PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

AN OPTIMAL ALGORITHM FOR STRICT CIRCULAR SERIATION

SANTIAGO ARMSTRONG CRUZ

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisors:
CRISTÓBAL ANDRÉS GUZMÁN PAREDES
CARLOS ALBERTO SING-LONG COLLAO

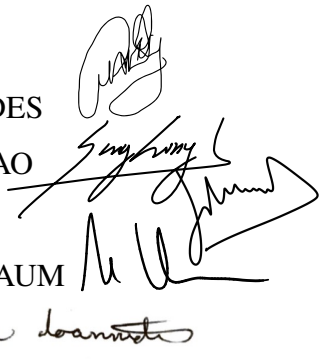
Santiago de Chile, January 2021

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

AN OPTIMAL ALGORITHM FOR STRICT CIRCULAR SERIATION

SANTIAGO ARMSTRONG CRUZ

Members of the Committee:
CRISTÓBAL ANDRÉS GUZMÁN PAREDES
CARLOS ALBERTO SING-LONG COLLAO
SEBASTIÁN VICUÑA DÍAZ
JOSÉ CLAUDIO VERSCHAE TANNENBAUM
ALEXANDER IOANNIDIS



Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, January 2021

«You can only find truth with logic
if you have already found truth
without it»
G. K. Chesterton

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a mis supervisores, Cristóbal Guzmán y Carlos Sing-Long, ya que sin su ayuda este trabajo no hubiera sido posible. De Cristóbal aprendí muchísimo en los últimos años. Le agradezco por su constante apoyo y buena disposición durante esta investigación, así como también por su buena onda y genuino interés por ayudar. Haber sido su alumno, ayudante y tesista ha sido un gran honor. A Carlos le agradezco por su orientación. Su capacidad de vincular problemas y de formular las preguntas correctas fueron clave en las distintas etapas de esta tesis. Además como alumno, le agradezco por su inagotable esfuerzo por enseñar de la mejor manera posible.

En segundo lugar, me gustaría agradecer a todos mis compañeros del Instituto de Ingeniería Matemática por sus valiosas discusiones. En particular a Pedro Izquierdo y Vicente Gómez.

En tercer lugar, por su participación y sugerencias, quiero agradecer a la comisión de mi tesis: Sebastián Vicuña, José Verschae y Alexander Ioannidis. Por las mismas razones también me gustaría agradecer a Alexandre d'Aspremont.

Finalmente, por su apoyo incondicional y cariño, me gustaría agradecer a María Luisa, a mis padres: Francisca y Jorge, a mis hermanos: Francisca, Andrés y Elisa, y a Emilio.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The seriation problem seeks to order a sequence of n objects when the only information we are given is a dissimilarity matrix between all pairs of objects. In linear seriation the goal is to find a *linear order* of the objects in a manner that is consistent with their dissimilarity. For this problem optimal $\mathcal{O}(n^2)$ algorithms are known. A generalization of the previous problem is circular seriation, where the goal is to find a *circular order* instead. In this thesis we study the circular seriation problem. Our contributions can be summarized as follows. First, we introduce *circular Robinson matrices* as the natural class of dissimilarity matrices for the circular seriation problem. Second, for the case of *strict circular Robinson dissimilarity matrices* we provide an optimal $\mathcal{O}(n^2)$ algorithm for the circular seriation problem. Finally, we propose a statistical model to analyze the well-posedness of the circular seriation problem for large n. In particular, we establish $\mathcal{O}(\log(n)/n)$ rates on the distance between any circular ordering found by solving the circular seriation problem to the underlying order of the model, in the Kendall-tau metric.

Keywords: Circular seriation, circular Robinson dissimilarities, PQ-trees, circular Robinsonian matrices, circular-arc hypergraphs, circular embeddings of graphs, generative model, unsupervised learning.

RESUMEN

El problema de la seriación busca ordenar una secuencia de n objetos cuando la única información que se nos da es una matriz de disimilitud entre todos los pares de objetos. En la seriación lineal, el objetivo es encontrar un em orden lineal de los objetos manera que sea consistente con su disimilitud. Para este problema se conocen los algoritmos óptimos $\mathcal{O}(n^2)$. Una generalización del problema anterior es seriación circular, donde el objetivo es encontrar un em orden circular. En esta tesis estudiamos el problema de la seriación circular. Nuestras contribuciones se pueden resumir de la siguiente manera. Primero, presentamos em matrices circulares de Robinson como la clase natural de matrices de disimilitud para el problema de seriación circular. En segundo lugar, para el caso de em matrices de disimilitud circular estrictas de Robinson proporcionamos un algoritmo $\mathcal{O}(n^2)$ óptimo para el problema de seriación circular. Finalmente, proponemos un modelo estadístico para analizar el buen planteamiento (*well-posedness* en el sentido de Hadamard) del problema de seriación circular para grandes valores de n. En particular, establecemos tasas del orden $\mathcal{O}(\log(n)/n)$ para la distancia entre cualquier orden circular encontrado al resolver el problema de seriación circular al orden subyacente del modelo, en la métrica de Kendall-tau.

Palabras Claves: Seriación circular, matrices de Robinson, árboles PQ, hipergrafos de arco, modelo generativo, aprendizaje no supervisado

1. INTRODUCTION

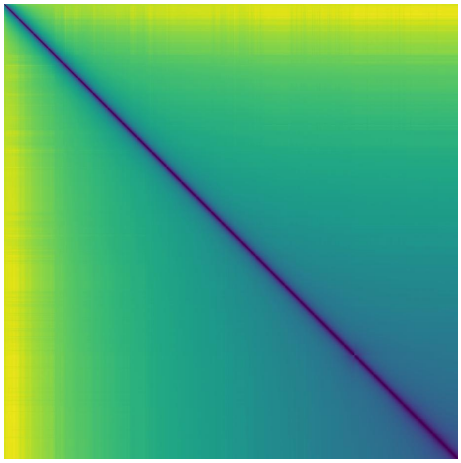
The seriation problem seeks to recover a latent ordering from dissimilarity information (Recanati, Brüls, & d'Aspremont, 2017). The input for this problem is a matrix measuring pairwise dissimilarity between a set of n elements. Liiv defines seriation as "an exploratory data analysis technique to reorder objects into a sequence along a one-dimensional continuum so that it best reveals regularity and patterning among the whole series" (Liiv, 2010). In seriation, one typically assumes that the data can be ordered along a chain where the dissimilarity between elements increases with respect to their distance within this chain. In practice, we observe a random permutation of this dissimilarity matrix, where the elements are not indexed according to that latent ordering. Seriation then seeks to find that global latent ordering using only pairwise dissimilarity (Recanati, Kerdreux, & d'Aspremont, 2018).

Originating from the field of archaeology, where it was used to infer the chronological order of a set of graves based on the artifacts recovered from them (Robinson, 1951), seriation has found applications in several areas such as sociology and psychology (Liiv, 2010), and gene sequencing and bioinformatics (Recanati et al., 2017). Although the applications of this mathematical problem are several, in this introduction we are going to have a special application in mind to give context to the problem, which we introduce in the next subsection.

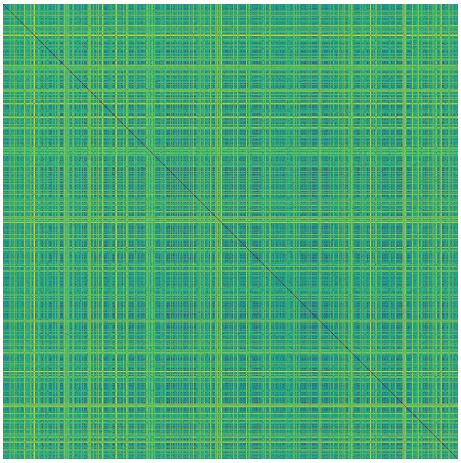
1.1. Seriation in the context of computer vision

A dissimilarity over a set $\mathcal{X}$ of n objects is a symmetric function $\mathbf{d} : \mathcal{X}^2 \to \mathbb{R}_+$ vanishing on the diagonal. Given any enumeration of the elements in $\mathcal{X}$, we can construct a dissimilarity matrix D whose entry (i, j) is $\mathbf{d}(x_i, x_j)$, the dissimilarity between the i-th and j-th objects. In seriation we are given the matrix D for an arbitrary enumeration of the elements and we want to infer a 'natural' ordering of the data points $\mathcal{X}$. To give context to the problem, suppose that the points in $\mathcal{X}$ correspond to the frames of a movie. Hence,

n corresponds to the number of frames and each element $x \in \mathcal{X}$ corresponds to a specific frame of the movie. Since each frame is an image, we can represent them with a vector in $\mathbb{R}^r$ were r corresponds to the number of pixels in the frames (assuming a black and white movie for simplicity). An example of a dissimilarity in $\mathbb{R}^r$ would be the Euclidean distance $\mathbf{d}(x, y) = \|x - y\|_2$. The natural intrinsic order of the frames would be the chronological order and we would expect consecutive frames in the movie to have smaller dissimilarity (Euclidean distance), than frames occurring at distinct instants of the movie (temporally separated). As an example, Figure 1.1 shows the Euclidean distance matrix between the images in a 31 second movie of a kiwi slowly rotting for several days. Since this process develops slowly, we expect consecutive frames in the movie to be very similar. Consistent with our intuition, in Figure 1.1a we see that when the indices of the dissimilarity matrix are chronologically ordered, the dissimilarity tends to increase as we move away from the diagonal, i.e., when the time interval between the kiwi images increases. Dissimilarity matrices satisfying this property are called *linear Robinson matrices*, which we will formally introduce in Section 1.4. In 1.1b, the frames are given in a random order, which is what is observed in practice. The goal of seriation is to recover 1.1a given 1.1b. In our time-lapse video example there is an important thing to observe: the last frames are very different from the initial frames of the video. If the video sequence were a closed-loop sequence instead, we would expect that the dissimilarity decreases at the end of the sequence as in the case of a distance matrix of points embedded in the unit circle. In contrast, Figure 1.1 shows the Euclidean distance matrix between the images in a 13 second movie of a face slowly spinning in 360 degrees. The last frame of the clip is very similar to the initial frame. In a matrix representation, this can be visualized as a symmetric matrix of pairwise dissimilarities where entries of each row (column) increase monotonically while moving to the right (bottom) until some specific element and then decrease again monotonically until the end of each row (column) and fold back from the left (top) of the matrix. Dissimilarity matrices satisfying this property are called *circular Robinson matrices*, which we will also introduce in section 1.4 (Evangelopoulos, Brockmeier, Mu, & Goulermas, 2020).

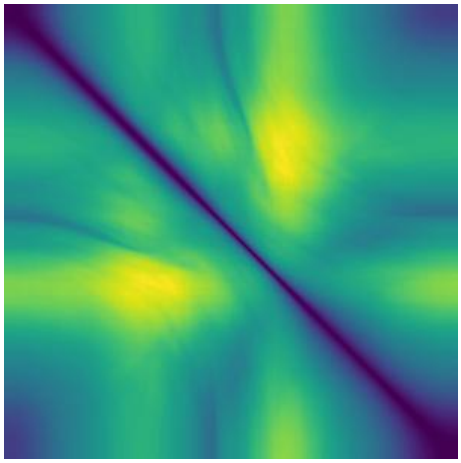


(a) movie-ordered

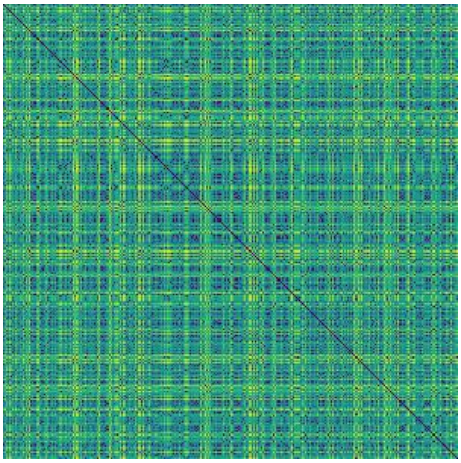


(b) permuted

Figure 1.1. Dissimilarity matrix between kiwi images using the euclidean distance, when the subscripts follow the chronological ordering of the movie 1.1a, and when it is randomly permuted 1.1b. Yellow entries correspond to higher distance values whereas bluer entries of the matrix correspond to smaller values of the distance.



(a) movie-ordered



(b) permuted

Figure 1.2. Dissimilarity matrix between images of faces using the euclidean distance, when the subscripts follow the chronological ordering of the movie 1.2a, and when it is randomly permuted 1.2b. Yellow entries correspond to higher distance values whereas bluer entries of the matrix correspond to smaller values of the distance.

1.2. Applications of circular seriation in real-world problems

There are several real-world applications that motivate the development of algorithms for circular seriation. In this section we review two examples that have been addressed recently in the literature.

1.2.1. Tomography from unknown random projections (Coifman, Shkolnisky, Sigworth, & Singer, 2008)

The problem in tomography is to reconstruct an object from samples of its projections. The object is characterized by its density function $\rho(x, y)$. The goal is to recover the density ρ from its projections $\mathcal{R}\rho(t, \theta)$ given by the line integral of ρ along parallel lines L at an angle θ with respect to an axis of reference, and at a distance t from the origin (i.e. its Radon transform) (see, e.g. (Deans, 2007)). However, there are cases in which the projection angles θ are unknown, for example, when reconstructing certain biological proteins or moving objects. By constructing a dissimilarity matrix from the random projections, circular seriation can be used to sort the projections cyclically. After sorting the projections it is possible to reconstruct the object's density if enough samples are obtained.

1.2.2. DNA sequencing (Recanati, 2018)

DNA sequencing refers to the process of determining the nucleotide order of a given DNA fragment. In a sequencing experiment, we can only 'read' small fragments (reads) of DNA due to physical limitations, whose location on the genome is unknown. *De novo* assembly aims to put them together to retrieve the full DNA sequence. A common approach to this problem is to construct a similarity matrix based on pairwise overlaps between reads and leave the task of sequencing to (linear) seriation algorithms. There are cases in which the DNA fragments are subsampled from a circular genome (e.g. bacterial plasmids and mitochondria), yielding an instance for the circular seriation problem instead (Recanati et al., 2017).

1.3. Notation and preliminaries

In order to formulate the mathematical problem, let us introduce some notations, basic definitions and folk results from several areas of discrete mathematics.

1.3.1. Matrix and vector notation

All arrays start at index 0. The symbol $\mathbb{R}$ denotes the set of real numbers, whereas $\mathbb{R}_+$ denotes the set of non-negative real numbers. Given a set $\mathcal{X}$, we write $\mathcal{X}^d$ to denote the set of d-dimensional vectors taking values in $\mathcal{X}$. Given a vector $x \in \mathcal{X}^d$, we write $x(i)$ as the element at index i. We write as $\mathcal{X}^{n\times m}$ the set of matrices of n rows and m columns taking values in $\mathcal{X}$. Given $M \in \mathcal{X}^{n\times m}$, $M(i,j)$ is the element at row i and column j. Given an array M, we write M^T to denote the transpose of M, i.e. the array that satisfies $M(i,j) = M^T(j,i)$. We endow the set $\mathbb{R}^d$ with the p-norm $\|x\|_p \triangleq \left(\sum_{i\in[d]} |x(i)|^p\right)^{\frac{1}{p}}$ and the corresponding induced distances. Given a numeric array M, we write $\|M\|_\infty$ to denote the maximum element among all its entries. The $n \times n$ identity matrix is denoted as I_n. The null vector and the 'all ones' vector are respectively denoted as $\mathbf{0}_n$ and $\mathbf{1}_n$. We may omit the subscripts when the dimensions are clear from the context.

1.3.2. Permutations

The set of all integers from 0 to $n-1$ is denoted as $[n]$.[1] We denote as $\mathrm{Sym}(n)$ the permutation set, i.e. the set of all bijections $\pi : [n] \to [n]$. The elements in $\mathrm{Sym}(n)$ are called permutations. Given any permutation $\pi \in \mathrm{Sym}(n)$ we represent it either with a vector π such that $\pi(i) = j$ whenever π maps i to j, or with a $n \times n$ matrix $\Pi \in \{0,1\}^{n\times n}$ (in upper case) such that $\Pi(i,j) = 1 \leftrightarrow \pi(i) = j$. We call such matrices, permutation matrices.

[1]Notice that this differs form the convention of writing $[n] = \{1, 2, \ldots, n\}$ but the proposed definition simplifies using modular arithmetic in the indices

1.3.3. Kendall-tau's metric

Kendall-tau's correlation coefficient (Kendall, 1938) has been introduced to measure discrepancy between permutations. In this work, we use instead Kendall-tau's metric (Ma, Tony Cai, & Li, 2020), which is defined as follows

$$\tau_K\left(\pi_1, \pi_2\right) \triangleq \frac{\left|\mathcal{G}\left(\pi_1, \pi_2\right)\right|}{\binom{n}{2}}$$

where, for two permutations π_1 and π_2, $\mathcal{G}\left(\pi_1, \pi_2\right)$ corresponds to the set of discordant pairs defined as

$$\mathcal{G}\left(\pi_1, \pi_2\right) \triangleq \{(i, j) : i < j, [\pi_1(i) < \pi_1(j) \wedge \pi_2(i) > \pi_2(j)] \vee [\pi_1(i) > \pi_1(j) \wedge \pi_2(i) < \pi_2(j)]\}$$

The denominator $\binom{n}{2}$ ensures that $\tau_K\left(\pi_1, \pi_2\right) \in [0, 1]$ where $\tau_K\left(\pi_1, \pi_2\right) = 0$ corresponds to $\pi_1 = \pi_2$ (Ma et al., 2020).

Definition 1. *Given a set $S \subset \mathrm{Sym}(n)$ of permutations, the diameter of the S is defined as*

$$\mathrm{diam}(S) \triangleq \max_{\pi_1, \pi_2 \in S} \tau_K(\pi_1, \pi_2)$$

1.3.4. Order theory

Definition 2. *(Burris & Sankappanavar, 2012) A binary relation $\leqslant$ defined on a set $\mathcal{X}$ is a partial order on the set $\mathcal{X}$ if the following conditions hold identically in $\mathcal{X}$:*

(i) $a \leqslant a$ *(reflexivity)*

(ii) $a \leqslant b$ *and* $b \leqslant a$ *imply* $a = b$ *(antisymmetry)*

(iii) $a \leqslant b$ *and* $b \leqslant c$ *imply* $a \leqslant c$ *(transitivity).*

If, in addition, for every a, b in $\mathcal{X}$

(iv) $a \leqslant b$ *or* $b \leqslant a$ *(totality)*

then we say $\leqslant$ is a linear order on $\mathcal{X}$.

Definition 3 (Interval)**.** *A subset $\mathcal{I}$ of an linearly ordered set $(\mathcal{X}, <)$ is said to be an interval if there are some $m, M \in \mathcal{X}$ such that $\mathcal{I} = \{i : m \leq i \leq M\}$. We refer to such elements m and M as the borders of $\mathcal{I}$ and often write $\mathcal{I} = [m, M]$.*

Definition 4. *A cyclic order on a set $\mathcal{X}$ is a relation $\mathscr{C} \subset \mathcal{X}^3$, that satisfies the following axioms:*

(i) *Cyclicity: If $(a, b, c) \in \mathscr{C}$ then $(b, c, a) \in \mathscr{C}$;*

(ii) *Asymmetry: If $(a, b, c) \in \mathscr{C}$ then $(c, b, a) \notin \mathscr{C}$;*

(iii) *Transitivity: If $(a, b, c) \in \mathscr{C}$ and $(a, c, d) \in \mathscr{C}$ then $(a, b, d) \in \mathscr{C}$; and*

(iv) *Totality: If a, b, and c are distinct, then either $(a, b, c) \in \mathscr{C}$ or $(c, b, a) \in \mathscr{C}$.*

We endow the set $[n]$ with the usual linear order $<$, and with induced cyclic order $\mathscr{C}_n$ defined by

$$(i, j, k) \in \mathscr{C}_n \quad \Longleftrightarrow \quad (i < j < k) \ \vee \ (j < k < i) \ \vee \ (k < i < j). \tag{1.1}$$

Figure 1.3 displays an example of the cyclic order in $\mathscr{C}_{11}$ over $[11] = \{0, 1, \ldots, 10\}$.

Definition 5 (Arc)**.** *A subset $\mathcal{I}$ of a cyclically ordered set $(\mathcal{X}, \mathscr{C})$ is said to be an arc if there are some $m, M \in \mathcal{X}$ such that for every $(m, k, M) \in \mathscr{C}$ it holds that $k \in \mathcal{I}$. We refer to such elements m and M as the borders of $\mathcal{I}$ and often write $\mathcal{I} = [m, M]$.*

REMARK 1. *Let $\mathcal{I}$ be an arc in $([n], \mathscr{C}_n)$, then either $\mathcal{I}$ or $\mathcal{I}^c$ (the complement) is an interval in $([n], <)$.*

Example 1. *In $([n], <, \mathscr{C}_n)$ the sets $\mathcal{I}_1 := \{0, 1, 2, 3\}$ and $\mathcal{I}_2 := \{n-3, n-2, n-1\}$ are both arc and interval. The set $\mathcal{I}_3 := \{n-2, n-1, 0, 1, 2\}$ is and arc but not an interval. We notice that $\mathcal{I}_3^c = \{3, 4, \ldots, n-4, n-3\}$ is an interval.*

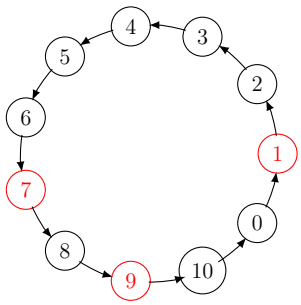


Figure 1.3. The cyclically ordered set $([11], \mathscr{C}_{11})$ represented as a directed cycle graph. In red the tern $(7, 9, 1)$ is present in the relation $\mathscr{C}_{11}$ since by cyclically permuting the tern we obtain the increasing sequence $(1, 7, 9)$. The tern $(7, 10, 9)$ is not in the relation since for every cyclic permutation the sequence is not increasing.

1.3.5. Group theory

A group is a set G endowed with a binary operation, called the product and denoted by $\cdot$, such that

(i) $a, b \in G$ implies that $a \cdot b \in G$ (closed)

(ii) $a, b, c \in G$ implies that $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (associative law).

(iii) There exists an element $e \in G$ such that $a \cdot e = e \cdot a = a$ for all $a \in G$ (the existence of an identity element in G).

(iv) For every $a \in G$ there exists an element $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$ (the existence of inverses in G)

Example 2. *The set* $\mathrm{Sym}(n)$ *together with the operation* $\circ$ *(function composition) is a group. The group* $(\mathrm{Sym}(n), \circ)$ *is called the symmetric group. The set of permutation matrices endowed with matrix multiplication as binary operation is a group. The identity permutation is denoted as* id

Definition 6. *(Herstein, 1975) A nonempty subset* H *of a group* G *is said to be a subgroup of* G *if, under the product in* G, H *itself forms a group.*

Definition 7. *(Herstein, 1975) A mapping* ϕ *from a group* G *into a group* $\bar{G}$ *is said to be a homomorphism if for all* $a, b \in G, \phi(a \cdot b) = \phi(a) \cdot \phi(b)$.

Definition 8. *(Herstein, 1975) A homomorphism ϕ from G into $\bar{G}$ is said to be an isomorphism if ϕ is one-to-one.*

The mapping ψ from the set of permutation matrices to the symmetric group such that $\psi(\Pi) = \pi$ with $\Pi(i,j) = 1 \leftrightarrow \pi(i) = j$ is in isomorphism.

Definition 9. *(Herstein, 1975) Two groups G, G^* are said to be isomorphic if there is an isomorphism of G onto G^*. In this case we write $G \cong G^*$.*

Definition 10. *(Herstein, 1975) Given a subset W of a group G. We write $\langle W \rangle$ to denote the set of all elements of G representable as a product of elements of W raised to positive, zero, or negative integer exponents.*

PROPOSITION 1. *(Herstein, 1975) $\langle W \rangle$ forms a subgroup of G and is the smallest subgroup of G containing W. In fact, $\langle W \rangle$ is the intersection of all the subgroups of G which contain W (this intersection is not vacuous since G is a subgroup of G which contains W). This group is called the group generated by W.*

Definition 11. *(Dummit & Foote, 2004) A group action of a group G on a set A is a map from $G \times A$ to A (written as $g \cdot a$, for all $g \in G$ and $a \in A$) satisfying the following properties:*

- $g_1 \cdot (g_2 \cdot a) = (g_1 g_2) \cdot a, \quad$ *for all* $g_1, g_2 \in G, a \in A$, *and*
- $e \cdot a = a, \quad$ *for all* $a \in A$

In that case we say 'G acts over the set A'

Example 3. *The symmetric group* $\mathrm{Sym}(n)$ *acts over the set $\mathbb{R}^{n \times n}$ of matrices through the following action*

$$\pi \cdot M \mapsto \Pi M \Pi^T$$

Notice that $M_\pi \triangleq \Pi M \Pi^T$ satisfies $M_\pi(i,j) = M(\pi(i), \pi(j))$. We call this special action 'conjugation'.

Definition 12. *(Dummit & Foote, 2004) Given a group G acting on a set A, and given some element $x \in A$, the set $\mathcal{O}_G(x) \triangleq \{g \cdot x | g \in G\}$ is called the orbit of x under the action of g*

Definition 13. *We write* $\mathbf{r}$ *to denote the reversing permutation defined by* $\mathbf{r}(i) := n - 1 - i$. *Also we write* $\mathbf{s}$ *to denote the cyclic shift permutation defined by* $\mathbf{s}(i) := i + 1 \bmod n$. *The matrix representations of these permutations will be denoted by* Π_r *and* Π_s, *respectively.*

We denote by Dih_n the dihedral group of $2n$ different symmetries of a regular polygon with n sides. We denote the cyclic group over n elements as $\mathbb{Z}/n\mathbb{Z}$. The particular case Dih_1 is defined as $\mathbb{Z}/2\mathbb{Z}$.

REMARK 2. *Note that the generated groups* $\langle \mathbf{r} \rangle$ *and* $\langle \mathbf{r}, \mathbf{s} \rangle$ *are isomorphic to the dihedral subgroups* Dih_1 *and* Dih_n *respectively. And similarly* $\mathbb{Z}/n\mathbb{Z}$ *is isomorphic to* $\langle \mathbf{s} \rangle$.

1.3.6. Graph theory

Given any set $\mathcal{X}$, a graph is a pair $G = (\mathcal{X}, \mathcal{E})$ such that $\mathcal{E}$ is a collection of pairs $\{x, y\}$ where $x, y \in \mathcal{X}$. We often denote as $V(G)$ the vertex set $\mathcal{X}$ (resp. $E(G)$ the edge set $\mathcal{E}$) when it is not given by context. The elements in $\mathcal{X}$ are called nodes (or vertices) and the elements in $\mathcal{E}$ are called edges. Two nodes x and y are said to be adjacent whenever there is an edge $\{x, y\} \in \mathcal{E}$. The set of adjacent nodes to x is denoted as $\mathcal{N}_G(x) \triangleq \{y \in \mathcal{X} : \{x, y\} \in \mathcal{E}\}$. The function $x \mapsto \mathcal{N}_G(x)$ is called neighbourhood. An hypergraph $\mathcal{H} = (\mathcal{X}, \mathcal{E})$ is a collection of nodes $\mathcal{X}$ together a collection of *hyperedges* $\mathcal{E}$, which are non empty subsets of $\mathcal{X}$.

Definition 14. *Given some graph G, a graph H with vertex set $V(G)$ is said to be a subgraph of G whenever $E(H) \subset E(G)$.*

Example 4. *(Bac, 1997) The ring graph $R_n = ([n], \mathcal{E})$ is the graph with edge set $\mathcal{E} = \{\{i, (i+1) \mod n\} : i \in [n]\}$. The path graph $P_n = ([n], \mathcal{E})$ is the graph with edge set $\mathcal{E} = \{\{i, i+1\} : i \in [n-1]\}$.*

Definition 15. *(Hartmanis, 1982) In graph theory, an homomorphism of graphs G and H is a function between the node sets of G and H*

$$f\colon V(G) \to V(H)$$

such that any two vertices u and v of G are adjacent in G whenever $f(u)$ and $f(v)$ are adjacent in H. Moreover, if a homomorphism $f : G \to H$ is a bijection whose inverse function is also a graph homomorphism, then f is said to be a graph isomorphism. Formally, $\{u, v\} \in E(G) \leftrightarrow \{f(u), f(v)\} \in E(H)$, for all pairs of vertices $u, v \in V(G)$

PROPOSITION 2. *Let H and G be two graphs. If there is a bijection f from $V(H)$ to $V(G)$ such that $f(\mathcal{N}_H(x)) \subset \mathcal{N}_G(x)$ for every $x \in V(G)$, then H is isomorphic to a subgraph of G*

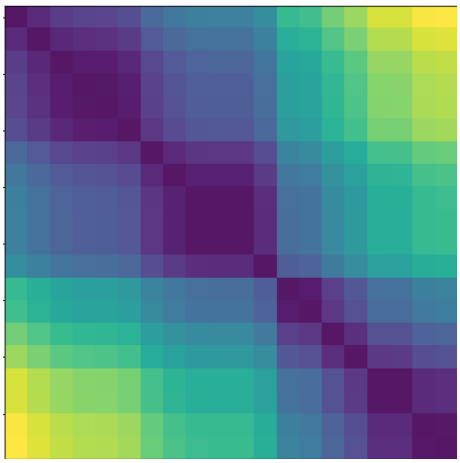
Definition 16. *(Atkins, Boman, & Hendrickson, 1998) Given symmetric, non-negative matrix A, the Laplacian matrix of A is defined as $L_A \triangleq D_A - A$, where D_A is a diagonal matrix with $D_A(i,i) \triangleq \sum_{j\in[n]} A(i,j)$.*

1.4. Mathematical formulation

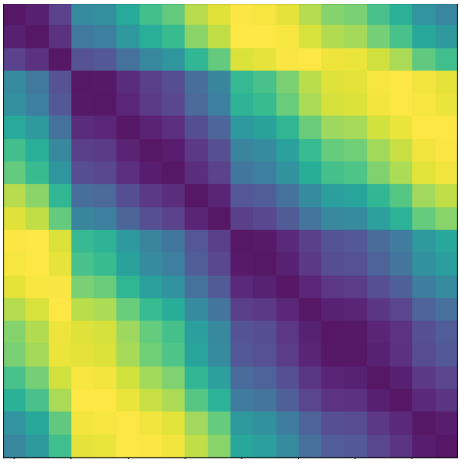
The seriation problem is traditionally modeled as a recognition problem. We assume that the observed dissimilarity matrix is a permuted version of a dissimilarity matrix belonging to a family of 'ordered matrices'. The task is to find a permutation under which the observed matrix is ordered. It is possible that this happens for more than one permutation. In those cases we consider all such permutations as solutions of the problem without making preferences between them as one would do in an optimization problem for instance, since all this permutations are consistent with our hypothesis *a priori*.

1.4.1. Abstract seriation problem

Given a class of 'ordered' matrices $\mathcal{M} \subseteq \mathbb{R}^{n\times n}$, define the class pre-$\mathcal{M}$ as the set of matrices A such that for some permutation matrix $\Pi \in \mathrm{Sym}(n)$ it holds that the matrix



(a) (Linear) Robinson dissimilarity matrix



(b) Circular Robinson dissimilarity matrix

$\Pi A \Pi^T$ (whose entry (i, j) is $A(\pi(i), \pi(j))$) is in $\mathcal{M}$. In terms of group theory, pre-$\mathcal{M}$ is the orbit of $\mathcal{M}$ under the action of $\mathrm{Sym}(n)$ by conjugation. The seriation problem can be stated as (Recanati et al., 2018)

$$\text{Given } A \in \text{pre-}\mathcal{M}, \text{ find } \Pi \in \mathrm{Sym}(n) \text{ such that } \Pi A \Pi^T \in \mathcal{M}.$$

The set $S_{\mathcal{M}}(A) \subset \mathrm{Sym}(n)$ of such permutations will be referred as the set of *solutions*. The classes of matrices where seriation have been mostly studied are called *Robinson* matrices.

1.4.2. Robinsonian dissimilarities: linear and circular seriation

There are two common classes of matrices that naturally arise when there is either a linear or circular underlying order: linear and circular Robinson dissimilarities. Despite that in the first section we implicitly defined dissimilarity matrices, we formalize it in the following definition.

Definition 17. *We say $D \in \mathbb{R}^{n \times n}$ is a dissimilarity matrix if it is*

(i) *Symmetric,*

(ii) *Non-negative and*

(iii) $D(i,i) = 0$ *for all* $i \in [n]$.

Similarly, a matrix $A \in \mathbb{R}^{n\times n}$ *is said to be a similarity (or affinity) matrix if there is non-increasing non-negative function* f *such that* $A(i,j) = f(D(i,j))$ *for some dissimilarity matrix* D

Definition 18. *(Linear Robinson Matrix) We say that a dissimilarity matrix* $D \in \mathbb{R}^{n\times n}$ *is (linear) Robinson iff for every* $i < j < k$

$$\begin{aligned} D(i,j) &\leq D(i,k) \\ D(j,k) &\leq D(i,k). \end{aligned} \tag{1.2}$$

If all inequalities hold strictly, we say D *is strictly linear Robinson.*

The set of linear Robinson dissimilarity matrices will be denoted as $\mathcal{L}_R$ whereas the set of strict linear Robinson dissimilarity matrices will be denoted as $\mathcal{L}_R^*$. To introduce circular Robinson dissimilarities, first we need to define unimodality

Definition 19. *(Unimodality) Let* $(\mathcal{X}, <)$ *be a linearly ordered set. A sequence* $\{x_j\}_{j=0}^{n-1} \subseteq \mathcal{X}$ *is said to be unimodal if there exists* $k \in [n]$ *such that*

$$\begin{aligned} x_i &\leq x_j \quad \forall i < j \leq k \\ x_l &\geq x_m \quad \forall k \leq l < m. \end{aligned} \tag{1.3}$$

Any k *that satisfies the properties above is called a mode. A unimodal sequence is said to be strictly unimodal if it has at most two consecutive modes, and before the mode it is strictly increasing, and after the mode it is strictly decreasing.*

REMARK 3. *If a sequence is strictly unimodal, then every subsequence is also strictly unimodal*

Definition 20. *(Circular Robinson Matrix) We say that a dissimilarity matrix* $D \in \mathbb{R}^{n\times n}$ *is circular Robinson iff for all* $i \in [n], \{D(i,(i+j) \bmod n)\}_{j=0}^{n-1}$ *is unimodal. If all such sequences are strictly unimodal, we say* D *is strictly circular Robinson. The set of*

circular and strict circular Robinson dissimilarity matrices will be denoted as $\mathcal{C}_R$ and $\mathcal{C}_R^*$, respectively.

Definition 18 states that when moving away from the diagonal in a given row or column of D, the entries are non-decreasing, whereas in Definition 20, a sequence of non-decreasing values is followed by a sequence of non-increasing values. For instance, the distance matrix of points embedded on a circle follows Definition 20 (Recanati et al., 2018). Figure 1.4.2 displays examples of such matrices.

It is easy to see that every linear Robinson matrix is also circular Robinson, hence $\mathcal{L}_R \subset \mathcal{C}_R$ and $\mathcal{L}_R^* \subset \mathcal{C}_R^*$. Elements in pre-$\mathcal{L}_R$ and pre-$\mathcal{C}_R$ are be said to be *Robinsonian* matrices. We call *(strict) linear seriation* to the seriation problem when the matrix class involved is the set of (strict) linear Robinson matrices and *(strict) circular seriation* when the matrix class is the set of (strict) circular Robinson matrices. The solutions in the context of linear and circular seriation are called *Robinson orderings*, which are all orderings consistent with the data.

Definition 21. *Let A be an affinity matrix A and let $D \triangleq I \cdot \|A\|_\infty - A$ be the dissimilarity matrix associated to A. We say a matrix A is a linear Robinson affinity matrix if D is a linear Robinson dissimilarity matrix. Similarly, A is said to be a circular Robinson affinity matrix if D is a circular Robinson dissimilarity matrix.*

REMARK 4. *Clearly, the seriation problems involving either affinity matrices or dissimilarity matrices are equivalent.*

1.5. State of the art

In this section, we present the main techniques that can be found in the literature to solve linear and circular seriation

1.5.1. Spectral embedding

Consider an affinity matrix A. The linear seriation problem can be addressed with the following combinatorial problem,

$$\text{minimize} \sum_{i,j\in[n]} A(i,j)\,|\pi(i)-\pi(j)|^2 \quad \text{such that} \quad \pi \in \text{Sym}(n) \tag{1.4}$$

The intuition is that in the optimum π^*, high values of $A(i,j)$ are compensated with small values of $|\pi^*(i)-\pi^*(j)|^2$, thus laying similar elements nearby. This problem is NP-hard (Barnard, Pothen, & Simon, 1995), thus a straightforward approach is not possible in practice. By a simple algebraic manipulation, the objective in (1.4) can be replaced for any $f \in \mathbb{R}^n$ with the quadratic form

$$\sum_{i,j\in[n]} A(i,j)|f(i)-f(j)|^2 = f^T L_A f \tag{1.5}$$

where L_A is the Laplacian of A (Recanati et al., 2018) (see Definition 16). Notice that $\mathbf{1} = (1,\dots,1)^T$ is an eigenvector of L_A associated to the eigenvalue $\lambda_0 = 0$. The spectral method (Atkins et al., 1998) consists in relaxing (1.4) by replacing the constraint $\pi \in \text{Sym}(n)$ with norm and orthogonality constraints, $\|\pi\|_2 = 1, \pi^T\mathbf{1} = 0$, to avoid the trivial solutions $\pi = 0$ and $\pi \propto \mathbf{1}$, yielding,

$$\text{minimize} \quad f^T L_A f \quad \text{such that} \quad \|f\|_2 = 1, f^T\mathbf{1} = 0 \tag{1.6}$$

This is an eigenvalue problem on L_A solved by f_1, the eigenvector associated to $\lambda_1 \geq 0$ the second smallest eigenvalue of L_A, called the *Fiedler vector* of A (Recanati et al., 2018). To retrieve an ordering from this optimization problem we have this key theorem:

Theorem 1 ((Atkins et al., 1998))**.** *Let A be a linear Robinson affinity matrix. Then, A has a monotone Fiedler vector.*

Recall that if (f, λ) are respectively an eigenvector and eigenvalue of L_A then

$$\begin{aligned} L_A f &= \lambda f \\ \iff \Pi L_A f &= \lambda \Pi f \\ \iff \Pi L_A I_n f &= \lambda \Pi f \\ \iff \Pi L_A \Pi^T \Pi f &= \lambda \Pi f \\ \iff \Pi (D_A - A) \Pi^T \Pi f &= \lambda \Pi f \\ \iff (D_{A_\pi} - A_\pi) \Pi f &= \lambda \Pi f \\ \iff L_{A_\pi} f^\pi &= \lambda f^\pi \end{aligned} \tag{1.7}$$

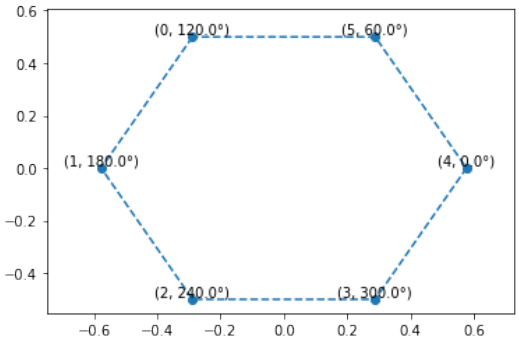
where $f_\pi \triangleq \Pi f$ satisfies $f^\pi(i) = f(\pi(i))$. Thus if $A \in$ pre-$\mathcal{L}_R$ and A has a simple Fiedler value, the permutation obtained by sorting the entries of the Fielder vector of A is a Robinson ordering. Theorem 1 is a consequence of the Perron-Frobenious theorem[2]. This result can also be exploited to obtain all Robinson orderings. The details can be found in (Atkins et al., 1998).

On the aim of generalizing the above procedure to circular seriation, one could consider computing the two largest non-trivial eigenvectors of the Laplacian matrix L_A of a circular Robinson affinity matrix A. This is exactly what is proposed in (Coifman et al., 2008) and (Recanati et al., 2018). Consider the following optimization problem,

$$\begin{array}{ll} \text{minimize} & \sum_{i,j \in [n]} A(i,j) \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|_2^2 \\ \text{such that} & \Phi = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T \right)^T \in \mathbb{R}^{n \times d}, \Phi^T \Phi = \mathbf{I}_d, \Phi^T \mathbf{1}_n = \mathbf{0}_d \end{array} \tag{1.8}$$

Like before, the objective in (1.8) can be written as $\operatorname{trace}\left(\Phi^T L_A \Phi\right)$ (see (Belkin & Niyogi, 2001)), yielding a multidimensional eigenvector problem. Once again the interpretation is that similar elements are mapped nearby in $\mathbb{R}^d$ and dissimilar placed apart. If

[2]For further information about this theorem see (Pillai, Suel, & Cha, 2005)



(a) f_1 vs f_2

$$\begin{bmatrix} 3 & 2 & 1 & 0 & 1 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1 \\ 1 & 2 & 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 & 2 & 1 \\ 1 & 0 & 1 & 2 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 \end{bmatrix}$$

(b) Strict circular R-matrix

Figure 1.4. Ideal example for the spectral method. In the right, a strict circular Robinson affinity matrix $A \in \mathbb{R}^{6\times 6}$. In the left, f_1 and f_2, the two first non-trivial eigenvectors of the Laplacian matrix of A. By computing the angles of the points $(f_1(i), f_2(i)) \in \mathbb{R}^2$ its possible to retrieve the circular order.

the original objects where nearby a closed continuum, by taking $d = 2$ one could expect the embedded points to lie in (close to) a circle. Therefore, the cyclic order could be retrieved by computing the angles from the points and sorting them[3]. There are asymptotic justifications of this procedure due to the connection of the Laplacian matrix to the continuous Laplace-Beltrami operator over a 1-manifold (Coifman et al., 2008). In spite of this, the theoretical guarantees for this procedure in a finite sample case are very limited. Only for highly structured affinity matrices this procedure has been proved recover a correct ordering (for instance see (Recanati et al., 2018)). Hence, characterizing instances where the spectral method recovers the true ordering is an open problem. An ideal example together with its spectral embedding is presented in Figure 1.4.

Negative evidence for the spectral method in the circular case

In this section we present some examples that suggest that the spectral method does not always recover circular orders successfully in the general circular-Robinson setting.

[3]If f_1 and f_2 are the two first non-trivial eigenvectors of L_A, then sort the values of $\theta(i) \triangleq \operatorname{atan2}(f_2(i), f_1(i))$

Given an affinity matrix $A \in \mathbb{R}^{n\times n}$, consider the mapping $\theta_A : [n] \to [0, 2\pi]$ such that each the i-th element is mapped to $\theta_A(i) \triangleq \mathrm{atan2}(f_2(i), f_1(i))$ where f_1 and f_2 are the two first non-trivial eigenvectors of L_A. We endow $[0, 2\pi]$ with the usual cyclic order $\mathscr{C}_{[0,2\pi]}$ such that

$$(\alpha, \beta, \gamma) \in \mathscr{C}_{[0,2\pi]} \quad \iff \quad (\alpha < \beta < \gamma) \ \vee \ (\beta < \gamma < \alpha) \ \vee \ (\gamma < \alpha < \beta). \tag{1.9}$$

For instance, the vector $(60°, 120°, 10°)$ is in $\mathscr{C}_{[0,2\pi]}$. Ideally, one would expect that if A is circular Robinson, then

$$(\theta_A(i), \theta_A(j), \theta_A(k)) \in \mathscr{C}_{[0,2\pi]} \leftrightarrow (i, j, k) \in \mathscr{C}_n \tag{1.10}$$

This means that ordering angles yields a Robinson ordering and, conversely, Robinson orderings yields ordered angles. A simple exercise is provided in the next section that shows this cannot always be true.

Existence of non-trivial symmetries

It is intuitive that since reversing an ordered sequence produces a new ordered sequence, then reversing the entries of a Robinson matrix should produce a new Robinson matrix. This is also valid for circular Robinson matrices. In the circular case we would also expect that circular permutations should produce new circular Robinson matrices. The following result formalizes this intuition.

Theorem 2. *(Armstrong et al., 2021) The class of linear Robinson matrices is invariant under the action of* Dih_1 *by conjugation and the class of circular Robinson matrices is invariant under the action of* Dih_n *by conjugation.*

Therefore, given any $A \in \mathcal{C}_R$ we have that $\mathrm{Dih}_n \subset S_{\mathcal{C}_R}(A)$. We call this solutions the *trivial symmetries* of the affinity matrix, since they are always present, whereas any $\pi \in S_{\mathcal{C}_R}(A) \setminus \mathrm{Dih}$ is called a *non-trivial symmetry* of A. It is not immediately clear that even in the strict case non trivial symmetries may exist (see for instance Figure 1.5).

By (1.7), given some permutation π we have that $\theta_A(\pi(i)) = \theta_{A_\pi}(i)$ for every $i \in [n]$. It is clear that if $\pi \in \mathrm{Dih}_n$, then ordering θ_A yields a Robinson ordering whenever ordering $\theta_A \circ \pi$ yields a Robinson ordering. However, if π is a non-trivial symmetry, both orderings cannot be represented in $[0, 2\pi]$ unless they are both collapsed somehow by θ_A. For instance, consider the affinity matrix A available in Figure 1.5. The matrix $A \in \mathbb{R}^{6\times 6}$ has only one non-trivial symmetry: $\pi \triangleq (0, 1, 2, 5, 4, 3)$. Since is π is a Robinson ordering, assuming (1.10) we would get that for every $(i, j, k) \in \mathscr{C}_n$ it holds that $(\theta_A(i), \theta_A(j), \theta_A(k)) \in \mathscr{C}_{[0,2\pi]}$ and $(\theta_A(\pi(i)), \theta_A(\pi(i)), \theta_A(\pi(i))) \in \mathscr{C}_{[0,2\pi]}$. In our example, $(\theta_A(3), \theta_A(4), \theta_A(5)) \in \mathscr{C}_{[0,2\pi]} \wedge (\theta_A(5), \theta_A(4), \theta_A(3)) \in \mathscr{C}_{[0,2\pi]}$, which implies that $\theta_A(3) = \theta_A(4) = \theta_A(5)$.

More generally, suppose that given $A \in \mathbb{R}^{2n\times 2n}$ it holds that $A, A_\pi \in \mathcal{C}_R$ for

$$\pi = (0, 1, \ldots, n-2, n-1, 2n-1, 2n-2, \ldots, n+1, n)$$

Then, the only way that in which θ_A can be consistent with both orderings is if $\theta_A(i)$ is constant for all $i = n, \ldots, 2n-1$. But, this implies that the elements $n, \ldots, 2n-1$ are not ordered which is not necessarily true. For instance in Figure 1.6, there are only two orderings and arbitrary permutations of the 'second half' of the element do not yield Robinson orderings.

Therefore, if A possess multiple non-trivial symmetries at most we can expect three things:

- Ordering θ_A yields to one of the Robinson orderings

- If A possess multiples non-trivial symmetries, then θ_A collapses in some values in order to represent the orderings
- Let λ_1, λ_2 be the minimum positive eigenvalues, then $\dim(\textit{ker}(A - \lambda_1 \cdot I)) + \dim(\textit{ker}(A - \lambda_2 \cdot I)) > 2$. Which implies, that the embedding is not 2-dimensional, but somehow more complex in order to represent all orderings.

In either case, this shows an inconvenience for choosing this method for circular seriation. In the first case, we loose possible solutions. In the second case, its possible that we obtain wrong solutions. In the third case, it is not clear how to design a rule (such as computing the angles in the 2D case) that correctly recovers the orderings.

To summarize, we have shown that if there non-trivial symmetries, then the spectral method is likely to fail. This is not a necessary condition for the method to fail since the example in Figure 1.7 does not have non-trivial symmetries and sorting the values of θ_A does not yield a Robinson ordering.

1.5.2. Seriation as an instance of the QAP

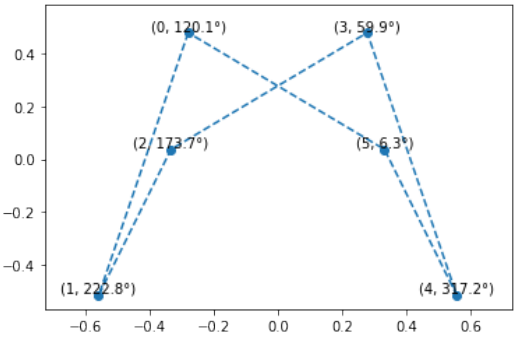
Some recent works aim to solve circular seriation also by starting from (1.4) but by taking a different direction. In (Evangelopoulos et al., 2020) the authors propose to solve the following optimization problem,

$$\begin{aligned} \text{minimize} \quad \text{QAP}(A, B) &\triangleq \text{trace}\left(\Pi A \Pi^T B\right) \\ &= \textstyle\sum_{i,j \in [n]} A(\pi(i), \pi(j)) B(i, j) \end{aligned} \tag{1.11}$$

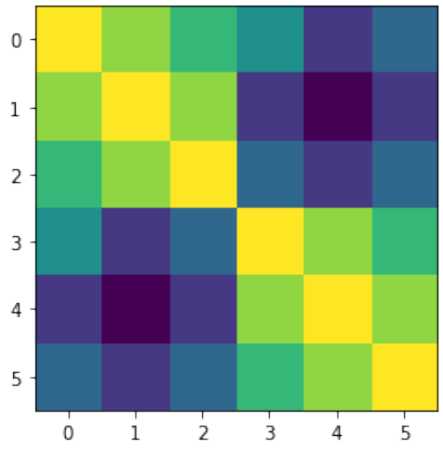
for a template

$$B(i,j) = \begin{cases} |i-j|, & \text{if } |i-j| \leq \left\lfloor \frac{n-1}{2} \right\rfloor \\ n - 1 - |i-j|, & \text{if } |i-j| > \left\lfloor \frac{n-1}{2} \right\rfloor \end{cases} \tag{1.12}$$

The above acts as a circular seriation template where the elements of the first row (column) increase monotonically while moving to the right (bottom) until the $\left\lfloor \frac{n-1}{2} \right\rfloor$-th element and then decrease again monotonically until the end of the row (column), and fold



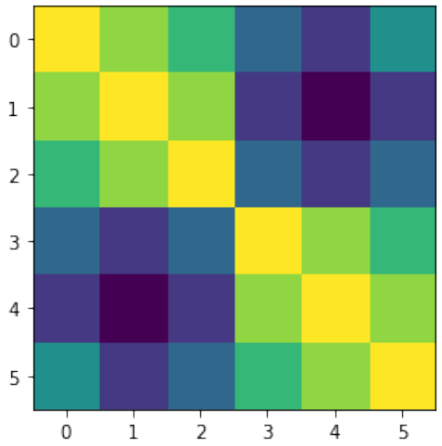
(a) f_1 vs f_2



(b) Color representation of A

$$\begin{bmatrix} 12 & 11 & 10 & 9 & 7 & 8 \\ 11 & 12 & 11 & 7 & 6 & 7 \\ 10 & 11 & 12 & 8 & 7 & 8 \\ 9 & 7 & 8 & 12 & 11 & 10 \\ 7 & 6 & 7 & 11 & 12 & 11 \\ 8 & 7 & 8 & 10 & 11 & 12 \end{bmatrix}$$

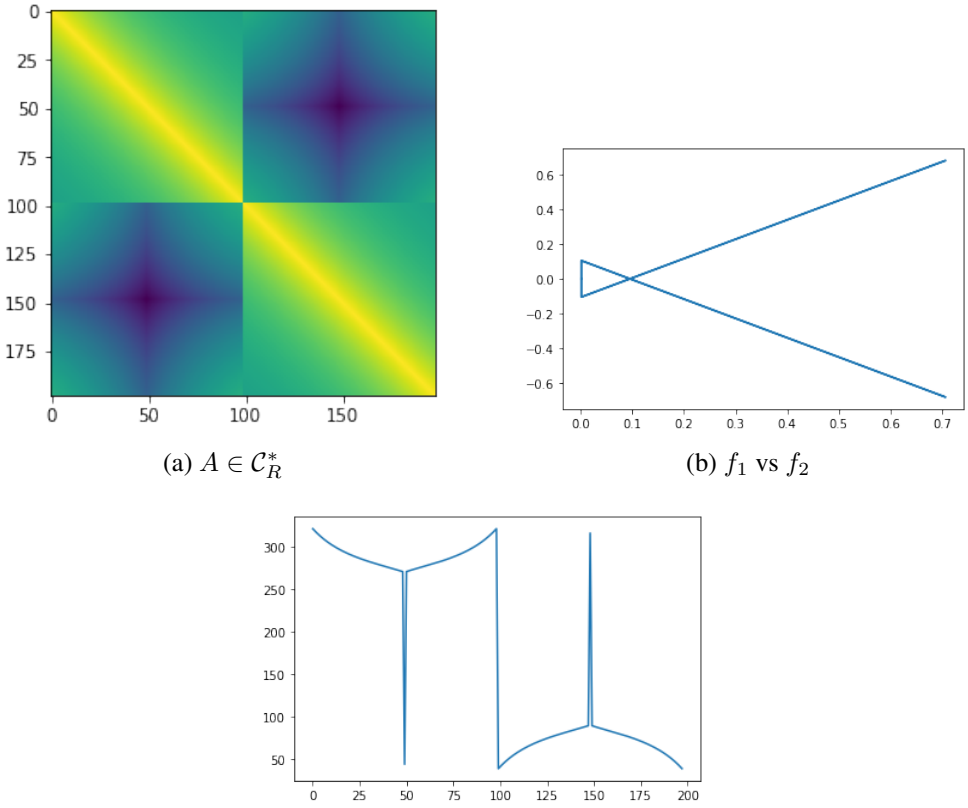
(c) Strict circular R-matrix



(d) Color representation of A_π

Figure 1.5. A strict-circular Robinson matrix which posses non-trivial symmetries: $A \in \mathcal{C}_R^* \wedge A_\pi \in \mathcal{C}_R^*$ for $\pi = (0, 1, 2, 5, 4, 3)$

back from the left (top) of the matrix (Evangelopoulos et al., 2020). A first inconvenience of this approach is that it requires to set the turning point in advance (in this example at $\lfloor \frac{n-1}{2} \rfloor$). Correctly fixing this parameter requires information *a priori* about the shape of the continuum where the points lie nearby. A second inconvenience of this approach is that (1.11) is an instance of the quadratic assignment problem (QAP) which is NP-hard. To this day, no exact algorithm can solve general QAP-instances of size $n > 20$ in reasonable computational time (Pitsoulis & Pardalos, 2009). This is why in (Evangelopoulos et al.,



(a) $A \in \mathcal{C}_R^*$

(b) f_1 vs f_2

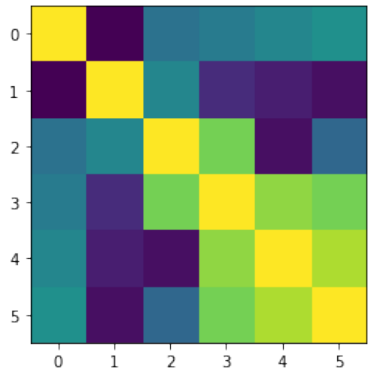
(c) θ_A in degrees

Figure 1.6. A 200×200 strict circular Robinson affinity matrix which has a unique non-trivial symmetry.

2020) an heuristic approximation is proposed for solving (1.11) which has no theoretical guarantees.

1.5.3. The definition of circular Robinson dissimilarities

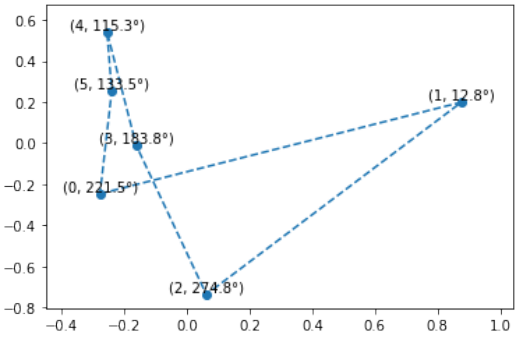
There is general agreement on the definition of linear Robinson dissimilarities in the literature. However, although most of the generalizations to the circular case follow the same intuition, the mathematical formulations present subtle, nevertheless important, differences. The intuition in all cases is that the entries of each row increase monotonically while moving to the right until some specific element and then decrease again monotonically until the end of each row and fold back from the left of the matrix, just as the



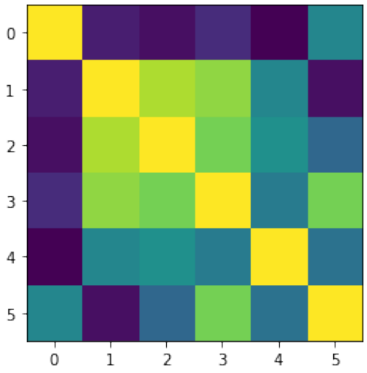
(a) Color representation

$$\begin{bmatrix} 25 & 1 & 10 & 11 & 12 & 13 \\ 1 & 25 & 12 & 4 & 3 & 2 \\ 10 & 12 & 25 & 20 & 2 & 9 \\ 11 & 4 & 20 & 25 & 21 & 20 \\ 12 & 3 & 2 & 21 & 25 & 22 \\ 13 & 2 & 9 & 20 & 22 & 25 \end{bmatrix}$$

(b) Strict circular R-matrix



(c) f_1 vs. f_2



(d) Permuted matrix obtained by ordering θ_A

Figure 1.7. An instance where the spectral method fails. In 1.7b, a strict circular Robinson affinity matrix $A \in \mathbb{R}^{6\times 6}$. In 1.7a, a color representation of the matrix. In 1.7c, f_1 and f_2, the two first non-trivial eigenvectors of the Laplacian matrix of A. By computing the angles of the points $(f_1(i), f_2(i)) \in \mathbb{R}^2$ we obtain the sequence $\theta_A = (221.5°, 12.7°, 274.7°, 183.8°, 115.3°, 133.4°)$ which is not ordered in $\mathscr{C}_{[0,2\pi]}$. Finally, in 1.7d the matrix $A_\pi \notin \mathcal{C}_R$ where π is the permutation obtained by sorting θ_A in increasing order. This affinity matrix has no non-trivial symmetries, i.e. $S_{\mathcal{C}_R}(A) = \mathrm{Dih}_6$

distance matrix of points embedded in a circle (this condition also holds for the columns by the symmetry of the matrix). The first mathematical formalization is due to Hubert and is as follows (Hubert, Arabie, & Meulman, 1998). For $1 \leq i \leq n-3$ and $i+1 < j \leq n-1$

$$\text{if } D(i+1,j) \leq D(i,j+1) \text{ then } D(i+1,j) \leq D(i,j)$$

$$\text{and } D(i+1,j) \leq D(i+1,j+1)$$

$$\text{if } D(i+1,j) \geq D(i,j+1) \text{ then } D(i,j) \geq D(i,j+1)$$

$$\text{and } D(i+1,j+1) \geq D(i,j+1)$$

and, for $2 \leq i \leq n-2$

$$\text{if } D(i+1,n) \leq D(i,1) \text{ then } D(i+1,n) \leq D(i,n)$$

$$\text{and } D(i+1,n) \leq D(i+1,1)$$

$$\text{if } D(i+1,n) \geq D(i,1) \text{ then } D(i,n) \geq D(i,1)$$

$$\text{and } D(i+1,1) \geq D(i,1)$$

In addition to being complex, this definition has two main problems. The first problem is that it allows bimodality within each row (modulo n). For instance consider the following dissimilarity matrix

$$Q = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 3 & 1 & 1 \\ 1 & 3 & 0 & 3 & 1 \\ 1 & 1 & 3 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}. \tag{1.13}$$

The matrix Q satisfies the conditions proposed by (Hubert et al., 1998). Notice that in the middle row we have two *turning points* corresponding to the entries $(2,1)$ and $(2,3)$, where the dissimilarity attains the values 3 (recall that arrays start at index 0). This bimodality is inconsistent with a circular embedding (where there is a unique turning point). A second problem with the definition (which is a consequence of the first problem) is that balls $B_r(i) \triangleq \{j \in [n] : D(i,j) \leq r\}$ do not correspond to arcs in Robinson

orderings (*disconnected classes* in the terminology of (Brucker & Osswald, 2008)). For instance the ball of radius $r = 2$ centered at the element at index 2 corresponds to $B_2(2) = \{0, 2, 4\}$ which is not an arc nor an interval.

A second definition available in the literature is due to Recanati et al. In (Recanati et al., 2018), a dissimilarity matrix D is said to be circular Robinson iff for all $i \in [n], \{D(i,j)\}_{j=0}^{i}$ and $\{D(i,j)\}_{i=j}^{n-1}$ are unimodal. Again the intuition that motivated this definition is correct, but this definition not only has the same problems that the previous definition (since for instance the matrix Q satisfies the definition), but also there is an additional contradictory feature: the lack of invariance to cyclic permutations. For instance consider the cyclic permutation defined by $\pi(i) = i + 2 \bmod n$. By permuting Q by π we obtain

$$Q_\pi = \begin{bmatrix} 0 & 1 & 1 & 1 & 3 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 3 \\ 3 & 1 & 1 & 3 & 0 \end{bmatrix} \tag{1.14}$$

and notice that $\{D_\pi(4,j)\}_{j=0}^{4}$ is not an unimodal sequence. This shows that if D is circular Robinson in the sense of (Recanati et al., 2018), then D_π is not necessarily circular Robinson when π is a circular permutation. This is contradicting since cyclic permutations preserve cyclic orders (by shifting the circle, it remains ordered).

In (Brucker & Osswald, 2008), dissimilarities whose balls correspond to arcs are studied. A definition which we show to be equivalent to Definition 20 is mentioned but this paper mostly focuses on *precircular* dissimilarities which is a particular case of the previous definition.

1.6. Main objectives and contributions

The main contributions of this thesis can be enumerated as follows:

- *Establish a tractable and meaningful definition for the circular seriation problem.* As opposed to the linear case, where a well established definition has been accepted (Chepoi & Fichet, 1997) and optimal algorithms have been designed (Préa & Fortin, 2014), the circular seriation problem has several different definitions, with all of them having some advantages and some disadvantages. Therefore, our first goal is to define a simple definition for circular seriation that captures the intuitive properties of a circular embedding, and that is suitable for efficient algorithms. The first contribution is to show that circular seriation can be solved in polynomial time under the proposed definition.
- *Optimal algorithm in the strict case*. When the data is continuous, a natural assumption for seriation is that inequalities hold strictly, i.e. strict seriation. Under this setting, the complexity is substantially reduced. This has been known for many years in the linear case, since an optimal algorithm for strict linear seriation was introduced 17 years before the first optimal algorithm for the general non-strict case (Chepoi & Fichet, 1997). However, in the circular case this problem was still open. Our second contribution is an $\mathcal{O}(n^2)$ time algorithm that solves both linear and circular strict seriation: the *Recursive Seriation Algorithm*.
- *Generative model*. The seriation problem has traditionally been modelled as a recognition problem, where all Robinson orderings are considered as solutions without establishing preferences between them. A philosophically different way of looking into the problem is to assume that the observed data was sampled from a continuous closed curve where the correct ordering is given by the parametrization of the curve. In this case, one would be interested in finding such ordering or, at least, being capable to bound the error of the obtained ordering. Our third contribution is designing a generative model for circular seriation where it is possible to bound the diameter of the solution set and obtain sample complexity bounds for the reconstruction of the ordering.

- *Numerical validation*. We present numerical experiments where the *Recursive Seriation Algorithm* successfully reconstructs the original order of the data in both synthetic and real-world data. This experiments are available at the Appendix.

1.7. Future Challenges

Along this thesis, a series of unsolved problem are raised. These constitute future lines of investigation. We point out the main ones:

- *Optimal algorithm in the circular (non-strict) case*. It is known in the literature that the linear (non strict) seriation problem can be solved in optimal $\mathcal{O}(n^2)$ time complexity. In (Préa & Fortin, 2014), an algorithm is presented for linear seriation which solves the problem by computing the ball hypergraph partially, and then carefully modifying the PQ-tree resulting from this hypergraph. There are some key procedures in this paper such us *partition refinement* for which a generalization to the circular case is not obvious. Therefore, the question whether applying a similar idea in the circular case is to the best of our knowledge still open. A future challenge would be finding an $\mathcal{O}(n^2)$ time complexity algorithm for (non-strict) circular seriation or (in the negative) proving a superquadratic lower complexity bound for general circular seriation.
- *Necessary and sufficient condition for the spectral method*. For some highly structured circular Robinson matrices such as circulant Toeplitz matrices the spectral method is guaranteed to successfully recover a Robinson ordering (Recanati et al., 2018). However, as we saw in section 1.5.1 there are many cases in which the method does not work. Another line of future research would be to completely characterize the set of circular Robinson matrices in which the embedding θ_A preserves cyclic orders.

REFERENCES

Armstrong, S., Guzmán, C., & Sing-Long, C. (2021). An optimal algorithm for strict circular seriation. *Preprint, under review in SIMODS*.

Atkins, J. E., Boman, E. G., & Hendrickson, B. (1998). A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, *28*(1), 297–310.

Bac, R. (1997). *Structure of graph homomorphisms* (Unpublished doctoral dissertation). Ph. D. Thesis, Simon Fraser University.

Barnard, S. T., Pothen, A., & Simon, H. (1995). A spectral algorithm for envelope reduction of sparse matrices. *Numerical linear algebra with applications*, *2*(4), 317–334.

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, *14*, 585–591.

Brucker, F., & Osswald, C. (2008). Hypercycles and dissimilarities. *Journal of Classification, accepté*.

Burris, S., & Sankappanavar, H. (2012). A course in universal algebra. *URl: https://www. math. uwaterloo. ca/˜ snburris/htdocs/UALG/univ-algebra2012. pdf (visited on 01/04/2016)(cit. on pp. 1, 5–8)*.

Chepoi, V., & Fichet, B. (1997). Recognition of robinsonian dissimilarities. *Journal of Classification*, *14*(2), 311–325.

Coifman, R. R., Shkolnisky, Y., Sigworth, F. J., & Singer, A. (2008). Graph laplacian tomography from unknown random projections. *IEEE Transactions on Image Processing*, *17*(10), 1891–1899.

Deans, S. R. (2007). *The radon transform and some of its applications*. Courier Corporation.

Derbyshire, S. (2016). *This image shows how points in the cantor set correspond to binary sequences.* Retrieved from `https://commons.wikimedia.org/wiki/File:Cantor_set_binary_tree.svg` (File under CC BY-SA 4.0

license: Cantor set binary tree.svg)

Dummit, D. S., & Foote, R. M. (2004). *Abstract algebra* (Vol. 3). Wiley Hoboken.

Evangelopoulos, X., Brockmeier, A. J., Mu, T., & Goulermas, J. Y. (2020). Circular object arrangement using spherical embeddings. *Pattern Recognition*, *103*, 107192.

Hartmanis, J. (1982). Computers and intractability: a guide to the theory of np-completeness (michael r. garey and david s. johnson). *Siam Review*, *24*(1), 90.

Herstein, I. (1975). Topics in algebra-john wiley & sons. *Inc., USA*.

Hubert, L., Arabie, P., & Meulman, J. (1998). Graph-theoretic representations for proximity matrices through strongly-anti-robinson or circular strongly-anti-robinson matrices. *Psychometrika*, *63*(4), 341–358.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93.

Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(2), 70–91.

Ma, R., Tony Cai, T., & Li, H. (2020). Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, 1–15.

Pillai, S. U., Suel, T., & Cha, S. (2005). The perron-frobenius theorem: some of its applications. *IEEE Signal Processing Magazine*, *22*(2), 62–75.

Pitsoulis, L. S., & Pardalos, P. M. (2009). *Quadratic assignment problem.*

Préa, P., & Fortin, D. (2014). An optimal algorithm to recognize robinsonian dissimilarities. *Journal of Classification*, *31*(3), 351–385.

Recanati, A. (2018). *Relaxations of the seriation problem and applications to de novo genome assembly* (Unpublished doctoral dissertation). Universitè de recherche Paris Sciences Lettres.

Recanati, A., Brüls, T., & d'Aspremont, A. (2017). A spectral algorithm for fast de novo layout of uncorrected long nanopore reads. *Bioinformatics*, *33*(20), 3188–3194.

Recanati, A., Kerdreux, T., & d'Aspremont, A. (2018). Reconstructing latent orderings by spectral clustering. *arXiv preprint arXiv:1807.07122*.

Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *American antiquity*, *16*(4), 293–301.

Royden, H. L., & Fitzpatrick, P. (1988). *Real analysis* (Vol. 32). Macmillan New York.

Wikimedia Commons. (2007). *Cantor ternary set, in seven iterations.* Retrieved from https://commons.wikimedia.org/wiki/File:Cantor_set_in_seven_iterations.svg (Public domain file: Cantor set in seven iterations.svg)

APPENDIX

A. NUMERICAL EXPERIMENTS

A.1. Synthetic experiment: the Cantor set

The algorithm presented in (Armstrong et al., 2021) performs seriation by merging nearest-neighbours. A limit case for such algorithm is when we only merge two neighbours at a time. This is exactly what happens when we consider the boundary of the intermediate Cantor sets, which we present in what follows (see Figure A.2). The Cantor ternary set $\mathcal{C}_\infty$ consists of all numbers in $[0, 1]$ that have a ternary expansion (Royden & Fitzpatrick, 1988). Consider the sets defined recursively

$$C_n \triangleq \frac{C_{n-1}}{3} \cup \left(\frac{2}{3} + \frac{C_{n-1}}{3} \right) \text{ for } n \geq 1, \text{ and } C_0 \triangleq [0, 1].$$

Then the cantor set is defined as $\mathcal{C}_\infty \triangleq \bigcap_{n=1}^{\infty} C_n$. For every integer $n \geq 1$, we define the intermediate Cantor sets as follows

$$\mathcal{C}_n \triangleq \bigcap_{k=1}^{n} C_k.$$

The sets $\mathcal{C}_n$ are illustrated in Figure A.1 for $n = 1, \ldots, 6$.

Consider the sequence $x_1 < x_2 < \cdots < x_{2^k}$ in $[0, 1]$ such that $\partial \mathcal{C}_k = \bigcup_{i=1}^{2^k} \{x_i\}$ (i.e. the elements at the boundary of the intervals at obtained at the k-th recursion of the construction of the Cantor set). In the optimal algorithm presented in (Armstrong et al., 2021), the maximum recursion depth is bounded by $\log_2(n)$ where n is the number of initial points. For instance, if the input matrix is $\mathbf{d}(x_i, x_j) \in \mathbb{R}^{2^k \times 2^k}$, then this algorithm correctly orders the sequence in at most $\log_2(2^k) = k$ iterations. A circular Robinson analog for such dissimilarity is given by $D^{\text{circ}}(i, j) \triangleq \mathbf{d}(f(x_i), f(x_j))$ where f corresponds to the mapping $\theta \mapsto (\cos(2\pi\theta), \sin(2\pi\theta))$. To test the tightness of such bound, in Figure A.3 we present the intermediate steps of the algorithm with a random permutation of D^{circ} as input.

Figure A.1. First six iterations of the construction of the Cantor set (i.e. the intermediate Cantor sets) (Wikimedia Commons, 2007).

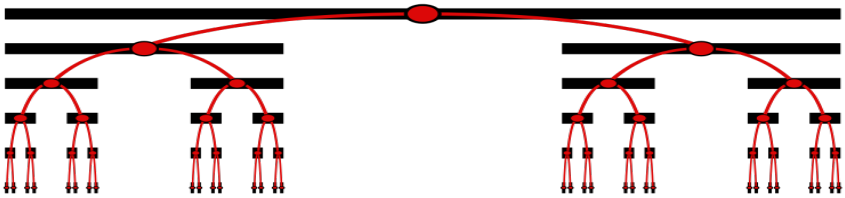
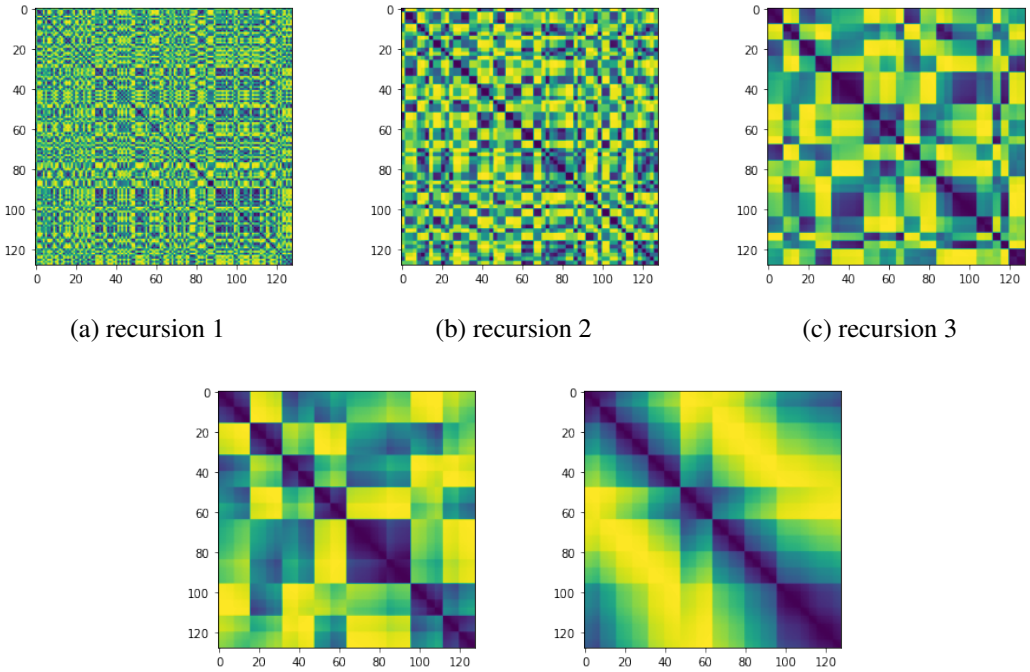


Figure A.2. Nearest neighbours relation between the the elements in $\partial\mathcal{C}_6$ (Derbyshire, 2016).



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

(d) recursion 4 (e) recursion 5

Figure A.3. Iterations of the algorithm from (Armstrong et al., 2021) with input $\Pi D^{\text{circ}}\Pi^T \in \mathbb{R}^{2^7\times 2^7}$ for some random permutation matrix Π

A.2. *Real-world* application: tomographic reconstruction

We present an example of the tomographic reconstruction problem introduced in Section 1.2.1. The objective of this experiment is to reconstruct a virus's density $\rho(x, y)$ from its projections. The density is available at Figure A.4. Let $\mathcal{R}\rho(t, \theta)$ be the Radon transform at angle $\theta \in \{1^\circ, 2^\circ \ldots, 180^\circ\}$ (see Figure A.5). Given some unknown permutation π, let $\mathcal{R}\rho(t, \theta_\pi)$ be the permuted collection of projections (the permuted Radon transform). An example is displayed at Figure A.6 for some random permutation π. A *naive* approach for the reconstruction is to invert $\mathcal{R}\rho(t, \theta_\pi)$ by assuming that the angle vector θ_π is $\theta_{\text{id}} \triangleq (1^\circ, 2^\circ \ldots, 180^\circ)$. The obtained density from this (wrong) approach in our example can be found at Figure A.7. To correctly solve this problem we first need to find the latent circular ordering of the projections. Let D_π be the dissimilarity matrix given by $D_\pi(i, j) \triangleq \|\mathcal{R}\rho(\cdot, \theta_\pi(i)) - \mathcal{R}\rho(\cdot, \theta_\pi(j))\|_1$. Such dissimilarity matrix is displayed at Figure A.8. Let σ be the ordering (permutation) obtained from the algorithm presented in (Armstrong et al., 2021) with input D_π. In Figure A.9 we present the dissimilarity matrix $\Sigma D_\pi \Sigma^T$ where Σ is the matrix representation of the permutation σ. By inverting the sorted Radon transform $\mathcal{R}\rho(t, \theta_{\sigma\circ\pi})$ we obtain the correct objects density as in Figure A.10.

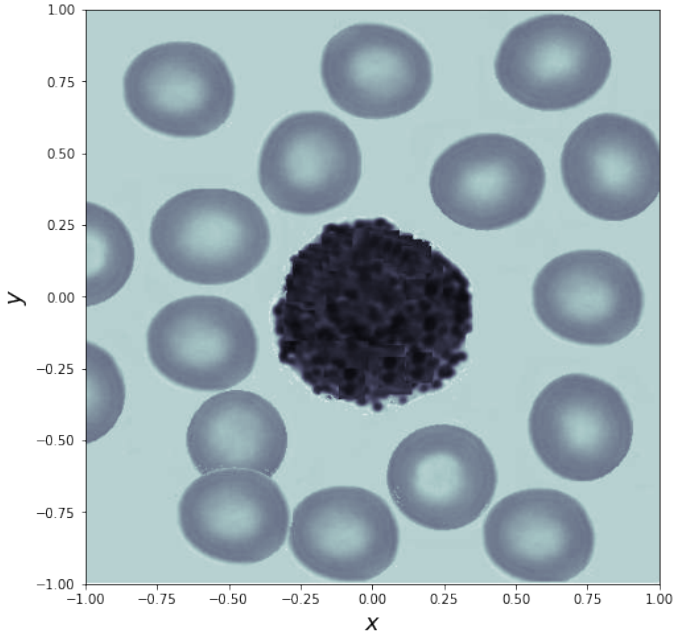


Figure A.4. Original object's density

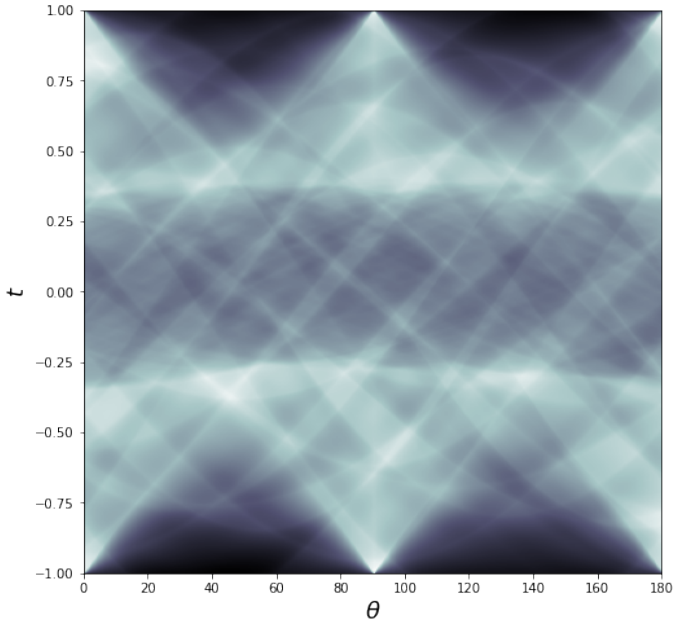


Figure A.5. Radon transform $\mathcal{R}\rho(t, \theta)$ of the density ρ

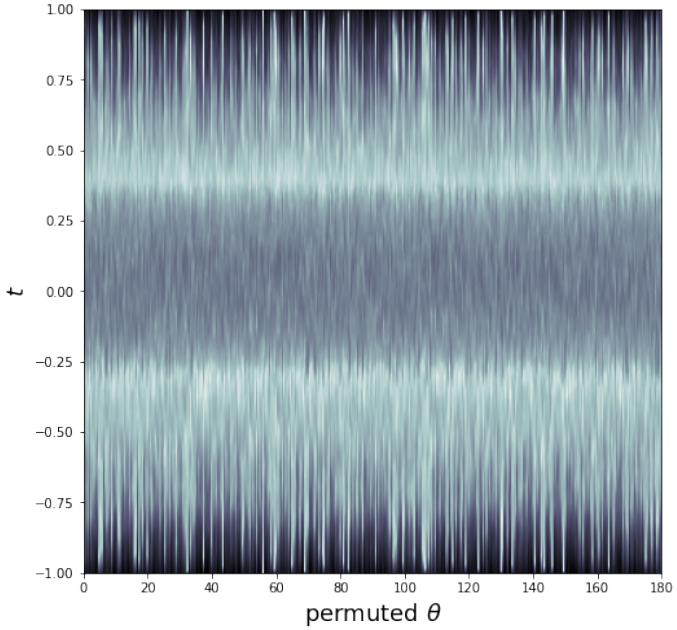


Figure A.6. Permuted Radon transform $\mathcal{R}\rho(t, \theta_\pi)$

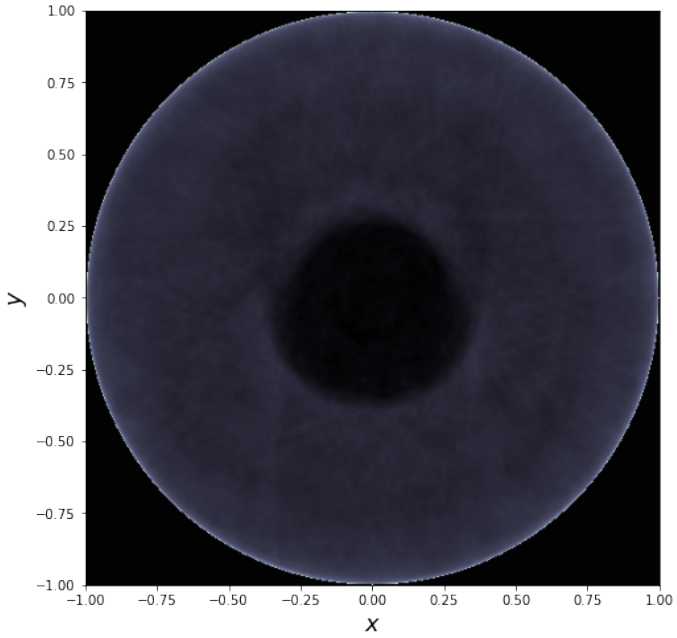


Figure A.7. Density obtained by a straightforward inversion $\mathcal{R}\rho(t, \theta_\pi)$

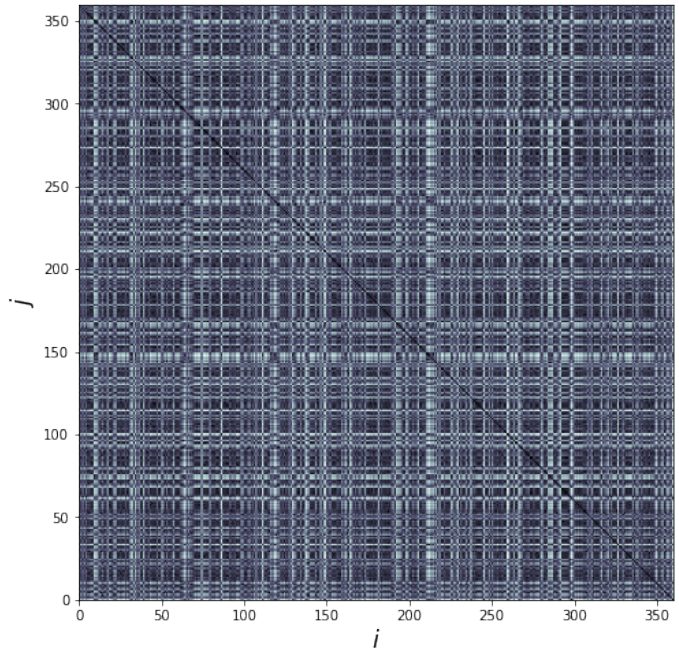


Figure A.8. Dissimilarity matrix between all pairs of projections

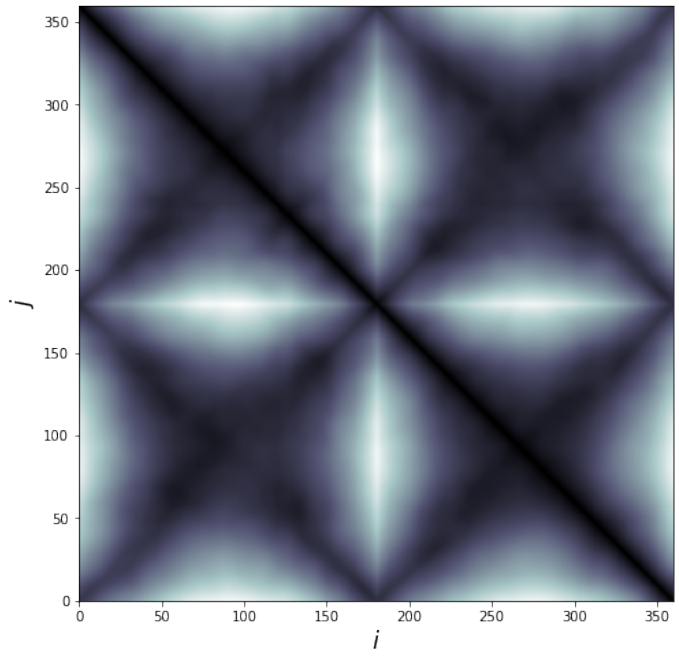


Figure A.9. Sorted dissimilarity matrix obtained from seriation

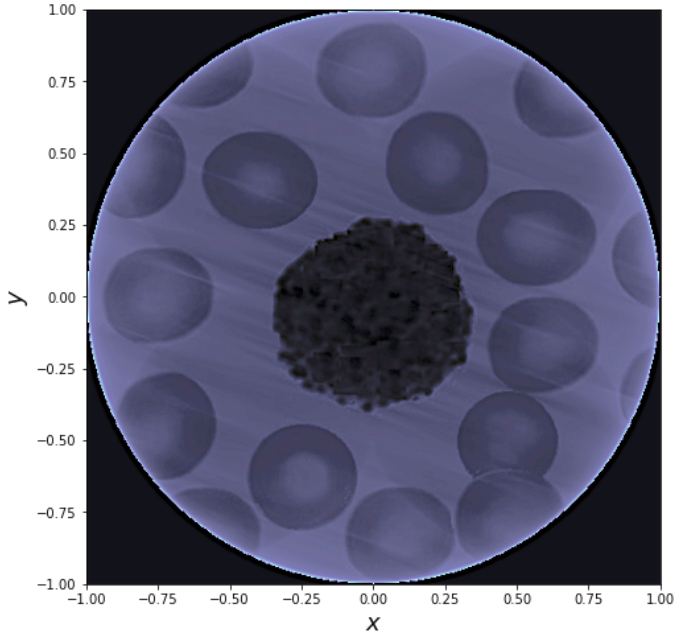


Figure A.10. Density obtained by inverting $\mathcal{R}\rho(t, \theta_{\sigma\circ\pi})$ where σ is the ordering of the projections found by seriation

An optimal algorithm for strict circular seriation*

Santiago Armstrong†, Cristóbal Guzmán†, and Carlos A. Sing Long†‡

Abstract. We study the problem of circular seriation, where we are given a matrix of pairwise dissimilarities between n objects, and the goal is to find a *circular order* of the objects in a manner that is consistent with their dissimilarity. This problem is a generalization of the classical *linear seriation* problem where the goal is to find a *linear order*, and for which optimal $O(n^2)$ algorithms are known. Our contributions can be summarized as follows. First, we introduce *circular Robinson matrices* as the natural class of dissimilarity matrices for the circular seriation problem. Second, for the case of *strict circular Robinson dissimilarity matrices* we provide an optimal $O(n^2)$ algorithm for the circular seriation problem. Finally, we propose a statistical model to analyze the well-posedness of the circular seriation problem for large n. In particular, we establish $O(log(n)/n)$ rates on the distance between any circular ordering found by solving the circular seriation problem to the underlying order of the model, in the Kendall-tau metric.

Key words. Circular seriation, circular Robinson dissimilarities, PQ-trees, circular Robinsonian matrices, circular-arc hypergraphs, circular embeddings of graphs, generative model

AMS subject classifications. 68R01, 05C85, 05C50, 05C25, 65C20

1. Introduction.

The seriation problem seeks to order a sequence of n objects from pairwise dissimilarity information. The goal is for the objects to be linearly ordered according to their dissimilarity [17, 25, 26]. Seriation has found applications in several areas such as archaeology [27], sociology and psychology [17], and gene sequencing and bioinformatics [25, 19]. However, in many applications the objects may be arranged along a closed continuum, resulting instead in a circular order. For instance, in *de novo* genome assembly of bacterial plasmids, the goal is to reorder DNA fragments sampled from a circular genome [25, 16]. In some problems in planar tomography, an object's density is to be reconstructed from projections taken at unknown angles between 0 and 2π. Reordering the projections according to their angle enables the reconstruction of the density [9]. In this case, the matrix representation of the pairwise dissimilarities is symmetric, with entries that increase monotonically starting from the diagonal along each row until they reach a maximum and then decrease monotonically, when the columns are wrapped around (see Figure 1). Matrices of this form are called circular Robinson [10, 12] in contrast to linear Robinson dissimilarities, where the entries are monotone non-decreasing along rows and columns when moving toward the diagonal [15].

*Submitted to the editors DATE.

Funding: This work was partially supported by INRIA through the INRIA Associate Teams project, CORFO through the Clover 2030 Engineering Strategy - 14ENI-26862, and ANID – Millennium Science Initiative Program – NCN17_059. C.A.SL. was partially supported by ANID – Millennium Science Initiative Program – NCN17_129.

†Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile and ANID – Millennium Science Initiative Program – Millennium Nucleus Center for the Discovery of Structures in Complex Data, Santiago, Chile (sarmstrong@uc.cl, crguzmanp@uc.cl, casinglo@uc.cl).

‡Institute for Biological and Medical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile and ANID – Millennium Science Initiative Program – Millennium Nucleus Center for Cardiovascular Magnetic Resonance, Santiago, Chile.

1.1. Our Contributions. In this work, we address the problem of circular seriation, both in its algorithmic and well-posedness for large n. Some of our results also apply to the linear case. Our first contribution is to provide a tractable and natural definition of circular Robinson matrices by leveraging unimodality (cf. Proposition 3.7). Various definitions of circular ordering have been proposed in the literature (see Subsection 1.2 below), but we believe this one captures intuitively the behavior of circular data.

Our second contribution is to provide the first optimal algorithm, with $\mathcal{O}(n^2)$ time and space complexity, for the seriation problem for strict Robinson dissimilarity matrices. Our algorithm is based on known techniques and data structures used in combinatorial seriation, but by virtue of the strict Robinson property our algorithm is substantially simpler. At a high level, the algorithm follows a divide-and-conquer approach, where we recursively detect nearest neighbors between chains of consecutive elements, and then resolve the orientations of such chains by comparing elements from their borders.

Our third contribution is a statistical model to analyze the large n regime. In this model, points are sampled from a closed curve, which without loss we assume is the unit circle, with a continuous and strict circular Robinson dissimilarity. Our main result here is a $\mathcal{O}(\log(n)/n)$ bound on the expected Kendall-tau distance of any strict circular Robinson ordering of the data. This result is based on an observation we make that in the continuous model, there is essentially[1] a unique ordering which makes the dissimilarity continuous and strictly circular Robinson. This analysis bridges the gap between solutions to the seriation problem, and their accuracy when data is naturally embedded in a continuous circular-like structure.

1.2. Related Work. Linear seriation is a classical problem in unsupervised learning and exploratory data analysis. As such, it has been thoroughly studied, and optimal algorithms for combinatorial seriation are known, as well as spectral methods. In contrast, circular seriation is substantially less understood. Next we summarize some results from the literature.

Linear Seriation. The first polynomial time algorithm for retrieving a linear order from permuted linear Robinson matrices was due to Mirkin and Rodin [19]. It is based on the connection between linear Robinson matrices and interval hypergraphs. It uses an algorithm introduced in [11] as a core subroutine, with an overall running time of $\mathcal{O}(n^4)$. Chepoi and Fichet [6] later introduced a simpler algorithm using a divide-and-conquer strategy. By recursively performing a partition refinement the algorithm computes an ordering in $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ space. Using similar techniques, Seston [28] improved the complexity to $\mathcal{O}(n^2 \log(n))$. Atkins [1] presented an entirely different strategy based on Laplacian eigenmaps (see [2]) with running time of $O(n(T(n) + n \log n))$, where $T(n)$ is the complexity of (approximately) computing the leading eigenvector of a $n \times n$ symmetric matrix. Prea and Fortin in [22] presented an optimal $\mathcal{O}(n^2)$ algorithm, using an algorithm from [4] to first compute a PQ-tree which is then updated by the algorithm. For the sparse case, Laurent and Seminarotti [15] present the Similarity-First Search algorithm in $\mathcal{O}\left(n^2 + nm \log n\right)$ operations where m is the number of nonzero entries of the dissimilarity matrix.

A natural question is how to perform seriation under noisy measurements of a dissimilarity.

[1] In an infinite set, permutations can be identified with bijections. However, given that for any finite sample we would only observe permutations of finitely many elements, we can substantially reduce the number of relevant permutations for this question. See Section 6 for further details.

Here, it is known that projecting a dissimilarity on the class of Robinsonian dissimilarities (in ℓ_∞-norm) is an NP-hard problem [7], and constant factor approximation algorithms exist [8].

Circular Seriation. In contrast to the linear case, where there is a common consensus for the definition of linear Robinson dissimilarities, in the circular case many definitions have been proposed that, in spite that they follow the same intuition, have mathematical formulations that are not equivalent. The first generalization of Robinson dissimilarities to the circular case was introduced in [12]. On top of being quite involved, this definition allows bimodality within each row (modulo n), which is incompatible with a circle embedding. The approach proposed for circular seriation is an instance of the quadratic assignment problem, which is NP-hard. A recent work following a similar line is [10]. The authors propose an optimization framework where they employ a spherical embedding together with a spectral method for circular ordering in order to recover circular arrangements of the embedded objects. This heuristic has no theoretical guarantees. A different approach in [9] aims to generalize Atkins' spectral approach by considering two eigenvectors. This methodology has asymptotic guarantees due to the connection between the Laplacian operator and the continuous Laplace-Beltrami operator over a manifold. Using the same idea, in [25] theoretical guarantees for a spectral method are introduced for the particular case in which the circular Robinson matrix is circulant, which is an idealized setting. In the same work, numerical experiments are presented to illustrate how the spectral method gains robustness by leveraging higher (> 2) Laplacian eigenvectors. In [5] dissimilarities whose ball, 2-ball and cluster hypergraph correspond to an arc hypergraphs are studied. Such dissimilarities can be considered as generalizations of Robinson dissimilarities to the circular case. We build upon this work by considering dissimilarities whose ball hypergraph corresponds to arcs and connect it to other definitions by showing that this definition is equivalent to requiring that the map $j \mapsto D(i, j + i \bmod n)$ is unimodal. Brucker and Osswald in [5] mainly focus in what they call circular dissimilarities which are a particular case of the previous definition.

1.3. Outline.

The paper is organized as follows. Section 2 introduces the notation and preliminaries. In Section 3 we formally introduce the seriation problem and the crucial concept of Robinson dissimilarities and matrices. In Section 4 we present some classical results on the *consecutive ones problem* and its connection to seriation, including the *PQ-tree data structure*, which is critical for our optimal algorithm. In Section 5 we present our optimal algorithm for strict circular seriation. Finally, in Section 6 we provide the generative model of sampling from a *continuous strictly Robinson curve*.

2. Preliminaries.

Throughout this work, arrays are indexed starting from 0 and are real unless it is explicitly stated otherwise. We let $[n] \triangleq \{0, 1, \ldots, n-1\}$ and denote as $\mathrm{Sym}(n)$ the group of permutations of $[n]$. A permutation is represented either by a vector π with entries in $[n]$ or by an $n \times n$ orthogonal and binary matrix Π. We denote as π_r the permutation that *reverses* the elements of $[n]$, i.e., $\pi_r(i) = n - 1 - i$, and π_s the *cyclic (right) shift* on $[n]$, i.e., $\pi_s(i) = i + 1 \bmod n$. We consider the *action by conjugation* of $\mathrm{Sym}(n)$ over the set of $n \times n$ matrices, which is defined by $(\Pi, A) \mapsto \Pi A \Pi^T$. If $S \subset \mathrm{Sym}(n)$ we denote $\langle S \rangle$ the subgroup generated by the elements of S. Finally, we denote the dihedral group of $2n$ different symmetries of a regular polygon with n sides as Dih_n.

For a countable set $\mathcal{X}$ and an enumeration $x : i \to x(i)$ we write x_i to denote $x(i)$ and let

$\#x$ be the integer such that $x(\#x) = x$. In this work, we consider finite sets of cardinality n. An enumeration becomes a bijection $\mathcal{X} \mapsto [n]$ with inverse $\# : \mathcal{X} \to [n]$.

The notion of and *ordered set* will play a crucial role. A *linear order* on $\mathcal{X}$ is a relation $\leq$ on $\mathcal{X}^2$ that is reflexive, antisymmetric, transitive and total. The pair $(\mathcal{X}, \leq)$ is a *linearly ordered set*. We say $x_0, \dots, x_{N-1}$ are *linearly ordered* if $x_i \leq x_{i+1}$ for $i \in [N]$. A *cyclic order* on $\mathcal{X}$ is a relation $\mathscr{C}$ on $\mathcal{X}^3$ that is cyclic, antisymmetric, transitive and total. The pair $(\mathcal{X}, \mathscr{C})$ is a *cyclically ordered set*. A cyclic order induces a linear order on $\mathcal{X}$. For $x_0 \in \mathcal{X}$ we define the linear order $\leq_{\mathscr{C},x_0}$ as $x \leq_{\mathscr{C},x_0} y$ if and only if $(x_0, x, y) \in \mathscr{C}$. Finally, we say $x_0, \dots, x_{N-1} \in \mathcal{X}$ are *cyclically ordered* if $x_i \leq_{\mathscr{C},x_0} x_{i+1}$ for $i \in [N]$. See [20] for more details.

3. The seriation problem and Robinson dissimilarities.

We introduce the seriation problem. Given a set $\mathcal{M}$ of $n \times n$ real matrices, let the *pre-$\mathcal{M}$ class* be the orbit of $\mathcal{M}$ under the action of $\mathrm{Sym}(n)$ by conjugation. The *abstract seriation problem* can be stated as [26]

Given A in pre-$\mathcal{M}$ find Π in $\mathrm{Sym}(n)$ such that $\Pi A \Pi^T$ is in $\mathcal{M}$.

The seriation problem is completely determined by the class $\mathcal{M}$. A *solution* to the seriation problem for A is any permutation Π satisfying the above. The set of all solutions is $\mathcal{S}_{\mathcal{M}}(A)$.

The seriation problem arises in applications when we consider a finite set $\mathcal{X}$ and we know the *dissimilarity* between its elements $x_1, \dots, x_n$. Ideally, a solution π to the seriation problem induces a linear order $\leq$ on $\mathcal{X}$ such that $x_{\pi(1)} \leq \dots \leq x_{\pi(n)}$. In this case, π and the order $\leq$ *orders* or *ranks* the elements of $\mathcal{X}$ in a way that is consistent with their dissimilarities.

We study two questions about this problem: *how to construct a suitable class $\mathcal{M}$ for which such a solution exists?* and, given this class, *is there an efficient algorithm to solve the seriation problem for any A?*. Our goal is to provide an answer when we allow for both linear and cyclic orders. For this reason, we explicitly distinguish between the *linear seriation problem* and the *circular seriation problem*; the *seriation problem* refers to either of them.

To answer the first question, in Subsection 3.1 we characterize dissimilarities that admit such linear or cyclic orders, and in Subsection 3.2 we discuss how these induce a suitable class of matrices for the seriation problem. We defer the answer to the second question to Section 4 and 5.

3.1. Robinson dissimilarities.

A *dissimilarity* or *premetric* $\mathbf{d} : \mathcal{X}^2 \to \mathbb{R}$ on $\mathcal{X}$ is a non-negative and symmetric function that is identically zero on the diagonal. *Robinson dissimilarities* are dissimilarities to which we can associate a linear or cyclic order on $\mathcal{X}$.

3.1.1. Linear Robinson dissimilarities.

Linear Robinson dissimilarities admit a family of linear orders on $\mathcal{X}$.

Definition 3.1 (The linear Robinson property). A dissimilarity $\mathbf{d}$ on $\mathcal{X}$ is *linear Robinson* if there exists a linear order $\leq_{\mathbf{d}}$ on $\mathcal{X}$ such that

$$\forall \text{ linearly ordered } x, y, z \in \mathcal{X} : \ \mathbf{d}(x, z) \geq \max\{\mathbf{d}(y, x), \mathbf{d}(y, z)\}. \tag{3.1}$$

It is *strictly linear Robinson* if all the inequalities are strict. We say $\leq_{\mathbf{d}}$ is *consistent* with $\mathbf{d}$ and that $\mathbf{d}$ is linear Robinson with respect to $\leq_{\mathbf{d}}$.

Linear Robinson dissimilarities preserve the *intervals* defined by *any* consistent order [19]. From Definition 3.1 it follows that for any $r > 0$ and $x \in \mathcal{X}$ the *(closed) balls* $B_r^{\mathbf{d}}(x) \triangleq \{y \in$

$\mathcal{X} : \mathbf{d}(x, y) \leq r\}$ are intervals in $(\mathcal{X}, \leq_{\mathbf{d}})$. In fact, this property uniquely characterizes linear Robinson dissimilarities. To prove this converse, the appropriate structure to analyze is the *hypergraph* $\mathcal{H}_{\mathbf{d}}$ with vertex set $\mathcal{X}$ and hyperedge set $\mathbf{B}_{\mathbf{d}} \triangleq \{B_r^{\mathbf{d}}(x) : x \in \mathcal{X}, r > 0\}$. This hypergraph is called an *interval hypergraph* if every hyperedge is an interval [14].

Proposition 3.2 ([19]). *Let* $\mathbf{d}$ *be a dissimilarity on* $\mathcal{X}$*. The following are equivalent:*

1. $\mathbf{d}$ *is linear Robinson.*
2. *The hypergraph* $\mathcal{H}_{\mathbf{d}}$ *is an interval hypergraph.*

Preserving intervals is not enough to yield uniqueness of the consistent order. This follows from the *natural symmetries* of Robinson dissimilarities. Let $\leq_{\mathbf{d}}$ be consistent with respect to $\mathbf{d}$. Its *reversal* $\leq'_{\mathbf{d}}$ is the linear order defined by $x \leq'_{\mathbf{d}} y$ if and only if $y \leq_{\mathbf{d}} x$. It is clear that $\mathbf{d}$ is linear Robinson with respect to $\leq_{\mathbf{d}}$ if and only if it is so with respect to $\leq'_{\mathbf{d}}$. Therefore, for Robinson dissimilarities the consistent order in $\mathcal{X}$ is not unique.

3.1.2. Circular Robinson dissimilarities. Circular Robinson dissimilarities arise naturally when we allow for cyclic orders.

Definition 3.3 (The circular Robinson property). A dissimilarity $\mathbf{d}$ on $\mathcal{X}$ is *circular Robinson* if there exists a cyclic order $\mathscr{C}_{\mathbf{d}}$ such that

$$\forall \text{ cyclically ordered } x, y, w, z \in \mathcal{X} : \quad \mathbf{d}(y, w) \geq \min\{\mathbf{d}(y, x), \mathbf{d}(y, z)\}.$$

We say it is *strict circular Robinson* if the inequality is strict. We say $\mathscr{C}_{\mathbf{d}}$ is *consistent* with $\mathbf{d}$ and that $\mathbf{d}$ is linear Robinson with respect to $\mathscr{C}_{\mathbf{d}}$.

Circular Robinson dissimilarities preserve the *arcs* of *any* compatible order, i.e., sets of the form $\{x \in \mathcal{X} : (m, x, M) \in \mathscr{C}_{\mathbf{d}}\}$ for $m, M \in \mathcal{X}$ called the *borders* of the arc. Arcs are the natural analogues of intervals for a cyclic order. Consequently, we say $\mathcal{H}_{\mathbf{d}}$ is an *arc hypergraph* if all its hyperedges are arcs. The analog of Proposition 3.2 for a cyclic order is the following.

Proposition 3.4. *([5, Proposition 5]) Let* $\mathbf{d}$ *be a dissimilarity. The following are equivalent:*

1. $\mathbf{d}$ *is circular Robinson.*
2. *The hypergraph* $\mathcal{H}_{\mathbf{d}}$ *is an arc hypergraph.*

Similarly to the linear case, preserving arcs is not sufficient to yield uniqueness of the consistent order. Let $\mathscr{C}_{\mathbf{d}}$ be consistent with respect to $\mathbf{d}$. In this case, its *reversal* $\mathscr{C}'_{\mathbf{d}}$ is the cyclic order such that $(x, y, z) \in \mathscr{C}'_{\mathbf{d}}$ if and only if $(z, y, x) \in \mathscr{C}_{\mathbf{d}}$. By definition, $\mathbf{d}$ is circular Robinson with respect to $\mathscr{C}_{\mathbf{d}}$ if and only if it is so with respect to $\mathscr{C}'_{\mathbf{d}}$.

3.2. Robinson matrices. Let $\mathbf{d}$ be a dissimilarity on $\mathcal{X}$. To any enumeration $x : [n] \to \mathcal{X}$ we can associate the $n \times n$ *dissimilarity matrix* D with entries $D(i, j) := \mathbf{d}(x_i, x_j)$. It is always non-negative, symmetric, and with zero-diagonal. However, some enumerations will endow D with additional properties. This leads us to *Robinson matrices*.

3.2.1. Linear Robinson matrices. If $\mathbf{d}$ is consistent with respect $\leq_{\mathbf{d}}$ there exists an enumeration of $\mathcal{X}$ such that for $i, j \in [n]$ we have $i \leq j$ if and only if $x_i \leq_{\mathbf{d}} x_j$. In this case, it follows that D induces a linear Robinson dissimilarity on $[n]$.

Definition 3.5 (Linear Robinson matrix). A dissimilarity matrix D is *linear Robinson* if

$$\forall \text{ linearly ordered } i, j, k \in [n] : \; D(i, k) \geq \max\{D(j, i), D(j, k)\}. \tag{3.2}$$

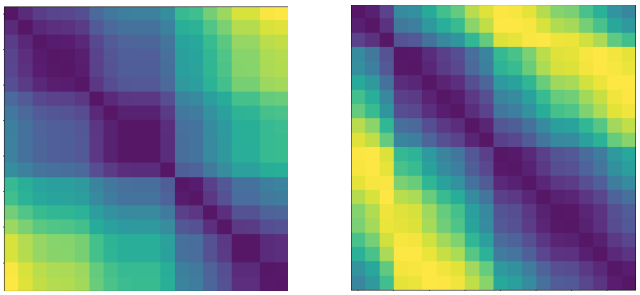


Figure 1: Example of a linear Robinson dissimilarity matrix (in the left) and a circular Robinson dissimilarity matrix (in the right)

It is *strictly linear Robinson* if all the inequalities are strict.

This implies D is consistent with the standard order on $[n]$ and Proposition 3.2 holds for D when $\mathcal{X} = [n]$. From the definition, we also deduce that

$$\forall \text{ linearly ordered } i, j, k \in [n] : \; D(i, j) \leq D(i, k) \text{ and } D(j, k) \leq D(i, k).$$

When the dimension is understood from context, the set of linear and strictly linear Robinson dissimilarity matrices will be denoted $\mathcal{L}_R$ and $\mathcal{L}_R^*$ respectively. Considering each one of these sets as $\mathcal{M}$ leads to the *linear seriation problem* and the *strict linear seriation* respectively.

Note linear Robinson matrices inherit the symmetries from the dissimilarity. In fact, it can be verified that D is linear Robinson if and only if $\Pi_r D \Pi_r^T$ is linear Robinson. Remark that $\mathrm{Dih}_1 \cong \langle \pi_r \rangle$ and consequently linear Robinson matrices are invariant under the action of Dih_1 by conjugation.

3.2.2. Circular Robinson matrices. For the circular case, we endow the set $[n]$ with the standard cyclic order $\mathscr{C}_n$

$$(i, j, k) \in \mathscr{C}_n \iff (i < j < k) \; \vee \; (j < k < i) \; \vee \; (k < i < j). \tag{3.3}$$

We still denote the standard linear order in $[n]$ as $\leq$.

If $\mathbf{d}$ is consistent with respect $\mathscr{C}_{\mathbf{d}}$ there exists an enumeration of $\mathcal{X}$ such that $(i, j, k) \in \mathscr{C}_n$ if and only if $(x_i, x_j, x_k) \in \mathscr{C}_{\mathbf{d}}$. Similarly to the linear case, this implies D induces a circular Robinson dissimilarity on $[n]$.

Definition 3.6 (Circular Robinson matrix). A dissimilarity matrix D is *circular Robinson* if

$$\forall \text{ cyclically ordered } i, j, k, \ell \in [n] : \quad D(j, k) \geq \min\{D(j, i), D(j, \ell)\}.$$

Therefore, D is consistent w.r.t. $\mathscr{C}_n$ and Proposition 3.4 holds for D when $\mathcal{X} = [n]$. When the dimension is understood from context, the set of circular and strictly circular Robinson matrices will be denoted $\mathcal{C}_R$ and $\mathcal{C}_R^*$, respectively. Considering each one of these sets leads to the *circular seriation problem* and the *strict circular seriation* respectively. Comparing Definition 3.5 and Definition 3.6 it is apparent that every linear Robinson matrix is also circular

Robinson. In this sense, the notion of circular Robinson extends that of linear Robinson. The difference between linear and circular Robinson matrices is illustrated in Figure 1.

We provide an alternative definition for circular Robinson matrices that will be useful in what follows. Let $f : [n] \to \mathbb{R}$. A *mode* is any $m \in [n]$ such that

$$\forall\, i, j \in [n] :\ i \le j \le m \text{ or } m \le j \le i \ \Rightarrow\ f_i \le f_j \le f_m.$$

We say f is *unimodal* if it has a mode. We say f is *strictly unimodal* if it has at most two distinct, consecutive modes $m_1 \le m_2$ with $f_{m_1} = f_{m_2}$ and

$$\forall\, i, j \in [n] :\ i < j < m_1 \ \Rightarrow\ f_i < f_j < f_{m_1} \text{ and } i > j > m_2 \ \Rightarrow f_i < f_j < f_{m_2}.$$

From the definition it is clear that every subsequence of a strictly unimodal sequence is also strictly unimodal. The proof of the following is deferred to Appendix A.1.

Proposition 3.7. *Let D be a dissimilarity matrix. The following are equivalent:*

1. *D is circular Robinson (resp. strict circular Robinson).*
2. *For any $i \in [n]$ the function $j \to D(i, i+j \bmod n)$ is unimodal (resp. strict unimodal).*

This property is naturally invariant under cyclic permutations.

Proposition 3.8. *A dissimilarity matrix D is circular Robinson if and only if $\Pi_r D \Pi_r^T$ and $\Pi_s D \Pi_s^T$ are circular Robinson matrices.*

Proof. First, notice that the (i, j) entry of $\Pi_s D \Pi_s^T$ and $\Pi_r D \Pi_r^T$ are $D(i + 1 \bmod n, j + 1 \bmod n)$ and $D(n-1-i, n-1-j)$, respectively. Noticing that $\{D(i \bmod n, i+j \bmod n)\}_{j=0}^{n-1}$ is unimodal for all i, we have that $\{D(i + 1 \bmod n, i + j + 1 \bmod n)\}_{j=0}^{n-1}$ and $\{D(n - 1 - i \bmod n, n - 1 - i + j \bmod n)\}_{j=0}^{n-1}$ are unimodal. Therefore $\Pi_r D \Pi_r^T$ and $\Pi_s D \Pi_s^T$ are also circular Robinson. ∎

Since $\mathrm{Dih}_n \cong \langle \pi_r, \pi_s \rangle$ it follows that circular Robinson matrices are invariant under the action of Dih_n by conjugation. This invariance is particular to the definition and should not be taken for granted. Other definitions proposed in the literature, e.g., [26], do not enjoy this property. We believe that cyclic invariance makes the definition arguably more natural.

3.3. Robinson orderings. The seriation problem does not assume we observe a linear or circular Robinson dissimilarity matrix, but instead its image under conjugation by an unknown permutation matrix. In other words, we observe matrices in pre-$\mathcal{L}_R$ and pre-$\mathcal{C}_R$. We call such matrices *Robinsonian matrices*.

Given a Robinsonian matrix and an algorithm for the corresponding seriation problem, the set of solutions may not be a singleton. In fact, the symmetries of linear and circular Robinson dissimilarity matrices *ensures* they will never be a singleton. We call *Robinson orderings* all the orderings represented by the elements in the set of solutions.

Although there will never be a unique Robinson ordering, we can at least distinguish which ones are due to the natural symmetries of the problem. Therefore, for the linear seriation problem we call solutions in the same orbit under the action Dih_1 the *trivial solutions* whereas those in different orbits *non-trivial solutions*. The same criteria applies for the circular seriation problem when the action of Dih_n is considered instead.

4. The consecutive ones problem and PQ-trees. Robinson matrices turn out to be natural to formulate the seriation problem. We now review the connection between this problem and the *consecutive ones problem*. This connection yields polynomial time algorithms for solving the seriation problem, and allows us to introduce PQ-trees, which will be extensively used in Section 5.

4.1. The consecutive ones problem. The linear seriation problem is deeply connected to a combinatorial problem known as the *consecutive ones (C1) problem*. To introduce this problem, consider a $n \times n$ binary matrix M. The C1 problem is to find a permutation Π such that the entries of $M\Pi$ equal to one appear consecutively along rows. We say M has the *consecutive ones property (C1) property* if the C1 problem has a solution for M. An example of such matrix can be found in Figure 2a. The first linear time algorithm for the C1 problem was introduced by Booth and Lueker in [3]. If f is the number of ones in M then their result states the C1 problem can be decided in $\mathcal{O}(n+f)$ time.

An extension to this problem is the *circular ones (Cr1) problem*. The Cr1 problem is to find a permutation Π such that the entries of $M\Pi$ equal to one appear consecutively *modulo* n along rows. We say M has the *circular ones (Cr1) property* if the Cr1 problem has a solution for M [29]. This problem can also be solved efficiently as it can be reduced to the C1 problem. Let $\overline{M}$ be the matrix such that every row with a 1 on its first entry is complemented. Then M satisfies the C1 property if an only if $\overline{M}$ satisfies the Cr1 property [29, Theorem 1]. Therefore, by forming the complement, the $Cr1$ problem can be decided in polynomial time.

Both problems are connected to the seriation problem through the hypegraph $\mathcal{H}_D$. In fact, interval and arc hypergraphs are precisely those for which their incidence matrices respectively satisfy the C1 and Cr1 properties [14]. This suggests how to efficiently solve the seriation problem for Robinson matrices.

Theorem 4.1. *([6, 19]) The linear and circular seriation problem can be reduced in polynomial time and space to deciding respectively the C1 and Cr1 problem. Robinson matrices can be recognized in $\mathcal{O}\left(n^3\right)$ time and with $\mathcal{O}\left(n^3\right)$ space.*

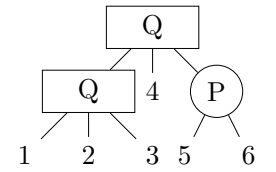
The bounds above follow from the worst case in which $\mathcal{H}_D$ has $\mathcal{O}\left(n^2\right)$ different hyperedges. In this case, for each of the n possible centers and each row $i \in [n]$ the matrix can take $\mathcal{O}(n)$ possible values. In this case, the incidence matrix has $\mathcal{O}(n^3)$ entries.

4.2. PQ-Trees. The algorithmic structure underlying the algorithm to solve the C1 problem is the PQ-tree. A PQ-tree $\mathcal{T}$ on a set $\mathcal{X}$ is a rooted tree with two types of internal nodes denoted by P, represented as circles, and Q, represented as rectangles, and where the leaves represent the elements in $\mathcal{X}$. The type of node represents admissible permutations on $\mathcal{X}$: children of a P-node can be permuted arbitrarily, whereas children of a Q-node can only be reversed. Figure 2b shows an example of a PQ-tree.

PQ-trees are related to the C1 problem as follows. Let Y_i be the indices of the columns of M such that its i-th entry equal to one. Then $\mathbf{Y} = \{Y_i\}_{i \geq 0}$ is a collection of subsets of $[n]$. The C1 problem can be solved if we can permute the elements of $[n]$ so that every Y_i becomes an interval. The algorithm starts with a single set $\mathbf{Y}_1 = \{Y_{i_1}\}$ and determines the set of *admissible permutations* such that Y_{i_1} becomes an interval. These can be represented by a PQ-tree $\mathcal{T}_1$ (see [3] and [4]). The algorithm proceeds by adding a Y_{i_2} to form $\mathbf{Y}_2 = \{Y_{i_1}, Y_{i_2}\}$

1	2	3	4	5	6
1	1	0	0	0	0
0	1	1	0	0	0
1	1	1	1	0	0
0	0	0	1	1	1
0	0	0	0	1	1

(a) $M \in \{0,1\}^{5\times 6}$ with the consecutive ones property.



(b) PQ-tree $\mathcal{T}$ representing the solutions of the (C1P) for M.

$\{(1,2,3,4,5,6), \quad (5,6,4,1,2,3),$
$(3,2,1,4,5,6), \quad (5,6,4,3,2,1),$
$(1,2,3,4,6,5), \quad (6,5,4,1,2,3),$
$(3,2,1,4,6,5), \quad (6,5,4,3,2,1)\}$

(c) The set $S(\mathcal{T})$ of permutations represented by the tree.

Figure 2: A PQ-tree of all solutions to the (C1P) for a $\{0,1\}$-matrix.

and update the PQ-tree accordingly. The main contribution of [4] is an algorithm for updating $\mathcal{T}_k$ in a way that given any subset $Y_k \subseteq [n]$, the set of permutations represented by the updated tree $\mathcal{T}_{k+1}$ is precisely the set of admissible permutations of $\mathbf{Y}_{k+1} \cup \{Y_k\}$. This is done in time linear in the size of Y_k. The algorithm finishes when $\mathbf{Y}$ is attained.

As an example, by considering all rows of the binary matrix in Figure 2a, the resulting PQ-tree at the final step would be the one in Figure 2b, and the solution set would be the one in Figure 2c.

5. Optimal Algorithm for Strict Circular Seriation. In this section, we present an optimal algorithm for circular seriation in the strict Robinson case (this algorithm would also work for the strict linear case, but we will omit this). Our algorithm runs in $\mathcal{O}(n^2)$ time and space, which is obviously optimal, since it is the time required to read the input and the space required to provide a strict Robinson dissimilarity.[2] The core algorithm relies in two main ideas: merging nearest neighbors, and discarding forbidden arc reversals. We recursively merge nearest neighbors, using the fact that nearest neighbors are guaranteed to be consecutive elements in a strict Robinsonian ordering. Exploiting this fact we can obtain chains of consecutive elements which are stored in Q-nodes of PQ-trees. The most complicated part of the algorithm consists in (efficiently) determining if each Q-node can be *uniquely oriented*. In that case such Q-node can be deleted and its children merged to the parent Q-node.

The process of building chains of consecutive elements and deciding their orientation can be done in several ways. The advantage of our algorithm is that the total number comparisons to decide each orientation is bounded by $\mathcal{O}(n)$, which leads to a running time $\mathcal{O}(n^2)$.

5.1. Preliminaries part I: nearest neighbours graph in strict Robinson dissimilarities. Given $D \in \text{pre-}\mathcal{C}_R$, a collection of subsets $\mathcal{P} = \{\mathcal{I}_i\}_{i\geq 0}$ of $\mathcal{X}$ is said to be an arc partition if $\mathcal{P}$ is a partition of $\mathcal{X}$ (i.e. $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ when $i \neq j$ and $\bigcup_{i\geq 0} \mathcal{I}_i = \mathcal{X}$) and every set $\mathcal{I}_i$ is an arc of consecutive elements in any Robinson ordering. A crucial part of the algorithm consists in building arc partitions by merging nearest neighbours. The set of nearest neighbours of $x \in \mathcal{X}$ is defined as $\text{NN}(x) \triangleq \arg\min_{y\in\mathcal{X}\setminus\{x\}} \mathbf{d}(x,y)$. The nearest-neighbours graph is an undirected

[2] Given the strict Robinson property, it is clear that the underlying matrix is dense, and therefore $\mathcal{O}(n^2)$ memory is required to even provide the input.

graph $G_{\mathrm{NN}}(\mathcal{X}, \mathbf{d}) = (\mathcal{X}, \mathcal{E})$ such that $\{x, y\} \in \mathcal{E}$ iff $x \in \mathrm{NN}(y)$ or $y \in \mathrm{NN}(x)$. An essential condition of strict Robinson dissimilarities is what we called the *nearest-neighbour condition*, which implies that the connected components of the nearest-neighbours graph correspond to arcs of consecutive elements, and since connected components form a partition, such collection corresponds to an arc partition.

Definition 5.1. (Nearest-neighbour condition) A dissimilarity matrix $D \in \mathbb{R}^{n\times n}$ is said to have the *nearest-neighbour condition* if it holds that $\mathrm{NN}(i) \subseteq V_i^C \triangleq \{i-1 \bmod n, i+1 \bmod n\}$.[3]

It is immediate to verify that *strict* circular Robinson dissimilarities satisfy the nearest-neighbour condition, which is not necessarily true in the non-strict case.

We also recall from graph theory that given a graph $G = (\mathcal{X}, \mathcal{E})$ and a node $x \in \mathcal{X}$, the set of adjacent nodes to x is denoted as $\mathcal{N}_G(x) \triangleq \{y \in \mathcal{X} : \{x, y\} \in \mathcal{E}\}$. The function $x \mapsto \mathcal{N}_G(x)$ is called the neighbourhood. The ring graph $R_n = ([n], \mathcal{E})$ is the graph with edge set $\mathcal{E} = \{\{i, (i+1) \mod n\} : i \in [n]\}$. If a graph G is a subgraph of R_n, then it is clear that its connected components correspond to arcs of $([n], \mathscr{C}_n)$. A direct consequence of the nearest-neighbour condition is that the nearest-neighbours graph of *strict* circular Robinson dissimilarity matrices correspond to subgraphs of R_n.

Our algorithm relies crucially on the fact that strict dissimilarities must respect nearest neighbors in any Robinson ordering. This is not necessarily true in the non-strict case.

Lemma 5.2. *Let $D \in$ pre-$\mathcal{C}_R^*$ and $i \in [n]$. Suppose that $j \in \mathrm{NN}(i)$, then in any Robinson ordering σ, the elements i and j are consecutive.*

Proof. Let σ be any Robinson ordering. Let $j \in \mathrm{NN}(i)$ and let $r \triangleq D(i, j)$. This implies that for any $k \in B(i, r) \setminus \{i\}$, $D(i, k) = D(i, j)$. Suppose by contradiction that there exist k_1, i, k_2 consecutive in σ, with $j \neq k_1, k_2$. Since in any Robinson ordering balls are arcs, this implies that either $k_1 \in B(i, r)$ or $k_2 \in B(i, r)$. Any of the two cases is a contradiction with the nearest-neighbour condition, proving the result. ∎

Since nearest-neighbours must be consecutive, we get that the connected components of the nearest-neighbours graph of a *strict* circular Robinson dissimilarity correspond to arcs of any Robinson ordering. Hence, the set of connected components constitute an arc partition. The fact that this graph is a subgraph of the ring graph makes computationally efficient finding the order intrinsic to each component, and the task is divided in two steps:

1. Find all degree 1 nodes. These correspond to the borders of the components.
2. Perform a Depth-First Search (Algorithm DFS) starting at each non visited degree 1 node. The order in which the nodes are visited will follow the Robinson ordering (or backwards).

If there are no degree one nodes, then $G_{\mathrm{NN}} \cong R_n$ and therefore we can start at any node. For an algorithmic implementation, tuples can be used to represent the local fragments of Robinson orderings (Q-nodes). A tuple is an ordered set $\alpha = (a_0, a_1, a_2, \ldots, a_{k-1})$. We write $\alpha(i)$ to denote a_i, the i-th element of α. Each connected component will be stored in a tuple α, where $\alpha(j)$ is the j-th element visited by performing a DFS. The procedure is summarized in Algorithm AP (`Arc Partition`), whose correctness is stated in the following Proposition

[3]Given some enumeration # and $i \in [n]$, when we write $\mathrm{NN}(i)$, we refer to the set $\#(\mathrm{NN}(x_i))$

(the proof of the next result is omitted for brevity).

Proposition 5.3. *Given any dissimilarity matrix $D \in$ pre-$\mathcal{C}_R^*$, by performing Algorithm AP with input $([n], D)$ the resulting tuples follow an arc ordering for every Robinson ordering.*

5.2. Preliminaries part II: orienting arcs. The previous section tell us that nearest-neighbours must be consecutive in the strict Robinson case. By exploiting this idea we can obtain ordered sequences of elements stored in Q-nodes of a PQ-tree. Notice however that Q-nodes are allowed to be reversed, which at this point of the algorithm is not guaranteed to lead to Robinson orderings. If this is not the case, the inconsistent ordering must be discarded, which corresponds to removing the Q-node and merging the its children directly to the parent Q-node. We call this process orienting. In this section we provide computationally efficient conditions to determine when a Q-node must be oriented. Each Q-node α in a tree $\mathcal{T}$ can be associated with an arc $\mathcal{I}_\alpha$ in $\mathcal{X}$: the arc of all leaves in $\mathcal{X}$ which are descendants of α. Reversing α corresponds to reversing $\mathcal{I}_\alpha$. The first relevant concept to determine when it is possible to reverse each arc is the strictly overlapping condition, which has been studied for instance in [23] and in [14]. An example of the property can be seen in Figure 3 (a).

Definition 5.4. Two arcs $\mathcal{I}$ and $\mathcal{J}$ are said to strictly overlap, denoted by $\mathcal{I} \between^* \mathcal{J}$, if

$$1.\ \mathcal{I} \not\subset \mathcal{J}; \qquad 2.\ \mathcal{J} \not\subset \mathcal{I}; \qquad 3.\ \mathcal{I}^c \not\subset \mathcal{J}; \quad \text{and} \quad 4.\ \mathcal{J} \not\subset \mathcal{I}^c.$$

Observation 5.5. The relation $\between^*$ is symmetric and equivalent to

$$1.\ \mathcal{I} \cap \mathcal{J}^c \neq \emptyset; \qquad 2.\ \mathcal{J} \cap \mathcal{I}^c \neq \emptyset; \qquad 3.\ \mathcal{I}^c \cap \mathcal{J}^c \neq \emptyset; \quad \text{and} \quad 4.\ \mathcal{I} \cap \mathcal{J} \neq \emptyset.$$

Lemma 5.6. *Let $\mathcal{I}$ and $\mathcal{J}$ be two arcs. Let a, b and a', b' be the borders of $\mathcal{I}$ and $\mathcal{I}^c$, respectively, where a (b) and a' (b') are consecutive in the cyclic order. Then $\mathcal{I} \between^* \mathcal{J}$ if and only if one of the following conditions holds*

$$(i)\ \{a, a'\} \subset \mathcal{J} \text{ and } \{b, b'\} \subset \mathcal{J}^c; \qquad \text{or} \qquad (ii)\ \{b, b'\} \subset \mathcal{J} \text{ and } \{a, a'\} \subset \mathcal{J}^c.$$

Proof. We first prove ($\Leftarrow$). Suppose (i) holds (the other case follows analogously). Then since $a \in \mathcal{I}$ and $a' \in \mathcal{I}^c$ we get conditions 2 and 4 of Observation 5.5. Now, since $b \in \mathcal{I}$ and $b' \in \mathcal{I}^c$ we get conditions 1 and 3 of Observation 5.5.

Next we prove ($\Rightarrow$). First we notice that there are at least two elements in $\mathcal{I}$ and two elements in $\mathcal{I}^c$ (otherwise containing a single element of these arcs would imply containing the whole set, contradicting one of the conditions in Definition 5.4). Hence, the elements a, b, a' and b' exist and are distinct. Suppose $a \in \mathcal{J}$ (the case $a \in \mathcal{J}^c$ is analogous), and let $z \in \mathcal{J} \setminus \mathcal{I}$ (exists by hypothesis). Since $\mathcal{J}$ is an arc, it must contain one of the two paths connecting a and z. Since it does not contain the whole $\mathcal{I}$ it must be the path that covers a', therefore $\{a, a'\} \subset \mathcal{J}$ and $b \notin \mathcal{J}$. On the other hand, since it does not contain the whole $\mathcal{I}^c$, $b' \in \mathcal{J}$. It follows that $\{b, b'\} \subset \mathcal{J}^c$. ∎

Given an arc $\mathcal{I} = \{a_0, \ldots, a_{k-1}\}$ (where elements are indexed following the cyclic order), we define the *permutation that reverses* $\mathcal{I}$ as the permutation σ s.t. $\sigma(a_j) = a_{k-j-1}$ for $j \in \{0, \ldots, k-1\}$, and $\sigma(x) = x$ if $x \notin \mathcal{I}$.

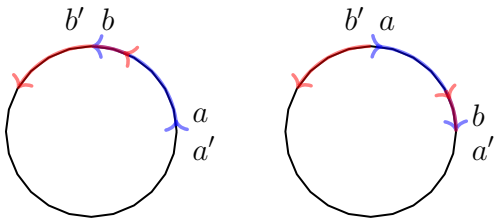


Figure 3: In the left, two strictly overlapping arcs $\mathcal{I}$ (in blue) and $\mathcal{J}$ (in red). In the right, $\sigma(\mathcal{I})$ (in blue) and $\sigma(\mathcal{J})$ (in red), where σ is the permutation that reverses the elements of $\mathcal{I}$.

Lemma 5.7. *Let $\mathcal{I}$ and $\mathcal{J}$ be two arcs and let σ be the permutation that reverses the elements of $\mathcal{I}$. Then $\mathcal{I} \between^* \mathcal{J}$ iff the permutation of $\mathcal{J}$ by σ is not an arc.*

Proof. We first prove ($\Leftarrow$). By contraposition, assume any of the conditions in Definition 5.4 do not hold, then it is easy to see that $\sigma(\mathcal{J}) = \mathcal{J}$, which is an arc. Now we prove ($\Rightarrow$). If $\mathcal{I} \between^* \mathcal{J}$, then at least one of the conditions of Lemma 5.6 hold. Since $\sigma(a) = b$, $\sigma(b) = a$, $\sigma(a') = a'$ and $\sigma(b') = b'$ then $\sigma(\mathcal{J})$ is not connected and thus it is not an arc. ∎

As an example, consider the two strictly overlapping arcs $\mathcal{I}$ and $\mathcal{J}$ in Figure 3. By reversing the blue arrow (the arc $\mathcal{I}$) the red arrow (the arc $\mathcal{J}$) gets ripped apart into two disconnected pieces. Recall from Proposition 3.4 that a dissimilarity matrix is circular Robinson iff each ball $\mathcal{J}$ is an arc. Therefore, any arc $\mathcal{I}$ cannot be arbitrarily reversed to produced a new Robinson ordering iff there is some ball that strictly overlaps with $\mathcal{I}$. In terms of PQ-trees, a necessary and sufficient condition for a Q-node α to be orientable is the existence of some $z \in \mathcal{X}$ and $r > 0$ such that the ball $B_r(z)$ strictly overlaps with $\mathcal{I}_\alpha$. In such case, one of the two orientations of the node is not compatible with a Robinson ordering since in one of these orientations the ball gets disconnected. By Lemma 5.6, to determine the orientation of the arc $\mathcal{I}$ one could equivalently check whether there exists $z \in \mathcal{X}$ and $r > 0$ such that

$$\Big[\{a, a'\} \subset B_r(z) \ \wedge \ \{b, b'\} \subset B_r(z)^c\Big] \quad \vee \quad \Big[\{b, b'\} \subset B_r(z) \ \wedge \ \{a, a'\} \subset B_r(z)^c\Big],$$

If none of this conditions hold, then we say the arc is *not orientable* which means that the Q-node in the tree must be preserved. Notice that this requires knowing that a, b (a', b') are the borders of $\mathcal{I}$ (resp. $\mathcal{I}^c$) in advance. Algorithm BCO is an efficient way for orienting the arc $\mathcal{I}$ with respect to the dissimilarity $\mathbf{d}$ when we have *border candidates* but the actual borders within the candidates are unknown.

Definition 5.8 (Border candidates *of an arc*). An 4-tuple of sets $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$ are said to be border candidates of the arc $\mathcal{I}$, if the following properties hold:

1. $\mathcal{A}, \mathcal{B} \subset \mathcal{I}$ and $\mathcal{A}', \mathcal{B}' \subset \mathcal{I}^c$.
2. The sets are pairwise disjoint.
3. If a, b are the borders of $\mathcal{I}$ and a', b' are the borders of $\mathcal{I}^c$, then: $a \in \mathcal{A}, b \in \mathcal{B}, a' \in \mathcal{A}'$ and $b' \in \mathcal{B}'$.
4. Either $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$ or $(\mathcal{A}', \mathcal{B}, \mathcal{A}, \mathcal{B}')$ is cyclically ordered.[4]

[4]Formally, the ordered collection is a consistent cyclic quasi order, see Definition 5.12

The next result provides correctness for Algorithm BCO (see its proof in Appendix A.2).

Lemma 5.9. *Let $\mathcal{I} \subset \mathcal{X}$ be an arc in any Robinson ordering. Suppose a, b are the borders of $\mathcal{I}$ and a', b' are the respective borders of $\mathcal{I}^c$. Additionally suppose that $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$ are border candidates for $\mathcal{I}$. Then, for every $z \in \mathcal{X}$, the following statements are equivalent:*

1. *Both $\{a, a'\} \subset B_r(z)$ and $\{b, b'\} \subset B_r(z)^c$ hold.*
2. *There exist $(x', x, y, y') \in \mathcal{A}' \times \mathcal{A} \times \mathcal{B} \times \mathcal{B}'$ s.t. $\max\{f_z(x), f_z(x')\} < \min\{f_z(y), f_z(y')\}$.*
3. *It holds that $\max\left\{\min\{f_z(\mathcal{A})\}, \min\{f_z(\mathcal{A}')\}\right\} < \min\left\{\max\{f_z(\mathcal{B})\}, \max\{f_z(\mathcal{B}')\}\right\}$.*

Above $f_z(\cdot) \triangleq \mathbf{d}(z, \cdot)$ and given any $U \subset \mathcal{X}$, $\min\{f_z(U)\} \triangleq \min_{y \in U} f_z(y)$.

Corollary 5.10. *Let $\mathcal{I} \subset \mathcal{X}$ be an arc in any Robinson ordering and suppose the sets $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$ are border candidates for the arc $\mathcal{I}$. Then, Algorithm BCO correctly determines if $\mathcal{I}$ must be fixed, reversed or if it is not orientable.*

Algorithm BCO `Border Candidates Orientation`

Input: A sequence of sets $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$
Let $f_z(x) \triangleq \mathbf{d}(z, x)$ for every $z \in \mathcal{X}$
Let $O_i : \mathcal{X} \to \{True, False\}$ for $i = 1, 2, 3, 4$ be defined by
$O_1(z) \triangleq \max\{\min\{f_z(\mathcal{A})\}, \min\{f_z(\mathcal{A}')\} < \min\{\max\{f_z(\mathcal{B})\}, \max\{f_z(\mathcal{B}')\}$
$O_2(z) \triangleq \max\{\min\{f_z(\mathcal{B})\}, \min\{f_z(\mathcal{B}')\} < \min\{\max\{f_z(\mathcal{A})\}, \max\{f_z(\mathcal{A}')\}$
$O_3(z) \triangleq \max\{\min\{f_z(\mathcal{A})\}, \min\{f_z(\mathcal{B}')\} < \min\{\max\{f_z(\mathcal{A}')\}, \max\{f_z(\mathcal{B})\}$
$O_4(z) \triangleq \max\{\min\{f_z(\mathcal{B})\}, \min\{f_z(\mathcal{A}')\} < \min\{\max\{f_z(\mathcal{B}')\}, \max\{f_z(\mathcal{A})\}$
for $z \in \mathcal{X}$ **do**
  if $O_1(z) \vee O_2(z)$ **then**
    return `'correct'`
  else if $O_3(z) \vee O_4(z)$ **then**
    return `'reverse'`
  end if
end for
return `'not orientable'`
Output: A string determining the orientation of the input

Observation 5.11. The time complexity of Algorithm BCO with input $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$ is $\mathcal{O}(|\mathcal{X}| \cdot \max\{|\mathcal{A}'|, |\mathcal{A}|, |\mathcal{B}|, |\mathcal{B}'|\})$, thus it is an efficient way of orienting a Q-node α whenever the sets of border candidates for $\mathcal{I}_\alpha$ is not too big.

5.3. The Recursive Seriation Algorithm. The general idea of the algorithm is first: to merge nearest neighbours into Q-nodes, and second: to orient these nodes afterwards whenever is possible. For this recursive algorithm to work, we need an appropriate data structure that maintains arcs certified by nearest neighbour conditions. We find convenient for this purpose to use trees, which keep track of the nearest neighbours obtained at different steps of the recursion. The starting family of trees are singletons indexed by the elements of $\mathcal{X}$

Consider a family $\mathbf{T}$ of Q-trees (which are PQ-trees composed solely by Q-nodes). For each $\mathcal{T} \in \mathbf{T}$ we write as $\partial\mathcal{T}$ the set of leaves of $\mathcal{T}$. Along this section we assume that $\{\partial\mathcal{T}\}_{\mathcal{T} \in \mathbf{T}}$ is an arc partition. Moreover, for every Robinson ordering of $\mathcal{X}$ we assume that

there is a configuration of each $\mathcal{T} \in \mathbf{T}$ in a way that the leaves of $\mathcal{T}$ follow an arc ordering. We endow each $\mathcal{T} \in \mathbf{T}$ with a set $\mathcal{B}(\mathcal{T})$ of *border candidates*[5], which are all leaves of $\mathcal{T}$ that appear in the extreme left or right under some configuration of the tree. Whenever $|\partial\mathcal{T}| \geq 2$, the set of border candidates $\mathcal{B}(\mathcal{T})$ can be split in two: left and right. The set of left border candidates, denoted as $\mathcal{B}^L(\mathcal{T})$, are all elements in $\mathcal{B}(\mathcal{T})$ that appear in the extreme left under some configuration of the tree, subject to fixing the Q-node in the root. Similarly, $\mathcal{B}^R(\mathcal{T})$ denotes the set of all right border candidates, which are are all elements in $\mathcal{B}(\mathcal{T})$ that appear in the extreme right. For instance, in the tree $\mathcal{T}$ appearing in Figure 5a, we have $\mathcal{B}^L(\mathcal{T}) = \{a_3, b_3, b_2, b_1\}$ and $\mathcal{B}^R(\mathcal{T}) = \{b_0\}$. Finally depth$(\mathcal{T})$ denotes the tree-depth of $\mathcal{T}$.

Next we present a high level pseudocode and describe its main steps. For simplicity, we describe the orienting steps at the end.

Algorithm 5.2 `Recursive Seriation`

Input: A family $\mathbf{T}$ of Q-trees
step 1: Compute $\mathbf{d}^{\min}$ and $\mathbf{d}^{\arg\min}$ over $\mathbf{T}$
step 2: Perform Algorithm EO of each $\mathcal{T} \in \mathbf{T}$
step 3: Compute a new family trees $\mathbf{T}'$ using Algorithm AP with input $(\mathbf{d}^{\min}, \mathbf{T})$
if $|\mathbf{T}'| = 1$ **then**
  step 4: Perform Algorithm FO in $\mathcal{T} \in \mathbf{T}'$
  return $\mathcal{T} \in \mathcal{T}'$
else
  step 4: Perform Algorithm CIO in each $\mathcal{T} \in \mathbf{T}'$
  Recurse over $\mathbf{T}'$
end if
Output: The unique Q-tree $\mathcal{T} \in \mathbf{T}'$ containing all Robinson orderings

5.3.1. Initialization. Initially we have an abstract set $\mathcal{X}$ (or indices) and some dissimilarity $\mathbf{d}$ among its elements. Given this input, we initialize `Recursive Seriation` with a collection of single element trees $\mathbf{T} = \mathcal{X}$. In such case, $\mathcal{B}(x) = \partial x \triangleq \{x\}$ for every $x \in \mathcal{X}$.

5.3.2. Computing the minimum pairwise dissimilarity among trees. The following step of each iteration consists in computing a dissimilarity matrix among the input trees. Given two trees $\mathcal{T}_1, \mathcal{T}_2$ define the dissimilarity $\mathbf{d}^{\min}(\mathcal{T}_1, \mathcal{T}_2) \triangleq \min\{\mathbf{d}(x, y) : x \in \mathcal{B}(\mathcal{T}_1), y \in \mathcal{B}(\mathcal{T}_2)\}$. Also let $\mathbf{d}^{\arg\min}(\mathcal{T}_1, \mathcal{T}_2)$ be the collection of all minimizers in $\mathcal{B}(\mathcal{T}_1) \times \mathcal{B}(\mathcal{T}_2)$. Computing dissimilarities among trees allows us to solve the problem recursively. In each step, the objects we reorder will be Q-trees and by doing so we obtain quasi-orders among our original set of objects. The following Lemmata justifies this procedure.

Definition 5.12 (Quasi-order[6]). Let $(\mathcal{X}, \mathscr{C})$ be a cyclically ordered set. An ordered partition $\{A_0, \ldots, A_{m-1}\}$ is a (consistent) cyclic quasi order if for all $(i, j, k) \in \mathscr{C}_m$, $x \in A_i$, $y \in A_j$ and $z \in A_k$ we have that $(x, y, z) \in \mathscr{C}$.

[5]We emphasize the distinction of the border candidates *of a tree* and the border candidates *of an arc*, introduced in Definition 5.8

[6]This extends the definition introduced in [6] for linear orders to cyclic orders.

Lemma 5.13. *Let $D \in \mathbb{R}^{n\times n}$ be a (strict) circular Robinson dissimilarity and let $\{A_i\}_{i\in[m]}$ be a cyclic quasi-order in $[n]$. The matrix $D^{\min}(A_i, A_j) \triangleq \min\{D(k,l) : k \in A_i, l \in A_j\}$ is a (strict) circular Robinson dissimilarity.*

Notice that since $\{\partial\mathcal{T}\}_{\mathcal{T}\in\mathbf{T}}$ is an arc partition, by computing the minimum dissimilarity among all pairs of leaves we can solve the problem recursively. However, this could be computationally expensive. The next result, which is a direct consequence of Proposition A.1 in Appendix A.1, implies that we can reduce this search by only using border candidates.

Lemma 5.14. *Suppose $D \in \mathcal{C}_R^*$ and let $\mathcal{I} = [a, b]$ be an arc in $([n], \mathscr{C}_n)$. Then for every $i \notin \mathcal{I}^c$, all minimizers of $\min\{D(i,j) : j \in \mathcal{I}\}$ are contained in $\{a, b\}$.*

Therefore, given a family of trees $\mathbf{T} = \{\mathcal{T}_i\}$, we have that $D'(i,j) \triangleq \mathbf{d}^{\min}(\mathcal{T}_i, \mathcal{T}_j) \in$ pre-$\mathcal{C}_R^*$ (pre-$\mathcal{C}_R$) whenever the original dissimilarity D is in pre-$\mathcal{C}_R^*$ (pre-$\mathcal{C}_R$) and a Robinson ordering for D' yields a quasi order for D.

Observation 5.15. Computing $\mathbf{d}^{\min}(\mathcal{T}_1, \mathcal{T}_2)$ takes $\mathcal{O}(|\mathcal{B}(\mathcal{T}_1)|\cdot|\mathcal{B}(\mathcal{T}_2)|)$ operations and storing $\mathbf{d}^{\arg\min}(\mathcal{T}_1, \mathcal{T}_2)$ requires $\mathcal{O}(1)$ space, since by Lemma 5.14 we have that $|\mathbf{d}^{\arg\min}(\mathcal{T}_1, \mathcal{T}_2)| \leq 4$.

5.3.3. Connected components of the nearest-neighbours graph. Since for any enumeration of $\mathbf{T}$ we have that $D'(i,j) \triangleq \mathbf{d}^{\min}(\mathcal{T}_i, \mathcal{T}_j) \in$ pre-$\mathcal{C}_R^*$, Proposition 5.3 guarantees that Algorithm AP with input $(\mathbf{T}, \mathbf{d}^{\min})$ returns an arc partition of $\mathbf{T}$ stored in ordered tuples. The elements in each tuple α must be consecutive. Therefore, for each α we build a new tree $\mathcal{T}_\alpha$ with a Q-node in the root whose i-th children corresponds to $\alpha(i)$. In each iteration we repeat this process until we end up with a unique connected component, yielding a Q-tree.

Now we introduce the main procedure required for orienting Q-nodes within the trees.

5.3.4. Consecutive Q-nodes orientation. Recall from Corollary 5.10 that in order to orient the root of a Q-tree $\mathcal{T}_2$, it suffices to find border candidates for the arc $\partial\mathcal{T}_2$. If $\mathcal{T}_2$ is the children of a Q-node α in which $\mathcal{T}_2$ succeeds some tree $\mathcal{T}_1$ and precedes another tree $\mathcal{T}_3$, then we propose the procedure `Consecutive Orientation` (Algorithm CO).

Algorithm CO `Consecutive Orientation`

Input: Three consecutive subtrees $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$ of a Q-node α
Let $\mathcal{A}' \triangleq \mathcal{B}(\mathcal{T}_1)$, $\mathcal{B}' \triangleq \mathcal{B}(\mathcal{T}_3)$, $\mathcal{A} \triangleq \mathcal{B}^L(\mathcal{T}_2)$ and $\mathcal{B} \triangleq \mathcal{B}^R(\mathcal{T}_2)$.
Run Algorithm BCO with input $(\mathcal{A}', \mathcal{A}, \mathcal{B}, \mathcal{B}')$. Let x be the output
if $x =$`correct` **then**
    Fix the the root of $\mathcal{T}_2$
else if $x =$`reverse` **then**
    Reverse the the root of $\mathcal{T}_2$, then fix it
else
    Label $\mathcal{T}_2$'s root as non-orientable and continue the algorithm as if the root of $\mathcal{T}_2$ where fixed. This node is an actual Q-node of the tree of Robinson orderings.
end if
Result: The root of $\mathcal{T}_2$ is oriented

Since $\mathcal{A}' \triangleq \mathcal{B}(\mathcal{T}_1)$, $\mathcal{B}' \triangleq \mathcal{B}(\mathcal{T}_3)$, $\mathcal{A} \triangleq \mathcal{B}^L(\mathcal{T}_2)$ and $\mathcal{B} \triangleq \mathcal{B}^R(\mathcal{T}_2)$ are border candidates

for $\partial\mathcal{T}_2$, the correctness of the procedure is due to the correctness of Algorithm BCO. By Observation 5.11, the complexity is given by $\mathcal{O}(n \cdot \max\{|\mathcal{B}(\mathcal{T}_1)|, |\mathcal{B}(\mathcal{T}_2)|, |\mathcal{B}(\mathcal{T}_3)|\})$.

5.3.5. Complete internal orientation of elements in a connected component. Given some Q-tree $\mathcal{T}$ with $\text{depth}(\mathcal{T}) > 1$, let $\mathcal{T}^L$ (resp. $\mathcal{T}^R$) be the subtree of $\mathcal{T}$ whose root is the first (resp. last) Q-node among the direct descendants of the root of $\mathcal{T}$. A complete internal orientation of a Q-tree $\mathcal{T}$, is a process in which we determine the orientation of all nodes present a tree except from the nodes present in $\mathcal{T}^L$ and $\mathcal{T}^R$. For this task we propose the procedure `Complete Internal Orientation` (Algorithm CIO).

Algorithm CIO `Complete Internal Orientation`

Input: A Q-tree $\mathcal{T}$ with root α with $\text{depth}(\mathcal{T}) > 1$
Let $\{\mathcal{T}_i\}_{i\in[k]}$ be all subtrees children of α, excluding the first and last trees ($\mathcal{T}^L$ and $\mathcal{T}^R$)
for $i \in [k]$ **do**
    Run Algorithm CO with input $(\mathcal{T}_{i-1}, \mathcal{T}_i, \mathcal{T}_{i+1})$, where $\mathcal{T}_{-1} \triangleq \mathcal{T}^L$ and $\mathcal{T}_k \triangleq \mathcal{T}^R$
end for
Recursively repeat this process until the only unoriented node in $\mathcal{T}$ are in $\mathcal{T}^L$ and $\mathcal{T}^R$
Result: All elements in $\partial\mathcal{T} \setminus \partial\mathcal{T}^L \cup \partial\mathcal{T}^R$ are directly connected to α

The recursion is done in a Breadth-first search fashion[7]. Once again, the correctness of the procedure is due to the correctness of Algorithm BCO. As an example, consider Figure 4. Here, the tree at the top was constructed at the second recursion of the algorithm and represents a connected component of the nearest-neighbours graph over a family of Q-nodes. $\mathcal{T}^L$ corresponds to the tree with root in Q_1 and $\mathcal{T}^R$ corresponds to the tree with root in Q_k. The tree in the bottom corresponds to the tree after Algorithm CIO.

In the final recursion of Recursive Seriation Algorithm we obtain a unique connected component from `Arc Partition`, from which we construct a unique tree $\mathcal{T}$ such that $\partial\mathcal{T} = \mathcal{X}$. Here, the cyclic order of $\mathcal{X}$ implies that the subtrees $\mathcal{T}^L$ and $\mathcal{T}^R$ are consecutive. Hence, to orient these trees we make a slight variation in the procedure `Final Orientation` (Algorithm FO). This is equivalent to consider the Q-node as a ring rather than as a list.

5.3.6. External orientation of trees. Since for each $\mathcal{T} \in \mathbf{T}$, the set $\partial\mathcal{T}$ corresponds to an arc and as a consequence of Lemma 5.14, $\mathbf{d}^{\min}(\mathcal{T}, \mathcal{T}')$ is attained at some $x \in \mathcal{B}(\mathcal{T})$ and $y \in \mathcal{B}(\mathcal{T}')$ which are guaranteed to be borders of $\partial\mathcal{T}$ and $\partial\mathcal{T}'$, respectively. Therefore, we must arrange some of their internal nodes in a way that x and y lie at the borders. We propose the procedure `External Orientation` (Algorithm EO).

An important observation is that in the tree $\mathcal{T}$ resulting from the first part of this procedure we have that $\mathcal{T}^L = \{x\}$ (assuming for simplicity that $x \in \mathcal{B}^L(\mathcal{T})$). In the second part we execute Algorithm CIO with input $\mathcal{T}$. Since $\mathcal{T}^L = \{x\}$ at the end the only Q-nodes remaining to be oriented are the ones present in $\mathcal{T}^R$. As an example, we consider the Q-trees in Figure 5. Let $\mathcal{T}$ be the tree in Figure 5a. In this example, $\mathcal{T}^L$ is the subtree with root in Q_1 and $\mathcal{T}^R$ is the singleton $\{b_0\}$. Suppose by computing $\mathbf{d}^{\min}(\mathcal{T}, \mathcal{T}')$ for some other $\mathcal{T}'$ we get that $\mathbf{d}^{\min}$ is attained at $b_2 \in \mathcal{B}^L(\mathcal{T})$. In that case, we must fix $\mathcal{T}^L$ following the algorithm. Since b_2 is

[7]This way, trees of same depth are compared in Algorithm CO (excluding comparisons with $\mathcal{T}^L$ and $\mathcal{T}^R$).

Algorithm FO `Final Orientation`

Input: A Q-tree $\mathcal{T}$ with root α.
Let $\mathbf{T} \triangleq \{\mathcal{T}_i\}_{i\in[k]}$ be all subtrees of α (including $\mathcal{T}^L$ and $\mathcal{T}^R$)
if $|\mathbf{T}| > 2$ **then**
  for $i \in [k]$ **do**
    Run Algorithm CO with input $(\mathcal{T}_{i-1}, \mathcal{T}_i, \mathcal{T}_{i+1})$, where $\mathcal{T}_{-1} \triangleq \mathcal{T}^R$ and $\mathcal{T}_k \triangleq \mathcal{T}^L$
  end for
else
  In this case we have that $\mathbf{T} = \{\mathcal{T}_1, \mathcal{T}_2\}$
  To orient the root of $\mathcal{T}_1$, run Algorithm CO with input $(\mathcal{T}_2^R, \mathcal{T}_1, \mathcal{T}_2^L)$ {Notice that suffices to orient the root of $\mathcal{T}_1$ to determine the orientation of both roots}
end if
Recursively repeat this process until all nodes are oriented
Result: All Q-nodes in $\mathcal{T}$ are oriented.

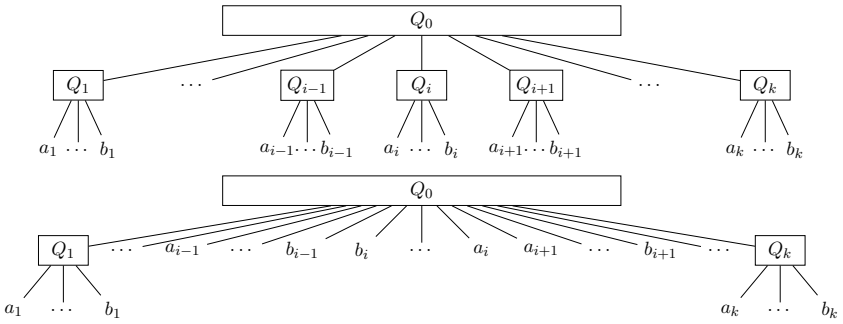
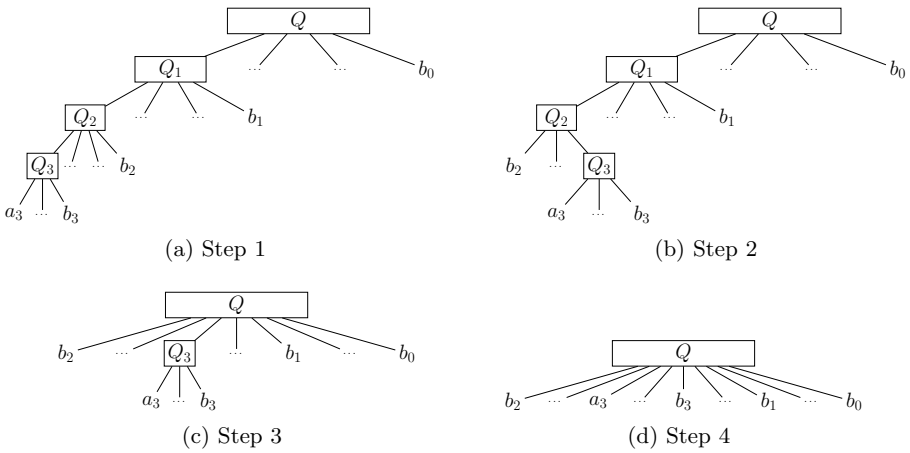


Figure 4: Example of a connected component of the nearest neighbours graph at the second recursion of the Recursive Seriation Algorithm before and after `Complete Internal Orientation`.

Algorithm EO `External Orientation`

Input: Q-trees $\mathcal{T}$ and $\mathcal{T}'$. The set $\mathbf{d}^{\arg\min}(\mathcal{T}, \mathcal{T}')$
for $(x, y) \in \mathbf{d}^{\arg\min}(\mathcal{T}, \mathcal{T}')$ **do**
  if $x \in \mathcal{B}^L(\mathcal{T})$ **then**
    Fix every Q-node in $\mathcal{T}^L$ containing x as a descendant from the root until the Q-node α where x lies in a way such that x is placed on the left
  else if $x \in \mathcal{B}^R(\mathcal{T})$ **then**
    Fix every Q-node in $\mathcal{T}^R$ containing x as a descendant from the root until the Q-node α where x lies in a way such that x is placed on the right
  end if
end for
Run Algorithm CIO with input $\mathcal{T}$
Repeat the same procedure with $\mathcal{T}'$ and y
Result: Either all Q-nodes in $\mathcal{T}^L$ (resp.$\mathcal{T}'^L$) or $\mathcal{T}^R$ (resp.$\mathcal{T}'^R$) are oriented



(a) Step 1 (b) Step 2

(c) Step 3 (d) Step 4

Figure 5: External orientation

a left border in Q_1, this node is correctly oriented. However, since b_2 appears in the right of Q_2, we must reverse Q_2 as in Figure 5b. The resulting tree is the one in Figure 5c. Next, we perform a complete orientation and the resulting tree is the one in Figure 5d.

Notice that excluding the running time of Algorithm CIO, the number of operations required for this procedure is bounded by $\mathcal{O}(\text{depth}(\mathcal{T}^L))$.

Observation 5.16. If $\mathcal{T}$ is a tree built at the k-th recursion of the algorithm, then clearly $\text{depth}(\mathcal{T}) \leq k$. We claim that since after this process either $\mathcal{T}^L$ or $\mathcal{T}^R$ gets completely oriented, then it holds that $|\mathcal{B}(\mathcal{T})| \leq k+1$. We prove this by induction on k. Notice that if $\mathcal{T}$ is composed by a single Q-node in the root, then $|\mathcal{B}(\mathcal{T})| \leq 2$. Now let $\mathcal{T}$ be a tree instantiated at the k-th recursion of the algorithm. W.l.o.g. assume $\mathcal{T}^L$ gets completely oriented. Then $\mathcal{B}(\mathcal{T}) = \mathcal{B}(\mathcal{T}^R) \cup \{x\}$. Hence, $|\mathcal{B}(\mathcal{T})| = |\mathcal{B}(\mathcal{T}^R)| + 1$. The claim follows by inducting on $\mathcal{T}^R$.

5.4. Analysis of the Recursive Seriation Algorithm.

Theorem 5.17. *Given $D \in$ pre-$\mathcal{C}_R^*$, let $\mathcal{T}$ be the PQ-tree obtained from the Recursive Seriation Algorithm with input D. Let $S(\mathcal{T})$ the set of all ordering of $\mathcal{X}$ (permutations) represented by the tree. Then, $S_{\mathcal{C}_R^*}(D) = \text{Dih}_n \circ S(\mathcal{T})$, i.e. it solves the strict circular seriation problem.*

Proof Sketch. For simplicity, suppose in the Recursive Seriation Algorithm we omit the orientation steps and leave them to the end of the process. This does not affect the set of solutions but may increase the time complexity. Denote $\mathcal{T}^{pre}$ and $\mathcal{T}$ the trees before and after orientation, respectively. Also let $\mathbf{T}_k$ the family of trees instantiated at the k-th recursive step. Notice that by Lemma 5.13, evaluating $\mathbf{d}^{\min}$ over $\mathbf{T}_k$ yields a dissimilarity matrix $D_k \in$ pre-$\mathcal{C}_R^*$ (a (permuted) submatix of D). Due to Proposition 5.3, we have that $S_{\mathcal{C}_R^*}(D) \subset \text{Dih}_n \circ S(\mathcal{T}^{pre})$ (at least all Robinson orderings are considered at this point). To complete the proof, it remains to show that in $\mathcal{T}$ all orientable Q-nodes originally in $\mathcal{T}^{pre}$ had been correctly fixed. To see this notice that the orientation of each Q-node in $\mathcal{T}^{pre}$ is tested either by Algorithm EO or

Algorithm CO. The correctness of Algorithm EO is due to Lemma 5.14. The correctness of Algorithm CO is due to Corollary 5.10. ∎

Theorem 5.18. *The Recursive Seriation Algorithm runs in $\mathcal{O}(n^2)$ time.*

Proof. We count the number of operations required by the procedure Algorithm CIO and Algorithm FO separately from the rest. At the i-th recursion let $\mathbf{T}(i)$ be the input Q-trees, let $k(i) \triangleq |\mathbf{T}(i)|$ and let $b(i) = \max_{\mathcal{T}\in\mathbf{T}(i)} |\mathcal{B}(\mathcal{T})|$. Then, by Observation 5.15, computing $\mathbf{d}^{\min}$, takes $\mathcal{O}(k(i)^2 \cdot b(i)^2)$ operations. By Observation 5.16, the complexity of the procedure Algorithm EO takes $\mathcal{O}(k(i)^2)$ operations. Computing G_{NN} takes $\mathcal{O}(k(i)^2)$ operations. The procedure Algorithm DFS takes $\mathcal{O}(k(i))$ operations.

On the other hand, notice that in each step of the recursion, every tree is merged to its nearest neighbour. This implies that $k(i) \leq \frac{n}{2^i}$ and, therefore, the depth of the recursion is bounded by $\log_2(n)$. Since by Observation 5.16 $b(i) \leq i+1$ then, there is some constant $C_1 > 0$ such that the the total number of operations of this procedure is bounded by $C_1 \sum_{i=0}^{\log_2(n)} \left(\frac{n}{2^i}\right)^2 (i+1)^2 + \left(\frac{n}{2^i}\right)^2 + \left(\frac{n}{2^i}\right) = \mathcal{O}(n^2)$.

It remains to consider Algorithm CIO and Algorithm FO. In this procedures, all Q-nodes α are oriented through Algorithm CO with input $(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$ where α is the root of $\mathcal{T}_2$. To count the operations of this procedure we consider two cases. The first (and most common) case is when $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ are trees instantiated at the same recursive step. In this case, if they where instantiated at the i-th recursion then by Observation 5.16 and Observation 5.11 the orientation takes $\mathcal{O}((i+1)\cdot n)$ operations.

By counting on the recursion where each node was instantiated, the total number of operations involving first case Q-nodes can be bounded by $C_2 \cdot \sum_{i=0}^{\log_2(n)} \left(\frac{n}{2^i}\right)(i+1)\cdot n = \mathcal{O}(n^2)$.

A second case to consider is during the complete orientation of a connected component. Let $\mathcal{T}$ be a tree generated from a connected component of the nearest-neighbours graph at the i-th recursion of the algorithm. Then, in Algorithm CIO (or Algorithm FO) with input $\mathcal{T}$, some of the internal Q-nodes will be oriented by having as border candidates $\mathcal{B}(\mathcal{T}^L)$ and $\mathcal{B}(\mathcal{T}^R)$. Since $\mathrm{depth}(\mathcal{T}) \leq i$, this can occur for $\mathcal{B}(\mathcal{T}^L)$ (resp. $\mathcal{B}(\mathcal{T}^R)$) for at most i internal Q-nodes of $\mathcal{T}$. If $C(i)$ be the number of connected component found in the i-th recursion, then the number of second case Q-nodes is at most $C(i)\cdot i\cdot 2$. Since by Observation 5.16, $|\mathcal{B}(\mathcal{T}^L)| \leq i$ and $|\mathcal{B}(\mathcal{T}^R)| \leq i$, the number of operations required for orienting all this nodes is bounded by $\mathcal{O}(C(i)\cdot i^2 \cdot n)$. Again, by counting through the recursion levels and considering that $C(i) \leq n/2^i$, the total cost of orienting second case Q-nodes is bounded by $C_3 \cdot \sum_{i=0}^{\log_2(n)} \left(\frac{n}{2^i}\right) \cdot i^2 \cdot n = \mathcal{O}(n^2)$, which proves the result. ∎

5.5. *PQ*-tree of solutions in the strict Robinson case. It is clear that if a sequence is strictly monotone the only permutation that preserves this property is the one that reverses the sequence. Therefore if $D \in \mathcal{L}^*_R$, we have that $S_{\mathcal{L}^*_R}(D) = \{\mathbf{e}, \mathbf{r}\} \cong \mathrm{Dih}_1$. However, it is not immediately clear which permutations are the ones that preserve the strict unimodality. The next Lemma will let us conclude that there is at most one *non trivial* ordering for $D \in \mathcal{C}^*_R$.

Lemma 5.19. *Let $D \in \mathcal{C}^*_R$ and let $\mathcal{I}_1, \ldots \mathcal{I}_k$ be disjoint arcs of $[n]$. Let $\sigma_{\mathcal{I}_i}$ be the permutation that reverses $\mathcal{I}_i$. Then, at most one of the $\sigma_{\mathcal{I}_i}$'s produces a new Robinson ordering.*

Proof. Suppose $\sigma_{\mathcal{I}}$ is a Robinson ordering for some arc $\mathcal{I}$. For every $i \in [n]$, let $M(i) = \arg\max_j D(i,j)$. We claim that for every $i \notin \mathcal{I}$ it holds that $M(i) \subset \mathcal{I}$. Otherwise, given

$m^* \in M(i)$, by the connectivity of $\mathcal{I}$, we must have that $\mathcal{I}$ must be strictly contained in one of the two paths connecting i and m^*. Also notice that $D(i,\cdot)$ is strictly monotone in such path. Hence, reversing $\mathcal{I}$ would violate the monotonicity of such sequence (an thus the unimodality of the whole sequence). This proves the claim. Since $\mathcal{I}^c$ is an arc, by the same argument we have that $i \in \mathcal{I}$ implies $M(i) \subset \mathcal{I}^c$. Hence, the only arcs that can be reversed are $\mathcal{I}$ and $\mathcal{I}^c$. ■

6. Behavior for large n. The literature on the seriation problem has mostly focused on finite ordered sets, either linearly or cyclically ordered, on suitable classes of matrices encoding properties of this order, such as Robinson matrices, and on efficient algorithms for its solution. However, typically the use of seriation algorithms is motivated by the interpretation of data as embedded in a closed curve, and it is unclear how these combinatorial solutions relate to the underlying order of a continuous object.

To bridge this gap, we provide a simple generative model of sampling from a continuous and periodic structure. That sample, and more specifically the dissimilarities between pairs of points from the sample, will be the input of our strict seriation algorithm. The question we want to answer is: To which extent the solution obtained by the seriation algorithm applied to a random sample reflects the underlying ordering of the periodic structure? We will answer this question by proving that as the sample size n grows, the expected Kendall-tau distance from the strict circular seriation algorithm solution to the order inherited from the continuous model decreases at a rate $\mathcal{O}(\log(n)/n)$.

6.1. Reduction to $\mathbb{S}^1$. We will consider our periodic continuous structure as parameterized by the unit circle. Equivalently, we will use the set $[0,1)$ as the set of points, where we topologically identify 0 and 1, making it a circular-like structure. This set is endowed with the *natural cyclic order*, which results from embedding $[0,1)$ into $\mathbb{S}^1$. We assume the set $[0,1)$ is endowed with a dissimilarity $\mathbf{d}$. We will make some assumptions that relate the circular ordering to the circular Robinson property.

Assumption 6.1. $\mathbf{d}$ *is continuous, and* strict circular Robinson*, i.e.,*

$$\forall \text{ cyclically ordered } x,y,z,w \in [0,1): \ \mathbf{d}(y,w) > \min\{\mathbf{d}(y,x), \mathbf{d}(y,z)\}. \tag{6.1}$$

One natural question is how general this continuous model is. We claim that the assumption that our sample space is the unit circle is without loss of generality. For example, if the sample space is a one dimensional compact manifold of $\mathbb{R}^d$, we can parameterize the manifold by its arc-length $\gamma : [0,1) \mapsto \mathbb{R}^d$, and let $\mathbf{d}(t,s) := \|\gamma(t) - \gamma(s)\|$, which is clearly continuous. Notice however that the validity of the strict circular Robinson property is not guaranteed in this example: such assumption depends on the relative positions of points in space.

6.2. Solutions in the limit. To understand the set of solutions in the limit we first need to characterize the natural symmetries of the strict Robinson dissimilarity $\mathbf{d}$. To do so, we consider the family of *cyclic shifts* $\{\pi_s : s \in [0,1)\}$ defined by $\pi_s(t) = t + s \bmod 1$, and the *reversal* $\pi_r(t) = 1 - t$. We let $\mathrm{Dih}_\infty := \langle \pi_s, \pi_r : s \in [0,1)\rangle$. In addition, given an arc $\mathcal{I} := (t,s) \subsetneq [0,1)$, we let $\sigma_{\mathcal{I}}$ be the bijection that reverses $\mathcal{I}$ and fixes $\mathcal{I}^c$. Since in the finite case all solutions can be expressed as compositions of such permutations, in the continuous case we look for solutions in $\mathrm{Sym}(\infty) \triangleq \mathrm{Dih}_\infty \circ \langle \sigma_{\mathcal{I}} : \mathcal{I} \text{ arc}\rangle$.

Theorem 6.2. *Suppose* $\mathbf{d}$ *satisfies Assumption 6.1, and let* $\pi \in \mathrm{Sym}(\infty)$. *If* $\mathbf{d} \circ \pi$ *is strict circular Robinson then* $\pi \in \mathrm{Dih}_\infty$.

This result can be seen as a well-posedness statement of the seriation problem in the continuous limit. Our next goal is to study its consequences for large (but finite) sample size.

6.3. Approximate well-posedness of seriation in the large n regime. We now propose a sampling model from the continuous model. We uniformly at random extract a size n sample from $[0,1)$. We denote this sample by $\mathcal{X}_n := \{x_0, \dots, x_{n-1}\}$. If we let λ be the Lebesgue measure on $[0,1)$, then our sampling is distributed as λ^n. Let $D_{\mathcal{X}_n}$ denote the dissimilarity matrix associated to $\mathcal{X}_n$. In particular, if $x_0, \dots, x_{n-1}$ are cyclically ordered, then the dissimilarity matrix is strict circular Robinson (cf. Assumption 6.1).

Despite that in the continuous case there is a unique Robinson ordering, with finitely many samples there might exists non-trivial orderings (cf. Lemma 5.19). In what follows we study conditions under which for a large sample, any ordering in $S_{\mathcal{C}_R^*}(D_{\mathcal{X}_n})$ is close to the one induced by the curve. Our closeness measure is given by the Kendall-tau's metric τ_K and the goal is to bound the expected value of the diameter of the set of solutions:

Definition 6.3 (Kendall-tau's metric [13, 18]). We define the Kendall-tau distance between permutations π_1 and π_2 as $\tau_K(\pi_1, \pi_2) \triangleq |\mathcal{G}(\pi_1, \pi_2)| / \binom{n}{2}$, where $\mathcal{G}(\pi_1, \pi_2)$ corresponds to the set of discordant pairs defined as

$$\mathcal{G}(\pi_1, \pi_2) \triangleq \{(i,j) : i < j, [\pi_1(i) < \pi_1(j) \wedge \pi_2(i) > \pi_2(j)] \vee [\pi_1(i) > \pi_1(j) \wedge \pi_2(i) < \pi_2(j)]\}.$$

The denominator $\binom{n}{2}$ ensures that $\tau_K(\pi_1, \pi_2) \in [0,1]$. The next definition of diameter takes into account that for seriation cyclic permutations provide the same ordering.

Definition 6.4. Given a set $S \subset \mathrm{Sym}(n)$, the diameter of the S is defined as $\mathrm{diam}(S) \triangleq \max_{\pi_1, \pi_2 \in S} \min_{\hat{\pi}_1 \in \mathrm{Dih}_n \circ \pi_1} \tau_K(\hat{\pi}_1, \pi_2)$.

Let $\mathrm{Arc} : [0,1) \times [0,1) \to [0, \frac{1}{2}]$ be the length of the shortest arc connecting two points in the unit circle, i.e. $\mathrm{Arc}(\theta_1, \theta_2) = \min\{|\theta_1 - \theta_2|, 1 - |\theta_1 - \theta_2|\}$. To prove rates on the Kendall-tau distance we make a final assumption. This condition allows us to avoid making overly restrictive metric assumptions on the dissimilarity, but still enjoying a weaker form of distance.

Assumption 6.5. *The dissimilarity* $\mathbf{d}$ *satisfies the following bi-Lipschitz property:*

$$(\exists L \geq \ell > 0)(\forall s, t \in [0,1)) \qquad \ell \cdot Arc(s,t) \leq \mathbf{d}(s,t) \leq L \cdot Arc(s,t). \tag{6.2}$$

We conclude this Section by providing a rate on the expected Kendall-tau diameter of the set of solutions of the circular Robinson algorithm. Hence, all these solutions must be close to the underlying order of the continuous model. Its proof is deferred to Appendix A.4.

Theorem 6.6. *Let* $\mathcal{X}_n = \{x_0, x_1, \dots x_{n-1}\} \overset{iid}{\sim} \mathrm{Unif}[0,1)$. *Then given any* $\mathbf{d}$ *satisfying Assumption 6.1 and Assumption 6.5 we have that*

$$\mathbb{E}_{\mathcal{X}_n}\Big[\mathrm{diam}(S_{\mathcal{C}_R^*}(D_{\mathcal{X}_n}))\Big] = \mathcal{O}\Big(\frac{(L+\ell)}{\ell} \cdot \frac{\log(n)}{n}\Big). \tag{6.3}$$

Appendix A. Proofs.

A.1. Proof of Proposition 3.7.

We need two auxiliary results first.

Proposition A.1. *f is unimodal (resp. strictly unimodal) if and only if for $i \le j \le k$ we have $f_j \ge \min\{f_i, f_k\}$ (resp. $f_j > \min\{f_i, f_k\}$).*

Proof. Suppose f is unimodal, let m be a mode, and suppose there are i, j, k, not all equal, such that $i \le j \le k$ and $x_j < \min\{x_i, x_k\}$. Then $x_i > x_j$ and $x_j < x_k$. This implies $m \le j$ and $m \ge j$. Hence $m = j$. This is a contradiction. Now, suppose f satisfies the inequality but has no mode. Then j is not a mode, and there is $i < j$ and $k > j$ such that $f_i > f_j$ and $f_k > f_j$. This is a contradiction. The proof for the strictly unimodal case follows from the same arguments. ∎

Proposition A.2. *If $(i_{-1}, i_0, i_1), (i_0, i_1, i_2) \in \mathscr{C}_n$ then for each $k \in \{-1, 1, 2\}$ there is $q_k \in [n]$ such that $i_k = i_0 + q_k \mod n$. Furthermore, $q_1 \le q_2 \le q_{-1}$.*

Proof. Consider $q_k = i_k - i_0 \mod n$. Then $q_0 = 0$. Since cyclic shifts do not change cyclic orderings, this implies $q_1 \le q_2 \le q_{-1}$. This proves the proposition. ∎

Proof of Proposition 3.7. For simplicity we define $d^i_j \triangleq D(i, i + j \mod n)$. $(2 \Rightarrow 1)$ From Proposition A.2 we can write $i = j + q_i$, $k = j + q_k$ and $\ell = j + q_\ell$ with $q_k \le q_\ell \le q_i$. Since d^j is unimodal, from Proposition A.1 we deduce $d^j_{q_\ell} \ge \min\{d^j_{q_k}, d^j_{q_i}\}$. $(1 \Rightarrow 2)$ If d^j is not unimodal, by Proposition A.1 there are $q_k \le q_\ell \le q_i$ with $d^j_{q_\ell} < d^j_{q_k}$ and $d^j_{q_\ell} < d^j_{q_i}$. If we define $i = j + q_i \mod n$, $k = j + q_k \mod n$ and $\ell = j + q_\ell \mod n$ we see that $(i, j, k), (j, k, \ell) \in \mathscr{C}_n$. This contradicts 1. ∎

A.2. Proofs for Section 5.

Proof of Lemma 5.9. Let $z \in \mathcal{X}$ and denote $f_z(\cdot) \triangleq \mathbf{d}(z, \cdot)$. First, notice that

$$(\exists r > 0).\{a, a'\} \subset B_r(z) \ \wedge \ \{b, b'\} \subset B_r(z)^c \\ \Leftrightarrow \max\{f_z(a), f_z(a')\} < \min\{f_z(b), f_z(b')\}. \tag{A.1}$$

$(1. \Rightarrow 2.)$ By (A.1), this implication is direct from the fact that $a \in \mathcal{A}, b \in \mathcal{B}, a' \in \mathcal{A}'$ and $b' \in \mathcal{B}'$. $(2. \Rightarrow 1.)$ Let $r \triangleq \max\{f_z(x), f_z(x')\}$, thus $\{x, x'\} \subset B_r(z)$ and $\{y, y'\} \subset B_r(z)^c$. By Proposition 3.4 this ball is an arc, and therefore is connected in any Robinson ordering. This implies that all elements in between x and x' (in all Robinson orderings), including a and a' must also be present in $B_r(z)$. Similarly, all elements in between y and y', including b and b' must not be present in $B_r(z)$. The implication follows from (A.1). $(3. \Rightarrow 2.)$ Direct. $(2. \Rightarrow 3.)$ Notice that given any $z \in \mathcal{X}$ and any $t > 0$ we have that if there is some $a \in \mathcal{A}$ and $a' \in \mathcal{A}'$ such that $\max\{f_z(a), f_z(a')\} < t$. Then, $f_z(a) < t \wedge f_z(a') < t$. Which implies $\max\{\min f_z(\mathcal{A}), \min f_z(\mathcal{A}')\} < t$. Similarly, the existence of $b \in \mathcal{B}$ and $b' \in \mathcal{B}'$ such that $\min\{f_z(b), f_z(b')\} > t$ implies that $\min\{\max f_z(\mathcal{B}), \max f_z(\mathcal{B}')\} > t$. This proves the final implication, and hence the result. ∎

Proof of Lemma 5.13. Let $x_d \in B_d$ and $x_b \in B_b$ be such that $D^{\min}(B_b, B_d) = D(x_b, x_d)$, and let $x_c \in B_c, x_a \in B_a$ be arbitrary. We notice that x_a, x_b, x_c, x_d is cyclically ordered, hence

$$D^{\min}(B_b, B_d) = D(x_b, x_d) \ge (>) \min\{D(x_b, x_a), D(x_b, x_c)\}.$$

On the other hand,

$$\min\{D(x_b, x_a), D(x_b, x_c)\} \geq \min\{D^{\min}(B_b, B_a), D^{\min}(B_b, B_c)\}$$

by definition of $D^{\min}$, proving the result. ∎

A.3. Proof of Theorem 6.2.

Proof. Suppose by contradiction that there exist an arc $\mathcal{I} = [a, b)$ such that $\mathbf{d} \circ \sigma_{\mathcal{I}}$ is strict Robinson. For $\epsilon > 0$ small, $a - \epsilon, a, b, b + \epsilon$ are cyclically ordered. By hypothesis,

$$\mathbf{d}(\sigma(a), \sigma(b + \epsilon)) > \min\{\mathbf{d}(\sigma(a), \sigma(a - \epsilon)), \mathbf{d}(\sigma(a), \sigma(b))\}, \tag{A.2}$$

and since $b + \epsilon, a - \epsilon \notin \mathcal{I}$, we get that $\sigma(a - \epsilon) = a - \epsilon$ and $\sigma(b + \epsilon) = b + \epsilon$. On the other hand, $\sigma(a) = b$ and $\sigma(b) = a$. Therefore, we can rewrite (A.2) as

$$\mathbf{d}(b, b + \epsilon) > \min\{\mathbf{d}(b, a - \epsilon), \mathbf{d}(b, a)\}. \tag{A.3}$$

Let $\delta := \mathbf{d}(b, a) > 0$. By continuity we get that $\mathbf{d}(\sigma(a)), \sigma(b + \epsilon)) \to 0$ and $\mathbf{d}(b, a - \epsilon) \to \delta$ as $\epsilon \to 0$. For sufficiently small ϵ, this is a contradiction with (A.3). ∎

A.4. Proof of Theorem 6.6. Given $x_0, \ldots, x_{n-1} \in [0, 1)$, the *order statistics* correspond to the variables $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$ obtained by sorting the samples by increasing order. The *gaps* of the sample correspond to the variables $w_i \triangleq x_{(i+1)} - x_{(i)}$. Let $\epsilon_n \triangleq \max_{i \in [n]} w_i$. The following result can be found in [21, Theorem 1.2].

Proposition A.3. *Suppose* $x_0, x_1, \ldots x_{n-1} \overset{iid}{\sim} \mathrm{Unif}[0, 1)$. *Then,*

$$\mathbb{P}(\epsilon_n \geq z) \leq \sum_{j=1}^{n+1} (-1)^{j-1} \binom{n+1}{j} (1 - jz)_+^n$$

Proposition A.4. *Given any* $\mathcal{I} = [x_i, x_j]$, *we write* $\mu(\mathcal{I})$ *to denote* $\mathrm{Arc}(x_i, x_j)$. *Suppose that Assumption 6.1 and Assumption 6.5 hold. Then the inequality* $\epsilon_n < \ell\delta/(L + \ell)$ *implies that any arc* $\mathcal{I} \subset \mathcal{X}_n$ *such that* $\mu(\mathcal{I}) > \delta$ *has a unique orientation in any circular Robinson ordering of* $D_{\mathcal{X}_n}$.

Proof. Let $s \triangleq x_{(i)}$ and $t \triangleq x_{(j)}$ for some $i < j$. Let $\delta \in (0, \frac{1}{2})$ and consider the arc $\mathcal{I} = [s, t]$ in $\mathcal{X}_n$. Suppose $\epsilon_n < \ell\delta/(L + \ell)$ and $\mu(\mathcal{I}) > \delta$. Let $s^+ = x_{(i-1 \bmod n)}$ and $t^+ = x_{(j+1 \bmod n)}$. We claim that $B_s(\mathbf{d}(s, s^+)) \between^* \mathcal{I}$. To prove the claim, it suffices to prove that

$$\mathbf{d}(s, s^+) < \min\{\mathbf{d}(s, t), \mathbf{d}(s, t^+)\}. \tag{A.4}$$

First, notice that $\mathbf{d}(s, s^+) \leq L \cdot \epsilon_n$. Second, notice that since $\mathrm{Arc}(s, t) > \delta$, then

$$\mathbf{d}(s, t^+) \geq \ell \cdot \mathrm{Arc}(s, t^+) \geq \ell \cdot (\mathrm{Arc}(s, t) - \epsilon_n) \geq \ell \cdot (\delta - \epsilon_n),$$

and therefore $\min\{\mathbf{d}(s, t), \mathbf{d}(s, t^+)\} \geq \ell \cdot (\delta - \epsilon_n)$. Joining this two results with the fact that $\epsilon_n < \ell\delta/(L + \ell)$ a proves the claim. ∎

Lemma A.5. *Let $n \geq \log(1/\delta)/\delta$, and let $x_0, x_1, \dots x_{n-1} \overset{iid}{\sim} \mathrm{Unif}[0,1)$. Let $E_n(\delta) = \{\epsilon_n < \ell\delta/(L+\ell)\}$, then*

$$\int_{E_n(\delta)} \operatorname{diam} S_{\mathcal{C}_R^*}(D_{\mathcal{X}_n}(\omega))\ d\lambda^n(\omega) = \mathcal{O}\Big(\delta^2 + \frac{\delta \log(1/\delta)}{n}\Big).$$

Proof. Recall from Lemma 5.19 that there is at most one non-trivial ordering $\sigma_{\mathcal{I}}$ of $D_{\mathcal{X}_n}$ which corresponds to the permutation that reverses the arc $\mathcal{I} \cap \mathcal{X}_n$. Therefore, it suffices to bound the integral of the random variable $\tau_K(\mathrm{id}, \sigma^*)$, where $\sigma^* \in \arg\min_{\hat{\sigma} \in \{\pi_{\mathbf{r}} \circ \sigma_{\mathcal{I}}, \sigma_{\mathcal{I}}\}} \tau_K(\mathrm{id}, \hat{\sigma})$.

The number of discordant pairs between σ^* and id is bounded by $\frac{1}{2}\min\{|\mathcal{I} \cap \mathcal{X}_n|, n - |\mathcal{I} \cap \mathcal{X}_n|\}^2$. Therefore, we will focus in bounding this expression. Let $\omega \in E_n(\delta)$. By Proposition A.4, $\mu(\mathcal{I}(\omega)) \leq \delta$. This implies that either $\lambda(\mathcal{I}) \leq \delta$ or $\lambda(\mathcal{I}) \geq 1 - \delta$. Which leads to the bound

$$\begin{aligned}\int_{E_n(\delta)} &\min\{|\mathcal{I} \cap \mathcal{X}_n|, n - |\mathcal{I} \cap \mathcal{X}_n|\}^2\ \mathrm{d}\lambda^n(\omega) \\ &\leq \int_{E_n(\delta)} \max_{\mathcal{J}\text{interval},\ \lambda(\mathcal{J}) \leq \delta} |\mathcal{J} \cap \mathcal{X}_n|^2\ \mathrm{d}\lambda^n(\omega) \leq \int_{\Omega} \max_{\mathcal{J}\text{interval},\ \lambda(\mathcal{J}) \leq \delta} |\mathcal{J} \cap \mathcal{X}_n|^2\ \mathrm{d}\lambda^n(\omega).\end{aligned}$$ ∎

We bound the random variable inside the integral using a balls and bins argument. W.l.o.g. $1/\delta$ is an integer. Let $(\mathcal{J}_i)_{i \in [1/\delta]}$ be a partition of $[0,1)$ by disjoint intervals of length δ. For any $\omega \in \Omega$, the maximizer in the integral above lies in at most two of the partition intervals. Therefore, $\max_{\lambda(\mathcal{J}) \leq \delta} |\mathcal{J} \cap \mathcal{X}_n| \leq 2 \max_{i \in [n]} |\mathcal{J}_i \cap \mathcal{X}_n|$. Next, we can estimate $\max_{i \in [n]} |\mathcal{J}_i \cap \mathcal{X}_n|$ by looking into the problem of throwing n balls into $1/\delta$ bins (see [24] for further details); since we further assumed that $n \geq 1/\delta \log(1/\delta)$, then w.h.p., the maximum occupancy is bounded by $n\delta + \Theta(\sqrt{n\delta \log(1/\delta)})$. Plugging this bound above yields the result.

Proof of Theorem 6.6. Let $E_n(\delta)$ denote the event $\{\epsilon_n < \ell\delta/(L+\ell)\}$. Denote the random variable $Z = \operatorname{diam}(S_{\mathcal{C}_R^*}(D_{\mathcal{X}_n}))$. Then,

$$\mathbb{E}[Z] = \textstyle\int_{E_n(\delta)} Z(\omega)\mathrm{d}\lambda(\omega) + \int_{E_n(\delta)^c} Z(\omega)\mathrm{d}\lambda(\omega) \leq \mathcal{O}\Big(\delta^2 + \frac{\delta \log(1/\delta)}{n}\Big) + \mathbb{P}[E_n(\delta)^c], \tag{A.5}$$

where in the inequality we used Lemma A.5 and $\tau_K \leq 1$. By Proposition A.3 we have

$$\mathbb{P}[E_n(\delta)^c] \leq \textstyle\sum_{j=1}^{n+1} (-1)^{j-1} \binom{n+1}{j} (1 - jx)_+^n \leq \sum_{j=1}^{n+1} \Big(\frac{e(n+1)}{j}\Big)^j \exp\{-xjn\}, \tag{A.6}$$ ∎

where $x = \ell\delta/(L+\ell)$. By taking $\delta(n) = (L+\ell)\cdot\log(e(n+1)^2)/(n\cdot\ell)$ we obtain that (A.6) can be bounded by $\sum_{j=1}^{n+1} \frac{1}{(n+1)^j} \in \mathcal{O}(1/n)$. By (A.5) and (A.6) we conclude that $\mathbf{E}[Z] = \mathcal{O}(\frac{L+\ell}{\ell}\frac{\log n}{n})$.

Appendix B. Auxiliary subroutines.

Algorithm AP `Arc Partition`

Input: A dissimilarity $\mathbf{d}$ and a set $\mathbf{T}$
$\mathcal{N}_G(x) \triangleq \{y \in \mathbf{T} : x \in \mathrm{NN}(y) \vee y \in \mathrm{NN}(x)\}$ {Compute the neighbourhood function}
$\mathcal{B} \triangleq \{x \in \mathbf{T} : |\mathcal{N}_G(x)| = 1\}$
{Find all degree 1 nodes (if there are no such nodes pick any)}
$i = 0$ {Run DFS starting at every non visited degree 1 node}
for $x \in \mathcal{B} \setminus \cup_{j<i}\alpha_i$ **do**
  $\alpha_i = \mathrm{DFS}(\mathcal{N}_G, \emptyset, x)$
  $i = i + 1$
end for
Output: An arc partition stored into tuples $\mathcal{P} \triangleq \{\alpha_i\}_{i\in[k]}$

Algorithm DFS `Depth-First Search`

Input: The neighbourhood function of a graph $\mathcal{N}_G(\cdot)$, a tuple α of visited nodes and a starting node x
$\alpha(n) = x$ {Set x as n-th visited node where n is the size of α}
for $y \in \mathcal{N}_G(x) \setminus \alpha$ **do**
  $\alpha = \mathrm{DFS}(\mathcal{N}_G, \alpha, y)$ {Recurse over all adjacent nodes that have not been visited}
end for
return α
Output: A tuple of visited nodes α

Acknowledgments. We would like to thank Alexandre d'Aspremont for valuable discussions at different stages of this work.

REFERENCES

[1] J. E. Atkins, E. G. Boman, and B. Hendrickson, *A spectral algorithm for seriation and the consecutive ones problem*, SIAM Journal on Computing, 28 (1998), pp. 297–310.
[2] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural computation, 15 (2003), pp. 1373–1396.
[3] K. S. Booth and G. S. Lueker, *Linear algorithms to recognize interval graphs and test for the consecutive ones property*, in Proceedings of the seventh annual ACM symposium on Theory of computing, 1975, pp. 255–265.
[4] K. S. Booth and G. S. Lueker, *Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms*, Journal of computer and system sciences, 13 (1976), pp. 335–379.
[5] F. Brucker and C. Osswald, *Hypercycles and dissimilarities*, Journal of Classification, accepté, (2008).
[6] V. Chepoi and B. Fichet, *Recognition of robinsonian dissimilarities*, Journal of Classification, 14 (1997), pp. 311–325.
[7] V. Chepoi, B. Fichet, and M. Seston, *Seriation in the presence of errors: Np-hardness of ℓ^∞-fitting robinson structures to dissimilarity matrices*, Journal of classification, 26 (2009), pp. 279–296.
[8] V. Chepoi and M. Seston, *Seriation in the presence of errors: A factor 16 approximation algorithm for ℓ^∞-fitting robinson structures to distances*, Algorithmica, 59 (2011), pp. 521–568.

[9] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, *Graph laplacian tomography from unknown random projections*, IEEE Transactions on Image Processing, 17 (2008), pp. 1891–1899.
[10] X. Evangelopoulos, A. J. Brockmeier, T. Mu, and J. Y. Goulermas, *Circular object arrangement using spherical embeddings*, Pattern Recognition, 103 (2020), p. 107192.
[11] D. Fulkerson and O. Gross, *Incidence matrices and interval graphs*, Pacific journal of mathematics, 15 (1965), pp. 835–855.
[12] L. Hubert, P. Arabie, and J. Meulman, *Graph-theoretic representations for proximity matrices through strongly-anti-robinson or circular strongly-anti-robinson matrices*, Psychometrika, 63 (1998), pp. 341–358.
[13] M. G. Kendall, *A new measure of rank correlation*, Biometrika, 30 (1938), pp. 81–93.
[14] J. Köbler, S. Kuhnert, and O. Verbitsky, *Circular-arc hypergraphs: Rigidity via connectedness*, Discrete Applied Mathematics, 217 (2017), pp. 220–228.
[15] M. Laurent and M. Seminaroti, *Similarity-first search: a new algorithm with application to robinsonian matrix recognition*, SIAM Journal on Discrete Mathematics, 31 (2017), pp. 1765–1800.
[16] Y.-C. Liao, H.-W. Cheng, H.-C. Wu, S.-C. Kuo, T.-L. Lauderdale, and F.-J. Chen, *Completing circular bacterial genomes with assembly complexity by using a sampling strategy from a single minion run with barcoding*, Frontiers in Microbiology, 10 (2019), p. 2068.
[17] I. Liiv, *Seriation and matrix reordering methods: An historical overview*, Statistical Analysis and Data Mining: The ASA Data Science Journal, 3 (2010), pp. 70–91.
[18] R. Ma, T. Tony Cai, and H. Li, *Optimal permutation recovery in permuted monotone matrix model*, Journal of the American Statistical Association, (2020), pp. 1–15.
[19] B. G. Mirkin and S. N. Rodin, *Graphs and genes. Biomathematics*, Springer-Verlag). Springer, 1984.
[20] V. Novák, *Cyclically ordered sets*, Czechoslovak Mathematical Journal, 32 (1982), pp. 460–473.
[21] I. Pinelis, *Order statistics on the spacings between order statistics for the uniform distribution*, arXiv preprint arXiv:1909.06406, (2019).
[22] P. Préa and D. Fortin, *An optimal algorithm to recognize robinsonian dissimilarities*, Journal of Classification, 31 (2014), pp. 351–385.
[23] A. Quilliot, *Circular representation problem on hypergraphs*, Discrete mathematics, 51 (1984), pp. 251–264.
[24] M. Raab and A. Steger, *"balls into bins"—a simple and tight analysis*, in International Workshop on Randomization and Approximation Techniques in Computer Science, Springer, 1998, pp. 159–170.
[25] A. Recanati, T. Brüls, and A. d'Aspremont, *A spectral algorithm for fast de novo layout of uncorrected long nanopore reads*, Bioinformatics, 33 (2017), pp. 3188–3194.
[26] A. Recanati, T. Kerdreux, and A. d'Aspremont, *Reconstructing latent orderings by spectral clustering*, arXiv preprint arXiv:1807.07122, (2018).
[27] W. S. Robinson, *A method for chronologically ordering archaeological deposits*, American antiquity, 16 (1951), pp. 293–301.
[28] M. Seston, *Dissimilarités de Robinson: algorithmes de reconnaissance et d'approximation*, PhD thesis, Aix-Marseille 2, 2008.
[29] A. Tucker, *Matrix characterizations of circular-arc graphs*, Pacific Journal of Mathematics, 39 (1971), pp. 535–545.