Bioinformatics and Statistical Genetics

Elective specialization for MDS/MIRI/MAI students

Marta Castellano

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

marta.castellano@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics
2. Hardy-Weinberg equilibrium
3. Linkage disequilibrium and haplotype estimation
4. Population substructure 28 November 2023
5. Genetic association analysis
6. Relatedness analysis (allele sharing)

Tuesdays 5-8pm

Hardy-Weinberg Equilibrium (HWE)

RECAP

When a population is in Hardy-Weinberg equilibrium for a gene, it is not evolving, and allele frequencies will stay the same across generations….in the absence of **disturbing factors**.

Statistical testing allows to decide whether the sample genotype frequencies adhere to the frequencies observed in HWE. The null hypothesis states that $f(AA) = p^2, f(AB) = 2pq, f(BB) = q^2$. Most frequent statistical tests:

- Pearson's test.
- Fisher's exact test (based on).
- Permutation test.

Measures of disequilibrium (like the inbreeding coefficient f) quantify how far from HWE the sample is.

Generalizations of the HWE are required for multiple alleles, the variant has multiple copies, the organism studied is tetraploid or the variant is found on the X-chromosome.

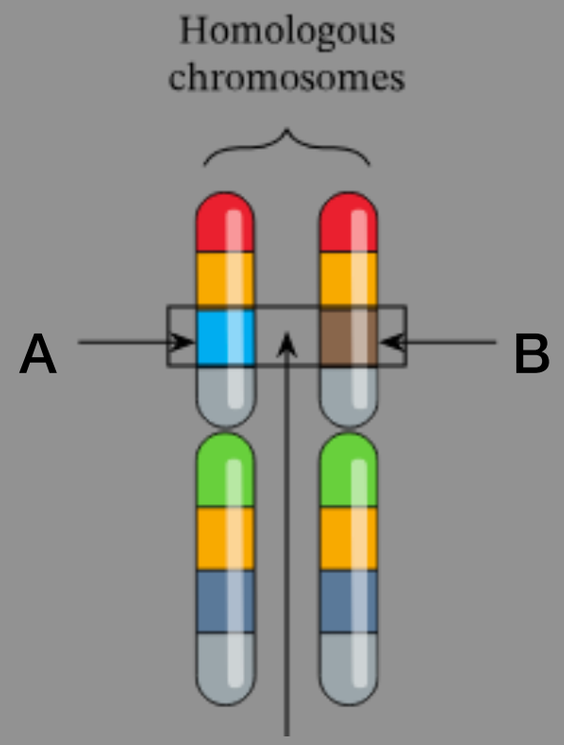
		Females	
		A (p)	B (q)
Males	A (p)	AA (p^2)	AB (pq)
	B (q)	AB (qp)	BB (q^2)

frequency of homozygous dominant genotype

frequency of heterozygous recessive genotype

$$p^2 + 2pq + q^2 = 1$$

frequency of heterozygous genotype



Locus of interest

Linkage Disequilibrium & Haplotype Estimation

RECAP

When two alleles are in **linkage disequilibrium**, their co-ocurrence in the population is not independent and thus….LD analyses the on-random association of alleles at different loci in a given population.

Several metrics quantify LD. MLE required for haplotype estimation. The coefficient of linkage disequilibrium (D) is the most frequent, together with its standardized D' and the squared correlation between variants (R2).

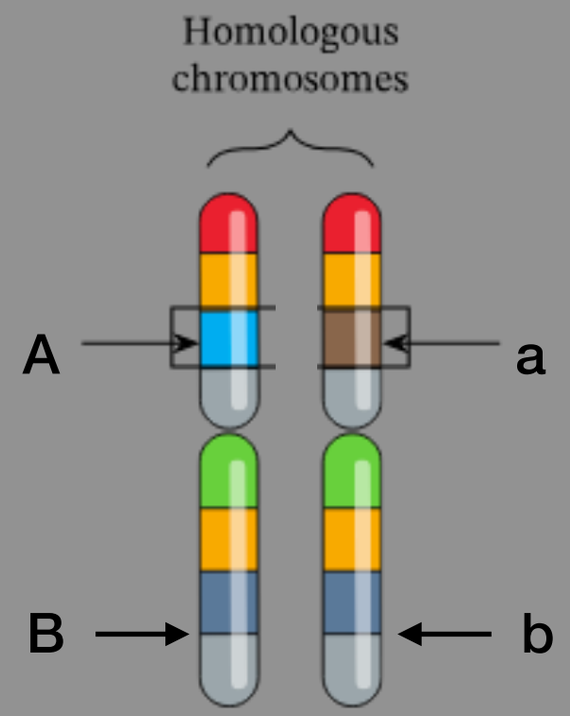
		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

Hypothesis testing for LD states that $H_0 : D = 0$

Haplotype refer to a set of variants at different loci within the same chromosome that are statistically associated.

Statistical phasing exploit inter-variant correlation to find set of haplotypes either to….

- Find haplotype constitution of an individual
- Estimate haplotype frequency of a population



Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Classical MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

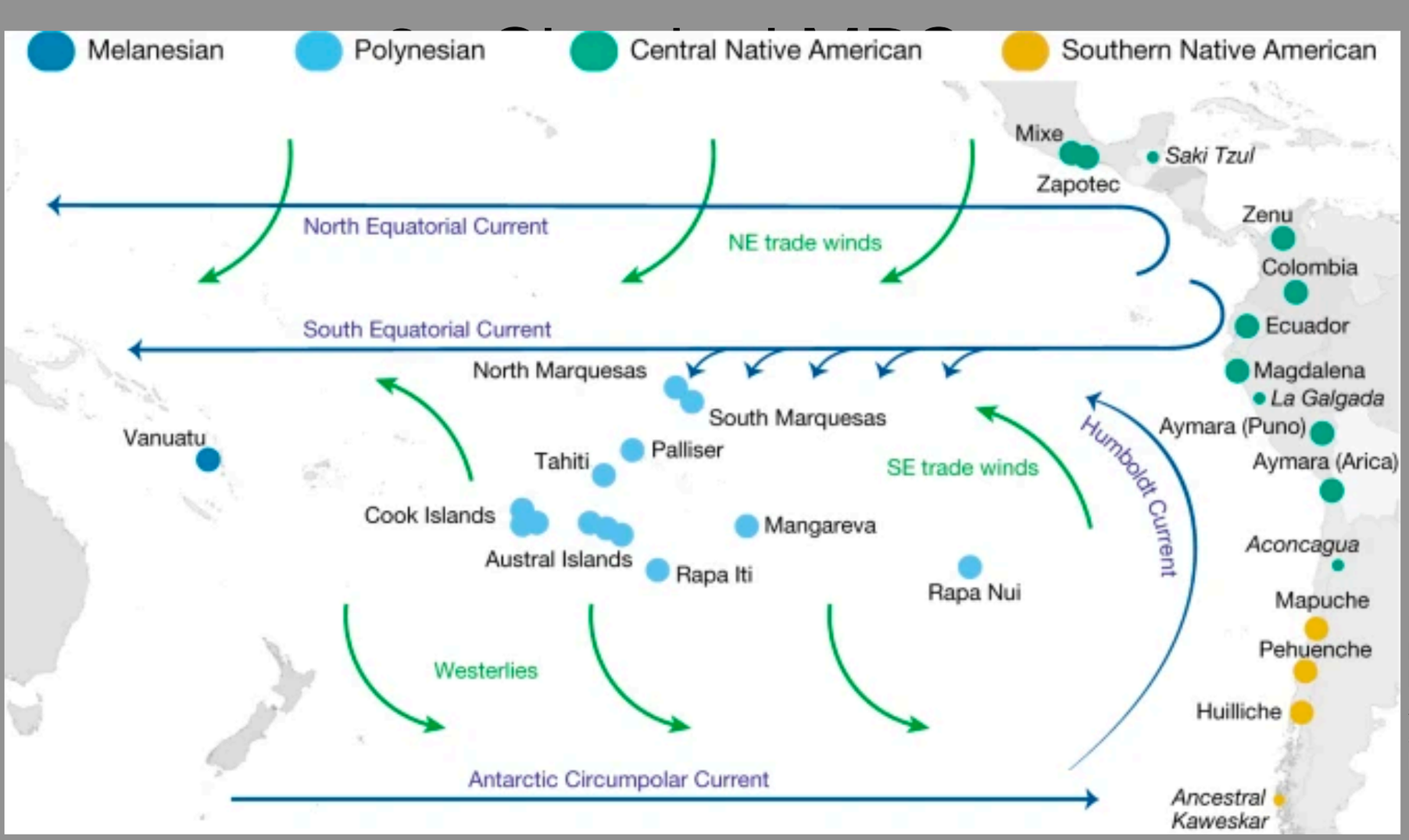
Population structure (also called genetic structure or population stratification) is the presence of a systematic difference in allele frequencies between subpopulations.

Content

Population substructure

Population is often defined as a set of organisms in which any pair of members can breed together….living in a defined geographic region at a specific point in time.

1. Introduction to population substructure
2. Introduction to MDS



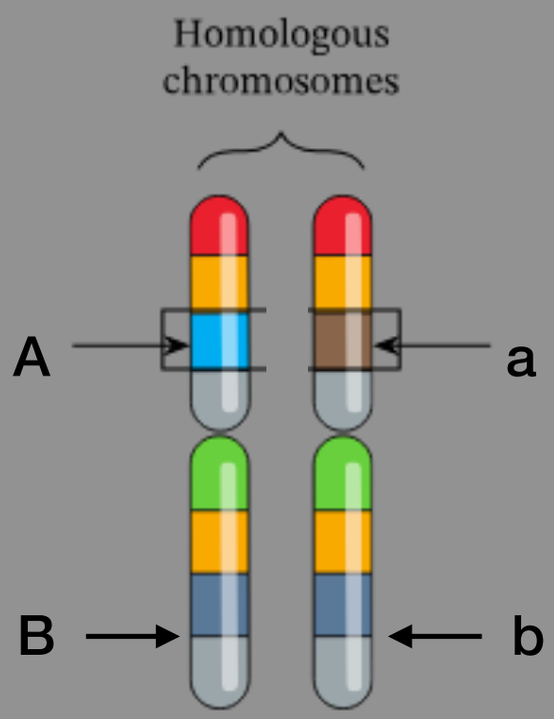
Ioannidis, A.G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Miquel-Poblete, J.F., Moreno-Mayar, J.V., Rodríguez-Rodríguez, J.E., Quinto-Cortés, C.D., Auckland, K., Parks, T. and Robson, K., 2020. **Native American gene flow into Polynesia predating Easter Island settlement.** Nature, 583(7817), pp.572-577.

Population substructure

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in population genetics: ~~Linkage disc~~equilibrium

~~Haplotype~~ estimation



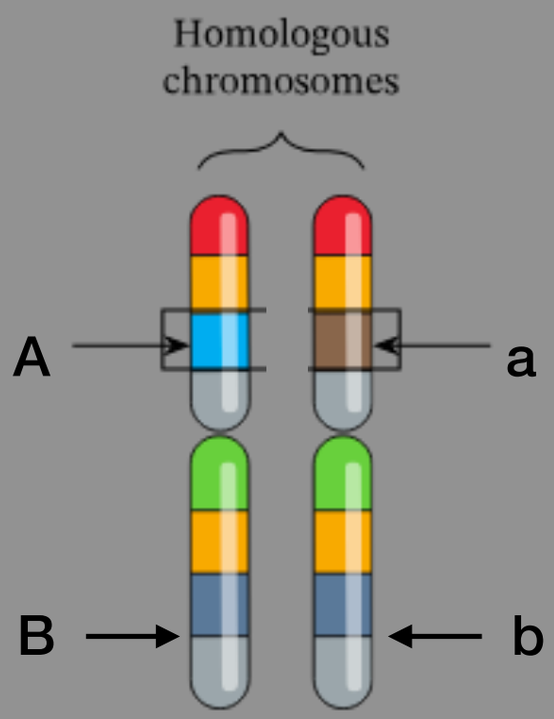
- What factors can affect the methods we have presented so far? Taking HWE assumptions as starting point…
  - The organism under study is diploid.
  - There is sexual reproduction.
  - Non-overlapping generations.
  - Random mating (w.r.t the trait under study).
  - Allele frequencies are equal in the sexes.
  - Population size is very (infinitely) large.
  - Migration is negligible.
  - Mutation can be ignored (i.e. no genetic drift).
  - Natural selection does not affect the trait under study.
  - There is no genotyping error.

Population substructure

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in population genetics: ~~Linkage disc~~equilibrium

~~Haplotype~~ estimation



- What factors can affect the methods we have presented so far? Taking HWE assumptions as starting point…
  - The organism under study is diploid.
  - There is sexual reproduction.
  - Non-overlapping generations. ← Lecture 6
  - Random mating (w.r.t the trait under study).
  - Allele frequencies are equal in the sexes.
  - Population size is very (infinitely) large.
  - Migration is negligible (i.e. gene flow).
  - Mutation can be ignored (i.e. no genetic drift).
  - Natural selection does not affect the trait under study.
  - There is no genotyping error.

(the change in frequency of an existing gene variant in the population due to random chance)

Population substructure

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in population genetics: ~~Linkage disc~~equilibrium

~~Haplotype estimation~~

- Genotype frequencies and allele frequencies may vary over human (sub)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?**

Population substructure

Explain Like I'm 5

The ~~Hardy-Weinberg~~ Equilibrium is fundamental in population genetics: ~~Linkage disc~~equilibrium

~~Haplotype~~ estimation

	id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633
1	NA18939	T/G	C/T	G/A	T/G	T/G
2	NA18940	G/G	T/T	A/A	G/G	T/G
3	NA18941	G/G	T/T	A/A	G/G	T/G
4	NA18942	G/G	T/T	A/A	G/G	T/T
5	NA18943	G/G	T/T	A/A	G/G	T/T
6	NA18944	T/T	C/C	G/G	T/G	G/G
7	NA18945	G/G	T/T	A/A	G/G	G/G
8	NA18946	T/G	C/T	G/A	G/G	G/G
9	NA18947	T/G	C/T	G/A	G/G	T/G
10	NA18948	G/G	T/T	A/A	G/G	G/G
11	NA18949	T/G	C/T	G/A	T/G	T/G
12	NA18950	G/G	T/T	A/A	G/G	T/G
13	NA18951	G/G	T/T	A/A	G/G	T/G
14	NA18952	T/G	C/C	G/G	T/G	T/G

- **How to detect population substructure?**
- Set of methods that allow to detect population substructures, which refers to the existence of groups of individuals in a genetic database that come from **different human populations**.

Pemberton, T.J., Wang, C., Li, J.Z. and Rosenberg, N.A., 2010. Inference of unexpected genetic relatedness among individuals in **HapMap** Phase III. The American Journal of Human Genetics, 87(4), pp.457-464.

Population substructure

Introduction

- 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 studies
  - ...

Population substructure

Introduction - effects on the HWE

		Females	
		A (p)	B (q)
Males	A (p)	AA (p^2)	AB (pq)
	B (q)	AB (qp)	BB (q^2)

frequency of homozygous dominant genotype

frequency of heterozygous recessive genotype

$$p^2 + 2pq + q^2 = 1$$

frequency of heterozygous genotype

- 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 observe Hardy-Weinberg equilibrium within each population.

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

```
> library(HardyWeinberg)
> x1
[1] 27 126 147
> out1 <- HWChisq(x1,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 = 7.097959e-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= 4.470212e-30  p-value= 1       D= 0

> x3 <- x1+x2
[1] 219 222 159
> out3 <- HWChisq(x3,cc=0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium
Chi2 = 38.2614 p-value = 6.187448e-10 D = -18.75
```

We have FAILED to reject the null hypothesis and we can't refute H0, that our alleles are in equilibrium

We reject the null hypothesis that our alleles are in equilibrium

Population substructu

		SNP2		
		B	b	
SNP1	A	$p_A p_B + D$	$p_A p_b - D$	p_A
	a	$p_a p_B - D$	$p_a p_b + D$	p_a
		p_B	p_b	1

Introduction - effects on the LD

- Let there be two populations, and consider two polymorphisms, A/a and B/b.
  - In the first population we have pA = 0.7 and pB = 0.6. In the second population we have pA = 0.3 and pB = 0.9.
  - Let there be 100 individuals (200 haplotypes) in each population (n1 = n2 = 100).
- We observe linkage equilibrium within each population.

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

```
> 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
```

We have FAILED to reject the null hypothesis and we can't refute H0, that our alleles are in linkage disequilibrium

Population substructure

Introduction - effects on the LD

- Let there be two populations, and consider two polymorphisms, A/a and B/b.
  - In the first population we have pA = 0.7 and pB = 0.6. In the second population we have pA = 0.3 and pB = 0.9.
  - Let there be 100 individuals (200 haplotypes) in each population (n1 = n2 = 100).
- We observe linkage equilibrium within each population.

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

```
> X4 <- (X1+X2)
> > X4
     [,1] [,2]
[1,]  138   62
[2,]  162   38

> out <- chisq.test(X4,correct=FALSE) >
print(out)
Pearson's Chi-squared test
data: X4
X-squared = 7.68, df = 1, p-value = 0.005584
```

We *reject* the null hypothesis that our alleles are in linkage disequilibrium

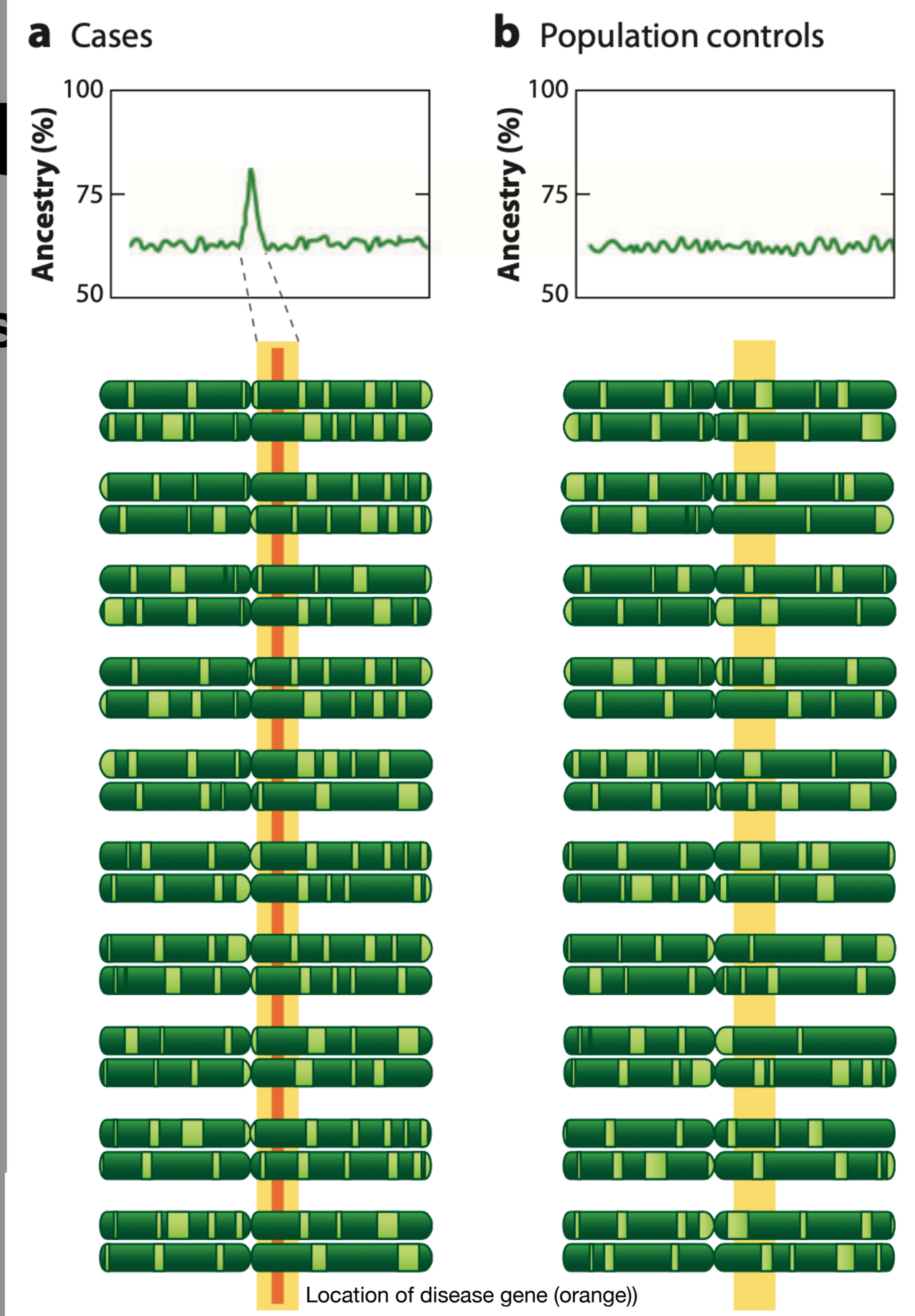
Population subst

Introduction - effects on marker-association studies

- Population structure (i.e. admixture mapping) could be applied to case-control studies to increase statistical power.

Example (right):

- Pattern of chromosomal admixture around a disease locus is shown
- Cases show an excess of ancestry blocks around the location of the disease locus
- In case-control studies, the locus-specific ancestry at ancestry-informative markers is compared between cases and controls.

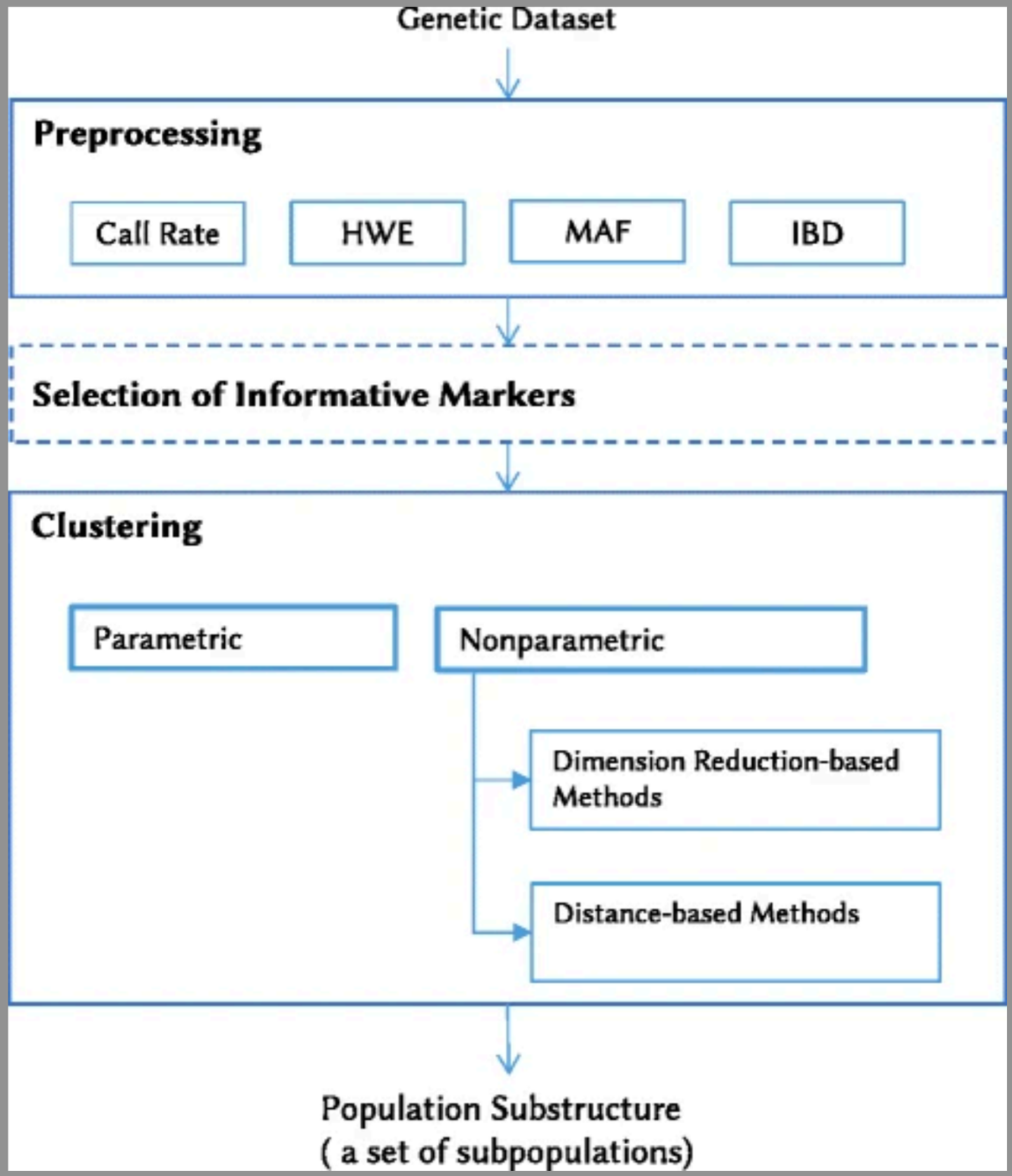


Winkler, C.A., Nelson, G.W. and Smith, M.W., 2010. Admixture mapping comes of age. Annual review of genomics and human genetics, 11, pp.65-89.

Population substructure

Introduction - general workflow

- Call rate: amount of missing data, usually poorly genotyped SNPS are removed (95% threshold)
- HWE: if a variant deviates from equilibrium, it's removed, assuming to comes from genotyping errors.
- MAF: SNPs with low Minor Allele Frequency are excluded, and a threshold of 1–2% is typically applied.
- IBD: Identity by descent assess which individuals are related (lecture 6)



Alhusain, L. and Hafez, A.M., 2018. Nonparametric approaches for population structure analysis. Human genomics, 12(1), pp.1-12.

Population substructure

How to detect substructure?

- Population structure is a complex phenomenon and no single measure can capture all aspects of it. **Many methods** address different aspects of population substructure.

- **Local ancestry:** also known as admixture mapping or chromosome painting, aim to classify ancestry in small chromosomal regions

- **Global ancestry**: estimate a global parameter that summarizes the ancestry of study participants given genotype data.

  - Principal component analysis of the marker data

  - Multidimensional scaling (MDS) of distance matrix computed from the marker data

  - ….

Hellwege, J.N., Keaton, J.M., Giri, A., Gao, X., Velez Edwards, D.R. and Edwards, T.L., 2017. Population stratification in genetic association studies. Current protocols in human genetics, 95(1), pp.1-22.

Popula

How to de

- Population
  capture all a
- **Many meth**

Table 3 Programs for admixture mapping

Program	Ancestry inference method	Disease association test	Markers	Number of parental populations	Allows background LD?	Quantitative trait?
STRUCTURE/ MALDsoft	HMM-MCMC	Z score	Microsatallites or SNPs	Any number	No	No
ADMIXMAP	HMM-MCMC	Regression tests	Microsatellites or SNPs	Any number	No	Yes
ANCESTRYMAP	HMM-MCMC	LOD score (genome-wide and local)	SNP AIMs	2	No	No
ADMIXPROGRAM	HMM-ML	Z score	SNPs	2	No	No
SABER	HMM-ML	None	Any SNPs	Any number	Yes	N/A
HAPMIX	MHMM-MCMC	None	Dense set of SNPs	2	Yes	N/A
LAMP LAMP-ANC	Moving window	None	Dense set of SNPs	Any number	Yes	N/A
HAPAA, uSWITCH	Hierarchical HMM	None	Dense set of SNPs	Any number	Yes	N/A

HMM, hidden Markov model;
LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo;
MHMM, Markov hidden Markov model;
ML, maximum likelihood;

Winkler, C.A., Nelson, G.W. and Smith, M.W., **2010**. Admixture mapping comes of age. Annual review of genomics and human genetics, 11, pp.65-89.

Populatio

How to detec

- Population struct
  capture all aspec
- **Many methods**

Table 1.22.2 Global Ancestry Methods and Software

Program	Method	Function	Link for download
Eigensoft	PCA	Calculate PCA from genotype data	*https://reich.hms.harvard.edu/software*
LASER	PCA	Calculate PCA from sequencing data (low pass)	*http://laser.sph.umich.edu/*
FlashPCA	PCA	Rapid calculation of PCA	*https://github.com/gabraham/flashpca*
PC-AiR	PCA	PCA in samples that may contain cryptically related participants	*http://bioconductor.org/packages/release/bioc/html/GENESIS.html*
PCAmask	PCA	PCA in highly structured populations	*https://github.com/armartin/ancestry_pipeline*
PLINK	MDS	Calculation of multidimensional scaling variables from IBD distance matrix	*http://zzz.bwh.harvard.edu/plink/*
EMMA	Mixed model	Perform linear mixed model analysis for quantitative traits	*http://mouse.cs.ucla.edu/emma/*
GEMMA	Mixed Model	Perform linear mixed model analysis for quantitative traits	*http://www.xzlab.org/software.html*
EMMAX	Mixed Model	Perform linear mixed model analysis for quantitative traits more quickly than EMMA	*http://genetics.cs.ucla.edu/emmax/*
LD score regression	LD score regression	Calculate genomic inflation parameters accounting for LD	*https://github.com/bulik/ldsc*
PC loading regression	PC loading regression	Improved PS control compared with PCA	Not yet available
GAP, SCGAP	Geographic ancestry positioning	Probabilistic spatial genetic model and ancestry localization algorithm, as well as the related population stratification correction procedure for genome-wide association studies, SCGAP	*https://github.com/anand-bhaskar/gap*
SNPweights	SNPweights	Inferring genome-wide genetic ancestry using SNP weights precomputed from large external reference panels	*https://www.hsph.harvard.edu/alkes-price/software/*

In the remainder of this module we focus on MDS

PCA: Principal component analysis
MDS: Multidimensional scaling
LD: linkage disequilibrium

Hellwege, J.N., Keaton, J.M., Giri, A., Gao, X., Velez Edwards, D.R. and Edwards, T.L., 2017. Population stratification in genetic association studies. Current protocols in human genetics, 95(1), pp.1-22.

Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Classical MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

Population substructure

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:

distances (km between cities)

	Albacete	Alicante	Almeria	Avila	Badajoz	Barcelona	Bilbao	Burgos	···
Albacete	0	171	369	366	525	540	646	488	···
Alicante	171	0	294	537	696	515	817	659	···
Almeria	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	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Population substructure

Multidimensional scaling

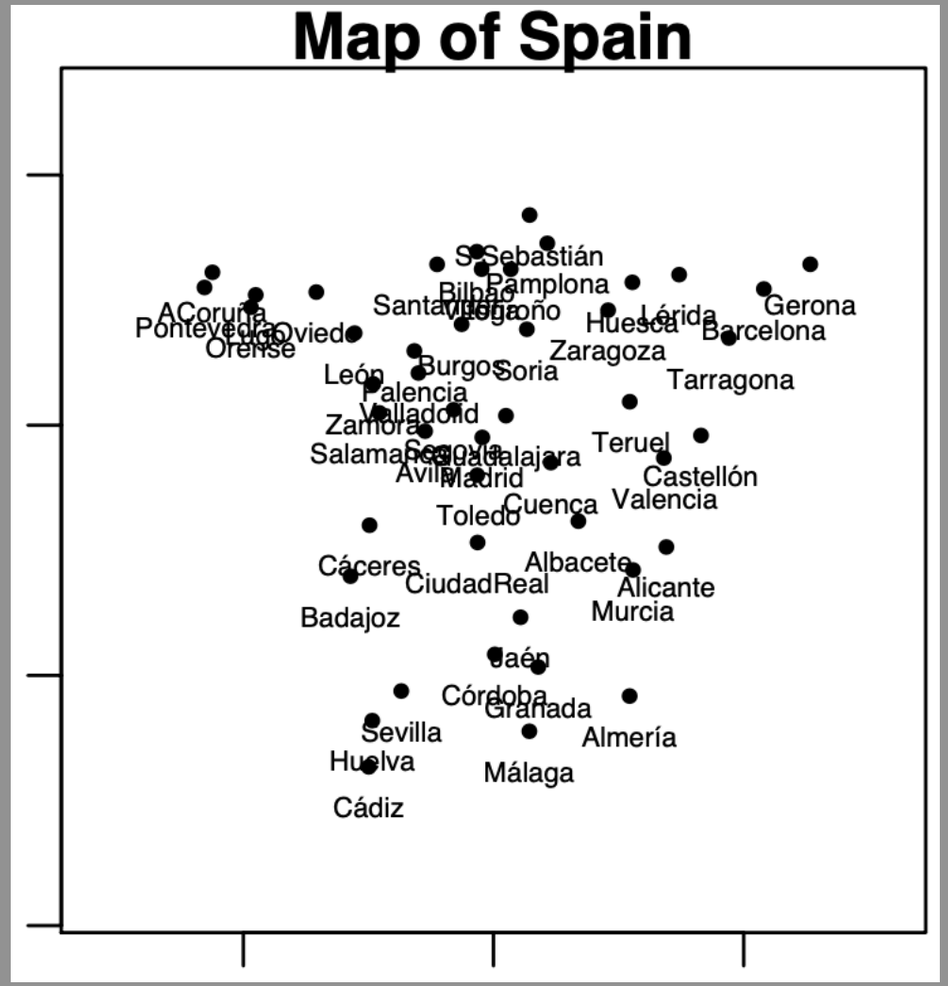
Objective:

- 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:

Given a distance matrix with the distances between each pair of objects in a set, and a chosen number of dimensions, N, an MDS algorithm places each object into N-dimensional space (a lower-dimensional representation) such that the **between-object distances** are preserved as well as possible.

For N = 1, 2, and 3, the resulting points can be visualized on a scatter plot (right).



Population substructure

Multidimensional scaling

Terminology:

- 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, δ(A, B) satisfies:
  - δ(A, B) = δ(B, A)
  - δ(A,B) ≥ 0
  - δ(A, A) = 0

Symmetry: the distance from A to B is always the same as the distance from B to A

Non-negativity: the distance between to distinct points is always positive

Identity: the distance between something and itself is 0. Alternately, two things that have a distance measure of 0 are identical.

- The distance function δ(A,B) called a **metric** if also
  - δ(A,B) = 0 iff A = B
  - the triangle inequality holds: δ(A, B) ≤ δ(A, C) + δ(C, B).

Triangle inequality

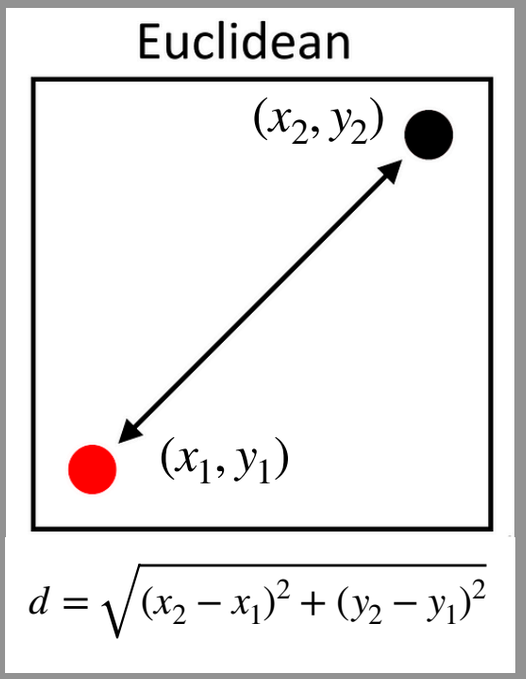
Population substructure

Multidimensional scaling

- Sometimes data are given in the form of similarities (s_{rs}). Similarity data violates some of the axioms that are required on the distance metric, like identity and triangle inequality.
- A similarity matrix C has
  - $s_{rs} = s_{sr}$ **Symmetry**: the distance from A to B is always the same as the distance from B to A
  - $s_{rr} \geq s_{rs}$ Similarity increases as the similarity between A and B increases
  - $s_{rs} > 0$ **Non-negativity:** the distance between to distinct points is always positive
- Similarities can be transformed into distances with the transformation:
  - $d_{rs} = \sqrt{s_{rr} - 2s_{rs} + s_{ss}}$
- If C is positive semidefinite (psd), then the obtained distance matrix will be Euclidean.
  - An nxn symmetric real matrix C is positive semi-definite $\iff x^T Cx \geq 0$ for all $x \in R^n$
- Examples of similarity data:
  - People might agree that North Korea is similar to South Korea, and that North Korea is also similar to Iran, but feel that South Korea and Iran are very different.
  - Individual's evaluation of similarity between words, products, ideas…

Population substructure

Multidimensional scaling



Some distance measures for x_r and x_s two points in p dimensional space

- Euclidean distance

L_2 **norm**: the distance between two points in euclidean space is the length of the segment between them.

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

- 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

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

Population substructure

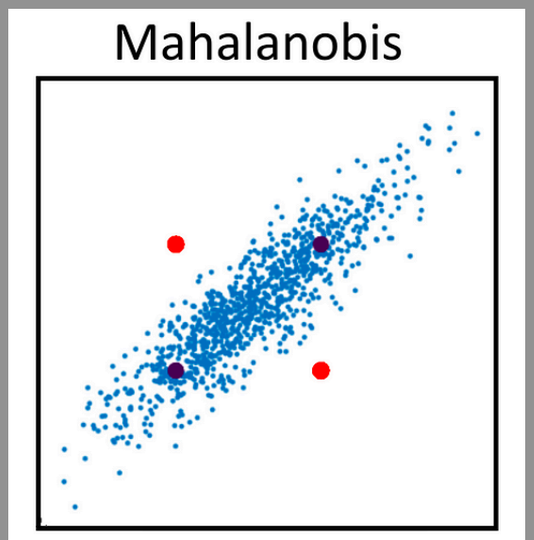
Multidimensional scaling

Some distance measures for x_r and x_s two points in p dimensional space

- Euclidean distance

L_2 **norm**: the distance between two points in euclidean space is the length of the segment between them.

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



- Mahalanobis distance

- Mahalanobis takes into account variance of data points and thus, will be able to represent correlations

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

S = covariance matrix of the data

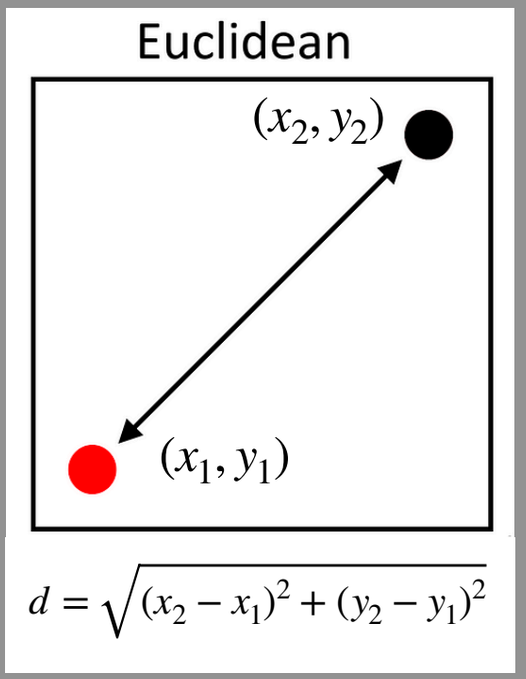
$$cov(X, Y) = E[(X - E[X])(Y - E[Y])^T]$$
for $X = (x_1, x_2, \ldots, x_n)$
and $Y = (y_1, y_2, \ldots, y_n)$

- Minkowski distance

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

Population substructure

Multidimensional scaling



Some distance measures for x_r and x_s two points in p dimensional space

- Euclidean distance

L_2 **norm**: the distance between two points in euclidean space is the length of the segment between them.

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

- Mahalanobis distance

- Mahalanobis takes into account variance of data points and thus, will be able to represent correlations

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

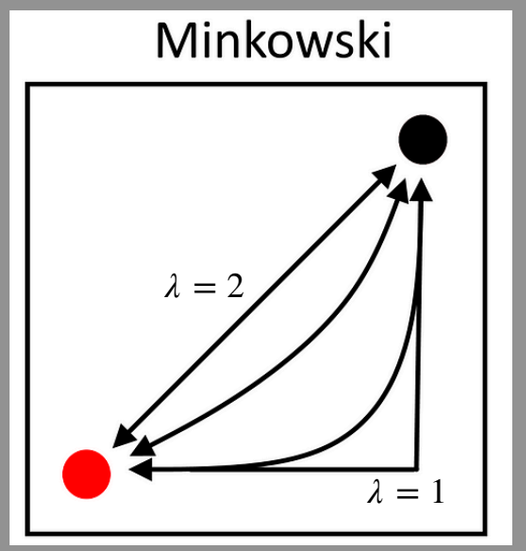
S = covariance matrix of the data

- Minkowski distance

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

L_p **norm**: generalized distance metric.

λ = order of the distance.
If $\lambda = 2$ –> Euclidean distance
If $\lambda = 1$ –> Manhattan distance



Population substructure

Multidimensional scaling

Types of MDS:

- $d_{rs} \approx \delta_{rs}$: Classical scaling or Principal Coordinate Analysis (PCoA)
- $d_{rs} \approx f(\delta_{rs})$ 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

Distance measure between points r and s on the solution map

Sample distances

- In classical MDS and metric MDS, the configuration of points is directly obtained from the distances. Both assume that the distances between data points in the original space can be preserved in the lower-dimensional space.
- Metric MDS is a generalization of classical MDS that generalizes the optimization procedure by making it iterative.
- In non-metric MDS, only the rank order of the distances is important.

On the impact of the similarity/dissimilarity measures in clustering algorithms:
Shirkhorshidi, A.S., Aghabozorgi, S. and Wah, T.Y., 2015. A comparison study on similarity and dissimilarity measures in clustering continuous data. PloS one, 10(12), p.e0144059.

Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Classical MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

Population substructure

Classical MDS

- Classical MDS is best applied to metric variables.
- The method was initially developed by Torgerson (1958).
- It assumes that the data obey distance axioms–they are like a proximity or distance matrix on a map.
- Input to the algorithm is something that behaves like a distance measure. If you have n items, then the input is an nxn matrix where each entry specifies pairwise (and non-negative) distances between items.
- It uses eigendecomposition of the distance to identify major components and axes, and represents any point as a linear combination of dimensions.
- Note that this is very similar to Principal Components Analysis, but it uses the distance matrix rather than a correlation matrix as input.

Population substructure

Classical MDS - simplified algorithm

It takes an input matrix giving dissimilarities between pairs of n items and outputs a coordinate matrix:

Let X be the matrix of coordinates with the solution where x_r and x_s are two rows of X and define $B = XX'$

1. Compute a distance or dissimilarity metrics and set up the squared proximity matrix $D = [d_{rs}^2]$
2. As we assume the solution to be centered at the origin…apply double centering to $D = [d_{rs}^2]$ to obtain B
   - Double centering : $B = -\frac{1}{2}HDH$ using the centering matrix $H = I - \frac{1}{n}J_n$ where I is the n x n identity matrix and J_n is an n x n matrix of all ones.
3. Compute eigenvalues $\mathbf{V} = [\mathbf{v_1}, \ldots, \mathbf{v_n}]$ and eigenvectors $\mathbf{D}_\lambda = diag(\lambda_1, \ldots, \lambda_n)$ of the centered matrix B, so that $B = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$
4. Considering $B = XX' = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$, obtain the coordinates of the solution as $X = \mathbf{V}\mathbf{D}_\lambda^{\frac{1}{2}}$

Population substructure

Classical MDS - simplified algorithm

NOTES:

- Typically a two-dimensional representation is made (i.e. using the first two eigenvalues and eigenvectors)
- $\hat{B} = \mathbf{V}_{(,\mathbf{1}:\mathbf{k})}\mathbf{D}_{\lambda(\mathbf{1}:\mathbf{k},\mathbf{1}:\mathbf{k})}\mathbf{V}'_{(,\mathbf{1}:\mathbf{k})}$ gives the rank k least square approximation to B

3. Compute eigenvalues $\mathbf{V} = [\mathbf{v_1}, \ldots, \mathbf{v_n}]$ and eigenvectors $\mathbf{D}_\lambda = diag(\lambda_1, \ldots, \lambda_n)$ of the centered matrix B, so that $B = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$

4. Considering $B = XX' = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$, obtain the coordinates of the solution as $X = \mathbf{V}\mathbf{D}_\lambda^{\frac{\mathbf{1}}{\mathbf{2}}}$

Population substructure

Classical MDS - simplified algorithm

It takes an input matrix giving dissimilarities between pairs of n items and outputs a coordinate matrix:

Let X be the matrix of coordinates with the solution where x_r and x_s are two rows of X and define $B = XX'$

1. Compute a distance or dissimilarity metrics and set up the squared proximity matrix $D = [\delta_{rs}^2]$

2. As we assume the solution to be centered at the origin…apply double centering to $D = [\delta_{rs}^2]$ to obtain B

   - Double centering : $B = -\frac{1}{2}HDH$ using the centering matrix $H = I - \frac{1}{n}J_n$ where I is the n x n identity matrix and J_n is an n x n matrix of all ones.

3. Compute eigenvalues $\mathbf{V} = [\mathbf{v_1}, \ldots, \mathbf{v_n}]$ and eigenvectors $\mathbf{D}_\lambda = diag(\lambda_1, \ldots, \lambda_n)$ of the centered matrix B, so that $B = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$

4. Considering $B = XX' = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}'$, obtain the coordinates of the solution as $X = \mathbf{V}\mathbf{D}_\lambda^{\frac{1}{2}}$

Population substructure

Classical MDS - simplified algorithm

It takes an input matrix giving dissimilarities between pairs of n items and outputs a coordinate matrix:

Let X be the matrix of coordinates with the solution where x_r and x_s are two rows of X and define $B = XX'$

1. Compute a distance or dissimilarity metrics and set up the squared proximity matrix $D = [\delta_{rs}^2]$

NOTES:

- Classical MDS assumes that the distance used to compute the proximity matrix $D = [\delta_{rs}^2]$ is Euclidean distance.
- A distance matrix D is Euclidean if and only if B (= HAH, as previously defined) is positive semi definite.
    - By definition, there will always be at least on eigenvalue equal to zero
- If a negative eigenvalue appears, then the matrix is not positive-semidefinite

Population substructure

Classical MDS - simplified algorithm

It takes an input matrix giving dissimilarities between pairs of n items and outputs a coordinate matrix:

Let X be the matrix of coordinates with the solution where x_r and x_s are two rows of X and define $B = XX'$

1. Compute a distance or dissimilarity metrics and set up the squared proximity matrix $D = [\delta_{rs}^2]$

NOTE 2:

Sometimes data are given in the form of similarities (s_{rs}).

- A similarity matrix C has $s_{rs} = s_{sr}$ and $s_{rs} \leq s_{rr}$.
- Similarities can be transformed into distances with the transformation $\delta_{rs} = \sqrt{s_{rr} - 2s_{rs} + s_{ss}}$
- If C is positive semidefinite, then the obtained distance matrix will be Euclidean, and by definition, there will always be at least on eigenvalue equal to zero.

Population substructure

Classical MDS - Goodness of fit

How well do we manage to approximate the distance matrix?

A goodness of fit measure typically summarize discrepancy between observed values and the expected values of the particular model under study. Here we work on the eigenvalue space:

$g = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n-1} \lambda_i}$ for a k-dimensional map and a n dimensional dataset

If B is not positive semi-definite (i.e. has negative eigenvalues λ_i):

$g = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|}$ for a k-dimensional map and a n dimensional dataset OR

$g = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{\lambda_i > 0}^{n-1} |\lambda_i|}$

Population substructure

Metric MDS - example in R

Classical application: given a distance matrix (in km or in travel time) between cities, contruct a map of the cities:

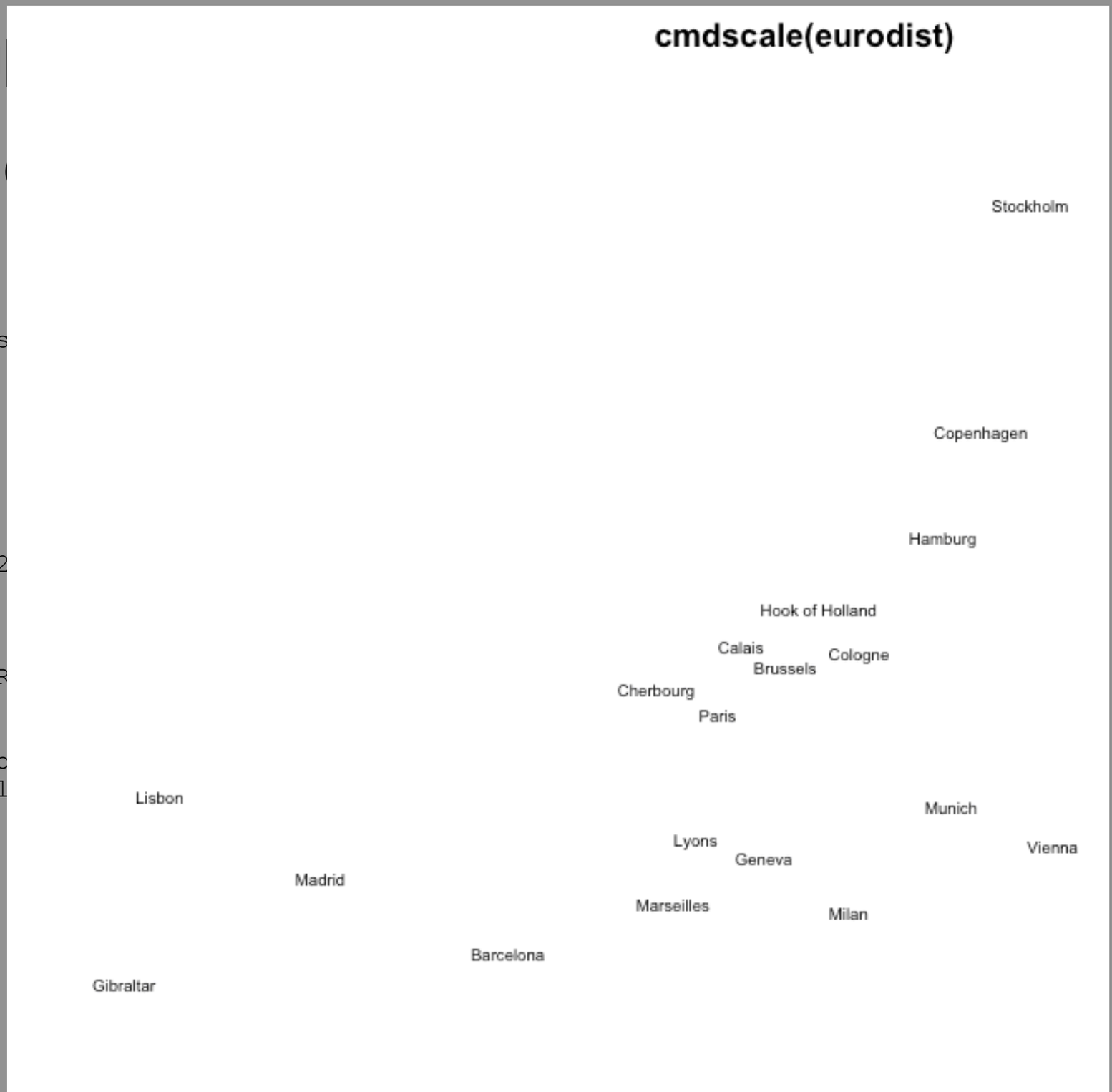
```
> eurodist

               Athens Barcelona Brussels Calais Cherbourg Cologne Copenhagen Geneva …
Barcelona        3313
Brussels         2963      1318
Calais           3175      1326      204
Cherbourg        3339      1294      583    460
Cologne          2762      1498      206    409       785
Copenhagen       3276      2218      966   1136      1545     760
Geneva           2610       803      677    747       853    1662        1418
Gibraltar        4485      1172     2256   2224      2047    2436        3196   1975
Hamburg          2977      2018      597    714      1115     460         460   1118
…

> mds.out <- cmdscale(eurodist, eig=TRUE)

> x <- mds.out$points[,1]
> y <- -mds.out$point[,2] # reflect so North is at the top
> plot(x, y, type = "n", xlab = "", ylab = "", asp = 1, axes = FALSE,
     main = "cmdscale(eurodist)")
> text(x, y, rownames(mds.out), cex = 0.6)

> ev <- mds.out$eig
> gof <- mds.out$GOF
> print(round(gof,digits=4))
[1] 0.7538 0.8679
```

Population substructure

Metric MDS - exam

Classical application: given a distance matrix (
cities:

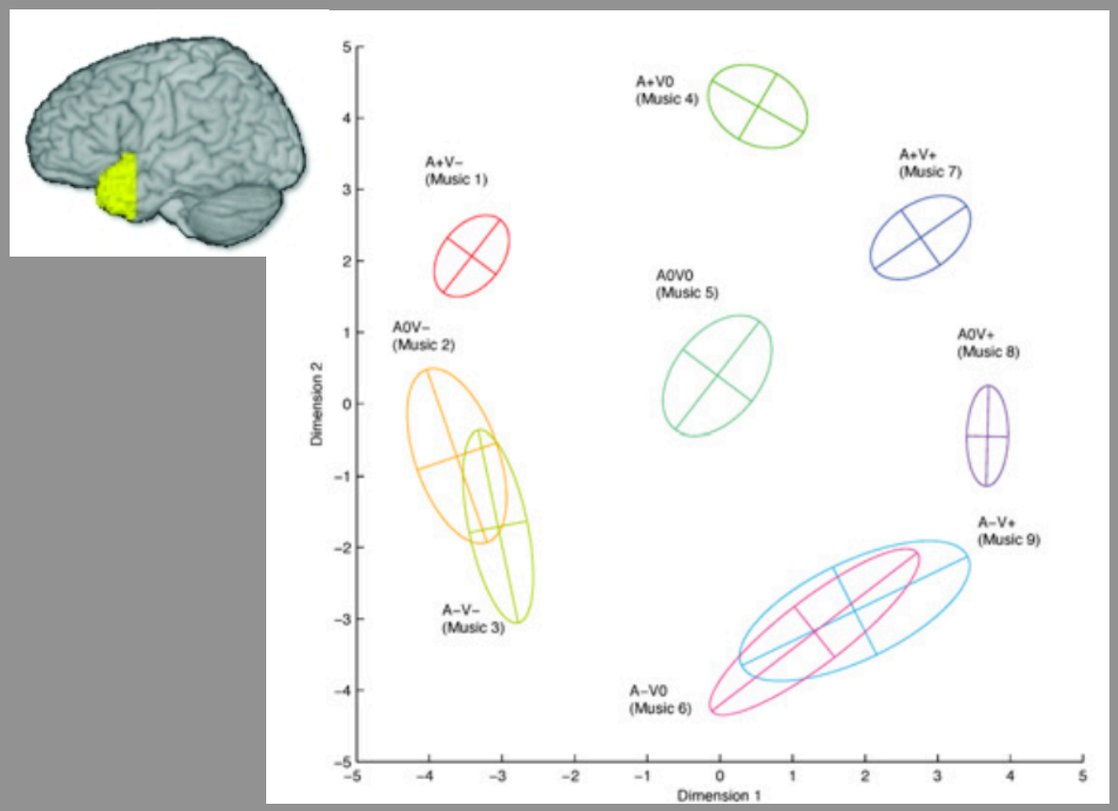
```
> eurodist

               Athens Barcelona Bruss
Barcelona        3313
Brussels         2963      1318
Calais           3175      1326
Cherbourg        3339      1294
Cologne          2762      1498
Copenhagen       3276      2218
Geneva           2610       803
Gibraltar        4485      1172     2
Hamburg          2977      2018
…

> mds.out <- cmdscale(eurodist, eig=TR

> x <- mds.out$points[,1]
> y <- -mds.out$point[,2] # reflect sc
> plot(x, y, type = "n", xlab = "", yl
     main = "cmdscale(eurodist)")
> text(x, y, rownames(mds.out), cex =

> ev <- mds.out$eig
> gof <- mds.out$GOF
> print(round(gof,digits=4))
[1] 0.7538 0.8679
```



- Axes are meaningless
- Orientation of the picture is arbitrary
- The important aspect on MDS plots is…which point is closer to which points

Population substructure

Classical MDS - Examples of distances

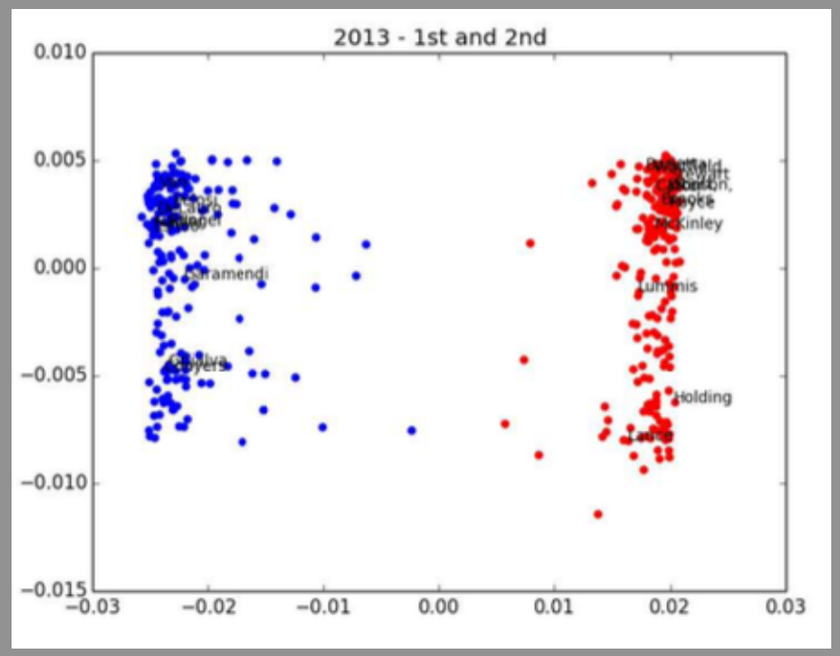


Population: HC (n=12) vs TemporalLobe resection patients (n=14) who lack amygdala, uncus, hypocampus and various amounts of surrounding cortices.

Hypothesis: TL patients show impaired emotional judgement as compared to HC

Distance metrics: How emotionally similar these two songs are?

Conclusion: not impaired implicit emotions vs explicit emotional labelling



Population: republicans (R, red) and democrat (D, blue) members of the House of Representatives.

Hypothesis: Voting patterns of R and D differ

Distance metrics: how many bills legislator i votes differently from legislator j.

Conclusion: distinct voting patterns emerge

Dellacherie, D., Bigand, E., Molin, P., Baulac, M. and Samson, S., 2011. Multidimensional scaling of emotional responses to music in patients with temporal lobe resection. Cortex, 47(9), pp.1107-1115.

Duong-Trong, H., 2014. Multidimensional Scaling. BSc thesis

Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Classical MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

Population substructure

Non-metric MDS

RECALL types of MDS:

- $d_{rs} \approx \delta_{rs}$: 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

Distance measure between points r and s on the solution map

Sample distances

- In classical MDS and metric MDS, the configuration of points is directly obtained from the distances. Both assume that the distances between data points in the original space can be preserved in the lower-dimensional space.
- In non-metric MDS, only the rank order of the distances is important.
- Non-metric MDS do not assume Euclidean distances
- Non-metric MDS is typically implemented as an iterative process that minimise an objective function numerically (STRESS), starting from an initial configuration.

Population substructure

Non-metric MDS - simplified algorithm

Non-metric MDS is typically implemented as an iterative process that minimise an objective function numerically (STRESS), starting from an initial configuration.

Function of the input data

Distance measure between points r and s on the solution map

$$STRESS(x_1, \ldots, x_n; f) = \sqrt{\frac{\sum_{r \neq s}^{n} (f(\delta_{rs}) - d_{rs})^2}{\sum_{r \neq s}^{n} d_{rs}^2}}$$

n = dimensions of the solution

Scaling factor that keeps stress values between 0 and 1

- Goodness of fit statistic that MDS tries to minimize.

- The larger the STRESS, the worse the MDS map compresses the data and thus, the distances in the MDS map are distortions of the input data.
  - If STRESS = 0 the MDS map reproduces perfectly the input data
  - Rule of thumb: anything under STRESS<0,1 it's acceptable.

- Sources of stress
  - Insufficient dimensionality
  - Measurement error
  - Adjustment of the monotonic transformation f

Population substructure

Non-metric MDS - simplified algorithm

Non-metric MDS is typically implemented as an iterative process that minimise an objective function numerically (STRESS), starting from an initial configuration.

$$STRESS(x_1, \ldots, x_n; f) = \sqrt{\frac{\sum_{r \neq s}^{n} (f(\delta_{rs}) - d_{rs})^2}{\sum_{r \neq s}^{n} d_{rs}^2}}$$

Function of the input data

Distance measure between points r and s on the solution map

n = dimensions of the solution

Scaling factor that keeps stress values between 0 and 1

Simplified procedure

1. Choose a distance measure δ_{rs}
2. Choose a monotone transformation f
3. Choose an algorithm to minimize Stress.

stress minimization is usually done iteratively by gradient descent or other methods

Popula

Non-metri

Non-metric MDS
objective functio

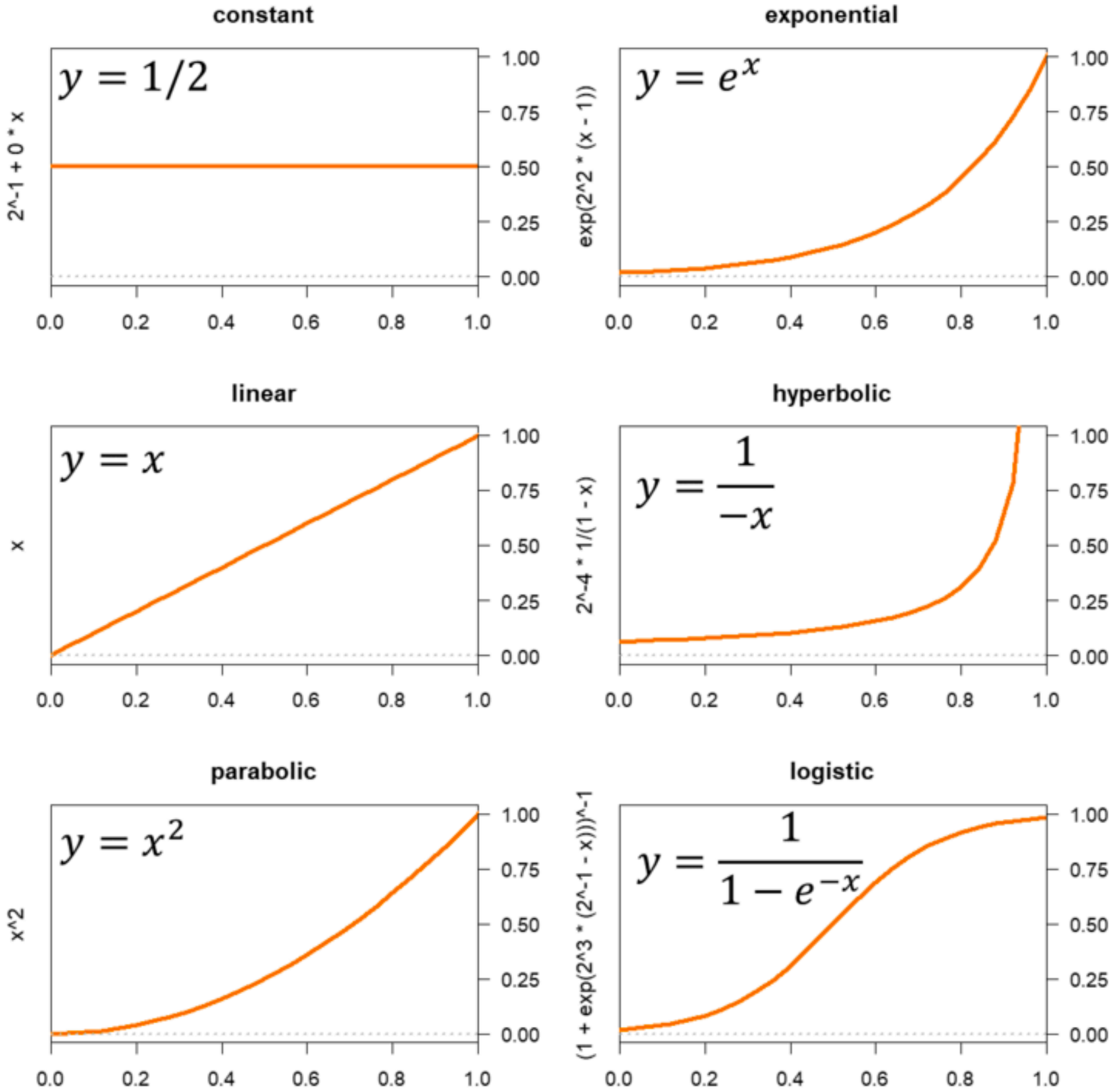
STRESS

Simplified proced

1. Choose a dist
2. Choose a mo
3. Choose an alg

stress mini
descent o

Example of monotonic growth functions



Population substructure

Non-metric MDS

As to better approach the global minima (and avoid getting stuck into a local minima):

- Use different initial configurations (i.e. Initialize randomly by sampling from a normal distribution).

- Compare stress over 1, 2, 3,… dimensional solutions

Population substructure

Non-metric MDS - example in R

Classical application: given a distance matrix (in km or in travel time) between cities, contruct a map of the cities:

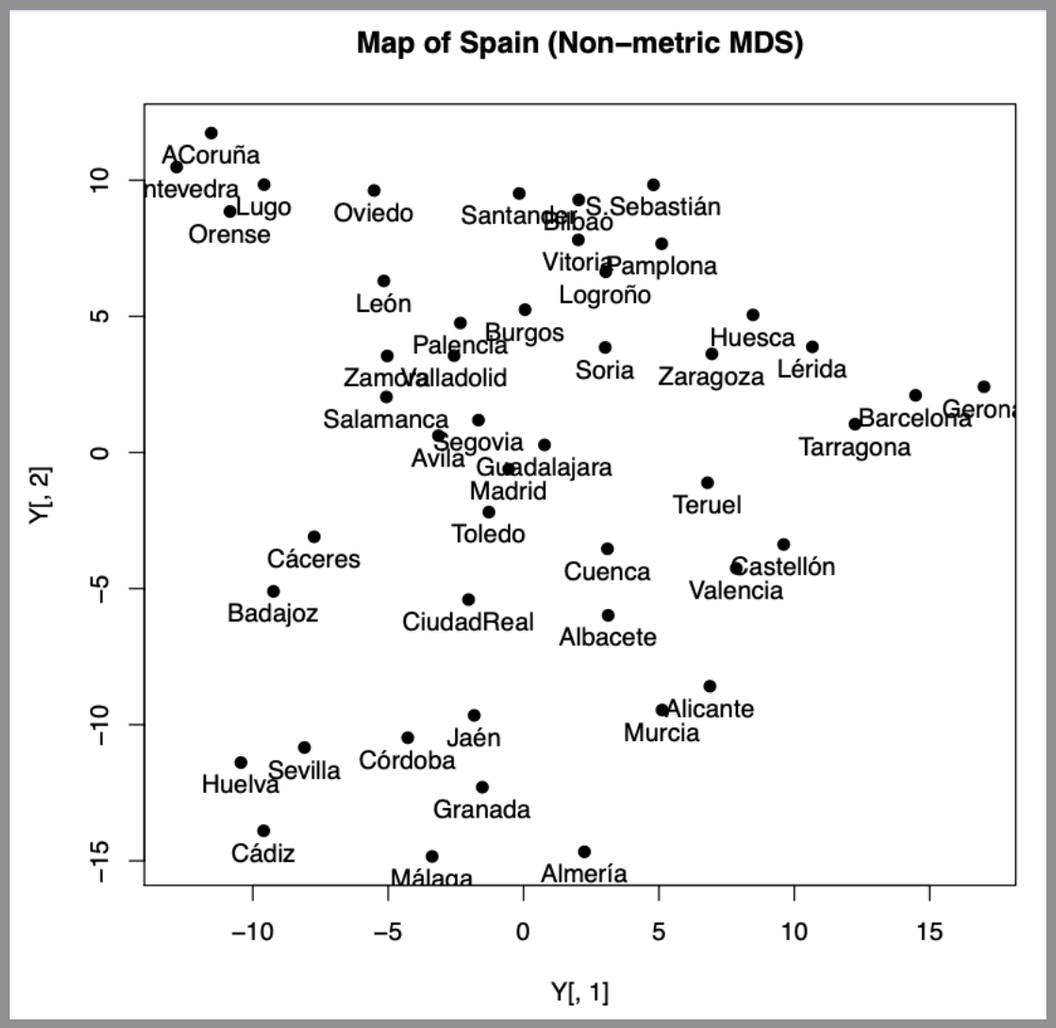
```
> Spain <- as.matrix(read.table("http://www-eio.upc.es/~jan/data/SpainDist.dat"))
> rownames <- Spain[1,]
> Spain <- Spain[-1,]
> n <- nrow(Spain)

> init <- scale(matrix(runif(n*2),ncol=2),scale=FALSE)
> nmmds.out <- isoMDS(Spain,y=init,k=2)

initial  value 41.354230
iter   5 value 38.829957
iter  10 value 28.065912
…
stopped after 50 iterations

OR

> nmmds.out <- isoMDS(Spain,y=init,k=2,maxit=100)
initial  value 41.354230
iter   5 value 38.829957
iter  10 value 28.065912
…
final value 5.057439 ← STRESS achieved in percentage
converged
```

Population substructure

Non-metric MDS - example in R

Classical application: given a distance matrix (in km or in travel time) between cities, contruct a map of the cities:

```
> Y <- nmmds.out$points
> plot(Y[,2],Y[,1],pch=19)
> text(Y[,2], Y[,1], rownames, cex=0.5,pos=1)
```



Some notes:

- Axes are ‘meaningless’, as we loose the ‘distance’ information
- Orientation of the picture is arbitrary
- All that matters is the relative separation between the points.

Population substructure

Non-metric MDS

Diagnostics of MDS:

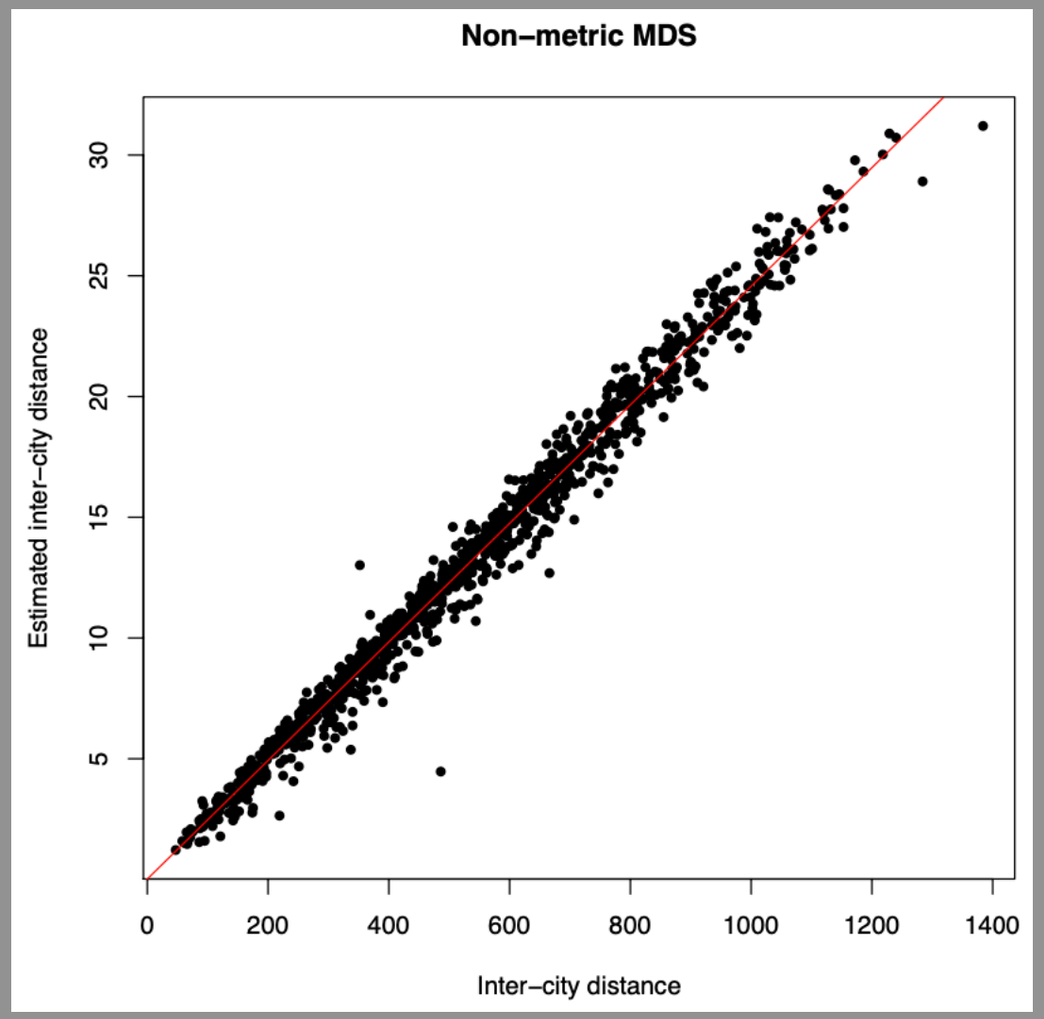
δ_{rs}: distance measure between points r and s on the input data
d_{rs}: distance measure between points r and s on the solution map

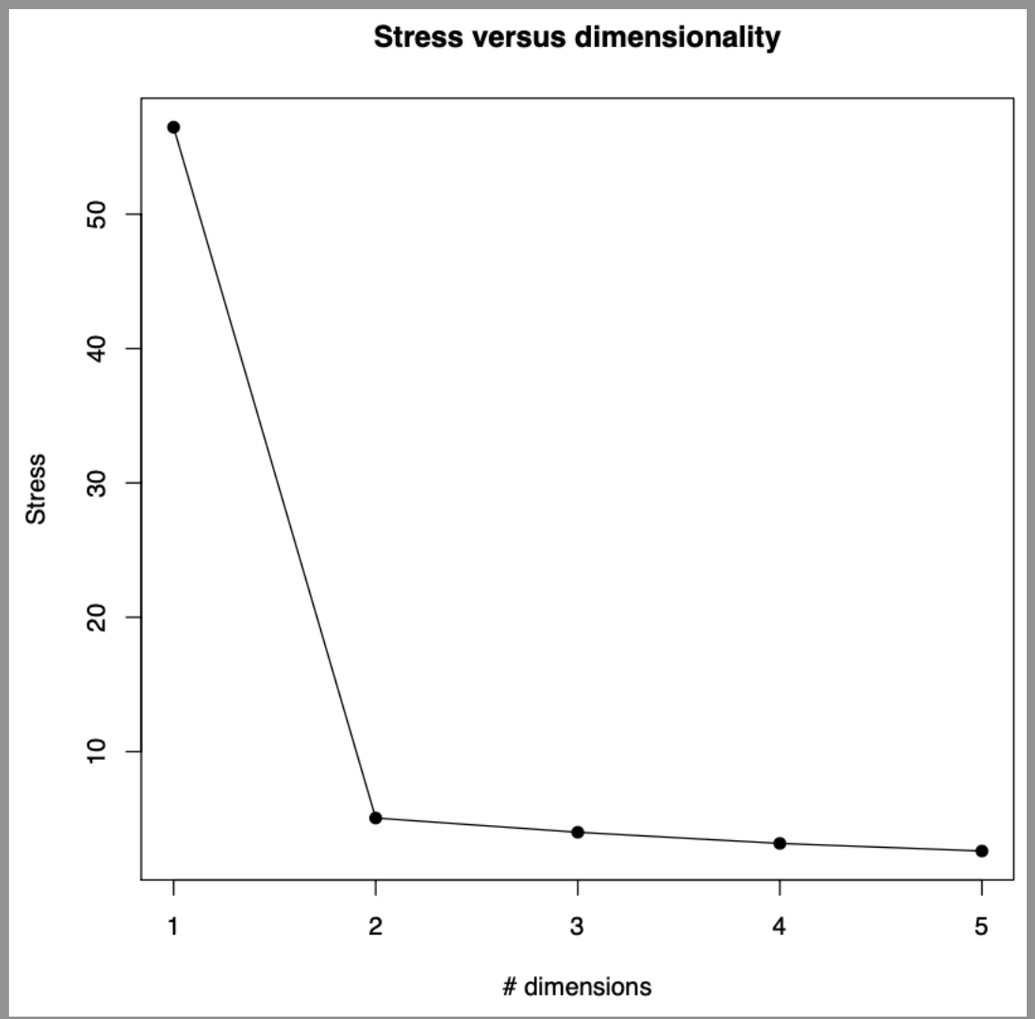
- Scatter plot of δ_{rs} versus d_{rs}
- Plot STRESS versus number of dimensions n
- Degeneracy (how many points with the same d_{rs})
- Compute residuals ($d_{rs} - f(\delta_{rs})$)

Population substructure

Non-metric MDS - example in R

Diagnostics of non-metric MDS solutions:





Population substructure

Metric vs Non-metric MDS

RECALL types of MDS:

- $d_{rs} \approx \delta_{rs}$: 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

- Capture the linear relationships between the data points more accurately than nonmetric MDS
- Easier to interpret
- Sensitive to outliers and noise in the data
- Sensitive to the choice of the distance measure

- Capture the non-linear relationships between the data points more accurately than nonmetric MDS
- More robust to outliers and noise in the data (since it relies on the rank order)
- Difficult to interpret the dimensions of the map
- Sensitive to the choice of the monotonic transformation f
- Computationally intensive

Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Metric MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

Population substructure

Example with genetic data

RECALL MDS 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).

There is a rich literature on how to measure **genetic distance**

SIDE NOTE:
Between species or in populations within species.
Example: length of the shared DNA segments

- Basic idea: populations with many similar alleles have small genetic distances
- Genetic distance measures genetic divergence between two loci, and can mean the degree of differentiation or the divergence time from ancestor.
- The **allele sharing distance** is an often used measure that assigns a numerical value to each allele.

	id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633	rs12534908
1	NA18939	T/G	C/T	G/A	T/G	T/G	G/A
2	NA18940	G/G	T/T	A/A	G/G	T/G	A/A
3	NA18941	G/G	T/T	A/A	G/G	T/G	A/A
4	NA18942	G/G	T/T	A/A	G/G	T/T	A/A
5	NA18943	G/G	T/T	A/A	G/G	T/T	A/A
6	NA18944	T/T	C/C	G/G	T/G	G/G	G/G
7	NA18945	G/G	T/T	A/A	G/G	G/G	A/A
8	NA18946	T/G	C/T	G/A	G/G	G/G	G/A
9	NA18947	T/G	C/T	G/A	G/G	T/G	G/A
10	NA18948	G/G	T/T	A/A	G/G	G/G	A/A
11	NA18949	T/G	C/T	G/A	T/G	T/G	G/A
12	NA18950	G/G	T/T	A/A	G/G	T/G	A/A
13	NA18951	G/G	T/T	A/A	G/G	T/G	A/A
14	NA18952	T/G	C/C	G/G	T/G	T/G	G/G

Population substructure

Example with genetic data

- The **allele sharing distance** is an often used measure:

- Let x_{ijk} be the number of shared alleles of individual i and j for variant k
- Define $d_{ijk} = 2 - x_{ijk}$
- Typically averaged over K genetic variants:

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

- The obtained matrix $D = [d_{ij}]$ is used as input for the MDS

	id	rs34684677	rs1839115	rs4727804	rs4727805	rs200888633	rs12534908
1	NA18939	T/G	C/T	G/A	T/G	T/G	G/A
2	NA18940	G/G	T/T	A/A	G/G	T/G	A/A

Population substructure

Example with genetic data

- The **allele sharing distance:**
- Let x_{ijk}be the number of shared alleles of individual i and j for variant k
- Define $d_{ijk} = 2 - x_{ijk}$
- Typically averaged over K genetic variants

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

- $d_{12,1} = 2 - 1 = 1$ and $d_{12,2} = 2 - 2 = 0$
- $d_{13,1} = 2 - 2 = 0$ and $d_{13,2} = 2 - 1 = 1$
- $d_{12} = \frac{1}{2}(d_{12,1} + d_{12,2}) = \frac{1}{2}(1 + 0)$ and
- $d_{13} = \frac{1}{2}(d_{13,1} + d_{13,2}) = \frac{1}{2}(0 + 1)$
- $d_{23} = \frac{1}{2}(d_{23,1} + d_{23,2}) = \frac{1}{2}(1 + 1) = 1$

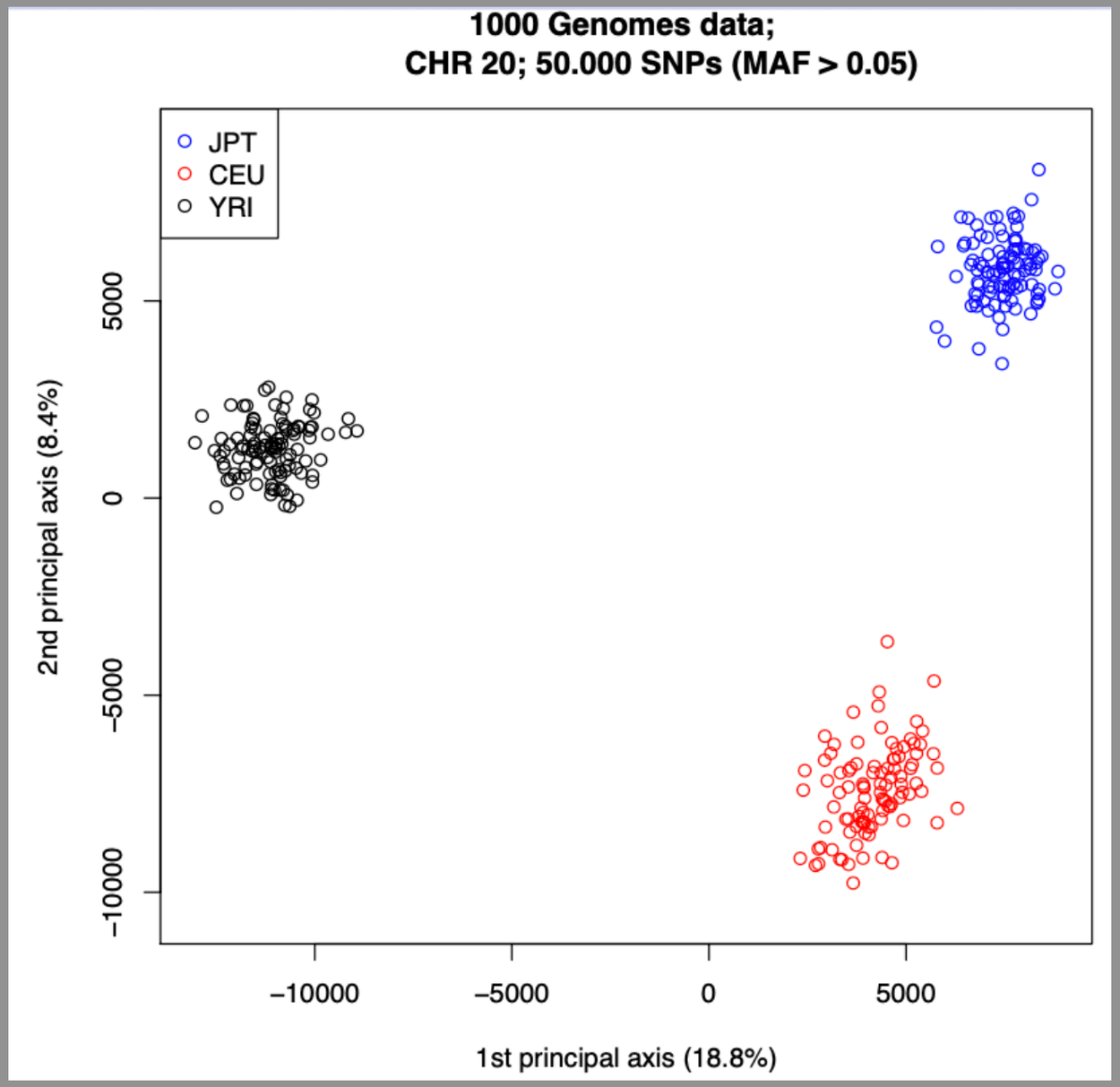
	SNP1	**SNP2**
Ind1	AB	AA
Ind2	BB	AA
Ind3	AB	AB

	Ind1	**Ind2**	**Ind3**
Ind1	0		
Ind2	1/2	0	
Ind3	1/2	1	0

Population substructure

Example with genetic data

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



Things to look at:

- **Clusters**: groups of items that are closer to each other than to other items

Is there a population structure?

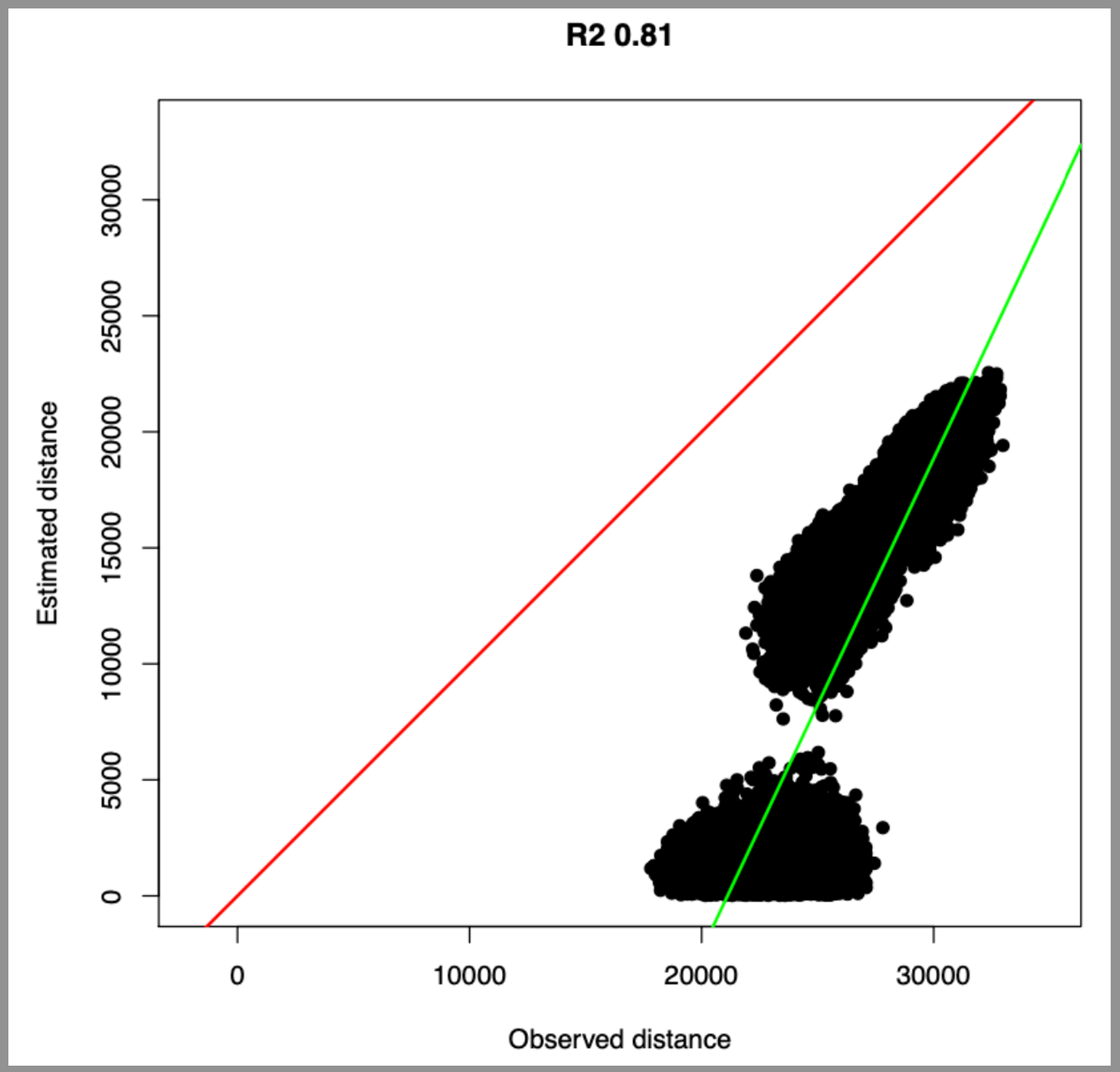
when really tight, highly separated clusters occur in data, it may suggest that each cluster is a domain or subdomain which should be analyzed individually

JPT = Japanese in Tokyo, Japan
CEU = Northern and Western European Ancestry in Utah, US
YRI = Yoruba in Ibadan, Nigeria

Population substructure

Example with genetic data

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



Goodness of fit:

Scatter plot of δ_{rs} versus d_{rs}

Same clusters appear. There is a systematic bias in under-estimating short distances.

Population substructure

Example with genetic data - notes

- In genetics, some pairwise similarity or distance measure between populations is calculated (e.g. distance measures calculated from allele frequencies).
- MDS can also be applied to between-population distances.
- MDS maps made on more homogeneous populations typically show more variability, often revealing overlap between populations.
    - Distances are representations of the relationships given by the data itself.
    - Larger distances are going to be more accurate (recall stress function).
- MDS widely used for detecting the existence of population substructure.

$$STRESS(x_1, \ldots, x_n; f) = \sqrt{\frac{\sum_{r \neq s}^{n} (f(\delta_{rs}) - d_{rs})^2}{\sum_{r \neq s}^{n} d_{rs}^2}}$$

Function of the input data

Distance measure between points r and s on the solution map

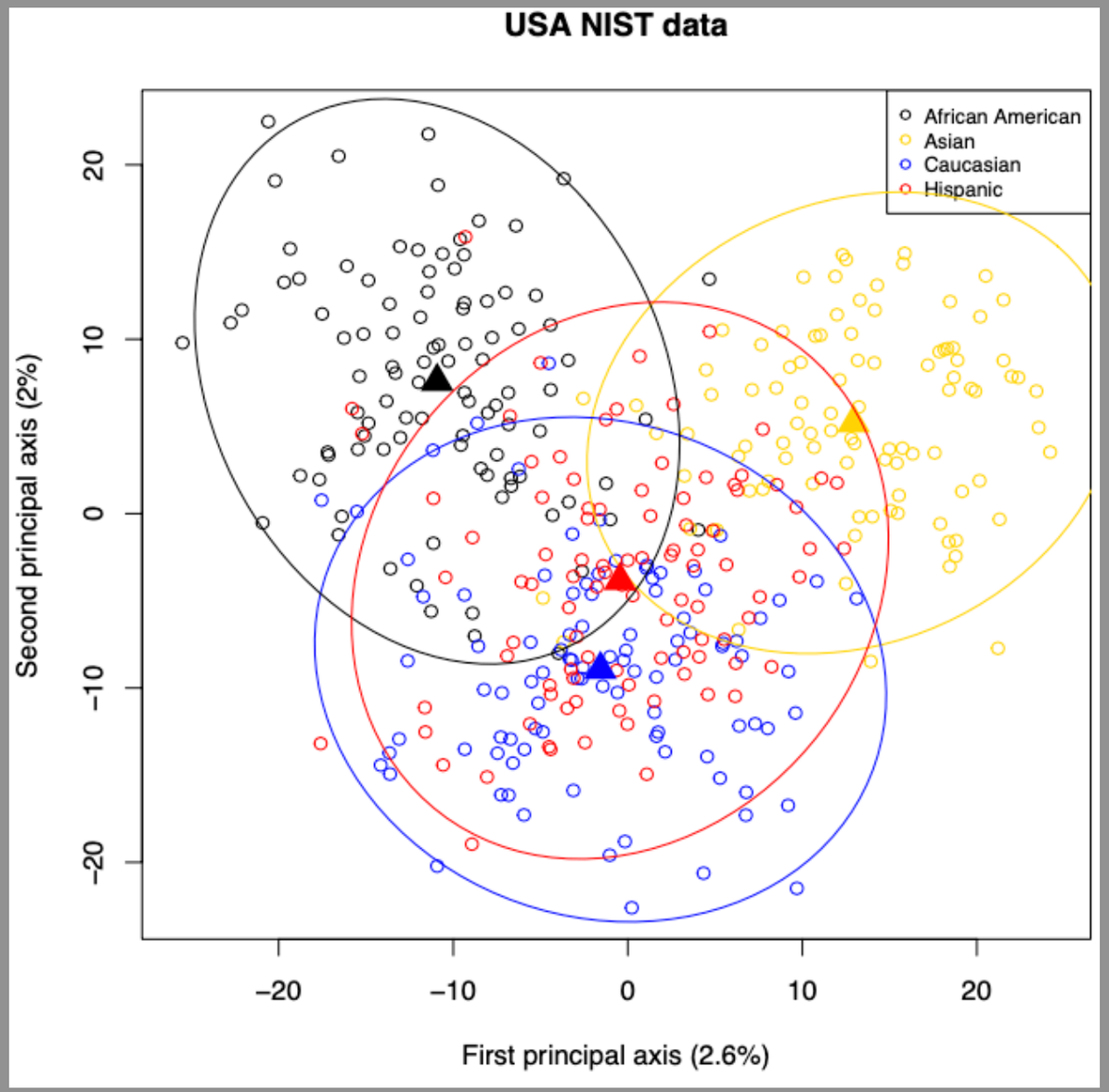
n = dimensions of the solution

Scaling factor that keeps stress values between 0 and 1

Population substructure

Example with genetic data

MDS with US NIST STR data



Short Tandem Repeat database by the National Institute of Standards and Technology

- 23 sequence-based STRs used
- Sample balanced with 97 individuals of four ethnicities: African American, Asian, Caucasian and Hispanics.
- Data coded in (0,1,2) format
- Manhattan distance used (equivalent to allele sharing distance)

Content

Population substructure

1. Introduction to population substructure
2. Introduction to MDS
3. Metric MDS
4. Non-metric MDS
5. Example with genetic data
6. Computer exercise

Population substructure

References

- Visualizing MNIST: An Exploration of Dimensionality Reduction: https://colah.github.io/posts/2014-10-Visualizing-MNIST/

- 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.
- Winkler, C.A., Nelson, G.W. and Smith, M.W., 2010. Admixture mapping comes of age. Annual review of genomics and human genetics, 11, pp.65-89.