compo-

nentes de $\boldsymbol{\mu}$ verifican una cierta relación de orden, por ejemplo, $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{n-1} \leq \mu_n$. Un vector se dice que es isotónico con respecto a un orden, si verifica dicho orden. Asumir un orden entre los parámetros es equivalente a restringir el espacio paramétrico a un cono convexo y cerrado. Un cono $C \in \mathbb{R}^n$ es un conjunto que cumple la siguiente propiedad,

$$\text{Si } \boldsymbol{x} \in C, \text{ entonces } \lambda \boldsymbol{x} \in C \quad \forall \lambda \geq 0.$$

Exponemos a continuación algunos de los conos de orden más usados en la práctica.

- Orden simple: tendencia monótona creciente.

$$C_{os} = \{\boldsymbol{x} \in \mathbb{R}^n : x_1 \leq \ldots \leq x_n\} \tag{2.3}$$

  Un ejemplo de aparición de este cono se encuentra en los estudios de dosis-respuesta, en donde se asume a menudo una relación creciente, del tipo que se puede observar en la Figura 2.5a.

- Árbol simple.

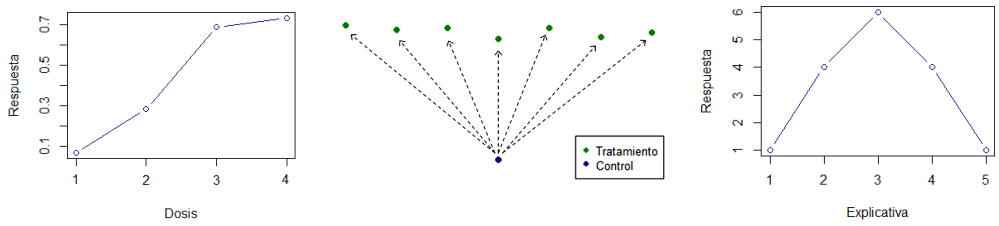
$$C_{st} = \{\boldsymbol{x} \in \mathbb{R}^n : x_0 \leq x_i, i = 1, \ldots, n-1\}$$

  Aparece, por ejemplo, cuando se comparan poblaciones con un control. Puede verse la forma de árbol en la Figura 2.5b.

- Unimodalidad.

$$C_u = \{\boldsymbol{x} \in \mathbb{R}^n : x_1 \leq \ldots \leq x_r \geq x_{r+1} \geq \ldots \geq x_q\}$$

  Se encuentra en los perfiles de la expresión de algunos genes, como puede verse en algunos ejemplos en Lin et al. (2012) y tiene la forma que se observa en la Figura 2.5c.



(a) Orden simple (b) Orden árbol simple (c) Orden unimodal

2.3.2. Regresión isotónica y PAVA

A continuación definimos la regresión isotónica que juega un papel muy importante en la ICR porque es la solución a un problema de optimización que trata de buscar el vector ordenado que más se acerca a las observaciones. En concreto, supongamos que $\boldsymbol{\mu} \in C$, siendo C un cono de orden. Entonces la regresión isotónica de $\boldsymbol{Y}$ con pesos $\boldsymbol{\omega}$, respecto al orden representado por C es,

$$\boldsymbol{Y}^* = \arg\min_{\boldsymbol{Z}\in C} \sum_{i=1}^{n} \omega_i (Z_i - Y_i) = \arg\min_{\boldsymbol{Z}\in C} (\boldsymbol{Y} - \boldsymbol{Z})' W (\boldsymbol{Y} - \boldsymbol{Z}), \tag{2.4}$$

donde W es una matriz definida positiva de la forma $W = diag(\omega_1, \ldots, \omega_n)$ siendo ω_i, $i = 1, \ldots, n$ un peso positivo.

Si consideramos la métrica definida por el siguiente producto escalar $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}' W \boldsymbol{v}$, la regresión isotónica viene dada por la proyección de Y sobre el cono C,

$$\boldsymbol{Y}^* = P_W(\boldsymbol{Y}/C).$$

La regresión isotónica se obtiene mediante promedios de componentes. Puede probarse que existe una partición $\{(l)\}_{l=1}^{m}$ de los subíndices $\{1, \ldots, n\}$ tal que $Y_i^* = Av(G_{(l)})$ $\forall i \in (l)$, donde $G_{(l)} = \{Y_i\}_{i\in(l)}$ y $Av(G_{(l)}) = \frac{\sum_{i\in(l)} \omega_i Y_i}{\sum_{i\in(l)} \omega_i}$. A estos conjuntos, dentro de los cuales los valores del estimador son iguales, se les conoce como conjuntos de nivel. Denominaremos m al número de conjuntos de nivel de $\boldsymbol{Y}^*$.

En el caso particular del orden simple, $\boldsymbol{\mu} \in C_{os}$ (2.3), la solución al problema (2.4) se puede obtener mediante el algoritmo PAVA (*pool adyacent violator algorithm*), (Robertson y Wright (1980)). Este algoritmo se basa en promediar las observaciones próximas que violan las restricciones de orden. Se puede ver paso a paso en el Algoritmo 1.

Algoritmo 1: PAVA: Pool Adjacent Violator Algorithm

entrada: $\boldsymbol{Y} = Y_1, \ldots, Y_n$.

salida : $\boldsymbol{Y}^*$.

if $\boldsymbol{Y}$ *isotónico según el orden simple* **then**

  $\boldsymbol{Y}^* = \boldsymbol{Y}$;

if $\boldsymbol{Y}$ *no isotónico* **then**

  $\boldsymbol{Y}^* = \boldsymbol{Y}$;

  while $\exists\ i / Y_i^* > Y_{i+1}^*$ **do**

    Se reemplazan ambos valores por su promedio:

$$Y_i^* = Y_{i+1}^* = Av(i, i+1) = \frac{\omega_i Y_i^* + \omega_{i+1} Y_{i+1}^*}{\omega_i + \omega_{i+1}}$$

    y los dos pesos por su suma $\omega_i + \omega_{i+1}$;

    Cada vez que haya que promediar se incluirían en dicho promedio todos los valores del bloque con el que ya ha sido promediado cada elemento i e $i+1$ y se actualizará para todo el bloque;

return $\boldsymbol{Y}^*$;

La propiedad del valor medio de Cauchy para la media es imprescindible para que el PAVA funcione adecuadamente. Esta propiedad dicta que la media de dos valores se encuentra acotada por dichos valores. La media aritmética usada en el caso euclídeo cumple esta propiedad. Sin embargo, como veremos en la siguiente sección, la media circular no verifica dicha propiedad y no se puede aplicar el PAVA directamente.

En el caso de otro tipo de restricciones de orden diferentes al orden simple, existen otros algoritmos que no presentamos aquí por no ser de interés en este trabajo, (Dykstra (1981), Lee (1983), Pardalos y Xue (1999)). Existen también extensiones de la regresión isotónica para resolver problemas con restricciones más generales (ver por ejemplo, Robertson y Wright (1975), Sasabuchi et al. (1983) y Bacchetti (1989)).

2.3.3. Inferencias en el modelo Normal

La mayoría de los métodos se han desarrollado para el modelo Normal. Así, en este apartado, $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$, donde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ es el vector de medias y Σ la matriz de covarianzas, $\Sigma = diag(\frac{\sigma_1^2}{n_1}, \ldots, \frac{\sigma_n^2}{n_n})$. Con el fin de simplificar la exposición, presentamos el caso de Σ conocida. De forma general, suponemos $\boldsymbol{\mu} \in C$. Exponemos a continuación la estimación máximo verosímil restringida seguida de los contrastes de hipótesis donde se revisan el test de razón de verosimilitudes y el test condicional.

Estimador máximo verosímil restringido

La función de verosimilitudes viene dada por,

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

Se puede comprobar que el estimador máximo verosímil restringido (EMVR) del parámetro $\boldsymbol{\mu}$ es la regresión isotónica para el orden representado por el cono C, (Robertson et al. (1988)), $\widehat{\boldsymbol{\mu}} = \boldsymbol{Y}^* = P_{\Sigma^{-1}}(\boldsymbol{Y}/C)$.

El EMVR es sesgado y tiene un error cuadrático medio (ECM) menor que el EMV sin restringir (Robertson et al. (1988)),

$$E[(\boldsymbol{Y}^* - \boldsymbol{\mu})'(\boldsymbol{Y}^* - \boldsymbol{\mu})] \leq E[(\boldsymbol{Y} - \boldsymbol{\mu})'(\boldsymbol{Y} - \boldsymbol{\mu})].$$

Esta propiedad no siempre se mantiene al estimar funciones del parámetro. Las primeras referencias sobre este asunto en el campo de la estimación con restricciones son las de Lee (1981a) y Lee (1981b) y Kelly (1989). Esta cuestión ha sido estudiada posteriormente con detalle en trabajos como Rueda et al. (1997a,b)), Fernández (1995), Fernández et al. (1997, 1998, 1999, 2000).

Contrastes de hipótesis con restricciones de orden

Los contrastes de hipótesis con restricciones involucran a dos de las tres hipótesis que definimos a continuación,

$$\begin{aligned} H_h: &\quad \mu_1 = \mu_2 = \cdots = \mu_n \\ H_0: &\quad \boldsymbol{\mu} \in C \\ H_1: &\quad \boldsymbol{\mu} \in \mathbb{R}^n \equiv \boldsymbol{\mu} \notin C \end{aligned} \tag{2.5}$$

donde C es un cono convexo, cerrado, que muy habitualmente es el cono del orden simple pero que puede representar otro tipo de relación de orden de interés. Los dos contrastes de mayor interés son: por un lado contrastar la homogeneidad de medias (H_h) frente a que los parámetros sigan el orden representado por el cono C (H_0) y por otro lado contrastar el cono C (H_0) frente a todo el espacio paramétrico (H_1).
En este apartado nos centramos en este último por ser el simétrico del que se presenta en la ICR para modelos circulares (Sección 2.4.3) y que resultará de interés posteriormente. Describimos brevemente tanto el test de razón de verosimilitudes (TRV) como un test condicional.
Las dos propuestas que presentamos para hacer el contraste de H_0 contra H_1 se basan en el estadístico razón de verosimilitudes,

$$T = ||\boldsymbol{Y} - \boldsymbol{Y}^*||^2_{\Sigma^{-1}} = \sum_{i=1}^{n} \omega_i (Y_i - Y_i^*)^2. \tag{2.6}$$

donde $\boldsymbol{Y}^* = P_{\Sigma^{-1}}(\boldsymbol{Y}/C)$ y $\omega_i = n_i/\sigma_i^2$.

Test razón de verosimilitudes (TRV) para H_0 contra H_1.

Para calcular el nivel del test es necesario identificar primero la configuración más desfavorable del parámetro bajo H_0 a lo que denominamos μ^0. Como valor de μ^0 puede elegirse cualquiera de los elementos de $S^0 = \{\boldsymbol{\mu} \in \mathbb{R}^n / \mu_1 = \cdots = \mu_n\}$. La distribución de T bajo dicha configuración es una mixtura de distribuciones χ^2, conocida como *Chi-Bar-Squared*,

$$\sup_{\mu \in H_0} pr_{\mu}(T \geq t) = pr_{\mu^0}(T \geq t) = \sum_{l=0}^{n} \nu(m, n; \boldsymbol{\omega}) pr(\chi_l^2 \geq t),$$

donde $l = n - m$ y $\nu(m, n; \boldsymbol{\omega})$ son las denominadas probabilidades de nivel o pesos de la distribución. $\nu(m, n; \boldsymbol{\omega})$ es el valor de la probabilidad bajo μ^0 de que el número de conjuntos de nivel de $\boldsymbol{Y}^*$ sea exactamente m. Es decir, que si definimos,

$$S_m = \{\boldsymbol{Y} \in \mathbb{R}^n : \boldsymbol{Y}^* \text{ tiene } m \text{ conjuntos de nivel}\},$$

tenemos que $\nu(m, n; \boldsymbol{\omega}) = pr_{\mu^0}(\boldsymbol{Y} \in S_m)$. El cálculo de $\nu(m, n; \boldsymbol{\omega})$ no es trivial y precisamente una ventaja del test condicional es que se elimina el cálculo de estos valores. Para todos los detalles de lo anterior ver Robertson et al. (1988, p.69) y Silvapulle y Sen (2005, p.78-81).

Cabe comentar que en caso de Σ desconocida y $(\sigma_1 = \cdots = \sigma_n)$, la distribución del estadístico correspondiente bajo la configuración más desfavorable de H_0 es la conocida como *E-Bar-Squared* y se trata de una mixtura de distribuciones beta, ver Robertson et al. (1988, p.63-74).

El test de razón de verosimilitudes (TRV) es consistente y mas potente que el test de razón de verosimilitudes para alternativas no restringidas (Perlman (1969) y Mukerjee et al. (1986)). El comportamiento de estos tests ha sido ampliamente estudiado en la literatura, Rueda (1989), Menéndez y Salvador (1991), Menéndez et al. (1991a,b, 1992), Menéndez

y Salvador (1992).

Test condicional para H_0 contra H_1.

El test condicional se basa también en el estadístico razón de verosimilitudes T definido en (2.6). El test condicional de nivel α rechaza H_0 cuando $T \geq c(l)$ donde $c(l)$ viene dado por,

$$\alpha' = pr(\chi^2_l \geq c(l)) = \frac{\alpha}{1 - pr_{\mu^0}(T = 0)}.$$

Está demostrado que, asintóticamente, la configuración más desfavorable bajo H_0 es μ^0 (Wollan y Dykstra (1986), Menéndez et al. (1991a)). Esto garantiza que el test condicional es un test de nivel α y permite obtener el p-valor mediante una χ^2_l. Cuando $pr_{\mu^0}(T = 0)$ es un valor pequeño, para valores grandes o moderados de n, $c(m)$ se define simplemente como el percentil $1 - \alpha$ mediante una χ^2_l.

En comparación con el test razón de verosimilitudes donde el cálculo de los pesos $\nu(m, n; \boldsymbol{\omega})$ es tedioso, el test condicional es mucho más simple y también se beneficia de un incremento en la potencia para algunas alternativas interesantes.

El uso de tests condicionales no es nuevo en Inferencia con Restricciones (ver por ejemplo, Bartholomew (1961), Iverson y Harp (1987), Hu y Wright (1994)). En los últimos años, el test condicional ha sido propuesto también para modelos circulares en Fernández et al. (2012). La idea de usar el test condicional como un procedimiento con un estadístico con *grados de libertad que dependen de los datos* ha sido tomada de la Inferencia con Restricciones y usada en un contexto más general en Susko (2013). Recientemente, en el trabajo de Rueda et al. (2014b) se define un test condicional para otro problema con restricciones y se ilustra sus múltiples ventajas.

2.4. Inferencia con restricciones para datos circulares

La ICR se ha desarrollado casi exclusivamente en el espacio euclídeo hasta hace unos años, cuando surgen los trabajos de Rueda et al. (2009) y Fernández et al. (2012), donde por primera vez se considera la incorporación de restricciones en el análisis de datos circulares. Estos trabajos surgen también para resolver problemas en el ámbito del análisis de expresiones de genes asociados al ciclo celular.

En esta sección la notación será la misma que en el resto de la memoria, $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$ es el vector de observaciones circulares y $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_n)'$ el vector de direcciones medias poblacionales. El interés es realizar inferencias sobre $\boldsymbol{\Phi}$ bajo restricciones de orden, en este caso restricciones de orden circular. El desarrollo de la metodología de la ICR para datos circulares comienza con la definición de los nuevos órdenes de interés en el espacio circular.

2.4.1. Restricciones de orden en el círculo

Como ya se ha comentado en la definición de las medidas de asociación circular (ver Sección 2.2.1) el número mínimo de elementos circulares que mantiene una relación de asociación es tres. Así, el orden circular se define para un mínimo de tres elementos, es decir, $\phi_1 \leq \phi_2 \leq \phi_3 \leq \phi_1$.

Definición 2.8. *El conjunto que define un orden circular* $\boldsymbol{O}$ *en dimensión* n *viene dado por,*

$$C_O = \{\boldsymbol{\Phi} \in [0, 2\pi)^n : \boldsymbol{\Phi} \circlearrowleft \boldsymbol{O}\}.$$

En otras palabras, $\boldsymbol{\Phi} \in C_O \Leftrightarrow$ las componentes de $\boldsymbol{\Phi}$ verifican el orden circular $\boldsymbol{O}$, $\phi_{o_1} \leq \phi_{o_2} \leq \cdots \leq \phi_{o_n} \leq \phi_{o_1}$. El orden se denota por $\boldsymbol{O} = (o_1, o_2, \ldots, o_n)$. En algunas aplicaciones puede ser de interés un orden circular parcial, es decir,

un orden entre grupos de elementos. De forma general un orden circular parcial $\boldsymbol{O}_P$ se puede escribir como sigue,

$$\boldsymbol{O}_P = \left\{ \begin{array}{c} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{l_1} \end{array} \right\} \leq \cdots \leq \left\{ \begin{array}{c} \phi_{l_1+\ldots+l_{L-1}+1} \\ \phi_{l_1+\ldots+l_{L-1}+2} \\ \vdots \\ \phi_{l_1+\ldots+l_L} \end{array} \right\} \leq \left\{ \begin{array}{c} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{l_1} \end{array} \right\}, \tag{2.7}$$

donde L es el número de grupos, teniendo l_c elementos en el grupo c de tal forma que $n = \sum_{c=1}^{L} l_c$.

Recordar que como se mencionó en la Sección 2.1 características importantes de estos órdenes son su invariancia frente a la rotación y su independencia de la dirección inicial.

2.4.2. Regresión isotónica circular

El estimador de regresión isotónica (CIRE, del inglés Circular Isotonic Regression Estimator) es la generalización al espacio circular del concepto de regresión isotónica, se presenta en Rueda et al. (2009) y se define como el vector más cercano a $\boldsymbol{\Theta}$ que verifique el orden como vemos a continuación.

Definición 2.9. *El estimador de regresión isotónica circular (CIRE) de* $\boldsymbol{\Phi}$ *con respecto a* C_O *se define como,*

$$\widetilde{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\alpha}} SCE(\boldsymbol{\alpha}, \boldsymbol{\Theta}), \tag{2.8}$$

La existencia y unicidad (casi seguro) del CIRE, así como otras propiedades, se demuestran en Rueda et al. (2009). La obtención del CIRE no es posible mediante ningún algoritmo para obtención de estimadores restringidos en contextos euclídeos, ni mediante una adaptación directa de los mismos. Como ya se ha comentado, el PAVA (Algoritmo 1) no puede ser usado con la media circular ya que esta no cumple la propiedad del valor medio de Cauchy. La solución $\widetilde{\boldsymbol{\Phi}}$ se obtiene mediante un algoritmo específico presentado en Rueda et al. (2009),

que está inspirado en el PAVA y que tiene en cuenta las peculiaridades del espacio circular. La implementación de dicho algoritmo se encuentra disponible tanto en código SAS como en el lenguaje R. Este último está incluido como parte del paquete **isocir** cuyo uso se explica con detalle en el Capítulo 6.

2.4.3. Inferencias en el modelo de von Mises

De la misma forma que el modelo Normal es el habitual en el caso euclídeo, el modelo de von Mises lo es cuando los datos son circulares. A lo largo de esta memoria cuando sea necesario hacer hipótesis distribucionales supondremos distribuciones de von Mises, $\theta_i \sim M(\phi_i, \kappa)\ \forall i = 1, \ldots, n$ independientes (ver distribución de von Mises, Sección 2.2.2). Inicialmente se supone que κ es conocido.

En Rueda et al. (2009) se demuestra que bajo la suposición de distribución de von Mises, el CIRE es el estimador máximo verosímil restringido (EMVR) de $\mathbf{\Phi}$, cuando $\mathbf{\Phi} \in C_O$.

En este apartado presentamos el contraste de un orden circular prefijado resuelto en Fernández et al. (2012) mediante un test condicional. Recordamos que la principal ventaja de los test condicionales es la eficiencia computacional. Este contraste se puede formular de la siguiente forma,

$$\begin{aligned} H_0 :&\ \ \mathbf{\Phi} \in C_O \\ H_1 :&\ \ \mathbf{\Phi} \notin C_O. \end{aligned}$$

El estadístico razón de verosimilitudes se formula como sigue,

$$T = 2\kappa SCE(\mathbf{\Theta}, \widetilde{\mathbf{\Phi}}^{(\boldsymbol{C_O})}), \tag{2.9}$$

donde $\widetilde{\mathbf{\Phi}}^{(\boldsymbol{C_O})}$ es el CIRE con respecto al orden C_O. Sea $l = n - m$ donde m el número de conjuntos de nivel del CIRE bajo H_0.

Este test condicional se resuelve siguiendo la misma idea que en el caso euclídeo presentado en la Sección 2.3. El test condicional de nivel α viene dado por,

$$\text{Rechazar } H_0 \text{ si } T \geq c(l).$$

donde el valor de $c(l)$ se obtiene como sigue,

$$pr(\chi^2_l \geq c(l)) = \frac{\alpha}{1 - pr_{\phi^0}(C_O)},$$

donde ϕ^0, verificando $\phi^0_1 = \phi^0_2 = \cdots = \phi^0_n$, es la configuración más desfavorable bajo la hipótesis nula. Es fácil demostrar que, $pr_{\phi^0}(C_O) = \frac{1}{(n-1)!}$. La distribución asintótica del estadístico T (2.9 bajo H_0 para valores altos de κ es χ^2_l. Esto garantiza que el test condicional es de nivel α y que se puede obtener el p-valor mediante una χ^2_l.

En el caso de suponer κ desconocido, la expresión del estadístico razón de verosimilitudes es la siguiente:

$$T = \frac{2\hat{\kappa} SCE(\boldsymbol{\Theta}, \widetilde{\boldsymbol{\Phi}}^{(\boldsymbol{C})})}{n},$$

donde $\hat{\kappa}$ es el estimador de κ (2.2). El valor de $c(l)$ se obtiene para este caso de manera que,

$$pr(F_{l,n-1} \geq c(l)) = \frac{\alpha}{1 - pr_{\phi^0}(C_O)},$$

donde $F_{l,n-1}$ es una variable aleatoria F centrada y con $(l, n-1)$ grados de libertad.

Los estudios de simulación llevados a cabo en Fernández et al. (2012) apuntan a una potencia razonable para estos test. Para más detalles, consultar el material suplementario de Fernández et al. (2012) donde entre otros aspectos se incluyen estudios de la potencia de estos tests. La extensión al caso de que la hipótesis nula sea un orden circular parcial es inmediata. Todos los métodos aquí expuestos están implementados en el paquete de R llamado **isocir**, explicado en el Capítulo 6 de esta memoria.

2.5. Métodos de agregación de rankings en la línea

En esta sección se considera el problema de encontrar el ranking agregado entre un conjunto de n elementos a partir de la información de diferentes fuentes. El problema de agregación de rankings es un problema que ha sido ampliamente tratado en la literatura y en muy diferentes ámbitos. De forma general se puede describir como el problema de ordenar un conjunto de n items a partir de la información de p observadores. Las palabras item y observador se traducen en palabras más específicas en las diferentes aplicaciones. En particular, en nuestro caso, trataremos con genes y con experimentos respectivamente. Las primeras aplicaciones de la agregación de rankings surgieron en el ámbito de las ciencias sociales para resolver problemas de elección social o de preferencias sociales (Bartholdi et al. (1989), Caplin y Nalebuff (1991), Chevaleyre et al. (2007), Hassanzadeh (2013)). Un problema típico en este contexto es el de ordenar candidatos en base a las valoraciones de un conjunto de votantes o el de hacer un ranking de instituciones en base a los rankings que generan un conjunto de índices. Otros ámbitos de aplicación son las competiciones deportivas (Sizemore (2013)) donde se busca el ranking entre los participantes, la biología (DeConde et al. (2006), Pihur et al. (2008), Simko y Pechenick (2010), Kadota y Shimizu (2011)), evaluación de la calidad (Xu et al. (2012)), etc.
En la última década el tema de agregación de rankings ha sufrido un gran desarrollo impulsado fundamentalmente por las aplicaciones en los entornos Web, relacionadas con las meta-búsquedas, la reducción de spam, las técnicas de asociación de palabras, etc. (Dwork et al. (2001a,b), Shishkin et al. (2013), Chen et al. (2013)). De hecho, el algoritmo PageRank desarrollado por Langville y Meyer (2006) se ha convertido en una de las soluciones más populares de este tipo de problemas por ser el que usa el buscador Google.

El problema se aborda de manera general como la búsqueda del ranking mas cercano al conjunto de observaciones en base a una función objetivo o un

criterio. En la literatura existe un abanico muy amplio de técnicas para la resolución de este problema. Esta amplitud de técnicas puede clasificarse acorde a diferentes aspectos como son: a) la función objetivo a optimizar o criterio; b) el tipo de información usada (ordinal o cardinal); c) la representación matemática de la información disponible, mediante vectores (cuando se usan las observaciones individuales) o matrices (cuando se usan las relaciones entre pares de elementos que bien pueden ser observadas directamente u obtenidas de las observaciones individuales); d) si existen suposiciones subyacentes y de qué tipo; e) si se hace uso de información adicional (métodos supervisados). Como es de suponer, resulta difícil clasificar todos los métodos existentes (ver Schalekamp y Zuylen (2009) y Lin (2010)). Por ello, en la Sección 2.5.2 realizamos una revisión de los mismos mediante una selección representativa clasificada según el aspecto c).

Sea $V = \{1, \ldots, n\}$ el conjunto de elementos a ser ordenados. Sea $\boldsymbol{T}_j = (\tau_{1j}, \ldots, \tau_{nj})'$ un vector de posiciones que define un ranking entre n elementos donde τ_{ij} es la posición del elemento i en el ranking dado por el experimento j. Denominamos $\mathcal{T}$ al conjunto de todos los posibles vectores de posiciones de n elementos. En ocasiones, se observan directamente los valores cardinales, entonces denominaremos $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$ al vector de observaciones para los n elementos del experimento j.
En caso de abordar el problema mediante la minimización de una función objetivo se suele hacer uso de una distancia entre un conjunto de datos y un ranking para lo cual existen diversas propuestas en la literatura, (vemos en la Sección 2.5.1 las dos más habituales), y se formularía de forma general como sigue,

$$\widehat{\boldsymbol{T}} = \arg\min_{\mathrm{T}\in\mathcal{T}} \sum_{j=1}^{p} d(\boldsymbol{x}_j, \mathrm{T}),$$

donde $\widehat{\boldsymbol{T}}$ es el ranking agregado y $d(\boldsymbol{x}_j, \mathrm{T})$ es una distancia entre un ranking T y las observaciones.
En otros casos se busca que la solución verifique al menos uno de los dos

principios o reglas fundamentales de la agregación de rankings en la línea que exponemos a continuación.

Principio de Condorcet (Condorcet (1785)). La idea en la que se basa es que si un elemento aparece por delante del resto en todos los rankings observados, deberá ser el primero también en el ranking agregado. La extensión de este criterio la realizó Truchon (1997) y ha dado lugar a diversas reglas para la agregación de rankings que pueden verse recopiladas en el trabajo de Felsenthal y Tideman (2014). El principio de Condorcet extendido dice que para cada par de elementos (i, h) si i está delante de h en la mayoría de los rankings, entonces en el ranking agregado también.

Regla de Borda (Borda (1781)). La idea en este caso es ordenar los elementos según las puntuaciones atribuidas a cada posición de cada ordenación observada de la siguiente forma: 1 punto para el último clasificado, 2 puntos para el penúltimo, 3 para el antepenúltimo etc. Este sistema conlleva que no siempre el elemento colocado más veces en primer lugar es el vencedor en el ranking agregado pudiendo incumplir por tanto el principio de Condorcet.

Tanto si se usa una función objetivo o un criterio como los anteriores, el concepto de distancia entre rankings es imprescindible para abordar este problema, porque aparece directamente en la función objetivo o porque se utiliza para evaluar el método. Por tanto, en la Sección 2.5.1 presentamos las distancias más usadas. En la Sección 2.5.2 presentamos una selección de métodos que consideramos de interés. En la Sección 2.5.3, presentamos con mayor detalle la metodología basada en la teoría de Hodge por el interés que tiene para el desarrollo de una de las técnicas presentadas en el capítulo siguiente.

2.5.1. Distancias entre rankings en la línea

Existe una amplia variedad de definiciones de distancias entre un conjunto de datos y un ranking, además se pueden encontrar múltiples estudios comparativos de este tipo de distancias (ver Kumar y Vassilvitskii (2010) y Cook (2006)). En este apartado presentamos las dos más usadas.
La distancia Spearman footrule (Definición 2.10) mide las diferencias entre rankings a través de las diferencias en las posiciones de cada elemento. La distancia Tau de Kendall (Definición 2.11) mide el número de pares discordantes, es decir, cuantas veces el orden entre pares de elementos es diferente de un ranking a otro.

Definición 2.10. *Sean $\boldsymbol{T}_1$ y $\boldsymbol{T}_2$ dos vectores de posiciones, la distancia Spearman footrule $F(\boldsymbol{T}_1, \boldsymbol{T}_2)$ se define como,*

$$F(\boldsymbol{T}_1, \boldsymbol{T}_2) = \sum_{i=1}^{n} |\tau_{i1} - \tau_{i2}|,$$

donde τ_{ij} contiene la posición del elemento i en ranking observado en el experimento j, $j = 1, 2$.

Definición 2.11. *Sean $\boldsymbol{T}_1$ y $\boldsymbol{T}_2$ dos vectores de posiciones, la distancia Tau de Kendall $K(\boldsymbol{T}_1, \boldsymbol{T}_2)$ se define como,*

$$K(\boldsymbol{T}_1, \boldsymbol{T}_2) = \sum_{i<k} I_{i,k},$$

donde,

$$I_{i,k} = \begin{cases} 1 & si\ (\tau_{i1} < \tau_{k1} \wedge \tau_{i2} > \tau_{k2}) \vee (\tau_{i1} > \tau_{k1} \wedge \tau_{i2} < \tau_{k2}) \\ 0 & en\ otro\ caso \end{cases}$$

Estas distancias se definen también cuando se tiene información cardinal, ver Kumar y Vassilvitskii (2010) para más detalles.

2.5.2. Aproximaciones al problema de agregación de rankings en la línea

En este apartado presentamos una selección de los métodos desarrollados para el problema de agregación de rankings agrupándolos dependiendo de si usan la información de los valores observados individualmente para cada elemento o la información de las medidas de la intensidad de relación entre pares de elementos.

Métodos que usan la información individual

Estos métodos se caracterizan porque la agregación de las observaciones de distintos experimentos se realiza elemento a elemento independientemente del resto.

- **Método de Borda**, Borda (1781) es un método que usa la información ordinal. Este método y todos aquellos inspirados en él (Ho et al. (1994), Aslam y Montague (2001), Lumini y Nanni (2006)), se basan en la regla de Borda. En el método básico de Borda, cada elemento i tiene un valor $B(i)$ definido según las posiciones de i en cada uno de los p rankings,

$$B(i) = \sum_{j=1}^{p} \tau_{ij}.$$

  La ordenación de dichos $B(1), \ldots, B(i), \ldots, B(n)$ es el ranking agregado. Existen numerosos métodos inspirados en este usando otras medidas como la mediana, la media geométrica o la p-norma. Se pueden ver comparaciones de estas últimas para ejemplos típicos en Lin (2010). Este se trata de uno de los métodos más usados en la agregación de rankings (Dym et al. (2002), Klamler (2004), Mc Donald y Smeaton (2005), Baltrunas et al. (2010), Mekonnen (2014)).

- **Método de agregación de Footrule**, Diaconis y Graham (1977), este método busca minimizar la distancia de *Sperman footrule* de la Defini-

ción 2.10, haciendo uso de la información ordinal mediante las posiciones en los órdenes observados.

- **Algoritmos de búsqueda local**, también denominados algoritmos *single vertex moves*, en cada iteración mueven un elemento a una posición adyacente si eso mejora el valor de la función objetivo. Normalmente, se comienza con una ordenación aleatoria, pero también puede tomarse un ranking que proceda de otro método. Un ejemplo muy conocido es el algoritmo *Local Kemenization* que se presenta en el trabajo de Dwork et al. (2001a) y cuya adaptación al círculo realizamos en este trabajo (ver Sección 3.6).

- **Métodos basados en algoritmos de comparación de ordenaciones.** Se basan principalmente en cumplir el principio de Condorcet extendido. Un ejemplo de estos métodos es el llamado **InsertionSort** (Schalekamp y Zuylen (2009)) que resulta interesante porque combina la minimización de una distancia con el cumplimiento del principio de Condorcet.

- **Método de Copeland.** Este método presentado por Copeland (1951) y todos los que le han seguido se denominan híbridos debido a que tratan de conciliar la regla de Borda y el principio de Condorcet.

- **Métodos que usan modelos de probabilidad.** Existen varios métodos que se basan en estimar la probabilidad de que cada elemento aparezca en una cierta posición y de esa manera ofrecer un ranking agregado. Algunos ejemplos son el algoritmo **Cross-Entropy Montecarlo** (Rubinstein y Kroese (2004)) y el modelo basado en permutaciones presentado por Mallows (1957).

Métodos que usan la relación entre pares

El uso de las relaciones entre pares de elementos puede surgir al tener directamente medidas del grado de preferencia o de la intensidad de relación entre pares de elementos (Rajkumar y Agarwal (2014)) o al medir la intensidad de

relación entre pares a partir de las observaciones individuales (ver por ejemplo, Gleich y Lim (2011)). En cualquier caso la relación entre pares se representa matemáticamente mediante matrices.
Se pueden encontrar en la literatura numerosos métodos de agregación de rankings que hacen uso de este tipo de información. Se presenta a continuación una selección de los más populares.

- **Métodos basados en cadenas de Markov.**
  La idea principal es el uso de cadenas de Markov de forma que los estados de la cadena son los elementos a ser ordenados y la probabilidad de transición entre un elemento i y otro h, denotada por p_{ih}, se calcula en base a las posiciones de los elementos en los rankings observados. En Dwork et al. (2001a) se presentan cuatro criterios para definir las probabilidades de transición denominados: MC1, MC2, MC3, MC4.
  Si denominamos M a la matriz de transición compuesta por las probabilidades p_{ih} y denotamos por $\boldsymbol{y}$ a la distribución estacionaria de la cadena de Markov, es sabido que $\boldsymbol{y}$ es el autovector principal por la izquierda de M, es decir que, $\boldsymbol{y}M = \lambda\boldsymbol{y}$. El ranking agregado viene dado por la ordenación de los estados en la distribución estacionaria $\boldsymbol{y}$.
  Una ventaja importante es la rapidez de computación para encontrar la solución. Un ejemplo de los muchos métodos y variantes que hacen uso de las cadenas de Markov es el algoritmo PageRank, Langville y Meyer (2006), que usa el buscador Google. La extensión de estos métodos al espacio circular se ha realizado en este trabajo y se presenta en el siguiente capítulo.

- **Métodos basados en modelos de probabilidad por pares.** Estos modelos hacen uso de las observaciones para calcular la probabilidad de que un par de elementos mantenga un cierto orden. Los tres modelos más conocidos y que han servido de base de otros muchos métodos son los siguientes.

Modelo Thrustone. Este modelo presentado por Thurstone (1927) y aplicado posteriormente a diferentes problemas (Daniels (1950) y Mosteller (1951)), asume para los valores del ranking agregado una distribución subyacente Normal multivariante de manera que cada par de valores seguirá una distribución Normal bivariante. Se calcula la probabilidad de que un cierto elemento esté colocado en el ranking por encima o por debajo de otro elemento. El ranking agregado viene dado por la ordenación del vector de medias estimado.

Bradley-Terry. Este modelo presentado por Bradley y Terry (1952) funciona de forma similar al anterior con el añadido de que asume independencia en las probabilidades de cada par del resto de elementos. Esto es un poco irreal y no obtiene resultados muy buenos, entre otras cosas produce demasiados empates.

Modelo de preferencia multinomial. Este modelo introducido por Volkovs y Zemel (2012) tiene como novedad el uso del grado de confianza de la ordenación de cada par.

- **Método *HodgeRank*.** Este método presentado por Jiang et al. (2011) hace uso de la teoría de Hodge para reducir el problema de agregación de rankings a un problema de mínimos cuadrados que puede ser resuelto de manera inmediata. En el siguiente apartado se presenta en más detalle esta técnica por la relevancia que tiene en esta tesis.

2.5.3. Teoría de Hodge aplicada a la agregación de rankings

Esta metodología para agregación de rankings en la línea ha sido presentada por Jiang et al. (2011) y se denomina *HodgeRank*. La teoría de Hodge es una técnica potente que se encuentra entre los campos del análisis matemático, el álgebra y la topología y que ha sido introducida recientemente en el análisis de datos y en concreto en la agregación de rankings. Los métodos basados en

la teoría de Hodge se caracterizan porque a partir de la información de la intensidad de relación entre pares de elementos, se obtiene el vector que genera el ranking agregado, mediante las relaciones entre los espacios de matrices y vectores que establece la teoría de Hodge.
Como ventajas más importantes de esta técnica señalamos la eficiencia computacional o la flexibilidad a la hora de medir y agregar la información. Además, la teoría de Hodge permite definir índices de inconsistencia para valorar la bondad del ranking obtenido. Por otro lado, una de las desventajas que podemos señalar es que se requiere el manejo de terminología de áreas alejadas de la estadística. En el Apéndice A se exponen los conceptos de la teoría de Hodge necesarios, tanto para introducir la metodología en la línea, como para desarrollar la nueva metodología en el círculo que se presenta en este trabajo (Sección 3.5).

Sea $Y^j \in \mathbb{R}^{n\times n}$ la matriz que recoge la intensidad de relación entre pares en el experimento j, $j = 1, \ldots, n$. Se proponen en la literatura diferentes opciones para definir Y^j_{ih} pero generalmente se define Y^j_{ih} a partir de las observaciones como $Y^j_{ih} = x_{hj} - x_{ij}$, $i, h = 1, \ldots, n$. Denominamos $\overline{Y} \in \mathbb{R}^{n\times n}$ a la matriz que recoge la información agregada, para su construcción existe una variedad de opciones siendo la más usual la media aritmética,

$$\overline{Y}_{ih} = \frac{\sum_j \omega^j_{ih} Y^j_{ih}}{\sum_j \omega^j_{ih}}, \qquad i, h = 1, \ldots, n,$$

donde ω^j_{ih} es el peso para el par (i, h) en el experimento j, normalmente con valor 0 en caso de ser *missing*.

Se considera el producto interno $\langle X, Y\rangle_{\boldsymbol{\omega}} = \sum_{ih} \omega_{ih} X_{ih} Y_{ih}$, donde $\omega_{ih} = \sum_j \omega^j_{ih}$.

El subespacio que contiene las matrices que representan un ranking se define como, $\mathcal{M}_G = \{X \in \mathbb{R}^{n\times n} : X_{ih} = s_h - s_i, \boldsymbol{s} : V \to \mathbb{R}\}$, donde el ranking que representa $X \in \mathcal{M}_G$ se tiene según la siguiente regla:
$i_1 \leq, \ldots, \leq i_n \Leftrightarrow s_{i_1} \leq, \ldots, \leq s_{i_n}$.

De la definición de $\mathcal{M}_G$ el problema de optimización para resolver la agregación de rankings se formula como,

$$\widehat{Y} = \arg \min_{X \in \mathcal{M}_G} ||X - \overline{Y}||^2_{\boldsymbol{\omega}}.$$

Este problema se resuelve en el Teorema 3 de Jiang et al. (2011) y la solución viene dada por,

$$\widehat{Y}_{ih} = s_h - s_i \text{ tal que } s_i = -\frac{1}{n}\sum_h \overline{Y}_{ih} \quad \forall i, h = 1, \ldots, n.$$

$\widehat{Y}$ así definida verifica $\widehat{Y} \in \mathcal{M}_G$ y el ranking agregado lo determina el vector $\boldsymbol{s} = (s_1, \ldots, s_n)$.

Este procedimiento conocido como *HodgeRank* funciona especialmente mejor que otros métodos cuando se tienen datos incompletos y no balanceados. Si se elije convenientemente la definición de Y^j_{ih}, *HodgeRank* es equivalente al método de Borda. Por otro lado, este método se adapta a las características de cada aplicación por su flexibilidad en la definición de Y^j_{ih} y como consecuencia el resultado es más coherente, en muchos casos, tanto con el objetivo concreto perseguido como para la aplicación. Para más detalles, ver Jiang et al. (2011). Esta metodología ha sido usada en diversas aplicaciones como los problemas del entorno web (Dalal et al. (2012)), en la búsqueda de rankings en las puntuaciones de diferentes jueces en el deporte (Sizemore (2013)) o en otros campos (Xu et al. (2012)).

Capítulo 3

Agregación de Órdenes Circulares

An approximate answer to the right question
is worth a great deal more than
a precise answer to the wrong question.
John Tukey

En este capítulo trataremos el problema de la búsqueda del orden circular entre un conjunto de n elementos en base a la información de p experimentos. Esta cuestión es de interés en diversas situaciones, una de ellas es la que resolvemos en esta memoria en relación a un problema biológico. En situaciones como esta, las observaciones provienen de experimentos heterogéneos que hacen imposible la agregación directa de la información. Sin embargo, sí es posible determinar el orden circular agregado que es biológicamente interpretable y será de utilidad en el contraste de hipótesis diseñado en el Capítulo 4.

Se tiene el siguiente conjunto de datos $\boldsymbol{\Theta} = (\boldsymbol{\Theta_1}, \ldots, \boldsymbol{\Theta_j}, \ldots, \boldsymbol{\Theta_p})'$, $j = 1, \ldots, p$ donde $\boldsymbol{\Theta_j} = (\theta_{1j}, \ldots, \theta_{nj})'$, es el vector de observaciones angulares del experimento j para el conjunto de elementos $V = \{1, \ldots, n\}$.
El problema de agregación de órdenes circulares se enfoca en este trabajo como un problema de optimización que se formula matemáticamente de la siguiente forma,

$$\widetilde{\boldsymbol{O}} = arg \min_{\mathbf{O} \in \mathcal{O}} d(\boldsymbol{\Theta}, \mathbf{O}), \tag{3.1}$$

donde la función objetivo $d(\boldsymbol{\Theta}, \mathbf{O})$ es una medida de la distancia entre el conjunto de datos y un orden circular y que definimos en la Sección 3.1. En la Sección 3.2 se hace una presentación general del problema y de las diferentes técnicas de resolución diseñadas en este trabajo. En la Secciones 3.3, 3.4 y 3.5 se detallan diferentes técnicas de agregación de órdenes circulares. En la Sección 3.6 se describe un algoritmo que se usa en una segunda etapa y que mejora la solución obtenida con cualquiera de las técnicas propuestas y en las Secciones 3.7 y 3.8 comparamos los métodos mediante ejemplos y simulaciones. Finalmente, en la Sección 3.9 se resumen las ventajas e incovenientes de cada técnica propuesta.

3.1. Distancias entre un orden circular y un conjunto de datos

Hasta lo que nosotros conocemos no hay en la literatura definiciones de distancias entre un orden circular y un conjunto de datos aunque sí que hay distintas definiciones de distancias entre un ranking y un conjunto de datos en el espacio euclídeo tal y como vimos en la Sección 2.5.1.

En este trabajo proponemos una definición de distancia entre un orden circular y un conjunto de datos, $d_1(\boldsymbol{\Theta}, \mathbf{O})$ (Definición 3.1). Esta medida varía entre 0 y 2, donde 0 implica concordancia total entre el conjunto de datos y el orden circular mientra que 2 significa discordancia total. $d_1(\boldsymbol{\Theta}, \mathbf{O})$ es una media ponderada de la suma de errores circulares. Se trata de una medida muy razonable si se dispone de información cardinal (*scores*). Sin embargo, en el caso de que la información de partida sea ordinal será mas adecuado el uso de la distancia $d_2(\boldsymbol{\Theta}, \mathbf{O})$ (Definición 3.2). Esta medida toma el valor 0 en caso de no asociación, es igual a 1 si existe concordancia y vale -1 en caso de discordancia. Las distancia análoga a $d_2(\boldsymbol{\Theta}, \mathbf{O})$ en el espacio euclídeo es una de las más usadas en la agregación de rankings (ver Definición 2.11).

Recordamos que la notación principal usada a lo largo de esta memoria en

relación a los órdenes circulares está en la Sección 2.1.

Sea ω_j, $j = 1, \ldots, p$, los pesos asociados a cada experimento con $\sum_{j=1}^{n} \omega_j = 1$.

Definición 3.1. *La Media de la Suma de Errores Circulares (MSCE, del inglés: Mean Sum of Circular Errors) entre un conjunto de datos* $\boldsymbol{\Theta}$ *que proviene de* p *experimentos y un orden circular* $\boldsymbol{O}$ *se define como:*

$$d_1(\boldsymbol{\Theta}, \boldsymbol{O}) = MSCE(\boldsymbol{\Theta}, \boldsymbol{O}) = \sum_{j=1}^{p} \omega_j SCE(\boldsymbol{\Theta_j}, \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{j}}^{(\boldsymbol{O})}), \tag{3.2}$$

donde SCE es la suma de errores circulares (Definición 2.5).

Definición 3.2. *La Tau Circular de Kendall media entre un conjunto de datos* $\boldsymbol{\Theta}$ *y un orden circular dado* $\boldsymbol{O}$ *se define como,*

$$d_2(\boldsymbol{\Theta}, \boldsymbol{O}) = \sum_{j=1}^{p} \omega_j d_2(\boldsymbol{\Theta}_j, \boldsymbol{O}) = \sum_{j=1}^{p} \omega_j d_2(\boldsymbol{O}_j, \boldsymbol{O}) = \sum_{j=1}^{p} \omega_j \hat{\Delta}(\boldsymbol{T}_j, \boldsymbol{T}) \tag{3.3}$$

donde $\widehat{\Delta}(\boldsymbol{T}_j, \boldsymbol{T})$ *es la tau circular de Kendall (Definición 2.7).*

El Ejemplo 3.1 abajo ilustra la diferencia entre usar el $SCE(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ o la tau circular de Kendall $\widehat{\Delta}(\boldsymbol{T_1}, \boldsymbol{T_2})$ (en las que se basan $d_1(\boldsymbol{\Theta}, \mathbf{O})$ y $d_2(\boldsymbol{\Theta}, \mathbf{O})$ respectivamente) para medir distancias entre experimentos.

Ejemplo 3.1. Consideramos dos conjuntos de datos que representan dos situaciones reales ambos con $n = 5$ y $p = 2$.

Caso 1: Datos de dos especies distintas con órdenes claramente diferenciados, (ver Figura 3.1).

(a) Experimento 1 (b) Experimento 2

Figura 3.1: Representación de los datos del Caso 1 (Ejemplo 3.1)

Caso 2: Datos de la misma especie donde los elementos están claramente divididos en tres grupos. Aunque el orden dentro de cada grupo no está tan claramente determinado debido a la variabilidad de los experimentos, (ver Figura 3.2).

(a) Experimento 1 (b) Experimento 2

Figura 3.2: Representación de los datos del Caso 2 (Ejemplo 3.1)

Para un biólogo estos experimentos están próximos y representan situaciones similares porque las diferencias ocurren dentro de los grupos y se deben al azar. Mientras que los experimentos del caso 1 recogen dos situaciones biológicamente muy diferentes.

Tabla 3.1: Distancias entre experimentos en cada caso del Ejemplo 3.1

	Caso 1	Caso 2
$SCE(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$	2.58	0.17
$\widehat{\Delta}(\boldsymbol{T}_1, \boldsymbol{T}_2)$	-0.2	-0.2

La distancia entre experimentos medida con el $SCE(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ es mayor en el caso 1 que en el caso 2 mientras que medida usando la tau circular de Kendall $\widehat{\Delta}(\boldsymbol{T_1}, \boldsymbol{T_2})$ es igual en ambos casos. Como ya se ha comentado, consideramos oportuno, para el problema que nos atañe, hacer uso de una distancia que diferencie entre casos del tipo de los aquí expuestos y por tanto se ha elegido $d_1(\boldsymbol{\Theta}, \mathbf{O})$ como distancia entre un conjunto de datos y un orden circular.

3.2. Planteamiento del problema y propuesta de resolución

Utilizando $d_1(\boldsymbol{\Theta}, \boldsymbol{O})$ como función objetivo el problema de optimización (3.1) queda como sigue,

$$\widetilde{\boldsymbol{O}} = \arg\min_{\mathbf{O}\in\mathcal{O}} d_1(\boldsymbol{\Theta}, \mathbf{O}) = \arg\min_{\mathbf{O}\in\mathcal{O}} \sum_{j=1}^{p} \omega_j SCE(\boldsymbol{\Theta}_j, \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{j}}^{(\mathbf{O})}). \tag{3.4}$$

Este problema es del tipo NP-hard (Karp (1972)), lo cual significa que no se puede asegurar que se encuentre el óptimo en tiempo polinómico.
En este trabajo se han diseñado una serie de métodos que ofrecen una buena aproximación al problema y tienen en común una estructura general en dos pasos. En el primer paso se obtiene un orden global inicial $\widehat{\boldsymbol{O}}^0$, este orden se aproximará al optimo en el segundo paso mediante intercambios de indices vecinos. El paso 1 es mucho más determinante para el resultado final que el paso 2, sobretodo para valores moderados o grandes de n. Para el primer paso proponemos diferentes técnicas y dentro de cada técnica varios métodos. El segundo paso es el mismo para todos los métodos del paso 1 y consiste en la ejecución de un algoritmo de búsqueda local denominado *Circular Local Minimization* (CLM) cuyo objetivo es realizar mejoras locales en la solución inicial $\widehat{\boldsymbol{O}}^0$ en términos de la función objetivo (3.4). La idea general de este algoritmo es realizar permutaciones de pares de elementos consecutivos de manera que se cumpla $d_1(\widehat{\boldsymbol{O}}, \boldsymbol{\Theta}) < d_1(\widehat{\boldsymbol{O}}^0, \boldsymbol{\Theta})$. Los detalles de este algoritmo se ven en la Sección 3.6. El resultado del procedimiento completo es el orden circular agre-

gado denominado $\widehat{\boldsymbol{O}}$.

Vamos a suponer que disponemos de la información dada por una observación para cada elemento y experimento. Matemáticamente tendremos p vectores de $[0, 2\pi)^n$. Algunas de las técnicas que proponemos hacen uso directo de los vectores, lo que sería utilizar un primer nivel de la información. Otras técnicas hacen uso de relaciones entre pares, que sería usar un segundo nivel de información (en este caso se representa matemáticamente la información con matrices en $\mathbb{R}^{n\times n}$). Un tercer nivel de información viene dado por las relaciones entre tripletas de elementos que matemáticamente se representan mediante hipermatrices en $\mathbb{R}^{n\times n\times n}$.
Las técnicas propuestas en este capítulo para el paso 1 se pueden clasificar según el nivel de información tal y como se ilustra en la Figura 3.3. La principal novedad frente a los métodos en la línea, es el uso del tercer nivel de información, el de las tripletas, que aparece de forma natural en el círculo ya que tres es el conjunto mínimo de elementos que mantiene una relación de asociación en el círculo (ver Sección 2.2.1).

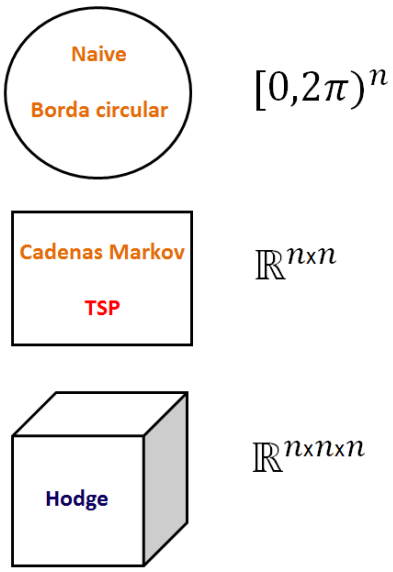


Figura 3.3: Clasificación de las técnicas de agregación de órdenes circulares según el espacio

Agruparemos los métodos derivados de la línea y adaptados al círculo en la Sección 3.3 (Naive, Borda Circular y cadenas de Markov). Dedicaremos por su importancia la Sección 3.4 a los métodos basados en el problema del viajante (TSP) y la Sección 3.5 a los basados en la teoría de Hodge.

3.3. Técnicas adaptadas de la agregación de rankings

En esta sección presentamos tres técnicas para resolver el problema (3.4). La primera de ellas es la más simple y por eso se denomina Naive. La segunda técnica es una adaptación al círculo del método más usado en la línea, Borda (ver Sección 2.5.2); ambas hacen uso de la representación de la información en $[0, 2\pi)^n$. En tercer lugar presentamos la técnica basada en cadenas de Markov que tiene en cuenta la información de las relaciones entre pares de elementos y por tanto hace uso del segundo nivel de información (matrices en $\mathbb{R}^{n\times n}$).

3.3.1. Método Naive ($[0, 2\pi)^n$)

La idea de este método es muy sencilla, se escoge de entre los órdenes que generan los datos aquel que hace mínima la función objetivo. Según este criterio el orden global viene dado por,

$$\widehat{\boldsymbol{O}}^0 = \arg \min_{\mathbf{O}\in\{\boldsymbol{O}_1,\ldots,\boldsymbol{O}_j,\ldots,\boldsymbol{O}_p\}} MSCE(\boldsymbol{\Theta}, \mathbf{O})$$

donde $\{\boldsymbol{O}_1, \ldots, \boldsymbol{O}_j, \ldots, \boldsymbol{O}_p\}$ es el conjunto de los p órdenes circulares generados por los datos de los p experimentos $\{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_j, \ldots, \boldsymbol{\Theta}_p\}$.

Aunque este método ofrece una buena aproximación, en casos concretos, en general no es una buena opción. Por ejemplo, si n es grande y p pequeño este método no funciona correctamente.

3.3.2. Métodos Borda circular ($[0, 2\pi)^n$)

En este apartado presentamos la adaptación al circulo del método de Borda (Borda (1781)) que a pesar de ser uno de los primeros que surgió para resolver el problema de agregación de rankings en la línea sigue siendo uno de los más usados (Baltrunas et al. (2010), Mekonnen (2014)). Para adaptar el método al caso circular hay que tener en cuenta que las observaciones no se pueden agrupar directamente porque los diferentes experimentos pueden tener distintos puntos de inicio. La propuesta que hacemos es rotar previamente las observaciones de forma que denotamos por θ_{ij}^l a las observaciones rotadas,

$$\theta_{ij}^l = (\theta_{ij} - \theta_{lj}) \pmod{2\pi} \ \forall i = 1, \ldots, n, j = 1, \ldots, p \tag{3.5}$$

donde l varia de 1 a n.
El método se ejecuta para cada conjunto de datos $\boldsymbol{\Theta}^l$, obteniendo entonces un orden circular agregado de cada ejecución.
Sea $\{\overline{\boldsymbol{O}}^1, \ldots, \overline{\boldsymbol{O}}^l, \ldots, \overline{\boldsymbol{O}}^n\}$ el conjunto de órdenes agregados donde $\overline{\boldsymbol{O}}^l$ es el orden circular de los elementos dado por el vector de valores agregados obtenido usando $\boldsymbol{\Theta}^l$, entonces, se define,

$$\widehat{\boldsymbol{O}}^0 = \arg \min_{\mathbf{O} \in \{\overline{\boldsymbol{O}}^1, \ldots, \overline{\boldsymbol{O}}^l, \ldots, \overline{\boldsymbol{O}}^n\}} MSCE(\boldsymbol{\Theta}, \mathbf{O}).$$

Las diferentes variantes de la técnica Borda circular provienen de la distintas formas de usar (las posiciones o *scores*) y de agregar (medias o medianas) la información una vez rotados los datos, en la Tabla 3.2 se definen tres propuestas que valoramos en esta memoria.
El método **BCpos** (Borda Circular Posiciones) hace uso de la información de las posiciones en el orden de los datos rotados mediante $\tau_{ij}^l(2\pi/n)$, el valor circular correspondiente a la posición de cada elemento i en el orden del experimento j rotado en el elemento l.
Los métodos **BCmean** (Borda Circular Medias) y **BCmed** (Borda Circular Medianas) hacen uso de los datos observados rotados (3.5).
Estos tres métodos son simples pero tienen un tiempo de ejecución elevado. Además hemos detectado que en varios ejemplos reales las soluciones con esta

Tabla 3.2: Métodos Borda Circular

Etiqueta	Información	Agregación
BCpos	Posiciones	Media circular
BCmean	*Scores*	Media circular
BCmed	*Scores*	Mediana circular

técnica no son biológicamente interpretables, incluso en escenarios muy simples como el del Ejemplo 3.2, expuesto en la Sección 3.7. A lo largo de esa sección se presentan ejemplos que ilustran las características de cada método. Se presentan una vez que se han introducido todas las técnicas.

3.3.3. Métodos basados en cadenas de Markov ($\mathbb{R}^{n\times n}$)

Esta técnica es una extensión del método presentado por Dwork et al. (2001a) para agregación de rankings en la línea (ver Sección 2.5.2). En esta formulación recordamos que los estados representan a los elementos y las probabilidades de transición entre dichos elementos cuantifican la relación entre pares. De esta manera se representa la información de los órdenes observados en el espacio $\mathbb{R}^{n\times n}$. El orden agregado $\widehat{\boldsymbol{O}}^0$ viene dado por el orden de la distribución estacionaria calculada para la matriz de transición que contiene las probabilidades de transición.

La adaptación al círculo de esta técnica mantiene su estructura principal salvo en la definición de las probabilidades de transición entre estados p_{hk} que deben tener en cuenta la geometría particular de las observaciones circulares. Las diferentes alternativas que presentamos son extensiones al círculo de las definiciones propuestas por Dwork et al. (2001a) y se concretan en la Tabla 3.3.

En esta tabla se hace uso de las cantidades,

$$\begin{aligned} D^j_{hk} &= (\tau_{kj} - \tau_{h-1j}) \text{ (mód } n) \\ F^j_{hk} &= (\tau_{hj} - \tau_{k-1j}) \text{ (mód } n), \quad \forall j = 1, \ldots, p, \forall h, k = 1, \ldots, n, \end{aligned} \tag{3.6}$$

donde D^j_{ih} es el número de elementos que se encuentra entre h y k en el experimento j en la dirección de rotación y equivalentemente F^j_{hk} para la dirección contraria a la rotación.

Tabla 3.3: Métodos basados en cadenas de Markov

Etiqueta	p^j_{hk}, $h, k = 1, \ldots, n$, $j = 1, \ldots, p$.
CMC1	$\frac{1}{\#\{h:p_{hk}>0\}}$ si $\exists\ j : D^j_{hk} \leq F^j_{hk}$ 0 si $D^j_{hk} > F^j_{hk}$
CMC2	$\frac{1}{p}\sum_{j=1}^{p} I_j(h,k) \cdot \frac{2}{(n-1)}$ donde $I_j(h,k) = \begin{cases} 1 & \text{si } D^j_{hk} \leq F^j_{hk} \\ 0 & \text{si } D^j_{hk} > F^j_{hk}. \end{cases}$
CMC3	$\frac{1}{p}\sum_{j=1}^{p} I_j(h,k)\frac{1}{n}$
CMC4maj	$\frac{1}{n}$ si $\#M \geq \frac{p}{2}$ 0 si $\#M < \frac{p}{2}$, $1 - \sum_{k=1}^{n} p_{hk}$ si $h = k$ donde $M = \{j : D^j_{hk} \leq F^j_{hk}\}$
CMC4num	$\frac{1}{n}$ si $\sum_j D^j_{hk} \leq \sum_j F^j_{hk}$ 0 si $\sum_j D^j_{hk} > \sum_j F^j_{hk}$, $1 - \sum_{k=1}^{n} p_{hk}$ si $h = k$

El comportamiento de esta técnica basada en cadenas de Markov en la linea es muy satisfactorio (Liu et al. (2007)). Esa es la principal razón que nos ha

llevado a introducirla en esta memoria y compararla con el resto de propuestas a pesar de haber detectado que en el círculo no ofrece buenos resultados como se muestra posteriormente en las simulaciones.

3.4. Técnica basada en el problema del viajante ($\mathbb{R}^{n\times n}$)

El enfoque de resolución de la agregación de órdenes circulares que mostramos en esta sección es una idea original que surge a partir de detectar, en el caso del círculo, que la distribución estacionaria asociada a las cadenas de Markov, definidas en la Sección 3.3.3, no es el objetivo razonable y si lo es el buscar la *ruta* más probable entre elementos. De hecho, utilizando la representación de los elementos (nodos) en un grafo se puede establecer una relación uno a uno entre un orden circular y una ruta que pase por todos los nodos una sola vez y vuelva al nodo inicial. Así llegamos a plantear un problema del viajante (TSP) que será la aproximación al problema (3.4) que proponemos en esta sección.

El problema del viajante (**TSP**, del inglés *Traveling Salesperson Problem*) es uno de los problemas más famosos, y quizás el mejor estudiado, en el campo de la optimización combinatoria computacional y la teoría de grafos (Flood (1956), Lawler et al. (1985), Reinelt (1994), Hahsler y Hornik (2011)). El objetivo es, dado un conjunto de localizaciones y distancias (o costes de desplazamiento) entre ellas, encontrar una ruta que, comenzando y terminando en la misma localización, pase exactamente una vez por cada una de las demás localizaciones, minimizando la distancia total recorrida por el viajante.

En esta propuesta cada experimento j se representa mediante un grafo dirigido donde los nodos son los elementos. Cada par de nodos (h, k) está conectado por una arista dirigida cuya longitud E^j_{hk} se corresponde con una medida de in-

tensidad o *preferencia* de h a k. E^j_{hk} se define más adelante en la Sección 3.4.1, utilizando diferentes distancias dirigidas, es decir distancias que tienen en cuenta la dirección de la arista que se esta midiendo. Notar que pueda ocurrir que $E^j_{hk} \neq E^j_{kh}$.

Una ruta que pasa exactamente una vez por todos los nodos del grafo, empezando y terminando en el mismo nodo, se corresponde con un orden circular entre los respectivos elementos. Por ejemplo, la ruta que siguen las flechas verdes en el grafo dirigido de la Figura 3.4 se corresponde con el orden circular (1,3,2,4,6,5,7).

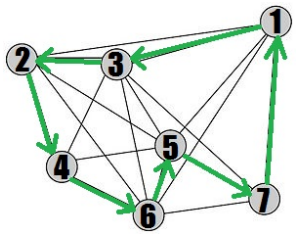


Figura 3.4: Ejemplo de ruta en un grafo dirigido

El problema de optimización del viajante de interés en nuestro caso se formula como sigue,

$$\widehat{X} = \arg\min_{X\in\mathcal{X}} \sum_{hk} X_{hk} E^*_{hk} = \arg\min_{X\in\mathcal{X}} \sum_{j=1}^{p} \omega_j \left(\sum_{hk} X_{hk} E^j_{hk} \right). \tag{3.7}$$

Sujeto a

$$(1) \quad \sum_{h=1}^{n} X_{hk} = 1 \quad \forall k = 1, \ldots, n$$

$$(2) \quad \sum_{k=1}^{n} X_{hk} = 1 \quad \forall h = 1, \ldots, n$$

$$(3) \quad \sum_{h\in S, k\in S} X_{hk} \leq |S| - 1 \ \ \forall S \subset V, |S| > 1,$$

donde ω_j es un peso asociado al experimento j y E^* es la matriz agregada con la información de los p experimentos tal que, $E^*_{hk} = \sum_{j=1}^{p} \omega_j E^j_{hk}$. Una ruta viene representada por una matriz binaria X donde $X_{hk} = 1$ si la artista (h, k)

está activa en la ruta y se corresponde con el orden circular de la siguiente forma,

$$X_{hk} = 1 \Leftrightarrow \exists i, 1 \leq i \leq n : o_i = h, o_{i+1} = k.$$

Por ejemplo, en el caso de la Figura 3.4, $X_{13} = 1$ y $X_{1k} = 0\ \forall k \neq 3$.
Así, el orden circular $\widehat{\boldsymbol{O}}^0$ correspondiente a la ruta $\widehat{X}$ que resuelve (3.7) es la solución propuesta como aproximación de (3.4) haciendo uso de la técnica TSP.

El TSP es un problema inicialmente del tipo NP-hard (Papadimitriou y Steiglitz (1998)) debido al tiempo que se precisa para calcular el óptimo. Sin embargo, si existe un máximo L de longitud total de ruta, el problema de decisión es NP-completo (Orponen y Mannila (1987)) y por tanto más sencillo de resolver. En concreto, el TSP ha sido ampliamente estudiado y existen heurísticas que aproximan adecuadamente la solución, véase Reinelt (1994). Podemos adelantar que la rapidez computacional será una de las ventajas de esta solución al hacer uso de las heurísticas.

En la Sección 3.4.1 presentamos diferentes alternativas para definir las distancias que miden la longitud de arista dirigida entre pares de elementos. En la Sección 3.4.2 se presenta un algoritmo de resolución del problema (3.7) haciendo uso de diversas heurísticas.

3.4.1. Definición de E^j_{hk}

Recordamos que E^j_{hk} se corresponde con una medida de intensidad y se va a definir usando una distancia dirigida, es decir que, como ya se ha comentado, permita $E^j_{hk} \neq E^j_{kh}$, de manera que se tenga en cuenta si se usa una dirección de rotación o la contraria. Proponemos en esta sección diferentes definiciones para E^j_{hk} dependiendo de si usamos información cardinal u ordinal y el tipo del distancia. Las propiedades básicas necesarias que debe cumplir E^j_{hk} para usar las heurísticas ya conocidas que resuelven (3.7), son que sea positiva y tenga un máximo; es decir, que no tiene que cumplir las propiedades de una métrica

necesariamente, sino que puede tratarse simplemente de una medida de la cercanía o lejanía entre los elementos. La continuidad en todo $[0, 2\pi)$ será también una propiedad deseable.

A excepción de la distancia que hace uso de los tiempos que se define al final de este apartado, el resto de las distancias consideradas se pueden formular de acuerdo a la siguiente definición general:

$$E^j_{hk} = \text{mín}(d_R(\theta_{hj}, \theta_{kj}),\ \ \alpha \cdot d_C(\theta_{hj}, \theta_{kj})), \tag{3.8}$$

donde d_R y d_C son a su vez distancias en el sentido de la dirección de rotación y en el contrario respectivamente y $\alpha \geq 1$ es una constante de penalización que aporta flexibilidad como veremos en las propuestas concretas (Tabla 3.4). Cabe comentar que d_R y d_C pueden ser distancias asimétricas ($d_R(\theta_{hj}, \theta_{kj}) \neq d_R(\theta_{kj}, \theta_{hj})$) o simétricas ($d_R(\theta_{hj}, \theta_{kj}) = d_R(\theta_{kj}, \theta_{hj})$).
La idea que motivó esta definición procede de un problema denominado *directed circular arrangement* presentado por Naor y Schwartz (2010) donde se hace uso de dos formulaciones diferentes de la distancia dependiendo de la dirección de rotación. En el mismo artículo se propone la incorporación de la penalización α que aquí incluimos en el caso de hacer uso de la dirección contraria a la rotación del círculo.

Todas las distancias que resultan de las definiciones de la Tabla 3.4 y de usar (3.8) cumplen las características imprescindibles, ser positivas y tener un máximo, para ser usadas en la resolución del problema (3.7). Además, las distancias usadas para definir **TSPcho**, **TSP1**, **TSP2**, **TSP3** y **TSP4** cumplen la propiedad de continuidad. Las tres últimas vienen dadas por definiciones de E^j_{hk} asimétricas por lo que tienen en cuenta la dirección de rotación del círculo. Así, cumplen todas las características deseables para E^j_{hk}. Estas últimas definiciones son casos particulares de la distancia asociada a TSPα que denominamos $d_\alpha(\theta_h, \theta_k)$ y que mantiene las características vistas para **TSP2**, **TSP3** y **TSP4** siempre que $1 < \alpha < \infty$. En la Proposición 3.1 se realiza un estudio más detallado de las propiedades de $d_\alpha(\theta_h, \theta_k)$.

Tabla 3.4: Métodos basados en la técnica TSP según las distancias dirigidas con penalización α

Etiqueta		α	$d_R(\theta_{hj},\theta_{kj})\ \forall j=1,\dots,p;\ h,k=1,\dots,n.$	$d_C(\theta_{hj},\theta_{kj})\ \forall j=1,\dots,p;\ h,k=1,\dots,n.$
TSPbin		∞	1 si $\tau_{kj}=\tau_{hj}+1\ (\text{mód } 2\pi)$ 0 si $\tau_{kj}\neq\tau_{hj}+1\ (\text{mód } 2\pi)$	1 si $\tau_{hj}=\tau_{kj}+1\ (\text{mód } 2\pi)$ 0 si $\tau_{hj}\neq\tau_{kj}+1\ (\text{mód } 2\pi)$
TSPpos		∞	$\tau_{kj}-\tau_{hj}\ (\text{mód } 2\pi)$	$\tau_{hj}-\tau_{kj}\ (\text{mód } 2\pi)$
TSParc		∞	$\theta_{kj}-\theta_{hj}\ (\text{mód } 2\pi)$	$\theta_{hj}-\theta_{kj}\ (\text{mód } 2\pi)$
TSPcho		1	$\sqrt{2-2\cos(\theta_{kj}-\theta_{hj})}$	$\sqrt{2-2\cos(\theta_{hj}-\theta_{kj})}$
TSPα	**TSP1**	1		
	TSP2	2	$1-\cos(\theta_{kj}-\theta_{hj})$ si $0\leq\theta_{kj}-\theta_{hj}\leq\pi$	$3-\cos(\theta_{kj}-\theta_{hj}-\pi)$ si $0\leq\theta_{kj}-\theta_{hj}\leq\pi$
	TSP3	3	$3-\cos(\theta_{kj}-\theta_{hj}-\pi)$ si $\pi<\theta_{kj}-\theta_{hj}<2\pi$	$1-\cos(\theta_{kj}-\theta_{hj})$ si $\pi<\theta_{kj}-\theta_{hj}<2\pi$
	TSP4	4		
	TSP∞	∞		

Proposición 3.1. *$d_\alpha(\theta_h, \theta_k)$ verifica las siguientes propiedades:*

- ***1.*** *Positividad.* $d_\alpha(\theta_h, \theta_k) \geq 0 \ \forall \alpha \geq 1$.
- ***2.*** *Reflexividad.* $d_\alpha(\theta_h, \theta_h) = 0 \ \forall \alpha \geq 1$.
- ***3.*** *Continuidad en* $[0, 2\pi) \times [0, 2\pi) \ \forall \ 1 \leq \alpha < \infty$.
- ***4.*** *Identidad de los indiscernibles.* $d_\alpha(\theta_h, \theta_k) = 0 \Leftrightarrow \theta_h = \theta_k$,
- ***5.*** *Acotada superiormente.* $d_\alpha(\theta_h, \theta_k) \leq L(\alpha)$ *con* $L < 4 \ \forall \alpha \geq 1$.
- ***6.*** *Desigualdad triangular relajada* $d_\alpha(\theta_h, \theta_k) \leq 2(d_\alpha(\theta_h, \theta_i) + d_\alpha(\theta_i, \theta_k))$ $\forall \theta_h, \theta_k, \theta_i \in [0, 2\pi), \alpha \geq 1$.

Demostración. La demostración de las propiedades de la 1 a la 5 es trivial y por tanto pasamos a demostrar la propiedad 6.

En términos generales podemos escribir la distancia de la forma siguiente,

$$d_\alpha(\theta_h, \theta_k) = \begin{cases} 1 - \cos(\theta_k - \theta_h) & \text{si } \theta_k - \theta_h \in A \\ 3 - \cos(\theta_k - \theta_h - \pi) & \text{si } \theta_k - \theta_h \in B \\ \alpha(1 - \cos(\theta_k - \theta_h)) & \text{si } \theta_k - \theta_h \in C \end{cases}$$

donde $A = \{\theta_k, \theta_h : 0 \leq \theta_k - \theta_h \leq \pi\}$, $B = \{\theta_k, \theta_h : \pi < \theta_k - \theta_h \leq \delta(\alpha)\}$, $C = \{\theta_k, \theta_h : \delta(\alpha) < \theta_k - \theta_h < 2\pi\}$ y $\delta(\alpha)$ es el valor $x \in [\pi, 2\pi]$ para el que $3 - \cos(x - \pi) = \alpha(1 - \cos x)$. Notar que $\alpha = \infty$ es equivalente a $C = \varnothing$. Puesto que esta distancia depende de θ_h, θ_k sólo a través de $\theta_h - \theta_k$ podemos definir $y = \theta_i - \theta_k$, $z = \theta_h - \theta_k$ lo que es equivalente a suponer $\theta_k = 0$ y nos permitirá trabajar solamente con dos ángulos. En otras palabras, para demostrar este resultado es suficiente probar que $d_\alpha(z, 0) \leq 2(d_\alpha(z, y) + d_\alpha(y, 0))$.

Teniendo en cuenta que la distancia está definida de forma diferente en varios sectores vamos a considerar varios casos en esta prueba. Denotamos

$$f(y, z, \tau) = \tau(d_\alpha(z, y) + d_\alpha(y, 0)) - d_\alpha(z, 0)$$

y probaremos que $f(y, z, \tau)$ es positiva para $\tau = 2$.

1. Supongamos que $0 \leq y \leq z \leq 2\pi$.

   a) $z \in A$. En este caso $f(y,z,\tau) = 2\tau - 1 + \cos z - \tau\,(\cos y + \cos(z-y))$. Derivando con respecto a y y z y discutiendo el correspondiente sistema de ecuaciones obtenemos que los mínimos locales de esta función aparecen cuando $y = z = 0$, cuando $y = z = 2\pi$ y cuando $y = \arccos(\tau/2)$, $z = 2\arccos(\tau/2)$ (este último obviamente sólo aparece si $\tau \leq 2$). Para el primer caso tenemos que $f(0,0,\tau) = 0$, para el segundo $f(\pi,\pi,\tau) = 2\tau - 2$, y para el tercero $f(\arccos(\tau/2), 2\arccos(\tau/2), \tau) = \frac{-(\tau-2)^2}{2}$. Podemos concluir por tanto que para que $f(y,z,\tau)$ sea positiva es necesario que $\tau \geq 2$.

   b) $z \in B$, $y, z-y \in A$. Tenemos que $f(y,z,\tau) = 2\tau - 3 - \cos z - \tau\,(\cos y + \cos(z-y))$. Si efectuamos un análisis similar al anterior encontramos que en este caso no hay mínimos y que la función toma siempre valores positivos para $\tau \geq 1$.

   c) $z \in B$, $y \in B$, $z - y \in A$ (ó $z - y \in B$, $y \in A$). Es obvio que los dos casos son equivalentes con lo que sólo trataremos el primero. Ahora $f(y,z,\tau) = 4\tau - 3 - \cos z - \tau\,(-\cos y + \cos(z-y))$. Los mínimos aparecen cuando $y = z = 2\pi$ o cuando $y = \pi + \arccos(\tau/2)$, $z = \pi + 2\arccos(\tau/2)$ (de nuevo este último obviamente sólo aparece si $\tau \leq 2$). Para el primer caso tenemos que $f(2\pi, 2\pi, \tau) = 4\tau - 4$ y para el segundo $f(\pi + \arccos(\tau/2), \pi + 2\arccos(\tau/2), \tau) = -4 + 4\tau - (\tau^2)/2$. Podemos concluir por tanto que, en este caso, para que $f(y,z,\tau)$ sea positiva es necesario que $\tau \geq 2(2 - \sqrt{2}) = 1.17$.

   d) $z \in C$. Este caso puede reducirse a los anteriores teniendo en cuenta que si $z \in C$, $d_\alpha(z,0) = \alpha(1 - \cos z) \leq 3 - \cos(z - \pi)$ y que al ser $\alpha \geq 1$ se tiene que $\alpha(1 - \cos y) \geq (1 - \cos y)$.

2. Supongamos ahora que $0 \leq z < y \leq 2\pi$.

   a) $y \in A \cup B$. En este caso $d_\alpha(y,0) \geq d_\alpha(z,0)$ luego $f(y,z,\alpha) \geq 0$ para cualquier $\tau \geq 1$.

b) $y \in C$, $z + 2\pi - y \in A \cup B$. En esta otra situación $d_\alpha(z, y) \geq d_\alpha(z, 0)$ luego $f(y, z, \alpha) \geq 0$ para cualquier $\tau \geq 1$.

c) $y \in C$, $z + 2\pi - y \in C$. Si se tiene en cuenta que $\alpha \geq 1$ puede verse que, dependiendo de si $z \in A, B$ o C, este caso se reduce al (1.a), (1.b) o (1.d) respectivamente.

Recopilando todos los casos anteriores comprobamos que hemos demostrado que $f(y, z, 2) \geq 0, \forall y, z \in [0, 2\pi]$.

■

De todos los posibles valores de α hay uno desacado por su interpretación geométrica que es $\alpha = 3$ cuando se usa la distancia en el sentido contrario al de rotación $d_C(\theta_{hj}, \theta_{kj})$. En estos casos se puede interpretar como el camino que hay que recorrer cuando una ruta ha llegado a h sin visitar una localización anterior k y debe retroceder hasta ella recorriendo en total 3 veces el camino entre h y k (Figura 3.5).

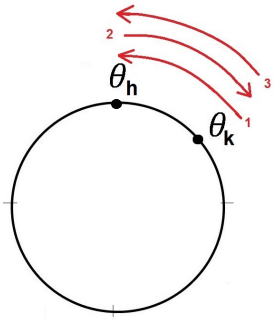


Figura 3.5: Interpretación geométrica de la penalización $\alpha = 3$

TSPtime es la última propuesta para E^j_{hk} y se define como,

$$E^j_{hk} = t(\theta_{hj}, \theta_{kj}) = \begin{cases} \frac{2 \cdot (1 - \cos(\theta_{kj} - \theta_{hj}))}{3 \cdot \pi} & \text{si } \theta_{kj} - \theta_{hj} \leq \pi \ \ (\text{mód } 2\pi) \\ \\ \frac{2 \cdot (1 - \cos(\theta_{kj} - \theta_{hj}))}{\pi} & \text{si } \theta_{kj} - \theta_{hj} > \pi \ \ (\text{mód } 2\pi), \end{cases}$$

donde $t(\theta_{hj}, \theta_{kj})$ es el tiempo que ocupa el recorrido entre θ_{hj} y θ_{hj}, Si la dirección es la correcta (contraria a las agujas del reloj) se produce una aceleración

y en el caso contrario una deceleración. Dicha aceleración se representa por α (con valores negativos en caso de deceleración) y se calcula mediante las ecuaciones (3.9) del Movimiento Circular Uniformemente Acelerado (MCUA) (ver Kane et al. (1989)).
E^j_{hk} viene dada por $t_{kj} - t_{hj}$, el tiempo entre el punto inicial θ_{hj} y el punto final θ_{kj} obtenido según las ecuaciones del MCUA que tienen la siguiente forma,

$$\begin{aligned} \omega_{kj} - \omega_{hj} &= \alpha(t_{kj} - t_{hj}) \\ d(\theta_{hj}, \theta_{kj}) &= \omega_{hj}(t_{kj} - t_{hj}) + \tfrac{1}{2}\alpha(t_{kj} - t_{hj})^2 \end{aligned} \tag{3.9}$$

para todo $h, k = 1, \ldots, n$ y $j = 1, \ldots, p$, donde $d(\theta_{hj}, \theta_{kj}) = 1 - \cos(\theta_{kj} - \theta_{hj})$ es la distancia entre el punto inicial y el punto final, $\omega_{hj} = \pi\, m/sg$ es la velocidad angular inicial y ω_{kj} es la velocidad angular final que se fija en $\omega_{kj} = 2\pi\, m/sg$ en caso de seguir la dirección de rotación del círculo (acelerando) y en $\omega_{kj} = 0\, m/sg$ si se sigue la dirección contraria a la de rotación (decelerando).

Esta definición de la distancia hace uso de los tiempos para penalizar la dirección contraria del círculo de forma similar a ciertas variantes generales del TSP que consideran el tiempo necesario para realizar un recorrido (véase Gong et al. (2007), Tsitsiklis (1992), Malandraki y Daskin (1992)).

3.4.2. Algoritmo para resolver el problema de optimización

Como ya comentamos al inicio de la sección, no es sencillo encontrar el óptimo del TSP debido a que se trata de un problema de optimización NP-hard. La gran ventaja es que existen múltiples heurísticas que ofrecen una muy buena aproximación. En concreto, nosotros usamos aquellas programadas en el paquete de R llamado **TSP**, Hahsler y Hornik (2011), en particular: *nearest neighbor*, *repetitive nearest neighbor*, *nearest insertion*, *farthest insertion*, *cheapest insertion* y *arbitrary insertion*. En el caso del uso de estas heurísticas

se requiere al menos el cumplimiento de la desigualdad triangular relajada de la definición de la longitud de la arista (ver Moret y Shapiro (1991), Bender y Chekuri (2000), Andreae (2001), Forlizzi et al. (2005), Abdullah et al. (2012)). Esta propiedad se cumple con coeficiente máximo de 2 en el caso que nos ocupa (ver Proposión 3.1).

En general, no está comprobado que exista alguna heurística con mejor comportamiento que el resto. En nuestro problema particular hemos comprobado en muchos ejemplos que no existe una heurística ganadora. Para ello hemos simulado diferentes situaciones comprobando si existía alguna tendencia de mejor aproximación al óptimo del problema (3.4). Con el fin de subsanar esta desventaja, hemos diseñado un algoritmo para obtener la solución (ruta) que más se aproxime al óptimo de (3.4). El Algoritmo 2 que se presenta abajo obtiene las soluciones del TSP ejecutando repetidas veces las diferentes heurísticas y se queda con aquella ruta que hace mínimo el MSCE.

El número de rutas entre las que se obtiene la solución depende del valor de una constante denominada c que se introduce como parámetro de entrada al algoritmo, siendo consideradas como posibles soluciones finales las $c \cdot n$ rutas de menor longitud. En las simulaciones (Sección 3.8.3) se trata la elección del valor de c.

Para finalizar esta sección mostramos un esquema con todos los pasos del procedimiento completo siguiendo el enfoque TSP en la Figura 3.6.

Algoritmo 2: Agregación mediante el enfoque TSP

entrada: $\boldsymbol{\Theta}$; E^*; c.

salida : $\widehat{\boldsymbol{O}}^0$.

R: conjunto de rutas resultado del TSP y distintas entre sí.;

H={*nearest neighbor*, *repetitive nearest neighbor*, *nearest insertion*, *farthest insertion*, *cheapest insertion*, *arbitrary insertion*};

for *i in 1:n* **do**

  for *j in 1:#H* **do**

    X_i^j: solución del TSP con la heurística $H_{[j]}$;

    if $X_i^j \notin R$ **then**

      R=R$\cup X_i^j$;

$\mathcal{L}$: conjunto de las c·n rutas de menor longitud de R, $\mathcal{L} \subset R$;

$\widehat{\boldsymbol{O}}^0 = \arg\min_{\mathbf{O}\in\mathcal{L}} MSCE(\boldsymbol{\Theta}, \mathbf{O})$;

return $\widehat{\boldsymbol{O}}^0$;

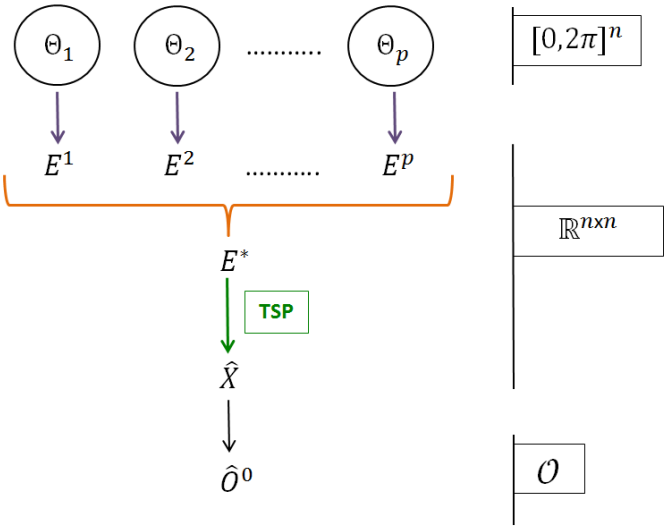


Figura 3.6: Esquema del procedimiento de agregación de órdenes circulares usando la técnica basada en el TSP

3.5. Técnica basada en la teoría de Hodge ($\mathbb{R}^{n\times n\times n}$)

En esta sección se propone una técnica basada en la teoría de Hodge para la resolución del problema de agregación de órdenes circulares. La novedad principal que introducimos con esta técnica es el usar la información de la intensidad entre tripletas de elementos. Hasta ahora, tanto en los procedimientos estudiados en esta memoria, como en los precedentes basados en la teoría de Hodge para la agregación de rankings (ver Sección 2.5.3), se hacía uso de la información individual o como mucho de la información sobre intensidades de relación entre pares. Las ventajas del enfoque que aquí proponemos son múltiples, comenzando por que las tripletas aparecen de manera natural al ser el conjunto mínimo con una relación de orden en el círculo y permiten agregar la información de forma directa independientemente del punto de inicio del círculo. Otra de las ventajas es la flexibilidad tanto en las diferentes formas de introducir la información, definición directa de las tripletas o transformación de la información individual (vectores) o de la información entre pares (matrices), así como de las diversas formas de realizar la agregación de la información. Aunque desde un punto de vista formal esta técnica requiere de una base teórica importante, desde un punto de vista computacional los cálculos son muy simples y por tanto el tiempo de ejecución es muy corto.

La propuesta que hacemos es representar la intensidad de la relación entre tripletas para cada experimento j mediante una hipermatriz denominada $\Psi^j \in \mathbb{R}^{n\times n\times n}$, $j = 1, \ldots, p$ que debe cumplir la propiedad de ser antisimétrica ($\psi^j_{ihk} = \psi^j_{hki} = \psi^j_{kih} = -\psi^j_{ikh} = -\psi^j_{khi} = -\psi^j_{hki}$, $i, h, k = 1, \ldots, n; j = 1, \ldots, p$) por ser una condición básica para poder utilizar resultados teóricos relevantes para este problema. En este trabajo proponemos la definición general de sus componentes dada por,

$$\psi^j_{ihk} = signo^j(i, h, k) \cdot \lambda^j_{ihk}, \quad i, h, k = 1, \ldots, n, \;\; j = 1, \ldots, p,$$

donde $signo^j(i, h, k)$ es el signo de la tripleta (i, h, k), $1 = signo^j(i, h, k) = signo^j(h, k, i) = signo^j(k, i, h) = -signo^j(i, k, h) = -signo^j(k, h, i) = -signo^j(h, k, i)$,

siendo (i,h,k) concordante con la dirección de rotación del círculo y λ^j_{ihk} es una medida de la intensidad de relación de la tripleta (i,h,k) que toma valores máximos si los 3 elementos están equiespaciados en el círculo, el valor de λ^j_{ihk} no depende de la permutación.

Esta definición es independiente del punto de inicio y por tanto permite agregar directamente la información de los experimentos que tienen diferentes puntos de inicio. En este trabajo, definiremos la hipermatriz agregada $\overline{\Psi} \in \mathbb{R}^{n\times n\times n}$ como sigue,

$$\overline{\psi}_{ihk} = \sum_{j=1}^{p} \omega_j \psi^j_{ihk},$$

donde ω_j es el peso del experimento j.

El resto de esta sección se divide en un primer apartado, Sección 3.5.1, donde se expone el procedimiento que proponemos para obtener el orden circular agregado a partir de $\overline{\Psi}$ y un segundo apartado, Sección 3.5.2, donde concretamos las diferentes definiciones que proponemos para λ^j_{ihk}.

3.5.1. Obtención del orden circular agregado

La terminología que se utiliza en esta sección es similar a la que se usa en Jiang et al. (2011) y puede consultarse en el Apéndice A.
Como veremos a continuación la solución a partir de $\overline{\Psi}$ se obtiene transformando la hipermatriz en una matriz y esta en un vector. Estas transformaciones que se llevan a cabo precisan de ciertos operadores entre diferentes espacios, ($[0,2\pi)^n$, $\mathbb{R}^{n\times n}$ y $\mathbb{R}^{n\times n\times n}$) y para definirlos se necesita previamente definir productos escalares en cada espacio.
Así, se define el producto escalar en $\mathbb{R}^{n\times n\times n}$ como: $\langle \Psi^1, \Psi^2 \rangle_{\boldsymbol{\omega}} = \sum_{ihk} w_{ihk}\psi^1_{ihk}\psi^2_{ihk}$, donde $\boldsymbol{w}$ es el peso de la tripleta, que puede ser usado para contabilizar los datos *missing* o para incorporar el peso del experimento o de las observaciones (eliminamos $\boldsymbol{\omega}$ a lo largo de la exposición por simplicidad), se define el pro-

ducto escalar en $\mathbb{R}^{n\times n}$ como: $\langle X, Y\rangle = \sum_{ih} X_{ih}Y_{ih}$ y el producto escalar en $\mathbb{R}^n$ como: $\langle \boldsymbol{r}, \boldsymbol{s}\rangle = \sum_{i=1}^{n} r_i s_i$.
Siguiendo el desarrollo de Jiang et al. (2011), los operadores δ_1^*, δ_1, δ_0^* y δ_0 se definen tal que verifiquen,

$$< \delta_k f_k, g_{k+1} >_{k+1} = < f_k, \delta_k^* g_{k+1} >_k, \qquad k = 0, 1.$$

Las definiciones formales de cada uno de ellos se muestran en el Apéndice A y la formulación específica de la que haremos uso en esta sección, en la Tabla 3.5. En concreto, el operador adjunto rotacional lleva una hipermatriz antisimétrica $\Psi \in \mathbb{R}^{n\times n\times n}$ a una matriz antisimétrica $Y \in \mathbb{R}^{n\times n}$ mientras que el operador rotacional hace la transformación contraria. El operador adjunto del gradiente transforma una matriz antisimétrica ($Y_{ih} = -Y_{hi}$) en un vector en $\mathbb{R}^n$ y para la transformación contraria se usa el operador gradiente.

Tabla 3.5: Operadores de Hodge

Adjunto rotacional:	$\delta_1^*(\Psi) = Y,$ donde $Y_{ih} = \sum_k \psi_{ihk}$
Rotacional:	$\delta_1(Y) = \Psi,$ donde $\psi_{ihk} = Y_{ih} + Y_{hk} + Y_{ki}$
Adjunto del gradiente:	$\delta_0^*(Y) = \boldsymbol{s},$ donde $s_i = \sum_h Y_{ih}$
Gradiente:	$\delta_0(\boldsymbol{s}) = Y,$ donde $Y_{ih} = s_h - s_i$

Además, introducimos algunos términos nuevos como el uso del superíndice (l) que indica, en cualquier subespacio o subconjunto, que se ha eliminado dicho elemento l y por tanto la dimensión se ha reducido una unidad.
A continuación, definimos el conjunto $\mathcal{H}_C$ como el conjunto de hipermatrices que inducen un orden circular. La justificación se verá una vez introducidos los diferentes subconjuntos que intervienen en la definición de $\mathcal{H}_C$ que tiene la siguiente forma,

$$\mathcal{H}_C = \{\Psi \in \mathcal{B} \subseteq \mathbb{R}^{n\times n\times n} : \Psi = \delta_1(Y), Y \in \mathcal{M}_C\}, \tag{3.10}$$

donde $\mathcal{B} = \{\Psi \in \mathbb{R}^{n\times n\times n} : \psi_{ihk} = \psi_{hki} = \psi_{kih} = -\psi_{ikh} = -\psi_{khi} = -\psi_{hki}\}$ y,

$$\mathcal{M}_C = \{Y \in \mathcal{A} \subseteq \mathbb{R}^{n\times n} : \exists l : Y^{(l)} \in \mathcal{M}_G^{(l)}\}, \tag{3.11}$$

donde $\mathcal{A} = \{Y \in \mathbb{R}^{n\times n} : Y^\top = -Y\}$ y,

$$\mathcal{M}_G^{(l)} = \{X \in \mathcal{A}^{(l)} \subseteq \mathbb{R}^{(n-1)\times(n-1)} : X_{ih} = \delta_0(\boldsymbol{s}), \boldsymbol{s} : V^{(l)} \to \mathbb{R}\}, \tag{3.12}$$

siendo $\mathcal{M}_G^{(l)}$ el conjunto de matrices antisimétricas que inducen un ranking en el subconjunto $V^{(l)}$.
Así, una matriz en $\mathcal{M}_G^{(l)}$ induce un orden circular en V según la siguiente regla: $l \leq i_1 \leq \ldots, \leq, i_{n-1} \leq l \Leftrightarrow s_{i_1} \leq, \ldots, \leq s_{i_{n-1}} \Leftrightarrow i_1 \leq, \ldots, \leq i_{n-1}$ con $\{i_1, \ldots, i_{n-1}\} \in V^{(l)}$.

Por tanto, acabamos de ver que cualquier $\Psi \in \mathcal{H}_C$ induce un orden circular. Por otro lado, dado un conjunto de n valores circulares $\{\phi_i\}_{i=1}^n$, se obtiene fácilmente la hipermatriz $\Psi \in \mathcal{H}_C$ de la siguiente forma. Sea $l \subset V$ entonces, para cada $i, h \in V, i, h \neq l$ se define $s_i = \phi_i - \phi_l$ $i \in V$ y $Y_{ih}^{(l)} = s_h - s_i$, $Y_{il} = -\sum_{h\neq l} Y_{ih}$. Por construcción, $Y \in \mathcal{M}_C \subseteq \mathbb{R}^{n\times n}$ y $\Psi = \delta_1(Y) \in \mathcal{H}_C \subseteq \mathbb{R}^{n\times n\times n}$. La expresión de las componentes de Ψ en términos de los valores iniciales viene dada para $l \in V$ de la siguiente forma,

$$\begin{aligned}
&\psi_{ihk} = 0; && i, h, k \in V^{(l)}\\
&\psi_{lih} = \phi_h - \phi_i; && i, h, \in V^{(l)}\\
&\psi_{ihk} = \psi_{hki} = \psi_{kih} = -\psi_{ikh} = -\psi_{khi} = -\psi_{hki}; && i, h, k \in V
\end{aligned}$$

De todo lo anterior, el problema de agregación de órdenes circulares se reduce entonces al siguiente problema de mínimos cuadrados,

$$\widehat{Y} = \arg \min_{\Psi\in\mathcal{H}_C} ||\overline{\Psi} - \Psi||^2,$$

que por la propia definición de $\mathcal{H}_C$ (3.10) es inmediatamente equivalente a,

$$\widehat{Y} = \arg \min_{Y\in\mathcal{M}_C} ||\overline{\Psi} - \delta_1(Y)||^2. \tag{3.13}$$

En el Teorema 3.2, que se introduce más adelante, se obtiene $\widehat{Y}$. En particular, se demuestra que el problema de optimización (3.13), definido en términos del producto interno en $\mathbb{R}^{n\times n\times n}$, se reduce a un problema de optimización en $\mathbb{R}^{n\times n}$ y que la solución a este último problema pasa por resolver en $\mathbb{R}^{n-1\times n-1}$ un

problema muy estudiado de la teoría de Hodge, Teorema 3 y Ecuación (7) de (Jiang et al. (2011)).

Teorema 3.2. *Sea* $\overline{\Psi} \in \mathbb{R}^{n\times n\times n}$ *una hipermatriz antisimétrica,* $\overline{Y} = \frac{1}{n}\delta_1^*(\overline{\Psi})$ *y* $l_0 = \arg\max_l \sum_h \overline{Y}_{lh}^2$*, entonces, si* $\widehat{Y} = \arg\min_{Y\in\mathcal{M}_C} ||\overline{\Psi} - \delta_1(Y)||^2$ *se tiene que:*

(a) $\widehat{Y} = \arg\min_{Y\in\mathcal{M}_C} ||\overline{Y} - Y||^2$,

(b) $\widehat{Y} = \delta_0(\boldsymbol{s})$*, donde:*

$$s_i = -\frac{1}{n-1}\sum_{h\neq l_0} \overline{Y}_{ih} \quad \forall i \neq l_0, \quad \text{con } s_{l_0} = \sum_{i\neq l_0} |s_i|,$$

Demostración

(a) De la equivalencia enunciada en la teoría de Hodge (ver Definición A.7),

$$< \delta_k(f_k), g_{k+1} >_{k+1} = < f_k, \delta_k^*(g_{k+1}) >_k,$$

se tiene por un lado que,

$$\begin{aligned} ||\overline{\Psi} - \delta_1(Y)||^2 &= < \overline{\Psi}, \overline{\Psi} > -2 < \overline{\Psi}, \delta_1(Y) > + < \delta_1(Y), \delta_1(Y) > \\ &= < \overline{\Psi}, \overline{\Psi} > -2 < \tfrac{1}{n}\delta_1^*(\overline{\Psi}), Y > + < Y, \tfrac{1}{n}\delta_1^*(\delta_1(Y)) > \\ &= < \overline{\Psi}, \overline{\Psi} > -2 < \overline{Y}, Y > + < Y, Y > . \end{aligned} \tag{3.14}$$

y por otro lado, se tiene,

$$\begin{aligned} ||\overline{Y} - Y||^2 = \; ||\tfrac{1}{n}\delta_1^*(\overline{\Psi}) - Y||^2 &= < \tfrac{1}{n}\delta_1^*(\overline{\Psi}), \tfrac{1}{n}\delta_1^*(\overline{\Psi}) > -2 < \tfrac{1}{n}\delta_1^*(\overline{\Psi}), Y > \\ &\quad + < Y, Y > \\ &= < \overline{\Psi}, \overline{\Psi} > -2 < \overline{Y}, Y > + < Y, Y >, \end{aligned} \tag{3.15}$$

donde además se usa que $\overline{Y} = \frac{1}{n}\delta_1^*(\overline{\Psi})$.
Entonces, de (3.14) y (3.15), se tiene que
$\widehat{Y} = \arg\min_{Y\in\mathcal{M}_C} ||\overline{\Psi} - \delta_1(Y)||^2 = \arg\min_{Y\in\mathcal{M}_C} ||\overline{Y} - Y||^2$.

(b) De la definición de $\mathcal{M}_C$ (3.11)) y $\mathcal{M}_G^{(l)}$ (3.12), se tiene que (3.13) es equivalente a,

$$\widehat{Y} = \arg\min_{l} \min_{Y^{(l)}\in\mathcal{M}_G^{(l)}} ||\overline{Y}^{(l)} - Y^{(l)}||^2,$$

esta última igualdad de nuevo por la definición de $\mathcal{M}_G^{(l)}$ es equivalente a,

$$\widehat{Y} = \arg\min_{l} \min_{\boldsymbol{v}\in\mathbb{R}^{n-1}} ||\overline{Y}^{(l)} - \delta_0(\boldsymbol{v})||^2.$$

Entonces, por la Ecuación (7) de Jiang et al. (2011) se tiene que,

$$\begin{array}{l}\arg\min_{\boldsymbol{v}\in\mathbb{R}^{n-1}} ||\overline{Y}^{(l)} - \delta_0(\boldsymbol{v})||^2 = \boldsymbol{v}^l \quad \text{donde,}\\ \boldsymbol{v}^l = -\frac{1}{n-1}\delta_0^*(\overline{Y}^{(l)}),\ \text{con } v_i^l = -\frac{1}{n-1}\sum_h \overline{Y}_{ih}^{(l)}, \quad \forall i\in V^{(l)}, l\in V.\end{array}$$

Por otro lado, es inmediato de la propiedad $\sum_i \overline{Y}_{ih} = 0$ que,

$$v_i^l = \frac{1}{n-1}\overline{Y}_{il}, \forall i\in V^{(l)}, l\in V \tag{3.16}$$

Ahora, sea $C = \sum_{ih}\overline{Y}_{ih}^2$, entonces se tiene la siguiente cadena de igualdades,

$$\begin{aligned}
||\overline{Y}^{(l)} - \delta_0(\boldsymbol{v}^l)||^2 &= \sum_{ih\neq l}(v_h^l - v_i^l - \overline{Y}_{ih})^2\\
&= \sum_{ih\neq l}(v_h^l - v_i^l)^2 - 2\sum_{ih\neq l}\overline{Y}_{ih}(v_h^l - v_i^l) + \sum_{ih\neq l}\overline{Y}_{ih}^2\\
&= 2(n-1)\sum_{h\neq l} v_h^{l2} - 2\sum_{h\neq l} v_h^{l2}(\sum_{i\neq l}\overline{Y}_{ih}) + 2\sum_{i\neq l} v_i^l(\sum_{h\neq l}\overline{Y}_{ih}) + \sum_{ih\neq l}\overline{Y}_{ih}^2\\
&= \tfrac{2}{n-1}\sum_{h\neq l}\overline{Y}_{lh}^2 - 2\sum_{h\neq l} v_h^l(-\overline{Y}_{lh}) + 2\sum_{i\neq l} v_i^l(-\overline{Y}_{il}) +\\
&\quad + \sum_{ih}\overline{Y}_{ih}^2 - \sum_h \overline{Y}_{lh}^2 - \sum_i \overline{Y}_{il}^2\\
&= \tfrac{2}{n-1}\sum_{h\neq l}\overline{Y}_{lh}^2 - 2\textstyle\sum_{h\neq l}\frac{1}{n-1}\overline{Y}_{lh}(-\overline{Y}_{lh}) + 2\sum_{i\neq l}\frac{1}{n-1}\overline{Y}_{il}(-\overline{Y}_{il}) +\\
&\quad + C - \sum_h \overline{Y}_{lh}^2 - \sum_i \overline{Y}_{il}^2\\
&= \tfrac{2}{n-1}\sum_{h\neq l}\overline{Y}_{lh}^2 + \tfrac{2}{n-1}\sum_{h\neq l}\overline{Y}_{lh}^2 - \tfrac{2}{n-1}\sum_{i\neq l}\overline{Y}_{il}^2 + C - \sum_h \overline{Y}_{lh}^2 - \sum_i \overline{Y}_{il}^2\\
&= \tfrac{2}{n-1}\sum_{h\neq l}\overline{Y}_{lh}^2 - 2\sum_h \overline{Y}_{lh}^2 + C\\
&= (\tfrac{2}{n-1} - 2)\sum_h \overline{Y}_{lh}^2 + C,
\end{aligned} \tag{3.17}$$

donde la primera igualdad se obtiene usando el operador δ_0, la tercera igualdad viene de que $\sum_{i\in V^{(l)}} v_i^l = 0$, $\forall l \in V$, la cuarta igualdad de la propiedad antisimétrica de la matriz $\overline{Y}$, la quinta igualdad de (3.16) y por ser C una constante que no depende de l, y la última igualdad de nuevo por la propiedad antisimétrica de $\overline{Y}$.

Finalmente, de la equivalencia $(\frac{2}{n-1} - 2) \leq 0 \Leftrightarrow n \geq 2$, a partir de (3.17), se tiene que,

$$\begin{aligned} \arg\min_l ||\overline{Y}^{(l)} - \delta_0(\boldsymbol{v}^l)||^2 &= \arg\min_l (\tfrac{2}{n-1} - 2) \sum_h \overline{Y}_{lh}^2 + C \\ &= \arg\max_l \textstyle\sum_h \overline{Y}_{lh}^2 = l_0, \end{aligned}$$

que es lo que queríamos demostrar.

■

Del Teorema 3.2 se obtiene una matriz $\widehat{Y}$. Esta matriz $\widehat{Y}$ perteneciente a $\mathcal{M}_C$ determina el orden agregado, denominado $\widehat{O}^0$ como sigue.

El ranking en $\mathbb{R}^{n-1}$ asociado a $\widehat{Y}$ viene dado por, $i_1 \leq, \ldots, \leq i_{n-1} \Leftrightarrow s_{i_1} \leq , \ldots, \leq s_{i_{n-1}}$ y entonces el orden circular correspondiente en V viene dado por: $l_0 \leq i_1 \leq \ldots, \leq, i_{n-1} \leq l_0$, que es por construcción el orden que define $\widehat{Y}$ y el orden circular agregado $\widehat{\boldsymbol{O}}^0$.

En la Figura 3.7 mostramos un esquema del procedimiento completo de esta técnica basada en la teoría de Hodge.

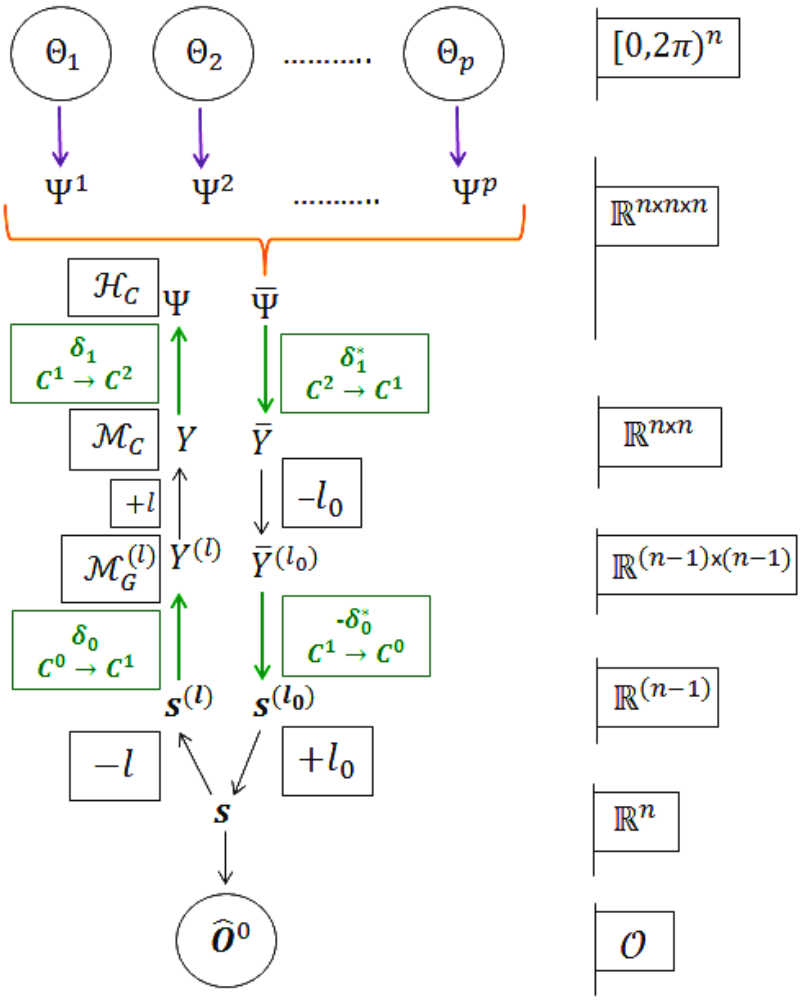


Figura 3.7: Esquema del procedimiento de agregación de órdenes circulares usando la técnica basada en la teoría de Hodge

3.5.2. Definición de λ_{ihk}^j

Como ya se ha comentado previamente, una de las mayores ventajas de esta técnica es la flexibilidad en la definición de la intensidad de las tripletas Así, λ_{ihk}^j debe tomar valores máximos cuando $(\theta_{ij}, \theta_{hj}, \theta_{kj})$ estén equiespaciados en el círculo, debe tomar valores intermedios en caso de que θ_{ij}, θ_{hj} estén muy juntos pero θ_{kj} muy alejado de ambos y debe tomar valores pe-

queños cuando $(\theta_{ij}, \theta_{hj}, \theta_{kj})$ muy cercanos entre sí. Estas propiedades implican que la definición de λ^j_{ihk} no sea sencilla. Por ejemplo, una definición que hace uso de la distancia angular mas utilizada en la literatura sería, $\lambda^j_{ihk} = (1 - \cos(\theta_{hj} - \theta_{ij})) + (1 - \cos(\theta_{kj} - \theta_{hj})) + (1 - \cos(\theta_{ij} - \theta_{kj}))$, pero esta definición de λ^j_{ihk} no cumple las propiedades anteriores.

Se definen distintas propuestas para λ^j_{ihk} en la Tabla 3.6. Estas definiciones cubren una amplia variedad de opciones que usan información ordinal o cardinal, diferentes descriptivos circulares así como distancias.

Tabla 3.6: Métodos basados en la técnica de Hodge según las definiciones de λ^j_{ihk}

Etiqueta	$\lambda^j_{ihk}\ \forall i, h, k \in V, j = 1, \ldots, p.$
CHsig	$1\ \forall i, h, k, j$
CHpos	$\lvert(F^j_{ih} - D^j_{ih}) + (F^j_{hk} - D^j_{hk}) + (F^j_{ki} - D^j_{ki})\rvert$
CHcos	$(1 + \cos(\theta_{hj} - \theta_{ij})) + (1 + \cos(\theta_{kj} - \theta_{hj})) + (1 + \cos(\theta_{ij} - \theta_{kj}))$
CHcmean	$(\cos(\theta_{ij} - \overline{\theta}^j_{ihk})) + (\cos(\theta_{hj} - \overline{\theta}^j_{ihk})) + (\cos(\theta_{kj} - \overline{\theta}^j_{ihk}))$
CHmrl	$\overline{R}^j_{ihk}$
CHe3	$E^j_{ih} + E^j_{ik} + E^j_{hk} + E^j_{hi} + E^j_{ki} + E^j_{kh} =$ $= d_3(\theta_{ij}, \theta_{hj}) + d_3(\theta_{ij}, \theta_{kj}) + d_3(\theta_{hj}, \theta_{kj}) +$ $+ d_3(\theta_{hj}, \theta_{ij}) + d_3(\theta_{kj}, \theta_{ij}) + d_3(\theta_{kj}, \theta_{hj})$
CHave	$Ave(arc(\theta_{ij}, \theta_{hj}), arc(\theta_{hj}, \theta_{kj}), arc(\theta_{kj}, \theta_{ij}))$

En particular, **CHsig** usa únicamente el signo de la tripleta. **CHpos** utiliza las posiciones mediante los valores definidos en (3.6) y usados en las cadenas de

Markov. La definición **CHcos** es una modificación de la distancia más habitual en el círculo de manera que cumpla las características que este procedimiento requiere. En la definición **CHcmean**, se tienen en cuenta las diferencias entre cada elemento y la media de la tripleta donde $\overline{\theta}^{j}_{ihk}$ es la media circular de $(\theta_{ij}, \theta_{hj}, \theta_{kj})$. En la definición **CHmrl** se hace uso de la longitud media resultante (Definición 2.3) para medir la intensidad de cada tripleta. En la definición **CHe3** se ha hecho uso de la distancia definida para el TSP de la longitud de las aristas entre elementos $d_3(\theta_{ij}, \theta_{hj})$, es decir la del método **TSP3** (ver Tabla 3.4). En la definición **CHave** se hace uso de los caminos entre los tres elementos a través de su media circular denotada por *Ave* y siendo cada camino el arco mínimo denotado por *arc*.

3.6. Algoritmo de búsqueda local: CLM

El algoritmo que presentamos en esta sección tiene como objetivo mejorar la solución $\widehat{\boldsymbol{O}}^{0}$ dada por el primer paso del procedimiento, en términos de la función objetivo del problema de optimización (3.4). La propuesta que hacemos es un algoritmo de búsqueda local. Este tipo de algoritmos son una de las estrategias más usadas para mejorar la aproximación a la solución en problemas de optimización del tipo NP-hard. Se caracterizan por comparar soluciones vecinas a una dada con el objetivo de mejorar el valor de la función objetivo.

El algoritmo denominado *Circular Local Minimization* (CLM) que presentamos aquí está inspirado en el conocido método en la línea llamado *local Kemenization* (Dwork et al. (2001a)). CLM realiza mejoras locales del orden circular $\widehat{\boldsymbol{O}}^{0}$, comprobando para cada tripleta de elementos consecutivos si existe una permutación que mejore la función objetivo (3.4), que en nuestro caso es $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}})$. Por tanto el algoritmo buscará aquellos órdenes circulares que cumplan $MSCE(\widehat{\boldsymbol{O}}, \boldsymbol{\Theta}) < MSCE(\widehat{\boldsymbol{O}}^{0}, \boldsymbol{\Theta})$. Su funcionamiento se muestra paso a paso en el Algoritmo 3. El orden circular así obtenido se denomina $\widehat{\boldsymbol{O}}$.

Algoritmo 3: CLM: Circular Local Minimization

entrada: $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_p$; $\widehat{\boldsymbol{O}}^0$.

salida : $\widehat{\boldsymbol{O}}$.

1 **for** *i desde 1 hasta n* **do**

2 Se toma la tripleta de elementos consecutivos (i,h,k) donde h=i+1; k=i+2;

3 Se tiene $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^1)$: la distancia entre el orden circular usando (i,h,k) con los datos;

4 Se calcula $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^2)$: la distancia entre el orden circular correspondiente a la tripleta permutada (i,k,h) con los datos;

5 **if** $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^2) < MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^1)$ **then**

6 Se permuta la tripleta $\rightarrow$ (i,k,h);

7 **if** $i > 1$ **then**

8 Se comprueba la tripleta inmediatamente anterior (i-1,i,k) actualizando los valores de $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^2)$ y $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^1)$;

9 **while** $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^2) < MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^1)$ **do**

10 Se permuta la tripleta quedándonos con $\widehat{\boldsymbol{O}}^2$;

11 **if** *primer elemento de la tripleta >1* **then**

12 Se comprueba la tripleta inmediatamente anterior actualizando los valores de $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^2)$ y $MSCE(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}^1)$;

13 **else**

14 Continuar;

15 **return** $\widehat{\boldsymbol{O}}$;

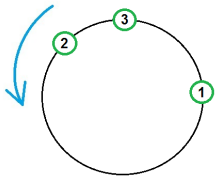
3.7. Ejemplos ilustrativos del funcionamiento de los métodos

En esta sección se presentan tres ejemplos que ilustran los métodos propuestos en este capítulo para obtener $\widehat{\boldsymbol{O}}^0$ (ver Tabla 3.7). El Ejemplo 3.2 es un caso muy sencillo con 3 elementos y 2 experimentos que pone de manifiesto que técnicas tan sencillas como Borda Circular y cadenas de Markov no funcionan incluso en casos muy simples. En el Ejemplo 3.3 se ilustra la importancia de tener en cuenta la dirección de rotación al definir las relaciones entre pares de elementos. Y el Ejemplo 3.4 refleja una situación real donde se pone de manifiesto que algunos métodos no dan soluciones coherentes.

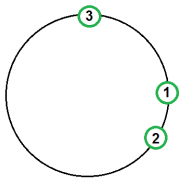
Tabla 3.7: Etiquetas de los métodos de agregación de órdenes circulares

Etiqueta	Significado
Naive	Método Naive
BCpos	Método Borda Circular Posiciones
BCmean	Método Borda Circular Medias
BCmed	Método Borda Circular Medianas
CMC1	Cadenas de Markov Circular Opción 1
CMC2	Cadenas de Markov Circular Opción 2
CMC3	Cadenas de Markov Circular Opción 3
CMC4maj	Cadenas de Markov Circular Opción 4 por mayoría
CMC4num	Cadenas de Markov Circular Opción 4 números
TSPbin	Método TSP ceros y unos
TSPpos	Método TSP posiciones
TSP1	Método TSP $\alpha = 1$
TSP2	Método TSP $\alpha = 2$
TSP3	Método TSP $\alpha = 3$
TSP4	Método TSP $\alpha = 4$
TSPinf	Método TSP $\alpha = \infty$
TSPtime	Método TSP tiempos
TSParc	Método TSP Arco
TSPcho	Método TSP Cuerda
CHsig	Método Hodge en el Círculo. Signos
CHpos	Método Hodge en el Círculo. Posiciones
CHcos	Método Hodge en el Círculo. Cosenos
CHcmean	Método Hodge en el Círculo. Cosenos con Medias
CHmrl	Método Hodge en el Círculo. Longitud media resultante
CHe3	Método Hodge en el Círculo. Basado en TSP3
CHave	Método Hodge en el Círculo. Medias

Ejemplo 3.2. En este caso tenemos dos experimentos con tres elementos cada uno y con el mismo orden. Es obvio que un método razonable debería encontrar dicho orden común a los dos. Las observaciones están representadas en la Figura 3.8 y recogidas en la Tabla 3.8.



(a) Experimento 1



(b) Experimento 2

Figura 3.8: Representación de los datos del Ejemplo 3.2

Tabla 3.8: Datos del Ejemplo 3.2

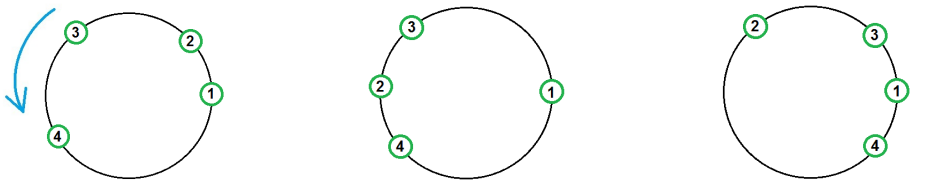
	Elemento 1	Elemento 2	Elemento 3
Experimento 1	0	$3\frac{\pi}{4}$	$\frac{\pi}{2}$
Experimento 2	0	$11\frac{\pi}{6}$	$\frac{\pi}{2}$

Se observan los resultados de la agregación de estos dos experimentos en la Tabla 3.9. Sorprendentemente hay algunos métodos como **BCmean, BCmed** y los basados en cadenas de Markov que no encuentran el orden circular común a los dos experimentos.

Tabla 3.9: Resultados del Ejemplo 3.2

Orden circular	MSCE	Métodos
(1,3,2) *(Óptimo)*	0	**Naive, BCpos, TSPbin, TSPpos, TSParc, TSPcho, TSP1, TSP2, TSP3, TSP4, TSPinf, TSPtime, CHsig, CHpos, CHcos, CHcmean, CHmrl, CHe3, CHave.**
(1,2,3)	0.0367	**BCmean, BCmed, CMC1, CMC2, CMC3, CMC4maj, CMC4num.**

Ejemplo 3.3. En este caso tenemos 3 experimentos con 4 elementos cada uno. Las observaciones se muestran en la Figura 3.9 y en la Tabla 3.10. Es un caso sencillo donde no todos los experimentos tienen el mismo orden pero donde hay un orden óptimo utilizando la función objetivo (3.4) que es (1,3,2,4) que además es el orden de 2 de los 3 experimentos.



(a) Experimento 1 (b) Experimento 2 (c) Experimento 3

Figura 3.9: Representación de los datos del Ejemplo 3.3

Tabla 3.10: Datos del Ejemplo 3.3

	Elemento 1	Elemento 2	Elemento 3	Elemento 4
Experimento 1	0	$\frac{\pi}{4}$	$3\frac{\pi}{4}$	$5\frac{\pi}{4}$
Experimento 2	0	π	$3\frac{\pi}{4}$	$5\frac{\pi}{4}$
Experimento 3	0	$3\frac{\pi}{4}$	$\frac{\pi}{4}$	$7\frac{\pi}{4}$

Los resultados de aplicar los diferentes métodos en el conjunto de datos de la Tabla 3.10 aparecen en la Tabla 3.11. En esta tabla se observa que el óptimo es alcanzado por aquellos métodos que tienen en cuenta la dirección del círculo independientemente de la técnica que se use. Los métodos basados en cadenas de Markov y aquellos que dentro de la técnica del TSP hacen uso de distancias simétricas dan lugar a soluciones alejadas del óptimo.

Tabla 3.11: Resultados Ejemplo 3.3

Orden circular	MSCE	Métodos
(1, 3, 2, 4) *(Óptimo)*	0.0488	**TSP2, TSP3, TSP4**, **TSPinf, TSPtime** **Naive**, **BCpos, BCmean, BCmed** **CHsig, CHpos, CHcos, CHcmean** **CHmrl, CHe3, CHave**
(1, 2, 3, 4)	0.0615	**CMC1** **TSP1**
(1, 4, 3, 2)	0.1230	**TSPcho, CMC2**
(1, 3, 4, 2)	0.1310	**CMC3, CMC4maj, CMC4num**

Ejemplo 3.4. En este caso tenemos dos experimentos con información sobre 5 elementos cuyas observaciones están representadas en la Figura 3.10 y recogidas en la Tabla 3.12.

(a) Experimento 1 (b) Experimento 2

Figura 3.10: Representación de los datos del ejemplo 3.4

Tabla 3.12: Datos del Ejemplo 3.4

	Elemento 1	Elemento 2	Elemento 3	Elemento 4	Elemento 5
Experimento 1	0	2	2,6	5	5,2
Experimento 2	6,2	2,3	1,6	4,8	0

Este ejemplo puede corresponder en nuestra aplicación, donde las observaciones son momentos de máxima expresión de genes, a dos experimentos en los

que se miden 5 genes que se activan, por grupos, en dos fases diferentes: por un lado los genes 1,4,5 y por otro los genes 2 y 3. La variabilidad experimental provoca que en los dos experimentos se observen los patrones de órdenes tan diferentes como los que ilustra la Figura 3.10; sin embargo, una solución, biológicamente razonable, de un orden agregado no debería mezclar los dos grupos.

Ejecutamos todos los métodos expuestos en este capítulo y los resultados obtenidos se muestran en la Tabla 3.13 donde los órdenes circulares agregados resultados de los distintos métodos están ordenados de menor a mayor MSCE. Las ultimas cuatro filas corresponden con resultados de órdenes que mezclan genes de los dos grupos. Todos los métodos que hacen uso de la técnica basada en Hodge dan soluciones interpretables. Cabe destacar el mal comportamiento que encontramos en el caso de **TSPinf** y también de la mayoría de los métodos basados en cadenas de Markov.

Tabla 3.13: Resultados del Ejemplo 3.4

Orden Circular	MSCE	Métodos
(1,3,2,4,5)	0.0091	**CHave**
(Óptimo)		**TSP2, TSP3, TSP4, BCmean, BCmed**
(1,2,3,4,5)	0.0123	**CHmrl, CHmean, CHcos**
		BCpos, Naive
(1,5,3,2,4)	0.0375	**CHe3**
(1,5,2,3,4)	0.0407	**CHpos, CHsig**
(1,2,3,5,4)	0.0656	**CMC1**
(1,4,2,3,5)	0.1172	**CMC4maj**
(1,5,4,2,3)	0.1231	**TSP1**
(1,2,5,4,3)	0.2648	**TSPinf**
(1,2,4,3,5)	0.2105	**CMC2**
(1,3,5,4,2)	0.2714	**CMC3**
(1,4,2,5,3)	0.3516	**CMC4num**

Esta sección podría haberse extendido con muchos más ejemplos pero por falta de espacio hemos expuesto los más representativos y que mostraban con más claridad las diferencias entre las técnicas desarrolladas.

3.8. Estudio de simulación

En esta sección compararemos los resultados de los diferentes métodos propuestos mediante la generación de datos aleatorios. Utilizaremos la distribución de von Mises así, supondremos $\theta_{ij} \sim M(\phi_{ij}, \kappa_j)$ $i = 1, \ldots, n, j = 1, \ldots, p$, independientes. Con la finalidad de generar diversas situaciones de interés tanto teórico como práctico se considerarán casos artificiales y reales. Los diferentes escenarios se describen a partir de los valores de los parámetros en la Tabla 3.14.

Tabla 3.14: Configuraciones de los parámetros para las simulaciones

Parámetros		
n	5, 7, 11, 34	
p	4, 6, 10, 20	
$\mathbf{\Phi}$	$\phi_{ij} = i\frac{2\pi}{n}$	$\forall j \in S_1, i = 1, \ldots, n$
	$\phi_{ij} = i\frac{\pi/4}{n}$	$\forall j \in S_2, i = 1, \ldots, n$
	$\phi_{ij} = (i\frac{\pi/8}{n}) + \pi$	$\forall j \in S_3, i = 1, \ldots, n$
	datos reales	
κ	0, 1, 2, 10	
	estimados, bajos, medios, altos	

donde S_1, S_2, S_3 son grupos de diferentes tamaños dependiendo en cada caso de p. Las medias circulares procedentes de datos reales se tomaron de experimentos de las levaduras *S. pombe* y *S. cerevisiae* (ver Tabla B.5) y se han escogido de forma que estén representados diferentes tipos de patrones.

En el caso del parámetro de concentración κ_j, los escenarios artificiales usan

valores $\kappa_j = \kappa_{j'}\ \forall j, j'$ con variaciones entre 0 y 10. Por otro lado, en los escenarios reales se han usado valores obtenidos de los experimentos considerados en el Capítulo 5 para los humanos y las dos levaduras *S. pombe* y *S. cerevisiae*, en algunos casos agrupados según observamos en la Tabla 3.15.

Tabla 3.15: Grupos de valores de κ según número de experimentos

	1	2	3	4	5	6	7	8	9	10
Bajos	0.628	0.8534	1.1104	1.3764	1.4284	1.5954	1.6044	1.6843	2.2952	2.5627
Medios	2.7225	3.599	4.7507	6.9018						
	2.7225	3.599	4.7507	6.9018	8.9412	9.1249				
	1.6044	1.6843	2.2952	2.5627	2.7225	3.599	4.7507	6.9018	8.9412	9.1249
Altos	19.1917	23.8777	26.2811	26.558						

De todas las posibles combinaciones se han ejecutado un total de 215 escenarios. En cada escenario se ejecutaron 1000 repeticiones.
La evaluación se llevará a cabo mediante dos criterios, el MSCE que corresponde con el valor de la función objetivo a minimizar (3.4) y el tiempo de ejecución de cada método medido en segundos. Es habitual en problemas de optimización de este tipo que no exista un método óptimo con ambos criterios como ocurre en este caso. Tampoco encontraremos un método que sea ganador según uno de los criterios en todos los escenarios como veremos a continuación. Además, el análisis de las simulaciones no es sencillo por la diversidad de escenarios y métodos e incluso porque valoramos la bondad de los métodos con diferentes criterios. Para organizar este análisis hemos dividido esta sección en tres partes, en la Sección 3.8.1 se comparan todos los métodos mediante la ejecución del paso 1 del procedimiento en una selección de escenarios representativos utilizando el MSCE medio y el tiempo medio para las 1000 simulaciones. Con los resultados de este primer análisis se eliminan un conjunto de métodos que sistemáticamente funcionan peor. En la Sección 3.8.2 se analiza más detalladamente el comportamiento de los métodos seleccionados considerando ya el

segundo paso del procedimiento (el algoritmo CLM, ver Sección 3.6). De estos análisis se puede concluir como veremos que la técnica TSP es claramente ganadora y por tanto se dedica la Sección 3.8.3 específicamente a analizar resultados únicamente de esta técnica.

3.8.1. Comparaciones globales de los métodos

El objetivo de esta primera comparación es realizar una selección de aquellos métodos que funcionen mejor. Además, de los 215 escenarios se muestra únicamente una selección, aquellos con $n = 5$ y valores medios del parámetro de concentración κ, que es representativa en cuanto a que las conclusiones serían equivalentes a la gran mayoría del resto de escenarios. La Figura 3.11 muestra los valores del MSCE medio y el tiempo medio para estos casos, y los diferentes métodos, mediante un diagrama de dispersión.

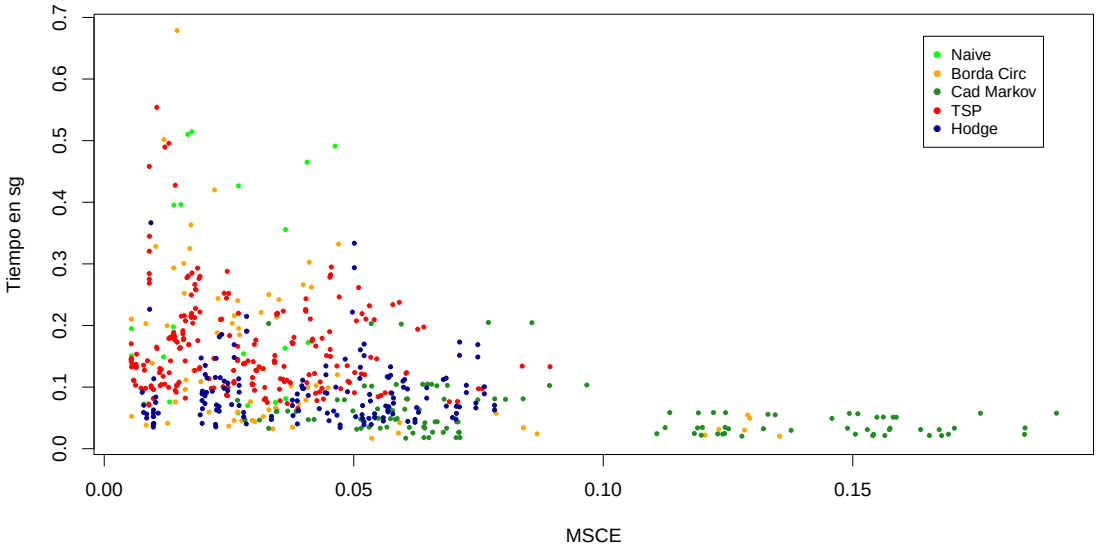


Figura 3.11: MSCE y tiempo de ejecución. Escenarios con n=5 y κ medio

Este gráfico sirve para detectar aquellos métodos que están funcionando sistemáticamente mal. Se puede apreciar como los valores más altos del MSCE se alcanzan mediante el uso de algunos métodos basados en cadenas de Markov y Borda circular. En relación al tiempo de ejecución los métodos que peor funcionan son algunas opciones de Borda Circular, el Naive y también algunos de los basados en el TSP.

Con la finalidad de realizar un análisis más detallado se han realizado diagramas de cajas tanto para el MSCE como para los tiempos en cada escenario y así observar la variabilidad de los resultados según cada método. Hemos comprobado que el patrón de comportamiento es similar en todos escenarios por lo que mostramos aquí un escenario representativo con valores para los parámetros: $n = 11$, $p = 6$ y valores de κ altos (Figura 3.12).

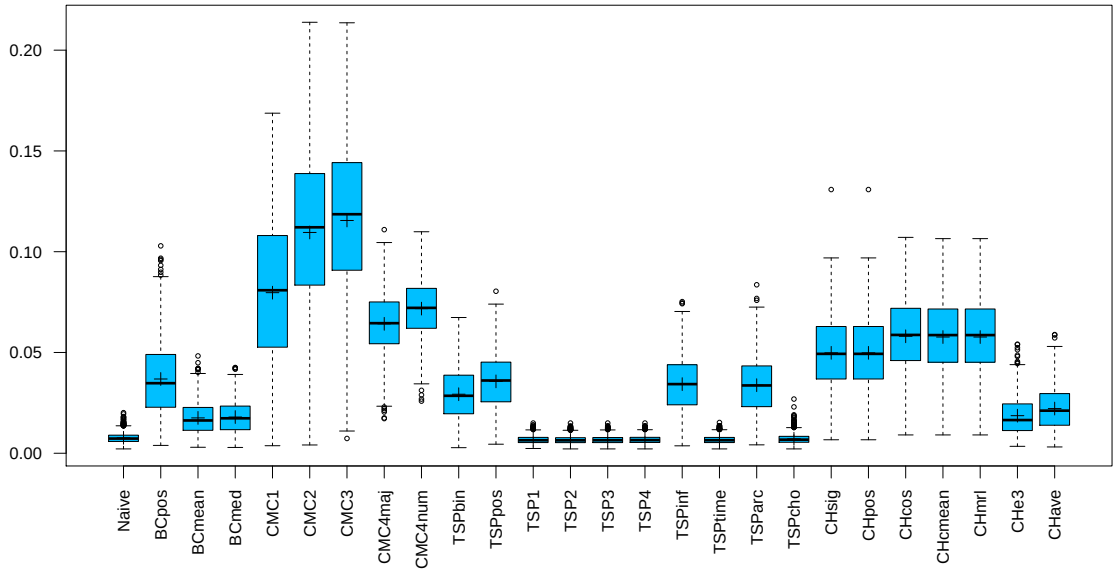


Figura 3.12: Diagramas de cajas de los valores del MSCE obtenidos en el escenario con parámetros n=11 p=6 y kappas altos

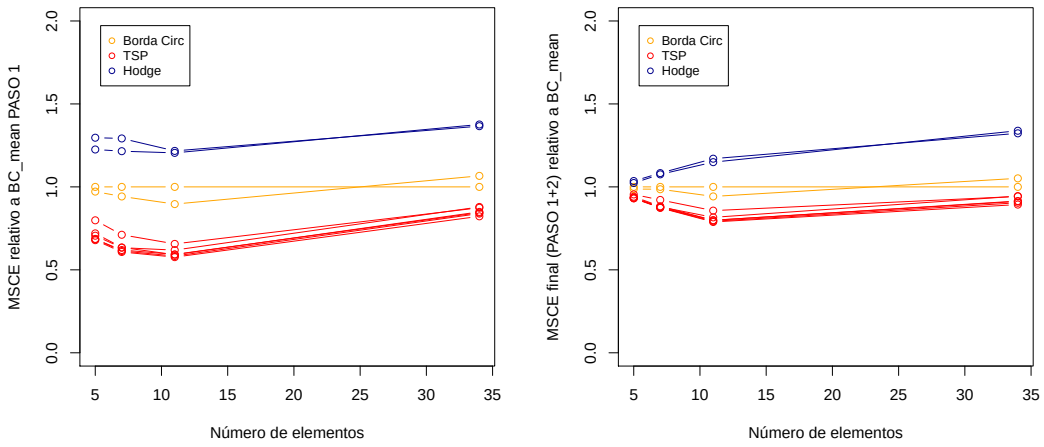
Como conclusiones de este análisis podemos hacer una clasificación de los métodos en cuatro grupos. En un primer grupo consideramos los métodos que peor se comportan en cualquier situación, son aquellos que usan la técnica ba-

sada en cadenas de Markov. El segundo grupo en el que se tienen métodos que proceden de diversas técnicas: **BCpos**, **TSPbin**, **TSPpos**, **TSPinf**, **TSParc**, **CHsig**, **CHpos**, **CHcos**, **CHcmean** y **CHmrl** y **Naive**, se caracteriza por obtener resultados poco favorables en la mayoría de las situaciones. Estos dos grupos son los que descartamos para futuros análisis ya que ofrecen peores resultados en la mayoría de los casos. Respecto al método **Naive** decir que, como ya esperábamos, funciona bien cuanto menor es el número de elementos y mayor el número de experimentos. Además, en casos con valores reales hemos obtenido resultados muy poco coherentes y el tiempo de ejecución es bastante alto en la mayoría de los escenarios.

Por otra parte, los dos siguientes grupos de métodos serán estudiados posteriormente con más detalle. El tercer grupo compuesto por los métodos **BCmean**, **BCmed**, **CHe3** y **CHave** se caracteriza por ofrecer buenas aproximaciones o soluciones biológicamente interpretables en casi todos los casos, en un tiempo muy razonable, aunque no suelen alcanzar la mejor solución en términos de MSCE. En un segundo grupo se encuentran los métodos **TSP1**, **TSP2**, **TSP3**, **TSP4**, **TSPtime** y **TSPcho**, que tienen un comportamiento mejor que el resto en casi todos los escenarios.

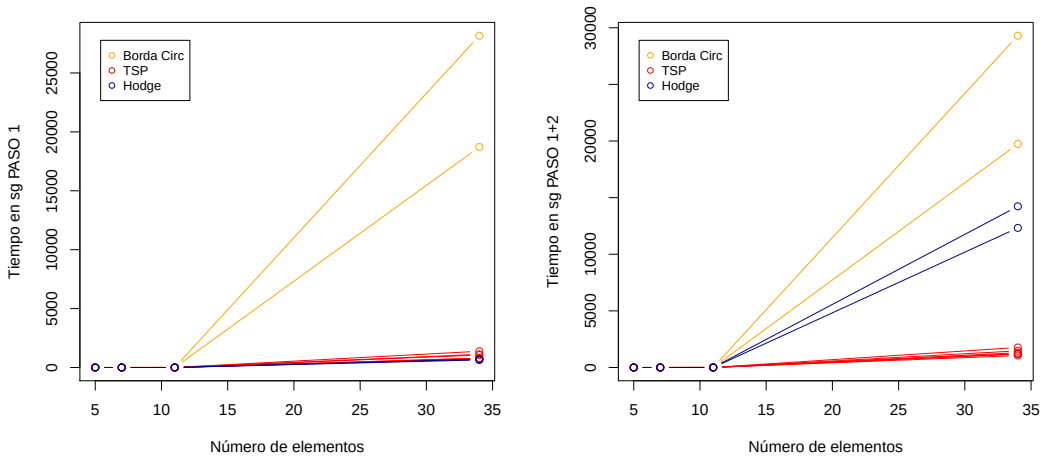
3.8.2. Comparaciones entre métodos seleccionados

En este apartado presentamos un resumen de los análisis de las simulaciones realizadas para el procedimiento completo de búsqueda de un orden circular propuesto en dos pasos. Los métodos seleccionados que se estudiarán con más detalle en este apartado son los siguientes: **BCmean**, **BCmed**, **TSP1**, **TSP2**, **TSP3**, **TSP4**, **TSPtime**, **TSPcho**, **CHe3** y **CHave**.
Para este análisis consideramos únicamente 4 escenarios con $n = 5, 7, 11$ y 34 donde se tiene $p = 6$ y los valores de κ son aquellos que aparecen en la Tabla 3.15 como valores medios. Los patrones que se usan para las medias circulares son los descritos en la Tabla 3.14 teniendo $\#S_1 = \#S_2 = \#S_3 = 2$.



(a) Paso 1 del procedimiento de agregación (b) Paso 1+2 del procedimiento

Figura 3.13: Incremento del MSCE con n



(a) Paso 1 del procedimiento de agregación (b) Paso 1+2 del procedimiento

Figura 3.14: Incremento del tiempo de ejecución con n

En la Figura 3.13 se representan los valores del MSCE relativos al obtenido con el método **BCmean** para todos los métodos y para cada valor de n

diferenciando el paso 1 del paso 1+2. En ambos casos, las mejores soluciones vienen dadas por métodos que hacen uso de la técnica basada en el TSP.
En la Figura 3.14 se tienen representados los tiempos medios para todos los métodos en segundos y diferenciando los dos pasos. En el primer paso, el incremento del tiempo al aumentar n es más acentuado en los métodos de la técnica Borda circular mientras que en el paso 1+2 en los de la técnica de Hodge. El tiempo de ejecución de los métodos basados en el TSP es razonable tanto en el primer paso como después de ejecutar el procedimiento completo.

Se concluye, al igual que en el apartado anterior, que los métodos TSP tienen comportamientos mejores tanto en términos de MSCE como de tiempos de ejecución. Y además ese comportamiento se mantiene aunque el número de elementos aumente.

3.8.3. Análisis de la técnica TSP

En este apartado se realiza una comparativa entre los métodos que mejores resultados ofrecen según los análisis anteriores: **TSP1**, **TSP2**, **TSP3**, **TSP4**, **TSPtime** y **TSPcho**. Además se consideran alternativas en el Algoritmo 2 de optimización. En particular, consideramos el efecto del valor del parámetro c, donde $c \cdot n$ es el número máximo de rutas distintas que se validan para escoger una solución final. Se comparan los resultados haciendo uso de los valores $c = 2, 1, 1/2, 1/4, 1/8, 0,$ (ver Tabla 3.16).

Tabla 3.16: Etiquetas según los valores de c

$c =$	2	1	1/2	1/4	1/8	0
Etiqueta	_68	_34	_17	_8	_4	_1

Para los análisis de esta sección se consideran dos escenarios (A y B) con $n = 34$ y $p = 6$ en ambos casos que se diferencian en los valores de las medias circulares $\mathbf{\Phi}$ y el parámetro de concentración κ. El escenario A es el descrito en el apartado anterior con valores del vector de medias $\mathbf{\Phi}$ definidos en la

Tabla 3.14 con $\#S_1 = \#S_2 = \#S_3 = 2$ y valores de κ medio (Tabla 3.15). El escenario B tiene como valores de $\mathbf{\Phi}$ el CIRE (2.8) calculado con los datos originales observados en la levadura *S.cerevisiae* para el orden agregado de todos los experimentos (Apéndice B). Los valores de κ varían entre 1.6 y 23.88. Se trata de uno de los conjuntos más grandes de datos que se maneja en la aplicación y lo consideramos representativo de *n grande* para este problema.

Como resumen del análisis de los resultados se han realizado diagramas de cajas. Las conclusiones son similares en los dos escenarios, para simplificar mostramos sólo los resultados del escenario A en las Figuras 3.15 y 3.16 (en verde los resultados del paso 1 y en azul los del paso 1+2).

En el gráfico de la Figura 3.15 se puede observar el aumento del MSCE al disminuir el coeficiente c sea cual sea el método utilizado. La comparativa de los valores del MSCE para $c = 0$ permite hacer una comparación cruda del efecto de usar cada una de las distancias. En este sentido parece que las opciones que mejor funcionan en media para este escenario son **TSP1, TSP2, TSP3** y **TSPtime**. En el escenario B, cuyos gráficos no se muestran aquí, las conclusiones serían similares, los que mejor funcionan con $c = 0$ son **TSP2, TSP3, TSP4** y **TSPtime**. En cuanto al criterio del tiempo, en los diagramas de cajas de la Figura 3.16 se observa, como era de esperar, una disminución evidente del tiempo en segundos a medida que disminuye el valor de c, llegando incluso a décimas de segundo para $c = 0$.

Con el fin de evaluar la relación entre el MSCE medio y los tiempos medios de cada método se realizan los diagramas de dispersión que se muestran en la Figura 3.17 para el escenario A y en la Figura 3.18 para el escenario B. En ambas representaciones se tienen los valores del MSCE relativos de todos los métodos al obtenido con el método **TSP3** con $c = 2$ (**TSP3_68**) y la media de los segundos de ejecución, en ambos criterios se ha usado el resultado final después de los dos pasos propuestos para la de búsqueda del orden circular.

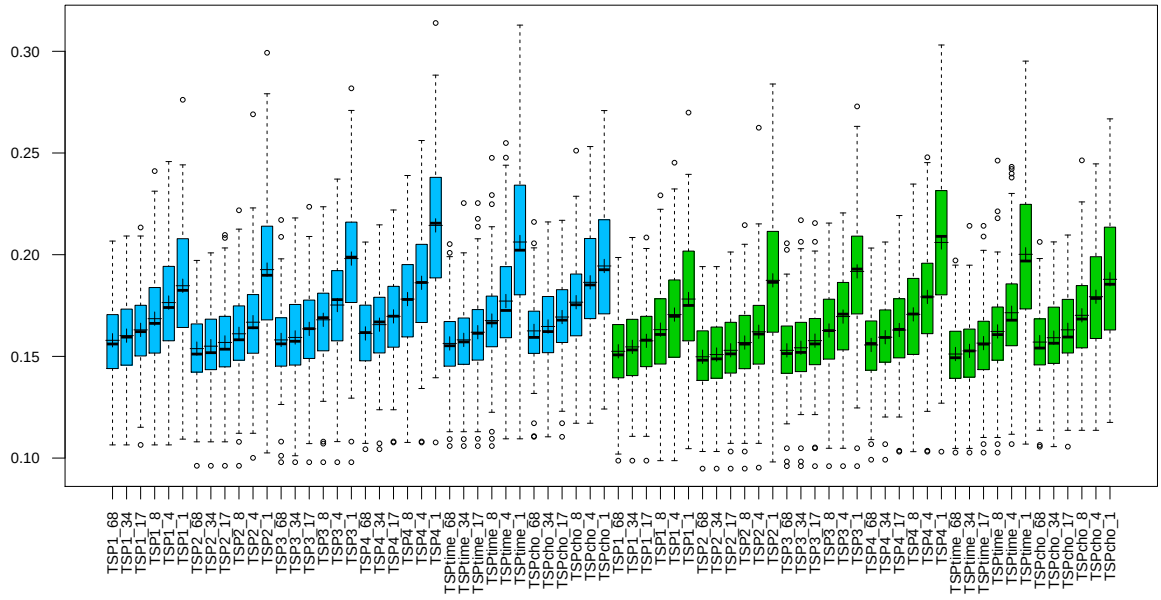


Figura 3.15: Diagramas de cajas para el MSCE. Escenario A.

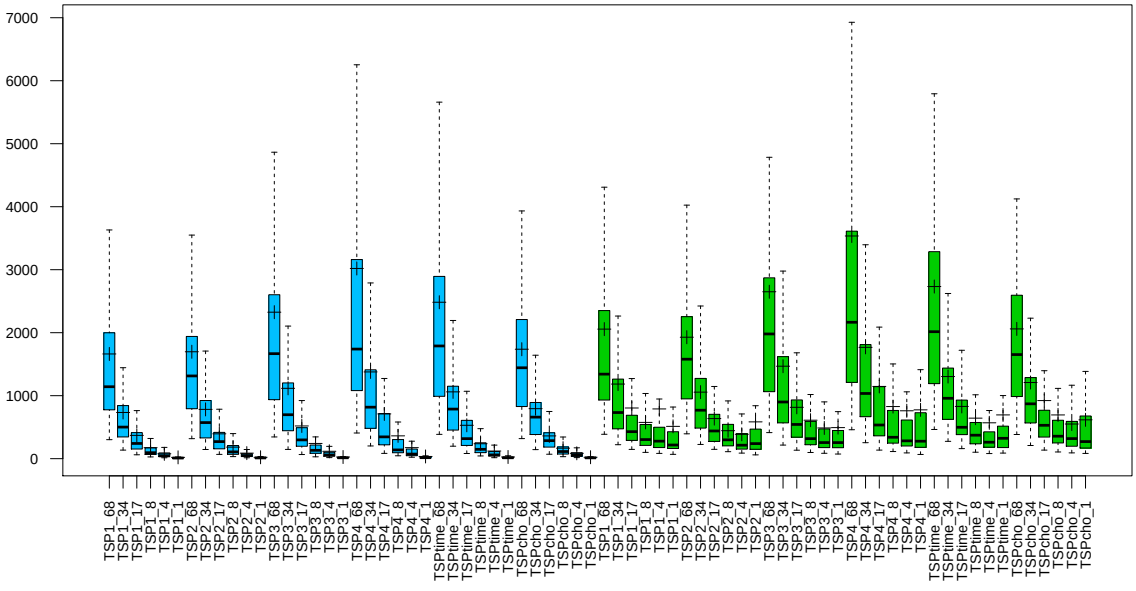


Figura 3.16: Diagramas de cajas para los tiempos de ejecución. Escenario A.

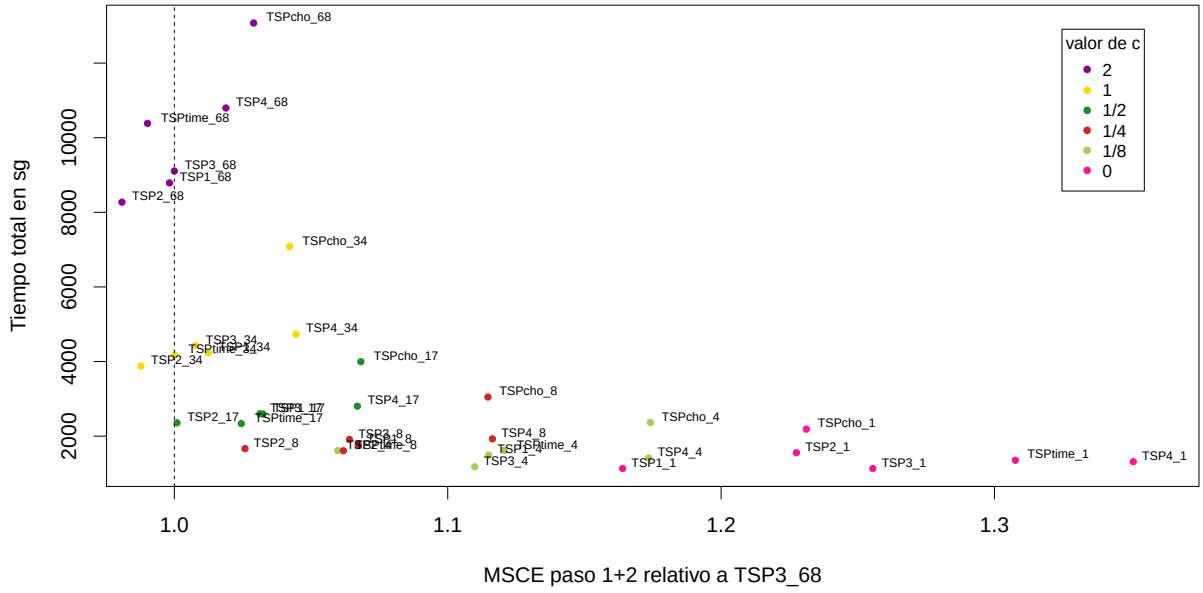


Figura 3.17: Escenario A. Relación MSCE medio y tiempo medio

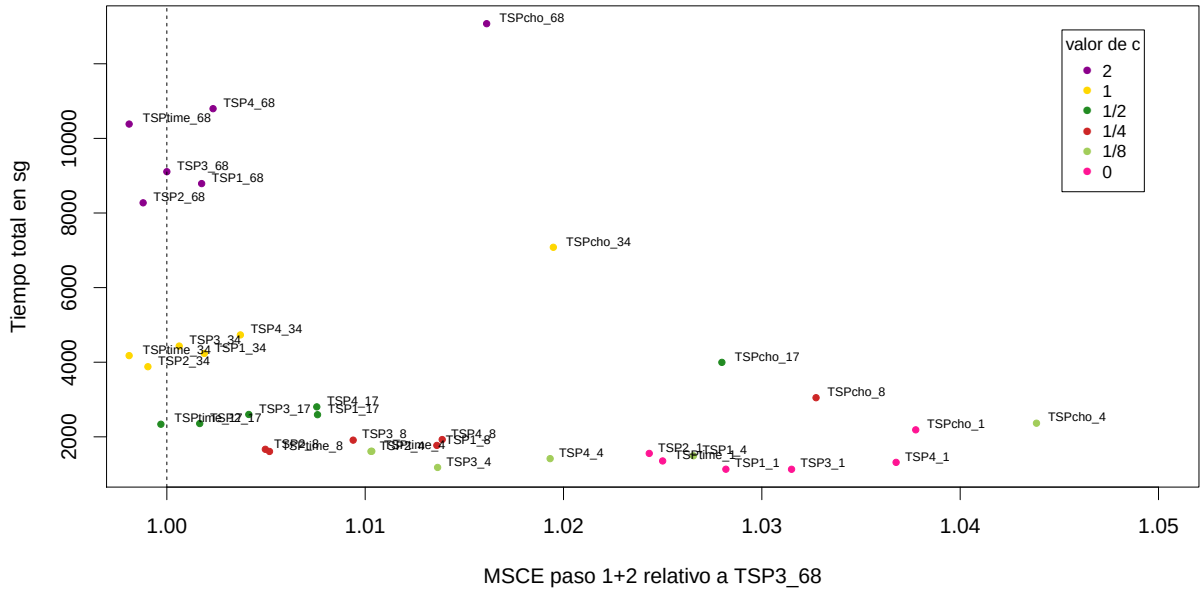


Figura 3.18: Escenario B. Relación MSCE medio y tiempo medio

Como conclusiones del gráfico correspondiente al escenario A cabe destacar que el método **TSP2** se encuentra en la frontera de interés (menor MSCE y menor tiempo de ejecución) para todos los valores de c. Los siguientes mejores métodos son **TSP3** y **TSPtime**. El método que peor comportamiento tiene es **TSPcho** salvo en el caso de $c = 0$ que se trata de **TSP4**.
En el gráfico que representa los resultados del escenario B se tienen como métodos con mejor comportamiento también **TSP2**, **TSP3** y **TSPtime**, aunque en este caso siendo ligeramente mejor este último. El peor comportamiento viene dado por **TSPcho** para todos los valores de c, seguido por **TSP4**.

Resumiendo, **TSP2, TSP3** y **TSPtime** ofrecen las soluciones con una mejor relación entre los dos criterios estudiados, aunque se ha observado que esta última en algunos ejemplos numéricos (Sección 3.7) ofrecía una solución no interpretable biológicamente. En cuanto a la elección del coeficiente c, podríamos recomendar hacer uso de $c = 1/4$ para asegurar simultaneamente una solución razonable en términos de MSCE y eficiencia computacional. Sin embargo, si el objetivo fuera obtener un resultado razonable en un tiempo muy corto se propone usar $c = 0$ o $c = 1/8$.

3.9. Conclusiones

A partir del estudio de las propiedades de las diferentes técnicas, de los resultados de las simulaciones, y del comportamiento en ejemplos y casos reales hemos elaborado una lista de fortalezas (⏡) y debilidades (⏠) para cada técnica. En esta lista hemos tenido en cuenta también las posibilidades de desarrollo y mejora en el futuro o la flexibilidad de intervención del usuario.

- **Método Naive**
  ⏡ Es un método sencillo e intuitivo. Funciona bien para aquellos casos con pocos elementos en relación al número experimentos y valores de κ medios o altos.

☹ En determinados escenarios llega a soluciones muy alejadas del optimo. Los tiempos de computación pueden ser muy elevados. Se han encontrado casos en los que las soluciones no son interpretables. En concreto, en la aplicación que se presenta en el Capítulo 5 el resultado de aplicar este método da una solución que no es biológicamente interpretable.

- **Técnica Borda Circular**

  ☺ Es sencilla en su construcción e interpretación.

  ☹ Esta técnica es muy poco eficiente computacionalmente, el tiempo de ejecución se dispara casi exponencialmente al aumentar el número de elementos a ordenar. En algunas situaciones la solución no es biológicamente interpretable, como la aplicación que se hace en este trabajo.

- **Técnica basada en cadenas de Markov**

  ☹ Las soluciones en la mayor parte de los escenarios tienen valores del MSCE lejos del optimo. Además, aunque los tiempos de ejecución del primer paso del procedimiento son muy buenos, al no serlo los resultados, los tiempos de ejecución del procedimiento total aumentan considerablemente.

- **Técnica basada en el TSP**

  ☺ Las soluciones que ofrece está técnica son una buena aproximación en términos del MSCE al óptimo del problema de optimización (3.4) siempre que se haga la elección razonable de la distancia. Los tiempos de ejecución son muy razonables incluso para valores elevados de n. Las soluciones a las que lleva son interpretables en todos los casos que hemos valorado. Por otro lado, esta técnica ofrece la posibilidad de intervención del usuario de forma que se puede controlar la relación del MSCE y tiempo con el parámetro c.

- **Técnica basada en la teoría de Hodge**

  ⌣̈ Esta técnica se caracteriza por su sencillez de cálculo y por tanto da lugar a tiempos de ejecución muy bajos en el paso 1. Las soluciones que ofrece son biológicamente interpretables. Se trata de una técnica con muchas posibilidades futuras aun no exploradas, que van desde la propuesta de otras formas de definir la intensidad de las tripletas a la posibilidad de definir medidas de la *confianza* del orden obtenido.

  ⌢̈ El MSCE de la solución obtenida es ligeramente mayor que el obtenido con otras técnicas en algunos escenarios.

Capítulo 4

Contraste de Igualdad de Órdenes Circulares

Our only certainty today is uncertainty.
Zygmunt Bauman.

En el presente capítulo se presenta la solución al problema del contraste de igualdad de órdenes circulares entre diferentes poblaciones. Hasta donde conocemos esta es la primera vez que este contraste se plantea en la literatura. El estudio de este problema viene motivado por el interés de los biólogos en la comparación del orden de activación de los genes entre diferentes especies. La estructura del capítulo es la siguiente. En la Sección 4.1 se formula el problema, en la la Sección 4.2 se presenta un procedimiento de remuestreo que consiste en un algoritmo de selección aleatoria, diferentes propuestas de estadísticos estandarizados y el cálculo de ciertos índices de confianza. En la Sección 4.3 se comparan con datos simulados los resultados del procedimiento dependiendo del estadístico.

4.1. Introducción

El problema que nos ocupa es el contraste de la hipótesis de igualdad de los ordenes circulares entre n elementos en S poblaciones. Sea $\boldsymbol{O}_s$ el orden circular para esos n elementos en la población s, donde $s = 1, 2, \ldots, S$. Entonces el problema de interés se formula de la manera siguiente,

$$\begin{aligned} &H_0 : \boldsymbol{O}_1 = \boldsymbol{O}_2 = \ldots = \boldsymbol{O}_S = \boldsymbol{O}_G \\ &H_1 : H_0 \text{ no es cierta.} \end{aligned} \tag{4.1}$$

donde $\boldsymbol{O}_G$ es el orden circular global desconocido.
Se dispone de un conjunto de datos $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_s, \ldots, \boldsymbol{\Theta}_S)'$ procedentes de las S poblaciones donde $\boldsymbol{\Theta}_s = (\boldsymbol{\Theta}_{1s}, \ldots, \boldsymbol{\Theta}_{js}, \ldots, \boldsymbol{\Theta}_{p_s s})'$, $j = 1, \ldots, p_s$, son las observaciones de los experimentos llevados a cabo en la población s, $s = 1, \ldots, S$ y $\boldsymbol{\Theta}_{js} = (\theta_{1js}, \ldots, \theta_{ijs}, \ldots, \theta_{njs})'$ $i = 1, \ldots, n$, es el vector que contiene las observaciones del experimento j. Se tienen $p = \sum_{s=1}^{S} p_s$ experimentos totales, donde p_s es el número de experimentos en la población s.

Aunque este problema es la primera vez que se plantea, existen dos posibles antecedentes por tener ciertas similitudes. Por un lado, el problema que compara dos órdenes en el espacio euclídeo resuelto en Berger (1984). Por otro lado, el problema de contraste de un orden circular prefijado para una muestra resuelto en Fernández et al. (2012). Ambos presentan un enfoque paramétrico y se refieren a un máximo de dos poblaciones.

Para el contraste que aquí se plantea no parece factible el uso de un test razón de verosimilitudes, entre otras razones debido a las dificultades que supone la formulación del propio estadístico. Como hay varios experimentos en cada población, es posible aleatorizar los experimentos entre especies y de esta forma surge el planteamiento de un test no paramétrico basado en permutaciones de los experimentos (Raj y Khamis (1958), Odén y Wedel (1975)).

En este trabajo se ha comenzado planteando un test de permutaciones

clásico sobre los experimentos. Se ha comprobado mediante simulaciones que dicho test no tiene un buen comportamiento. Se observó como en escenarios simulados bajo la hipótesis nula de igualdad de órdenes circulares, este test rechazaba en muchas ocasiones.

En la literatura varios autores han puesto de manifiesto problemas con el test de permutaciones en presencia de muestras no balanceadas y/o cuando las poblaciones a comparar difieran en otras características diferentes de la localización. En concreto, en el trabajo de Huang et al. (2006) se cuestionan las situaciones dónde aplicar un test de permutaciones y más concretamente, Kerr (2009) estudia el comportamiento de estos tests cuando se dispone de muestras no balanceadas. Para resolver este problema se han propuesto varias alternativas que principalmente consisten por un lado en diseñar diferentes procedimientos que utilicen un mecanismo de selección aleatoria más sofisticado basado en bootstrap en varias etapas (Mewhort et al. (2009)). Por otro lado, algunos autores ponen de manifiesto la importancia del estadístico test utilizado y en concreto se proponen diferentes estandarizaciones (ver Chung y Romano (2011, 2013)) que se demuestra que mejoran el error de tipo I.

Con estos antecedentes en mente hemos diseñado un procedimiento que se presenta en la Sección 4.2 y que considera mecanismos de selección aleatoria tipo bootstrap, que detallamos en la Sección 4.2.1, y una variedad de estadísticos que se presentan en la Sección 4.2.2. El procedimiento de remuestreo diseñado ofrece además otras salidas de interés que permiten el cálculo de índices de confianza tal y como se expone en la Sección 4.2.3. El estadístico que se usa en el test definitivo es elegido en base a los resultados presentados en la Sección 4.3 del estudio de simulación.

4.2. Procedimiento de remuestreo

Las propuestas de remuestreo que hemos desarrollado inicialmente tratan de imitar las ideas que otros autores habían probado. Por ejemplo, una de las alternativas planteadas inicialmente es un procedimiento en dos etapas donde la primera es tipo *bootstrap* y en la segunda se realizan permutaciones, inspirado en el propuesto por Mewhort et al. (2009) para muestras no balanceadas. Se comprobó que en situaciones extremas, caracterizadas por variabilidades muy diferentes entre experimentos, este procedimiento fallaba, además de ser computacionalmente muy costoso debido a las dos etapas. El procedimiento de selección de muestras para el test de permutaciones presentamos en detalle en la Sección 4.2.1, se ha diseñado teniendo en cuenta esencialmente la diferencias en la variabilidad de los experimentos y el factor de las muestra no balanceadas. Por otro lado se consideran en la Sección 4.2.2 diferentes versiones para el estadístico test.

4.2.1. Selección aleatoria

En cada selección r, se seleccionan p experimentos aleatoriamente con reemplazamiento. La probabilidad de que el experimento j de la población s sea elegido en la selección r es: $pr(\boldsymbol{\Theta}_{js}) = \frac{1}{(S \cdot p_s)}$. En cada selección se asignan los p_1 a un primer grupo, los siguientes p_2 al segundo y así sucesivamente. A estos grupos se les denomina poblaciones ficticias en el Algoritmo 4 donde se detalla este procedimiento para distinguirles de los grupos de experimentos originales asociados a las poblaciones reales.

En cada selección aleatoria se calcula el valor del estadístico test T. Así, se tiene una distribución de valores de T, $T^r \; \forall \; r = 1, \ldots, N$ donde N es el número de selecciones aleatorias realizadas. Sea T_{obs} el valor del estadístico T con los datos observados originales. Entonces, el p-valor correspondiente al contraste (4.1) se obtiene como sigue,

$$\text{p-valor} = \#\{T^r \geq T_{obs}\}/(N+1).$$

Algoritmo 4: Procedimiento de remuestreo

entrada: $\boldsymbol{\Theta}$.

salida : $\widehat{\boldsymbol{O}}_G^r$, T^r $\forall r = 1, \ldots, N$, p-valor.

Estimar los órdenes circulares de cada población $\widehat{\boldsymbol{O}}_s$ $\forall s = 1, \ldots, S$ y el global $\widehat{\boldsymbol{O}}_G$.

Calcular T_{obs}, el valor observado del estadístico T con el conjunto de datos originales.

for *r desde 1 hasta N* **do**

  Seleccionar p experimentos con reemplazamiento y probabilidad de extracción $pr(\boldsymbol{\Theta}_{js}) = \frac{1}{(S \cdot p_s)}$.

  Se tiene la muestra de experimentos $\boldsymbol{\Theta}^r = \{\boldsymbol{\Theta}^{1r}, \ldots, \boldsymbol{\Theta}^{fr}, \ldots, \boldsymbol{\Theta}^{pr}\}$, $f = 1, \ldots, p$, de la selección r, donde $\boldsymbol{\Theta}^{fr} = \boldsymbol{\Theta}_{js}$ para algún j de alguna población s.

  Asignar los experimentos elegidos en la selección r a cada población ficticia donde se tienen tantas poblaciones ficticias como reales. Sea F_t el conjunto de subíndices de los experimentos que pertenecen a la población ficticia t tal que $F_t = \{\sum_{s=1}^{s=t-1} p_s, \ldots, p_t + \sum_{s=1}^{s=t-1} p_s\}$, $t = 1, \ldots, S$.

  $\boldsymbol{\Theta}^{F_t r} = \{\boldsymbol{\Theta}^{fr} : f \in F_t \text{ en la selección } r\}$.

  Estimar los órdenes circulares para cada población ficticia $\widehat{\boldsymbol{O}}_t^r$ $\forall t = 1, \ldots, S$ y el global $\widehat{\boldsymbol{O}}_G^r$.

  Calcular T^r, el valor del estadístico T para la selección r.

Calcular el p-valor:

$$\text{p-valor} = \#\{T^r \geq T_{obs}\}/(N+1).$$

return $\widehat{\boldsymbol{O}}_G^r$, T^r $\forall r = 1, \ldots, N$, p-valor.

4.2.2. Estadísticos

Los estadísticos que hemos considerado están inspirados en las formulaciones de estadísticos clásicos en problemas de comparación de varias poblaciones tipo ANOVA en donde se mide la diferencia de la variabilidad entre grupos y dentro de los grupos o poblaciones. Para ello consideramos estimadores de los órdenes tanto para el global denominado $\widehat{\boldsymbol{O}}_G$ como para los de cada población denominados $\widehat{\boldsymbol{O}}_s$, $s = 1, \ldots, S$. Se toma como estimador del orden circular aquel orden circular agregado obtenido como resultado de una de las técnicas presentadas en el Capítulo 3.

El vector $\boldsymbol{\Theta}_{js}$ contiene los datos circulares observados en el experimento j de la población s, y $\widetilde{\boldsymbol{\Theta}}_{js}^{(\widehat{\boldsymbol{O}}_G)}$ es el CIRE (2.8) calculado con $\boldsymbol{\Theta}_{js}$ bajo el orden circular global estimado $\widehat{\boldsymbol{O}}_G$. Se hace uso del MSCE (3.2) como medida de distancia y haciendo uso de la misma se formulan dos estadísticos que son la base de los estadísticos test propuestos más adelante. Por un lado se denota por T_0 al estadístico que mide las diferencias entre el conjunto total de datos y el orden circular global estimado y se formula como,

$$T_0 = \sum_{s=1}^{S} \sum_{j=1}^{p_s} \omega_{js} SCE(\boldsymbol{\Theta}_{js}, \widetilde{\boldsymbol{\Theta}}_{js}^{(\widehat{\boldsymbol{O}}_G)}),$$

donde ω_{js} es el peso del experimento j de la población s y para tener en cuenta correctamente las diferencias en la variabilidad entre experimentos proponemos la siguiente definición de los mismos, $\omega_{js} = \frac{\kappa_{js}}{\sum_{j=1}^{p_s} \kappa_{js}}$, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$, donde κ_{js} es el valor del parámetro de concentración conocido o estimado para cada experimento. Por otra parte, denotamos por $\overline{T}$ al estadístico que mide las diferencias medias entre el conjunto de datos de cada población y del orden circular estimado para cada población, es decir,

$$\overline{T} = \sum_{s=1}^{S} d(\boldsymbol{\Theta}_s, \widehat{\boldsymbol{O}}_s) = \sum_{s=1}^{S} \sum_{j=1}^{p_s} \omega_{js} SCE(\boldsymbol{\Theta}_{js}, \widetilde{\boldsymbol{\Theta}}_{js}^{(\widehat{\boldsymbol{O}}_s)}).$$

A partir de T_0 y $\overline{T}$ se definen una serie de estadísticos que se diferencian en el tipo de estandarización utilizada y se presentan en la Tabla 4.1. T_1 es el

único estadístico propuesto sin estandarizar y a continuación T_2 y T_3 se basan en estandarizaciones del tipo de las utilizadas en ANOVA. Por otro lado, en T_4, T_5, T_6, T_7 y T_8 se hace uso de el parámetro de concentración de la distribución de von Mises, κ, cuyos valores pueden ser conocidos o estimados. Estos últimos estadísticos se caracterizan por tener en cuenta de diferentes fuentes la variabilidad dentro de cada población y entre poblaciones e incluso, como en el caso de T_5 y T_7, la diferencia entre dichas variabilidades.

Tabla 4.1: Definiciones de los estadísticos

$T_1 = \quad T_0 - \overline{T}$
$T_2 = \quad \frac{T_0 - \overline{T}}{T_0}$
$T_3 = \quad \frac{T_0 - \overline{T}}{\overline{T}}$
$T_4 = \quad \frac{T_0 - \overline{T}}{\sqrt{\sum_s \sum_{j \in s} (\kappa_{js} - \overline{\kappa})^2}}$
$T_5 = \quad \frac{T_0 - \overline{T}}{\sqrt{\sum_s \sum_{j \in s} (\kappa_{js} - \overline{\kappa}_s)^2}}$
$T_6 = \quad \frac{T_0 - \overline{T}}{\sqrt{\sum_{s=1}^{S} \sum_{j=1}^{p_s} \kappa_{js}^2}}$
$T_7 = \quad \frac{T_0 - \overline{T}}{\sqrt{\sum_s \sum_{j \in s} (\frac{1}{\kappa_j} - \frac{1}{\overline{\kappa}_s})^2}}$
$T_8 = \quad \frac{T_0 - \overline{T}}{\sqrt{\sum_s \sum_{j \in s} (\frac{1}{\kappa_j})^2}}$

4.2.3. Índices de confianza

El procedimiento propuesto de remuestreo para la resolución de este contraste (4.1) genera otras medidas de interés que presentamos en este apartado. Los índices que a continuación se muestran han sido diseñados por demanda específica de los especialistas del campo de aplicación para ayudar a la interpretación biológica de los resultados.
Las N selecciones aleatorias generadas en el proceso dan lugar a N estimadores del orden $\boldsymbol{O}_G$ y por tanto a una distribución de frecuencias para $\widehat{\boldsymbol{O}}_G$. Esta distribución es valiosa para dos cuestiones: la validación del orden global y el cálculo del coeficiente de confianza.

Orden estimado más frecuente. El orden circular más frecuente en la distribución puede considerarse un orden global razonable. Si además coincide con el orden circular estimado con el conjunto original de datos ($\widehat{\boldsymbol{O}}_G$) podemos considerarlo como un indicio adicional de validez de dicho estimador del orden global.

Confianza del orden estimado. Denotamos por $C(\boldsymbol{O})$ a la confianza del orden $\boldsymbol{O}$ definida a través de la frecuencia relativa en tanto por cierto de aparición como $C(\boldsymbol{O}) = fr(\boldsymbol{O}) \cdot 100/N$. De este modo la confianza de nuestro orden estimado $\widehat{\boldsymbol{O}}_G$ es

$$C(\widehat{\boldsymbol{O}}_G) = fr(\widehat{\boldsymbol{O}}_G) \cdot 100/N.$$

En algunas aplicaciones, puede ser de interés tener en cuenta no sólo un orden circular simple sino un orden circular parcial (2.7). Por ejemplo, en nuestra aplicación tienen sentido todos aquellos órdenes que difieren del orden global estimado en exactamente la permutación de los elementos i e $i+1$ (considerando $n+1=1$). Definimos la unión de todos estos órdenes como sigue,

$$\widehat{\boldsymbol{O}}_P = \bigcup_{r=1}^{N} I_{\{\widehat{\boldsymbol{O}}_G^r \equiv \widehat{\boldsymbol{O}}_G\}} \widehat{\boldsymbol{O}}_G^r,$$

donde en $\widehat{\boldsymbol{O}}_G^r \equiv \widehat{\boldsymbol{O}}_G$ el símbolo $\equiv$ significa que son iguales salvo en una permutación de dos elementos consecutivos. Entonces, el coeficiente de confianza de $\widehat{\boldsymbol{O}}_P$ es la suma de las frecuencias relativas de todos los órdenes circulares participantes en la unión tal que,

$$C(\widehat{\boldsymbol{O}}_P) = \big(\sum_{r=1}^{N} I_{\{\widehat{\boldsymbol{O}}_G^r \equiv \widehat{\boldsymbol{O}}_G\}} fr(\widehat{\boldsymbol{O}}_G^r)\big) \cdot 100/N \tag{4.2}$$

4.3. Estudio de simulación

En esta sección se evalúan mediante simulación los diferentes estadísticos propuestos. Supondremos $\theta_{ij} \sim M(\phi_{ij}, \kappa_j)$ $i = 1, \ldots, n, j = 1, \ldots, p$, independientes. Con la finalidad de generar diversas situaciones de interés tanto bajo la hipótesis nula como bajo la alternativa se consideran diferentes configuraciones de los parámetros $\boldsymbol{\Phi}_s = (\phi_{1s}, \ldots, \phi_{is}, \ldots, \phi_{ns})'$, $i = 1, \ldots, n$ y κ_{js}, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$, que se describen a continuación.
Los valores de κ usados son estimaciones realizadas con datos de experimentos reales que aparecen en el Capítulo 5 y se encuentran en la Tabla B.5. Cuando se habla de valores de κ reales se entiende que se han usado directamente los estimadores tal y como aparecen en la Tabla 5.2. Cuando se habla de valores agrupados es porque se han hecho grupos por bajos para la población 1, medios para la población 2 y altos para la población 3 (ver Tabla 3.15). En cada escenario se han simulado 100 conjuntos de datos.

En lo que respecta a los escenarios que representan situaciones bajo la hipótesis nula, las configuraciones seleccionadas para $\boldsymbol{\Phi}_s$ y κ_{js}, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$, se describen en la Tabla 4.2. Se han probado diferentes opciones en los vectores de medias manteniendo siempre la igualdad de órdenes circulares $\boldsymbol{O}_1 = \boldsymbol{O}_2 = \boldsymbol{O}_3$.
En el escenario N_1 los valores de $\boldsymbol{\Phi}_s$ se han tomado de los valores observados para 6 genes en el experimento *S. pombe elut 1* de Rustici et al (2004), que corresponde al experimento 8 de la Tabla B.5. Estos valores son agrupados en dos grupos cuyas medias circulares son usadas para el escenario N_2. En

el escenario N_3 los valores del vector de medias se encuentran uniformemente distribuidos en el intervalo $(0, 2\pi)$ para la primera población,en el intervalo $(0, \pi/4)$ para la segunda población y en el intervalo $(\pi, 9\pi/8)$ para la tercera población. En el escenario N_4 los valores para las medias circulares son los mismos que en el escenario N_1, sin embargo se introduce heterogeneidad entre poblaciones mediante variaciones en los valores de κ.

Tabla 4.2: Escenarios bajo la hipótesis nula

Escenario	$\boldsymbol{\Phi}$	κ
N_1	(1.57, 2.28, 2.51, 2.82, 2.85, 1.19)'	Reales
N_2	(1.37, 2.60, 2.60, 2.60, 2.60, 1.37)'	Reales
N_3 ($\boldsymbol{\Phi}_1$)	(1.05, 2.09, 3.14, 4.19, 5.23, 6.28)'	Reales
N_3 ($\boldsymbol{\Phi}_2$)	(0.13, 0.26, 0.39, 0.52, 0.65, 0.78)'	Reales
N_3 ($\boldsymbol{\Phi}_3$)	(3.21, 3.27, 3.34, 3.40, 3.47, 3.53)'	Reales
N_4	(1.57, 2.28, 2.51, 2.82, 2.85, 1.19)'	Agrupados

Por otro lado, en los escenarios bajo la hipótesis alternativa (Tabla 4.3) se tiene valores de los vectores de medias circulares que en cada población siguen un orden circular diferente, $\boldsymbol{O}_1 \neq \boldsymbol{O}_2 \neq \boldsymbol{O}_3$.

Tabla 4.3: Escenarios bajo la hipótesis alternativa

Escenario	$\boldsymbol{\Phi}$	κ
A_1 ($\boldsymbol{\Phi}_1$)	(1.57, 2.28, 2.51, 2.82, 2.85, 1.19)'	Reales
A_1 ($\boldsymbol{\Phi}_2$)	(4.02, 0.10, 2.15, 1.25, 2.09,4.68)'	Reales
A_1 ($\boldsymbol{\Phi}_3$)	(2.78, 1.96, 2.69, 1.33, 4.05, 0.69)'	Reales
A_2 ($\boldsymbol{\Phi}_1$)	(1.57, 2.28, 2.51, 2.82, 2.85, 1.19)'	Reales
A_2 ($\boldsymbol{\Phi}_2$)	(0.00, 0.75, 5.77, 0.68, 5.69, 1.12)'	Reales
A_2 ($\boldsymbol{\Phi}_3$)	(0.04, 6.27, 0.12, 3.84, 3.70, 1.01)'	Reales
A_3 ($\boldsymbol{\Phi}_1$)	(1.57, 2.28, 2.51, 2.82, 2.85, 1.19)'	Bajos
A_3 ($\boldsymbol{\Phi}_2$)	(4.02, 0.10, 2.15, 1.25, 2.09,4.68)'	Medios
A_3 ($\boldsymbol{\Phi}_3$)	(2.78, 1.96, 2.69, 1.33, 4.05, 0.69)'	Altos

En el escenario A_1 los valores de $\mathbf{\Phi}_s$ se han tomado de los datos del experimento *S. pombe elut 1* de Rustici et al (2004) ($s = 1$), del experimento 2 de *S. cerevisiae* realizado por De Lichtenberg et al. 2002 ($s = 2$) y del experimento *Humanos Thynoc* de Whitfield et al. 2002 ($s = 3$). Estos experimentos se han escogido de manera que los órdenes circulares sean distintos pero no muy alejados entre sí, $\Delta(O_1, O_2) = 0$; $\Delta(O_2, O_3) = 0$ y $\Delta(O_1, O_3) = -0.2$. En el escenario A_2 los valores de $\mathbf{\Phi}_s$ se han tomado de los datos del experimento *S. pombe elut 1* de Rustici et al (2004) ($s = 1$), del experimento *S. pombe cdc* de Oliva et al. 2005 ($s = 2$) y del experimento *S. pombe cdc2* de Rustici et al. 2004 ($s = 3$). Estos experimentos se han escogido de manera que los órdenes circulares estén muy alejados entre sí, $\Delta(O_1, O_2) = -0.2$; $\Delta(O_2, O_3) = -0.2$ y $\Delta(O_1, O_3) = -0.4$. En el escenario A_3 los valores usados para $\mathbf{\Phi}_s$ serán los mismos que en el patrón A_1 pero en este caso se introducen variaciones en los valores de los parámetros de concentración κ.

Tabla 4.4: Resultados de las simulaciones

	Escenario	T1	T2	T3	T4	T5	T6	T7	T8
Error Tipo I	N_1	0.01	0.01	0.01	0	0	0	0.02	0
	N_2	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
	N_3	0	0.055	0.055	0	0.018	0	0	0
	N_4	0.55	0.04	0.04	0.55	0.45	0.58	0.54	0.37
Potencia	A_1	0.73	1	1	0.46	0.43	0.41	0.46	0.24
	A_2	0.7	0.93	0.93	0.43	0.45	0.42	0.46	0.37
	A_3	0.92	0.85	0.85	0.97	0.74	0.98	0.64	0.34

La Tabla 4.4 muestra los resultados más relevantes de las simulaciones. Podemos decir que el test que hace uso del estadístico T_1 rechaza la hipótesis nula con una frecuencia muy superior al nivel establecido en caso de que exista heterogeneidad entre poblaciones, con lo que es una mala opción para el procedimiento que se está diseñando.

Los test que usan los estadísticos T_2 y T_3 funcionan de manera muy similar. Ambos estadísticos detectan correctamente la hipótesis nula con un error de

Tipo I estimado máximo de 0.055 y tienen una potencia estimada mínima de 0.85 por lo que detecta el rechazo adecuadamente.
En el caso de los test que usan los estadísticos estandarizados mediante los valores de κ (T_4, T_5, T_6, T_7 y T_8) tienen un comportamiento similar entre sí y similar en lo que se refiere a la hipótesis nula a lo descrito para T_1. Podemos observar en los resultados de los escenarios bajo la hipótesis nula que en caso de existir diferencias en la variabilidad dentro de cada población (N_1, N_2 y N_3) su comportamiento es adecuado. Sin embargo, si existe heterogeneidad entre poblaciones su uso no sería apropiado.

Resumiendo, los dos estadísticos cuyo uso tiene mejores resultados en cuanto a la potencia son T_2 y T_3. El test que hace uso de cualquiera de ellos tiene un error de Tipo I que no excede de 0.05 en más de una desviación típica. Por otro lado, después de un análisis más detallado de las distribuciones de estos dos estadísticos (Figura 4.1) vemos que el estadístico T_2 tiene un comportamiento más razonable.

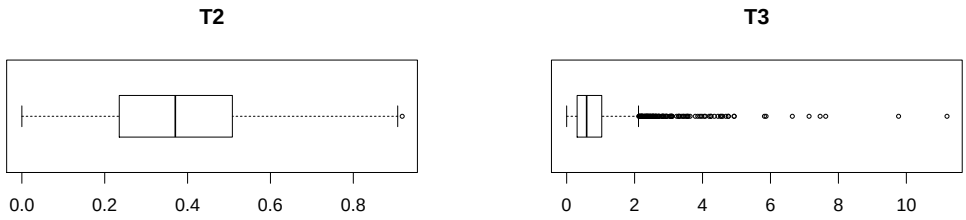


Figura 4.1: Diagramas de cajas para los valores de los estadísticos T_2 y T_3

Finalmente, se puede decir que se ha diseñado un test que resuelve adecuadamente el contraste de igualdad de órdenes circulares entre diferentes poblaciones (4.1). Este test hace uso del procedimiento de selección aleatoria del Algoritmo 4 y del estadístico estandarizado denominado T_2 que se formula como sigue,

$$T = \frac{d(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}_G) - \sum_{s=1}^{S} d(\boldsymbol{\Theta}_s, \widehat{\boldsymbol{O}}_s)}{d(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}_G)}.$$

Capítulo 5

Análisis de datos de expresiones de genes

The final test of a theory is its capacity
to solve the problems which originated it.
George Dantzig.

La cuestión biológica que en este capítulo resolvemos se plantea de forma general como la identificación de aquellos genes asociados al ciclo celular cuyas funciones biológicas se han conservado en el proceso evolutivo. La función biológica de un gen está directamente relacionada con el momento del ciclo celular en el que se activa. Por tanto, el problema se traduce estadísticamente en la comparación entre los órdenes de activación de los genes y para ello se consideran datos de dos levaduras y los humanos. A pesar de que estas tres especies consideradas han sido ampliamente estudiadas, todavía existe un desacuerdo entre investigadores sobre el orden en el que se expresan ciertos genes o cuál es su función biológica. Para resolver este dilema tiene mucho que aportar la inferencia con restricciones y en este caso, de la mano del análisis de datos circulares por la representación del ciclo celular en el círculo unidad.

Los datos disponibles, como veremos en detalle, proceden de experimentos heterogéneos con diferencias sustanciales incluso dentro de la misma especie. En concreto, el momento del ciclo en el que se inicia la observación de las células es diferente en cada experimento, esto se traduce en conjuntos de datos

circulares con diferentes puntos de inicio. Estas características impiden que los valores observados puedan ser agregados directamente para obtener las medias circulares u otros estadísticos resumen. De hecho, el problema que se presenta no puede resolverse correctamente con los métodos existentes en la literatura hasta el momento y hace falta hacer uso de la metodología desarrollada en esta tesis.

Con objeto de acercar al lector al campo de aplicación, se comienza en la Sección 5.1 repasando algunos conceptos biológicos necesarios para la comprensión del problema. Existen muchos manuales y libros de biología donde se pueden consultar mas detalles como pueden ser, Nelson et al. (2005) o Fox (2011). En la Sección 5.2 se presentan los datos originales de expresiones de genes y las manipulaciones necesarias que convierten las series de expresiones en observaciones angulares que son los datos de los que partimos y a los que aplicamos la metodología desarrollada. El problema de agregación de ordenes de diferentes experimentos para cada especie se resuelve en la Sección 5.3 y el problema de comparación de ordenes entre especies en la Sección 5.4.

5.1. Introducción a la biología molecular

La biología molecular se encarga del estudio de las estructuras, los procesos y el funcionamiento de los elementos de los seres vivos. En este contexto encontramos el estudio de las células y la evolución en el tiempo de las mismas a lo largo de diferentes procesos biológicos que se repiten periódicamente en el tiempo tales como el ciclo circadiano (ciclo de un día que regula algunos ritmos biológicos como el sueño), el ciclo celular (ciclo de vida de una célula), ciclos hormonales (variaciones cíclicas de hormonas como la progesterona, el estrogeno o la testosterona) u otros procesos cíclicos.

En este trabajo nos centramos específicamente en el análisis de la expresión de los genes asociados al ciclo celular. La expresión de un gen es un proceso

biológico en el cual la información de dicho gen es usada para la creación de una nueva proteína. Un gen es un segmento de ADN (Ácido Desoxirribonucleico). El ADN es la base de datos de la célula que contiene la información genética. En el proceso de expresión de un gen, el ADN realiza una copia de la información genética en un nuevo segmento que se denomina ARN (Ácido Ribonucleico) o más concretamente ARNm (ARN mensajero) cuando va a ser transportado para la creación de la proteína correspondiente. Por lo tanto, la cantidad de ARNm presente en un cierto momento en la célula es la intensidad de expresión del gen al que representa.

Figura 5.1: Dos microarrays

Las mediciones de los niveles/intensidades de expresión de los genes se realizan mediante los microarrays. En la Figura 5.1 vemos la apariencia de dos microarrays. El resultado final que ofrecen los microarrays es una imagen con los cuadros coloreados según la cantidad de ARNm como se ve en la Figura 5.2.

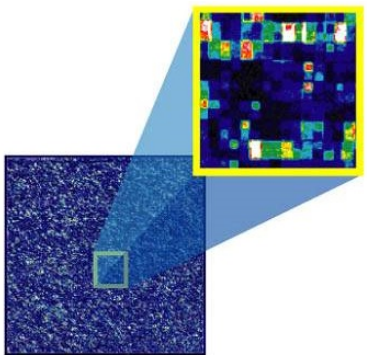


Figura 5.2: Imagen final obtenida de un microarray

Por lo tanto, para cada gen se tiene el nivel de expresión observado en un momento determinado. La expresión de los genes se lleva a cabo de manera continuada en la célula.

El ciclo celular es el ciclo de vida de cada célula y se trata de un conjunto ordenado de sucesos que culmina con la división en dos células hijas. Este

ciclo, como el propio nombre indica, se repite periódicamente y se divide en cuatro fases principales que transcurren en el siguiente orden: G1 (*Growth 1*), S (Síntesis), G2 (*Growth 2*) y M (Mitosis).

Tanto el tiempo de duración total del ciclo, como el de cada una de sus fases, es variable dependiendo de la especie. En la Figura 5.3 (extraída de Bähler (2005)) vemos los ciclos de dos levaduras con diferencias sustanciales. En *S.pombe* la fase G2 predomina considerablemente respecto a las demás mientras que en *S.cerevisiae* las fases con mayor protagonismo son G1 y M.

(a) *S.pombe* (b) *S.cerevisiae*

Figura 5.3: Ciclos celulares de dos especies de levaduras

Un gen se dice asociado al ciclo celular cuando su expresión se repite periódicamente en cada ciclo. En el problema biológico que nos atañe son de interés este tipo de genes y más concretamente el momento del ciclo celular en el cual la intensidad de expresión es máxima. Dicho momento está íntimamente relacionado con la función biológica de cada gen. Es de interés para los biólogos tanto el poder asociar un gen con la fase del ciclo celular donde tiene su máxima expresión como las comparaciones entre genes mediante el orden en el que ocurren dichos momentos.

5.2. Obtención y descripción de los datos

De acuerdo a lo que acabamos de comentar, los datos disponibles inicialmente son series de expresiones en el tiempo que son usadas para determinar el momento de máxima expresión para cada gen. Existe una base de datos realizada por biólogos denominada *Cyclebase.org* de Gauthier et al. (2008), que recoge información sobre los genes asociados al ciclo celular, desde los datos disponibles sobre momentos de máximas expresiones hasta algunos análisis y conclusiones obtenidas de diferentes estudios. Una información importante que podemos encontrar en dicha base de datos es el rango de periodicidad asignado a cada gen. A valores más pequeños de dicho rango, mayor periodicidad de ese gen en el ciclo celular, es decir que hay más certeza de que existe un único punto de máxima expresión que indica el momento en el que se activa ese gen. Nos valdremos de esta información para trabajar con aquellos genes con alta periodicidad en el ciclo celular, en concreto haremos uso de los genes que tienen rango menor que 500.
Como ejemplo del tipo de datos con el que trabajamos, la Figura 5.4 muestra una serie de los niveles medios de ARNm observados en un experimento realizado con un gran número de células en cultivo para el gen $CCNA2$ y medidos en diferentes momentos del ciclo celular.

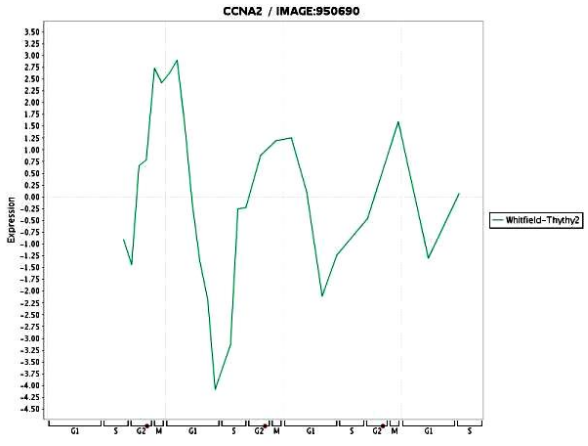


Figura 5.4: Niveles de expresión del gen $CCNA2$

Los datos iniciales de los que disponemos son las series de tiempo con las mediciones de los niveles de expresión para cada gen observados en los siguientes experimentos. Para la levadura *S.pombe*, 3 experimentos de Oliva et al. (2005), 2 experimentos de Peng et al. (2005) y 5 experimentos de Rustici et al. (2004). En el caso de la levadura *S.cerevisiae*, 1 experimento de Cho et al. (1998), otro de de Lichtenberg et al. (2005), dos experimentos de Pramila et al. (2006) y dos experimentos de Spellman et al. (1998). Y en el caso de los Humanos, 4 experimentos de Whitfield et al. (2002).

Antes de estimar el momento de máxima expresión hemos realizado una depuración de estas series recortando valores del inicio y del final necesarias por tratarse de células en cultivo. Las primeras observaciones se eliminan debido a que pueden reflejar un comportamiento irregular hasta que se estabiliza el ciclo celular. Las últimas observaciones son eliminadas debido a que aparece una atenuación en la expresión de los genes que produce un cierto sesgo. Con todo esto, se toma para cada experimento el primer ciclo y medio estabilizado.

Estas series ya depuradas, son usadas para estimar los puntos de máxima expresión de cada gen. Para ello se han desarrollado en la literatura diferentes modelos: Single-pulse Model (SPM) de Zhao et al. (2001), Random Periods Model (RPM) de Liu et al. (2004), transformada de Fourier Straume (2004), series temporales Wichert et al. (2004), Glynn et al. (2006), etc. De todos los modelos que encontramos en la literatura para estimar el momento de máxima expresión, se ha escogido el denominado *Random Periods Model* (RPM) porque es un modelo muy flexible al tener varios parámetros que reflejan diferentes fuentes de variabilidad.

El RPM es un modelo de regresión no lineal desarrollado por Liu et al. (2004). Las suposiciones de las que parte este modelo pueden encontrarse en este artículo y se verifican en nuestro caso.

El modelo se formula como sigue,

$$Y_g(t) = f(t, \eta_g) + \varepsilon_g(t), \quad \forall t = 1, \ldots, n_g,$$

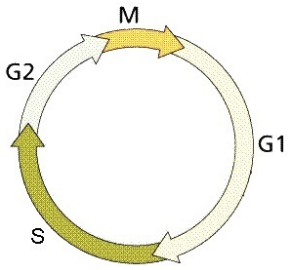
donde t es el tiempo, n_g es la longitud de la serie temporal del gen g y $\varepsilon_g(t)$ es un término de error que se supone que tiene media igual a cero. La función $f(t, \eta_g)$ del modelo es de la forma,

$$f(t, \eta_g) = a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos\Big(\frac{2\pi t}{T exp(\sigma z)} + \phi_g\Big) exp\Big(\frac{-z^2}{2}\Big) dz, \quad \forall t = 1, \ldots, n_g,$$

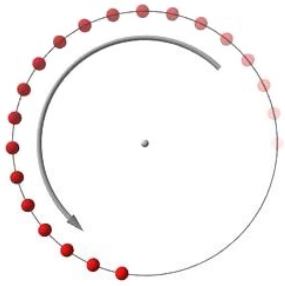
donde $\eta_g = (K_g, T, \sigma, \phi_g, a_g, b_g)$ es el vector de parámetros siendo K_g la amplitud inicial del patrón de expresión periódica, T la duración del ciclo celular, σ la atenuación de la amplitud y ϕ_g es el denominado *phase angle*, que se trata del punto en el círculo que representa el momento del ciclo celular donde el gen g tiene su máxima expresión. $\phi_g = 0$ índica el punto donde las células son liberadas, es decir donde se comienzan las mediciones. Y a_g y b_g son parámetros del ajuste lineal en el tiempo.

Como los datos iniciales han sido depurados podemos fijar la atenuación en cero ($\sigma = 0$) lo cual simplificará sustancialmente los cálculos. Se estiman estos parámetros haciendo uso de la norma L_1 para evitar la influencia de los outliers cuando la series de tiempo tienen pocos puntos.

En este punto cabe comentar que el ciclo celular se representa desde el punto de vista biológico girando en la dirección de las agujas del reloj (Figura 5.5a). Mientras que, el círculo unidad, matemáticamente hablando, hace uso de la dirección contraria a las agujas del reloj (Figura 5.5b). Teniendo en cuenta lo anterior, a los estimadores de los momentos de máxima expresión obtenidos con el RPM se les invierte la dirección de rotación para que los datos circulares con los que trabajaremos usen la dirección natural en el círculo unidad.



(a) Ciclo Celular.



(b) Círculo unidad.

Figura 5.5: Direcciones de rotación.

Del ajuste del modelo del RPM, para cada gen y cada experimento, se obtienen estimadores del parámetro ϕ_g. Así, se tiene el siguiente conjunto de datos circulares $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_s, \ldots, \boldsymbol{\Theta}_S)'$ donde $\boldsymbol{\Theta}_s = (\boldsymbol{\Theta}_{1s}, \ldots, \boldsymbol{\Theta}_{js}, \ldots, \boldsymbol{\Theta}_{p_s s})'$ y $\boldsymbol{\Theta}_{js} = (\theta_{1js}, \ldots, \theta_{ijs}, \ldots, \theta_{njs})'$, $i = 1, \ldots, n$, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$, donde θ_{ijs} es la observación del gen i en el experimento j de la especie s. Se tiene $S = 3$ especies y $p = \sum_s p_s = 20$ experimentos totales de los cuales $p_1 = 10$ son de *S.pombe*, $p_2 = 6$ de *S.cerevisiae* y $p_3 = 4$ de los Humanos. Por otro lado, el número de genes n en cada análisis será diferente. Los genes que estudiamos son ortólogos en las diferentes especies como se muestra en la Tabla 5.1. Los ortólogos son genes en distintas especies cuyas secuencias de ADN son muy similares entre sí y por tanto se sospecha que lo serán también sus funciones biológicas. Se realizan diferentes análisis que involucran diferente número de genes, por lo tanto el valor de n se especificará en cada análisis.

En muchos análisis que veremos a continuación se hace uso de una medida de la concentración de los datos que se utiliza como peso del experimento para lo que asumimos $\theta_{ijs} \sim M(\phi_{is}, \kappa_{js})$ y usamos el parámetro de concentración κ. Debido a la falta de réplicas en los datos, necesarias para estimar κ con el EMV (2.2), hacemos uso de un procedimiento de estimación presentado en Fernández et al. (2012). De este último análisis se obtienen los valores estimados de κ_{js} $j = 1, \ldots, p_s, s = 1, 2, 3,$, para el caso de $n = 11$ ver Tabla 5.2 y para $n = 34$ Tabla 5.3 que supondremos conocidos.

Tabla 5.1: Conjunto de genes ortólogos en las 3 especies

Genes	*S.pombe*	*S.cerevisiae*	*Humanos*
ϕ_1	*ace2*	*SWI5*	*ZNF367*
ϕ_2	*cdc18*	*CDC6*	*CDC6*
ϕ_3	*mik1*	*SWE1*	*PKMYT1*
ϕ_4	*hhf1*	*HHF1*	*HIST2H4B*
ϕ_5	*hta2*	*HTA2*	*H2AFX*
ϕ_6	*fkh2*	*FKH1*	*FOXM1*
ϕ_7	*klp5*	*KIP3*	*KIF10*
ϕ_8	*cig2*	*CLN2*	*CCNB1*
ϕ_9	*plo1*	*CDC5*	*PLK1*
ϕ_{10}	*slp1*	*CDC20*	*CDC20*
ϕ_{11}	*rad21*	*MCD1*	*RAD21*
ϕ_{12}	*mcp*1	*ASE*1	
ϕ_{13}	*mid*2	*BUD*4	
ϕ_{14}	*chs*2	*CHS*2	
ϕ_{15}	*sid*2	*DBF*2	
ϕ_{16}	*eng*1	*DSE*4	
ϕ_{17}	*hht*3	*HHT*1	
ϕ_{18}	*h*3_3	*HHT*2	
ϕ_{19}	*cdc*15	*HOF*1	
ϕ_{20}	*htb*1	*HTB*2	
ϕ_{21}	*pht*1	*HTZ*1	
ϕ_{22}	*fin*1	*KIN*3	
ϕ_{23}	*mob*1	*MOB*1	
ϕ_{24}	*mrc*1	*MRC*1	
ϕ_{25}	*msh*6	*MSH*6	
ϕ_{26}	*myo*3	*MYO*1	
ϕ_{27}	*pol*1	*POL*1	
ϕ_{28}	*pol*2	*POL*2	
ϕ_{29}	*SPAC1705_03C*	*PST*1	
ϕ_{30}	*rhp*51	*RAD*51	
ϕ_{31}	*ssb*1	*RFA*1	
ϕ_{32}	*cdc*22	*RNR*1	
ϕ_{33}	*psm*3	*SMC*3	
ϕ_{34}	*rgs*1	*SST*2	

Tabla 5.2: Estimadores del parámetro de concentración κ con $n = 11$

Especies	Experimentos									
	1	2	3	4	5	6	7	8	9	10
S.pombe	1.59	0.85	1.68	1.11	8.94	1.38	0.63	26.28	2.56	3.59
S.cerevisiae	9.12	1.60	19.19	23.88	6.90	4.75				
Humanos	1.43	2.72	26.56	2.29						

Tabla 5.3: Estimadores del parámetro de concentración κ con $n = 34$

Especies	Experimentos									
	1	2	3	4	5	6	7	8	9	10
S.pombe	1.75	1.01	1.24	1.25	1.88	1.71	0.85	30.34	2.29	2.85
S.cerevisiae	2.47	1.51	32.88	6.04	3.55	2.21				

5.3. Agregación de órdenes circulares en cada especie

Para obtener el orden de activación de los genes seleccionados en cada especie utilizamos el método **TSP3** ($c = 2$) por ser el método ganador en las simulaciones. Por otro lado, dicho orden se compara con el orden que ofrece la base de datos Cyclebase (Gauthier et al. (2008)) que sirve de guía para biólogos.

Para realizar dicha comparación se hace uso de dos criterios, el MSCE (Definición 3.1) y los p-valores obtenidos del test condicional propuesto en Fernández et al. (2012). Estos p-valores proceden de contrastes similares por lo que tiene sentido su combinación haciendo uso del método de Fisher (Fisher (1925)) como sigue. Sea $L = -\sum_{j=1}^{p_s} \log(p_j)$ donde p_j, $j = 1, \ldots, p_s$, es el p-valor obtenido del test condicional que contrasta el orden en el experimento j. Como estos p-valores son independientes, entonces asíntoticamente se verifica que $2L \sim \chi^2_p$, y siendo L_{obs} el valor observado de L se tiene que Fp-valor$= pr(\chi^2_p > 2T_{obs})$.

Veremos como con los criterios propuestos se obtienen, en cada especie, órdenes muy razonables y bastante mas cercanos a los datos que los que se obtiene con Cyclebase. De hecho, los órdenes obtenidos se utilizan para revisar los estimadores de los momentos de máxima expresión para los diferentes genes y especies que proporciona Cyclebase. Los estimadores así obtenidos utilizan, a través del orden, la información de todos los experimentos y en este sentido mejoran a los estimadores de Cyclebase que se obtienen a partir de los datos de un único experimento (que puede ser diferente para cada gen y especie). Estos análisis y otros similares de interés biológico están recopilados en el trabajo de Rueda et al. (2014a).

5.3.1. *S.pombe* (34 genes)

En el análisis de la levadura *S.pombe* se han usado los datos de las Tablas B.1 y B.2 que se corresponden con las observaciones de los 10 experimentos realizados en esta especie para los 34 genes que son ortólogos con la especie *S.cerevisiae* y periódicos en el ciclo celular. El orden circular estimado (**TSP Orden**) que se obtiene para estos 34 genes es el siguiente,

$$htb1 \preceq hta2 \preceq hhf1 \preceq hht3 \preceq h3\_3 \preceq mrc1 \preceq cdc22 \preceq ssb1 \preceq mcp1 \preceq plo1 \preceq \\ \preceq rgs1 \preceq rhp51 \preceq pol2 \preceq chs2 \preceq slp1 \preceq fkh2 \preceq myo3 \preceq SPAC1705\_03C \preceq \\ \preceq cig2 \preceq psm3 \preceq klp5 \preceq sid2 \preceq\preceq ace2 \preceq cdc15 \preceq mid2 \preceq msh6 \preceq cdc18 \preceq \\ \preceq mik1 \preceq fin1 \preceq eng1 \preceq rad21 \preceq mob1 \preceq pol1 \preceq pht1 \preceq htb1 \tag{5.1}$$

El orden que da Cyclebase (**CycleB Orden**) para estos mismos 34 genes es el siguiente,

$$htb1 \preceq hhf1 \preceq hta2 \preceq hht3 \preceq h3\_3 \preceq mcp1 \preceq plo1 \preceq fkh2 \preceq slp1 \preceq myo3 \preceq \\ \preceq pol2 \preceq SPAC1705\_03C \preceq ace2 \preceq chs2 \preceq cdc15 \preceq klp5 \preceq sid2 \preceq rgs1 \preceq \\ \preceq cdc18 \preceq rad21 \preceq fin1 \preceq rhp51 \preceq mik1 \preceq mob1 \preceq cig2 \preceq mrc1 \preceq msh6 \preceq \\ \preceq pol1 \preceq psm3 \preceq eng1 \preceq ssb1 \preceq cdc22 \preceq mid2 \preceq pht1 \preceq htb1$$

Tabla 5.4: Estimadores según Cyclebase y según el CIRE bajo **TSP Orden** para los 34 genes de *S.pombe*

Genes	Cyclebase	CIRE
htb1	6.22	6.22
hta2	0.00	0.00
hhf1	0.00	0.00
hht3	0.06	0.06
h3_3	0.06	0.06
mrc1	5.09	4.76
cdc22	5.22	4.76
ssb1	5.22	4.76
mcp1	**3.83**	**4.76**
plo1	4.27	4.76
rgs1	4.78	4.76
rhp51	4.96	4.76
pol2	4.65	4.76
chs2	4.71	4.76
slp1	4.65	4.76
fkh2	4.59	4.76
myo3	4.65	4.76
SPAC1705_03C	4.65	4.76
cig2	5.09	4.76
psm3	5.09	4.76
klp5	4.78	4.76
sid2	4.78	4.76
ace2	4.71	4.76
cdc15	4.71	4.76
mid2	5.40	5.07
msh6	5.09	5.07
cdc18	4.90	5.07
mik1	5.03	5.07
fin1	4.96	5.07
eng1	5.15	5.07
rad21	4.96	5.07
mob1	5.03	5.07
pol1	5.09	5.07
pht1	6.09	6.09

Tabla 5.5: Comparación de órdenes con Cyclebase para *S.pombe*

	MSCE	**Fp-valor**
TSP Orden	0.06169	0.8443
CycleB Orden	0.0914	3.62e-01

En la Tabla 5.5 se pueden observar los resultados de estos órdenes con los datos observados. El MSCE claramente es mejor para **TSP Orden** (5.1), pero además, con el criterio del Fp-valor, observamos que mientras el **TSP Orden** tiene asociado un p-valor muy alto, el **CycleB Orden** se rechaza a niveles muy bajos.
Calculando el CIRE para el **TSP Orden** sobre los estimadores de Cyclebase se obtienen nuevos estimadores de los momentos de activación de los genes. En la Tabla 5.4 se incluyen ambos conjuntos de estimadores y se obtiene que la mayor diferencia se encuentra en el gen *mcp*1 que es el gen con peor coeficiente de periodicidad. Esto implica que la estimación dada por Cyclebase no es tan fiable como los estimadores nuevos ya que Cyclebase no utiliza la información de todos los experimentos.

5.3.2. *S.cerevisiae* (34 genes)

En el análisis de la levadura *S.cerevisiae* se han usado los datos de las Tablas B.3 y B.4 que se corresponden con las observaciones de los 6 experimentos realizados en esta especie para los 34 genes que son ortólogos con la especie *S.pombe* siendo periódicos en el ciclo celular. El orden circular estimado (**TSP Orden**) que se obtiene para estos 34 genes es el siguiente,

$$HTZ1 \preceq HHF1 \preceq HTA2 \preceq HTB2 \preceq HHT2 \preceq HHT1 \preceq KIP3 \preceq FKH1 \preceq \\ \preceq SWI5 \preceq BUD4 \preceq CHS2 \preceq MYO1 \preceq CDC5 \preceq HOF1 \preceq MOB1 \preceq ASE1 \preceq \\ \preceq CDC20 \preceq KIN3 \preceq DBF2 \preceq DSE4 \preceq PST1 \preceq CDC6 \preceq RAD51 \preceq RFA1 \preceq \\ \preceq MSH6 \preceq MRC1 \preceq CLN2 \preceq RNR1 \preceq MCD1 \preceq POL2 \preceq SMC3 \preceq POL1 \preceq \\ \preceq SWE1 \preceq SST2 \preceq HTZ1. \tag{5.2}$$

El orden que da Cyclebase (**CycleB Orden**) para estos 34 genes de *S.cerevisiae* es el siguiente,

$$HHT2 \preceq HHF1 \preceq HHT1 \preceq HTA2 \preceq HTB2 \preceq KIP3 \preceq HTZ1 \preceq FKH1 \preceq$$
$$\preceq CDC5 \preceq SWI5 \preceq BUD4 \preceq MOB1 \preceq ASE1 \preceq CHS2 \preceq MYO1 \preceq HOF1 \preceq$$
$$\preceq CDC20 \preceq KIN3 \preceq DBF2 \preceq SST2 \preceq CDC6 \preceq PST1 \preceq DSE4 \preceq RFA1 \preceq$$
$$\preceq MRC1 \preceq MSH6 \preceq POL1 \preceq RNR1 \preceq SMC3 \preceq MCD1 \preceq RAD51 \preceq$$
$$\preceq CLN2 \preceq POL2 \preceq SWE1 \preceq HHT2.$$

Tabla 5.6: Comparación con Cyclebase para *S.cerevisiae*

	MSCE	**Fp-valor**
TSP Orden	0.0282	0.1659
CycleB Orden	0.0875	8.64e-28

Los resultados de la comparación del orden estimado con la metodología desarrollada en este trabajo y el orden dado por Cyclebase se muestran en la Tabla 5.6. En este caso, volvemos a encontrarnos con diferencias en el MSCE, incluso mayores que para la especie *S.pombe*, teniendo como mejor orden circular para estos 34 genes el dado por **TSP Orden** (5.2). Además, con el criterio del Fp-valor podemos observar de nuevo que **CycleB Orden** se rechaza a niveles muy bajos para estos datos mientras que **TSP Orden** no.
A continuación, calculamos los nuevos estimadores de los momentos de máxima expresión para cada gen en la especie *S.cerevisiae* mediante el CIRE para **TSP Orden** sobre los estimadores de Cyclebase. En la Tabla 5.7 se muestran ambos estimadores y se observa que la mayor diferencia en este caso se encuentra en el gen $SST2$ que se trata también de un gen con baja periodicidad en el ciclo celular, el segundo peor de los aquí estudiados. Podemos concluir que dado que estos estimadores son resultado de tener en cuenta la información de todos los experimentos disponibles y vista la mejora en el MSCE, son más adecuados que los ofrecidos por Cyclebase.

Tabla 5.7: Estimadores según Cyclebase y según el CIRE bajo **TSP Orden** para los 34 genes de *S.cerevisiae*

Genes	Cyclebase	CIRE
HTZ1	0.57	0.03
HHF1	6.16	0.03
HTA2	0.00	0.03
HTB2	0.00	0.03
HHT2	6.09	0.03
HHT1	6.22	0.03
KIP3	0.38	0.38
FKH1	0.63	0.63
SWI5	1.57	1.57
BUD4	1.57	1.57
CHS2	1.88	1.78
MYO1	1.88	1.78
CDC5	1.57	1.78
HOF1	1.95	1.88
MOB1	1.82	1.88
ASE1	1.88	1.88
CDC20	2.26	2.26
KIN3	2.58	2.58
DBF2	2.70	2.70
DSE4	4.15	3.83
PST1	3.77	3.83
CDC6	3.58	3.83
RAD51	5.09	5.01
RFA1	4.96	5.01
MSH6	5.03	5.01
MRC1	5.03	5.01
CLN2	5.15	5.01
RNR1	5.03	5.01
MCD1	5.09	5.01
POL2	5.15	5.01
SMC3	5.03	5.01
POL1	5.03	5.01
SWE1	5.47	5.01
SST2	**3.14**	**5.01**

5.3.3. Humanos (11 genes)

En el análisis de los Humanos se han usado los datos de la Tablas B.5 que se corresponden con las observaciones de los 4 experimentos realizados en esta especie para los 11 genes que son ortólogos con las ambas levaduras *S.pombe* y *S.cerevisiae*, siendo periódicos en el ciclo celular. El orden circular estimado (**TSP orden**) que se obtiene para estos 11 genes es el siguiente,

$$HIST2H4B \preceq ZNF367 \preceq PKMYT1 \preceq H2AFXCDC20 \preceq CCNB1 \preceq \\ \preceq PLK1 \preceq RAD21 \preceq KIF10 \preceq FOXM1 \preceq CDC6 \preceq HIST2H4B \tag{5.3}$$

El orden que da Cyclebase (**CycleB orden**) para estos mismos 11 genes es el siguiente,

$$HIST2H4B \preceq H2AFX \preceq FOXM1 \preceq KIF10 \preceq CCNB1 \preceq PLK1 \preceq \\ \preceq CDC20 \preceq RAD21 \preceq ZNF367 \preceq CDC6 \preceq PKMYT1 \preceq HIST2H4B$$

Tabla 5.8: Comparación con Cyclebase para Humanos

	MSCE	**Fp-valor**
TSP Orden	0.0085	0.4633
CycleB Orden	0.0995	6.12e-09

En este caso volvemos a encontrar diferencias sustanciales en la comparación de ambos órdenes. En la Tabla 5.8 se observan el MSCE y el Fp-valor para ambos órdenes, teniendo **TSP Orden** (5.3) considerablemente mejores resultados según ambos criterios.
A continuación, pasamos a calcular los nuevos estimadores de los momentos de máxima expresión para los Humanos mediante el CIRE para **TSP Orden** sobre los estimadores de Cyclebase.

En la Tabla 5.9 se muestran ambos estimadores observando la mayor diferencia en los genes $HIST2H4B$ y $ZNF367$. En este caso lo que resulta reseñable es que se puede observar que las dos histonas ($HIST2H4B$ y $H2AFX$) aparecen separadas en el **TSP Orden** con dos genes entre ellas. Es habitual que las histonas se expresen juntas y con una función biológica muy clara. Por lo que estos resultados pueden llevar a nuevas hipótesis de estudio para los biólogos respecto a la función biológica de los genes $ZNF367$ y *PKMYT1* que aparecen entre las histonas.

Tabla 5.9: Estimadores según Cyclebase y según el CIRE bajo **TSP Orden** para los 11 genes de los Humanos

Genes	Cyclebase	CIRE
HIST2H4B	**5.15**	**4.68**
ZNF367	**4.21**	**4.68**
PKMYT1	5.03	5.03
H2AFX	0.00	0.00
CDC20	1.38	1.27
CCNB1	1.32	1.27
PLK1	1.32	1.27
RAD21	1.51	1.27
KIF10	1.19	1.27
FOXM1	0.88	1.27
CDC6	4.65	4.65

5.4. Determinación del conjunto de genes que conservan el orden a lo largo de la evolución

En esta sección tratamos el problema principal que ha motivado todo nuestro trabajo que es el de identificar el mayor conjunto de genes que mantiene el orden de activación en las tres especies. Así, estadísticamente se trata de resolver el siguiente problema de contraste de hipótesis:

$$\begin{aligned} H_0 &: \boldsymbol{O}_P = \boldsymbol{O}_C = \boldsymbol{O}_H \\ H_1 &: H_0 \text{ no es cierta.} \end{aligned} \tag{5.4}$$

donde $\boldsymbol{O}_P$ es el orden circular subyacente en la especie *S.pombe*, $\boldsymbol{O}_C$ el correspondiente para la especie *S.cerevisiae* y $\boldsymbol{O}_H$ en los humanos. (En esta sección para denominar a los genes se hace uso del nombre correspondiente al ortólogo de los Humanos.)

Para resolver (5.4) se usa el procedimiento desarrollado en el Capítulo 4 que hace uso del procedimiento de selección aleatoria expuesto en el Algoritmo 4. Haciendo uso de dicha metodología comenzamos contrastando el mayor conjunto de genes ortólogos en estas tres especies y con alta periodicidad en el ciclo celular, es decir el conjunto de los 11 primeros genes de la Tabla 5.1. En el conjunto inicial de datos circulares existen observaciones ausentes que en este caso podrían influir en el análisis, por lo que se realiza imputación de los mismos. Para ello se hace uso del orden circular agregado obtenido en cada especie independientemente. Entonces, si θ_{ijs} es *missing*, el valor imputado se calcula mediante promedios circulares de los valores observados en el experimento j para los genes inmediatamente anteriores y posteriores a i en el orden. Se busca involucrar en el promedio al mínimo número de genes posible que haga que se cumpla el orden.

El conjunto de datos completo (después de la imputación) que se usa en

este análisis se muestra en la Tabla B.5. Para dicho conjunto se obtiene como resultado un p-valor=0.002 por lo que se rechaza la existencia de un orden circular común entre estos 11 genes.
A partir de este resultado, proponemos hacer uso de un procedimiento backward que se explica a continuación cuya finalidad será encontrar el conjunto máximo de genes cuyo orden de activación es común en las tres especies.

Procedimiento backward. En cada paso de este procedimiento se realizan n contrastes con las hipótesis formuladas en (5.4), para los $n-1$ genes, eliminando un gen distinto en cada contraste y siendo n el número de genes del paso previo. Al final de cada paso se elimina definitivamente aquel gen cuya ausencia da lugar a un mayor p-valor. El proceso finaliza cuando se obtiene un p-valor>0.05. Los resultados de cada paso realizando este procedimiento para la comparación entre las tres especies se muestran en la Tabla 5.10.

El resultado final es un conjunto de 7 genes con el que se obtiene un p-valor=0.3 y cuyo orden circular estimado es el siguiente,
$ZNF367 \preceq CDC6 \preceq PKMYT1 \preceq HIST2H4B \preceq H2AFX \preceq KIF10 \preceq FOXM1 \preceq ZNF367$,
con un coeficiente de confianza $C(\widehat{\boldsymbol{O}}_G)$=65 % (calculado según se presenta en la Sección 4.2.3). Este orden circular entre las máximas expresiones de estos 7 genes se mantiene en las 3 especies a estudio y por tanto se conserva a lo largo de la evolución desde las levaduras a los humanos. Este resultado es de utilidad para plantear nuevas hipótesis biológicas.

Por otro lado, se puede obtener un orden circular parcial (2.7) mediante la unión de ordenes circulares resultado del procedimiento de selección aleatoria. Los órdenes elegidos serán aquellos coherentes con el orden circular global, es decir que sólo difieran en una permutación de elementos consecutivos, como se explica en la Sección 4.2.3.

Tabla 5.10: Procedimiento *backward* paso a paso para las 3 especies

S.pombe - *S.cerevisiae* - Humanos									
Número de genes	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**
p-valor	0,002	0,0159	0,03696	0,03796	**0,2977**	0,42457	0,4975	0,36663	0,846
MSCE	0,06274451	0,04195264	0,03647254	0,02259195	0,01321005	0,01334834	0,01166004	0,0087984	0,003731633
C($\widehat{O}_G$)	6,9	45,9	45,6	48,8	64,6	64,8	76,7	74,3	97,3
C($\widehat{O}_P$)	19,7	57,2	50,3	56,5	90,3				
Gen eliminado		$CDC20$	$CCNB1$	$RAD21$	$PLK1$	$H2AFX$	$FOXM1$	$PKMYT1$	$ZNF367$

El orden circular parcial entre estos 7 genes para estas 3 especies es el siguiente,

$$\begin{aligned} &ZNF367 \preceq CDC6 \preceq PKMYT1 \preceq [HIST2H4B, H2AFX] \\ &\preceq [KIF10, FOXM1] \preceq ZNF367. \end{aligned} \tag{5.5}$$

Recordamos que los elementos que se encuentran entre corchetes no tienen un orden definido entre ellos. Se tiene para (5.5) un coeficiente de confianza (4.2) $C(\widehat{\boldsymbol{O}}_P)$=91 %.

Haciendo uso de estos dos últimos resultados y de la información que se encuentra en la literatura (Jensen et al. (2006)), respecto a las fases asociadas inicialmente a cada gen, se ha construido la Tabla 5.11 donde se muestran las fases del ciclo celular para estos 7 genes. A pesar de que tan sólo las fases de las histonas $(HIST2H4B, H2AFX)$ son exactamente comunes, se puede observar que para el resto de genes las fases asociadas tienen un orden coherente con los resultados aquí obtenidos. Por lo tanto, dichos resultados pueden servir para plantear nuevas hipótesis sobre la función biológica de estos genes.

Tabla 5.11: Fases del ciclo circular en las tres especies

Gen	Humanos	*S.pombe*	*S.cerevisiae*
ZNF367	G1	G2/M	G2
CDC6	G1/S	M	M
PKMYT1	S	M	G1/S
HIST2H4B	S	G1/S	S
H2AFX	S/G2	G1/S	S
KIF10	G2	S/G2	S
FOXM1	G2	G2/M	S/G2

Comparaciones dos a dos. También es interesante para conocer el proceso evolutivo realizar un análisis de las especies dos a dos.

En el caso de *S.pombe* y Humanos es interesante descubrir que a exceptión del gen $CDC20$ el orden para el resto de genes se mantiene obteniendo un p-valor=0.19. El orden circular parcial correspondiente es,

$$[ZNF367, CDC6] \preceq [PKMYT1, HIST2H4B, H2AFX]$$
$$\preceq [CCNB1, RAD21, PLK1, KIF10, FOXM1]$$
$$\preceq [ZNF367, CDC6],$$

con un coeficiente de confianza $C(\widehat{\boldsymbol{O}}_P)$=81 %.

En lo que respecta a Humanos y *S.cerevisiae* tan sólo 7 genes mantienen su orden con p-valor=0.13 siendo los mismos genes y el mismo orden que para la comparación entre las 3 especies (5.5), en este caso con un coeficiente de confianza $C(\widehat{\boldsymbol{O}}_P)$=72 %.

En la comparación entre las dos levaduras se mantiene el orden en 10 genes con un p-valor=0.06 y el siguiente orden circular parcial,

$$ZNF367 \preceq CDC20 \preceq CDC6 \preceq$$
$$[CCNB1, RAD21, PKMYT1] \preceq HIST2H4B \preceq$$
$$H2AFX \preceq [KIF10, FOXM1] \preceq ZNF367,$$

con un coeficiente de confianza $C(\widehat{\boldsymbol{O}}_P)$=90 %.

Tabla 5.12: Resumen de las 4 comparaciones.

	#genes	MSCE	p-valor	$C(\widehat{\boldsymbol{O}}_P)$
3 especies	7	0.013	0.30	91 %
S.pombe - S.cerevisiae	10	0.023	0.06	90 %
S.pombe - Humanos	10	0.029	0.19	81 %
S.cerevisiae - Humanos	7	0.009	0.13	72 %

La Tabla 5.12 resume los resultados para las comparaciones de las tres especies y dos a dos. En concreto, no resulta sorprendente que la comparación

entre *S.cerevisiae* y Humanos comparta menos genes. De hecho, parece posible que a lo largo de la evolución las funciones de los genes $CCNB1$ y $RAD21$ se hayan modificado ligeramente, manteniendo a *S.pombe* como eslabón común. Existe una teoría sobre la levadura *S.pombe* que la coloca como eslabón conector entre *S.cerevisiae* y los humanos en lo que se refiere a la regulación génica (Aravind et al. (2000), Roux et al. (2010)). Las dos levaduras comparten un ancestro común de hace millones de años. Mientras *S.pombe* y los humanos parecen haber mantenido algunas de las funciones de dicho ancestro común, *S.cerevisiae* parece haberlas perdido. Los resultados aquí obtenidos (Tabla 5.12) apoyan estas sospechas. Además, no es la primera vez que aparecen similitudes en las funciones biológicas entre *S.pombe* y otros animales (Forsburg (1999, 2007)). Para más detalle sobre los resultados de esta sección ver Barragán et al. (2014b).

Capítulo 6

Software desarrollado: El paquete de R isocir

An algorithm must be seen to be believed.
Donald Ervin Knuth.

La implementación de la metodología expuesta en los dos capítulos anteriores se organiza dentro del paquete de R **isocir** ("**iso**tonic inference for **cir**cular data"), Barragán et al. (2014a). Para comenzar este capítulo, en la Sección 6.1 se hace un repaso de los paquetes de R en la inferencia con restricciones y el análisis de datos circulares. En la Sección 6.2 se expone la estructura de la primera versión estable del paquete v1.1 cuyo contenido es la implementación de los métodos presentados en la Sección 2.4. En la Sección 6.3 mostramos la última versión del paquete **isocir** v2.0 que recoge además todos los métodos novedosos presentados en este trabajo (Capítulos 3 y 4). Finalizamos, en la Sección 6.4, con varios ejemplos de interés ejecutados paso a paso para una mejor comprensión del uso del código implementado.

6.1. Paquetes previos relacionados con isocir

Inferencia con Restricciones Existen diversos paquetes que implementan técnicas de este campo pero tal vez los más conocidos son aquellos rela-

cionados con la regresión isotónica (Sección 2.3.2).

- **isotone**, de Leeuw et al. (2009): Optimización isotónica y PAVA generalizado.
- **Iso**, Turner (2009): Funciones para llevar a cabo la regresión isotónica.
- **ordMonReg**, Balabdaoui et al. (2009): Calcula los estimadores de mínimos cuadrados de una o dos curvas de regresión isotónica.

Análisis de datos circulares A pesar de tratarse de un área que no es especialmente conocida, existen varios paquetes para analizar datos circulares en R, se puede encontrar una revisión en Pewsey et al. (2013). Algunos de los más usados son:

- **circular**, Agostinelli y Lund (2011): Este paquete es una extensión del paquete **CircStats** (Lund y Agostinelli (2009)), contiene la implementación de los métodos recogidos en el libro *Topics in circular Statistics*, Jammalamadaka y SenGupta (2001).
- **CircNNTSR**, Fernández-Durán y Gregorio-Dominguez (2013): paquete para el análisis estadístico de datos circulares haciendo uso de modelos de sumas trigonométricas no negativas.
- **NPCirc**, Oliveira et al. (2013): Este paquete recoge algunos métodos no paramétricos para datos circulares.

Sin embargo, no existía anteriormente a **isocir** (Barragán et al. (2014a)) ningún paquete en R que implementara métodos para analizar datos circulares con restricciones. Con esa finalidad se crea **isocir**, del inglés "**iso**tonic inference for **cir**cular data".

6.2. Estructura del paquete isocir 1.1

Le dedicamos este apartado a esta versión del paquete por contener la estructura principal y la base de futuras ampliaciones del mismo. En principio se

realizó una versión inicial denominada 1.0 que contenía los métodos de estimación y contraste expuestos en Rueda et al. (2009) y Fernández et al. (2012) y que fue subida al CRAN el 27 de abril del 2011. En dicha versión todo el código se encontraba en lenguaje R, pero por cuestiones de eficiencia computacional se tradujo el eje de la función principal del paquete a lenguaje C++. Este fue el primer cambio sustancial en la creación de la versión 1.1. La mejora en los tiempos de ejecución puede observarse en la Tabla 6.1. En dicha tabla se tienen los resultados (media de los segundos en negrita con las desviaciones típicas entre paréntesis) de ejecutar la función principal del paquete (CIREi en v1.0 y CIRE en v1.1) para r conjuntos de datos generados con n elementos y bajo órdenes circulares parciales (2.7) con L conjuntos (cuanto menor número de conjuntos mayor tiempo de ejecución).

Tabla 6.1: Comparación de tiempos de ejecución

r	n	L	R (`CIREi`)	R y C++ (`CIRE`)
50	8	8	**1.10** (0.27)	**0.03** (0.01)
50	8	4	**28.33** (6.30)	**0.67** (0.07)
50	8	3	**84.91** (14.66)	**2.00** (0.24)
20	10	10	**2.78** (0.71)	**0.06** (0.01)
20	10	4	**407.53** (91.01)	**7.88** (1.11)
20	10	3	**2269.86** (579.45)	**43.75** (6.44)
20	15	15	**20.03** (7.78)	**0.35** (0.18)
20	18	18	**55.44** (41.43)	**1.314** (1.58)
20	20	20	**84.89** (65.58)	**2.886** (6.91)
20	20	15	**3378.52** (2407.71)	**61.76** (72.52)

Otra modificación importante que dio lugar a la versión 1.1 fue la programación orientada a objetos, creando un objeto S3 llamado `isocir` que hace que el código sea mucho más amigable para el usuario. Con todo esto se lanzó la siguiente versión 1.1 en abril del 2012 en el CRAN. El artículo Barragán et al.

(2013) *isocir: An R Package for Constrained Inference Using Isotonic Regression for Circular Data, with an Application to Cell Biology* en *Journal of Statistical Software* contiene no sólo explicaciones sobre el manejo del paquete sino también ejemplos de su utilidad en el campo de la biología molecular. Los paquetes de los que depende son **circular** (Agostinelli y Lund (2011)) y **combinat** (Chasalow (2010)), que deben estar instalados en el ordenador antes de cargar **isocir**. Mostramos el esqueleto de esta versión en la Tabla 6.2 con todas las funciones y métodos que se incluyen.

Tabla 6.2: Resumen de los componentes del paquete **isocir** v1.1.

Funciones	Argumentos	Descripción
`sce`	`(arg1, arg2, meanrl)`	Calcula el SCE
`mrl`	`(data)`	Calcula la longitud media resultante
`CIRE`	`(data, groups, circular)`	Calcula el CIRE
`cond.test`	`(data, groups, kappa)`	Evalúa el test condicional
`isocir`	`(cirmeans,SCE,CIRE,pvalue,kappa)`	Objeto S3 de la clase `isocir`
`is.isocir`	`(x)`	Chequea la clase `isocir`
`print.isocir`	`(x,decCIRE,decpvalue,deckappa,...)`	Imprime un objeto de clase `isocir`
`plot.isocir`	`(x, option,...)`	Plot de un objeto de clase `isocir`

A continuación, describimos cada función con más detalle.

- **Funciones `sce()` y `mrl()`**

La función auxiliar `sce` calcula el suma de errores circulares entre un vector n-dimensional (`arg1`) y uno o más vectores de dimensión n (`arg2`).

La función `mrl` calcula las longitudes medias resultantes para el conjunto de datos que se introduzca.

- **Función `CIRE()`**

Esta función calcula el estimador de regresión isotónica circular (CIRE, del inglés, Circular Isotonic Regression Estimator) haciendo uso del algoritmo desarrollado en Rueda et al. (2009) para un orden dado. Se puede ver un resumen de la metodología en el Capítulo 2 de este trabajo. Los argumentos de esta función se resumen en la Tabla 6.3.

El argumento de entrada llamado `data` será una matriz donde cada columna

Tabla 6.3: Argumentos de la función `CIRE`

Argumentos	*Valores*
`data`	vector o matriz con los datos
`groups`	los conjuntos que definen el orden
`circular`	`=TRUE`(por defecto) / `=FALSE`

contiene las medias circulares sin restringir correspondientes a una réplica. En caso de no haber varias réplicas se permite que `data` sea un vector.

En cuanto al argumento `groups`, es un vector cuya posición i contiene el número del conjunto al que pertenece el parámetro ϕ_i. El argumento lógico `circular` fija si el orden dado da la vuelta al círculo siendo por tanto un orden circular (`circular = TRUE`) o no, siendo entonces un orden *abierto* en el círculo (`circular = FALSE`). Por ejemplo, un orden abierto sería de la forma $0 \leq \phi_1 \leq \phi_2 \leq \cdots \leq \phi_n \leq 2\pi$, mientras que el orden circular correspondiente sería $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_n \leq \phi_1$.

La salida de la función `CIRE` es un objeto S3 de clase `isocir` (se explicará más tarde en detalle) que contiene el estimador circular de regresión isotónica (CIRE) $\widetilde{\mathbf{\Phi}}$, las medias circulares sin restringir $\mathbf{\Theta}$ y la correspondiente suma de errores circulares $SCE(\widetilde{\mathbf{\Phi}}, \mathbf{\Theta})$.

- **Función** `cond.test()`

Esta función ejecuta el test condicional desarrollado en Fernández et al. (2012) calculando el p-valor correspondiente para la siguiente hipótesis:

$$H_0: \quad \phi_1, \ldots, \phi_n \in C_O$$
$$H_1: \quad H_0 \text{ no es cierta.}$$

donde C_O es el cono correspondiente a un orden circular simple (Definición 2.8) o parcial (2.7).

Los argumentos de esta función aparecen resumidos en la Tabla 6.4 y los explicamos a continuación.

Tabla 6.4: Argumentos de la función `cond.test`

Argumentos	κ *conocido*	κ *desconocido*
`data`	vector numérico	matriz (tantas columnas como réplicas)
`groups`	vector numérico con los conjuntos del orden a contrastar	
`kappa`	valor numérico	(NULL)
`biasCorrect`	(NULL)	`=TRUE`(por defecto) / `=FALSE`

Los argumentos `data` y `groups` son equivalente a los de la función `CIRE()` con la diferencia de que en este caso el orden representado por `groups` es aquel orden a ser contrastado.

El argumento `kappa` es el valor del parámetro de concentración κ correspondiente a la von Mises subyacente que se esta suponiendo para realizar este test condicional. Dicho argumento se necesita sólo si no hay réplicas, ya que entonces no es posible su estimación. Aunque haya réplicas, si es conocido y se introduce ya no será estimado. Si no es introducido, se estimará por máxima verosimilitud (2.2) y $\hat{\kappa}$ se mostrará en la salida. El argumento `biasCorrect` está relacionado con dicha estimación de κ. Si `biasCorrect=TRUE` se realiza la correción del sesgo tal y como puede verse en Mardia y Jupp (2000, p.87).

La salida de esta función es un objeto S3 de clase `isocir` (explicado más adelante) con los resultados del test condicional: el CIRE $\widetilde{\mathbf{\Phi}}$, las medias circulares sin restringir $\mathbf{\Theta}$, el SCE de $\mathbf{\Theta}$ respecto al orden, $SCE(\widetilde{\mathbf{\Phi}}, \mathbf{\Theta})$, el valor

de κ (estimado o introducido) y el p-valor del test condicional desarrollado en Fernández et al. (2012).

- **Clase** `isocir`

Describimos ahora la clase `isocir`. La función `isocir` crea objetos S3 de clase `isocir` cuya estructura es una lista que contiene los siguientes elementos:

`$cirmeans` es a su vez otra lista con las medias circulares sin restringir. Cuando el argumento `data` es un vector, coincidirá exactamente con los valores de esta lista. Sin embargo, en caso de la existencia de réplicas el argumento `data` será una matriz y `$cirmeans` contendrá las medias circulares de dichas réplicas.

`$SCE` es el valor de la suma de errores circulares entre el CIRE y las medias circulares sin restringir $SCE(\widetilde{\mathbf{\Phi}}, \mathbf{\Theta})$.

`$CIRE` es una lista con el estimador circular de regresión isotónica $\widetilde{\mathbf{\Phi}}$ obtenido bajo el orden circular definido por el argumento `groups`.

`$kappa` es el valor númerico de κ (bien introducido por el usuario cuando es conocido o bien estimado internamente).

`$pvalue` es el valor del p-valor del test condicional calculado con la función `cond.test`.

Cabe comentar que la razón del uso de listas en los argumentos `CIRE` y `cirmeans` se debe a poder representar correctamente los órdenes circulares parciales (2.7). De esta manera cada elemento de la lista contiene los elementos que forman parte de un mismo conjunto de dicho orden. Un caso particular es el orden circular simple en el cual cada conjunto contiene un sólo elemento, es decir cada elemento de la lista contendrá un sólo elemento del orden. En estos casos la representación vectorial sería suficiente y, en caso de que el usuario la prefiriera, su obtención es tan sencilla como hacer uso de la función de R llamada `unlist(miobjeto$CIRE)`.

Estos objetos S3 de clase `isocir` son la salida de las functiones `CIRE` y `cond.test`. Los últimos dos elementos (`$kappa` y `$pvalue`) serán `NULL` en caso de ser salida del la función `CIRE`. En caso de ser la salida de la función `cond.test` se tendrán los valores que son resultado específico del test condicional: `$kappa` y `$pvalue` así como un atributo asociado a `$kappa` llamado `estkappa` que informará de si dicho valor de κ fue introducido por el usuario o ha sido estimado internamente.

Se han definido algunos métodos S3 asociados a la clase `isocir`:

- `isocir(cirmeans = NULL, SCE = NULL, CIRE = NULL, pvalue = NULL, kappa = NULL)`: Esta función crea un objeto de clase `isocir`.
- `is.isocir(x)`: Esta función comprueba si un objeto `x` es de clase `isocir`.
- `print.isocir(x, decCIRE, decpvalue, deckappa, ...)`: Este método se usa para imprimir en pantalla un objeto de clase `isocir`. Se puede elegir el número de decimales a mostrar en cada item con los argumentos `decCIRE, decpvalue` y `deckappa`.
- `plot.isocir(x, option = c(''CIRE'', ''cirmeans''), ...)`: Este método S3 se usa para representar mediante un plot el objeto `x` de clase `isocir`. El argumento `option` da la posibilidad de representar los valores del CIRE (por defecto) o los valores de las medias circulares sin restringir.

6.3. Estructura del paquete isocir 2.0

La estructura básica del paquete de la versión 1.1 explicada en el apartado anterior no ha sufrido modificaciones en siguientes versiones. La versión más actualizada se explica en esta sección y contiene las funciones correspondientes a la implementación de los métodos detallados en los Capítulos 3 y 4 de esta tesis.

En cuanto a las dependencias, además de los paquetes **circular** y **combinat** se añade ahora el paquete **TSP** que contiene las heurísticas que se usarán para

la resolución mediante el enfoque del TSP de la agregación de órdenes circulares (ver Sección 3.4). En la Tabla 6.5 se ve el resumen de las funciones nuevas cuyo uso se detalla posteriormente.

Tabla 6.5: Resumen de las componentes nuevas del paquete **isocir**.

Funciones	**Argumentos**
`msce`	`(data, posorder, ws)`
`cirKendall`	`(phi1, phi2, test, control.test)`
`mcirktau`	`(data, posorder, ws)`
`ACO`	`(data, method, control.method, ws, coef)`
`CLM`	`(data, order0, ws)`
`eq.test`	`(data, popu, ws, method, control.method, output, coef, N)`

- **Funciones** `msce()`, `cirKendall()` **y** `mcirktau()`

La función `msce()` realiza el cálculo de la media de los errores circulares (*mean sum of circular errors*, MSCE) entre un conjunto de datos y un orden circular. Además pueden usarse pesos diferentes para cada experimento del conjunto de datos. Los datos se introducen en el argumento `data` en forma de matriz donde cada fila es un experimento. El orden circular se define en el argumento `posorder` de la misma manera que el argumento `groups` de las funciones `CIRE` y `cond.test`. Se trata de un vector donde en la posición i se encuentra el valor de la posición o conjunto al que pertenece en el orden el elemento i. Los pesos se definen mediante un vector de igual longitud al número de experimentos y se introducen en el argumento `ws`.

La salida es una lista de dos elementos `$msce` y `$msces`. Siendo el primero el valor medio de la suma de errores circulares (promedio por experimentos y por elementos) y el segundo el vector con la suma de errores circulares de cada experimento donde cada valor está promediado por el número de elementos.

La función `cirKendall()` realiza el cálculo de la tau circular de Kendall (Fisher (1993)) entre dos vectores de medidas circulares representados en los argumentos `phi1` y `phi2`. Permite realizar realizar también el test donde la hipótesis nula es $H_0 : \Delta(\phi_1, \phi_2) = 0$. Para realizar este test se fijará el argumento `test=TRUE` que por defecto es `FALSE`. Además se puede controlar la hipótesis alternativa mediante `control.test = c(''noteq'', ''upper'', ''lower'')`.

La salida será el valor numérico de la tau circular de Kendall. En caso de haber realizado el test, la salida será una lista donde el primer elemento es dicho valor numérico y el segundo el p-valor correspondiente al test obtenido como se detalla en Fisher (1993).

La función `mcirktau()` realiza el cálculo de la tau circular de Kendall media entre un conjunto de datos y un orden circular. Los argumentos se introducen de la misma manera que para la función `msce()`.

La salida es también una lista de dos elementos donde el primero (`$mtau`) es el valor de la tau circular media por experimentos y el segundo (`$ntaus`) es un vector con la tau circular de Kendall entre el orden dado y cada experimento.

- **Función** `ACO()`

La función `ACO()` realiza la agregación de órdenes circulares (*Aggregation of Circular Orders*). En la Tabla 6.6 se muestran los argumentos de entrada.

El argumento `data` al igual que en las funciones anteriores contendrá los datos. En este caso sólo tendrá sentido el formato matriz ya que se necesitan varios experimentos para ser agregados. Los argumentos `method` y `control.method` tienen las opciones que se describen en la Tabla 6.7.

Los únicos argumentos que no es obligatorio introducir son los siguientes. El argumento `ws` es un vector con los pesos por experimentos. Si se deja vacío

Tabla 6.6: Argumentos de la función `ACO`

Argumentos	*Valores*
`data`	matriz con los datos
`method`	método de agregación
`control.method`	opción escogida dentro del método
`ws`	pesos por experimentos
`coef`	coeficiente asociado al enfoque TSP

se entienden pesos iguales. El argumento `coef` es el coeficiente aplicado en caso del enfoque TSP denominado c de manera que se tomarán los (`coef`$\cdot n$) órdenes de menor longitud de ruta como posibles soluciones, se calculará el MSCE para todos ellos y el resultado será el que tenga menor MSCE (Algoritmo 2). Si no se introduce, por defecto se usará `coef=1`.

La salida de `ACO()` es una lista que tendrá unos elementos iniciales sea cual sea el método usado:

`$aggre_order` es el orden circular agregado resultado.

`$msce` es el MSCE entre el orden circular agregado y los datos.

`$mtau` es la tau circular de Kendall media entre el orden circular agregado y los datos.

En caso de usar el método Naive, Borda Circular o Cadenas de Markov la salida será una lista sólo con los tres elementos descritos anteriormente. En caso de usar el enfoque TSP se añaden los siguientes elementos:

`$mintour` es el orden circular correspondiente a la ruta de mínima longitud.

`$mt_msce` es el MSCE entre los datos y el orden circular correspondiente a la ruta de mínima longitud.

Tabla 6.7: Opciones de los argumentos `method` y `control.method`.

`method`	Descripción	`control.method`	Descripción
‘‘Naive’’	Método Naive	‘‘tau’’	Tau Circular de Kendall
		‘‘MSCE’’	criterio MSCE
‘‘CB’’	Borda Circular	‘‘pos’’	posiciones
		‘‘cirmean’’	media circular
		‘‘cirmed’’	mediana circular
‘‘CMC’’	Cadenas de Markov Circulares	‘‘1’’	tipo 1
		‘‘2’’	tipo 2
		‘‘3’’	tipo 3
		‘‘4m’’	tipo 4 por mayoría
		‘‘4c’’	tipo 4 contando
‘‘TSP’’	Traveling Salesman Problem	‘‘bin’’	unos y ceros
		‘‘pos’’	posiciones
		‘‘alpha1’’	mixta $\alpha = 1$
		‘‘alpha2’’	mixta $\alpha = 2$
		‘‘alpha3’’	mixta $\alpha = 3$
		‘‘alpha4’’	mixta $\alpha = 4$
		‘‘alphainf’’	mixta $\alpha = \infty$
		‘‘time’’	tiempos
		‘‘arc’’	arcos unidir.
		‘‘chord’’	cuerda
‘‘CH’’	Hodge Circular	‘‘sig’’	signos
		‘‘pos’’	posiciones
		‘‘cos’’	1+cos
		‘‘cmean’’	cos diferencia a la media
		‘‘mrl’’	long med result.
		‘‘e3’’	mixta $\alpha = 3$
		‘‘ave’’	media circular

`$tour_length` es la longitud de la ruta mostrada en `$mintour`.

`$scores` es un vector cuya posición i es la longitud del arco entre el elemento i del orden circular agregado y el siguiente elemento de dicho orden. La suma de los valores de este vector es la longitud de la ruta correspondiente al orden agregado.

En caso de usar el enfoque de Hodge se añaden los siguientes elementos a los tres iniciales:

`$scores` es un vector con los valores correspondientes al vector $\boldsymbol{s}^*$ del método de Hodge (ver Sección 3.5).

`$out` es el elemento l que fue eliminado en el procedimiento (ver Teorema 3.2).

- **Función** `CLM()`

La función `CLM()` ejecuta el Algoritmo 3 denominado *Circular Local Minimization* expuesto en el Capítulo 3 como paso del procedimiento de agregación de órdenes circulares. El conjunto de datos se introduce en `data` en forma de matriz y el orden circular inicial en el argumento `order0`. Este algoritmo obtiene una mejora local, si existe, de la función objetivo (3.4) con la que se aborda la agregación. Se pueden usar pesos por experimentos haciendo uso del argumento `ws`. En la Tabla 6.8 vemos el resumen de los argumentos de entrada de esta función.

Tabla 6.8: Argumentos de la función `CLM`.

Argumentos	*Valores*
`data`	matriz con los datos
`order0`	orden circular inicial
`ws`	vector de pesos por experimentos

Cabe comentar que el vector `order0` contiene el orden circular de los elementos, que en esta memoria se denomina $\boldsymbol{O}$. De esta manera se puede introducir como argumento de entrada de esta función la salida de la función `ACO()`

directamente mediante `$aggre_order`.

La salida de `CLM()` es una lista con los siguientes cuatro elementos:

`$order0` es el orden circular introducido en el argumento de entrada con el mismo nombre.

`$msce0` es el MSCE entre los datos y `$order0`.

`$final_order` es el orden circular final, resultado de la mejora local de `$order0` según el algoritmo CLM.

`$bestsce` es el mejor MSCE obtenido, aquel entre los datos y el orden circular final.

- **Función** `eq.test()`

La función `eq.test()` realiza el test de igualdad de órdenes circulares tal y como se ha explicado en el Capítulo 4. Un resumen de los argumentos de entrada aparece en la Tabla 6.9.

Tabla 6.9: Argumentos de la función `eq.test`.

Argumentos	*Valores*
`data`	matriz con los datos
`popu`	vector que define las poblaciones
`method`	método de agregación
`control.method`	opción del método de agregación
`ws`	pesos por experimentos
`output`	ruta de salida
`coef`	coeficiente asociado al enfoque TSP
`N`	número de selecciones

En la matriz de datos se tienen los experimentos por filas. El argumento `popu` es un vector numérico de la misma longitud que el número de experimentos y que define a que población pertenece cada experimento. De manera que la posición i-ésima de este vector contiene la población a la que pertenece el experimento i-ésimo.

Los argumentos `method`, `control.method` y `coef` definen las opciones de agregación de órdenes circulares y son equivalentes a los correspondientes de la función `ACO()`. Se pueden ver las posibilidades en la Tabla 6.7. Los pesos para cada experimento se introducen en forma de vector en el argumento `ws`, en este caso sin ponderar por la suma total de pesos de la población correspondiente, porque esto se realizará internamente cuando sea necesario. El argumento `output` es opcional e identifica la ruta de la carpeta dónde localizar los dos archivos denominamos *globalorders.csv* y *frequencydist.csv*. El primero contiene todos los órdenes circulares globales calculados así como el valor del estadístico en cada selección aleatoria, mientras que el segundo contiene la distribución de frecuencias de dichos órdenes. Estos archivos se crean sólo en caso de que `output` no sea `NULL`. El argumento `N` es el número de selecciones que se harán para la estimación de la distribución bajo H_0 del estadístico.
Los argumentos obligatorios son `data` y `popu`. Los valores por defecto del resto de argumentos son: `method=‘‘TSP’’`, `control.method=‘‘alpha3’’`, `ws` con pesos iguales, `output=NULL` y `N=500`.

La salida de esta función es una lista con los siguientes elementos:

`$allorders` es una matriz con los N órdenes circulares globales calculados en las N selecciones realizadas y los valores del estadístico en cada selección.

`$pvalue` es el p-valor del test de igualdad de órdenes circulares para los datos introducidos en `data` con las poblaciones definidas por `popu` mediante el procedimiento explicado en el Algoritmo 4.

`$global_order` es el orden circular estimado como global a todas las pobla-

ciones.

`$CC` es el valor del coeficiente de confianza para el orden circular estimado como global. Este coeficiente está calculado según se explica en la Sección 4.2.3.

`$MFO` es el orden más frecuente (*Most Frequent Order*) de todos los obtenidos como globales en las N selecciones aleatorias.

`$CCMFO` es el coeficiente de confianza del orden más frecuente.

6.4. Ejemplos

En este apartado se describen varios ejemplos básicos que ilustran el uso de las funciones contenidas en el paquete **isocir**.

Ejemplo 6.1. *CIRE bajo un orden circular parcial*

Suponemos las siguientes observaciones circulares de elementos:
$\theta_1 = 0.025$; $\theta_2 = 1.475$; $\theta_3 = 3.274$;
$\theta_4 = 5.518$; $\theta_5 = 2.859$;
$\theta_6 = 5.387$;
$\theta_7 = 4.179$; $\theta_8 = 1.962$.
Se trata de uno de los conjuntos de datos incluidos en `isocir` con el nombre de `cirdata`:

```
R> data("cirdata")
R> cirdata
[1] 0.025 1.475 3.274 5.518 2.859 5.387 4.179 1.962
```

Buscamos estimar los parámetros correspondientes a las direcciones medias haciendo uso de la información adicional que nos ofrece el siguiente orden circular parcial entre los mismos:

$$\left\{ \begin{array}{c} \phi_1 \\ \phi_2 \\ \phi_3 \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_4 \\ \phi_5 \end{array} \right\} \preceq \{\phi_6\} \preceq \left\{ \begin{array}{c} \phi_7 \\ \phi_8 \end{array} \right\} \preceq \left\{ \begin{array}{c} \phi_1 \\ \phi_2 \\ \phi_3 \end{array} \right\}$$

Este orden se representa como:

```
R> orderGroups <- c(1, 1, 1, 2, 2, 3, 4, 4)
```

El vector `orderGroups` tiene en la posición i el número del conjunto al que pertenece el elemento θ_i de `data`. Para obtener el CIRE mediante el algoritmo desarrollado en Rueda et al. (2009) usamos la función `CIRE` como sigue:

```
R> example1CIRE <- CIRE(cirdata, groups = orderGroups,
+  circular = TRUE)
```

Recomendamos guardar en un objeto los resultados para poder usarlos posteriormente, en este caso lo guardamos en `example1CIRE`. La salida de dicha función se muestra en pantalla de la siguiente manera:

```
R> example1CIRE
 Circular Isotonic Regression Estimator (CIRE):
    0.993 1.475 3.066
    5.056 3.066
    5.056
    5.056 0.993
 Sum of Circular Errors: SCE =  1.428
Invisible: Unrestricted circular means;
           these can be obtained via $cirmeans
```

Conseguido nuestro objetivo, tenemos los estimadores de las direcciones medias que satisfacen el orden circular parcial definido:

$$\left\{\begin{array}{l}\widetilde{\phi}_1 = 0.993\\ \widetilde{\phi}_2 = 1.475\\ \widetilde{\phi}_3 = 3.066\end{array}\right\} \preceq \left\{\begin{array}{l}\widetilde{\phi}_4 = 5.056\\ \widetilde{\phi}_5 = 3.066\end{array}\right\} \preceq \{\widetilde{\phi}_6 = 5.056\} \preceq \left\{\begin{array}{l}\widetilde{\phi}_7 = 5.056\\ \widetilde{\phi}_8 = 0.993\end{array}\right\}$$

donde $\widetilde{\mathbf{\Phi}} = (\widetilde{\phi}_1, \ldots, \widetilde{\phi}_n)'$ es el estimador circular de regresión isotónica (CIRE) de $\mathbf{\Phi}$ bajo el orden circular parcial dado. Se pueden obtener los resultados gráficos mediante `plot(example1CIRE)`. Por defecto se dibuja el CIRE, en caso de querer las medias circulares sin restringir se añadirá el argumento `option = ''cirmeans''` de la siguiente manera:

```
plot(example1CIRE, option = "cirmeans")
```

En este caso, al no haber réplicas las medias circulares sin restringir y los datos iguales.

Ejemplo 6.2. *Contraste de un orden circular dado (κ desconocido)*

En el caso de que el valor de κ sea desconocido necesitaremos tener réplicas para su estimación. Para ilustrar esta situación usaremos el conjunto de datos que se encuentra en el paquete con el nombre `datareplic`. Este conjunto tiene datos ficticios con réplicas contenidas en una matriz donde las columnas son las réplicas y cada fila tiene los datos de un elemento diferente. Tenemos 8 elementos correspondientes con los 8 parámetros desconocidos para los que queremos contrastar la hipótesis siguiente:

$$H_0: \quad \phi_1 \preceq \phi_2 \preceq \phi_3 \preceq \phi_4 \preceq \phi_5 \preceq \phi_6 \preceq \phi_7 \preceq \phi_8 \preceq \phi_1$$

$$H_1: \quad H_0 \text{ is not true.}$$

Procedemos a tomar los datos y a definir el orden circular a contrastar para ser introducido en el argumento `groups`.

```
R> data("datareplic")
R> orderGroups2 <- c(1:8)
```

Como desconocemos el valor de κ el argumento `kappa` quedará vacío. Pondremos `biasCorrect = TRUE` buscando que en la estimación se corrija el sesgo. Tenemos en la consola de R lo siguiente:

```
R> example2test <- cond.test(datareplic, groups = orderGroups2,
+  biasCorrect = TRUE)
R> example2test
 Circular Isotonic Regression Estimator (CIRE):
    1.223
    1.223
```

```
    1.223
    3.130
    4.194
    4.194
    5.541
    1.223
 Sum of Circular Errors: SCE =  1.532
Invisible: Unrestricted circular means;
           these can be obtained via $cirmeans
 pvalue =  0.0034
 kappa =  3.72
Kappa has been estimated
```

El resultado es el p-valor calculado como se detalla en Fernández et al. (2012). En este caso obtenemos un resultado de `p-value = 0.0034` por lo que se rechazaría la hipótesis nula y concluiríamos que los parámetros no satisfacen el orden circular especificado.

En este caso es interesante ver las medias circulares para cada elemento que se muestran con el comando `example2test$cirmeans`. El resultado es una lista donde cada conjunto del orden tiene un solo elemento.

```
R> round(unlist(example2test$cirmeans), digits = 3)
[1] 0.753 1.764 6.173 3.131 4.469 3.920 5.542 2.367
```

Puede verse que estas medias no verifican el orden representado en H_0.

Ejemplo 6.3. *Agregación de órdenes circulares*

En este ejemplo veremos como resolver un problema de agregación de órdenes circulares mediante el código implementado en R. Tenemos un conjunto de datos llamado `cirgenes` con las observaciones de 16 genes en 10 experimentos. El problema que planteamos en este ejemplo es encontrar el orden subyacente para los 6 primeros genes: *ssb*1, *cdc*22, *msh*6, *psm*3, *rad*21, *cig*2.

```
R> data("cirgenes")
R> datos <- cirgenes[,c(1:6)]
```

Para encontrar la solución se hace uso de la metodología expuesta en el Capítulo 3. Realizamos el procedimiento en dos pasos de forma que en un primer paso buscamos un orden circular agregado para estos 5 elementos mediante el enfoque del problema del viajante y la distancia con $\alpha = 3$. El código de R a ejecutar es el siguiente:

```
R> paso1 <- ACO(datos, method="TSP",
+                control.method="alpha3")
```

En el objeto denominado `paso1` se guarda el output de la función `ACO` que contiene el orden circular agregado en el elemento `$aggre_order`.

```
R> paso1$aggre_order
[1] 1 2 3 5 6 4
```

Se procede a ejecutar el paso 2 del procedimiento (*Circular Local Minimization*) mediante la función `CLM` para buscar posibles permutaciones de elementos que mejoren dicho orden.

```
R> paso2 <- CLM(datos, order0=paso1$aggre_order)
R> paso2
$order0
[1] 1 2 3 5 6 4
$msce0
[1] 0.007773727
$final_order
[1] 1 2 5 6 3 4
$bestsce
[1] 0.003361125
```

Se observa en la salida final que después de las permutaciones realizadas por el algoritmo CLM y la consecuente mejora del valor del MSCE, el orden circular agregado para estos 6 genes es $(1, 2, 5, 6, 3, 4)$.

Ejemplo 6.4. *Contraste de igualdad de órdenes circulares*

La función `eq.test` realiza un contraste de igualdad de órdenes circulares con la metodología expuesta en el Capítulo 4 como ilustramos a continuación.

Supongamos que, para los mismos datos del ejemplo anterior, los 6 primeros genes del conjunto `cirgenes` queremos contrastar si el orden circular es el mismo en los experimentos realizados con la técnica de *elutriación* (población 1) que con la técnica de *liberación de cdc* (población 2). Los experimentos 1, 4, 6 y 7 usaron *liberación de cdc*, mientras que los experimentos 2, 3, 5, 8, 9 y 10 usaron *elutriación*.

```
R> rownames(cirgenes)
 [1] "Oliva cdc" "Oliva elut1" "Oliva elut2" "Peng cdc" "Peng elut"
 [6] "Rust cdc1" "Rust cdc2" "Rust elut1" "Rust elut2" "Rust elut3"
```

Definimos los experimentos que proceden de cada población mediante el vector `populations` donde en la posición i está la población a la que pertenece el experimento i.

```
R> populations <- c(1, 2, 2, 1, 2, 1, 1, 2, 2, 2)
```

Entonces, ejecutamos el test:

```
R> equalitytest <- eq.test(cirgenes[,c(1:6)], popu=populations,
+                          method="TSP", control.method="alpha3",
+                          N=100)
1,2,3,4,5,...
pvalue=0.48
```

donde `N=100` indica que el procedimiento de remuestreo realizará 100 selecciones aleatorias. Podemos observar que va mostrando en pantalla el número de selección que se está ejecutando así como el p-valor final calculado, en este caso 0.48, por lo que no se puede rechazar que el orden circular para estos 6 genes es el mismo en ambas técnicas. Dicho orden circular estimado puede obtenerse de la siguiente manera:

```
R> equalitytest$global_order
[1] 1 2 5 6 3 4
```

En este caso además, como podemos ver en el resto de argumentos de salida, el coeficiente de confianza para el orden global es 42.42 % y el orden más frecuente es el mismo que el orden global. En la salida se encuentran también los órdenes circulares globales junto con el valor del estadístico, todo ello para cada selección aleatoria (`$allorders`). Esto queda recogido en una matriz que puede exportarse a un archivo con extensión .csv de manera tan sencilla como añadir en los argumentos de entrada `output=ruta` con la ruta completa donde colocar dicho archivo denominado *globalorders.csv*. La distribución de frecuencias para dichos órdenes se exporta al archivo denominado *frequencydist.csv*.

Capítulo 7

Conclusiones y trabajo futuro

In every end, there is also a beginning.
Libba Bray

Se podría decir que este trabajo genera dos tipos de aportaciones científicas, en el campo de la biología por un lado y en el campo de la inferencia estadística con restricciones, por el otro.

Las aportaciones biológicas tienen relación con el avance en el conocimiento de la función biológica de los genes y de las relaciones entre ellos. En concreto, a partir de los análisis de los datos de expresiones de los genes, se descubre un conjunto de 7 genes cuyo orden de activación en el ciclo celular se ha conservado desde las especies de levaduras, *S.pombe* y *S.cerevisiae* hasta los humanos. Así mismo, se descubre que entre ambas levaduras y entre *S.pombe* y los humanos hay un conjunto mayor de genes que conservan su orden de activación que entre *S.cerevisiae* y humanos. Lo que puede indicar que *S.pombe* sea un eslabón conector entre *S.cerevisiae*. Para poder llegar a estas conclusiones se han analizado series de expresiones de conjuntos de genes obtenidas en experimentos de diferentes laboratorios y se han comparado los órdenes de activación de los genes entre especies.

Las aportaciones metodológicas tienen que ver con el problema de comparación de órdenes circulares en base a observaciones de experimentos heterogéneos, y como parte de este, se resuelve el problema de la agregación de órdenes circulares. Ambos problemas, hasta lo que sabemos, inéditos en la literatura.

El problema de la agregación de ordenes circulares, se formula en esta tesis como un problema de optimización con una función objetivo que se define haciendo uso del CIRE. Para resolver el problema, se desarrollan y comparan métodos que utilizan tres niveles diferentes de procesar la información inicial: individual, por pares y por tripletas. La principal novedad en este sentido es el uso de las tripletas, que aparecen de forma natural para medir la asociación circular y son el instrumento básico que permite agregar la información de experimentos con diferentes puntos de inicio. Matemáticamente, la información de las tripletas se registra utilizando hipermatrices y la teoría de Hodge permite trasladar dicha información a una información más accesible sobre relaciones entre pares y a partir de esta se genera el orden circular más cercano. Esta técnica además de ser computacionalmente muy simple ofrece resultados, hasta lo que nosotros hemos comprobado, biológicamente interpretables. Además, tiene posibilidades de desarrollo aún no exploradas que tienen que ver con la elección de los valores iniciales para medir la relación entre tripletas en los experimentos y con la definición de índices que midan el nivel de inconsistencia del orden obtenido con los datos. Esto último es posible porque la teoría de Hodge formula el problema como el de minimizar una distancia sobre un subespacio y por tanto se puede utilizar el residuo como base para calcular dichos índices de una forma parecida a como se hace en el caso de la línea. Además de la propuesta basada en la teoría de Hodge, en esta memoria se diseña una técnica de agregación basada en el TSP que hace uso de las información de las intensidades de relación entre pares. A partir de los estudios numéricos realizados probamos que ofrece las mejores aproximaciones al óptimo en situaciones muy diversas. Como hemos puesto de manifiesto en este trabajo, la clave es

escoger, para definir la función objetivo del problema del TSP, una distancia adecuada para medir las relaciones entre pares. En concreto es determinante que la distancia elegida sea *dirigida*, es decir, que tenga en cuenta la dirección del círculo. De hecho, otra de las aportaciones relevantes de este trabajo es la definición de una distancia dirigida en el círculo que se demuestra que tiene varias propiedades deseables, en particular, garantiza el diseño de un algoritmo de resolución del problema de optimización que es computacionalmente eficiente.

El problema de agregación de órdenes, resuelve además el problema de estimación cuando se considera que el orden circular es un parámetro a estimar. A partir de aquí se resuelve el problema de contraste de hipótesis de la igualdad de órdenes circulares entre dos o más poblaciones. Por la naturaleza del problema se elige un procedimiento no paramétrico basado en la aleatorización de los experimentos entre poblaciones. La elección del estadístico test y el diseño del procedimiento de remuestreo es determinante para obtener una potencia razonable. En particular, el diseño elegido permite lidiar correctamente con las características del problema, a saber, muestras no balanceadas y experimentos heterogéneos.

Los nuevos métodos estadísticos y algoritmos diseñados en esta tesis se han implementado como parte de un paquete de R denominado **isocir** ("**iso**tonic inference for **cir**cular data") Barragán et al. (2014a) que se encuentra disponible en el CRAN para facilitar su ejecución por parte de otros investigadores.

Todos los métodos propuestos tienen aplicación a problemas que surgen en ámbitos diferentes del que aquí nos ocupa. Por ejemplo, la metodología desarrollada puede servir para dar solución a problemas relacionados con eventos a lo largo del ciclo circadiano (ciclo de un día que regula ritmos biológicos), Li et al. (2013), o de los ciclos de hormonas, Hendriks et al. (2014), donde es relevante la relación entre los momentos de máxima incidencia. También tienen

cabida en otros campos como en meteorología (Bowers et al. (2000)) donde puede ser de interés realizar comparaciones del orden de las direcciones del viento entre fenómenos atmosféricos o de manera similar cuando se analizan las direcciones de propagación de incendios (Albini y Forest (1976), Muñoz et al. (2013), García-Portugués et al. (2013)), el orden de dichas direcciones puede servir para realizar una clasificación de los incendios y más genéricamente en estudios donde el interés sea el orden de ocurrencia de las máximas incidencias de eventos cíclicos.

Los nuevos procedimientos que se presentan en esta tesis suponen un punto de partida para multitud de investigaciones. Uno de los más inmediatos es el estudio de las posibilidades todavía no exploradas en la técnica basada en la teoría de Hodge para la agregación de órdenes circulares. Esta técnica tiene un gran potencial por la flexibilidad de actuación en los distintos pasos de la misma e incluso en posibles propuestas novedosas de diferentes formas de proceder. Existe otro camino todavía sin estudiar que se trataría del uso de estas técnicas desarrolladas para el espacio circular para contribuir al avance de los métodos de agregación de rankings en la línea.

Con todo lo aquí expuesto podemos concluir que los métodos desarrollados en esta tesis suponen un avance tanto dentro del área estadística de los modelos circulares con restricciones de orden como en los diversos campos de aplicación donde son de interés para dar respuesta a múltiples cuestiones de relevancia.

Apéndices

Apéndice A

Definiciones básicas de la teoría de Hodge

Vemos brevemente aquella parte de la teoría de Hodge que es usada por Jiang et al. (2011) para la agregación de rankings y de la que hacemos uso en este trabajo para agregación de órdenes circulares (ver Sección 3.5). En la exposición simultaneamos notación propia de la teoría de grafos, del álgebra lineal y de la topología lo que puede ayudar a una mejor comprensión de los conceptos (ver Tabla A.1).
Comenzamos definiendo los conceptos más básicos que serán necesarios para las definiciones posteriores que culminan con el Teorema A.1 donde se presenta la descomposición de Helmholtz-Hodge, que se trata del resultado principal de esta teoría. Para terminar este apéndice se expone una visión matricial de la descomposición de Helmholtz-Hodge que es de utilidad para los métodos de agregación.

Sea $V = \{1, \ldots, n\}$ un conjunto de n elementos, entonces,

Definición A.1. *Un $k-$ simplejo es cualquier conjunto de $k+1$ elementos de V.*

Σ_k es el conjunto de todos los $k-$ *simplejos* en Σ.

Definición A.2. *Denominamos complejo simplicial al par $K = (V, \Sigma)$, donde Σ es la colección de subconjuntos de V, es decir $\Sigma = \bigcup_{k=0}^{n} \Sigma_k$.*

Sea $G = (V, E)$ un grafo donde V es el conjunto de nodos y E el conjunto de aristas entre pares de nodos. Se denota por T a la colección de triángulos con todas las aristas en E, es decir

$$T = \left\{ \{i, h, k\} \in \binom{V}{3} : \{i, h\}, \{h, k\}, \{k, i\} \in E \right\} \tag{A.1}$$

Entonces, se tiene que $\Sigma_0 = V$, $\Sigma_1 = E$ y $\Sigma_2 = T$.

Definición A.3. *Se denomina complejo k-clique al complejo simplicial de dimensión $k-1$, $K_G^k = (V, \Sigma_{k-1})$.*

Definición A.4. *Sea K_G aquel complejo k-clique de G donde k es máximo.*

Definición A.5. *Una cocadena de dimensión k (o* k*-cocadena) es una función $f : \Sigma_k \to \mathbb{R}$,*

$$f(i_{\sigma(0)}, \ldots, i_{\sigma(k)}) = signo(\sigma) f(i_0, \ldots, i_k),$$

para todo $\{i_0, \ldots, i_k\} \in \Sigma_k$ y toda permutación σ de dichos $k+1$ elementos, donde $signo(\sigma)$ es el signo de la permutación σ que viene dado por el determinante de la matriz que da lugar a dicha permutación.

En el caso que nos ocupa,

$$\begin{aligned} C^0 &= \mathbb{R}^n \\ C^1 &= \mathcal{A} = \{[X_{ij}] \in \mathbb{R}^{n \times n} : X_{ij} = -X_{ji}\} \\ C^2 &= \{[\psi_{ihk}] \in \mathbb{R}^{n \times n \times n} : \psi_{ihk} = \psi_{hki} = \psi_{kih} = -\psi_{ikh} = -\psi_{khi} = -\psi_{hik}\} \end{aligned} \tag{A.2}$$

Definición A.6. *El operador cofrontera $\delta_k : C^k \to C^{k+1}$ lleva una k-cocadena a una $(k+1)$-cocadena y se define como*

$$(\delta_k f)(i_0, i_1, \ldots, i_{k+1}) = \sum_{h=0}^{k+1} (-1)^h f(i_0, \ldots, i_{h-1}, i_{h+1}, \ldots, i_{k+1})$$

para todo $\{i_0, \ldots, i_{k+1}\} \in \Sigma_{k+1}$ y todo $f \in C^k$.

Dado un producto escalar $< \cdot, \cdot >_k$ en C^k,

Definición A.7. *El operador adjunto cofrontera* $\delta_k^* : C^{k+1} \to C^k$ *se define como*

$$< \delta_k f_k, g_{k+1} >_{k+1} = < f_k, \delta_k^* g_{k+1} >_k$$

donde $f_k \in C^k$ *y* $g_{k+1} \in C^{k+1}$.

En nuestro caso concreto el vector $\boldsymbol{s}$ es una $0-$cocadena, las matrices X e Y son $1-$cocadenas y la hipermatriz Ψ es una $2-$cocadena.

Definición A.8. *El operador gradiente* $\delta_0 : C^0 \to C^1$ *lleva una función de los vertices* $s : V \to \mathbb{R}$ *al flujos de aristas grad* s: $V \times V \to \mathbb{R}$ *a través de*

$$\delta_0(\boldsymbol{s})(i,h) = grad\ (\boldsymbol{s})(i,h) = s_h - s_i.$$

A un flujo de aristas que tiene esta forma se le denomina flujo gradiente.

Definición A.9. *El operador divergencia es menos el adjunto del operador gradiente,* $div := -\delta_0^*$, *donde* $\delta_0^* : C^1 \to C^0$, *por tanto se define como*

$$div\ (X)(i) = -\delta_0^*(X)(i) = \sum_h \omega_{ih} X_{ih}.$$

Definición A.10. *El operador rotacional* $\delta_1 : C^1 \to C^2$ *se define como*

$$\delta_1(X)(i,h,k) = X_{ih} + X_{hk} + X_{ki}$$

siendo $\delta_1 = curl$ *será equivalente a*

$$curl(X)(i,h,k) = \begin{cases} X_{ih} + X_{hk} + X_{ki} & si\ \{i,h,k\} \in T \\ 0 & si\ no \end{cases}$$

Definición A.11. *El flujo triangular* $\Psi : V \times V \times V \to \mathbb{R}$ *es una cocadena de dimensión 2, es decir* C^2 *(ver Definición A.5), que se representa mediante la hipermatriz* $\Psi \in \mathbb{R}^{n \times n \times n}$.

Definición A.12. *El operador adjunto del rotacional* $\delta_1^* : C^2 \to C^1$ *se define como,*

$$curl^*(\Psi)(i,h,k) = \delta_1^*(\Psi)(i,h,k) = \sum_{k=1}^{n} \psi_{ihk},$$

donde ψ_{ihk} *para cada* $i,h,k = 1,\ldots,n$ *son las componentes de la* $2-$*cocadena denominada* Ψ*.*

Vemos un esquema resumen de estos operadores:

$$C^0 \xrightarrow{\delta_0 \text{ (grad)}} C^1 \xrightarrow{\delta_1 \text{ (curl)}} C^2$$

$$C^0 \xleftarrow{\delta_0^* \text{ (grad*=-div)}} C^1 \xleftarrow{\delta_1^* \text{ (curl*)}} C^2$$

Teorema A.1. *Descomposición de Helmholtz-Hodge. Sea* G *un grafo y sea* K_G *su complejo clique, entonces* C^1 *admite la siguiente descomposición ortogonal*

$$C^1 = im(\delta_0) \oplus \left(\ker(\delta_1) \cap \ker(\delta_0^*)\right) \oplus im(\delta_1^*).$$

A continuación se expone una visión matricial de la descomposición de Helmholtz-Hodge. Sea $\mathcal{A} := \{X \in \mathbb{R}^{n\times n} : X^\top = -X\}$ el espacio de las matrices antisimétricas. Según la descomposición de Hodge (ver Figura A.1, extraída de Jiang et al. (2011)),

$$\mathcal{A} = \mathcal{M}_G \oplus \mathcal{M}_H \oplus \mathcal{M}_{T^\perp}.$$

El subespacio $\mathcal{M}_G \subseteq \mathcal{A}$ definido como,

$$\text{im}(\delta_0) = \text{im}(\text{grad}) = \mathcal{M}_G := \{X \in \mathbb{R}^{n\times n} : X_{ih} = s_h - s_i, s : V \to \mathbb{R}\}, \quad \text{(A.3)}$$

contiene las matrices que representan un orden en la línea o ranking según la siguiente regla:
$i_1 \leq,\ldots,\leq i_n \Leftrightarrow s_{i_1} \leq,\ldots,\leq s_{i_n}$. Aquellos rankings que surgen de esta manera se dice que son globalmente consistentes.

El subespacio $\mathcal{M}_T \subseteq \mathcal{A}$ definido como,

$$\ker(\delta_1) = \ker(curl) = \mathcal{M}_T := \{X \in \mathcal{A} : X_{ih} + X_{hk} + X_{ki} = 0 \,\forall\{i,h,k\} \in T\},$$

contiene las matrices que representan la ausencia de relación entre tripletas de elementos para todas las tripletas de V.

El subespacio $\mathcal{M}_H \subseteq \mathcal{A}$ contiene las matrices que representan tanto la ausencia de relación entre tripletas de elementos (ciclicidad local) como la información sobre la ciclicidad global entre todos ellos, es decir,

$$\mathcal{M}_H = \mathcal{M}_{G^\perp} \cap \mathcal{M}_T.$$

La última componente $\mathcal{M}_{T^\perp}$ representa las relaciones entre subconjuntos de elementos, es decir $X_{ih} + X_{hk} + \cdots + X_{pq} + X_{qi} \neq 0 \; \forall i, h, k, \ldots, p, q \in Z \subseteq V$, está componente se corresponde con, $\text{im}(\delta_1^*) = \text{im}(curl^*) = \mathcal{M}_{T^\perp}$.

Las relaciones entre componentes ortogonales correspondientes son,

$$\mathcal{A} = \mathcal{M}_G \oplus \mathcal{M}_{G^\perp}, \;\; \mathcal{A} = \mathcal{M}_T \oplus \mathcal{M}_{T^\perp}, \;\; \mathcal{M}_T = \mathcal{M}_G \oplus \mathcal{M}_H, \;\; \mathcal{M}_H = \mathcal{M}_T \cap \mathcal{M}_{G^\perp}.$$

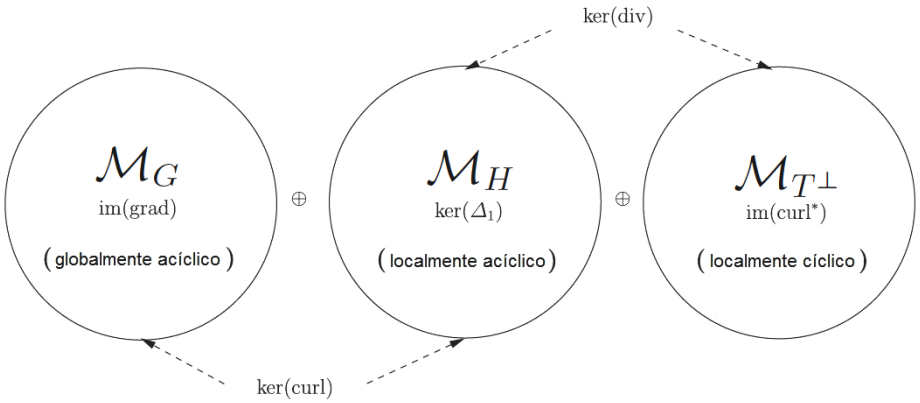


Figura A.1: Descomposición de Hodge para una cocadena C^1 en subespacios con sus diferentes denominaciones

En la Figura A.1 se representa la descomposición ortogonal a la conduce el Teorema A.1 en diferentes subespecios. El subespacio $\mathcal{M}_G$ es considerado globalmente acíclico porque no existe ningún subconjunto de elementos para los que $X_{ih} + X_{hk} + \cdots + X_{pq} + X_{qi} \neq 0$, $i, h, k, \ldots, p, q \in Z \subseteq V$. El subespacio

$\mathcal{M}_H$ es considerado localmente acíclico porque recoge la información sobre la ausencia de relación de las tripletas de elementos.
El subespacio $\mathcal{M}_{T^\perp}$ es localmente cíclico porque recoge la información sobre la relación entre todos los subconjuntos de cualquier tamaño de elementos.

Tabla A.1: Relación de conceptos básicos de la teoría de Hodge según la diferentes áreas

Teoría de grafos	Álgebra lineal	Topología	Ranking
Función en los vértices	Vector en $\mathbb{R}^n$	0-cocadena	Función de scores
Flujo de aristas	Matriz antisimétrica en $\mathbb{R}^{n\times n}$	1-cocadena	Ranking por pares
Flujo triangular	Hipermatriz antisimétrica en $\mathbb{R}^{n\times n\times n}$	2-cocadena	Ranking por tripletas

Apéndice B

Tablas de datos

- Tabla B.1: Datos *S. pombe* (34 genes Parte I).
- Tabla B.2: Datos *S. pombe* (34 genes Parte II).
- Tabla B.4: Datos *S. cerevisiae* (34 genes Parte I).
- Tabla B.4: Datos *S. cerevisiae* (34 genes Parte II).
- Tabla B.5: Estimadores sin restringir de la máxima expresión usando RPM en los 11 genes ortólogos a las 3 especies.

Tabla B.1: Datos *S.pombe* (34 genes Parte I)

Experimentos \ Genes	*ace2*	*cdc18*	*mik1*	*hhf1*	*hta2*	*fkh2*	*klp5*	*cig2*	*plo1*	*slp1*	*rad21*	*mcp1*	*mid2*	*chs2*	*sid2*	*eng1*	*hht3*	*h3_3*
Oliva cdc	5.6941	0.0014	5.7685	0.6836	1.1237	6.2108	1.2531	–	0.7512	5.4572	5.4413	4.9349	0.0234	0.0373	4.0369	6.0733	1.087	–
Oliva elut1	1.9283	1.8135	–	–	4.1017	0.843	–	–	3.3258	–	1.6942	4.2961	2.7998	0.3722	0.6215	2.7495	4.7873	4.8162
Oliva elut2	4.858	5.3213	6.199	1.0577	1.0249	3.6002	–	5.6399	5.0901	–	3.8476	2.4665	0.067	2.6417	4.8342	6.0967	0.9978	0.6728
Peng cdc	1.586	2.2256	1.7844	4.9202	5.0942	3.5096	3.0923	4.2484	0.6264	1.9551	2.9528	1.4944	3.7572	0.8624	4.2819	3.1983	5.149	–
Peng elut	4.3342	4.7114	5.3195	5.0629	5.2681	4.1704	4.1103	5.1237	3.3461	–	5.0332	5.2824	5.5535	4.4589	4.1325	5.0875	5.2609	5.2411
Rustici cdc1	6.0438	0.1251	0.5101	3.1787	3.145	6.1417	1.7776	0.0226	0.6655	1.3516	0.5658	0.1653	0.1342	0.7386	6.2692	1.6063	3.3921	3.4994
Rustici cdc2	6.2688	0.042	0.123	3.8413	3.7003	2.5596	3.3358	4.9447	1.0103	1.9827	1.6808	2.132	3.02	2.9872	3.3858	1.5691	4.0222	4.0755
Rustici elut1	1.5604	2.2753	2.5072	2.8164	2.8455	1.9516	1.7996	1.9721	1.1918	1.8311	2.2609	0.8546	1.8698	1.9917	–	2.4815	3.2324	3.1377
Rustici elut2	1.8271	1.6109	0.8609	2.9949	2.8002	1.4834	2.3849	1.6013	1.0148	1.7474	3.2001	0.5398	2.183	2.2423	1.643	1.9363	3.3776	2.9422
Rustici elut3	3.2471	2.907	4.0565	–	3.4699	3.1406	3.7113	2.5162	2.8796	2.3355	3.5536	1.5715	3.4959	2.2301	3.6768	3.1425	3.8845	4.4331

Tabla B.2: Datos *S.pombe* (34 genes Parte II)

Experimentos \ Genes	*cdc15*	*htb1*	*pht1*	*fin1*	*mob1*	*mrc1*	*msh6*	*myo3*	*pol1*	*pol2*	*SPAC1705.03C*	*rhp51*	*ssb1*	*cdc22*	*psm3*	*rgs1*
Oliva cdc	0.2263	–	5.7956	1.0644	5.32	6.0511	0.1352	0.7407	0.2922		1.6268	4.7392	0.0483	–	0.7855	2.1617
Oliva elut1	–	–	–	2.1228	3.19	–	–	1.4864	4.1269	–	2.0195	0.7854	2.8574	–	2.0386	0.5739
Oliva elut2	5.2881	1.1463	1.5073	3.3911	6.1698	5.6791	5.4055	3.297	6.0462	–	5.3788	1.3642	5.9642	0.5357	6.0886	5.5035
Peng cdc	2.1831	4.9116	4.7997	2.6099	4.1799	2.8583	4.3594	2.4036	4.0264	0.8493	3.3421	2.4669	1.1462	3.9575	3.9096	2.1137
Peng elut	4.1506	5.2538	–	4.712	4.0864	3.4752	5.1672	2.6313	4.8567	6.2447	3.4504	5.6088	3.5353	5.0193	4.1203	0.3062
Rustici cdc1	0.9487	2.7848	2.7734	1.8663	2.1113	6.2607	0.1077	0.7928	1.8776	–	1.1243	1.42	1.0582	0.9214	0.0271	0.6554
Rustici cdc2	2.0258	3.7731	3.788	0.8399	1.4539	2.3635	0.6337	2.9955	2.1179	3.5531	3.5219	2.1651	2.1438	1.4154	3.8309	3.9822
Rustici elut1	1.7315	2.8675	2.0647	2.3862	1.7974	3.1896	2.0292	1.7573	2.1879	1.8267	1.7977	1.6297	–	3.1637	–	1.1897
Rustici elut2	1.6508	2.7407	2.7257	2.165	3.2086	2.286	0.8307	2.0154	2.8142	–	1.4961	5.1016	3.3002	–	–	2.5864
Rustici elut3	3.3465	3.5894	3.9925	3.9163	3.185	2.9673	3.8828	–	4.0611	–	3.0354	3.2151	2.3721	–	3.75	4.7575

Tabla B.3: Datos *S.cerevisiae* (34 genes Parte I)

Genes / Experimentos	*SWI5*	*CDC6*	*SWE1*	*HHF1*	*HTA2*	*FKH1*	*KIP3*	*CLN2*	*CDC5*	*CDC20*	*MCD1*	*ASE1*	*BUD4*	*CHS2*	*DBF2*	*DSE4*	*HHT1*
Cho	–	2.6313	2.5689	3.0205	3.4359	3.0326	3.5034	1.6533	–	4.3966	–	4.6858	–	4.0394	4.6094	6.1875	–
Lichtenburg	4.0241	3.8237	–	1.247	2.0869	–	4.0002	–	5.0917	5.3424	2.9346	5.1929	–	4.6788	4.9828	–	2.4073
Pramilla30	4.4452	1.0683	3.0287	3.2772	3.3231	3.9774	3.7906	2.5863	4.8321	5.5761	2.5793	5.0798	4.5045	4.808	0.4138	5.9372	3.4916
Pramilla38	4.3735	1.8843	2.8864	3.5068	3.7854	4.0768	4.2793	2.6634	4.6247	5.5172	2.5613	5.3721	4.6404	4.7619	0.5854	2.6942	3.6616
Spellman alpha	4.3098	2.4911	2.7325	–	3.6375	3.738	3.9874	2.5127	4.4629	–	2.979	4.4201	4.3316	3.2701	5.76	0.1013	3.5579
Spellamn cdc	5.0686	0.8392	2.6909	3.5405	3.8744	4.4881	4.2105	2.7518	5.5488	6.0639	2.5083	5.4247	5.254	5.3986	0.3716	2.7657	3.0139

Tabla B.4: Datos *S.cerevisiae* (34 genes Parte II)

Experimentos \ Genes	*HHT2*	*HOF1*	*HTB2*	*HTZ1*	*KIN3*	*MOB1*	*MRC1*	*MSH6*	*MYO1*	*POL1*	*POL2*	*PST1*	*RAD51*	*RFA1*	*RNR1*	*SMC3*	*SST2*
Cho	2.2543	4.4105	3.05	2.3117	4.1053	4.5992	1.9462	2.033	4.3941	2.0856	1.6943	3.1014	1.7635	1.6435	1.4512	2.3644	1.2787
Lichtenburg	3.3366	4.9693	3.5319	–	–	3.1636	–	2.7073	3.2646	2.7605	1.9865	0.8506	2.856	1.6883	0.8587	2.9712	5.1997
Pramilla30	3.3833	4.9505	3.3687	2.5853	5.7109	5.1858	2.5287	2.5455	4.6696	2.982	2.7467	0.9581	2.1358	2.3889	–	2.7142	2.9048
Pramilla38	3.6451	4.908	3.2778	3.4932	5.3336	5.1097	2.8085	2.3902	4.8854	2.7921	2.7235	0.8357	2.4926	2.5433	2.4455	2.6578	2.8915
Spellman alpha	3.7083	4.6752	3.534	3.6791	0.847	4.7952	2.6023	2.6268	4.9998	2.3903	2.344	6.19	2.436	2.1759	2.6463	2.768	2.8743
Spellamn cdc	3.2281	5.0963	3.8346	–	5.6826	5.1074	2.321	–	4.668	0.5398	2.4471	2.0674	2.9231	2.3153	–	2.1899	5.1544

Tabla B.5: Estimadores sin restringir de la máxima expresión usando RPM en los 11 genes ortólogos a las 3 especies.

Especie	Experimento	*ZNF367*	*CDC6*	*PKMYT1*	*HIST2H4B*	*H2AFX*	*FOXM1*	*KIF10*	*CCNB1*	*PLK1*	*CDC20*	*RAD21*
S. pombe	1- Oliva et al., 2005 cdc	5.69	0.00	5.77	0.68	1.12	6.21	1.25	1.10	0.75	5.46	0.61
S. pombe	2- Oliva et al., 2005 elut1	1.93	1.81	0.59	4.34	4.10	0.84	1.25	2.99	3.33	1.49	3.26
S. pombe	3- Oliva et al., 2005 elut2	4.86	5.32	6.20	1.06	1.02	3.60	4.98	5.64	5.09	5.61	3.85
S. pombe	4- Peng et al., 2005 cdc	1.59	2.23	1.78	4.92	5.09	3.51	3.09	4.25	0.63	1.96	2.95
S. pombe	5- Peng et al., 2005 elut	4.33	4.71	5.32	5.06	5.27	4.17	4.11	5.12	3.35	3.93	5.03
S. pombe	6- Rustici et al., 2004 cdc1	6.04	0.13	0.51	3.18	3.15	6.14	1.78	0.02	0.67	1.35	0.57
S. pombe	7- Rustici et al., 2004 cdc2	6.27	0.04	0.12	3.84	3.70	2.56	1.50	3.81	1.01	1.98	1.68
S. pombe	8- Rustici et al., 2004 elut1	1.57	2.28	2.51	2.82	2.85	1.95	1.80	1.97	1.19	1.83	2.26
S. pombe	9- Rustici et al., 2004 elut2	1.83	1.61	0.86	2.99	2.80	1.48	2.38	1.60	1.01	1.75	3.20
S. pombe	10- Rustici et al., 2004 elut3	3.25	2.91	4.06	3.67	3.47	3.14	3.71	2.52	2.88	2.34	3.55
S. cerevisiae	1- Cho at al. 1998	3.74	3.20	2.57	3.00	3.44	3.03	3.50	1.65	4.30	4.40	1.99
S. cerevisiae	2- De Lichtenberg et al., 2002	4.02	0.10	2.15	1.25	2.09	3.66	4.00	1.15	4.68	5.34	2.93
S. cerevisiae	3- Pramila et al., 2006 30	4.45	1.07	3.03	3.28	3.32	3.98	3.79	2.59	4.83	5.58	2.58
S. cerevisiae	4- Pramila et al., 2006 38	4.37	1.88	2.89	3.51	3.79	4.08	4.28	2.66	4.62	5.52	2.56
S. cerevisiae	5- Spellman et al., 1998 alpha	4.31	2.49	2.73	3.34	3.64	3.74	3.99	2.51	4.46	5.86	2.98
S. cerevisiae	6- Spellman et al., 1998 cdc	5.07	0.84	2.69	3.54	3.87	4.49	4.21	2.75	5.55	6.06	2.51
Humanos	1- Whitfield et al., 2002 Thynoc	2.78	1.96	2.69	1.33	4.05	0.36	0.78	6.04	0.69	5.12	0.83
Humanos	2- Whitfield et al., 2002 Thy1	4.62	3.41	3.48	3.48	3.54	1.47	1.11	5.35	0.16	5.82	0.43
Humanos	3- Whitfield et al., 2002 Thy2	4.43	4.25	4.85	4.55	5.03	1.40	1.18	0.16	0.12	6.13	0.24
Humanos	4- Whitfield et al., 2002 Thy3	3.11	3.60	4.47	4.21	5.13	0.34	0.48	0.11	0.41	5.88	1.46

Apéndice C

Procedimiento backward de selección de genes

- Procedimiento *backward* paso a paso para las 3 especies (Tabla C.1):
  - Orden circular global (7 genes):
    $ZNF367\ CDC6\ PKMYT1\ HIST2H4B\ H2AFX\ KIF10\ FOXM1$.
  - Últimos 3 genes: $CDC6$, $HIST2H4B$, $KIF10$.
- Procedimiento *backward* paso a paso para *S.pombe-S.cerevisiae* (Tabla C.2):
  - Orden circular global (10 genes):
    $ZNF367\ CDC20\ CDC6\ CCNB1\ RAD21\ PKMYT1\ HIST2H4B$
    $H2AFX\ FOXM1\ KIF10$.
  - Últimos 3 genes: $ZNF367$, $CCNB1$, $HIST2H4B$.
- Procedimiento *backward* paso a paso para *S.pombe*-Humanos (Tabla C.3):
  - Orden circular global (10 genes):
    $ZNF367\ CDC6\ PKMYT1\ HIST2H4B\ H2AFX\ CCNB1\ RAD21$
    $PLK1\ KIF10\ FOXM1$.
  - Últimos 3 genes: $CDC6$, $KIF10$, $FOXM1$.

- Procedimiento *backward* paso a paso para *S.cerevisiae*-Humanos (Tabla C.4):
  - Orden circular global (7 genes):
    $ZNF367\,CDC6\,PKMYT1\,HIST2H4B\,H2AFX\,KIF10\,FOXM1$.
  - Últimos 3 genes: $H2AFX$, $KIF10$ ($CDC6$ ó $HIST2H4B$).

En las siguientes tablas, **OMF** es el Orden Más Frecuente en la distribución de órdenes del procedimiento de selección aleatoria y **OG** es el Orden Global circular estimado para el conjunto de todas las poblaciones.
$C(\widehat{O}_G)$ es el coeficiente de confianza para el orden circular global estimado y $C(\widehat{O}_P)$ es el coeficiente de confianza para el orden circular parcial que se genera a partir del orden circular global estimado.

Tabla C.1: Procedimiento *backward* paso a paso para las 3 especies

S.pombe - *S.cerevisiae* - Humanos									
Número de genes	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**
p-valor	0,002	0,0159	0,03696	0,03796	**0,2977**	0,42457	0,4975	0,36663	0,846
MSCE	0,06274451	0,04195264	0,03647254	0,02259195	0,01321005	0,01334834	0,01166004	0,0087984	0,003731633
OMF=OG	NO	SI	SI	SI	SI	SI	SI	SI	SI
C($\widehat{O}_G$)	6,9	45,9	45,6	48,8	64,6	64,8	76,7	74,3	97,3
C($\widehat{O}_P$)	19,7	57,2	50,3	56,5	90,3				
Gen eliminado		*CDC*20	*CCNB*1	*RAD*21	*PLK*1	*H2AFX*	*FOXM*1	*PKMYT*1	*ZNF*367

Tabla C.2: Procedimiento *backward* paso a paso para *S.pombe-S.cerevisiae*

S.pombe - Cerevisiae									
Número de genes	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**
p-valor	0,0119	**0,05794**	0,73526	0,881	0,984	0,904	0,906	0,903	0,977
MSCE	0,0298	0,02309524	0,03447382	0,02991312	0,02612439	0,03016329	0,02066734	0,0153265	0,001854516
OMF=OG	SI	SI	SI	SI	SI	SI	SI	SI	SI
C($\widehat{O}_G$)	41,1	33,4	41,8	82,5	81,1	77	67,1	73,2	99,9
C($\widehat{O}_P$)		59	90,1	98	97,4				
Gen eliminado		*PLK*1	*PKMYT*1	*KIF*10	*FOXM*1	*H2AFX*	*CDC*20	*CDC*6	*RAD*21

Tabla C.3: Procedimiento *backward* paso a paso para *S.pombe*-Humanos

S.pombe - Humanos									
Número de genes	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**
p-valor	0,0239	**0,186813**	0,3996	0,533466	0,50749	0,7582	0,6473	0,958	0,968
MSCE	0,03517	0,02981	0,0244	0,02076	0,02035	0,01239	0,008698	0,006907	0,000785513
OMF=OG	SI	SI	SI	SI	SI	SI	SI	SI	SI
C($\widehat{O}_G$)	10	15,3	22,2	23,4	25,6	44,6	54,6	70,2	89,8
C($\widehat{O}_P$)		39,8	49,2	49,4	51,4				
Gen eliminado		*CDC*20	*CCNB*1	*ZNF*367	*H2AFX*	*HIST2H4B*	*PKMYT*1	*PLK*1	*RAD*21

Tabla C.4: Procedimiento *backward* paso a paso para *S.cerevisiae*-Humanos

S.cerevisiae - Humanos									
Número de genes	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**
p-valor	0,0009	0,003996	0,006993	0,011988	**0,13086**	0,172827	0,35764	0,985	1
MSCE	0,06894	0,06786	0,03758	0,02714	0,009175	0,007191	0,001387	0,0006944	0
OMF=OG	SI	SI	SI	SI	SI	SI	SI	SI	SI
C($\widehat{O}_G$)	13,8	13,1	28	34	43,2	58,5	75	100	100
C($\widehat{O}_P$)	23,2	24,8	49	64,5	84				
Gen eliminado		*CCNB*1	*CDC*20	*PLK*1	*RAD*21	*PKMYT*1	*ZNF*367	*FOXM*1	*CDC*6 ó *HIST*2*H*4*B*

Apéndice D

English summary of this Ph.D thesis

1. Introduction

1.1. Motivation and objectives

Interest on biological systems is increasing in the last decades. DNA sequencing has played an essential role on this interest by discovering which part of DNA belongs to each gene. Nowadays, one of the key challenges is to know the biological function of each gene and the relations among them. Genomics studies on this line of research are interdisciplinary and they mainly combined biological, statistical and informatic knowledge. The contribution made by this thesis is on statistical methodology. We have developed new procedures as part of the analysis of circular data under order restrictions. This methods have been motivated by biological issues, in particular in the analysis of gene expression data.

The analysis of gene expressions has numerous goals depending on the study. For instance, an oncologyst may be interested in the expression patterns of genes related to some kind of cancer in order to detect it and to prescribe the appropiate treatment.
In this work, we focus on identifying those genes whose biological functions ha-

ve been conserved across evolution. We consider data of cell-cycle genes which are common (orthologs) to different species (*S.pombe*, *S.cerevisiae* and Humans). It is known (Bähler (2005)) that the biological function of a gene is closely related to its peak expression in the cell cycle. Then, if a set of genes keep the same order among their peak expressions, we may say that their biological functions are conserved evolutionarily.

The peak expression data are circular data as they can be seen as points in a circle representing the cell cycle. Circular data need specific methodology to deal properly with the circular geometry. The main references in this field are Mardia y Jupp (2000) and Fisher (1993). A revision of the basic methodology for circular data is done in the preliminaries (Section 2.2).
The other relevant issue is the order among genes in cell cycle. To deal with this question we use specific techniques. Consequently, ordered restricted inference techniques are also summarized in the preliminaries. From the union of these two fields, a new area called circular ordered restricted inference emerges. The first references in this new area are Rueda et al. (2009) and Fernández et al. (2012). In the first work, the circular isotonic regression estimator is defined while in the second work the problem of testing a given circular order is solved through a conditional test. The main results in this area are also summarized in preliminaries.

The comparison of peak expression orders among cell-cycle genes of different species can be solved through a testing problem. There is no procedure in the literature to solve a problem of these characteristics, then in this thesis we develop the suitable methodology as part of the circular ordered restricted inference. In the development of the methodology to solve the testing problem, the problem of circular order estimation from the information of heterogeneous experiments arised. This problem can be formulated as the search of the circular order that better represents all the available information, which is the same idea used to address the aggregation of orders in the Euclidean space. As

happened with the testing problem, the aggregation of circular orders has no specific methodology in the literature to be solved. The closest problem is rank aggregation. A revision of the methods to solve rank aggregation is presented in preliminaries. In fact, the adaptation to the circular geometry of some of these methods is our first attempt to solve the aggregation of circular orders.

All the approaches developed in this thesis to solve aggregation of circular orders are presented in Chapter 3 of this work. Some of them arised from original ideas from other areas. There are two main novel approaches. The first idea uses the traveling salesman problem (TSP) as an approximation to the aggregation problem. TSP is usually encountered in problems of operations research and logistic, with the idea of searching a tour of minimum length along a set of locations. The second novel method uses Hodge theory which comes from algebraic topology and is being recently applied to data analysis. We present briefly in preliminaries the procedure which uses Hodge theory to solve rank aggregation in the line, while the main concepts of Hodge theory are shown in the Appendix A. The different options inside each approach produce a set of methods which are evaluated through simulations and numerical examples. The results from the comparison reveal the main advantages and disadvantages and as a result, TSP is the winner approach.

Regarding the testing problem of this thesis, we present in Chapter 4 a test for the equality of circular orders among different (two or more) populations. We propose a non-parametric approach which takes into account unbalanced samples and differences in the characteristics of the populations apart from location. Due to the peculiarities of the problem, a clasical permutation test does not work properly. Then, we develop a specific randomization procedure which is bootstrap like and we propose a set of statistic tests. We validate the proccess through a simulation study that takes into account different settings under the null and the alternative hypotheses. The final procedure presents the best results regarding the power of the test. There are other outputs of interest

apart from the pvalue, such as the estimate of the global circular order and the confidence coefficient of that estimator.

The implementation of the methodology has been an important part of the research. We have chosen R language as it is free software which helps to the divulgation and facilitates the usage by other researchers. We have developed an R package called **isocir** (**iso**tonic inference for **cir**cular data), Barragán et al. (2013), which contains all the new methodology related to circular ordered restricted inference (the new methods developed in this thesis as well as the basis in Rueda et al. (2009) and Fernández et al. (2012)). Some algorithms have required the use of C++ to improve the computational efficiency. Then, we have taken advantage of the object `SEXP` and the function `.Call` in order to maintain the interface with the user in R so that the software we present is user-friendly. Last version of the package is available in the CRAN: http://cran.r-project.org/web/packages/isocir/.

As we have commented, the methodology presented here was motivated by a problem encountered in the analysis of gene expression data. Then, we solve that problem in Chapter 5 with the final aim to identify the maximum set of cell-cycle genes whose order is conserved along different species. We have analized data from two kinds of yeasts (*S.pombe* and *S.cerevisiae*) and Humans. We have discovered a set of 7 cell-cycle genes which follow the same circular order in the 3 species. This improves the knowledge of the biological function as well as the relation among genes along evolution.

1.2. Contributions of this thesis

The novel methodology presented in this thesis is a contribution to the area of circular ordered restrictions inference, mainly in the following:

- Methods of aggregation and estimation of circular orders.
- Procedure to test the equality of circular orders in different populations.

- Resolution using the statistical methodology developed of a biological problem related to the analysis of gene expression data.
- Implementation of all the methodology in R language, gathered in a package called isocir which is user-friendly.

2. Preliminaries

In this chapter, we briefly present a revision of the main areas used in this work to develop the new methods. We make a revision of the methods to analyze circular data (Mardia y Jupp (2000), Fisher (1993)), we summarize the main concepts of Ordered Restricted Inference (Robertson et al. (1988), Silvapulle y Sen (2005)), we present the basis of Circular Ordered Restricted Inference (Rueda et al. (2009), Fernández et al. (2012)) and finally, we give an overview to the Rank Aggregation problem (Borda (1781), Dwork et al. (2001a), Jiang et al. (2011)).

2.1. Notation and terminology

The datasets used in this work contain angular observations from n circular variables whose parameters of interest are their mean directions. Let $\mathbf{\Phi}_s = (\phi_{1s}, \ldots, \phi_{is}, \ldots, \phi_{ns})'$, $s = 1, \ldots, S$, $i = 1, \ldots, n$ be the vector of population mean directions where ϕ_{is} is the mean direction of the variable i in the population s.
Let $\mathbf{\Theta_s} = (\mathbf{\Theta}_{1s}, \ldots, \mathbf{\Theta}_{js}, \ldots, \mathbf{\Theta}_{p_s s})'$, $j = 1, \ldots, p_s$, $s = 1, \ldots, S$ be the observed set of data of p_s experiments from population s, where the vector $\mathbf{\Theta_{js}} = (\theta_{1js}, \ldots, \theta_{ijs}, \ldots, \theta_{njs})'$ contains the observations of the n variables in experiment j of population s. We denote as $p = \sum_{s=1}^{S} p_s$ to the total number of experiments.

In the application of interest, each population is a species where the variables are the moments of peak expression of the n cell-cycle genes considered in the analysis. We have circular data as those moments of peak expression

are angles located in the circumference whose pole is the initial point of the experiment. As one of the issues in this work is the order among genes by using the aggregation of circular orders and the corresponding problem in the Euclidean space deals with the concept of order among elements, we use *elements* to speak about the genes to be ordered.
Moreover, in order to simplify, we ommit the subindices s and j when there is no possibility of misunderstanding.

We denote as $\mathcal{O}$ to the set of all possible orders among n elements. Then, if $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ is a vector of angular observations, $\eta_i \in [0, 2\pi)$, $i = 1, \ldots, n$. The circular order followed by the elements of $\boldsymbol{\eta}$ is denoted as $\boldsymbol{O_\eta} = (o_1, \ldots, o_l, \ldots, o_n)' \in \mathcal{O}$, where $o_l = i$ if the element i is located in the l position in the order. In this case, we can say that $\boldsymbol{\eta}$ verifies the circular order $\boldsymbol{O_\eta}$ and this is denoted as $\boldsymbol{\eta} \circlearrowleft \boldsymbol{O_\eta}$. In relation to the aggregation of circular orders, there is some interest in the positions of the elements in the order. Then, we denote the vector of positions as $\boldsymbol{T_\eta} = (\tau_1, \ldots, \tau_n)'$ where $\tau_i = k$, if the element i represented by η_i is located in the position k of the circular order $\boldsymbol{O_\eta}$ (i.e. $o_k = i$). As an example, if $\boldsymbol{\eta} = (2.47, 0.56, 4.92, 1.23, 6.1)$, then, $\boldsymbol{O_\eta} = (2, 4, 1, 3, 5)'$ and $\boldsymbol{T_\eta} = (3, 1, 4, 2, 5)'$.

Due to the cyclic characterics of this application and other possible future applications, the circular orders defined are invariant to the rotation, i.e., $\boldsymbol{O}_\eta = (o_1, \ldots, o_l, \ldots, o_n)' \equiv \boldsymbol{O}_{\boldsymbol{\eta}}^{\boldsymbol{l}} = (o_l, \ldots, o_n, o_1, \ldots, o_{l-1})', \forall l \in \{1, \ldots, n\}$. Then, the circular order among the elements of $\boldsymbol{\eta}$ can be denoted also as $\eta_1 \leq \cdots \leq \eta_n \leq \eta_1$. Note that as a consequence of this equivalence relation, $\#\mathcal{O} = (n-1)!$.

3. Aggregation of Circular Orders

In this chapter, several techniques are proposed to deal with the problem of obtaining the circular order among n elements which is closest to a set of heterogeneous data with p experiments. Although there exist a huge literature

in ranking aggregation for Euclidean data, the problem is unexplored in the circular setting. We propose the use of a distance d_1 between a set of circular data and a circular order called MSCE (Mean Sum of Circular Errors) and the approach through the following optimization problem,

$$\widetilde{\boldsymbol{O}} = \arg\min_{\mathbf{O}\in\mathcal{O}} d_1(\boldsymbol{\Theta}, \mathbf{O}) = \arg\min_{\mathbf{O}\in\mathcal{O}} \sum_{j=1}^{p} \omega_j SCE(\boldsymbol{\Theta}_j, \widetilde{\boldsymbol{\Theta}}_{\boldsymbol{j}}^{(\mathbf{O})}). \tag{D.1}$$

As this problem cannot be solved in polinomical time (it is NP-hard), we have developed a two-steps procedure. In the first step we propose to use one of the several methods of aggregation of circular orders that we present here. As second step, we have designed a local-search algorithm to improve the solution in terms of the MSCE which is called Circular Local Minimization (CLM) and it can be applied to the solution of any aggregation method.

In this work, we present different techniques of aggregation of circular orders by using three levels of information. One may use the raw circular data (θ_{ij}), or convert such scores into pairwise information (Y_{ih}^j) measuring the degree of preference of the element i over the element h, or use triplewise information (ϕ_{ihk}^j) measuring the degree of circular preference of the triple (i, h, k).

Most techniques have different possibilities. This derives in different methods that share the same procedure. As a first approach, we have extended some ranking aggregation methods developed for Euclidean data to cope with the geometry of the circle. The simplest method is called *Naive* and it consists of taking the circular order closest to the data among all the circular orders generated by the experiments. Circular Borda is a technique based on the Borda method (Borda (1781)) which is widely known and studied in the line and is even still used for rank aggregation (Baltrunas et al. (2010), Mekonnen (2014)). The adaptation to the circle makes this technique computationally expensive due to the need of rotations to fix the same initial point in all experiments. There is another technique with good behaviour in the Euclidean space which is based on Markov chains. Its corresponding circular version maintains the sa-

me procedure (Dwork et al. (2001a)) apart from the definition of the transition probabilities.

We have developed two original proposals whose main idea arises from the circular geometry itself. One approach is 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 problem of searching the shortest tour in a graph where the vertex are the items to be ordered, the edges connect these items and the lengths of the edges measure pairwise relationships. One of the main issues of this approach that has been deeply study in this work is the suitable definition of the lengths of the edges.

The last technique proposed is based on Hodge theory (Jiang et al. (2011)). Although Hodge theory has been proposed in the literature to derive algorithms for rank aggregation using pairwise information, what we propose here is pioneer in dealing with the third level of information (triplewise relationships). In fact, triplewise relations seem to be the natural way to take into account the information from circular data since 3 is the minimum number of elements to be univocally ordered in the circle. Some theoretical results (Theorem 3.2) are developed under Hodge theory to prove that the proposed algorithm solves the optimization problem defined in $\mathbb{R}^{n\times n\times n}$ to approach (D.1).

Several interesting examples are included in this chapter to illustrate the weaknesses and strengths of the techniques and a very extensive simulation study is conducted to compare the different methods.

From the analyses of the simulations, we can conclude that the best technique is the one based on TSP. Moreover, more specific conclusions for each technique are summarized in the following.

- ***Naive* method**

  ⌣̈ It is an intuitive and simple method. It works properly in cases where

the number of elements is low in relation to the number of experiments and for medium or high values of κ.

☹ There are several cases where the solution given by this method is far from the optimum and other cases where the solutions are not biologically relevant. It can be computationally expensive.

- **Circular Borda technique**

  ☺ It is a simple procedure, easy to figure out.

  ☹ It is computationally expensive. The cost of execution increases exponentially with the number of elements. In some cases, such as in the application problem of this thesis, the solution is not interpretable.

- **Markov chains technique**

  ☹ In the majority of the cases, the solution is really far from the optimum. Moreover, the total execution time is high as the second step of the procedure (CLM) lasts a lot due to the bad solution given by the first step with this technique.

- **TSP based technique**

  ☺ The solution given by this technique is the best approximation to the optimum in terms of MSCE if the lengths of the edges are properly defined. The solution is biologically relevant in the cases we have studied. The time of execution is reasonable even for large values of n. Moreover, this techniques allows user intervention in the relation between time and MSCE value through the choice of the value of a parameter in the resolution procedure.

- **Hodge theory based technique**

  ☺ This technique gives a procedure that allows a simple calculation of the aggregated circular order. So, it is the most efficient computationally as

first step method. The solutions are biologically relevant. This technique has broad possibilities of future research. New definitions of the measure of triplewise relationships and the definition of a measure of uncertainty of the resulting order can be explored.

∴̈ In some cases, the MSCE of the aggregated circular order is slightly higher than the result from other techniques.

4. Testing equality of circular orders

In this chapter, we present a test based on a randomization procedure and a standardized statistic for the following hypotheses:

$$\begin{aligned} &H_0 : \boldsymbol{O}_1 = \boldsymbol{O}_2 = \ldots = \boldsymbol{O}_S = \boldsymbol{O}_G \\ &H_1 : H_0 \text{ not true.} \end{aligned} \tag{D.2}$$

where $\boldsymbol{O}_G$ is the unknown circular order common to all S populations and $\boldsymbol{O}_s$, $s = 1, \ldots, S$ is the circular order underlying in population s among the parameters $\boldsymbol{\Phi}_s = (\phi_{1s}, \ldots, \phi_{is}, \ldots, \phi_{ns})'$.

There is no precedent methodology in the literature to deal with this testing problem. As there are several experiments in each population, we first tried to use a clasical permutation test over the experiments. This approach does not yield good results when the samples are unbalanced or there are other differences among the populations apart from the order. Based on what is suggested in the literature to deal with these kinds of problems related to permutation tests, we propose a randomization procedure (Algorithm 4) and a standardized statistic formulated as follows,

$$T = \frac{d(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}_G) - \sum_{s=1}^{S} d(\boldsymbol{\Theta}_s, \widehat{\boldsymbol{O}}_s)}{d(\boldsymbol{\Theta}, \widehat{\boldsymbol{O}}_G)}.$$

We present in this work some alternatives to this statistic. All the possibilities are compared in a simulation study. The best results regarding the power of the test and the Type I error are obtained with the statistic T proposed above.

This procedure has some outputs of interest apart from the p-value. It yields an estimator of the global circular order and the confidence coefficient of that estimator can be calculated using the results from the random selections.

5. Analysis of gene expression data

The biological problem, that was the motivation of the statistical methodology presented in this thesis, is solved by using the new methods developed. There has been considerable interest among cell biologists to identify cell-cycle genes that are evolutionarily conserved in their functions across multiple species. Cell-cycle is a well-coordinated process where events must take place in an orderly fashion for a successful cell division. Hence genes participating in the cell division cycle express in an order according to their function. Thus a question of interest for biologists is to determine whether the order of peak expression among cell-cycle genes is evolutionarily conserved. This question was extensively discussed and debated during the past decade using gene expression data obtained from *S.cerevisiae*, *S.pombe* and human Hela cell. Conservative estimates of the number of genes that are periodic in both species of yeasts is about 34 and the number that are periodic in the two yeasts and Humans is about 11. Then, as a first approach we analyze separately each species by using data from those 34 genes in the two yeasts and the 11 genes in Humans. The aggregated circular order is obtained in each species by using the methodology presented in Chapter 3 of this thesis.

On the other hand, the main issue regarding the comparison of circular orders among species is also solved. That biological problem is formulated with

the following hypotheses,

$$\begin{aligned} H_0 &: \boldsymbol{O}_P = \boldsymbol{O}_C = \boldsymbol{O}_H \\ H_1 &: H_0 \text{ not true.} \end{aligned} \tag{D.3}$$

where $\boldsymbol{O}_P$ is the underlying circular order in *S.pombe*, $\boldsymbol{O}_C$ is the underlying circular order in *S.cerevisiae* and $\boldsymbol{O}_H$ in Humans.

This problem is solved by using the methodology proposed in Chapter 4. The hypothesis (D.3) is rejected for the initial set of 11 genes. Then, a backward procedure is applied in order to find the maximum set of genes which do not reject that hypothesis. Finally, we discover that the temporal program among the genes $ZNF367$, $CDC6$, $PKMYT1$, $HIST2H4B$, $H2AFX$, $KIF10$ and $FOXM1$ is evolutionarily conserved from yeast to Humans (with over 90 % confidence). Regarding comparisons two by two species, there are no additional genes in the circular order for *S.cerevisiae* and Humans. In the case of *S.pombe* and Humans, 3 additional genes, $CCNB1$, $RAD21$ and $PLK1$ conserve the order. Comparing the two yeasts, we discover that the following 3 additional genes are also conserved: $CDC20$, $CCNB1$ and $RAD21$.

Considering previous literature on the subject, it is not surprising the results obtained with this methodology in the pairwise comparisons that could suggests that perhaps evolutionarily the functions of $CCNB1$ and $RAD21$ may have changed between Humans and *S.cerevisiae* with *S.pombe* being the *connecting link*. The two yeasts shared a common ancestor nearly a billion years ago and neither of them is closer to human beings more than the other (Forsburg, 1999). However, according to Aravind et al. (2000) and Roux et al. (2010), while *S.pombe* and metazoan cell-cycle genes retained some of the functions from their common ancestor, the *S.cerevisiae* cell-cycle genes may have lost them. In fact, relative to *S.cerevisiae* there are proportionally more *S.pombe* genes conserved in metazoans. There are other similarities between *S.pombe* and higher order animals including stress response pathways. Thus our methodology provides new hypotheses for biologists to investigate further.

6. Software: R package isocir

All the methodology developed in this work have been implemented in R language as part of the package **isocir** ("**iso**tonic inference for **cir**cular data"), Barragán et al. (2014a). In this chapter, we present briefly some packages in R to execute methods of ordered restricted inference and circular data. However, before **isocir**, there was no package to execute circular ordered restricted inference. We present the functions and how to use them through some toy examples.

7. Conclusions

This thesis presents two main contributions in the field of ordered restricted inference and in the field of biology.

Regarding the biological problem solved in this work, we present results that help to improve the knowledge of biological function of cell-cycle genes. In particular, we discover 7 genes whose order of peak expression along the cell-cycle is conserved from yeasts (*S.pombe* and *S.cerevisiae*) to Humans. Moreover, we find more common genes when pairwise comparisons between yeats or between Humans and *S.pombe* are considered. These results come from the analyses that use gene expression data from different experiments and different laboratories to compare the circular order of the peak expressions.

Regarding the statistical methodology, the problem of comparison of circular orders is formulated as a testing problem and is solved in this thesis. The null hypothesis of this problem is the equality of circular orders from different populations. As part of this problem, we present and solve the aggregation of circular orders. To our best knowledge both problems had not been considered before in the literature.

The problem of aggregation of circular orders is formulated as an optimization problem whose aim is to find the circular order which is closest to the data. We have developed several methods to approach this problem by using three levels of information: individual, pairwise and triplewise. The main novelty is the use of triplewise relationships. The triples arise naturally in the circle to measure circular association and they are the tool to aggregate information from experiments with different initial points in the circle. The information of triplewise relationships is represented mathematically by using hypermatrices and Hodge theory is used to translate them to pairwise relationships and then to generate the circular order closest to the data. The technique that we present in this work using Hodge theory is computationally efficient and the result is coherent. Furthermore, this technique has wide possibilities of future research related to the way to measure the triplewise relationships as well as to the possible definition of uncertainty coefficients. The other great novelty in this work in relation to this problem is another technique to aggregate circular orders which is based in the Traveling Salesman Problem (TSP) and uses pairwise relationships. From the simulation study, we can say that this technique gives the best approximations to the optimum in many cases. The key is to define properly the objetive function of the TSP, which is a distance that measures the relations between pairs of elements. We have conclude that the best options come from directed distances, i.e. the ones that take into account the direction of the circle. In fact, one of the main contributions of this work is the definition of a directed distance in the circle with convenient properties to solve the corresponding TSP in a computationally efficient way.

The aggregation of circular orders solves the problem of estimation of a common circular order from a set of data. Then, we use this methodology in the problem of testing equality of circular orders between two or more populations. We propose a nonparametric procedure based on random selection over the experiments. The statistic is chosen among different standardized statistics using simulated data under the null and the alternative hypotheses. This

methodology is able to deal with heterogeneous experiments as well as with unbalanced samples.

The new statistical methodology and all the algorithms designed in this work are implemented as part of the R package **isocir** ("**iso**tonic inference for **cir**cular data") Barragán et al. (2014a) which is publicly available in the CRAN (R Core Team (2014)).

It is obvious that all the methodology presented can be applied to different problems apart from the application solved in this work. In particular, in the biological field, it can be used for solving problems related to circadian cycle (Li et al. (2013)) or hormones cycles (Hendriks et al. (2014)) where it is relevant the relation between the peak incidences. There are other fields where we can find possible applications of this methodology, such as meteorology (Bowers et al. (2000)) or wildfire propagation (Albini y Forest (1976), Muñoz et al. (2013), García-Portugués et al. (2013)).

As a summary we may say that all the methods developed in this work are a contribution to the statistical framework of circular models under ordered restrictions that can be used to solved very interesting problems in many applied fields.

As for future research, the most obvious line is related to the aggregation of circular orders based on Hodge theory. Another line to explore is the possible contribution of these methods to rank aggregation methodology in the Euclidean space.

Bibliografía

[1] Abdullah, A., Moeller, J., y Venkatasubramanian, S. (2012). Approximate Bregman near neighbors in sublinear time: Beyond the triangle inequality. In: *Proceedings of the twenty-eighth annual symposium on Computational geometry*, pages 31–40. ACM.

[2] Agostinelli, C. y Lund, U. (2011). *circular: Circular Statistics*. R package version 0.4-3. https://r-forge.r-project.org/projects/circular/.

[3] Albini, F. A. y Forest, I. (1976). Estimating wildfire behavior and effects. *USDA Forest Service General Technical Report. INT-30. Intermountain Forest and Range Experiment Station*.

[4] Andreae, T. (2001). On the traveling salesman problem restricted to inputs satisfying a relaxed triangle inequality. *Networks*, 38(2):59–67.

[5] Aravind, L., Watanabe, H., Lipman, D. J., y Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11319–11324.

[6] Aslam, J. A. y Montague, M. (2001). Models for Metasearch. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 276–284. ACM.

[7] Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association*, 84(405):289–294.

[8] Bähler, J. (2005). Cell-cycle control of gene expression in budding and fission yeast. *The Annual Review of Genetics*, 39:69–94.

[9] Balabdaoui, F., Rufibach, K., y Santambrogio, F. (2009). *OrdMonReg: Compute Least Squares Estimates of One Bounded or Two Ordered Isotonic Regression Curves*. R package version 1.0.2. http://CRAN.R-project.org/package=OrdMonReg.

[10] Baltrunas, L., Makcinskas, T., y Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 119–126. ACM.

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

[12] Barragán, S., Fernández, M., y Rueda, C. (2014a). *isocir: Isotonic Inference for Circular data*. R package version 1.2. http://CRAN.R-project.org/package=isocir.

[13] Barragán, S., Fernández, M., Rueda, C., y Peddada, S. (2013). isocir: An R package for constrained inference using isotonic regression for circular data, with an application to cell biology. *Journal of Statistical Software*, 54(4):1–17.

[14] Barragán, S., Rueda, C., Fernández, M., y Peddada, S. (2014b). Statistical framework for determining the temporal program in an oscillatory system. *Preprint*.

[15] Bartholdi, J., Tovey, C., y Trick, M. (1989). Voting schemes for which it can be difficult to tell who won the election. *Social Choice Welfare*, 6:157–165.

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

[17] Bender, M. A. y Chekuri, C. (2000). Performance guarantees for the TSP with a parameterized triangle inequality. *Information Processing Letters*, 73(1-2):17–21.

[18] Berger, R. (1984). Testing for the same ordering in several groups of means. Technical report, DTIC Document.

[19] Boles, L. y Lohmann, K. (2003). True navigation and magnetic maps in spiny lobsters. *Nature*, 421:60–63.

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

[21] Bowers, J., Morton, I., y Mould, G. (2000). Directional statistics of the wind and waves. *Applied Ocean Research*, 22:13–30.

[22] Bradley, R. A. y Terry, M. E. (1952). Rank analisis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345.

[23] Brunsdon, C. y Corcoran, J. (2005). Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems*, 30:300–319.

[24] Caplin, A. y Nalebuff, B. (1991). Aggregation and social choice: A mean voter theorem. *Econometrica*, 59(1):1–23.

[25] Chasalow, S. (2010). *combinat: Combinatorics Utilities*. R package version 0.0-8. http://CRAN.R-project.org/package=combinat.

[26] Chen, X., Bennett, P. N., Collins-Thompson, K., y Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In: *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 193–202.

[27] Chevaleyre, Y., Endriss, U., Lang, J., y Maudet, N. (2007). A short introduction to computational social choice. In: *SOFSEM 2007: Theory and Practice of Computer Science*, J. Leeuwen, G. Italiano, W. Hoek, C. Meinel, H. Sack, y F. Plášil, ed., volume 4362, pages 51–69. Springer.

[28] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., y Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73.

[29] Chung, E. y Romano, J. (2011). *Asymptotically Valid and Exact Permutation Tests Based on Two-sample U-Statistics*. Stanford University. Technical Report No. 2011-09.

[30] Chung, E. y Romano, J. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507.

[31] Cochran, W., Mouritsen, H., y Wikelski, M. (2004). Migrating songbirds recalibrate their magnetic compass daily from twilight cues. *Science*, 304:405–408.

[32] Condorcet, M. (1785). *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.

[33] Cook, W. D. (2006). Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of Operational Research*, 172(2):369 – 385.

[34] Copeland, A. (1951). A reasonable social welfare function. In: *Seminar on Mathematics in Social Sciences*. University of Michigan.

[35] Dalal, O., Sengemedu, S. H., y Sanyal, S. (2012). Multi-objective ranking of comments on web. In: *Proceedings of the 21st international conference on World Wide Web*, pages 419–428. ACM.

[36] Daniels, H. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(2):171–191.

[37] de Leeuw, J., Hornik, K., y Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24.

[38] de Lichtenberg, U., Wernersson, R., Jensen, T., Nielsen, H., Fausboll, A., Schmidt, P., Hansen, F., Knudsen, S., y Brunak, S. (2005). New weakly expressed cell cycle-regulated genes in yeast. *Yeasts*, 22(5):1191–1201.

[39] DeConde, R., Hawley, S., y Falcon, S. (2006). Combining results of microarray experiments: A rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1–23.

[40] Diaconis, P. y Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):262–268.

[41] Dwork, C., Kumar, R., Naor, M., y Sivakumar, D. (2001a). Rank aggregation methods for the web. In: *Proceedings of the 10th International World Wide Web Conference*, pages 613–622.

[42] Dwork, C., Kumar, R., Naor, M., y Sivakumar, D. (2001b). Rank aggregation revisited. *Manuscript*.

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

[44] Dym, C., Wood, W., y Scott, M. (2002). Rank ordering engineering designs: Pairwise comparison charts and borda counts. *Research in Engineering Design*, 13(4):236–242.

[45] Felsenthal, D. y Tideman, N. (2014). Weak condorcet winner(s) revisited. *Public Choice*, pages 1–14.

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

[47] Fernández, M., Rueda, C., y Peddada, S. (2012). Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research*, 40(7):2823–2832. http://nar.oxfordjournals.org/content/40/7/2823.

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

[49] Fernández, M., Rueda, C., y Salvador, B. (1998). Simultaneous estimation by isotonic regression. *Journal of Statistical Planning and Inference*, 70:111–119.

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

[51] Fernández, M., Rueda, C., y Salvador, B. (2000). Parameter estimation under orthant restrictions. *The Canadian Journal of Statistics*, 28:171–181.

[52] Fernández-Durán, J. J. y Gregorio-Dominguez, M. M. (2013). *CircNNTSR: An R package for the statistical analysis of circular data using nonnegative trigonometric sums (NNTS) models*. R package version 2.1. http://CRAN.R-project.org/package=CircNNTSR.

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

[54] Fisher, N. y Lee, A. (1982). Non-parametric measures of angular-angular association. *Biometrika*, 69:315–321.

[55] Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

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

[57] Forlizzi, L., Hromkovič, J., Proietti, G., y Seibert, S. (2005). On the stability of approximation for hamiltonian path problems. In: *SOFSEM 2005: Theory and Practice of Computer Science*, volume 3381 of *Lecture Notes in Computer Science*, pages 147–156. Springer Berlin Heidelberg.

[58] Forsburg, S. L. (1999). The best yeast? *Trends in genetics*, 15(9):340–344.

[59] Forsburg, S. L. (2007). The yeasts Saccharomyces cerevisiae and Schizosaccharomyces pombe: Models for cell biology research. *Gravitational and Space Research*, 18(2).

[60] Fox, S. (2011). *Fisiología Humana*. Mc Graw Hill interamericana.

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

[62] Gauthier, N., Larsen, M., Wernersson, R., de Lichtenberg, U., Jensen, L., Brunak, S., y Jensen, T. (2008). Cyclebase.org - a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Research*, 36:854–859.

[63] Gleich, D. F. y Lim, L.-h. (2011). Rank aggregation via nuclear norm minimization. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 60–68. ACM.

[64] Glynn, E. F., Chen, J., y Mushegian, A. R. (2006). Detecting periodic patterns in unevenly spaced gene expression time series using lomb-scargle periodograms. *Bioinformatics*, 22(3):310–316.

[65] Gong, J., Xu, C.-Z., y Holle, J. (2007). Predictive directional greedy routing in vehicular ad hoc networks. In: *27th International Conference on Distributed Computing Systems Workshops, 2007. ICDCSW '07.*, pages 2–2.

[66] Hahsler, M. y Hornik, K. (2011). *Traveling Salesperson Problem (TSP).* R package version 1.0-6. http://CRAN.R-project.org/.

[67] Haskey, J. (1988). The relative orientation of addresses of spouses before their marriage: An analysis of circular data. *Journal of Applied Statistics*, 15:183.

[68] Hassanzadeh, F. (2013). *Distances on Rankings: From Social Choice to Flash Memories.* PhD thesis, University of Illinois.

[69] Hendriks, G.-J., Gaidatzis, D., Aeschimann, F., y Grosshans, H. (2014). Extensive oscillatory gene expression during C. elegans larval development. *Molecular Cell*, 53(3):380 – 392.

[70] Ho, T. K., Hull, J. J., y Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75.

[71] Hu, X. y Wright, F. (1994). Monotonicity properties of the power functions of likelihood ratio tests for normal mean hypotheses constrained by a linear space and a cone. *The Annals of Statistics*, pages 1547–1554.

[72] Huang, Y., Xu, H., Calian, V., y Hsu, J. (2006). To permute or not to permute. *Bioinformatics*, 22(18):2244–2248.

[73] Iverson, G. J. y Harp, S. A. (1987). A conditional likelihood ratio test for order restrictions in exponential families. *Mathematical Social Sciences*, 14(2):141 – 159.

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

[75] Jammalamadaka, S. y SenGupta, A. (2001). *Topics in Circular Statistics.* World Scientific, New York.

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

[77] Jiang, X., Lim, L.-H., Yao, Y., y Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244.

[78] Jupp, P. y Mardia, K. (1980). A general correlation coeficient for directional data and related regression problems. *Biometrika*, 67:163–173.

[79] Kadota, K. y Shimizu, K. (2011). Evaluating methods for ranking differentially expressed genes applied to microarray quality control data. *BMC Bioinformatics*, 12(1):227.

[80] Kane, J., Sternheim, M., Vázquez, J., y Mirabent, D. (1989). *Física*. Reverté.

[81] Karp, R. (1972). Reducibility among combinatorial problems. In: *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. Springer US.

[82] Kelly, R. (1989). Stochastic reduction of loss in estimating normal means by isotonic regression. *The Annals of Statistics*, 17:937–940.

[83] Kerr, F. (2009). Comments on the analysis of unbalanced microarray data. *Bioinformatics*, 25(16):2035–2041.

[84] Kibiak, T. y Jonas, C. (2007). Applying circular statistics to the analysis of monitoring data. *European Journal of Psychological Assessment*, 23:227–237.

[85] Klamler, C. (2004). The dodgson ranking and the borda count: A binary comparison. *Mathematical Social Sciences*, 48(1):103 – 108.

[86] Kumar, R. y Vassilvitskii, S. (2010). Generalized distances between rankings. In: *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 571–580. ACM.

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

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

[89] Lee, C. (1981a). The quadratic loss of isotonic regression under normality. *The Annals of Statistics*, 9:686–688.

[90] Lee, C. (1981b). The quadratic loss of order restricted estimators for treatment means with a control. *The Annals of Statistics*, 16:751–758.

[91] Lee, C.-I. C. (1983). The min-max algorithm and isotonic regression. *The Annals of Statistics*, 11(2):467–477.

[92] Li, Y., Li, G., Wang, H., Du, J., y Yan, J. (2013). Analysis of a gene regulatory cascade mediating circadian rhythm in zebrafish. *PLoS computational biology*, 9(2):e1002940.

[93] Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., y Bijnens, L. (2012). *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*. Springer.

[94] Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570.

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

[96] Liu, Y., Liu, T., Qin, T., Ma, Z., y Li, H. (2007). Supervised rank aggregation. In: *Proceedings of the 2007 International World Wide Web Conference*, pages 481–489.

[97] Lumini, A. y Nanni, L. (2006). Detector of image orientation based on Borda count. *Pattern Recognition Letters*, 27(3):180–186.

[98] Lund, U. y Agostinelli, C. (2009). *CircStats: Circular Statistics, from "Topics in Circular Statistics" (2001)*. R package version 0.2-4. http://CRAN.R-project.org/package=CircStats.

[99] Malandraki, C. y Daskin, M. S. (1992). Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. *Transportation Science*, 26(3):185–200.

[100] Maldonado, P., Godecke, I., Gray, C., y Bonhoeffer, T. (1997). Orientation selectivity in pinwheel centers in cat striate cortex. *Science*, 276:1551–1555.

[101] Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, 44(1-2):114–130.

[102] Mardia, K. (1975). Statistics of directional data (with discussion). *Journal of the Royal Statistical Society - Series B.*, 37:349–393.

[103] Mardia, K. y Jupp, P. (2000). *Directional Statistics*. John Wiley & Sons.

[104] Mc Donald, K. y Smeaton, A. F. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval. In: *Image and video retrieval*, pages 61–70. Springer Berlin Heidelberg.

[105] Mekonnen, A. D. (2014). *Wind Farm Site Suitability Analysis in Lake Erie Using Web-Based Participatory GIS (PGIS)*. PhD thesis, Bowling Green State University.

[106] Menéndez, J., Rueda, C., y Salvador, B. (1991a). Conditional test for testing a face of the tree order cone. *Communications in statistics. Simulation and computation*, 20(2-3).

[107] Menéndez, J., Rueda, C., y Salvador, B. (1991b). Dominance of likelihood ratio tests under order constraints. *The Annals of Statistics*, 20:2087–2099.

[108] Menéndez, J., Rueda, C., y Salvador, B. (1992). Testing non-oblique hypothesis. *Communications in Statistics-Theory and Methods*, 21(2):471–484.

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

[110] Menéndez, J. y Salvador, B. (1992). Equivalence of likelihood ratio test and obliquity. *Statistics & probability letters*, 14:223–229.

[111] Mewhort, D., Kelly, M., y Johns, B. (2009). Randomization tests and the unequal-n/unequal variance problem. *Behavior Research Methods*, 41(3):664–667.

[112] Moret, B. y Shapiro, H. (1991). *Algorithms from P to NP: Design & efficiency*. Benjamin-Cummings Publishing Co.

[113] Mosteller, F. (1951). Remarks on the method of paired comparisons. *Psychometrika*, 16:203–218.

[114] Mukerjee, H., Robertson, T., y Wright, F. (1986). A probability inequality for elliptically countoured densities with applications in order restricted inference. *The Annals of Statistics*, 14(4):1544–1554.

[115] Muñoz, M. D., Mata, A., Corchado, E., y Corchado, J. M. (2013). (obifs) isotropic image analysis for improving a predicting agent based systems. *Expert Systems with Applications*, 40(12):5011 – 5020.

[116] Naor, J. S. y Schwartz, R. (2010). The directed circular arrangement problem. *ACM Transactions on Algorithms (TALG)*, 6(3):47.

[117] Nelson, D., Cox, M., y Lehninger, A. (2005). *Lehninger: Principios de Bioquímica*. Ediciones Omega.

[118] Odén, A. y Wedel, H. (1975). Arguments for Fisher's permutation test. *The Annals of Statistics*, pages 518–520.

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

[120] Oliveira, M., Crujeiras, R. M., y Rodríguez-Casal, A. (2013). *NPCirc: Nonparametric Circular Methods*. R package version 2.0.0. http://CRAN.R-project.org/package=NPCirc.

[121] Orponen, P. y Mannila, H. (1987). On approximation preserving reductions: Complete problems and robust measures. Technical report, C-1987-28, Department of Computer Science, University of Helsinki. (ND32, MP1, LO7).

[122] Papadimitriou, C. y Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. Dover Books on Computer Science Series. Dover Publications.

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

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

[125] Perlman, M. (1969). One-sided problems in multivariate analysis. *The Annals of Mathematical Statistics*, 40:549–567.

[126] Pewsey, A., Neuhäuser, M., y Ruxton, G. D. (2013). *Circular Statistics in R*. Oxford University Press.

[127] Pihur, V., Datta, S., y Datta, S. (2008). Finding common genes in multiple cancer types through meta-analysis of microarray experiments: A rank aggregation approach. *Genomics*, 92(6):400–403.

[128] Pramila, T., Wu, W., Miles, S., Noble, W., y Breeden, L. L. (2006). The forkhead transcription factor Hcm1 regulates chromosome segregation genes

and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes & Development*, 20(16):2266–2278.

[129] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[130] Raj, D. y Khamis, S. H. (1958). Some remarks on sampling with replacement. *The Annals of Mathematical Statistics*, 29(2):550–557.

[131] Rajkumar, A. y Agarwal, S. (2014). A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In: *Proceedings of The 31st International Conference on Machine Learning*, pages 118–126.

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

[133] Rivest, L. (1982). Some statistical methods for bivariate circular data. *Journal of the Royal Statistical Society - Series B*, 44(1):81–90.

[134] Robertson, T. y Wright, F. (1975). Consistency in generalized isotonic regression. *The Annals of Statistics*, pages 350–362.

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

[136] Robertson, T., Wright, F., y Dykstra, R. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons.

[137] Roux, A. E., Chartrand, P., Ferbeyre, G., y Rokeach, L. A. (2010). Fission yeast and other yeasts as emergent models to unravel cellular aging in eukaryotes. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 65(1):1–8.

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

[139] Rueda, C. (1989). *Contrastes de Hipótesis con Restricciones bajo Condiciones de Oblicuidad*. PhD thesis, Universidad de Valladolid.

[140] Rueda, C., Fernández, M., Barragán, S., y Peddada, S. (2014a). Circular order restricted inference with applications to molecular biology. *Preprint*.

[141] Rueda, C., Fernández, M., y Peddada, S. (2009). Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *Journal of the American Statistical Association*, 104(485):338–347.

[142] Rueda, C., Salvador, B., y Fernández, M. (1997a). A good property of the maximum likelihood estimator in a restricted normal model. *TEST*, 6(1):127–135.

[143] Rueda, C., Salvador, B., y Fernández, M. (1997b). Simultaneous estimation in a restricted linear model. *Journal of Multivariate Analysis*, 61(1):61–66.

[144] Rueda, C., Ugarte, M., y Militino, A. (2014b). Checking unimodality and locating the break-point: An application to breast cancer mortality trends. *Preprint*.

[145] Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C., Burns, G., Hayles, J., Brazma, A., Nurse, P., y Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, 36:809–817.

[146] Sasabuchi, S., Inutsuka, M., y Kulatunga, D. (1983). A multivariate version of isotonic regression. *Biometrika*, 70(2):465–472.

[147] Schalekamp, F. y Zuylen, A. (2009). Rank aggregation: Together we are strong. In: *Proceedings of 11th ALENEX*, pages 38–51.

[148] Shishkin, A., Zhinalieva, P., y Nikolaev, K. (2013). Quality-biased ranking for queries with commercial intent. In: *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 1145–1148.

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

[150] Simko, I. y Pechenick, D. (2010). Combining partially ranked data in plant breeding and biology: I. Rank aggregating methods. *Communications in Biometry and Crop Science*, 5(1):41–55.

[151] Simpson, S. L. y Edwards, L. J. (2013). A circular LEAR correlation structure for cyclical longitudinal data. *Statistical Methods in Medical Research*, 22(3):296–306.

[152] Sizemore, R. (2013). *HodgeRank: Applying Combinatorial Hodge Theory to Sports Ranking*. PhD thesis, Wake Forest University.

[153] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., y Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

[154] Straume, M. (2004). DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods in enzymology*, 383:149.

[155] Susko, E. (2013). Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika*, 100(4):1019–1023.

[156] Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

[157] Truchon, M. (1997). An extension of the Condorcet criterion and Kemeny orders. *cahier 98-15 du Centre de Recherche en Economieet Finance Appliquees*.

[158] Tsitsiklis, J. N. (1992). Special cases of traveling salesman and repairman problems with time windows. *Networks*, 22:263–282.

[159] Turner, R. (2009). *Iso: Functions to Perform Isotonic Regression*. R package version 0.0-8. http://CRAN.R-project.org/package=Iso.

[160] Volkovs, M. N. y Zemel, R. S. (2012). A flexible generative model for preference aggregation. In: *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 479–488.

[161] von Mises, R. (1918). Über die "Ganzzahligkeit"der Atomgewichte und verwandte Fragen. *Physikal Z.*, 19:490–500.

[162] Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., y Brown, P. O. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell*, (6):1977–2000.

[163] Wichert, S., Fokianos, K., y Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

[164] Wollan, P. y Dykstra, R. (1986). Conditional tests with an order restriction as a null hypothesis. In: *Advances in Order Restricted Statistical Inference*, volume 37 of *Lecture Notes in Statistics*, pages 279–295. Springer New York.

[165] Xu, Q., Huang, Q., Jiang, T., Yan, B., Lin, W., y Yao, Y. (2012). Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857.

[166] Zhao, L. P., Prentice, R., y Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10):5631–5636.