Population substructure

Jan Graffelman[1]

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

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

jan.graffelman@upc.edu

November 28, 2018

Contents

1. Population substructure
2. Metric MDS
3. Non-metric MDS
4. Example
5. Computer exercise

Population substructure

- Genotype frequencies and allele frequencies vary over human populations.
- If data is a mixture of individuals from different populations, spurious associations may result.
- If the subpopulations are known then
  - a stratified analysis may be more adequate
  - account for population substructure by defining a covariate
- How to detect population substructure?

Consequences of population substructure

- Population substructure can influence many types of analysis in statistical genetics.
- It can affect tests for HWE.
- It can affect tests for LD.
- It can affect marker-trait association tests.
- ...

Population substructure and HWE

Let there be two populations, and consider one polymorphism. The polymorphism has allele frequency $p_1 = 0.3$ in the first population, and allele frequency $p_2 = 0.8$ in the second population. Let there be 300 individuals in each population ($n_1 = n_2 = 300$). We assume Hardy-Weinberg equilibrium within each population. Then

Pop 1	A	B	
A	$300 \cdot 0.3^2 = 27$	$300 \cdot 0.3 \cdot 0.7 = 63$	90
B	$300 \cdot 0.3 \cdot 0.7 = 63$	$300 \cdot 0.7^2 = 147$	210
	90	210	300

Pop 2	A	B	
A	$300 \cdot 0.8^2 = 192$	$300 \cdot 0.2 \cdot 0.8 = 48$	240
B	$300 \cdot 0.2 \cdot 0.8 = 48$	$300 \cdot 0.2^2 = 12$	60
	240	60	300

Joint	A	B	
A	$27 + 192 = 219$	$63 + 48 = 111$	330
B	$63 + 48 = 111$	$147 + 12 = 159$	270
	330	270	600

Chi-square tests

```
library(HardyWeinberg)
> x1
[1]  27 126 147
> out1 <- HWChisq(x1,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  5.962667e-30 p-value =  1 D =  0
> x2
[1] 192  96  12
> out2 <- HWChisq(x2,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  0 p-value =  1 D =  0
> x3 <- x1+x2
> x4 <- x3/2
> x4
[1] 109.5 111.0  79.5
> out4 <- HWChisq(x4,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  19.1307 p-value =  1.220655e-05 D =  -18.75
```

Population substructure and LD

Let there be two populations, and consider two polymorphisms, A/a and B/b. In the first population we have $p_A = 0.7$ and $p_B = 0.6$. In the second population we have $p_A = 0.3$ and $p_B = 0.9$. Let there be 100 individuals (200 haplotypes) in each population ($n_1 = n_2 = 100$). We assume linkage equilibrium within each population. Then

Pop 1	B	b	
A	$200 \cdot 0.7 \cdot 0.6 = 84$	$200 \cdot 0.7 \cdot 0.4 = 56$	140
a	$200 \cdot 0.3 \cdot 0.6 = 36$	$200 \cdot 0.3 \cdot 0.4 = 24$	60
	120	80	200

Pop 2	B	b	
A	$200 \cdot 0.3 \cdot 0.9 = 54$	$200 \cdot 0.3 \cdot 0.1 = 6$	60
a	$200 \cdot 0.7 \cdot 0.9 = 126$	$200 \cdot 0.7 \cdot 0.1 = 14$	140
	180	20	200

Joint	B	b	
A	$84 + 54 = 138$	$56 + 6 = 62$	200
a	$36 + 126 = 162$	$24 + 14 = 38$	200
	300	100	400

Chi-square tests

```
> X1
     [,1] [,2]
[1,]   84   56
[2,]   36   24
> out <- chisq.test(X1,correct=FALSE)
> print(out)

Pearson's Chi-squared test

data:  X1
X-squared = 0, df = 1, p-value = 1
> X2
     [,1] [,2]
[1,]   54    6
[2,]  126   14
> out <- chisq.test(X2,correct=FALSE)
> print(out)

Pearson's Chi-squared test

data:  X2
X-squared = 0, df = 1, p-value = 1
```

Chi-square tests

```
> X4 <- (X1+X2)/2
> X4
     [,1] [,2]
[1,]   69   31
[2,]   81   19
> out <- chisq.test(X4,correct=FALSE)
> print(out)

Pearson's Chi-squared test

data:  X4
X-squared = 3.84, df = 1, p-value = 0.05004
```

How to detect substructure

- Principal component analysis of the marker data
- Multidimensional scaling of distance matrix computed from the marker data
- ...
- In the remainder of this module we focus on MDS.

Multidimensional scaling

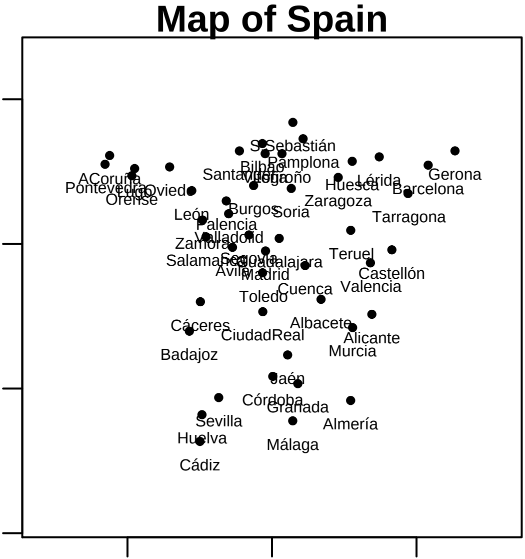
Objective

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**).

Example data set

	Albacete	Alicante	Almera	Avila	Badajoz	Barcelona	Bilbao	Burgos	· · ·
Albacete	0	171	369	366	525	540	646	488	· · ·
Alicante	171	0	294	537	696	515	817	659	· · ·
Almera	369	294	0	663	604	809	958	800	· · ·
Avila	366	537	663	0	318	717	401	243	· · ·
Badajoz	525	696	604	318	0	1022	694	536	· · ·
Barcelona	540	515	809	717	1022	0	620	583	· · ·
Bilbao	646	817	958	401	694	620	0	158	· · ·
Burgos	488	659	800	243	536	583	158	0	· · ·
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Download SpainDist.dat


Map of Spain
S.Sebastián
Bilbao
Pamplona
Santander
Vitoria
Logroño
ACoruña
Pontevedra
Lugo
Oviedo
Orense
Huesca
Lérida
Gerona
Barcelona
Zaragoza
Burgos
Soria
León
Palencia
Tarragona
Valladolid
Zamora
Teruel
Salamanca
Segovia
Guadalajara
Avila
Madrid
Castellón
Cuenca
Valencia
Toledo
Cáceres
Albacete
CiudadReal
Alicante
Badajoz
Murcia
Jaén
Córdoba
Granada
Sevilla
Almería
Huelva
Málaga
Cádiz

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: $\delta(A, B) \leq \delta(A, C) + \delta(C, B)$.

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

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

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.

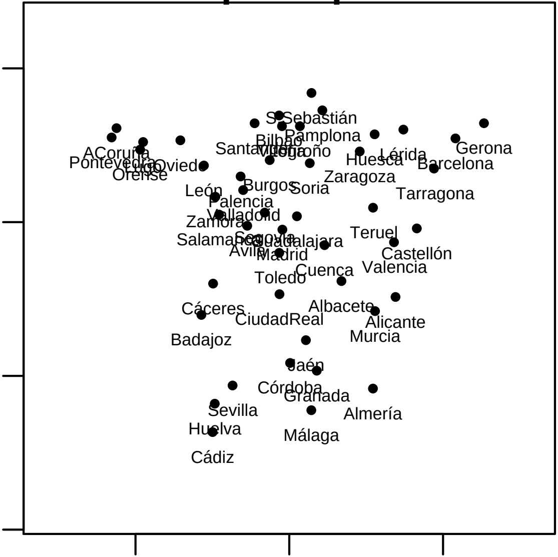
- $d_{rs} \approx \delta_{rs}$: Classical scaling.
- $d_{rs} \approx f(\delta_{rs})$ with $f(\delta_{rs}) = \alpha + \beta\delta_{rs}$: Metric scaling.
- $d_{rs} \approx f(\delta_{rs})$ with $f(\delta_{rs})$ arbitrary, monotone: Non-metric scaling.

Metric 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, contruct a map of the cities.

	Albacete	Alicante	Almera	Avila	Badajoz	Barcelona	Bilbao	Burgos	···
Albacete	0	171	369	366	525	540	646	488	···
Alicante	171	0	294	537	696	515	817	659	···
Almera	369	294	0	663	604	809	958	800	···
Avila	366	537	663	0	318	717	401	243	···
Badajoz	525	696	604	318	0	1022	694	536	···
Barcelona	540	515	809	717	1022	0	620	583	···
Bilbao	646	817	958	401	694	620	0	158	···
Burgos	488	659	800	243	536	583	158	0	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Map of Spain


S.Sebastián
Pamplona
Bilbao
Santander
Vitoria
Logroño
Huesca
Lérida
Gerona
Barcelona
ACoruña
Pontevedra
Lugo
Oviedo
Orense
Zaragoza
Burgos
Soria
León
Tarragona
Palencia
Valladolid
Zamora
Teruel
Salamanca
Segovia
Guadalajara
Avila
Madrid
Castellón
Cuenca
Valencia
Toledo
Albacete
Cáceres
CiudadReal
Alicante
Murcia
Badajoz
Jaén
Córdoba
Granada
Almería
Sevilla
Huelva
Málaga
Cádiz

Theory (1)

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

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

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 $\mathbf{A}$

$$\mathbf{B} = \mathbf{HAH} \quad \mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{11}'$$

and

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

We wish to approximate $\mathbf{B}$ in a low dimensional space.

Theory (4) Spectral Decomposition

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

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

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

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

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:

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

Note: there will always be at least one eigenvalue equal to zero.

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:

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

Euclidean Distance matrix

- Definition

  A distance matrix **D** is called **Euclidean** if there exists a configuration of points in Euclidean space whose interpoint distances are given by **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 **D** is Euclidean if and only if **B** $(=$ **HAH**, as previously defined) is positive semi definite.

Similarity data

- Sometimes data are given in the form of similarities (c_{rs}).
- A similarity matrix $\mathbf{C}$ has $c_{rs} = c_{sr}$ and $c_{rs} \leq c_{rr}$.
- Similarities can be transformed into distances with the transformation $d_{rs} = \sqrt{c_{rr} - 2c_{rs} + c_{ss}}$
- If $\mathbf{C}$ is psd, then the obtained distance matrix will be Euclidean.

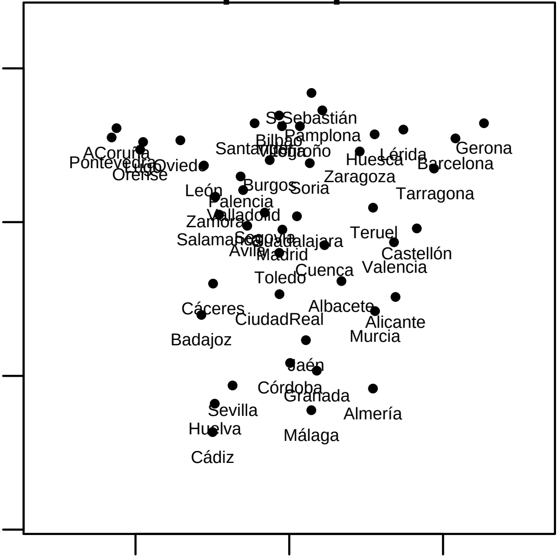
R code for classical scaling

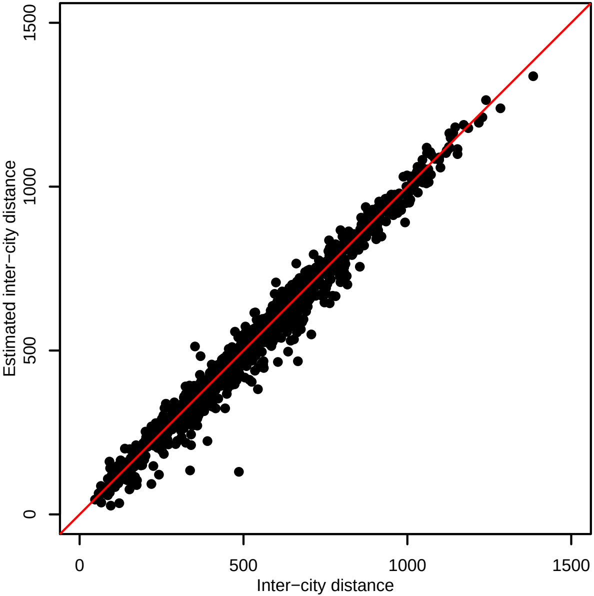
```
Spain <- as.matrix(read.table("http://www-eio.upc.es/˜jan/data/SpainDist.dat",
         header=TRUE))
rownames(Spain) <- colnames(Spain)
n <- nrow(Spain)

mds.out <- cmdscale(Spain,k=n-1,eig=TRUE)

X <- mds.out$points[,1:2]
plot(X[,2],X[,1],type="n", xlab="", ylab="", main="Map of Spain",asp=1,
     xlim=c(-800,800),ylim=c(-800,500))
points(X[,2],X[,1],pch=19,cex=0.5)
text(X[,2],X[,1],rownames(Spain), cex=0.5,pos=1)
```

R code for classical scaling

```
> ev <- mds.out$eig
> gof <- mds.out$GOF
> print(round(ev,digits=2))
 [1] 4419357.73 3710242.86  523390.06  222914.52  215904.45  143955.45
 [7]  128021.63  103602.38   92361.07   77669.80   67866.94   55724.33
[13]   51347.16   38327.38   32347.58   29609.07   18785.64   14974.46
[19]    9473.34    9317.99    6911.58    4219.73    1459.24     105.43
[25]       0.00    -854.58   -3724.49   -4557.54   -5306.92   -8958.67
[31]  -11879.05  -15217.83  -16867.79  -24417.22  -34120.67  -43608.19
[37]  -50334.85  -63916.60  -77134.54  -80754.15  -91612.38  -97422.06
[43] -120383.81 -125973.49 -179445.66 -253056.31 -340735.97
> print(round(gof,digits=4))
[1] 0.8581 1.0000
```

Map of Spain


S.Sebastián
Pamplona
Bilbao
Santander
Vitoria
Logroño
ACoruña
Pontevedra
Lugo
Oviedo
Orense
Huesca
Lérida
Gerona
Barcelona
Zaragoza
Burgos
Soria
León
Tarragona
Palencia
Valladolid
Zamora
Teruel
Salamanca
Segovia
Guadalajara
Avila
Madrid
Castellón
Cuenca
Valencia
Toledo
Albacete
Cáceres
CiudadReal
Alicante
Badajoz
Murcia
Jaén
Córdoba
Granada
Sevilla
Almería
Huelva
Málaga
Cádiz


Estimated inter−city distance
0
500
1000
1500
0
500
1000
1500
Inter−city distance

Non-metric MDS: objective function

- 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.

Procedure for Non-metric MDS

- Choose a distance measure (e.g. $\delta_{rs} = \left\{ \sum_{i=1}^{p} \mid x_{ri} - x_{si} \mid^{\lambda} \right\}^{\frac{1}{\lambda}}$)
- Choose a monotone transformation f
- Choose an algorithm to minimize Stress.

Global versus local minima

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

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 ($d_{rs} - f(\delta_{rs})$)

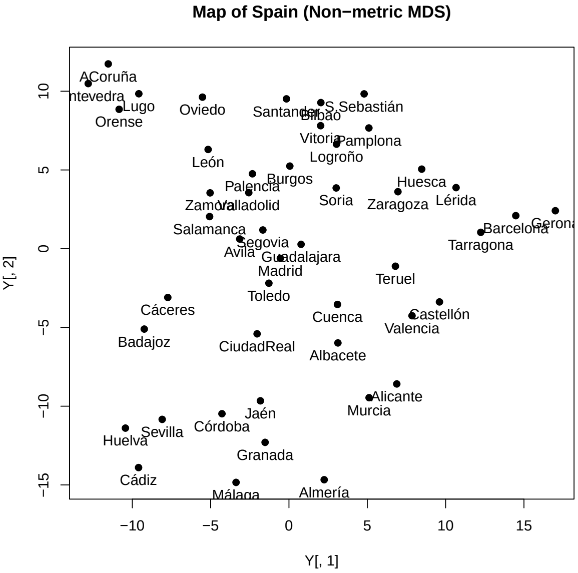
R code for non-metric MDS

```
> init <- scale(matrix(runif(n*2),ncol=2),scale=FALSE)
> nmmds.out <- isoMDS(Spain,y=init,k=2)
initial  value 41.659041
iter   5 value 40.219780
iter  10 value 37.286307
iter  15 value 30.177635
iter  20 value 22.661686
iter  25 value 14.483317
iter  30 value 10.703962
iter  35 value 7.756514
iter  40 value 6.116380
iter  45 value 5.360785
iter  50 value 5.145884
final  value 5.145884
stopped after 50 iterations
```

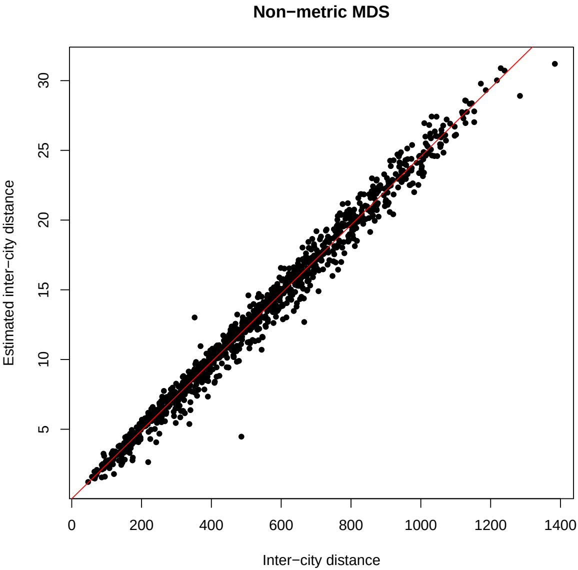
R code for non-metric MDS

```
> nmmds.out <- isoMDS(Spain,y=init,k=2,maxit=100)
initial  value 41.659041
iter   5 value 40.219780
iter  10 value 37.286307
iter  15 value 30.177635
iter  20 value 22.661686
iter  25 value 14.483317
iter  30 value 10.703962
iter  35 value 7.756514
iter  40 value 6.116380
iter  45 value 5.360785
iter  50 value 5.145884
iter  55 value 5.088756
final  value 5.057439
converged
> Y <- nmmds.out$points
> nmmds2.out <- isoMDS(Spain,y=X2,k=2) # PCO solution as initial configuration
initial  value 6.252429
final  value 6.252214
converged
> Y2 <- nmmds2.out$points
> plot(Y[,2],Y[,1],pch=19)
> text(Y[,2], Y[,1], rownames(Spain), cex=0.5,pos=1)
```

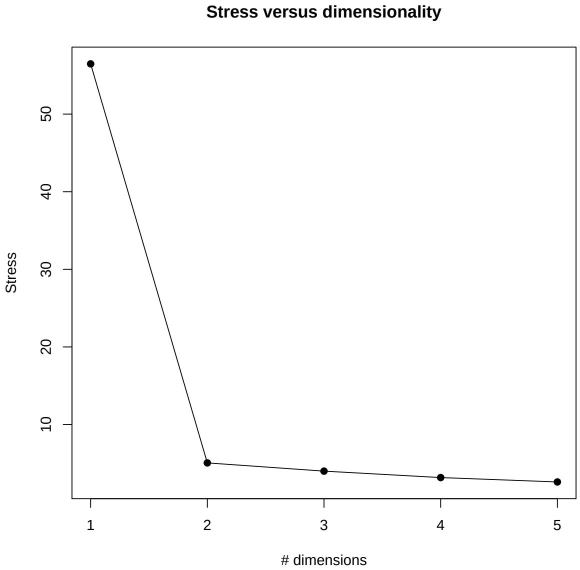
Non-metric MDS map of Spain



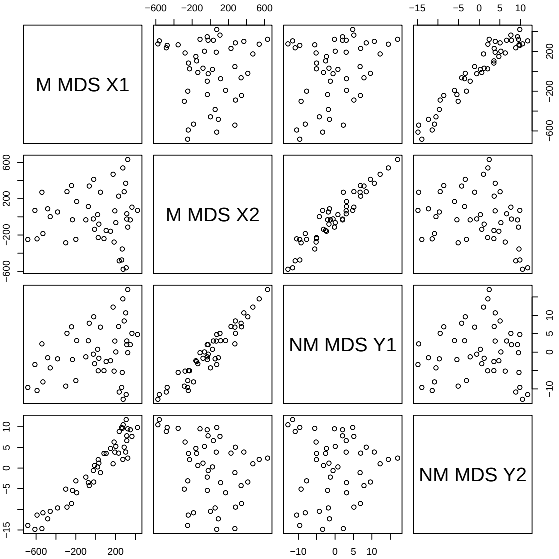
Diagnostics non-metric MDS



Diagnostics non-metric MDS



Relation metric MDS and non-metric MDS solutions



Correlation solutions MDS versus non-metric MDS

```
          M MDS X1 M MDS X2 NM MDS Y1 NM MDS Y2
M MDS X1      1.00     0.00      0.31      0.96
M MDS X2      0.00     1.00      0.95     -0.29
NM MDS Y1     0.31     0.95      1.00      0.02
NM MDS Y2     0.96    -0.29      0.02      1.00
```

MDS for 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

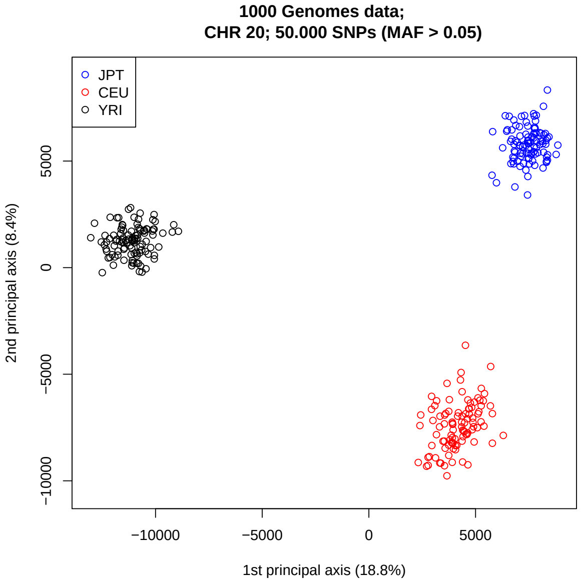
$i \backslash j$	AA	AB	BB
AA	0	1	2
AB	1	0	1
BB	2	1	0

- 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.

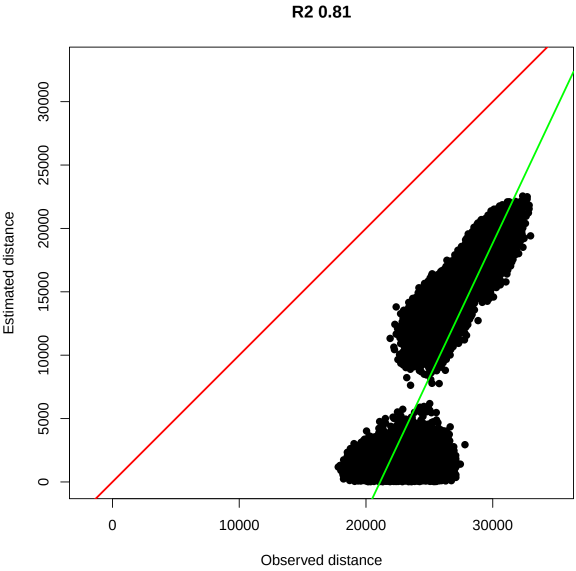
MDS with genetic data (CEU, JPT and YRI samples from 1,000 Genomes)



Goodness of fit

Dim.	λ	%	% Cum.
1	4.04	0.14	0.14
2	1.64	0.06	0.20
3	1.43	0.05	0.25
4	0.98	0.04	0.29
5	0.88	0.03	0.32
6	0.78	0.03	0.35
7	0.74	0.03	0.37
8	0.69	0.02	0.40
9	0.65	0.02	0.42
⋮	⋮	⋮	⋮
62	0.14	0.00	0.98
63	0.13	0.00	0.99
64	0.13	0.00	0.99
65	0.13	0.00	1.00
66	0.12	0.00	1.00
67	0.12	0.00	1.01
68	0.11	0.00	1.01
69	0.11	0.00	1.01
⋮	⋮	⋮	⋮
121	0.01	0.00	1.11
122	0.01	0.00	1.11
123	0.00	0.00	1.11
124	0.00	0.00	1.11
125	0.00	0.00	1.11
126	0.00	0.00	1.11
127	-0.00	-0.00	1.11
128	-0.00	-0.00	1.11
129	-0.00	-0.00	1.11
130	-0.01	-0.00	1.11
⋮	⋮	⋮	⋮
200	-0.08	-0.00	1.01
201	-0.08	-0.00	1.01
202	-0.09	-0.00	1.00
203	-0.10	-0.00	1.00

Goodness of fit



Computer exercise

- Load the database CHBChr2-200.rda
- Convert the genotype data into and $n \times n$ distance matrix.
- Produce a map of the individuals by metric multidimensional scaling. Is there evidence for the existence of groups?
- Make a graph of the fitted against the observed distances, and comment on the results.
- Produce a map of the individuals by non-metric multidimensional scaling. Are the results comparable not those obtained by metric MDS? Is there evidence for the existence of groups?

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
- Foulkes, A.S. (2009) Applied statistical genetics with R. Springer.
- Mardia, K.V. et al. (1979) Multivariate Analysis. Chapter 14. Academic press.