Module 18 Multivariate Analysis for Genetic data
Session 06: Multidimensional Scaling

Jan Graffelman

jan.graffelman@upc.edu

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

26th Summer Institute in Statistical Genetics (SISG 2021)

Contents

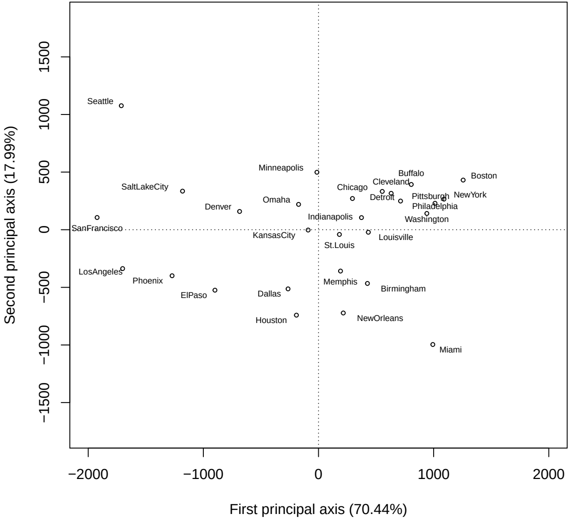
Multidimensional scaling

Objective

On the basis of information regarding the distances (or similarities) of n objects, construct a configuration of n points in a low-dimensional space (a map).

Classical application: inter-city distances (28 US cities)

	Birmingham	Boston	Buffalo	Chicago	Cleveland	Dallas	Denver	Detroit	ElPaso	Houston	Indianapolis	KansasCity	···
Birmingham	0	1194	947	657	734	653	1318	754	1278	692	492	703	···
Boston	1194	0	457	983	639	1815	1991	702	2358	1886	940	1427	···
Buffalo	947	457	0	536	192	1387	1561	252	1928	1532	510	997	···
Chicago	657	983	536	0	344	931	1050	279	1439	1092	189	503	···
Cleveland	734	639	192	344	0	1205	1369	175	1746	1358	318	815	···
Dallas	653	1815	1387	931	1205	0	801	1167	625	242	877	508	···
Denver	1318	1991	1561	1050	1369	801	0	1310	652	1032	1051	616	···
Detroit	754	702	252	279	175	1167	1301	0	1696	1312	290	760	···
ElPaso	1278	2358	1928	1439	1746	625	652	1696	0	756	1418	936	···
Houston	692	1886	1532	1092	1358	242	1032	1312	756	0	1022	750	···
Indianapolis	492	940	510	189	318	877	1051	290	1418	1022	0	487	···
KansasCity	703	1427	997	503	815	508	616	760	936	750	487	0	···
LosAngeles	2078	3036	2606	2112	2424	1425	1174	2369	800	1556	2096	1609	···
Louisville	378	996	571	305	379	865	1135	378	1443	981	114	519	···
Memphis	249	1345	965	546	773	470	1069	756	1095	586	466	454	···
Miami	777	1539	1445	1390	1325	1332	2094	1409	1957	1237	1225	1479	···
Minneapolis	1067	1402	955	411	763	969	867	698	1353	1211	600	466	···
NewOrleans	347	1541	1294	947	1102	504	1305	1101	1121	365	839	839	···
NewYork	983	213	436	840	514	1604	1780	671	2147	1675	729	1216	···
Omaha	907	1458	1011	493	819	661	559	754	1015	903	590	204	···
Philadelphia	894	304	383	758	432	1515	1698	589	2065	1586	647	1134	···
Phoenix	1680	2664	2234	1729	2052	1027	836	1986	402	1158	1713	1226	···
Pittsburgh	792	597	219	457	131	1237	1411	288	1778	1395	360	847	···
St.Louis	508	1179	749	293	567	638	871	529	1179	799	239	255	···
SaltLakeCity	1805	2425	1978	1458	1786	1239	512	1721	877	1465	1545	1128	···
SanFrancisco	2385	3179	2732	2212	2540	1765	1266	2475	1202	1958	2299	1882	···
Seattle	2612	3043	2596	2052	2404	2122	1373	2339	1760	2348	2241	1909	···
Washington	751	440	386	695	369	1372	1635	526	1997	1443	565	1071	···


Seattle
SaltLakeCity
SanFrancisco
LosAngeles
Phoenix
ElPaso
Denver
Minneapolis
Omaha
KansasCity
Dallas
Houston
Chicago
Indianapolis
St.Louis
Memphis
NewOrleans
Birmingham
Louisville
Cleveland
Detroit
Buffalo
Pittsburgh
Philadelphia
Washington
NewYork
Boston
Miami
Second principal axis (17.99%)
First principal axis (70.44%)
−2000
−1000
0
1000
2000
1500
1000
500
0
−500
−1000
−1500

Some basic terminology

Terminology

- proximity
- similarity (s_{rs})
- dissimilarity or distance (d_{rs})

A similarity measure satisfies:

- $s(A, B) = s(B, A)$
- $s(A, B) > 0$
- $s(A, B)$ increases as the similarity between A and B increases

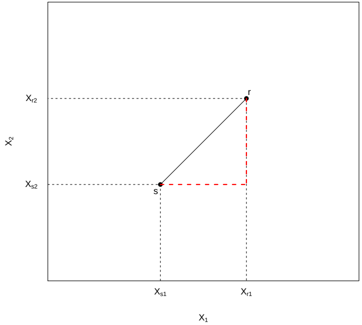
A distance measure, $\delta(A, B)$ satisfies:

- $\delta(A, B) = \delta(B, A)$
- $\delta(A, B) \geq 0$
- $\delta(A, A) = 0$

The distance function $\delta(A, B)$ called a metric if also

- $\delta(A, B) = 0$ iff $A = B$
- the triangle inequality holds: $\delta(A, B) \leq \delta(A, C) + \delta(C, B)$.

Euclidean Distance



$$\delta_{rs}^2 = (x_{r1} - x_{s1})^2 + (x_{r2} - x_{s2})^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)$$

Generalizes to p variables.

Some dissimilarity measures (quantitative data)

- Euclidean distance:

$$\delta_{rs} = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)} = \left\{ \sum_{i=1}^{p} (x_{ri} - x_{si})^2 \right\}^{\frac{1}{2}}$$

- Mahalanobis distance:

$$\delta_{rs} = \left\{ (\mathbf{x}_r - \mathbf{x}_s)' \mathbf{S}^{-1} (\mathbf{x}_r - \mathbf{x}_s) \right\}^{\frac{1}{2}}$$

- Minkowski distance

$$\delta_{rs} = \left\{ \sum_{i=1}^{p} |x_{ri} - x_{si}|^{\lambda} \right\}^{\frac{1}{\lambda}}$$

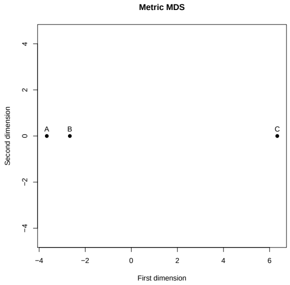
Metric versus Non-metric MDS

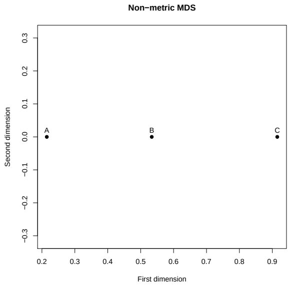
- In metric MDS, the configuration of points is directly obtained from the distances.
- In non-metric MDS, only the rank order of the distances is important.

- $d_{rs} \approx \delta_{rs}$: Classical scaling.
- $d_{rs} \approx f(\delta_{rs})$ with $f(\delta_{rs}) = \alpha + \beta\delta_{rs}$: Metric scaling.
- $d_{rs} \approx f(\delta_{rs})$ with $f(\delta_{rs})$ arbitrary, monotone: Non-metric scaling.

Metric versus Non-metric MDS

	A	B	C
A	0	1	10
B	1	0	9
C	10	9	0





Metric MDS

- Also known as: classical scaling, principal coordinate analysis (PCO).
- Given n objects with dissimiliarities (δ_{rs}) find a set of points in Euclidean space such that $d_{rs} \approx \delta_{rs}$.
- Classical application: given a distance matrix (in km or in travel time) between cities, construct a map of the cities.

Theory (1)

Let $\mathbf{X}$ be the matrix of coordinates with the solution.
$\mathbf{x}_r, \mathbf{x}_s$ two rows of $\mathbf{X}$.

$$\delta_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)$$

Let $\mathbf{B}$ be the inner product matrix with

$$b_{rs} = \mathbf{x}_r{}'\mathbf{x}_s$$

Assume the solution to be centered at the origin:

$$\sum_{r=1}^{n} x_{ri} = 0$$

Theory (2)

$$d_{rs}^2 = \mathbf{x}_r{}'\mathbf{x}_r + \mathbf{x}_s{}'\mathbf{x}_s - 2\mathbf{x}_r{}'\mathbf{x}_s$$

$$\frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 = \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r{}'\mathbf{x}_r + \mathbf{x}_s{}'\mathbf{x}_s$$

$$\frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 = \mathbf{x}_r{}'\mathbf{x}_r + \frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s{}'\mathbf{x}_s$$

$$\frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = \frac{2}{n}\sum_{r=1}^{n} \mathbf{x}_r{}'\mathbf{x}_r$$

Theory (3)

Let $b_{rs} = \mathbf{x}_r{}'\mathbf{x}_s = -\frac{1}{2}\left(d_{rs}^2 - \mathbf{x}_r{}'\mathbf{x}_r - \mathbf{x}_s{}'\mathbf{x}_s\right)$

$$b_{rs} = -\frac{1}{2}\left(d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right).$$

We define $a_{rs} = -\frac{1}{2}d_{rs}^2$ so that $b_{rs} = a_{rs} - a_{r.} - a_{.s} + a_{..}$
and build matrix **A**

$$\mathbf{B} = \mathbf{HAH} \quad \mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

and

$$\mathbf{B} = \mathbf{X}\mathbf{X}'$$

We wish to approximate **B** in a low dimensional space.
We approximate the distance matrix indirectly, via de matrix of scalar products.

Theory (4) Spectral Decomposition

Let $\mathbf{B}$ be any $n \times n$ symmetric matrix we want to approximate

$$\mathbf{B} = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}' = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

with $\mathbf{D}_\lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{V} = [\mathbf{v}_i, \ldots, \mathbf{v}_n]$

$$\tilde{\mathbf{B}} = \mathbf{V}_{(,1:k)}\mathbf{D}_{\lambda(1:k,1:k)}\left(\mathbf{V}_{(,1:k)}\right)'$$

gives the rank k least squares approximation to $\mathbf{B}$

Theory (5) Solution

$$\mathbf{B} = \mathbf{X}\mathbf{X}' = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$$

The coordinates of the solution are obtained as:

$$\mathbf{X} = \mathbf{V}\mathbf{D}_\lambda^{\frac{1}{2}}$$

Notes:

- There will always be at least one eigenvalue equal to zero.
- There is nesting of the solution.

Algorithm for Classical Scaling

- Compute a distance or dissimilarity matrix.
- Compute $[a_{rs}] = -\frac{1}{2}\delta_{rs}^2$
- Double center $\mathbf{A}$ to obtain $\mathbf{B} = \mathbf{HAH}$
- Compute eigenvalues and eigen vectors of $\mathbf{B}$
- Compute the solution as $\mathbf{X} = \mathbf{V}\mathbf{D}_{\lambda}^{\frac{1}{2}}$

Goodness of Fit

How well do we manage to approximate the distance matrix?

$$\frac{\sum_{i=1}^{P} \lambda_i}{\sum_{i=1}^{n-1} \lambda_i}$$

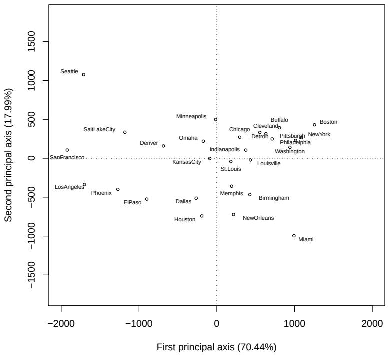
If **B** is not positive semi-definite:

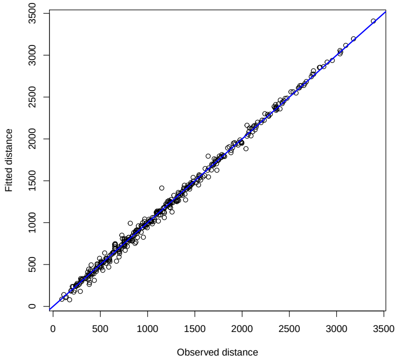
$$\frac{\sum_{i=1}^{P} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|} \quad \text{or} \quad \frac{\sum_{i=1}^{P} \lambda}{\sum_{\lambda_i > 0} \lambda_i}$$

Eigenvalues for US cities

	λ_i	$\lvert\lambda_i\rvert / \sum \lvert\lambda_i\rvert$	Cum.
1	22191639.31	0.70	0.70
2	5666889.03	0.18	0.88
3	631806.72	0.02	0.90
4	349072.27	0.01	0.92
5	320061.81	0.01	0.93
6	151990.13	0.00	0.93
7	135990.15	0.00	0.93
8	96223.40	0.00	0.94
9	41151.80	0.00	0.94
10	32504.38	0.00	0.94
11	25294.02	0.00	0.94
12	17104.35	0.00	0.94
13	11261.65	0.00	0.94
14	6491.34	0.00	0.94
15	0.00	0.00	0.94
16	-3667.46	0.00	0.94
17	-9367.48	0.00	0.94
18	-15622.89	0.00	0.94
19	-30669.91	0.00	0.94
20	-43411.51	0.00	0.95
21	-44714.98	0.00	0.95
22	-63959.28	0.00	0.95
23	-84957.57	0.00	0.95
24	-91580.01	0.00	0.95
25	-164686.47	0.01	0.96
26	-193735.06	0.01	0.97
27	-349548.74	0.01	0.98
28	-731051.32	0.02	1.00

Diagnostic plot





Euclidean Distance matrix

- Definition

  A distance matrix $\mathbf{D}$ is called Euclidean if there exists a configuration of points in Euclidean space whose interpoint distances are given by $\mathbf{D}$. That is, for some p there exists points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ such that $d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)$.

- Theorem

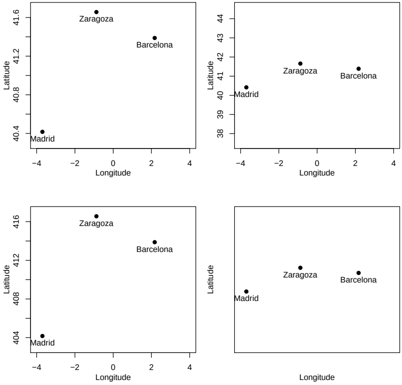
  A distance matrix $\mathbf{D}$ is Euclidean if and only if $\mathbf{B}$ ($= \mathbf{HAH}$, as previously defined) is positive semi definite.

About plotting maps

	Latitude (°)	Longitude (°)
Zaragoza	41.66	-0.88
Barcelona	41.39	2.17
Madrid	40.42	-3.70

	Zaragoza	Barcelona	Madrid
Zaragoza	0.00	158.67	170.25
Barcelona	158.67	0.00	313.74
Madrid	170.25	313.74	0.00

Distances in km.



Non-metric MDS: objective function

- $\text{STRESS} = \sqrt{\frac{\sum_{r \neq s}^{n} (f(\delta_{rs}) - d_{rs})^2}{\sum_{r \neq s} d_{rs}^2}}$
- $stress(\Delta, \hat{\mathbf{X}}) = \min_{all \mathbf{X}} \quad stress(\Delta, \mathbf{X})$
- We minimize the objective function numerically, starting from an initial configuration.
- Goodness-of-fit:

Stress	fit
20%	poor fit
10%	fair fit
5%	good fit
0%	perfect fit

Global and local minima; diagnostics

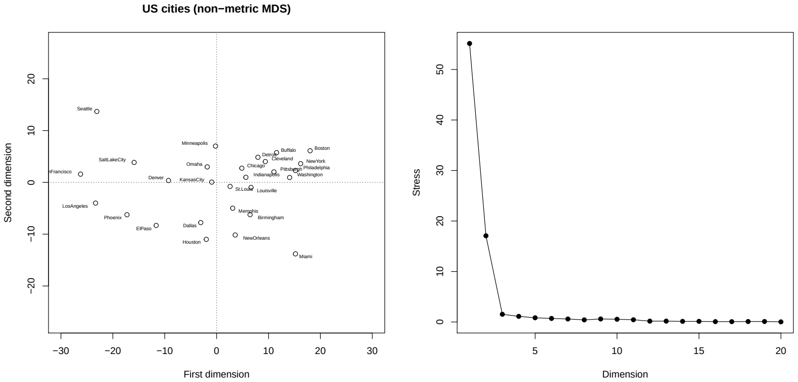
How do you know if the solution corresponds to a local or global minimum?

- Use different initial configurations
- Compare stress over 1,2,3,... dimensional solutions

General diagnostics:

- Scatter plot of δ_{rs} versus d_{rs}
- Plot stress versus number of dimensions
- Degeneracy (many points with the same d_{rs})
- Compute residuals $(d_{rs} - f(\delta_{rs}))$

Non-metric MDS map of US cities



MDS with genetic data

- There is a rich literature on how to measure genetic distance
- The allele sharing distance is an often used measure

$i \backslash j$	AA	AB	BB
AA	2	1	0
AB	1	2	1
BB	0	1	2

$i \backslash j$	AA	AB	BB
AA	0	1	2
AB	1	0	1
BB	2	1	0

- Let x_{ijk} be the number of shared alleles of individual i and j for variant k
- $d_{ijk} = 2 - x_{ijk}$
- Often scaled by multiplying by $\frac{1}{2}$
- Typically averaged over K genetic variants:

$$d_{ij} = \frac{1}{K} \sum_{k=1}^{K} d_{ijk}$$

- The so obtained $\mathbf{D} = [d_{ij}]$ is used as input for MDS.

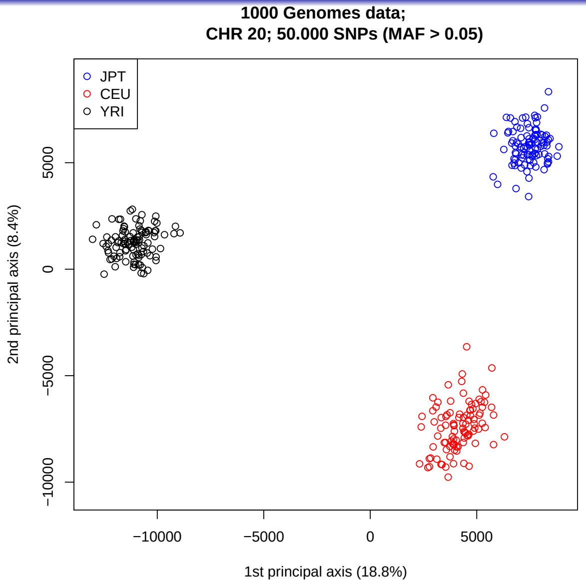
Genetic data of the 1000 Genomes project

	rs6078030	rs552482647	rs4814683	rs6076506	rs6139074	
1	0	0	0	0	0	⋯
2	2	0	2	0	2	⋯
3	0	0	1	0	0	⋯
4	0	0	0	0	0	⋯
5	0	0	0	0	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

	1	2	3	4	5	
1	0.00	0.43	0.44	0.47	0.43	⋯
2	0.43	0.00	0.44	0.46	0.48	⋯
3	0.44	0.44	0.00	0.46	0.44	⋯
4	0.47	0.46	0.46	0.00	0.45	⋯
5	0.43	0.48	0.44	0.45	0.00	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

We consider 310 individuals for 50,000 SNPs

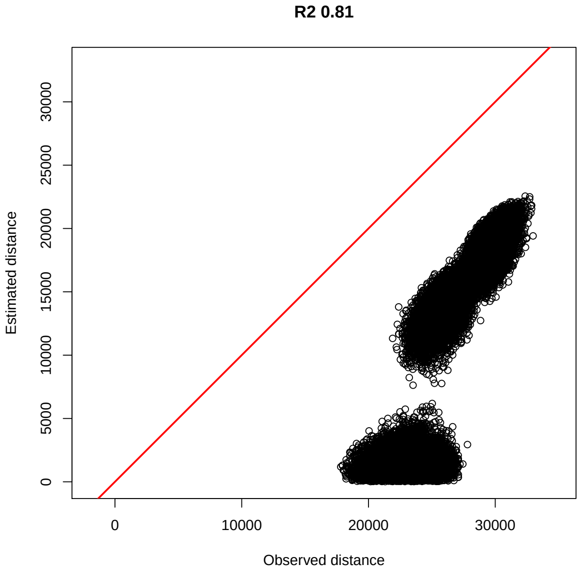
MDS with genetic data (CEU, JPT and YRI from the 1000 Genomes project)



Goodness of fit

Dim.	λ	%	% Cum.
1	8.3567	0.1816	0.1816
2	3.7353	0.0812	0.2628
3	0.6275	0.0136	0.2764
4	0.5274	0.0115	0.2879
5	0.4909	0.0107	0.2985
6	0.4675	0.0102	0.3087
7	0.4540	0.0099	0.3186
8	0.4493	0.0098	0.3283
9	0.4345	0.0094	0.3378
10	0.4211	0.0091	0.3469
.	.	.	.
.	.	.	.
.	.	.	.
260	0.0005	0.0000	0.9840
261	0.0001	0.0000	0.9840
262	0.0000	0.0000	0.9840
263	-0.0008	0.0000	0.9841
264	-0.0011	0.0000	0.9841
265	-0.0017	0.0000	0.9841
.	.	.	.
.	.	.	.
.	.	.	.
306	-0.0289	0.0006	0.9972
307	-0.0297	0.0006	0.9979
308	-0.0305	0.0007	0.9986
309	-0.0321	0.0007	0.9993
310	-0.0343	0.0007	1.0000

Goodness of fit



Final notes

- In genetics, MDS is often used at an aggregate level.
- Some pairwise similarity or distance measure between populations is calculated (e.g. F_{st}, distance measures calculated from allele frequencies, etc.).
- MDS applied to between-population distances.
- MDS maps made at an individual level typically show more variability, often revealing overlap between populations.
- MDS widely used for detecting the existence of population substructure.

Bibliography

- Borg, I. & Groenen, P. (1997) Modern Multidimensional Scaling. Theory and Applications. Springer.
- Cox, T.F. & Cox, M.A. (2001) Multidimensional Scaling. Second edition. Chapman & Hall