University of Cagliari
Department of Mathematics and Computer Science

Matlab implementation of a spectral algorithm for the seriation problem

Anna Concas

Joint work with Caterina Fenu and Giuseppe Rodriguez

NASCA 2018
Kalamata, Greece, July 2-6 2018

The seriation problem

Seriation is an important ordering problem:

- it seeks the best enumeration order of a set of units whose interrelationship is defined by a bipartite graph;
- the sought order can be characteristc of the data, a chronological order or any sequential structure of the data;
- it appears in various fields such as archaeology, antropology, psychology, biology, etc.

Aim of the archaeological investigation: *to date ancient objects and determine their relative chronology*.

The seriation problem in archaeology: the data matrix

The problem data can be represented by a matrix A (*data matrix*) in which the rows are the archaeological units and the columns represent the types of the archaeological finds.

The data matrix can be of two types:

- *incidence matrix*: binary representation, the element $a_{i,j}$ is 1 if the type j is present in unit i, 0 otherwise;
- *abundance matrix*: it reports the number of objects belonging to a certain type in a given unit or its percentage.

Example of data matrix

	G3	F27	S1	F26	N2	F24	P6	F25	P5	P4	N1	F23
Mollebakken 2	1	1	1	1	0	0	0	0	0	0	0	0
Kobbea 11	0	1	1	0	1	1	0	0	0	0	0	0
Mollebakken 1	1	1	0	1	1	0	1	1	0	0	0	0
Levka 2	0	1	1	0	1	0	0	1	1	0	0	0
Grodbygard 324	0	0	0	0	1	1	0	0	0	1	0	0
Melsted 8	0	0	1	1	0	0	1	1	0	1	0	0
Bokul 7	0	0	0	0	0	0	1	1	0	0	1	0
Heslergaard 11	0	0	0	0	0	0	0	1	0	1	0	0
Bokul 12	0	0	0	0	0	0	0	1	1	0	0	1
Slamrebjerg 142	0	0	0	0	0	0	0	0	0	1	0	1
Nexo 6	0	0	0	0	0	0	0	0	0	1	1	1

Table: *Adjacency matrix for archaeological data come from female burials of the Germanic Bornholm site (Denmark).*

Aim of seriation

Aim of seriation:

- arrange the locations in chronological order by assuming that the types were produced only for a limited period of time;
- the purpose of determining a relative chronology results in obtaining an ordering of the rows and columns of the data matrix that places the nonzero entries close to the main diagonal.

A *bipartite graph* G is a graph whose vertices can be divided into two disjoint sets U and V containing n and m nodes, respectively, such that every edge connects a node in U to one in V.

In our archaeological metaphor:

- U represents the excavation sites;
- V represents the found artifacts.

Then the associated adjacency matrix A, of size $n \times m$, is obtained by setting

$$[A]_{i,j} = \begin{cases} 1 & \text{if the unit } i \text{ contains at least one object of type } j \\ 0 & \text{otherwise} \end{cases}$$

If the element $a_{ij} \neq 1$, we denote A as the *abundance matrix*.

The similarity matrix

The first mathematical definition of seriation was based on the construction of the symmetric *similarity matrix* S which can be defined as

$$S = AA^T.$$

In this case:

- each s_{ij} is equal to the number of types shared between units i and j;
- the largest value on each row is the diagonal element;
- by permuting rows and columns of S, we obtain a permutation of the corresponding rows in A that places the units similar in types closer to each other.

A particular similiarity matrix: the Robinson form

A symmetric (similarity) matrix S is in *Robinson's form*, or is an R-matrix, if and only if

$$s_{ij} \leqslant s_{ik}, \quad \text{if } j \leqslant k \leqslant i,$$
$$s_{ij} \geqslant s_{ik}, \quad \text{if } i \leqslant j \leqslant k.$$

A symmetric matrix is *pre-R* if and only if there exists a simultaneous permutation of its rows and columns which transforms it into Robinson's form.

$$\begin{bmatrix} 6 & 4 & 2 & 2 \\ 4 & 8 & 5 & 3 \\ 2 & 5 & 9 & 4 \\ 2 & 3 & 4 & 7 \end{bmatrix} \text{ R} \qquad \begin{bmatrix} 6 & 4 & 9 & 2 \\ 4 & 8 & 5 & 3 \\ 9 & 5 & 9 & 4 \\ 2 & 3 & 4 & 7 \end{bmatrix} \text{ not R} \qquad \begin{bmatrix} 9 & 2 & 5 & 4 \\ 2 & 6 & 4 & 2 \\ 5 & 4 & 8 & 3 \\ 4 & 2 & 3 & 7 \end{bmatrix} \text{ pre-R}$$

A particular similiarity matrix: the Robinson form

A symmetric (similarity) matrix S is in *Robinson's form*, or is an R-matrix, if and only if

$$s_{ij} \leqslant s_{ik}, \quad \text{if } j \leqslant k \leqslant i,$$
$$s_{ij} \geqslant s_{ik}, \quad \text{if } i \leqslant j \leqslant k.$$

A symmetric matrix is *pre-R* if and only if there exists a simultaneous permutation of its rows and columns which transforms it into Robinson's form.

$$\begin{bmatrix} 6 & 4 & 2 & 2 \\ 4 & 8 & 5 & 3 \\ 2 & 5 & 9 & 4 \\ 2 & 3 & 4 & 7 \end{bmatrix} \text{ R} \qquad \begin{bmatrix} 6 & 4 & 9 & 2 \\ 4 & 8 & 5 & 3 \\ 9 & 5 & 9 & 4 \\ 2 & 3 & 4 & 7 \end{bmatrix} \text{ not R} \qquad \begin{bmatrix} 9 & 2 & 5 & 4 \\ 2 & 6 & 4 & 2 \\ 5 & 4 & 8 & 3 \\ 4 & 2 & 3 & 7 \end{bmatrix} \text{ pre-R}$$

PQ-trees

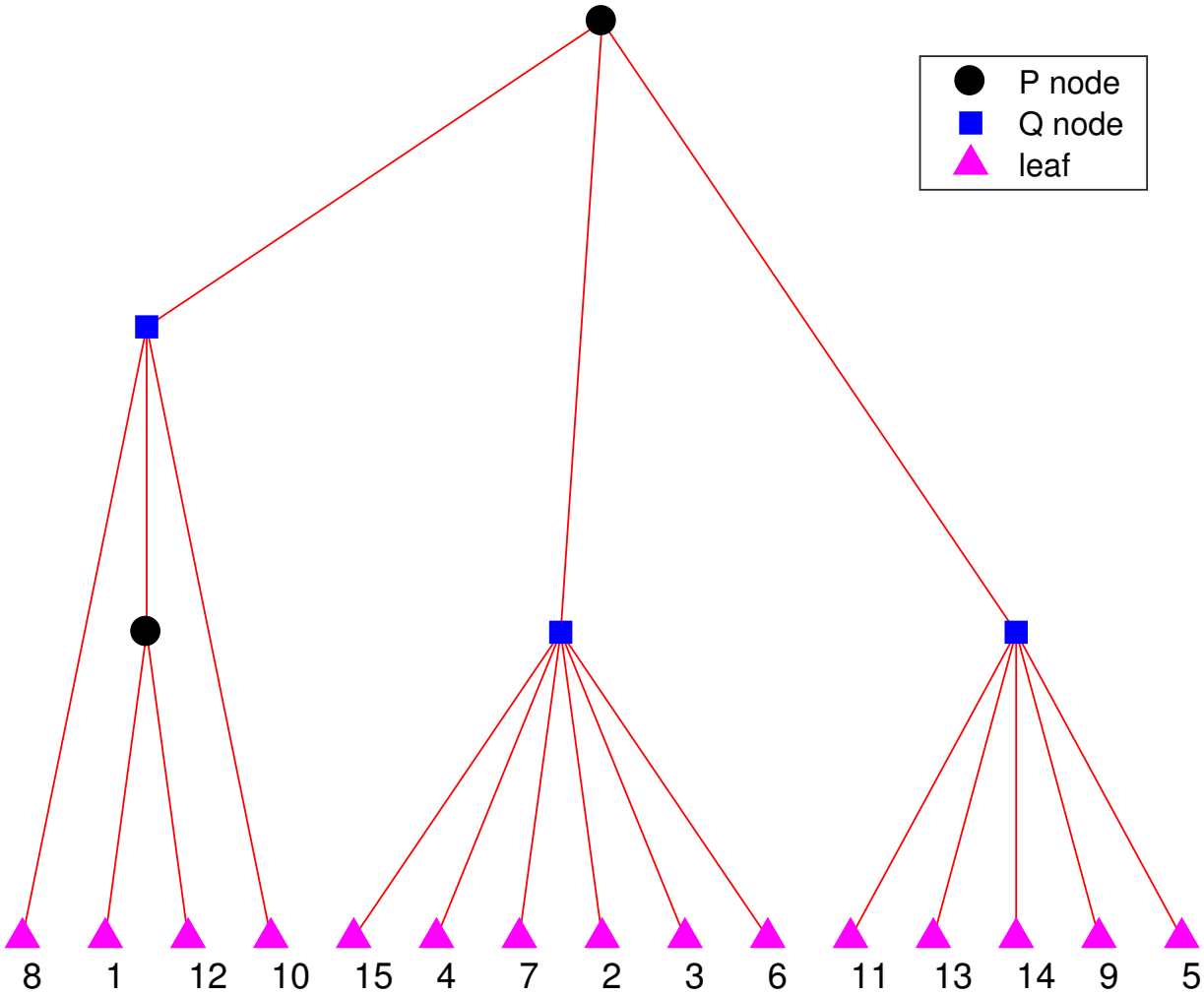
A *PQ-tree* is a data structure introduced to encode a family of permutations of a set of elements and solve problems connected to finding admissible permutations according to specific rules.

A PQ-tree T over a set $U = \{u_1, u_2, \ldots, u_n\}$ is a rooted tree whose leaves are elements of U and whose internal nodes are `P-nodes` and `Q-nodes`.

The only difference between them is the way in which their children are treated.

The root of the tree can be either a `P` or a `Q-node`.

PQ-trees (cont.)



Pq-tree over the set {1,2,...,15}

PQ-trees (cont.)

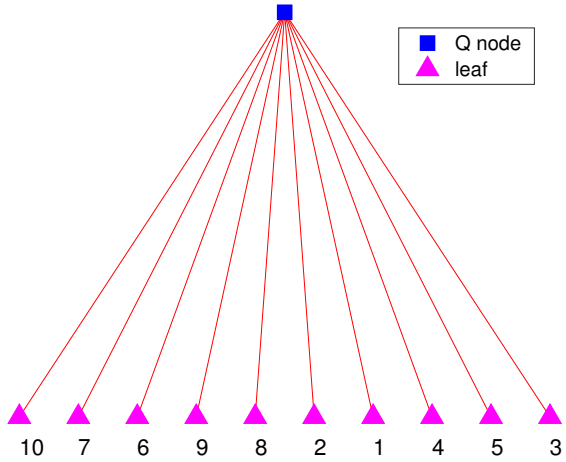
T is *proper* when:

- every $u_i \in U$ appears once as a leaf;
- every `P-node` has at least two children;
- every `Q-node` has at least three children.

Related to the sorting problem:

- for a `Q-node` only one order and its reverse are allowed;
- for a `P-node` all possible permutations of the children are permitted.

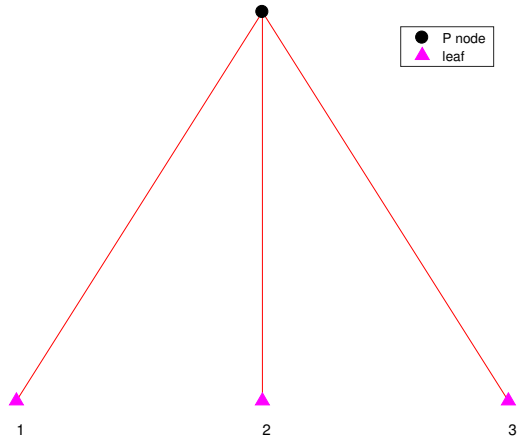
PQ-trees implementation



Pq-tree over the set {1,2,...,10}
with a Q-node as root

10	7	6	9	8	2	1	4	5	3
3	5	4	1	2	8	9	6	7	10

The 2 admissible permutations
encoded in the tree

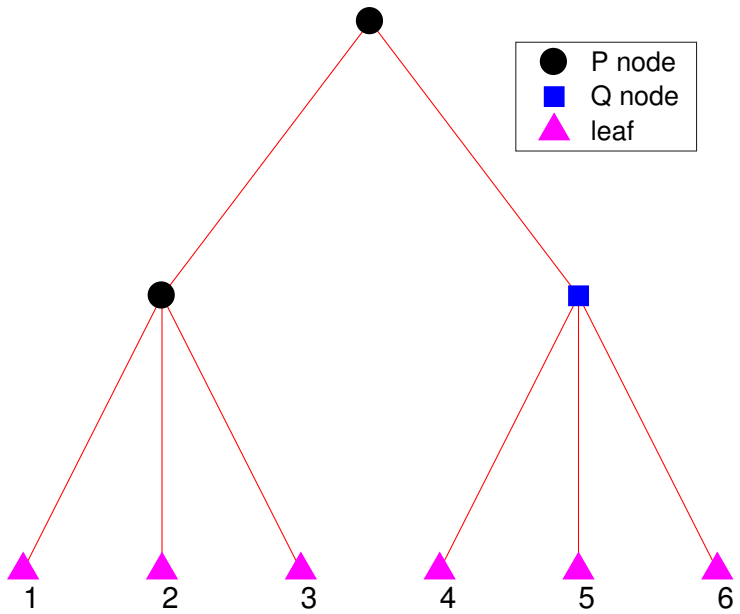


Pq-tree over the set {1,2,3}
with a P-node as root

3	2	1
3	1	2
2	3	1
2	1	3
1	3	2
1	2	3

The 3! admissible permutations
encoded in the tree

PQ-trees implementation (cont.)



PQ-tree over the set $U = \{1, \ldots, 6\}$ with a P-node as root

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

The 24 admissible permutations encoded in the tree

A spectral algorithm

Given the units set $U = \{u_1, \ldots, u_n\}$, introduce the notation

$$i \preccurlyeq j \quad \textit{if } u_i \textit{ precedes } u_j \textit{ in a desired order.}$$

Let f be a symmetric *correlation function* , reflecting the desire for units i and j to be near to each other in the sought sequence, the aim is to find all the index permutation vectors $\boldsymbol{\pi}$ s.t.

$$\pi_i \preccurlyeq \pi_j \preccurlyeq \pi_k \iff f(\pi_i, \pi_j) \geq f(\pi_i, \pi_k) \text{ and } f(\pi_j, \pi_k) \geq f(\pi_i, \pi_k) \quad (1)$$

Let F be a real symmetric matrix, s.t. $f_{ij} = f(i, j)$.

If A is the data matrix, then we will set $F = AA^T$.

If F is pre-R, then there exists a row/column permutation that takes it into R-form.

A spectral algorithm (cont.)

Following [Atkins et al., SJC 1998], the approach is to minimize the penalty function:

$$h(x) = \frac{1}{2} \sum_{i,j=1}^{n} f_{ij}(x_i - x_j)^2, \quad x \in \mathbb{R}^n$$

Computed $\mathbf{x}_{\min}$, it is sorted yielding $\mathbf{x}_{\boldsymbol{\pi}} = (x_{\pi_1}, \ldots, x_{\pi_n})^T$ and the permutation of the units $\boldsymbol{\pi}$ realizes (1).

The resulting constrained minimization problem is:

$$\begin{array}{rl} \text{minimize} & h(\mathbf{x}) = \frac{1}{2} \sum_{i,j=1}^{n} f_{ij}(x_i - x_j)^2 \\ \text{subject to} & \sum_i x_i = 0 \quad \text{and} \quad \sum_i x_i^2 = 1 \end{array}$$

A spectral algorithm (cont.)

The previous optimization problem can be rewritten as

$$\min_{x^T e=0,\ x^T x=1} x^T L x, \quad e = [1, \ldots, 1]^T \in \mathbb{R}^n \tag{2}$$

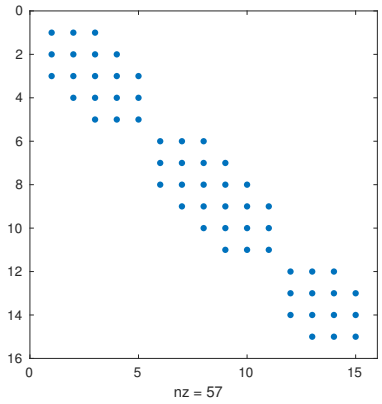
where $L = D - F$ is the (unnormalized) *graph Laplacian* of F with $D = \text{diag}(d_1, \ldots, d_n)$, called *degree matrix*, is such that $d_i = \sum_{j=1}^{n} f_{ij}$.

The (2) defines the *Fiedler value* or *algebraic connectivity* of the graph described by F and the corresponding eigenvector is the *Fiedler vector*.

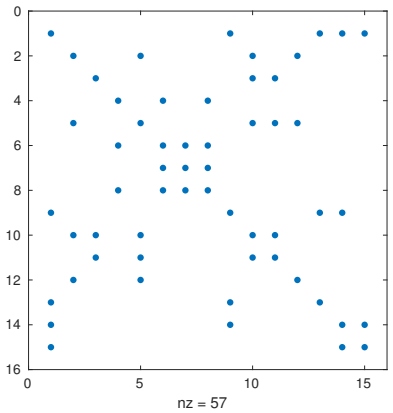
The problem is well posed only when F is pre-R, however

sorting the entries of the Fiedler vector generates an ordering that tries to keep highly correlated elements close to each other.

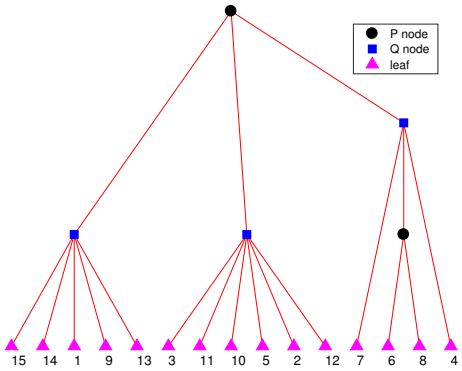
A numerical experiment



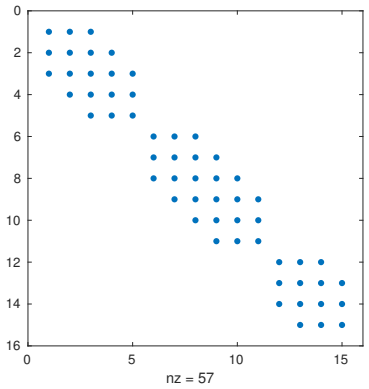
Starting matrix



Permuted matrix

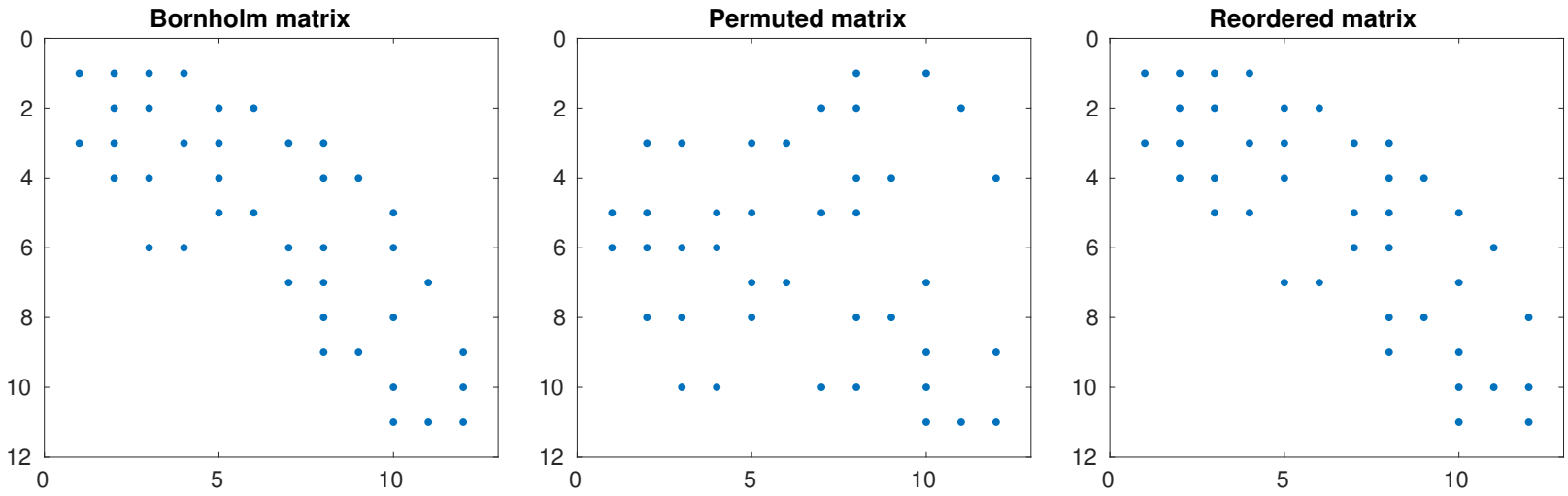


PQ-tree

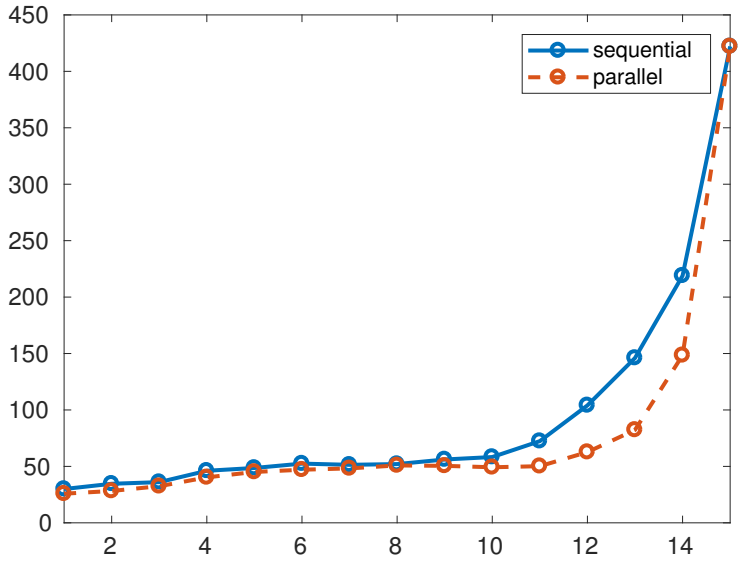


Reodered matrix

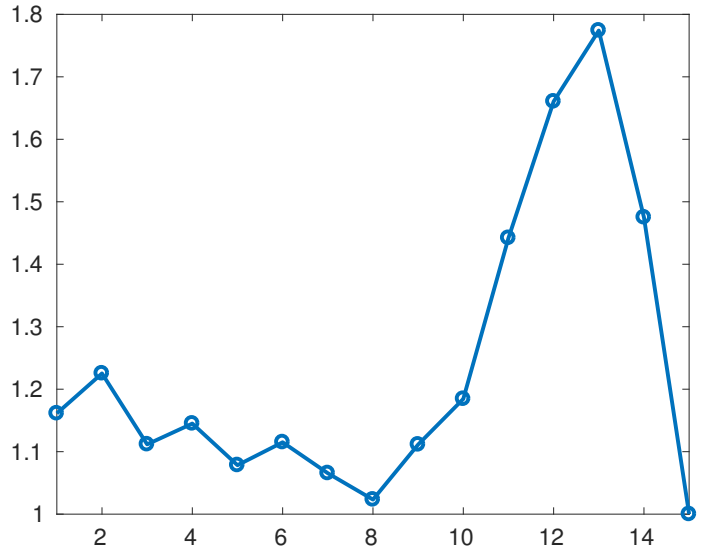
The Bornholm data set



Sequential vs parallel



Execution time of the sequential and the parallel implementations

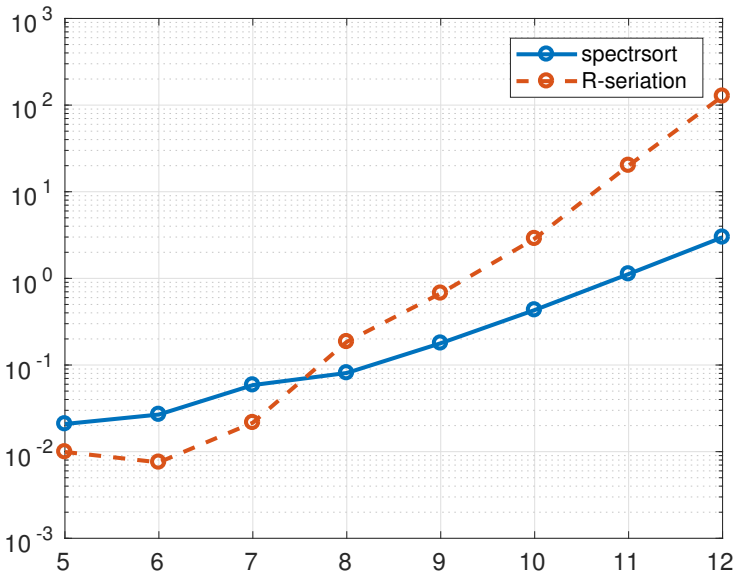


Ratio between the timings of the sequential and the parallel versions

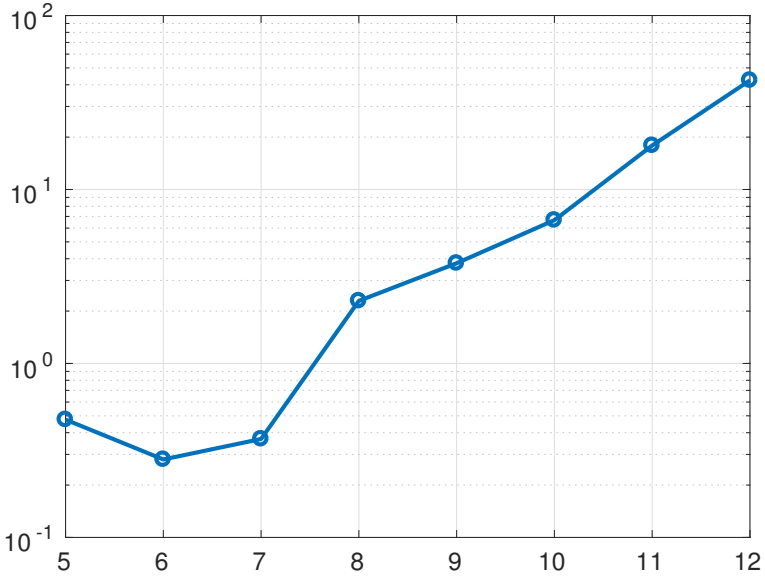
The size of the problem is $n = 2^{15} = 32768$ and for $j = 1, 2, \ldots, 15$ we generate a sequence of test matrices containing $n2^{-j}$ blocks of size 2^j.

The experiment was performed on a dual Xeon CPU E5-2620 system (12 cores), running the Debian GNU/Linux operating system and Matlab 9.2.

MATLAB vs R



Execution time of our implementation and the function seriate



Ratio between the timings of our implementation and the function seriate

Comparison with the `seriate` function (method="spectral"), from the R package `seriation` [Hahsler et al., J. Stat. Softw. 2008].

The case of a multiple Fiedler value

When the Fiedler value is a multiple root of the characteristic polynomial of the Laplacian L, there is no uniqueness in the choice of the Fiedler vector.

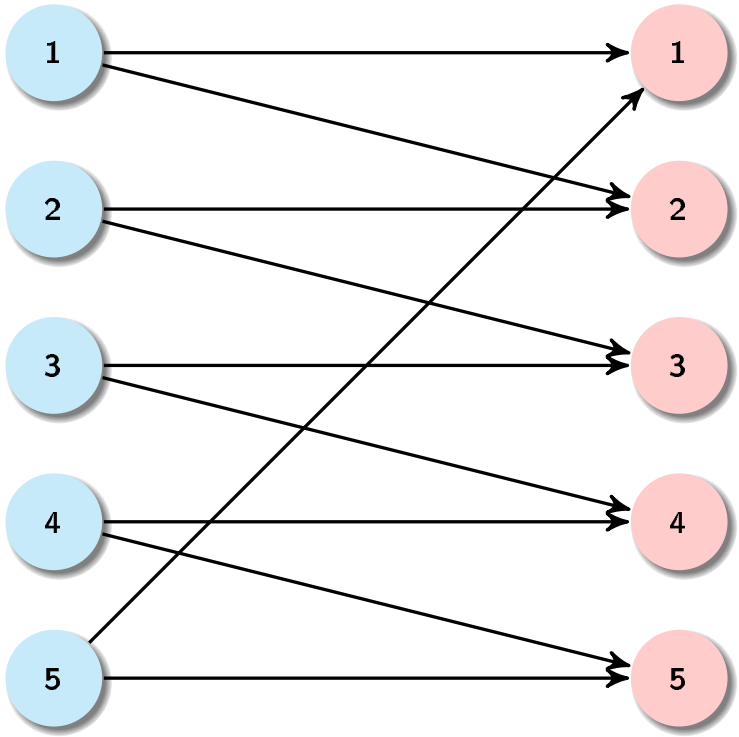
We conjecture that in this case, sorting the entries of a Fiedler vector, does not necessarly lead to all possible index permutations.

We experimentally observed that there may be some constraints to the number of permutations deriving from the Fiedler vector.

Our toolbox conventionally associates an `M-node` to the presence of a multiple Fiedler value.

The case of a multiple Fiedler value: the *cycle* graph

Consider the seriation problem described by the bipartite graph:



The *cycle* seriation problem: the units are on the left, the types on the right.

The case of a multiple Fiedler value: the *cycle* graph

The adjacency matrix associated to the graph is:

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then we compute the similarity matrix $F = AA^T$ and its Laplacian $L = D - F$:

$$F = \begin{bmatrix} 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 \end{bmatrix}, \qquad L = \begin{bmatrix} 2 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}$$

The *cycle* graph: a numerical experiment

We considered $10,000$ random linear combinations of an orthonormal basis for the eigenspace corresponding to the Fiedler value.

Each vector is sorted, and the corresponding permutations of indices are stored in the columns of a matrix.

Removing all the repeated permutations we obtain 10 ($\ll 5! = 120$) permutations

$$\begin{bmatrix} 1 & 4 & 3 & 5 & 1 & 5 & 2 & 2 & 4 & 3 \\ 5 & 5 & 2 & 4 & 2 & 1 & 1 & 3 & 3 & 4 \\ 2 & 3 & 4 & 1 & 5 & 4 & 3 & 1 & 5 & 2 \\ 4 & 1 & 1 & 3 & 3 & 2 & 5 & 4 & 2 & 5 \\ 3 & 2 & 5 & 2 & 4 & 3 & 4 & 5 & 1 & 1 \end{bmatrix}.$$

When the Fiedler value is multiple some constraints may be imposed on the number of the admissible permutations of the units.

References

- Jonathan E. Atkins, Erik G. Boman, and Bruce Hendrickson. *A spectral algorithm for seriation and the consecutive ones problem*. SIAM Journal on Computing, 28(1):297–310, 1998;
- K.S.Booth and G.S.Lueker. *Testing for the consecutive ones property,interval graphs and graph planarity using PQ-tree algorithms*. J. Comput. System Sci., 13(1976), pp. 333-379.
- A. Concas, C. Fenu, and G. Rodriguez. `PQser`: a Matlab package for spectral seriation. Numer. Algorithms, 2018;
- M. Fiedler. *Algebraic connectivity of graphs*. Czech.Math.Journal, 23(1973), pp. 298-305.

Thanks for your attention.

Grazie per l'attenzione.

References

- Jonathan E. Atkins, Erik G. Boman, and Bruce Hendrickson. *A spectral algorithm for seriation and the consecutive ones problem*. SIAM Journal on Computing, 28(1):297–310, 1998;
- K.S.Booth and G.S.Lueker. *Testing for the consecutive ones property,interval graphs and graph planarity using PQ-tree algorithms*. J. Comput. System Sci., 13(1976), pp. 333-379.
- A. Concas, C. Fenu, and G. Rodriguez. `PQser`: a Matlab package for spectral seriation. Numer. Algorithms, 2018;
- M. Fiedler. *Algebraic connectivity of graphs*. Czech.Math.Journal, 23(1973), pp. 298-305.

Thanks for your attention.

Grazie per l'attenzione.