CS 260: Foundations of Data Science

Prof. Thao Nguyen Fall 2024



Admin

Lab 5 grades & feedback posted on Moodle

- I will be away for a conference Nov 5-10
 - No lab on Nov 5
 - Prof. Mathieson will teach Nov 6 lecture
 - Will check my email but responses can be delayed

Outline for today

Dimensionality reduction

PCA for data visualization

Outline for today

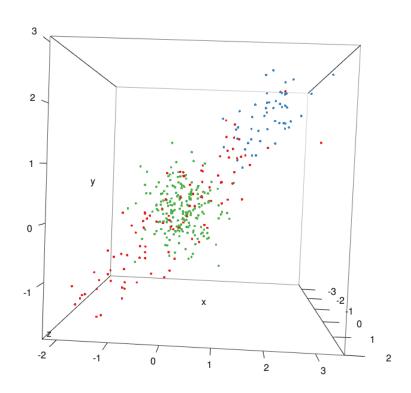
Dimensionality reduction

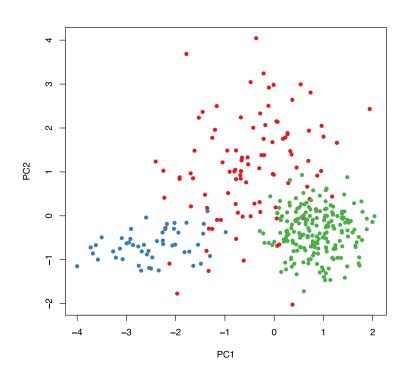
PCA for data visualization

Principal Component Analysis (PCA)

- Transforms p-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on
- PCA is a linear transformation
- Typically, we look at the first few dimensions of the transformed data as a means of dimensionality reduction and visualization
- PCA is often used for:
 - Data visualization
 - Infer qualitative relationships between groups

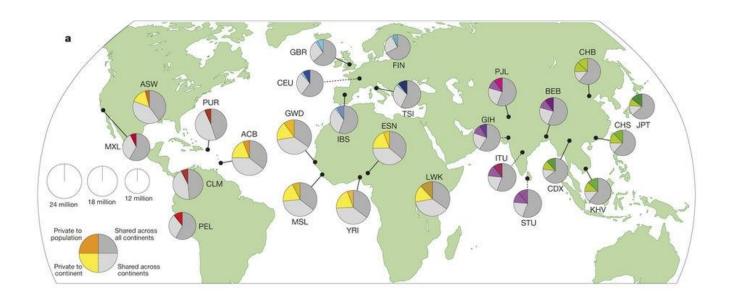
PCA Example





The 1000 Genomes project

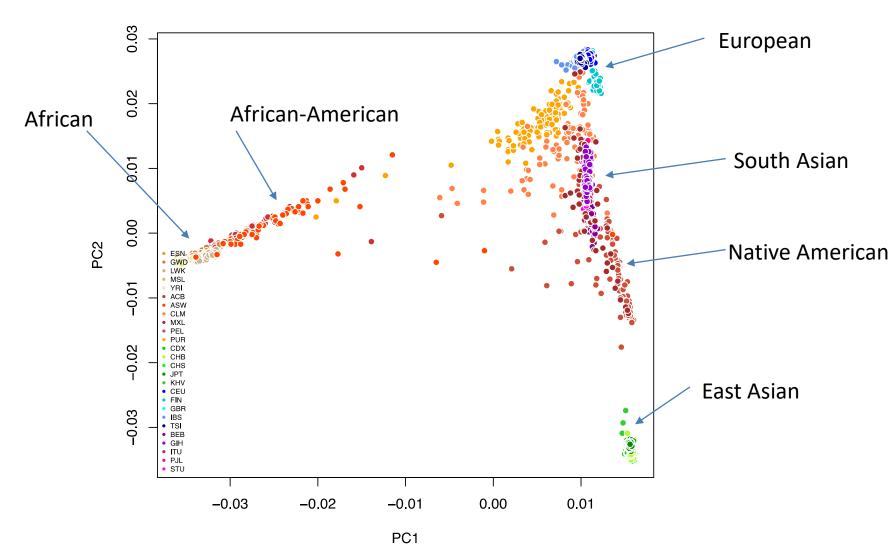
- Whole-genome sequence data from 2504 individuals from 26 populations
- A catalog of human genetic variation, useful as a reference or imputation panel
- Completely public. Download from ftp://ftp-trace.ncbi.nih.gov/1000genomes/



The 1000 Genomes Project Consortium, 2015; https://www.nature.com/articles/nature15393

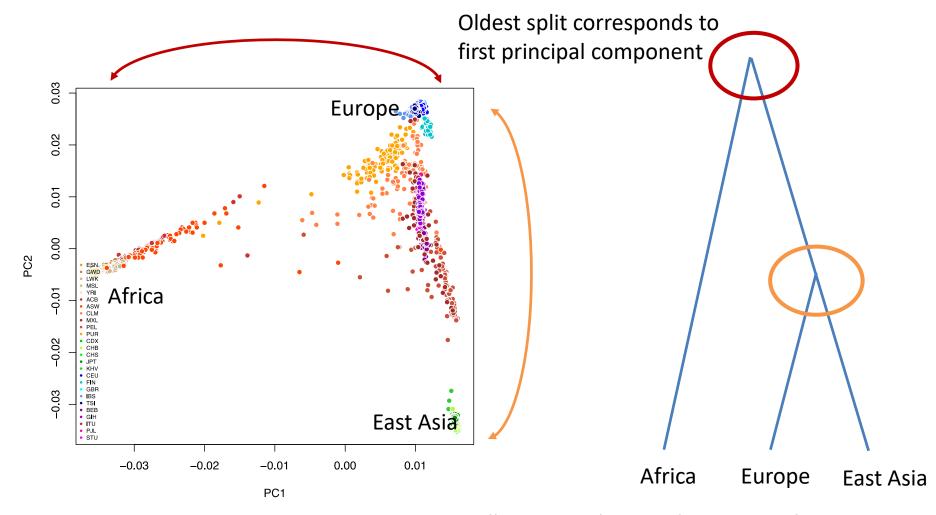
##ALT=<: ##ALT=<: ##ALT=<: ##FORMA	ID=CN121 ID=CN122 ID=CN123 ID=CN124 T= <id=gt< th=""><th>,Description="Co ,Description="Co ,Description="Co ,Description="Co ,Description="Co ,Number=1,Type=5</th><th>opy numbe opy numbe opy numbe opy numbe String,De</th><th>r allelor allelor allelor alleloscription</th><th>e: 121 co e: 122 co e: 123 co e: 124 co</th><th>opies"> opies"> opies"> opies"></th><th></th><th></th><th></th><th colspan="12"></th></id=gt<>	,Description="Co ,Description="Co ,Description="Co ,Description="Co ,Description="Co ,Number=1,Type=5	opy numbe opy numbe opy numbe opy numbe String,De	r allelor allelor allelor alleloscription	e: 121 co e: 122 co e: 123 co e: 124 co	opies"> opies"> opies"> opies">															
##bcftools_annotateVersion=1.6+htslib-1.6 ##bcftools_annotateCommand=annotate -x INFO 20130502_phase3_final/ALL.chr20.phase3_shapeit2_mvncall_integrated_v5.20130502.genotypes.vcf.gz; Date=Fri Jan 19 19:20:16 2018																					
#CHROM	POS	ID REF	ALT	QUAL	FILTER	_	FORMAT					HG00101									HG00111
20	60343	. G	A	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				010	0 0	0 0	0 0	0 0	0 0
20	60419	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	60479	rs149529999	Č	T	100	PASS		GT	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60522	rs150241001	Ť	TC	100	PASS		GT	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	60568	. A	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0	010		010	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60571	rs116145529	Č	A	100	PASS		GT	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	60579	. G	Ā	100	PASS	11133	GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60649	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	60778	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	010	0 0
20	60795	rs184056664	G	C	100	PASS		GT	0 0	0 0	0 0	0 0		0 0		0 0	0 0	0 0	0 0	0 0	0 0
20	60808	. G	Ā	100	PASS	11133	GT	010	0 0	0 0	0 0	0 0		0 0		0 0	0 0	0 0	0 0	0 0	0 0
20	60810	. G	GA	100	PASS		GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60826	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60828	rs187713677	T	G	100	PASS		GT	0 0	0 1	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60864	. G	Å	100	PASS	11133	GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60895	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	60916	. G	Ť	100	PASS		GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61044	Č	Å	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61070	. č	T T	100	PASS		GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61098	rs6078030	Ċ	T	100	PASS		GT	0 0	0 0	0 0	0 0		0 0	1 0	0 0	0 0	0 0	0 0	0 1	0 0
20	61118	. A	Ğ	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	61138	rs140305189	Č	CT	100	PASS		GT	0 0	0 1	0 0	0 0		0 0	0 0	0 1	0 0	0 0	0 0	0 0	0 0
20	61270	rs143291093	A	Č.	100	PASS		GT	0 0	0 0	0 0	0 0		0 0		0 0	0 0	0 0	0 0	0 0	0 0
20	61271	T	A	100	PASS	17133	GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61272	Ċ	A	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61279	rs189899941	Ċ	T	100	PASS		GT	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61329	rs182162684	Č	Ť	100	PASS		GT	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61388	rs146681064	Ť	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61409	rs139103017	Å	Ğ	100	PASS		GT	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61437	. A	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61450	i î	Č	100	PASS		GT	0 0	0 0	0 0	0 0	010		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61517	rs187280035	Č	T	100	PASS		GT	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61538	. A	G	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61638	. ĉ	T	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	61651	rs76553454	Ċ	A	100	PASS		GT	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61711	rs369824431	Ğ	Ï	100	PASS		GT	0 0	0 0	0 0	0 0		0 0		0 0	0 0	0 0	0 0	0 0	0 0
20	61724	rs142532139	A	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	61795	rs4814683	G	Ť	100	PASS		GT	1 0	0 0	0 0	0 0		0 0	1 0	0 1	1 0	0 0	0 0	0 1	0 0
20	61955	. C	Ť	100	PASS		GT	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	61972	Ť	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0	0 0			0 0	0 0	0 0	0 0	0 0	0 0	0 0
20	62100	rs6047235	Ť	C	100	PASS		GT	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20	62174	. AGATCAG	TCCTTT	Ā	100	PASS		GT	010	010	0 0					0 0	010	010	0 0	0 0	010
20		rs192879424	T	C	100	PASS		GT	0 0	0 0	0 0	- 1 -	-1-	- 1 -			0 0	0 0	0 0	0 0	0 0
20	62283	, T	Ċ	100	PASS		GT	0 0	0 0	0 0	0 0					0 0	0 0	0 0	0 0	0 0	0 0
20		rs141113228	A	G	100	PASS		GT	0 0	0 0	0 0	0 0					0 0	0 0	0 0	0 0	0 0
20	62387	. T	A	100	PASS		GT	0 0	0 0	0 0	0 0	0 0				0 0	0 0	0 0	0 0	0 0	0 0
20		rs185326153	A	G	100	PASS		GT	0 0	0 0	0 0					0 0	0 0	0 0	0 0	0 0	0 0
20	62461	. C	T	100	PASS		GT	0 0	0 0	0 0	0 0					0 0	0 0	0 0	0 0	0 0	0 0
20		rs188652106	Ġ	A	100	PASS		GT	0 0	0 0	0 0						0 0	0 0	0 0	0 0	0 0
20		rs192812899	Ā	G	100	PASS		GT	0 0	0 0	0 0					0 0	0 0	0 0	0 0	0 0	0 0
20		rs150267191	C	G	100	PASS		GT	0 0	0 0	0 0						0 0	0 0	0 0	0 0	0 0
20		rs114190700	Ť	Č	100	PASS		GT	0 0	0 0	0 0	0 0	0 0	0 0	0 0		0 0	0 0	0 0	0 0	0 0
	,,								- 1	- 1	- 1	- 1 -	- 1 -				- 1 -	- 1 -	- 1 -		

Global population structure



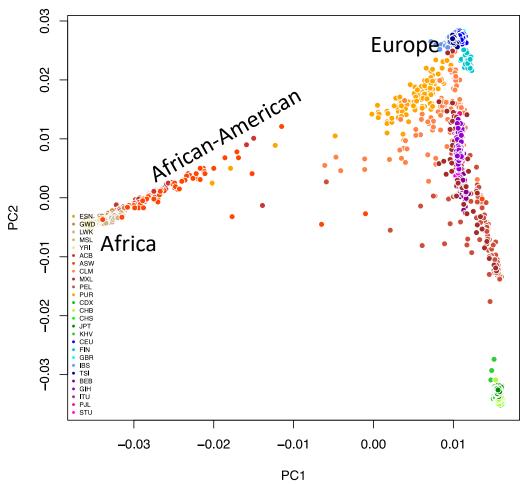
What causes these patterns?

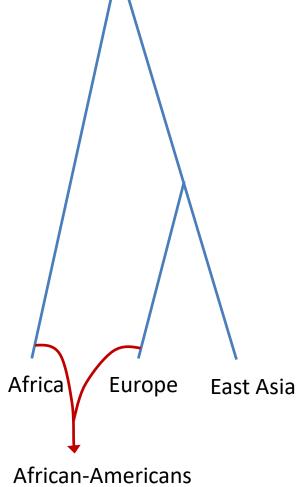
1. Populations splits separate populations



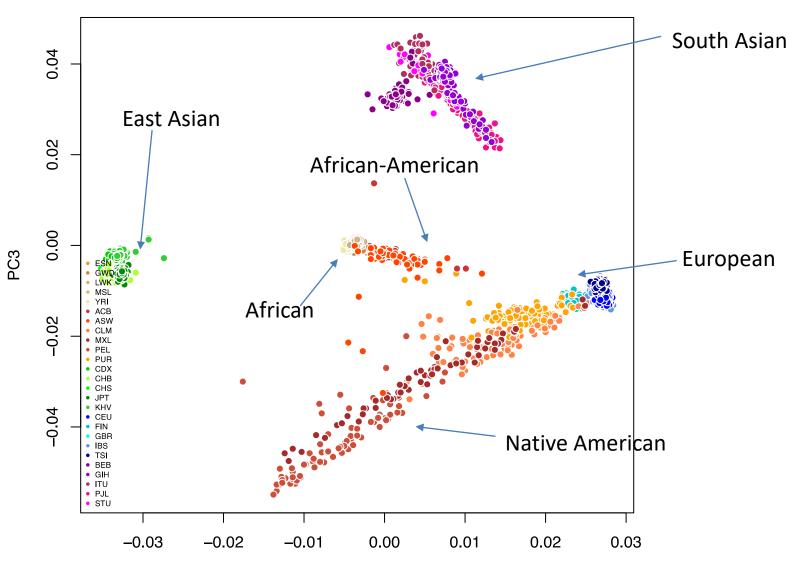
What causes these patterns?







Global population structure



Slide: Iain Mathieson PC2

Genes mirror geography within Europe

John Novembre ⊡, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101(2008) | Cite this article CH ITCH RU TR ES TR IT

PCA application: Eigenfaces









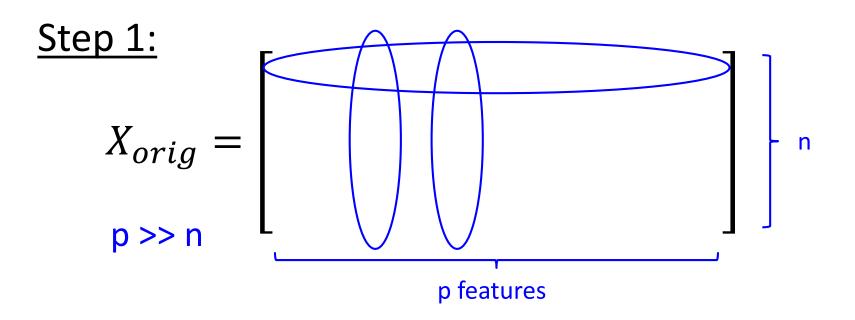
Wikipedia

- Low-dimensional representation of face images
- Used for face recognition/classification

Outline for today

Dimensionality reduction

PCA for data visualization



Goal: Create nx2 matrix for visualization

Step 2: Subtract off column-wise mean

$$X_{orig} = \begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix}$$

$$\overline{x_1} = 2.5 \quad \overline{x_2} = 2$$

$$X = \begin{bmatrix} -0.5 & -1 \\ 0.5 & 1 \end{bmatrix}$$

Step 3: Compute covariance matrix A

$$A = \begin{bmatrix} cov(f,f) & cov(f,g) \\ cov(g,f) & cov(g,g) \end{bmatrix} \text{ 2 features f, g}$$
square &
symmetric Runtime $O(np^2)$

$$cov(f,g) = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \bar{f})(g_i - \bar{g})$$

$$cov(f,f) = var(f) = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \bar{f})^2$$

Step 4: Compute eigenvalues and eigenvectors of A

eigenvalue
$$A\vec{v} = \lambda\vec{v}$$

$$\downarrow$$
eigenvector

$$\det(A - \lambda I) = \vec{0}$$

Solve for λ and plug into first equation to solve for \vec{v}

Step 5: Sort eigenvectors by eigenvalues (high->low)

$$W = \begin{bmatrix} \lambda_1 & \lambda_2 \\ \vdots & \vdots \\ \overrightarrow{v_1} & \overrightarrow{v_2} & \dots & \overrightarrow{v_r} \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{array}{c} pxr \\ \text{usually } r = 2 \\ \vdots & \vdots & & \vdots \end{array}$$
first eigenvector

And compute the transformed data:

$$T_{nxr} = X_{nxp} W_{pxr}$$

Handout 16

VO((S,) (O)(f,,fz) Handon't 16

pex (a b) = ad-bc $\left(\frac{3}{10}-\lambda\right)^2-\left(3/\sqrt{2}\right)^2=0$ 15-5-3/+/

--0 Tz=XV <u>ල</u>