The fifth international conference useR! 2009

Proximity data visualization with h-plots

Irene Epifanio

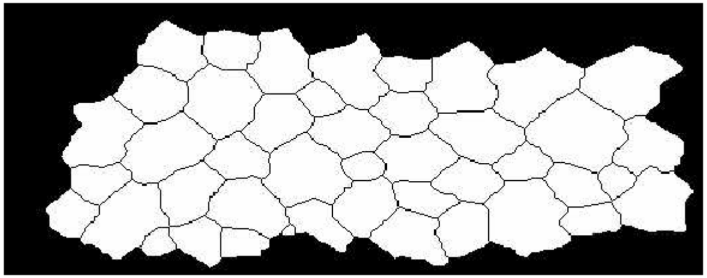
Dpt. Matemàtiques, Univ. Jaume I (SPAIN)

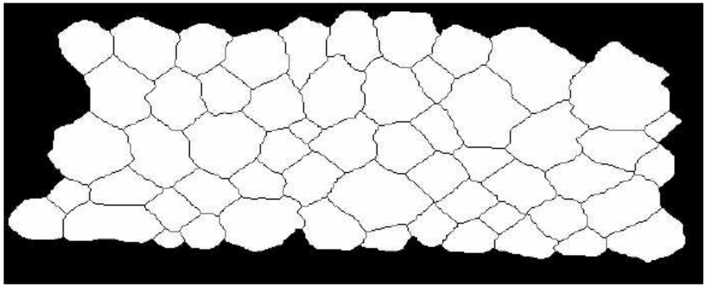
epifanio@uji.es; http://www3.uji.es/~epifanio

Outline

- Motivating problem
- Methodology
- Small-size examples
- Point patterns
- Conclusions

Motivating problem



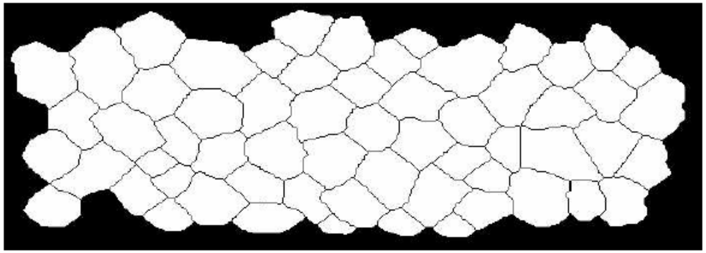


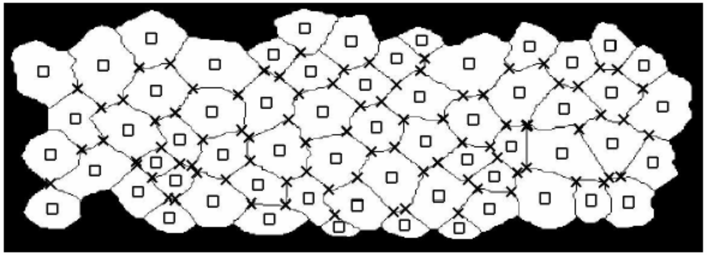
In Ayala et al. 2006: to find groups corresponding with different morphologies of the corneal endothelia

↓

Different dissimilarities (non-metric) between human corneal endothelia.

Motivating problem





Corneal endothelia described by bivariate point patterns (centroids and triple points).

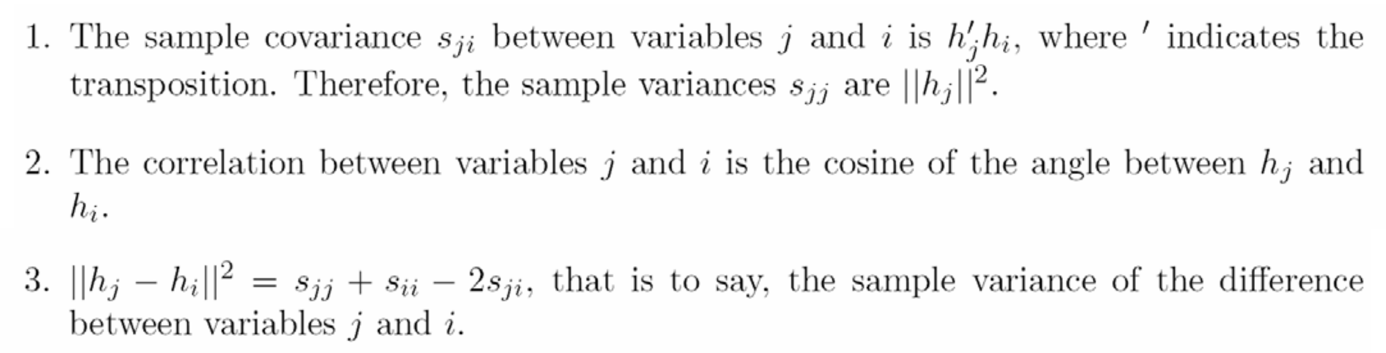
Different dissimilarities (triangle inequality is not hold) between point patterns.

Methodology: h-plot

X data matrix, S covariance matrix: λ_1, λ_2 largest eigenvalues, q_1, q_2 unit eigenvectors:

$$H_2 = (\sqrt{\lambda_1} q_1, \sqrt{\lambda_2} q_2)$$

Rows h_j of matrix H_2 have properties:



1. The sample covariance s_{ji} between variables j and i is $h'_j h_i$, where $'$ indicates the transposition. Therefore, the sample variances s_{jj} are $||h_j||^2$.

2. The correlation between variables j and i is the cosine of the angle between h_j and h_i.

3. $||h_j - h_i||^2 = s_{jj} + s_{ii} - 2s_{ji}$, that is to say, the sample variance of the difference between variables j and i.

Methodology: h-plot

We do not have a classical data matrix, but a dissimilarity matrix, *D*: d_{ij} represents the dissimilarity from the object *i* to object *j*.

Asymmetric relationship ($d_{ij} \neq d_{ji}$): we can consider the variable measuring dissimilarity from j to other objects ($d_{j.}$) and the dissimilarity to j ($d_{.j}$).

With a symmetric dissimilarity ($d_{j.} = d_{.j}$): variable *j* represents dissimilarity with respect *j.*

Euclidean distance between h_j and h_i in h-plot is sample standard deviation of difference between variables $d_{j.}$ and $d_{i.}$.

If these variables are similar, their difference, and therefore, its standard deviation will be small.

Comparison

- Classical Metric Multidimensional (cmdscale)
- Isomap (Tenenbaum et al., 2000)
- Kruskal's Non-metric Multidimensional Scaling (isoMDS) and Sammon's Non-Linear Mapping (sammon): Library MASS.

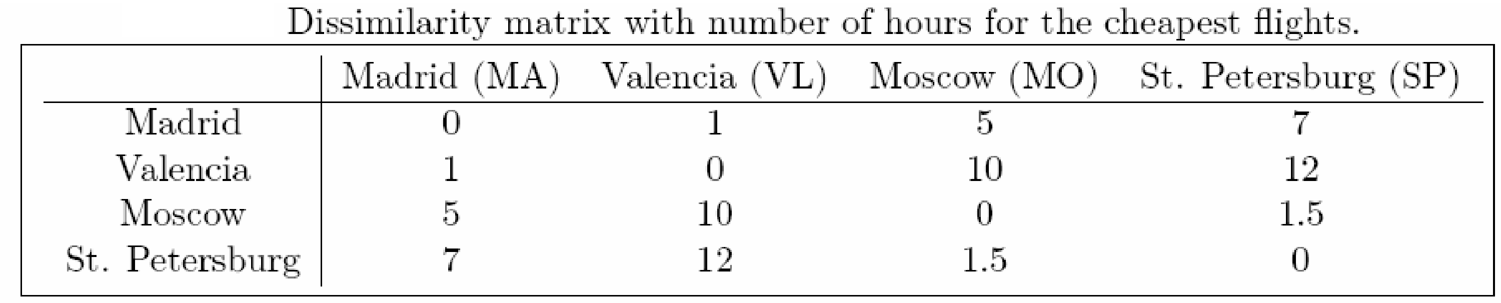
Congruence coefficient (0-1): similarity of two configurations X and Y.

$$c(X,Y) = \frac{\sum_{i<j} d_{ij}(X) d_{ij}(Y)}{(\sum_{i<j} d_{ij}^2(X))^{1/2} (\sum_{i<j} d_{ij}^2(Y))^{1/2}}$$

1 is achieved if X and Y are perfectly similar geometrically (match by rigid motions and dilations).

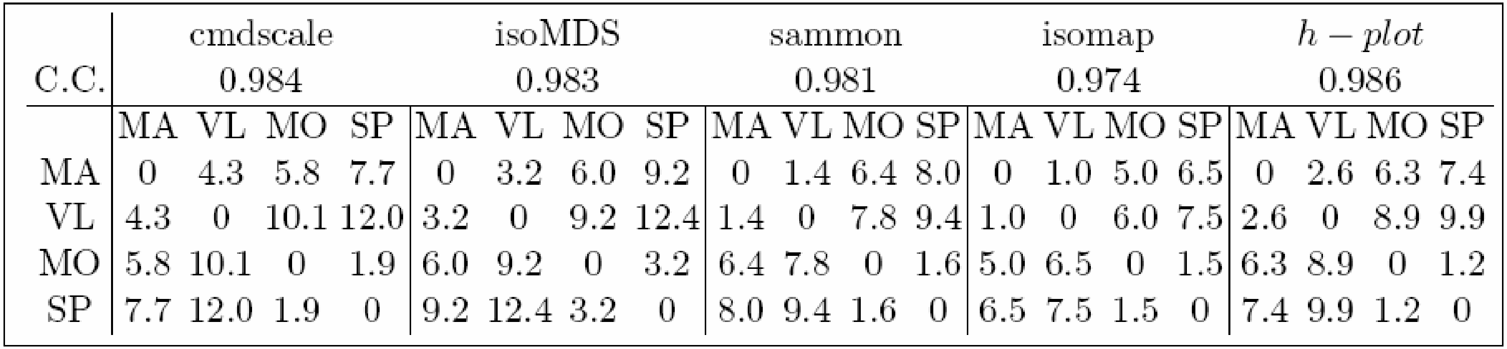
Example 1

If triangle inequality is not hold, although d_{ij} is small, variables $d_{j.}$ and $d_{i.}$ can be very different, and the objects *i* and *j* should not be represented near.



Dissimilarity matrix with number of hours for the cheapest flights.

	Madrid (MA)	Valencia (VL)	Moscow (MO)	St. Petersburg (SP)
Madrid	0	1	5	7
Valencia	1	0	10	12
Moscow	5	10	0	1.5
St. Petersburg	7	12	1.5	0

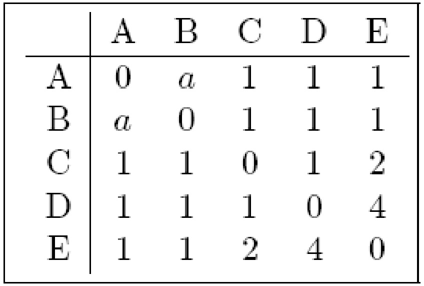


	cmdscale				isoMDS				sammon				isomap				$h-plot$			
C.C.	0.984				0.983				0.981				0.974				0.986			
	MA	VL	MO	SP	MA	VL	MO	SP	MA	VL	MO	SP	MA	VL	MO	SP	MA	VL	MO	SP
MA	0	4.3	5.8	7.7	0	3.2	6.0	9.2	0	1.4	6.4	8.0	0	1.0	5.0	6.5	0	2.6	6.3	7.4
VL	4.3	0	10.1	12.0	3.2	0	9.2	12.4	1.4	0	7.8	9.4	1.0	0	6.0	7.5	2.6	0	8.9	9.9
MO	5.8	10.1	0	1.9	6.0	9.2	0	3.2	6.4	7.8	0	1.6	5.0	6.5	0	1.5	6.3	8.9	0	1.2
SP	7.7	12.0	1.9	0	9.2	12.4	3.2	0	8.0	9.4	1.6	0	6.5	7.5	1.5	0	7.4	9.9	1.2	0

Example 2

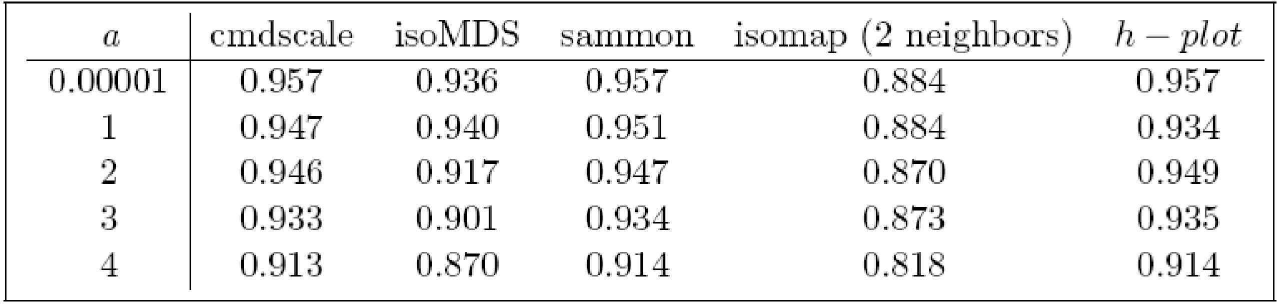
The observed values for variables $d_{j.}$ and $d_{i.}$ coincide, but d_{ij} is not zero, therefore the observed difference between $d_{j.}$ and $d_{i.}$ is zero for all the observed objects, except for the objects i and j.

Dissimilarity matrix between five brands.



	A	B	C	D	E
A	0	a	1	1	1
B	a	0	1	1	1
C	1	1	0	1	2
D	1	1	1	0	4
E	1	1	2	4	0

Congruence coefficients for the different methods with brands, varying a.

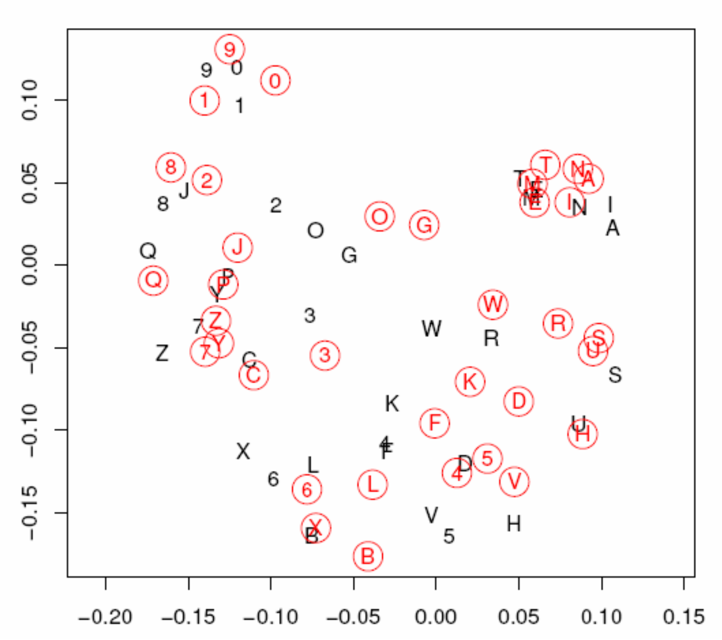


a	cmdscale	isoMDS	sammon	isomap (2 neighbors)	$h - plot$
0.00001	0.957	0.936	0.957	0.884	0.957
1	0.947	0.940	0.951	0.884	0.934
2	0.946	0.917	0.947	0.870	0.949
3	0.933	0.901	0.934	0.873	0.935
4	0.913	0.870	0.914	0.818	0.914

Example 3

Asymmetric data: d is not a distance. Even when $d_{jj} > 0$.

Dissimilarity formed by the variables giving the dissimilarity from each Morse code (i.e. $d_{i.}$, where code i-th is first presented), and the variables giving the dissimilarity to each Morse code (i.e. $d_{.i}$, where code i-th is second presented).



Point patterns: simulation

Same experiments considered in Ayala et al. (Clustering of spatial point patterns. Computational Statistics & Data Analysis. 50 (4) 1016-1032, 2006):

Three experiments for simulated Strauss processes with different parameters.

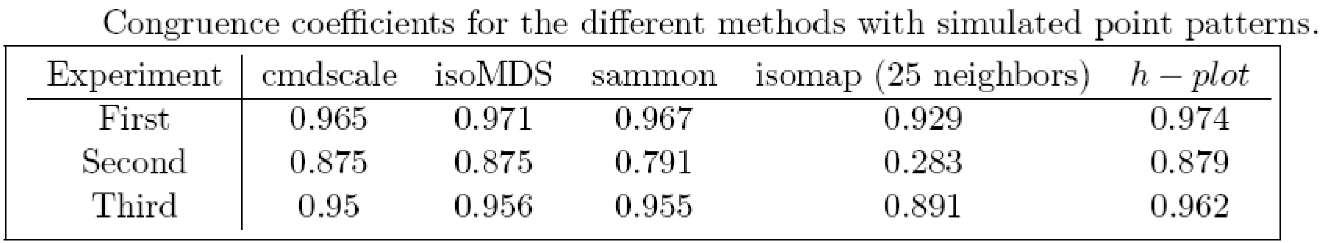
In each experiment, the same experimental setup: three different groups, each of them composed of 100 point patterns. Therefore, 3 dissimilarity matrices of 300x300.

Considered dissimilarity (based on the log rank statistic applied to the nearest-neighbor distances, Ayala et al. 2006) between point patterns is not a metric: triangle inequality is not hold.

Libraries of R used: Splancs; Spatstat and Survival.

UNIVERSITAT
JAUME•I

Point patterns: simulation

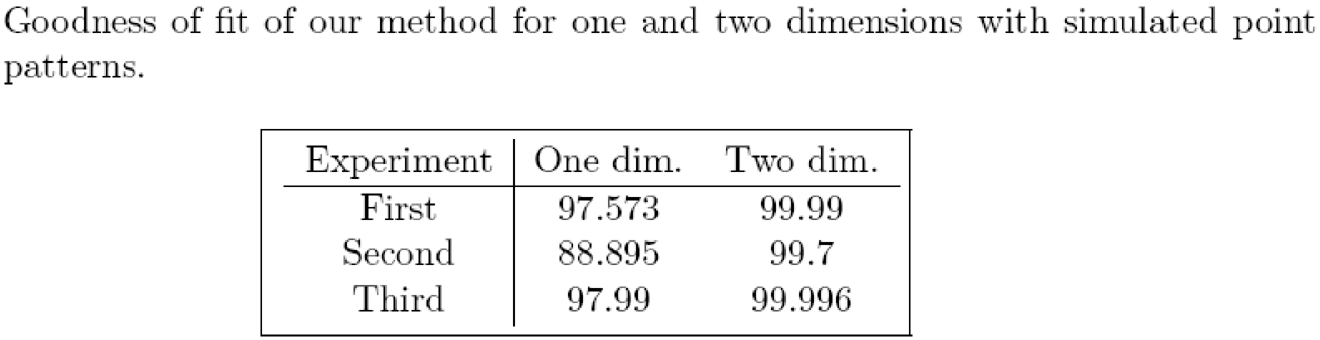


Congruence coefficients for the different methods with simulated point patterns.

Experiment	cmdscale	isoMDS	sammon	isomap (25 neighbors)	$h - plot$
First	0.965	0.971	0.967	0.929	0.974
Second	0.875	0.875	0.791	0.283	0.879
Third	0.95	0.956	0.955	0.891	0.962

Corsten and Gabriel (1976) goodness of fit for h-plotting in two dimensions:

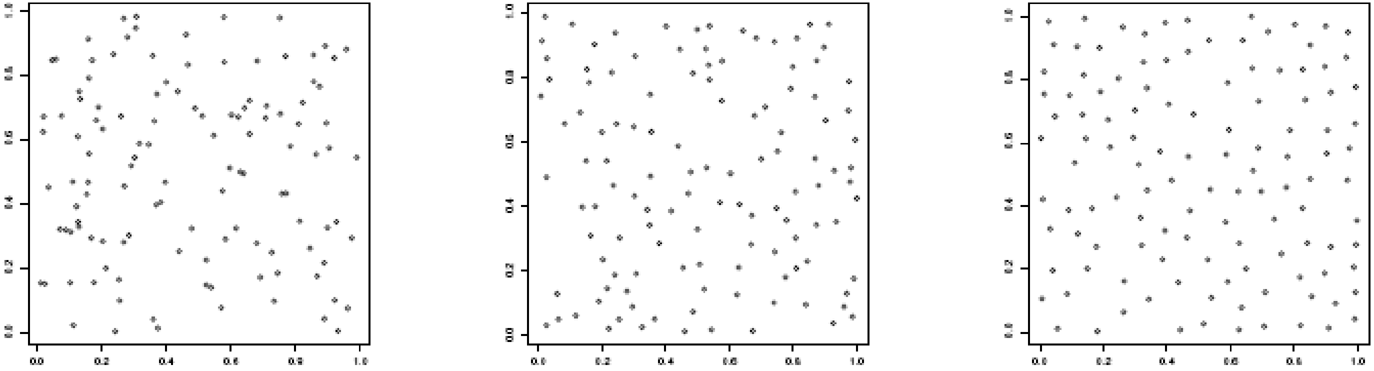
$$\boxed{(\lambda_1^2 + \lambda_2^2)/\sum_j \lambda_j^2}$$



Goodness of fit of our method for one and two dimensions with simulated point patterns.

Experiment	One dim.	Two dim.
First	97.573	99.99
Second	88.895	99.7
Third	97.99	99.996

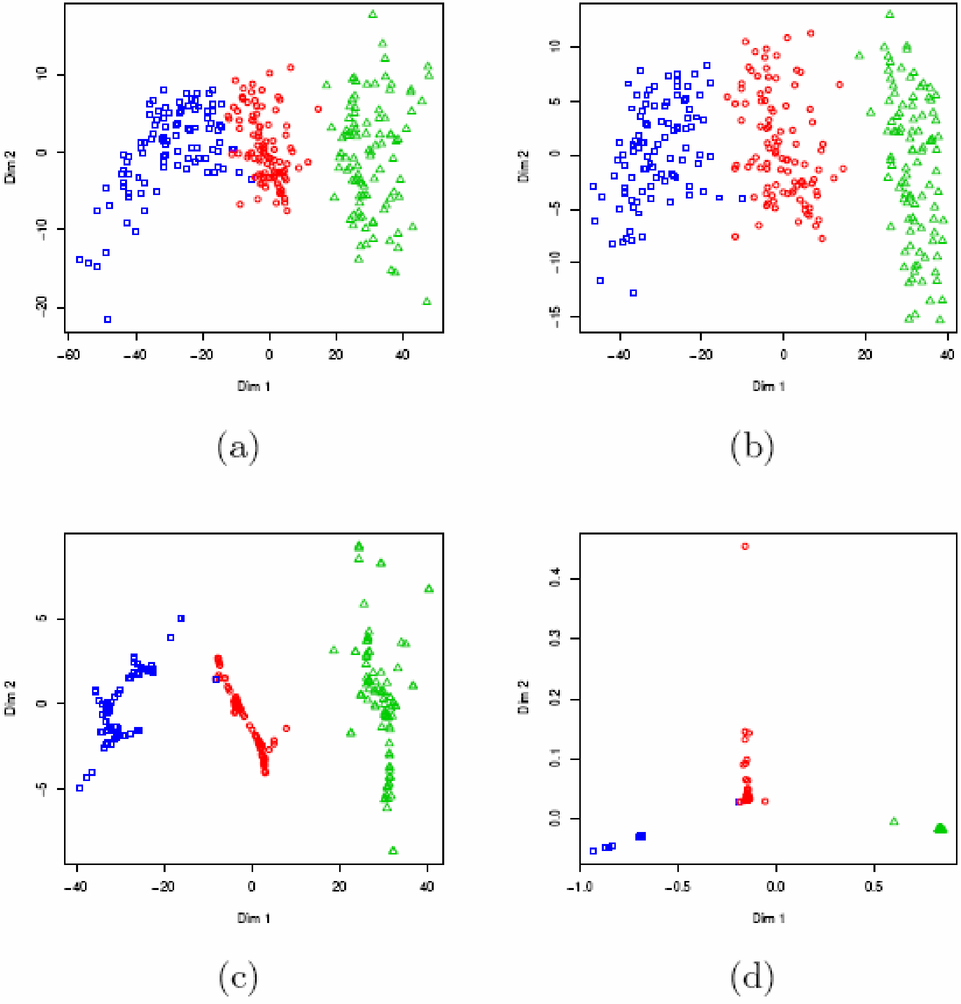
Point patterns: Experiment 1



One of the 100 point patterns generated for each group.

Note that we compute the dissimilarity between these point patterns, not inside them.

Point patterns: Experiment 1



(a)

(b)

(c)

(d)

a) Cmdscale

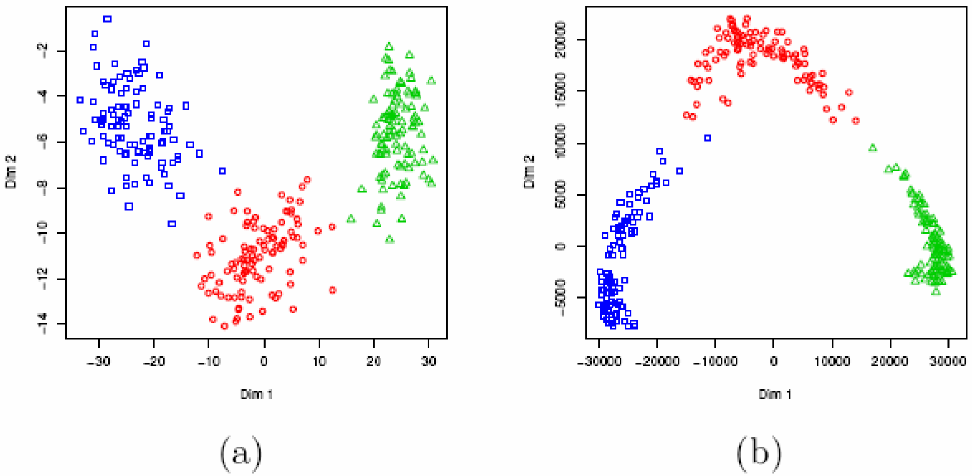
b) isoMDS

c) Sammon

d) Isomap (25 neighbors)

UNIVERSITAT JAUME•I

Point patterns: Experiment 1

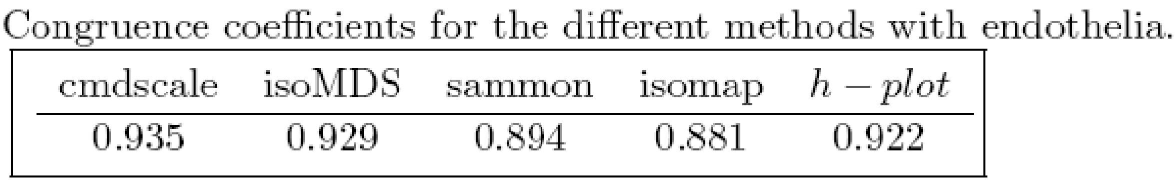


First experiment with $h-plot$ using: (a) the original dissimilarities, and (b) the dissimilarity ranks.

Besides the original dissimilarities, the ranking of the dissimilarities have been also considered (Seber 1984: if we have in mind cluster and pattern detection, then an expansion or contraction of the configuration could be more useful).

Point patterns: Endothelia

The dissimilarity matrix is made up of dissimilarities based on the log rank statistic applied to the nearest-neighbor distance between triple points (Ayala et al. 2006), for 153 individuals.

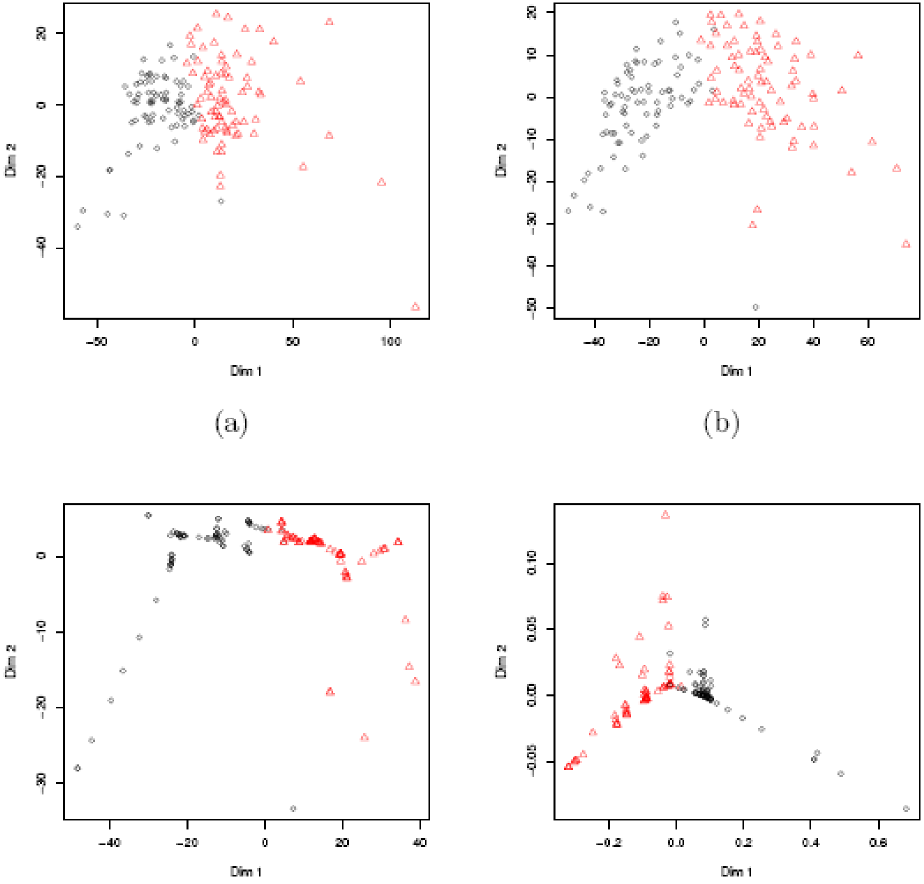


Congruence coefficients for the different methods with endothelia.

cmdscale	isoMDS	sammon	isomap	$h - plot$
0.935	0.929	0.894	0.881	0.922

The unhealthy cases obtained in (Ayala et al. 2006) are represented by red triangles, while black circles are healthy cases.

Point patterns: Endothelia



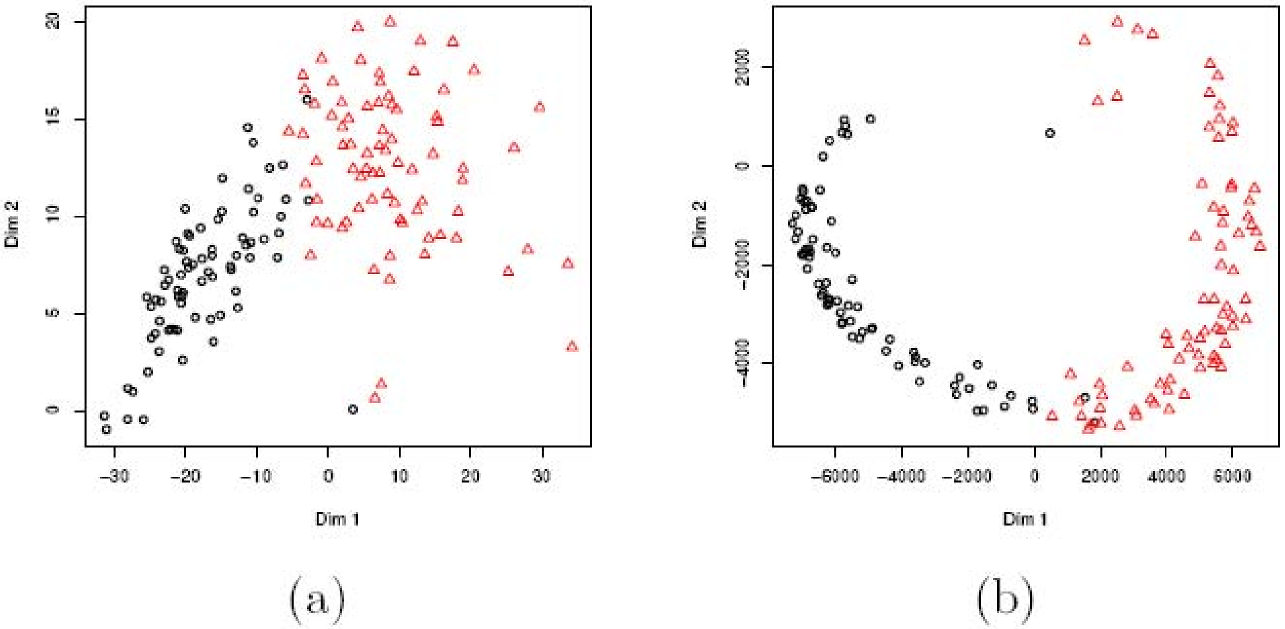
(a) (b)

(c) (d)

a) Cmdscale
b) isoMDS
c) Sammon
d) Isomap (25 neighbors)

UNIVERSITAT JAUME•I

Point patterns: Endothelia



(a) the original dissimilarities, and (b) the dissimilarity ranks.

Conclusions

- Alternative method for displaying dissimilarity matrices, based on h-plots.
- Good behavior through several examples (dissimilarity was not a metric).
- Non-iterative method, very simple to implement and computationally efficient.
- The representation goodness can also be easily assessed.
- It can also handle naturally asymmetric data.
- More illustrative results at:

  http://www3.uji.es/~epifanio/RESEARCH/hplot.pdf
- Future work: instead of second order differences between variables that indicates dissimilarity with respect to an object: higher order differences. Although the simplicity could be lost.

UNIVERSITAT JAUME•I

Thanks for your attention

UNIVERSITAT JAUME•I