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 in a circular order instead. 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 \( 2\pi \). 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...
Our contributions. In this work, we address the problem of circular seriation. In particular, we address algorithmic questions and study its 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, i.e., with \( O(n^2) \) time and space complexity, for the seriation problem on 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 for the large \( n \) regime. In this model, points are sampled from a closed curve, which without loss of generality we assume is the unit circle, with a continuous and strict circular Robinson dissimilarity. Our main result here is an \( O(\log(n)/n) \) bound on the expected Kendall-tau distance of any strict circular Robinson ordering of the data. This result is based on an observation 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, 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 \( O(n^4) \). 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 \( O(n^3) \) operations and \( O(n^2) \) space. Using similar techniques, Seston [28] improved the complexity to \( O(n^2 \log(n)) \). 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 \( O(n^2) \) algorithm, using an algorithm from [4] to first compute a \( PQ \)-tree which is then updated by the algorithm. For the sparse case, Laurent

\(^1\)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.
and Seminarotti [15] present the Similarity-First Search algorithm with $O(n^2 + nm \log n)$ 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. Here, it is known that projecting a dissimilarity on the class of Robinsonian dissimilarities (in $\ell_\infty$-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 [26] 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 $j \mapsto D(i, j + i \mod n)$ 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 $\text{Sym}(n)$ the group of permutations of $[n]$. A permutation is represented either by a vector $\pi$ with entries in $[n]$ or by an $n \times n$ orthogonal and $\{0,1\}$-matrix $\Pi$. We denote as $\pi_r$ the permutation that reverses the elements of $[n]$, i.e., $\pi_r(i) = n - 1 - i$, and $\pi_s$ the cyclic (right) shift on $[n]$, i.e., $\pi_s(i) = i + 1 \mod n$. We consider the action by conjugation of $\text{Sym}(n)$ over the set of $n \times n$ matrices, which is defined by $(\Pi, A) \mapsto \Pi A \Pi^T$. If $S \subset \text{Sym}(n)$ 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 $\text{Dih}_n$.

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 $[n] \to \mathcal{X}$ with inverse $\# : \mathcal{X} \to [n]$.

The notion of an 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 $x_0, \ldots, x_{N-1}$ are linearly ordered if $x_i \leq x_{i+1}$ for $i \in [N]$. A cyclic order on $\mathcal{X}$ is a relation $\mathcal{C}$ on $\mathcal{X}^3$ that is cyclic, antisymmetric, transitive and total. The pair $(\mathcal{X}, \mathcal{C})$ 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 $\leq_{\mathcal{C}, x_0}$ as $x \leq_{\mathcal{C}, x_0} y$ if and only if $(x_0, x, y) \in \mathcal{C}$. Finally, we say $x_0, \ldots, x_{N-1} \in \mathcal{X}$ are cyclically ordered if $x_i \leq_{\mathcal{C}, x_0} x_{i+1}$ 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 $\text{Sym}(n)$ by conjugation. The abstract seriation problem can be stated as [26]

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

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

We study two questions about this problem: for which class $\mathcal{M}$ can we ensure a solution exists? and, given this class, is there an efficient algorithm to solve the seriation problem for any $A$? In this work we focus on the case where the class $\mathcal{M}$ is induced by a dissimilarity on a finite set $\mathcal{X}$. Our goal is to provide an answer when this dissimilarity may induce a linear or cyclic order on this set. 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 Section 5.

3.1. Robinson dissimilarities. Most of the work on the seriation problem has focused on matrix dissimilarities. In this work, we primarily focus on an equivalent formulation based on abstract pairwise dissimilarities. This allows for a more transparent presentation of algorithms, as well as a more natural extension to the case of infinite continuous sets. Later, we provide a formal connection between the two concepts.

A dissimilarity or premetric $d : \mathcal{X}^2 \to \mathbb{R}$ 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 $d$ on $\mathcal{X}$ is linear Robinson if there exists a linear order $\leq_d$ on $\mathcal{X}$ such that

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

We denote the set of all solutions by $\mathcal{S}_\mathcal{M}(A)$.
It is strictly linear Robinson if all the inequalities are strict. We say $\leq_d$ is consistent with $d$ and that $d$ is linear Robinson with respect to $\leq_d$.

Linear Robinson dissimilarities preserve the intervals defined by any consistent order [19]. From Definition 3.1 it follows that for any $r > 0$ and $x \in X$ the (closed) balls $B^d_r(x) \triangleq \{y \in X : d(x, y) \leq r\}$ are intervals in $(X, \leq_d)$. In fact, this property uniquely characterizes linear Robinson dissimilarities. To prove this converse, the appropriate structure to analyze is the hypergraph $H^d$ with vertex set $X$ and hyperedge set $B^d \triangleq \{B^d_r(x) : x \in X, r > 0\}$. This hypergraph is called an interval hypergraph if every hyperedge is an interval [14].

**Proposition 3.2 ([19]).** Let $d$ be a dissimilarity on $X$. The following are equivalent:
1. $d$ is linear Robinson.
2. The hypergraph $H^d$ is an interval hypergraph.

It is easy to see that orderings consistent with a given linear Robinson dissimilarity are never unique. This follows from the natural symmetries of Robinson dissimilarities. Let $\leq_d$ be consistent with respect to $d$. Its reversal $\leq'_d$ is the linear order defined by $x \leq'_d y$ if and only if $y \leq_d x$. It is clear that $d$ is linear Robinson with respect to $\leq'_d$ if and only if it is so with respect to $\leq_d$. Hence, the reversal of an ordering consistent with a dissimilarity is always consistent with the dissimilarity. Furthermore, there could be other consistent orderings in the case the dissimilarity is not strict, as we will see in Section 4.

**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 $d$ on $X$ is circular Robinson if there exists a cyclic order $C^d$ such that
$$\forall \text{ cyclically ordered } w, x, y, z \in X : \quad d(y, w) \geq \min\{d(y, x), d(y, z)\}.$$ 

We say it is strict circular Robinson if the inequality is strict. We say $C^d$ is consistent with $d$ and that $d$ is linear Robinson with respect to $C^d$.

Circular Robinson dissimilarities preserve the arcs of any compatible order, i.e., sets of the form $\{x \in X : (m, x, M) \in C^d\}$ for $m, M \in X$ called the borders of the arc. Arcs are the natural analogues of intervals for a cyclic order. Consequently, we say $H^d$ 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 $d$ be a dissimilarity. The following are equivalent:
1. $d$ is circular Robinson.
2. The hypergraph $H^d$ is an arc hypergraph.

Similarly to the linear case, consistent orderings in the circular case are never unique. Let $C^d$ be consistent with respect to $d$. In this case, its reversal $C''^d$ is the cyclic order such that $(x, y, z) \in C''^d$ if and only if $(z, y, x) \in C^d$. By definition, $d$ is circular Robinson with respect to $C''^d$ if and only if it is so with respect to $C''^d$.

**3.2. Robinson matrices.** Let $d$ be a dissimilarity on $X$. To any enumeration $x : [n] \to X$ we can associate the $n \times n$ dissimilarity matrix $D$ with entries $D(i, j) := d(x_i, x_j)$. It is always non-negative, symmetric, and with zero-diagonal. However, some enumerations will endow $D$ with additional properties. This leads us to the concept of Robinson matrices.
3.2.1. Linear Robinson matrices. If $d$ is consistent with respect to $\leq_d$ there exists an enumeration of $X$ such that for $i, j \in [n]$ we have $i \leq j$ if and only if $x_i \leq_d x_j$. 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

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

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 $X = [n]$.

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 linear seriation and 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 $\Pi, D\Pi^T$ is linear Robinson. Since $\text{Dih}_1 \cong \langle \pi_r \rangle$ then linear Robinson matrices are invariant under the action of $\text{Dih}_1$ by conjugation.

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

\[ (i, j, k) \in \mathcal{C}_n \iff (i < j < k) \lor (j < k < i) \lor (k < i < j). \]

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

If $d$ is consistent with respect to $\mathcal{C}_d$ there exists an enumeration of $X$ such that $(i, j, k) \in \mathcal{C}_n$ if and only if $(x_i, x_j, x_k) \in \mathcal{C}_d$. 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

\[ \forall \text{ cyclically ordered } i, j, k, \ell \in [n] : D(i, k) \geq \min\{D(j, k), D(k, \ell)\}. \]
Therefore, $D$ is consistent w.r.t. $\mathcal{C}_n$ 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 is denoted by $\mathcal{C}_R$ and $\mathcal{C}_R^s$, 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

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

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 $m_1 \leq m_2$ with $f_{m_1} = f_{m_2}$ and

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

Notice that when $m_1 < m_2$ there are two modes, whereas when $m_1 = m_2$ there is only one mode. Our definition above allows us to treat both cases simultaneously. From the definition it is clear that every subsequence of a strictly unimodal sequence is also strictly unimodal. The proof of the following result 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 $j \to D(i, i + j \mod n)$ 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 $\Pi_r D^T \Pi_s$ and $\Pi_s D^T \Pi_s$ are circular Robinson matrices.

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

Since $\text{Dih}_n \cong \langle \pi_r, \pi_s \rangle$ it follows that circular Robinson matrices are invariant under the action of $\text{Dih}_n$ 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 ensure they will never be a singleton. We call Robinson orderings to 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 $\text{Dih}_1$ the trivial solutions whereas those in different orbits non-trivial solutions. The same criteria applies for the circular seriation problem when the action of $\text{Dih}_n$ 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 an $m \times n$ binary matrix $M$. The C1 problem is to find a permutation $\Pi$ such that the entries of $M\Pi$ equal to one appear consecutively along rows. We say $M$ has the consecutive ones (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 $O(m + n + f)$ time.

An extension to this problem is the circular ones (Cr1) problem. The Cr1 problem is to find a permutation $\Pi$ 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 and 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 ball hypergraph $H_D$, introduced in Subsection 3.1. 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 $O(n^3)$ time and with $O(n^3)$ space.

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

4.2. $PQ$-Trees. The algorithmic structure underlying the algorithm to solve the C1 problem is the $PQ$-tree. A $PQ$-tree $T$ 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 $Y = \{Y_i\}_{i \geq 0}$ 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 $Y_1 = \{Y_i\}$ and determines the set of admissible permutations such that $Y_i$ becomes an interval. These can be represented by a PQ-tree $T_1$ (see [3] and [4]). The algorithm proceeds by adding a $Y_i$ to form $Y_2 = \{Y_i, Y_{i+1}\}$ and update the PQ-tree accordingly. The main contribution of [4] is an algorithm for updating $T_k$ in a way that given any subset $Y_k \subseteq [n]$, the set of permutations represented by the updated tree $T_{k+1}$ is precisely the set of admissible permutations of $Y_k$. This is done in time linear in the size of $Y_k$. The algorithm finishes when $Y$ 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. We note in passing that our algorithm works as well for the strict linear case, but we omit this variant. Our algorithm runs in $O(n^2)$ 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. The core algorithm relies on 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 delicate part of the algorithm consists in efficiently deciding whether 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. We will indistinctly refer to this operation as the Q-node being fixed or oriented. 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 of comparisons

---

2Given the strict Robinson property, it is clear that the underlying matrix is dense, and therefore $O(n^2)$ memory is required to even provide the input.
to decide each orientation is bounded by \( \mathcal{O}(n) \), which leads to a running time \( \mathcal{O}(n^2) \).

We provide preliminary results in Subsection 5.1 and Subsection 5.2 about the arc structure of the nearest-neighbour graph, and the treatment of arc reversals. Then, in Subsection 5.3 we develop our algorithm, whose correctness and optimality are proved in Subsection 5.4.

5.1. Preliminaries part I: nearest neighbours graph in strict Robinson dissimilarities.

One of our algorithmic building blocks is based on the idea that in the case of strict dissimilarities, a pair of nearest neighbours must lie consecutively in any Robinson ordering. In this subsection, we prove this fact and use it to justify the first step of our algorithm, which is based on merging nearest neighbours to reduce the size of the instance.

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 \( \mathcal{X} \), and every set \( \mathcal{I}_i \) is an arc of consecutive elements in any Robinson ordering. The set of nearest neighbours of \( x \in \mathcal{X} \) is defined as \( \text{NN}(x) \triangleq \arg \min_{y \in \mathcal{X} \setminus \{x\}} d(x, y) \).

The nearest-neighbours graph is an undirected graph \( G_{\text{NN}}(\mathcal{X}, d) = (\mathcal{X}, \mathcal{E}) \) such that \( \{x, y\} \in \mathcal{E} \) iff \( x \in \text{NN}(y) \) or \( y \in \text{NN}(x) \). An essential condition of strict Robinson dissimilarities is what we call 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 \( \text{NN}(i) \subseteq V^C_i \triangleq \{ i-1 \mod n, i+1 \mod n \} \).

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 \( N_G(x) \triangleq \{ y \in \mathcal{X} : \{x, y\} \in \mathcal{E} \} \). The function \( x \mapsto N_G(x) \) is called the neighbourhood. The cycle graph \( C_n = ([n], \mathcal{E}) \) is the graph with edge set \( \mathcal{E} = \{(i, (i+1) \mod n) : i \in [n]\} \). If a graph \( G \) is a subgraph of \( C_n \), then it is clear that its connected components correspond to arcs of \( ([n], \mathcal{E}_n) \). A direct consequence of the nearest-neighbour condition is that the nearest-neighbours graphs of strict circular Robinson dissimilarity matrices correspond to subgraphs of \( C_n \).

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

**Proof.** Let \( \sigma \) be any Robinson ordering. Let \( j \in \text{NN}(i) \) 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 \( \sigma \), 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.

---

3Given some enumeration \# and \( i \in [n] \), when we write \( \text{NN}(i) \), we refer to the set \#(\text{NN}(x_i)) \).
The fact that this graph is a subgraph of the cycle 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 Depth-First Search (DFS) (Algorithm B.2) starting at each non visited degree 1 node. The order of visits will follow the Robinson ordering (or backwards).

If there are no degree one nodes, then $G_{\text{NN}} \cong C_n$ 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 $\alpha$. Each connected component will be stored in a tuple $\alpha$, where $\alpha(j)$ is the $j$-th element visited by performing a DFS. The procedure is summarized in the procedure $\text{Arc Partition}$ (Algorithm B.1), whose correctness is stated in the following Proposition (the proof of the next result is omitted for brevity).

**Proposition 5.3.** Given a matrix $D \in \text{pre-}C^*_R$, by performing $\text{Arc Partition}$ (Algorithm B.1) 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 tells 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 its children directly to the parent $Q$-node. We call this process orientation. In this section we provide computationally efficient conditions to determine when a $Q$-node must be oriented. Each $Q$-node $\alpha$ in a tree $T$ can be associated with an arc $I_\alpha$ in $X$: the arc of all leaves in $X$ which are descendants of $\alpha$. Reversing $\alpha$ corresponds to reversing $I_\alpha$. 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 $I$ and $J$ are said to strictly overlap, denoted by $I \not\prec_J$, if

1. $I \not\subset J$; 2. $J \not\subset I$; 3. $I^c \not\subset J$; and 4. $J^c \not\subset I$.

**Observation 5.5.** The relation $\not\prec_J$ is symmetric and equivalent to

1. $I \cap J^c \neq \emptyset$; 2. $J \cap I^c \neq \emptyset$; 3. $I^c \cap J^c \neq \emptyset$; and 4. $I \cap J \neq \emptyset$.

**Lemma 5.6.** Let $I$ and $J$ be two arcs. Let $a, b$ and $a', b'$ be the borders of $I$ and $I^c$, respectively, where $a$ and $a'$ ($b$ and $b'$) are consecutive in the cyclic order. Then $I \not\prec_J$ if and only if one of the following conditions holds

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

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

---

4 We use the symbol $\cong$ to either denote graph and group isomorphism
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 \(\mathcal{I}^c\), \(b' \notin \mathcal{J}\). 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 \textit{permutation that reverses} \(\mathcal{I}\) as the permutation \(\sigma\) s.t. \(\sigma(a_j) = a_{k-j-1}\) for \(j \in \{0, \ldots, k-1\}\), and \(\sigma(x) = x\) if \(x \notin \mathcal{I}\).

\textbf{Lemma 5.7.} Let \(\mathcal{I}\) and \(\mathcal{J}\) be two arcs and let \(\sigma\) be the permutation that reverses the elements of \(\mathcal{I}\). Then \(\mathcal{I} \not\preceq_{\sigma} \mathcal{J}\) iff the permutation of \(\mathcal{J}\) by \(\sigma\) is not an arc.

\textbf{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 \(\mathcal{I} \not\preceq_{\sigma} \mathcal{J}\), 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 \(\mathcal{I}\), the arc \(\mathcal{J}\) gets ripped apart into two disconnected pieces: namely its red nodes and its purple nodes. 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 produce 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 \(\alpha\) 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 \( I \) one could equivalently check whether there exists \( z \in X \) and \( r > 0 \) such that

\[
\left[ \{a, a'\} \subset B_r(z) \land \{b, b'\} \subset B_r(z)^c \right] \lor \left[ \{b, b'\} \subset B_r(z) \land \{a, a'\} \subset B_r(z)^c \right],
\]

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 \( I \) (resp. \( I^c \)) in advance. Border Candidates Orientation (Algorithm 5.1) is an efficient way for orienting the arc \( I \) with respect to the dissimilarity \( d \) 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 \((A', A, B, B')\) are said to be border candidates of the arc \( I \), if the following properties hold:

1. \( A, B \subset I \) and \( A', B' \subset I^c \).
2. The sets are pairwise disjoint.
3. If \( a, b \) are the borders of \( I \) and \( a', b' \) are the borders of \( I^c \), then: \( a \in A, b \in B, a' \in A' \) and \( b' \in B' \).
4. Either \((A', A, B, B')\) or \((A', B, A, B')\) is cyclically ordered.\(^5\)

The next result provides correctness for Border Candidates Orientation (Algorithm 5.1) (see its proof in Appendix A.2).

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

1. Both \( \{a, a'\} \subset B_r(z) \) and \( \{b, b'\} \subset B_r(z)^c \) hold for some \( r > 0 \).
2. There exist \((x', x, y, y') \in A' \times A \times B \times B'\) s.t. \( \max\{f_z(x), f_z(x')\} < \min\{f_z(y), f_z(y')\} \).
3. It holds that \( \max\{\min f_z(A), \min f_z(A')\} < \min\{\max f_z(B), \max f_z(B')\} \).

Above \( f_z(\cdot) \triangleq d(z, \cdot) \) and given any \( U \subset X \), \( \min f_z(U) \triangleq \min_{y \in U} f_z(y) \) (similar for max).

**Corollary 5.10.** Let \( I \subset X \) be an arc in any Robinson ordering and suppose the sets \((A', A, B, B')\) are border candidates for the arc \( I \). Then, Algorithm 5.1 correctly determines if \( I \) must be fixed, reversed or if it is not orientable.

**Observation 5.11.** The time complexity of Algorithm 5.1 with input \((A', A, B, B')\) is \( O(|X| \cdot \max\{|A'|, |A|, |B|, |B'|\}) \), thus it is an efficient way of orienting a Q-node \( \alpha \) whenever the sets of border candidates for \( I_\alpha \) is not too big.

### 5.3. The recursive seriation algorithm.

We devote the next subsection to describing our main algorithm: Recursive Seriation (Algorithm 5.2). This algorithm starts from the singleton elements of \( X \), and proceeds by recursively detecting nearest neighbors among previously computed arcs, merging them, and deciding whether the merged sets must be oriented. An important ingredient of our algorithm is the use of Q-trees: this is a useful data structure to merge arcs, as well as fixing the orientations when they are detected. In order to run in total quadratic time, great care is needed regarding the number of comparisons used to decide the orientations, for which we find useful to compute border candidates for Q-trees.

\(^5\)Formally, the ordered collection is a consistent cyclic quasi order, see Definition 5.12
Algorithm 5.1 Border Candidates Orientation

1: Input: A sequence of sets \((A', A, B, B')\)
2: Let \(f_z(x) \triangleq d(z, x)\) for every \(z \in X\)
3: Let \(O_i : X \rightarrow \{True, False\}\) for \(i = 1, 2, 3, 4\) be defined by
4: \(O_1(z) \triangleq \max \{\min f_z(A), \min f_z(A')\} < \min \{\max f_z(B), \max f_z(B')\}\)
5: \(O_2(z) \triangleq \max \{\min f_z(B), \min f_z(B')\} < \min \{\max f_z(A), \max f_z(A')\}\)
6: \(O_3(z) \triangleq \max \{\min f_z(A), \min f_z(B')\} < \min \{\max f_z(A'), \max f_z(B)\}\)
7: \(O_4(z) \triangleq \max \{\min f_z(B), \min f_z(A')\} < \min \{\max f_z(B'), \max f_z(A)\}\)
8: for \(z \in X\) do
9: if \(O_1(z) \lor O_2(z)\) then
10: return ‘correct’
11: else if \(O_3(z) \lor O_4(z)\) then
12: return ‘reverse’
13: end if
14: end for
15: return ‘not orientable’
16: Output: A string determining the orientation of the input

For the sake of readability, we dissect the main algorithm in terms of various subroutines, which are presented and analyzed separately. Hence, we structure our presentation as follows. In Subsection 5.3.1 we specify the initialization of Recursive Seriation. In Subsection 5.3.2, we describe the key operation of computing the border candidates of a \(Q\)-tree, summarized in the procedure Border Candidates (Algorithm 5.3). Next, in Subsection 5.3.3 we describe the operation of computing dissimilarities among \(Q\)-trees, and how this leads to the recursion behind Recursive Seriation. In Subsection 5.3.4 we study the process of deciding the orientation among consecutive children of a given \(Q\)-node, including a pseudocode of this procedure Consecutive Orientation (Algorithm 5.4). Next, in Subsection 5.3.5 we use the previous subroutine and complement it with the additional steps required to completely orient the internal nodes of a \(Q\)-tree; the associated subroutines here are in the procedures Complete Internal Orientation (Algorithm 5.5), and Final Orientation (Algorithm 5.6). In Subsection 5.3.6 we provide a method to decide the orientation between trees that have been detected to be nearest neighbours; the useful subroutine here is the External Orientation (Algorithm 5.7).

5.3.1. Initialization. Recursive Seriation receives as input a family \(T\) of \(Q\)-trees (which are \(PQ\)-trees composed solely by \(Q\)-nodes), with leaves corresponding to elements of \(X\). Given an instance \((X, d)\), we initialize the algorithm with \((X, d, T)\), where \(T = \{x : x \in X\}\). For each \(T \in T\) we write as \(\partial T\) the set of leaves of \(T\). Hence, in the initial case we get \(\partial x \triangleq \{x\}\).

5.3.2. Computing border candidates. We endow each \(T \in T\) with a set \(B(T)\) of border candidates\(^6\), which are all leaves of \(T\) that appear in the extreme left or right under some configuration of the tree. Whenever \(|\partial T| \geq 2\), and conditionally on an given orientation of the root, the set of border candidates \(B(T)\) can be split in two: left and right. The set of left border candidates, denoted as \(B^L(T)\), are all elements in \(B(T)\) that appear in the extreme

\(^6\)We emphasize the distinction of the border candidates of a tree, which we are about to introduce; and the border candidates of an arc, introduced in Definition 5.8.
Algorithm 5.2 Recursive Seriation

1: **Input**: $\{X, d, T\}$: $d$ dissimilarity over $X$; $T$ family of $Q$-trees with leaves given by elements of $X$
2: for $T \in T$ do
3: for $T' \in T$ do
4: $d_{\text{min}}(T, T'), d_{\text{arg min}}(T, T') = \text{Tree Dissimilarity}(T, T', d)$
5: $\text{External Orientation}(T, T', d_{\text{arg min}}(T, T'))$
6: end for
7: end for
8: $T' = \text{Arc Partition}(d_{\text{min}}, T)$
9: if $|T'| = 1$ then
10: $\text{Final Orientation}(T)$ {where $T' = \{T\}$}
11: return $T$
12: else
13: for $T \in T'$ do
14: $\text{Complete Internal Orientation}(T)$
15: end for
16: return Recursive Seriation($X, d, T'$)
17: end if
18: **Output**: A $Q$-tree $T$ containing all Robinson orderings

left under some configuration of the tree. The definition of $B^R(T)$ is entirely analogous with appearing on the right. For instance, in the tree $T$ appearing in Figure 5a, we have $B^L(T) = \{a_3, b_3, b_2, b_1\}$ and $B^R(T) = \{b_0\}$.

Let $\alpha$ denote the root of $T$. Whenever $\text{depth}(T) > 1$, where $\text{depth}(T)$ denotes the tree-depth of $T$, $T^L$ (resp. $T^R$) denotes the subtree of $T$ whose root is the first (resp. last) $Q$-node among the direct descendants of $\alpha$. Notice that for the computation of the border candidates of $T$ it is convenient to consider the relations: $B^L(T) = B(T^L)$ and $B^R(T) = B(T^R)$. Thus, to obtain $B(T) = B(T^L) \cup B(T^R)$ it suffices to recursively call for the border candidates of the subtrees $T^L$ and $T^R$. Border Candidates (Algorithm 5.3) is an straightforward implementation of this idea that runs in time $O(|B(T)|)$.

Algorithm 5.3 Border Candidates

1: **Input**: A $Q$-tree $T$
2: if $\text{depth}(T) = 0$ then
3: return $T$ {Return the single element from the tree}
4: else
5: return Border Candidates($T^L$) $\cup$ Border Candidates($T^R$)
6: end if
7: **Output**: The set $B(T)$ of border candidates of $T$

5.3.3. Computing the minimum pairwise dissimilarity among trees. A second key step is to define an appropriate dissimilarity between trees. This allows us to solve the problem recursively by decreasing the number of objects we need to sort in each iteration. In the initial case we have that this dissimilarity corresponds to $d$, and since it is strict circular Robinson, Arc Partition (available in Appendix B) with input $(T, d)$ returns an arc partition of $T$ stored
in tuples. The elements in each tuple \( \alpha \) must be consecutive. Therefore, for each \( \alpha \) we build a new tree \( T_\alpha \) with a \( Q \)-node in the root whose \( i \)-th child corresponds to \( \alpha(i) \). The recursion works by repeating the process over the smaller family of trees \( T' = \{ T_\alpha \} \), until \( |T'| = 1 \).

For this to work we need to define a dissimilarity \( d' \) over \( T' \) in a way that sorting \( T' \) with respect to \( d' \), help us in our goal of sorting \( T \) with respect to \( d \). We now present a dissimilarity that does exactly that: given two trees \( T_1, T_2 \), consider the dissimilarity \( d_{\text{min}}(T_1, T_2) \equiv \min \{ d(x, y) : x \in \partial T_1, y \in \partial T_2 \} \). Also let \( d'_{\text{arg min}}(T_1, T_2) \) be the collection of all minimizers of this problem in \( \partial T_1 \times \partial T_2 \). The following Lemma justifies this choice, as it ensures that by sorting \( T' \) we obtain a quasi-order among the elements of our original set \( T \).

**Definition 5.12 (Quasi-order\(^7\)).** Let \((\mathcal{X}, \mathcal{C})\) be a cyclically ordered set. An ordered partition \( \{A_0, \ldots, A_{m-1}\} \) is a (consistent) cyclic quasi order if for all \((i, j) \in \mathcal{C}_m\), \( x \in A_i \), \( y \in A_j \) and \( z \in A_k \) we have that \((x, y, z) \in \mathcal{C}\).

**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 \( D_{\text{min}}(A_i, A_j) \equiv \min \{D(k, l) : k \in A_i, l \in A_j\} \) is a (strict) circular Robinson dissimilarity.

Since nearest neighbours must be consecutive, the family of \( Q \)-trees \( T \) at each recursion is guaranteed to satisfy that \( \{\partial T\}_{T \in T} \) is an arc partition. Therefore, by the previous Lemma we have that \( D'(i, j) \equiv d_{\text{min}}(T_i, T_j) \in \text{pre-C}_R \) (pre-\( C_R \)) whenever the original dissimilarity matrix \( D \) is in \( \text{pre-C}_R \) (pre-\( C_R \)) and a Robinson ordering for \( D' \) yields a quasi order for \( D \).

A naïve computation of \( d_{\text{min}} \) does not lead to a global \( \mathcal{O}(n^2) \) time complexity. This is why our next goal is to find a workaround. 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 considering border candidates of each tree. In particular, \( d'_{\text{arg min}}(T_1, T_2) \subseteq \mathcal{B}(T_1) \times \mathcal{B}(T_2) \).

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

**Observation 5.15.** Let Tree Dissimilarity be the procedure which receives a pair \((T_1, T_2)\) of \( Q \)-trees and a dissimilarity \( d \) to return \( d_{\text{min}}(T_1, T_2) \) and \( d'_{\text{arg min}}(T_1, T_2) \) by brute force comparisons among border candidates. It takes \( \mathcal{O}(\max(|B(T_1)|, |B(T_2)|)^2) \) operations and \( \mathcal{O}(1) \) space, since by Lemma 5.14, \( |d'_{\text{arg min}}(T_1, T_2)| \leq 4 \). We omit its pseudocode for brevity.

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 \( T_2 \), it suffices to find border candidates for the arc \( \partial T_2 \). Let \( T_1, T_2, T_3 \) be children of a \( Q \)-node \( \alpha \), such that \( T_2 \) succeeds \( T_1 \) and precedes \( T_3 \), then we decide their orientation with the Consecutive Orientation procedure (Algorithm 5.4).

Since \( A' \equiv B(T_1), B' \equiv B(T_3), A \equiv B^L(T_2) \) and \( B \equiv B^R(T_2) \) are border candidates for the arc \( \partial T_2 \), the correctness of the procedure is due to the correctness of Algorithm 5.1. By Observation 5.11, the complexity is given by \( \mathcal{O}(n \cdot \max\{|B(T_1)|, |B(T_2)|, |B(T_3)|\}) \).

In a computational implementation of the algorithm, \( Q \)-nodes can be represented by tuples, whereas \( Q \)-trees are simply nested tuples. For clarity, we present an example of the \( Q \)-node fixing procedure mentioned in Algorithm 5.4.

\(^7\)This extends the definition introduced in [6] for linear orders to cyclic orders.
Algorithm 5.4 Consecutive Orientation

1: Input: Three consecutive subtrees \(T_1, T_2, T_3\) of a \(Q\)-node \(\alpha\)
2: Let \(A' \triangleq B(T_1), B' \triangleq B(T_3), A \triangleq B^L(T_2)\) and \(B \triangleq B^R(T_2)\).
3: \(x = \text{Border Candidates Orientation}(A', A, B, B')\)
4: if \(x = \text{correct}\) then
5: Fix the root of \(T_2\) in \(\alpha\)
6: else if \(x = \text{reverse}\) then
7: Reverse the root of \(T_2\), then fix it in \(\alpha\)
8: else
9: Label \(T_2\)'s root as non-orientable and continue the algorithm as if the root of \(T_2\) were fixed.
   This node is an actual \(Q\)-node of the tree of Robinson orderings.
10: end if
11: Result: The direct children of the root of \(T_2\) had been directly connected to \(\alpha\)

In this example we consider \(Q\)-node \(\alpha = ((0, 1, 2), (6, 5, (3, 4)), (7, (8, 9, 10)))\) taking values over \(X = \{11\}\). There are three \(Q\)-trees, which are consecutive in \(\alpha\), which are \(T_1 = (0, 1, 2)\), \(T_2 = (6, 5, (3, 4))\) and \(T_3 = (7, (8, 9, 10))\). An example of a possible outcome of Consecutive Orientation \((T_1, T_2, T_3)\) would be that the root of \(T_2\) is reversed and fixed into \(\alpha\).
In such case \(\alpha\) is modified into \((0, 1, 2), (3, 4), 5, 6, (7, (8, 9, 10)))\).

5.3.5. Complete internal orientation of a connected component and final orientation.
A complete internal orientation of a \(Q\)-tree \(T\), is a process in which we determine the orientation of all nodes present in \(T\) except from those present in \(T^L\) and \(T^R\). For this task we propose the procedure Complete Internal Orientation (Algorithm 5.5).

Algorithm 5.5 Complete Internal Orientation

1: Input: A \(Q\)-tree \(T\) with root \(\alpha\) and depth(\(T\)) \(> 1\)
2: For \(i \in [k]\), let \(\alpha(i)\) be the \(i\)-th children of \(\alpha\) and let \(T_i\) be the subtree whose root is \(\alpha(i)\)
3: {Notice that \(T_0 = T^L\) and \(T_{k−1} = T^R\)}
4: while \(\exists 1 \leq j \leq k − 2\) such that the node \(\alpha(j)\) has not been fixed into \(\alpha\) do
5: for \(i = 1...k − 2\) do
6: Consecutive Orientation \((T_{i−1}, T_i, T_{i−1})\)
7: end for
8: end while
9: Result: All elements in \(\partial T \setminus (\partial T^L \cup \partial T^R)\) are directly connected to \(\alpha\).

The orientation over all the nodes is done in a breadth-first search fashion\(^8\). Once again, the correctness of the procedure is due to the correctness of Algorithm 5.1. 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. \(T^L\) corresponds to the tree with root in \(Q_1\) and \(T^R\) corresponds to the tree with root in \(Q_k\). The tree in the bottom corresponds to the tree after Algorithm 5.5.

In the final recursion of Recursive Seriation we obtain a unique connected component from Arc Partition, from which we construct a unique tree \(T\) such that \(\partial T = X\). Here, the

\(^{8}\) This way, trees of same depth are compared in Algorithm 5.4 (excluding comparisons with \(T^L\) and \(T^R\)).
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 5.6). This is equivalent to consider the $Q$-node as a ring rather than as a list.

\begin{algorithm}[H]
\caption{Final Orientation}
\begin{algorithmic}[1]
\STATE \textbf{Input:} A $Q$-tree $\mathcal{T}$ with root $\alpha$.
\STATE Let $\alpha(i)$ be the $i$-th children of $\alpha$ and let $\mathcal{T}_i$ be the subtree whose root is $\alpha(i)$.
\WHILE{$\exists j \in [||\alpha||]$ such that the node $\alpha(j)$ has not been fixed into $\alpha$}
\IF{$||\alpha|| > 2$}
\STATE Let $\mathcal{T}_{i-1} \triangleq \mathcal{T}^R$ and $\mathcal{T}_{i+1} \triangleq \mathcal{T}^L$.
\FOR{$i \in [||\alpha||]$}
\STATE Consecutive Orientation ($\mathcal{T}_{i-1}, \mathcal{T}_i, \mathcal{T}_{i+1}$)
\ENDFOR
\ELSE\IF{$||\alpha|| = 2$}
\STATE Consecutive Orientation ($\mathcal{T}^R_1, \mathcal{T}_0, \mathcal{T}^L_1$)
\STATE Fix $\alpha(1)$ into $\alpha$.
\ENDIF
\ENDIF
\ENDWHILE
\STATE \textbf{Result:} All $Q$-nodes in $\mathcal{T}$ are oriented.
\end{algorithmic}
\end{algorithm}

5.3.6. External orientation of trees. Since for each $\mathcal{T} \in \mathcal{T}$, the set $\partial \mathcal{T}$ corresponds to an arc and as a consequence of Lemma 5.14, $d^\text{min}(\mathcal{T}, \mathcal{T}')$ is attained at some $x \in B(\mathcal{T})$ and $y \in B(\mathcal{T}')$ 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. With this purpose in mind, we propose the procedure External Orientation (Algorithm 5.7).

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 $x \in B^L(\mathcal{T})$). In the second part, we execute Complete Internal Orientation (Algorithm 5.5) 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

![Figure 4: Example of a connected component of the nearest neighbours graph at the second recursion of Recursive Seriation before and after Complete Internal Orientation.](image-url)
Algorithm 5.7 External Orientation

1: **Input:** $Q$-trees $T$ and $T'$. The set $d^{\text{arg min}}(T, T')$
2: **for** $(x, y) \in d^{\text{arg min}}(T, T')$ **do**
3:  **if** $x \in B^L(T)$ **then**
4:     {Fix every $Q$-node in $T^L$ containing $x$ as a descendant from the root until the $Q$-node $\alpha$ where $x$ lies in a way such that $x$ is placed on the left}
5:  **J** = $T^L$
6:  **while** depth($J$) > 0 **do**
7:      **if** $x \in B^L(J)$ **then**
8:         Fix the root of $J$ into its parent
9:      **else**
10:         Reverse and then fix the root of $J$ into its parent
11:   **end if**
12:   $J \leftarrow J^L$
13: **end while**
14: **else**
15:   Proceed analogously, with $B^R(T)$, instead of $B^L(T)$; and placing $x$ at the right, instead of the left
16: **end if**
17: **end for**
18: Repeat the same procedure with $T'$ and $y$
19: **Result:** Either all $Q$-nodes in $T^L$ or $T^R$ (resp. $T'^L$ or $T'^R$) are oriented

the subtree with root in $Q_1$ and $T^R$ is the singleton $\{b_0\}$. Suppose by computing $d^{\text{min}}(T, T')$ for some other $T'$ we get that $d^{\text{min}}$ is attained at $b_2 \in B^L(T)$. In that case, we must fix $T^L$ following the algorithm. Since $b_2$ is 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 5.5, the number of operations required for this procedure is bounded by $O(\text{depth}(T^L))$.

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

5.4. Analysis of the recursive seriation algorithm.

**Theorem 5.17.** Given $D \in \text{pre-}\mathcal{C}^*_R$, let $T$ be the $PQ$-tree obtained from Algorithm 5.2 with input $D$. Let $S(T)$ the set of all orderings of $X$ (permutations) represented by the tree. Then, $S^c_R(D) = \text{Dih}_n \circ S(T)$, i.e. it solves the strict circular seriation problem.

**Proof Sketch.** For simplicity, suppose we omit the orientation steps in Algorithm 5.2 and leave them to the end of the process. This does not affect the set of solutions but may
increase the time complexity. Denote \( T^{pre} \) and \( T \) the trees before and after orientation, respectively. Also let \( T_k \) the family of trees instantiated at the \( k \)-th recursive step. Notice that by Lemma 5.13, evaluating \( d^{\text{min}} \) over \( T_k \) yields a dissimilarity matrix \( D_k \in \text{pre-}C^*_{R} \). Due to Proposition 5.3, we have that \( S_{\text{C}}(D) \subset \text{Dih}_n \circ \text{S}(T^{pre}) \) (at least all Robinson orderings are considered at this point). To complete the proof, it remains to show that in \( T \) all orientable \( Q \)-nodes originally in \( T^{pre} \) had been correctly fixed. To see this notice that the orientation of each \( Q \)-node in \( T^{pre} \) is tested either by External Orientation (Algorithm 5.7) or Consecutive Orientation (Algorithm 5.4). The correctness of External Orientation is due to Lemma 5.14. The correctness of Consecutive Orientation is due to Corollary 5.10.

Theorem 5.18. Recursive Seriation runs in \( O(n^2) \) time.

Proof. We count the number of operations required by the procedure Complete Internal Orientation and Final Orientation separately from the rest. At the \( i \)-th recursion let \( T(i) \) be the input \( Q \)-trees, let \( k(i) \triangleq |T(i)| \) and let \( b(i) = \max_{T \in T(i)} |B(T)| \). Then, by Observation 5.15, computing \( d^{\text{min}} \), takes \( O(k(i)^2 \cdot b(i)^2) \) operations. By Observation 5.16, the complexity of the procedure External Orientation takes \( O(k(i)^2) \) operations. Computing \( G_{\text{NN}} \) takes \( O(k(i)^2) \) operations. The procedure Depth-First Search takes \( O(k(i)) \) operations.

On the other hand, notice that in each step of the recursion, every tree is merged to its nearest neighbour. This implies that \( k(i) \leq \frac{3n}{2} \) 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 total number of operations of this procedure is bounded by \( C_1 \sum_{i=0}^{\log_2(n)} \left( \frac{3n}{2} \right)^2 (i+1)^2 + \left( \frac{3n}{2} \right)^2 + \left( \frac{3n}{2} \right)^2 = O(n^2) \).

It remains to consider Complete Internal Orientation and Final Orientation. In this procedures, all \( Q \)-nodes \( \alpha \) are oriented through Consecutive Orientation (Algorithm 5.4) with input \((T_1, T_2, T_3)\) where \( \alpha \) is the root of \( T_2 \). To count the operations of this procedure we consider two cases. The first (and most common) case is when \( T_1, T_2 \) and \( T_3 \) are trees instantiated at the same recursive step. In this case, if they were 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 internal orientation of a connected component. Let \( T \) be a tree generated from a connected component of the nearest-neighbours graph at the \( i \)-th recursion of the algorithm. Then, in Complete Internal Orientation (or Final Orientation) with input \( T \), some of the internal \( Q \)-nodes will be oriented by having as border candidates \( B(T^L) \) and \( B(T^R) \). Since depth(\( T \)) \( \leq i \), this can occur for \( B(T^L) \) (resp. \( B(T^R) \)) for at most \( i \) internal \( Q \)-nodes of \( T \). Let \( C(i) \) 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, \( |B(T^L)| \leq i \) and \( |B(T^R)| \leq i \), the number of operations required for orienting all these 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. \( \blacksquare \)

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 \( S_{L}^*(D) = \{e, r\} \cong \text{Dih}_1 \). However, it is not immediately clear which permutations are the ones that preserve strict unimodality. The next Lemma will let us conclude that there is at most one non trivial ordering for \( D \in \mathcal{L}_R^* \).

Lemma 5.19. Let \( D \in \mathcal{L}_R^* \) and let \( I_1, \ldots, I_k \) be disjoint arcs of \([n]\). Let \( \sigma_I \) be the permutation that reverses \( I \). Then, at most one of the \( \sigma_I \)'s produces a new Robinson ordering.

Proof. Suppose \( \sigma_I \) is a Robinson ordering for some arc \( I \). For every \( i \in [n] \), let \( M(i) = \arg \max_j D(i, j) \). We claim that for every \( i \notin I \) it holds that \( M(i) \subset I \). Otherwise, given \( m^* \in M(i) \), by the connectivity of \( I \), we must have that \( I \) 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 \( I \) would violate the monotonicity of such sequence (an thus the unimodality of the whole sequence). This proves the claim. Since \( I^c \) is an arc, by the same argument we have that \( i \in I \) implies \( M(i) \subset I^c \). Hence, the only arcs that can be reversed are \( I \) and \( I^c \).\( \blacksquare \)

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 $S^1$. 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 $S^1$. We assume the set $[0, 1)$ is endowed with a dissimilarity $d$. We will make some assumptions that relate the circular ordering to the circular Robinson property.

Assumption 6.1. $d$ is continuous, and strict circular Robinson, i.e.,

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

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 $d(t, s) := \|\gamma(t) - \gamma(s)\|$, 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 $d$. To do so, we consider the family of cyclic shifts $\{\pi_s : s \in [0, 1)\}$ defined by $\pi_s(t) = t + s \mod 1$, and the reversal $\pi_r(t) = 1 - t$. We let $\text{Dih}_\infty := \langle \pi_s, \pi_r : s \in [0, 1) \rangle$. In addition, given an arc $I := (t, s) \subseteq [0, 1)$, we let $\sigma_I$ be the bijection that reverses $I$ and fixes $I^c$. Since in the finite case all solutions can be expressed as compositions of such permutations, in the continuous case we look for solutions in $\text{Sym}(\infty) \triangleq \text{Dih}_\infty \circ (\sigma_I : I \text{ arc})$.

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

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 $X_n := \{x_0, \ldots, x_{n-1}\}$. If we let $\lambda$ be the Lebesgue measure on $[0, 1)$, then our sampling is distributed as $\lambda^n$. Let $D_{X_n}$ denote the dissimilarity matrix associated to $X_n$. In particular, if $x_0, \ldots, x_{n-1}$ 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 exist non-trivial orderings (cf. Lemma 5.19). In what follows we study conditions under which for a large sample, any ordering in $\text{Sym}(\infty)_n$ is close to the one induced by the curve. Our closeness measure is given by the Kendall-tau’s metric $\tau_K$ 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 $\pi_1$ and $\pi_2$ as $\tau_K(\pi_1, \pi_2) = \frac{\lvert \mathcal{G}(\pi_1, \pi_2) \rvert}{\binom{n}{2}}$, where $\mathcal{G}(\pi_1, \pi_2)$ corresponds to the

\[ \forall \text{ cyclically ordered } x, y, z, w \in [0, 1) : d(y, w) > \min\{d(y, x), d(y, z)\}. \]
This contradicts 1. Since $d_i \leq f_j$ we have $f_j \geq \min\{f_i, f_k\}$ (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 $i \leq j \leq k$ and $x_j < \min\{x_i, x_k\}$. Then $x_i > x_j$ and $x_j < x_k$. This implies $m \leq j$ and $m \geq j$. 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 $(i_{-1}, i_0, i_1), (i_0, i_1, i_2) \in C_n$ 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, $q_1 \leq q_2 \leq q_{−1}$.

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

**Proof of Proposition 3.7.** For simplicity we define $d^j_i \triangleq D(i, i + j \mod n)$. (2 $\Rightarrow$ 1) From Proposition A.2 we can write $i = j + q_i, k = j + q_k$ and $\ell = j + q_\ell$ with $q_k \leq q_\ell \leq q_i$. Since $d^j_i$ is unimodal, from Proposition A.1 we deduce $d^j_{q_i} \geq \min\{d^j_{q_k}, d^j_{q_\ell}\}$. (1 $\Rightarrow$ 2) If $d^j_i$ is not unimodal, by Proposition A.1 there are $q_k \leq q_\ell \leq q_i$ with $d^j_{q_i} < d^j_{q_k}$ and $d^j_{q_\ell} < d^j_{q_\ell}$. 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 $(i, j, k), (j, k, \ell) \in C_n$. This contradicts 1.
A.2. Proofs for Section 5.

**Proof of Lemma 5.9.** Let \( z \in X \) and denote \( f_z(\cdot) \triangleq d(z, \cdot) \). First, notice that

\[
(\exists r > 0). \{a, a' \} \subset B_r(z) \land \{b, b' \} \subset B_r(z)^c
\]

\[
\iff \max\{f_z(a), f_z(a')\} < \min\{f_z(b), f_z(b')\}.
\]

(1. \( \Rightarrow \) 2.) By (A.1), this implication is direct from the fact that \( a \in A, b \in B, a' \in A' \) and \( b' \in B' \). (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 \( z \in X \) and any \( t > 0 \) we have that if there is some \( a \in A \) and \( a' \in A' \) such that \( \max\{f_z(a), f_z(a')\} < t \). Then, \( f_z(a) < t \land f_z(a') < t \). Which implies \( \max\{\min f_z(A), \min f_z(A')\} < t \). Similarly, the existence of \( b \in B \) and \( b' \in B' \) such that \( \min\{f_z(b), f_z(b')\} > t \) implies that \( \min\{\max f_z(B), \max f_z(B')\} > t \). This proves the final implication, and hence the result.

**Proof of Lemma 5.13.** We only prove the strict case, as the non-strict case follows an analogous argument. Let \( x_d \in B_d \) and \( x_b \in B_b \) be such that \( D^{\text{min}}(B_b, B_d) = D(x_b, x_d), \) 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

\[
D^{\text{min}}(B_b, B_d) = D(x_b, x_d) > \min\{D(x_b, x_a), D(x_b, x_c)\}
\]

On the other hand,

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

by definition of \( D^{\text{min}} \), proving the result.

A.3. Proof of Theorem 6.2.

**Proof.** Suppose by contradiction that there exist an arc \( I = [a, b] \) such that \( d \circ \sigma_I \) is strict Robinson. For \( \epsilon > 0 \) small, \( a - \epsilon, a, b, b + \epsilon \) are cyclically ordered. By hypothesis,

\[
d(\sigma(a), \sigma(b + \epsilon)) > \min\{d(\sigma(a), \sigma(a - \epsilon)), d(\sigma(a), \sigma(b))\},
\]

and since \( b + \epsilon, a - \epsilon \notin I \), 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

\[
d(b, b + \epsilon) > \min\{d(b, a - \epsilon), d(b, a)\}.
\]

Let \( \delta := d(b, a) > 0 \). By continuity we get that \( d(\sigma(a)), \sigma(b + \epsilon)) \to 0 \) and \( d(b, a - \epsilon) \to \delta \) as \( \epsilon \to 0 \). For sufficiently small \( \epsilon \), 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 $x_0, x_1, \ldots x_{n-1} \sim \text{Unif}[0,1)$. 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 $I \subset X$, we write $\mu(I)$ to denote 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 $I \subset X_n$ such that $\mu(I) > \delta$ has a unique orientation in any circular Robinson ordering of $D_{X_n}$.

**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 $I = [s,t]$ in $X_n$. Suppose $\epsilon_n < \ell \delta/(L + \ell)$ and $\mu(I) > \delta$. Let $s^+ = x_{(i-1 \mod n)}$ and $t^+ = x_{(j+1 \mod n)}$. We claim that $B_s(d(s^+)) \nsubseteq I$. To prove the claim, it suffices to prove that

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

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

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

and therefore $\min\{d(s,t), d(s, t^+)\} \geq \ell \cdot (\delta - \epsilon_n)$. 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 $x_0, x_1, \ldots x_{n-1} \sim \text{Unif}[0,1)$. Let $E_n(\delta) = \{\epsilon_n < \ell \delta/(L + \ell)\}$, then

$$
\int_{E_n(\delta)} \text{diam} S_{K_n}(D_{X_n}(\omega)) \cdot d\lambda^n(\omega) = O\left(\delta^2 + \frac{\delta \log(1/\delta)}{n}\right).
$$

**Proof.** Recall from Lemma 5.19 that there is at most one non-trivial ordering $\sigma_I$ of $D_{X_n}$ which corresponds to the permutation that reverses the arc $I \cap X_n$. Therefore, it suffices to bound the integral of the random variable $\tau_K(\text{id}, \sigma^*)$, where $\sigma^* \in \text{arg min}_{\sigma \in \{\text{id}, \sigma_I\}} \tau_K(\text{id}, \sigma)$.

The number of discordant pairs between $\sigma^*$ and id is bounded by $\frac{1}{2} \min\{[I \cap X_n], n - [I \cap X_n]\}^2$. Hence, we will focus in bounding this expression. Let $\omega \in E_n(\delta)$. By Proposition A.4, $\mu(I(\omega)) \leq \delta$. This implies that either $\lambda(I) \leq \delta$ or $\lambda(I) \geq 1 - \delta$. Therefore,

$$
\int_{E_n(\delta)} \min\{[I \cap X_n], n - [I \cap X_n]\}^2 \cdot d\lambda^n(\omega) \leq \int_{E_n(\delta)} \max_{\text{interval}, \lambda(J) \leq \delta} [J \cap X_n]^2 \cdot d\lambda^n(\omega) \leq \int_{E_n(\delta)} \max_{J \cap X_n} |J \cap X_n|^2 \cdot d\lambda^n(\omega).
$$

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 $\delta$. For any $\omega \in \Omega$, the maximizer in the integral above lies in at most two of the partition intervals. Therefore, $\max_{J \cap X_n} \leq 2 \max_{i \in [n]} |J_i \cap X_n|$. Next, we can estimate $\max_{i \in [n]} |J_i \cap X_n|$ 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 $Z = \text{diam}(Sc_{\delta}(D_{X_n})))$. Then,

$$\mathbb{E}[Z] = \int_{E_n(\delta)} Z(\omega) d\lambda(\omega) + \int_{E_n(\delta)^c} Z(\omega) d\lambda(\omega) \leq O\left(\delta^2 + \frac{\delta \log(1/\delta)}{n}\right) + \mathbb{P}[E_n(\delta)^c],$$

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

$$\mathbb{P}[E_n(\delta)^c] \leq \sum_{j=1}^{n+1} (-1)^{j-1} \binom{n+1}{j} (1 - jx)^n \leq \sum_{j=1}^{n+1} \left(\frac{e(n+1)}{j}\right)^j \exp\{-xjn\},$$

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 $\sum_{j=1}^{n+1} \frac{1}{(n+1)^j} \in O(1/n)$. By (A.5) and (A.6) we conclude that $\mathbb{E}[Z] = O\left(\frac{L + \ell \log n}{n}\right)$.

Appendix B. Auxiliary subroutines.

Algorithm B.1 Arc Partition

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

Algorithm B.2 Depth-First Search

1: Input: The neighbourhood function of a graph $N_G(\cdot)$, a tuple $\alpha$ of visited nodes and a starting node $x$
2: $\alpha(n) = x$ \{Set $x$ as $n$-th visited node where $n$ is the size of $\alpha$\}
3: for $y \in N_G(x) \setminus \alpha$ do
4: $\alpha = \text{DFS}(N_G, \alpha, y)$ \{Recurse over all adjacent nodes that have not been visited\}
5: end for
6: return $\alpha$
7: Output: A tuple of visited nodes $\alpha$

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. Kőhler, S. Kühnert, and O. Verbitsky, *Circular-arc hypergraphs: Rigidity via connectedness*, Discrete Applied Mathematics, 217 (2017), pp. 220–228.

[15] V. Novák, *Cyclically ordered sets*, Czechoslovak Mathematical Journal, 32 (1982), pp. 460–473.

[16] I. Pinelis, *Order statistics on the spacings between order statistics for the uniform distribution*, arXiv preprint arXiv:1909.06406, (2019).

[17] P. Préa and D. Fortin, *An optimal algorithm to recognize robinsonian dissimilarities*, Journal of Classification, 31 (2014), pp. 351–385.

[18] A. Quilliot, *Circular representation problem on hypergraphs*, Discrete mathematics, 51 (1984), pp. 251–264.

[19] A. Recanati, T. Kerdreux, and A. d’Aspremont, *Reconstructing latent orderings by spectral clustering*, arXiv preprint arXiv:1807.07122, (2018).

[20] M. Seston, *Dissimilarités de Robinson: algorithmes de reconnaissance et d’approximation*, PhD thesis, Aix-Marseille 2, 2008.

[21] A. Tucker, *Matrix characterizations of circular-arc graphs*, Pacific Journal of Mathematics, 39 (1971), pp. 535–545.