Robinsonian dissimilarities: a randomized analysis and extentions

Encadrants: V. Chepoi, G. Naves, P. Préa

Laboratoire: LIS, équipe ACRO

{victor.chepoi,guyslain.naves,pascal.prea}@lis-lab.fr

1. Context

A major issue in classification and data analysis is to visualize simple geometrical and relational structures between objects. Necessary for such an analysis is a similarity or a dissimilarity measure on a set of objects. Many applied algorithmic problems ranging from archeological dating to DNA sequencing and numerical ecology involve ordering a set of objects so that closely coupled elements are placed near each other. The rearranged data may then speak for themselves. For example, the classical *seriation problem* [10, 11] is to find a simultaneous ordering of the rows and the columns of the dissimilarity matrix with the objective of revealing an underlying one-dimensional structure, i.e., large values should be concentrated around the main diagonal as closely as possible, whereas small values should fall as far from it as possible. This goal is best achieved by considering the so-called *Robinson property* [17]. A dissimilarity matrix is said to have this property if its values increase monotonically in the rows and the columns when moving away from the main diagonal in both directions. Seriation is of importance in archeological dating [10, 11, 17], clustering hypertext orderings [2], sparse matrix ordering [1], DNA sequencing [14], and matrix visualization methods.

The most common methods for clustering provide a visual display of data in the hierarchical form of *dendrograms*. Dissimilarities which are in perfect agreement with dendrograms satisfy the Robinson property and are best known under the name of *ultrametrics*. Generalizing the correspondence between ultrametrics and dendrograms, it has been shown in [7] and [8] that there exists a one-to-one correspondence between the Robinson dissimilarities and *pyramids*. Analogously to dendrograms, the objects belonging to a single cluster in a pyramid appear in consecutive order in the Robinsonian matrix, however in pyramids two clusters may overlap.

Let X be a set of n elements to sequence, endowed with a dissimilarity function that reflects the desire for two elements to be near or far from each other in the sequence. Recall that a *dissimilarity* is a symmetric function d from X^2 to the nonnegative real numbers and vanishing on the diagonal, i.e. $d(x, y) = d(y, x) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$. We call $d(x, y)$ the *distance* between the objects $x, y \in X$. If d satisfies, in addition, the triangle inequality $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$, then d is called a *metric*. A dissimilarity d and a total order $\prec$ on a set X are said to be *compatible* if $x \prec z \prec y$ implies that $d(x, y) \geq \max\{d(x, z), d(z, y)\}$. A dissimilarity d on X is said to be *Robinsonian* if it admits a compatible order. Equivalently, d is Robinsonian if its matrix can be symmetrically permuted so that its elements do not decrease when moving away from the main diagonal along any row or column. Such a matrix is called *Robinson* or *linear*.

Due to their importance, several algorithms have been proposed to recognize Robinson (dis)similarities. Atkins et al. [1] showed that if S is a Robinson similarity matrix, then the coordinates of the eigenvector of its smallest nonzero eigenvalue of the Laplacian of S constitute a monotone sequence of numbers. They use this result and PQ-trees to design an algorithm of complexity $O(nT(n) + n^2 \log n)$ to recognize if a similarity matrix of size $n \times n$ is pre-Robinson, where $T(n)$ is the complexity of computing the respective eigenvector. Mirkin and Rodin [14] describe an $O(n^4)$ algorithm for testing if a dissimilarity d on n points is Robinsonian. For this, they build up the hypergraph of all balls of d and test using the PQ-tree algorithm if this hypergraph is an interval hypergraph. A simple divide-and-conquer $O(n^3)$-time algorithm for the same recognition problem has been designed in [4]. An $O(n^2 \log n)$ time algorithm was proposed in [12] and [19].

Finally, Préa and Fortin [13] presented an optimal $O(n^2)$ algorithm, using the algorithm of computing the first PQ-tree, which they update throughout the algorithm. Finally, [5] presented a factor 16 approximation algorithm for ℓ_∞-best fitting a dissimilarity by a Robinson dissimilarity.

2. Randomized approach

The transformation of a dissimilarity matrix into a Robinson (monotone) matrix, by ordering its rows and the columns, can be viewed as the 2-dimensional version of the classical algorithmic problem of *sorting of n numbers*. Quicksort is the most practical sorting algorithm. It uses the *divide-and-conquer* paradigm and perform a partition of the current list into *three parts* with respect to a randomly chosen *pivot*: the elements less or equal the pivot, the pivot, and the elements larger than the pivot. The main tool in the complexity analysis of Quicksort is the estimation of the probability that two elements are compared by the algorithm. One important thing is that a randomly chosen element has probability $\frac{1}{2}$ to be a good pivot, i.e., to provide a balanced partition. All this analysis works, if the input array is a random permutation of the input elements.

The divide-and-conquer algorithm of [4] selects as a pivot a special pair of points and partitions the set X into *five parts*. It is not clear how to perform a partition of X into a constant number of parts with respect to pivot consisting of a randomly chosen (1) element, (2) pair of elements, or (3) constant number of elements. What will be a good pivot in each of these cases? How to compute the probability that two entries of the distance matrix of X will be compared?

Therefore, the main goal will be to try to extend *the* Quicksort *algorithm and its analysis to the problem of computation of a compatible order of a Robinsonian dissimilarity.* The second goal is to use the *Ehrenfest model* to find a given compatible order of a Robinson martix. For this, consider a Markov chain with $n(n-1)/2$ states, where the final state correspond to a desired compatible order $\prec^*$ and state i correspond to all total orders $\prec$ which differs from $\prec^*$ by exactly $\frac{n(n-1)}{2} - i$ inversions. The transitions correspond to inversions. The goal will be to *evaluate the cover time of this Markov chain.* If this approach to finding any compatible order will work, it will be interesting to apply this approach to other problems, in particular to approximation problems.

3. Extensions

3.1. **Reconstruction problem.** In computational molecular biology, the aim of *restriction site mapping* is to locate the restriction sites of a given enzyme on a given DNA molecule. Determining the location of sites from restriction site data is a difficult algorithmic problem; see [16] for the introduction and discussion of many approaches. *Partial digest* is an approach to the restriction site mapping is an approach introduced by Skiena and Sundaram [20] and is based on the theory of *homeometric sets* on the line of Rosenblatt and Seymour [18]. Two noncongruent n-point sets of the line $\mathbb{R}$ are *homeometric* if the multisets of $\binom{n}{2}$ distances they determine are the same. The partial digest approach of [20] of reconstructing n-point sets of $\mathbb{R}$ from the multisets of $\binom{n}{2}$ interpoint distances. Skiena and Sundaram [20] proposed practical algorithms for reconstructing sets from noisy interpoint distances, but it was shown in [21] that such algorithms are exponential. Thus the status of the partial digest problem is still unresolved (see also [15] for the labeled version and a polynomial time algorithm for this version).

Since Robinson dissimilarities generalize distance martices of points on the line, in the PhD thesis, we plan investigate the homeometric sets in Robinson matrices and the reconstruction problem from interpoint distances for Robinson dissimilarities. For homeometric sets, we are planning to *investigate lower and upper bounds for the maximum possible number of mutually noncongruent and homeometric Robinson dissimilarities on n points.* The same type of questions can be raised for ultrametrics, for-tree metrics or for tree-Robinsonian dissimilarities. For the reconstruction problem, we intend to *develop algorithms solving this problem*, in particular, extending the algorithm of [20]. it should be emphasized that even if the reconstruction problem for Robinsonian

dissimilarities generalize the reconstruction problem for the line, in general the first problem may have much more solutions than the second one and thus maybe easier. A very challenging goal would be to extend the approximation algorithms of [5] and [9] to solve the reconstruction problem with noisy interpoint distances (for the line or for Robinsonian matrices).

3.2. **Robinson cubes.** Warrens and Heiser [22] generalized the notion of Robinson matrix to 3-dimensions and they defined two types of *Robinson cubes* and *regular Robinson cubes*; . A *cube* is a $n \times n \times n$ array $C = (c_{ijk}$. Whereas a matrix is characterized by rows and columns, a cube consists of rows, columns, and tubes. A cube C satisfies (a) the *three-way symmetry* if $c_{ijk} = c_{\pi(i)\pi(j)\pi(k)}$ for any permutation π of i, j, k and (b) the *diagonal-plane equality* if $b_{iji} = b_{ijj}$. A *Robinson cube* is a cube B satisfying conditions (a), (b) and (c) such that the lowest entries in each row, column, and tube of B are on the main diagonal (elements b_{iii}) and moving away from this diagonal, the entries neither decrease. A Robinson cube B is *regular* if (d) all matrices, which are formed by cutting the cube perpendicularly, are Robinson dissimilarities. The paper [22] presented several examples of (regular) Robinson cubes showing that they appear naturally in statistics and data analysis, but do not provided any algorithm of their recognition.

It is an interesting and challenging algorithmic question to *design algorithms for recognizing (regular) Robinson cubes*, i.e., algorithms for permuting rows, columns, and tubes of a cube satisfying the conditions (a) and (b) to a cube also satisfying the conditions (c) and (d), or establishing that such a transformation does not exist. The natural way would be to try to extend the known methods (described above) for recognizing Robinson dissimilarities.

References

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

[2] M.W. Berry, B. Hendrickson, and P. Raghavan, Sparse matrix reordering schemes for browsing hypertext, In J. Renegar, M. Shub, and S. Smale, editors, Lectures in Applied Mathematics, Vol. 32: The Mathematics of Numerical Analysis, AMS, 1996.

[3] M.J. Brusco, A branch-and-bound algorithm for fitting anti-Robinson structures to symmetric dissimilarity matrices, Psychometrica, 67 (2002), 459–471.

[4] V. Chepoi and B. Fichet, Recognition of Robinsonian dissimilarities, J. Classif. 14 (1997), 311–325.

[5] V. Chepoi, and M. Seston, Seriation in the presence of errors: a factor 16 approximation, Algorithmica 59 (2011) 521—568.

[6] F. Critchley and B. Fichet, The partial ordre by inclusion of the principal classe of dissimilarity on a finite set, and some properties of their basic properties, In B. van Cutsen (Ed.) Classification and Dissimilarity Analysis. (1994), 5–65.

[7] E. Diday, Orders and overlapping clusters by pyramids, in Multidimensional Data Analysis, Eds., J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley, Leiden: DSWO (1986), 201–234.

[8] C. Durand and B. Fichet, One–to–one correspondences in pyramidal representation: a unified approach, in Classification and Related Methods of Data Analysis, Ed., H.H. Bock, North–Holland (1988), 85–90.

[9] J. Hastad, L. Ivansson, and J. Lagergren, Fitting points on the real line and its application to RH mapping, Journal of Algorithms 49 (2003), no. 1, 42–62.

[10] L.J. Hubert, Some applications of graph theory and related nonmetric techniques to problems of approximate seriation: The case of symmetric proximity measures, British J. Math. Stat. Psych., 27 (1974), 133–153.

[11] D.G. Kendall, Seriation from abundance matrices, in Mathematics in the Archaeological and Historical Sciences, 1971. Edited by F. R. Hodson, D. G. Kendall, and P. Tautu, pp. 215–252. Edinburgh University Press.

[12] M. Laurent and M. Seminaroti, Similarity-First Search: a new algorithm with application to Robinsonian matrix recognition, SIAM J. Discret. Math. 31 (2017), 1765–1800.

[13] P. Préa, D. Fortin, An optimal algorithm to recognize Robinsonian dissimilarities, J. Classif. 31 (2014), 351–385.

[14] B. Mirkin and S. Rodin, Graphs and Genes, Springer-Verlag, Berlin, 1984.

[15] G. Pandurangan and H. Ramesh, The restriction mapping problem revisited, J. Comput. Syst. Sci. 65 (2002), 526–544.

[16] P.A. Pevzner, Computational Molecular Biology: An Algorithmic Approach, MIT Press, Cambridge, 2000.

[17] W. S. Robinson, A method for chronologically ordering archaeological deposits, American Antiquity **16** (1951), 293–301.

[18] J. Rosenblatt and P. Seymour, The structure of homeometric sets, SIAM J. Discr. Meth., 3 (1982), 343–350.
[19] M. Seston, Dissimilarités de Robinson: Algorithmes de Reconnaissance et d'Approximation, PhD thesis, Université de la Méditerranée, 2008.
[20] S. Skiena and G. Sundaram, A partial digest approach to restriction site mapping, Bull. Math. Biology, 56 (1994), 275-294.
[21] Z. Zang, An exponential example for a partial digest mapping algorithm, J. Comput. Biol. 1 (1994), 235–239.
[22] M.J. Warrens and W.J. Heiser, Robinson cubes, In: Brito P, Bertrand P, Cucumel G, Caravalho F de (eds) Selected contributions in data analysis and classification. Springer, Heidelberg, 2007, pp 515–523.