

5. Population substructure

Iván Galván-Femenía¹, Jan Graffelman¹

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

December 1, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Masters in **Computer Science** and **Engineering**

Contents

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

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

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 for HWE

```
> library(HardyWeinberg)
> x1 <- c(27, 126, 147)
> out1 <- HWChisq(x1,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 7.097959e-30 DF = 1 p-value = 1 D = 7.105427e-15 f = 0

> x2 <- c(192, 96, 12)
> out2 <- HWChisq(x2,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 4.470212e-30 DF = 1 p-value = 1 D = 0 f = 0

> x3 <- x1+x2
> out3 <- HWChisq(x3,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 19.5815 DF = 1 p-value = 9.639802e-06 D = -19.00249 f = 0.2550585
>
```

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 for LD

```
> X1 <- matrix(c(84,56,36,24),ncol=2,byrow = T)
> chisq.test(X1,correct=FALSE)
```

Pearson's Chi-squared test

```
data: X1
X-squared = 0, df = 1, p-value = 1
```

```
> X2 <- matrix(c(54,6,126,14),ncol=2,byrow = T)
> chisq.test(X2,correct=FALSE)
```

Pearson's Chi-squared test

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

```
> X3 <- (X1+X2)/2
> chisq.test(X3,correct=FALSE)
```

Pearson's Chi-squared test

```
data: X3
X-squared = 3.84, df = 1, p-value = 0.05004
```

```
>
```


How to detect substructure

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

Multidimensional scaling

Objective

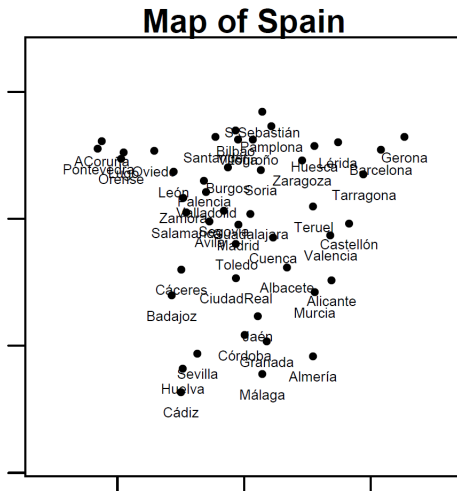
On the basis of information regarding the distances (or similarities) of n objects, construct a configuration of n points in a low-dimensional space (a **map**).

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 <http://www-eio.upc.es/~jan/data/SpainDist.dat>

Example data set



Some basic terminology

Terminology

- dissimilarity or distance (d_{sr})

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 $\delta(A, B)$ is 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:

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

- Mahanalobis distance:

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

- Minkowski distance

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

$\lambda = 1$ Manhattan distance

$\lambda = 2$ Euclidean distance

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

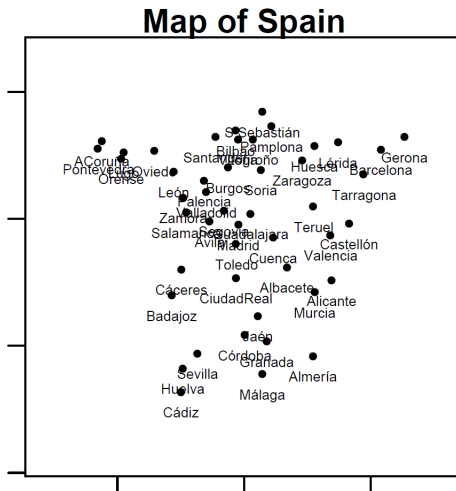
- Also known as: classical scaling, principal coordinate analysis (PCO).
- Given n objects with dissimilarities (δ_{rs}) find a set of points in Euclidean space such that $d_{rs} \approx \delta_{rs}$.
- Classical application: given a distance matrix (in km or in travel time) between cities, construct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

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 <http://www-eio.upc.es/~jan/data/SpainDist.dat>

Example data set



Algorithm for Classical Scaling

- Compute a distance or dissimilarity matrix \mathbf{A} of dimension $n \times n$.
- Let $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}'_n \mathbf{1}_n$ with \mathbf{I}_n identity matrix and $\mathbf{1}_n$ a n -row-vector of ones.
- Double center \mathbf{A} by $\mathbf{B} = \mathbf{H} \mathbf{A} \mathbf{H}$
- Compute the spectral decomposition (eigenvalues and eigenvectors) by $\mathbf{B} = \mathbf{V} \mathbf{D}_\lambda \mathbf{V}'$ where \mathbf{D}_λ is a diagonal n -matrix with the eigenvalues of \mathbf{B} $\lambda_1, \lambda_2, \dots, \lambda_n$ and \mathbf{V} is a n -matrix of the eigenvectors of \mathbf{B} $[v_1, v_2, \dots, v_n]$
- Compute a least square approximation of rank $m = 2$ to \mathbf{B} , $\mathbf{B}^* = \mathbf{V}_{[1:2]} \mathbf{D}_{\lambda[1:2,1:2]} \mathbf{V}'_{[1:2]}$, and plot the map by $\mathbf{X} = \mathbf{V}_{[1:2]} \mathbf{D}_{\lambda[1:2,1:2]}^{1/2}$.

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 \mathbf{B} is not positive semi-definite (there are $\lambda < 0$):

$$\frac{\sum_{i=1}^P \lambda_i}{\sum_{\lambda_i > 0} \lambda_i}$$

$P = 2$ for two-dimensional solution

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 .

- Theorem

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

R code for classical scaling

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

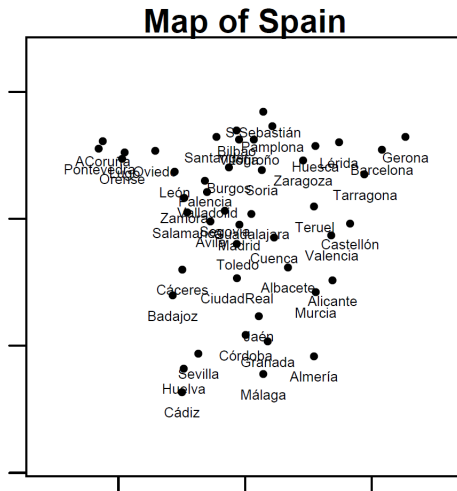
> mds.out <- cmdscale(Spain,k=n-1,eig=TRUE)
Warning message:
In cmdscale(Spain, k = n - 1, eig = TRUE) :
  only 24 of the first 46 eigenvalues are > 0

> 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)
```

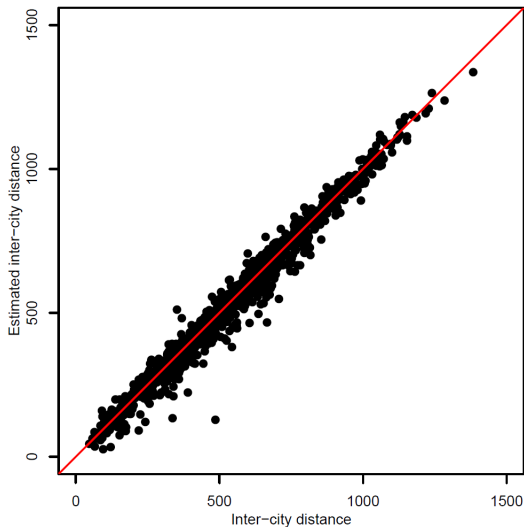
R code for classical scaling

```
> ev <- mds.out$eig
> gof <- mds.out$GOF
> print(round(ev,digits=2))
[1] 4419357.73 3710242.86 523390.06 222914.52 215904.45
[6] 143955.45 128021.63 103602.38 92361.07 77669.80
[11] 67866.94 55724.33 51347.16 38327.38 32347.58
[16] 29609.07 18785.64 14974.46 9473.34 9317.99
[21] 6911.58 4219.73 1459.24 105.43 0.00
[26] -854.58 -3724.49 -4557.54 -5306.92 -8958.67
[31] -11879.05 -15217.83 -16867.79 -24417.22 -34120.67
[36] -43608.19 -50334.85 -63916.60 -77134.54 -80754.15
[41] -91612.38 -97422.06 -120383.81 -125973.49 -179445.66
[46] -253056.31 -340735.97
> print(round(gof,digits=4))
[1] 0.8581 1.0000
```

PCO map of Spain



Observed (δ_{rs}) vs estimated distances (d_{rs})



Non-metric MDS: objective function

- $$\text{STRESS} = \sqrt{\frac{\sum_{r \neq s}^n (f(\delta_{rs}) - d_{rs})^2}{\sum_{r \neq s}^n d_{rs}^2}}$$
- We minimize the objective function numerically, starting from an initial configuration.

Procedure for non-metric MDS

- Choose a distance measure (e.g. $\delta_{rs} = \{\sum_{i=1}^P |x_{ri} - x_{si}|^\lambda\}^{\frac{1}{\lambda}}$).
- 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 ($d_{rs} - f(\delta_{rs})$)

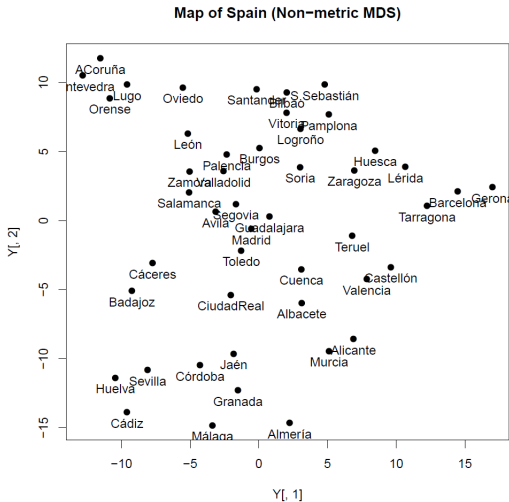
R code for non-metric MDS

```
> init <- scale(matrix(runif(n*2),ncol=2),scale=FALSE)
> nmmds.out <- isoMDS(Spain,y=init,k=2)
initial value 43.230124
iter 5 value 39.177603
iter 10 value 35.976246
iter 15 value 30.295813
iter 20 value 25.310568
iter 25 value 21.644328
iter 30 value 17.611273
iter 35 value 11.659683
iter 40 value 8.991240
iter 45 value 6.220859
iter 50 value 5.375842
final value 5.375842
stopped after 50 iterations
>
```

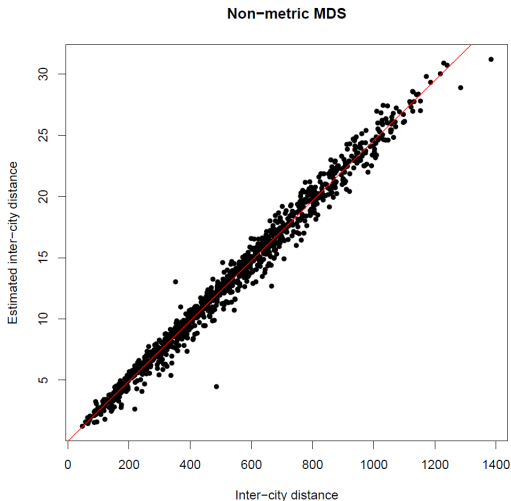
R code for non-metric MDS

```
> nmmds.out <- isoMDS(Spain,y=init,k=2,maxit=100)
initial value 43.230124
iter 5 value 39.177603
iter 10 value 35.976246
iter 15 value 30.295813
iter 20 value 25.310568
iter 25 value 21.644328
iter 30 value 17.611273
iter 35 value 11.659683
iter 40 value 8.991240
iter 45 value 6.220859
iter 50 value 5.375842
iter 55 value 5.174010
final value 5.124100
converged
> Y <- nmmds.out$points
> nmmds2.out <- isoMDS(Spain,y=X,k=2) # PCO solution as initial configuration
initial value 6.252429
final value 6.252214
converged
> Y <- nmmds2.out$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

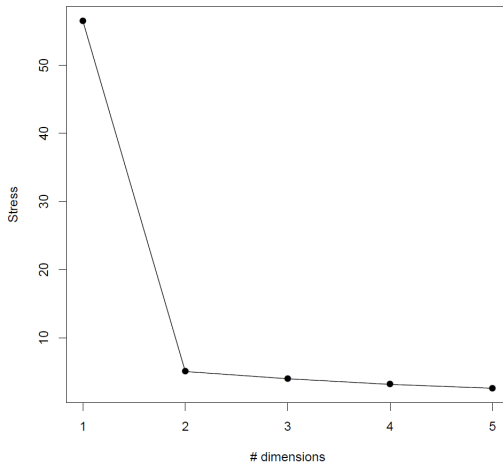


Diagnostics non-metric MDS

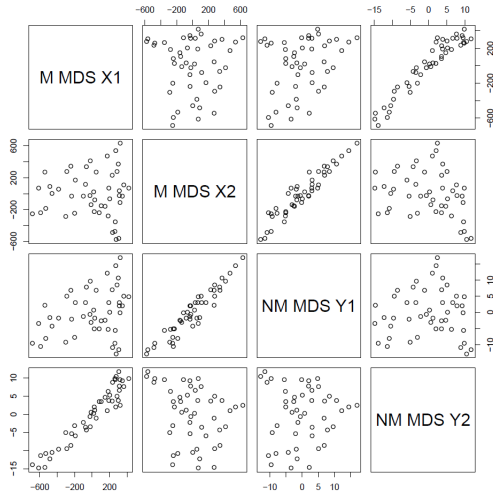


Diagnostics non-metric MDS

Stress versus dimensionality



Relation metric MDS and non-metric MDS solutions

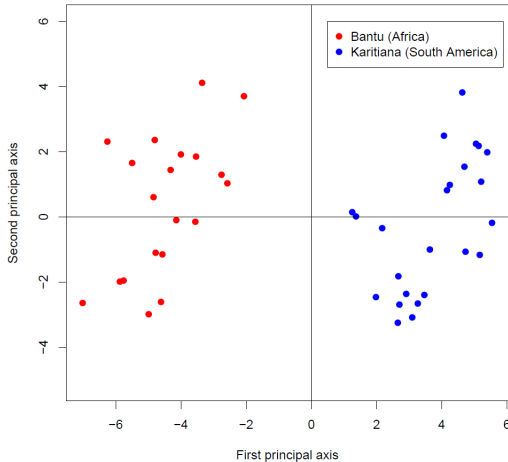


Correlation metric MDS and non-metric MDS solutions

	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 with genetic data

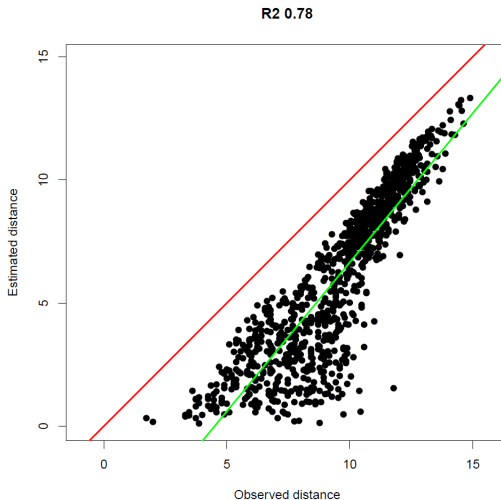
MDS of 144 SNPs (Chr 22) for two human populations



Goodness of fit

Dim.	λ	%	% Cum.	Dim.	λ	%	% Cum.
1	799.98	0.40	0.40	23	12.91	0.01	0.96
2	187.76	0.09	0.49	24	11.63	0.01	0.97
3	142.50	0.07	0.56	25	9.82	0.00	0.97
4	111.87	0.06	0.61	26	8.35	0.00	0.98
5	93.40	0.05	0.66	27	7.06	0.00	0.98
6	83.65	0.04	0.70	28	6.61	0.00	0.98
7	59.46	0.03	0.73	29	5.76	0.00	0.99
8	54.66	0.03	0.76	30	5.19	0.00	0.99
9	51.28	0.03	0.78	31	3.83	0.00	0.99
10	47.44	0.02	0.81	32	3.10	0.00	0.99
11	45.55	0.02	0.83	33	2.90	0.00	0.99
12	37.66	0.02	0.85	34	2.51	0.00	1.00
13	35.92	0.02	0.87	35	2.09	0.00	1.00
14	32.21	0.02	0.88	36	1.91	0.00	1.00
15	25.43	0.01	0.89	37	1.46	0.00	1.00
16	23.28	0.01	0.91	38	1.04	0.00	1.00
17	22.29	0.01	0.92	39	0.84	0.00	1.00
18	20.33	0.01	0.93	40	0.63	0.00	1.00
19	17.14	0.01	0.93	41	0.50	0.00	1.00
20	15.53	0.01	0.94	42	0.46	0.00	1.00
21	13.98	0.01	0.95	43	0.19	0.00	1.00
22	13.82	0.01	0.96	44	0.00	0.00	1.00





Goodness of fit



Computer exercises

- 1 Load the database
`http://www-eio.upc.es/~jan/data/bsg/CHBChr2-2000.rda`
- 2 Convert the genotype data into an $n \times n$ distance matrix.
- 3 Produce a map of the individuals by metric multidimensional scaling. Is there evidence for the existence of groups?
- 4 Make a graph of the fitted against the observed distances, and comment on the results.
- 5 Produce a map of the individuals by non-metric multidimensional scaling. Are the results comparable to those obtained by metric MDS? Is there evidence for the existence of groups?

References

-  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.