



HAL
open science

DISTANCES, ORDERS AND SPACES

Pascal Préa

► **To cite this version:**

Pascal Préa. DISTANCES, ORDERS AND SPACES. . 14-th Scientific Meeting. Classification and Data Analysis Group (CLADAG 2023), Carla Rampichini; Micheke La Rocca, Sep 2023, Salerne, Italy. hal-04159165

HAL Id: hal-04159165

<https://hal.science/hal-04159165v1>

Submitted on 11 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DISTANCES, ORDERS AND SPACES

Pascal Pr ea^{1,2}

¹ Aix-Marseille Universit e, CNRS, LIS, Marseille, France
(pascal.prea@lis-lab.fr)

²  cole Centrale de Marseille, Marseille, France

ABSTRACT: A dissimilarity d on a set X is said to be *Robinson* if there exists a total order, said *compatible*, on X such that $x < y < z \implies d(x, z) \geq \max\{d(x, y), d(y, z)\}$. Roughly speaking, d is Robinson if the points of X can be represented on a line *ie.* Robinson dissimilarities generalize line distances.

In this paper, we define k -dimensional Robinson dissimilarities, which generalize the possibility, for a metric set (X, d) , to be embedded into a k -dimensional Euclidean space. This generalization is more flexible than the classical embedding and we show that every dissimilarity on an n -set X is $(\log n)$ -dimensional Robinson. We give an $O(n^3)$ algorithm which builds such an embedding. This algorithm is based on an incremental algorithm to recognize Robinson dissimilarities.

KEYWORDS: Robinson dissimilarities, embeddings, incremental algorithms.

1 Introduction

Given a finite set X , a *dissimilarity* on X is a symmetrical function $X \times X \mapsto \mathbb{R}^+$ such that $\forall x \in X, d(x, x) = 0$ (we say that (X, d) is a dissimilarity space). Dissimilarities generalize distances (a distance is dissimilarity with the triangular inequality).

Given a dissimilarity on a set X , a fundamental problem is to derive “geometrical” properties of X from d , or to characterize dissimilarities from which such properties can be obtained. For instance Robinson dissimilarities (Robinson 1951) correspond to points on a line. These dissimilarities were invented to solve seriation problems in Archeology, but they are now a classical tool for seriation problems in any field. They are also linked with Pyramids (Diday 1986, Durand & Fichet 1988), the standard model with overlapping classes. Moreover, they play an important role to recognize tractable cases for TSP ( ela & al. 2023).

In this paper, we generalize Robinson dissimilarities to k -dimensional-*Robinson* dissimilarities, which represent the fact for X to be embedded into a k -dimensional space. This embedding is less strict than the usual Euclidean

embedding. We show that, if d is a dissimilarity on a set X with $|X| = n$, then d is $(\log n)$ -Robinson and we give a $O(n^3)$ algorithm which builds such an embedding. This algorithm is based on an incremental algorithm to recognize Robinson dissimilarities which is presented in the last section.

2 Robinson dissimilarities

A dissimilarity space (X, d) is *Robinson* if there exists a total order, which is said to be *compatible*, on X such that

$$x < y < z \implies d(x, z) \geq \max\{d(x, y), d(y, z)\} \quad (1)$$

Let (X, d) a dissimilarity space and $<$ be an order on X . Notice that $<$ is a compatible order of (X, d) (which is thus a Robinson space) if and only if:

$$x \leq y < z \leq t \implies d(y, z) \leq d(x, t) \quad (2)$$

Given a total order $<$ on X and $x, y, z \in X$, we say that y is *between* x and z for $<$ if $x < y < z$ or $z < y < x$. The set of the elements between x and z is an *interval* for $<$ and we denote it by $[x, z]_{<}$. Notice that $[x, z]_{<} = [z, x]_{<}$.

3 Multidimensional Robinson dissimilarities

Let (X, d) a dissimilarity space and $k \in \mathbb{N}_1$. We say that (X, d) is *k-Robinson* if there exist k orders $<_1, <_2, \dots, <_k$ such that:

$$\forall x, y, z, t \in X, (\forall 1 \leq i \leq k, y, z \in [x, t]_{<_i}) \implies d(y, z) \leq d(x, t)$$

We say that (X, d) is *k-quasi-Robinson* if there exist k orders $<_1, <_2, \dots, <_k$ such that:

$$\forall x, y, z \in X, (\forall 1 \leq i \leq k, y \in [x, z]_{<_i}) \implies d(x, z) \geq \max\{d(x, y), d(y, z)\}$$

If $k = 1$, it is equivalent for a dissimilarity space to be Robinson or 1-quasi-Robinson. For $k \geq 2$, then if (X, d) is k -Robinson, then (X, d) is k -quasi-Robinson, but the converse is false (see Figure 1). Notice in addition that, if (X, d) is k -(quasi-)Robinson, then (X, d) is $k + 1$ -(quasi-)Robinson. The smallest k such that (X, d) is the *Robinson dimension* of (X, d) .

If a metric space (X, d) can be embedded into a \mathbb{R}^k , then (X, d) is k -Robinson. But the Robinson dimension of (X, d) is generally smaller. For instance, if $|X| = n$ and d is the constant dissimilarity, then (X, d) is Robinson (its Robinson dimension is 1) although it needs an $n - 1$ -dimensional Euclidean space to be embedded. Moreover, we have:

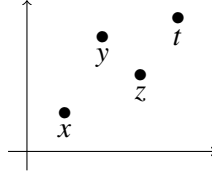


Figure 1. A set X with four points x, y, z, t . If (X, d) is 2-quasi-Robinson with the two orders represented by the two axis, then no condition is imposed on $d(y, z)$ and we can set $d(y, z) > d(x, t)$. If (X, d) is 2-Robinson (with the same orders), then $d(y, z) \leq d(x, t)$.

Proposition 1 The Robinson dimension of a dissimilarity space (X, d) with $|X| = n$ is $\leq \lceil \log_2 \lceil \frac{n}{3} \rceil \rceil + 1$.

Algorithm 1 returns an approximate value for the Robinson dimension of a dissimilarity space.

Algorithm 1: APPROXIMATE-ROBINSON-DIMENSION

Input: (X, d) , a dissimilarity space.

Output: An upper bound on the Robinson dimension of (X, d) .

begin

$X' \leftarrow X ; k \leftarrow 0 ;$

SORT-LINES(X, d) ;

while $X' \neq \emptyset$ **do**

$S \leftarrow$ MAXIMAL-ROBINSON-SUBSPACE(X', d) ;

$X' \leftarrow X' \setminus S ;$

$k \leftarrow k + 1 ;$

return $\lceil \log_2 k \rceil + 1 ;$

The function SORT-LINES(X, d), for every $x \in X$, sorts the points of X by increasing values of their distance from x . This function runs in $O(n^2 \log n)$ where $n = |X|$. The function MAXIMAL-ROBINSON-SUBSPACE returns a subset S of X' , maximal for inclusion and such that (S, d) is Robinson. This can be easily implemented by a greedy algorithm. We will see in Section 4 that, after SORT-LINES, such a greedy version of MAXIMAL-ROBINSON-SUBSPACE runs in $O(|X'|^2)$. So, as there is at most $n/3$ iterations of the **while** loop, Algorithm 1 runs in $O(n^3)$.

4 An incremental algorithm to recognize Robinson dissimilarities

In order to implement MAXIMAL-ROBINSON-SUBSPACE, we need a function ADD-AND-TEST which takes as entry a dissimilarity space (X, d) , a set $S \subset X$ such that (S, d) is Robinson, the PQ-tree $\mathcal{T}_P(S, d)$ and a point $x \in X \setminus S$. A PQ-tree (Booth & Lueker 1976) is a data structure which can encode all the compatible orders of a Robinson dissimilarity. ADD-AND-TEST returns the PQ-tree $\mathcal{T}_P(S \cup \{x\}, d)$ (If $(S \cup \{x\}, d)$ is not Robinson, then $\mathcal{T}_P(S \cup \{x\}, d) = \text{none}$). The algorithm of ADD-AND-TEST can be sketched as follows:

1. Compute the sets $B_\delta^S := B_\delta(x) \cap S$.
2. Insert the sets B_δ^S into $\mathcal{T}_P(S, d)$. We get a PQ-tree $\mathcal{T}_P^x(S, d)$.
3. Add the point x to $\mathcal{T}_P^x(S, d)$. We get the PQ-tree $\mathcal{T}_P(S \cup \{x\}, d)$. This will be done in two steps:
 - (a) Consider only the points of S the closest from x .
 - (b) Consider the other points of S .

Acknowledgements

This work was supported in part by ANR project DISTANCIA (ANR-17-CE40-0015).

References

- BOOTH, K.S. & LUEKER, G.S. 1976, Testing for the Consecutive Ones Property, Interval Graphs and Graph Planarity Using PQ-Tree Algorithm, *Journal of Computer and System Sciences* 13, 335–379.
- ÇELA, E., DEINEKO, V. AND WOENINGER G.J. 2023, Recognising permuted Demidenko matrices, ArXiv:2302.05191v1.
- DIDAY, E. 1986, Orders and overlapping clusters by pyramids in *Multidimensional Data Analysis*, J. de Leeuw, W. Heiser, J. Meulman and F. Critchley Eds., 201–234, DSWO.
- DURAND, C. & FICHET, B. 1988, One-to-one correspondences in pyramidal representation: an unified approach, in *Classification and Related Methods of Data Analysis*, H.H. Bock Ed., 85–90, North-Holland.
- ROBINSON, W.S. 1951, A method for chronologically ordering archeological deposits, *American Antiquity* 16, 293–301.