CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025

HAVERFORD COLLEGE

Materials by Sara Mathieson

Lab 7 notes

Logistic regression cost function:

$$J(\overrightarrow{w}) = -\sum_{i=1}^{n}[y_i \log(h_{\overrightarrow{w}}(\overrightarrow{x_i})) + (1 - y_i)\log(1 - h_{\overrightarrow{w}}(\overrightarrow{x_i}))]$$

if $h_{\overrightarrow{w}}(\overrightarrow{x_i}) = 0$ or $1 - h_{\overrightarrow{w}}(\overrightarrow{x_i}) = 0 \rightarrow$ skip log(0) or add 0

Outline for today

- Dimensionality reduction
- PCA for data visualization

Outline for today

- Dimensionality reduction
- PCA for data visualization

Principal Component Analysis (PCA)

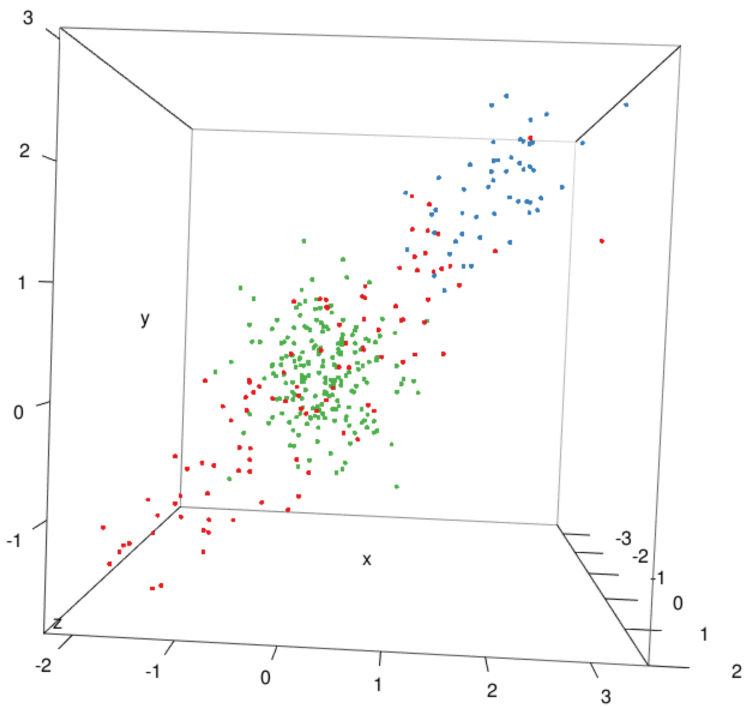
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

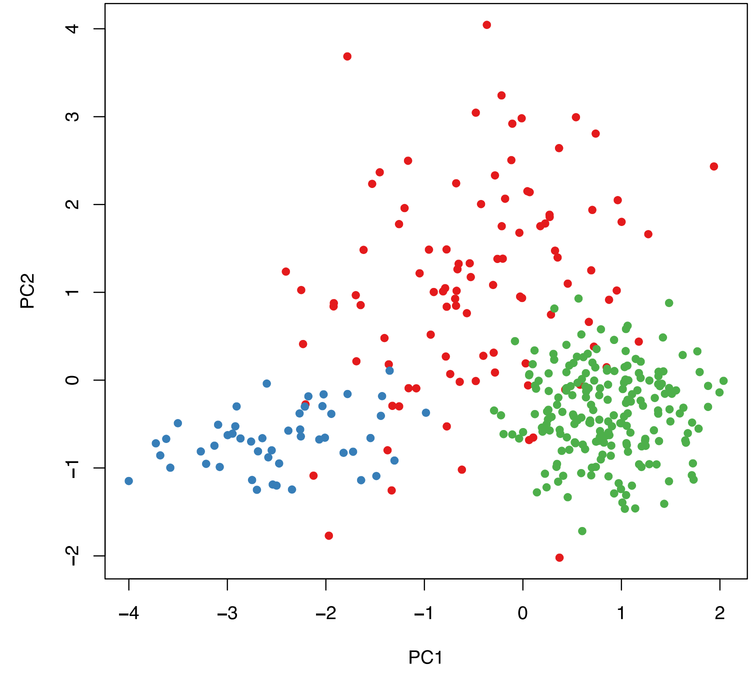
- PCA is a linear transformation

- 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

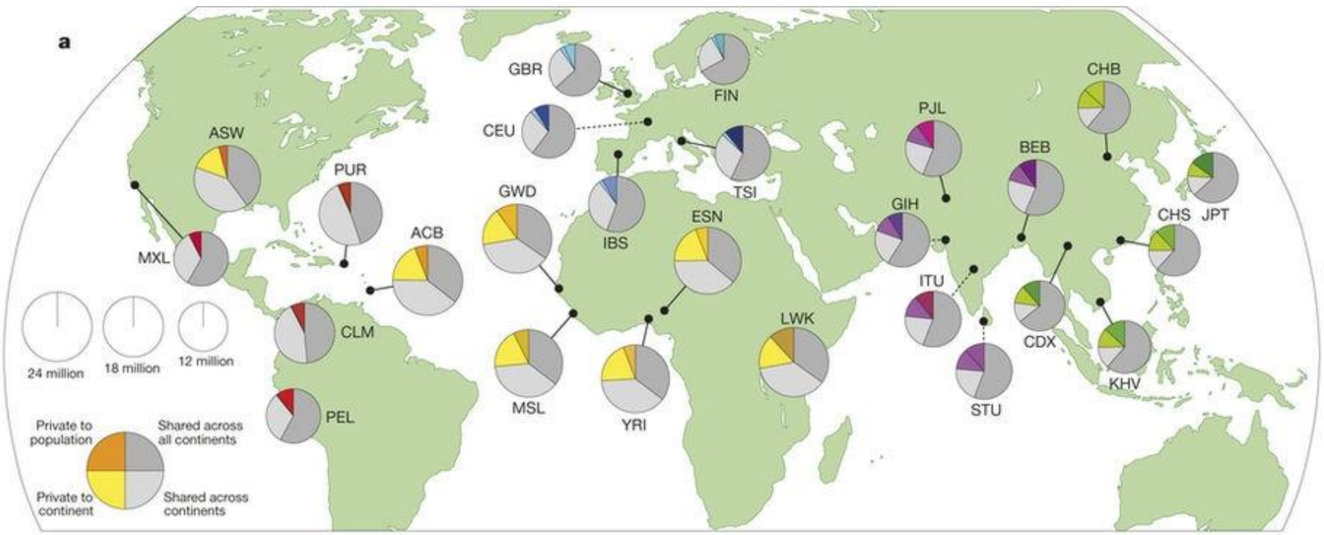




Slide: Iain Mathieson

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/

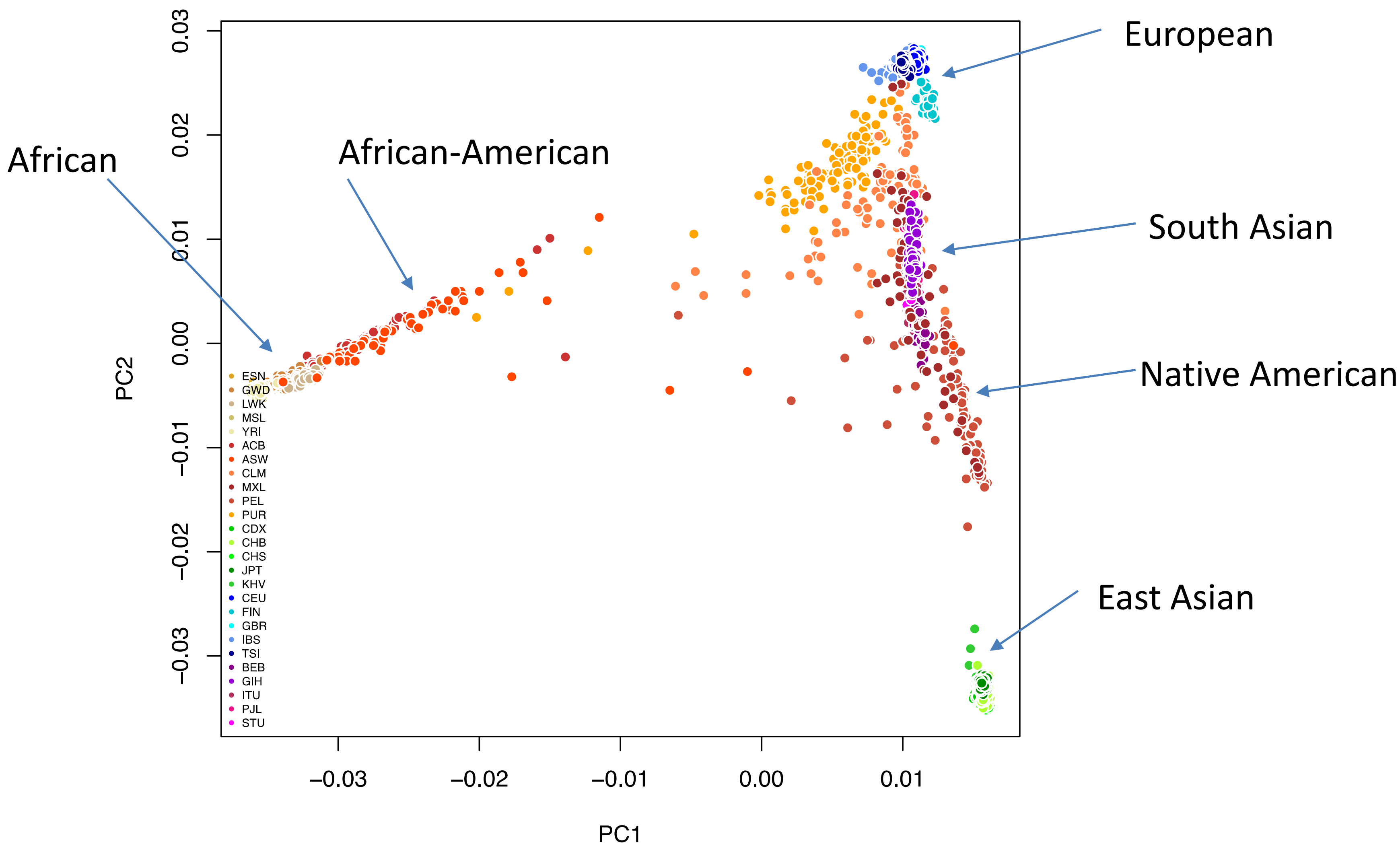


Slide: Iain Mathieson

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

```
##ALT=<ID=CN120,Description="Copy number allele: 120 copies">
##ALT=<ID=CN121,Description="Copy number allele: 121 copies">
##ALT=<ID=CN122,Description="Copy number allele: 122 copies">
##ALT=<ID=CN123,Description="Copy number allele: 123 copies">
##ALT=<ID=CN124,Description="Copy number allele: 124 copies">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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	INFO	FORMAT	HG00096	HG00097	HG00099	HG00100	HG00101	HG00102	HG00103	HG00105	HG00106	HG00107	HG00108	HG00109	HG00110	HG00111
20	60343	.	G	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	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	0|0	0|0	0|0
20	60479	rs149529999	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	0|0	0|0	0|0
20	60522	rs150241001	T	TC	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	60568	.	A	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	0|0	0|0	0|0
20	60571	rs116145529	C	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	60579	.	G	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	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	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	0|0	0|0	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	0|0	0|0
20	60808	.	G	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	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	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	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	0|0
20	60864	.	G	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	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	0|0
20	60916	.	G	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	0|0	0|0
20	61044	.	C	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	0|0
20	61070	.	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	0|0	0|0	0|0	0|0
20	61098	rs6078030	C	T	100	PASS	.	GT	0|0	0|0	0|0	0|0	0|1	0|0	1|0	0|0	0|0	0|0	0|0	0|1	0|0
20	61118	.	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	0|0
20	61138	rs140305189	C	CT	100	PASS	.	GT	0|0	0|1	0|0	0|0	0|1	0|0	0|0	0|1	0|0	0|0	0|0	0|0	0|0
20	61270	rs143291093	A	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	0|0	0|0
20	61271	.	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	0|0	0|0	0|0
20	61272	.	C	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	0|0
20	61279	rs189899941	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	0|0	0|0	0|0
20	61329	rs182162684	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	0|0	0|0	0|0
20	61388	rs146681064	T	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	0|0	0|0
20	61409	rs139103017	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	61437	.	A	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	0|0	0|0	0|0
20	61450	.	T	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	0|0	0|0	0|0
20	61517	rs187280035	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	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	0|0
20	61638	.	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	0|0	0|0	0|0	0|0
20	61651	rs76553454	C	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	61711	rs369824431	G	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	0|0
20	61724	rs142532139	A	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	0|0	0|0
20	61795	rs4814683	G	T	100	PASS	.	GT	1|0	0|0	0|0	0|0	0|1	0|0	1|0	0|1	1|0	0|0	0|0	0|1	0|0
20	61955	.	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	0|0	0|0	0|0	0|0
20	61972	.	T	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	0|0	0|0	0|0
20	62100	rs6047235	T	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	0|0	0|0
20	62174	.	AGATCAGTCCTTT	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	62255	rs192879424	T	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	0|0	0|0
20	62283	.	T	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	0|0	0|0	0|0
20	62348	rs141113228	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	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	0|0	0|0	0|0
20	62420	rs185326153	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	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	0|0	0|0	0|0	0|0
20	62471	rs188652106	G	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	62478	rs192812899	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	62545	rs150267191	C	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	62553	rs114190700	T	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	0|0	0|0
```

Global population structure

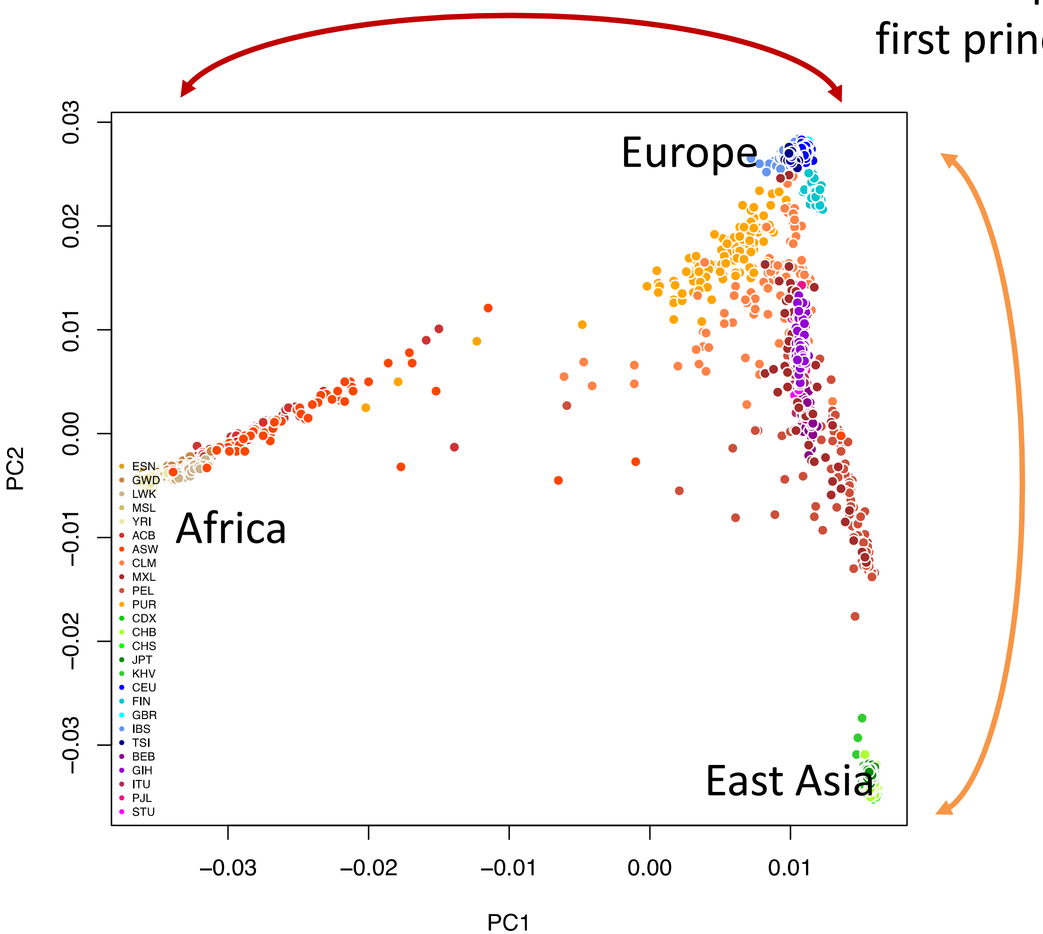


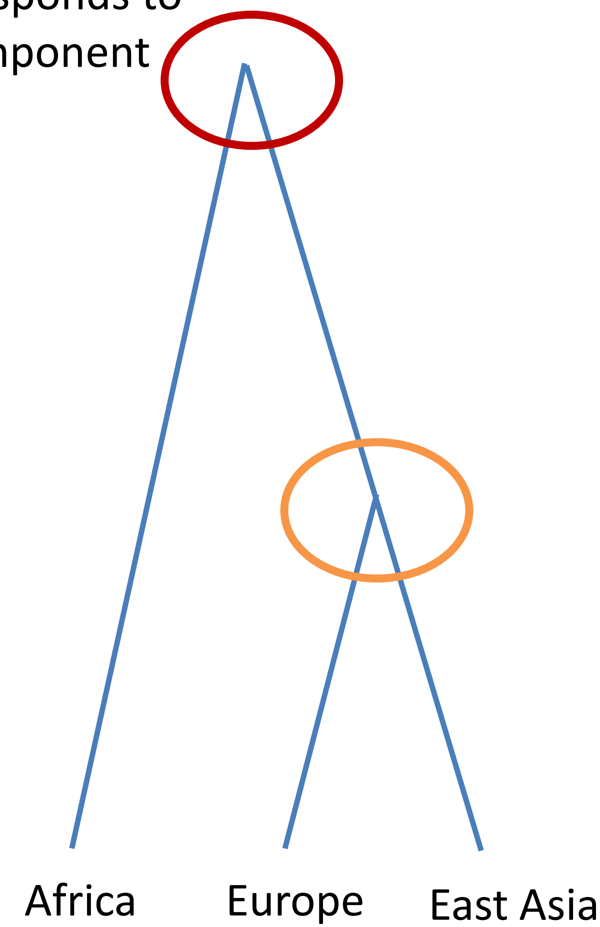
Slide: Iain Mathieson

What causes these patterns?

1. Populations splits separate populations

Oldest split corresponds to first principal component



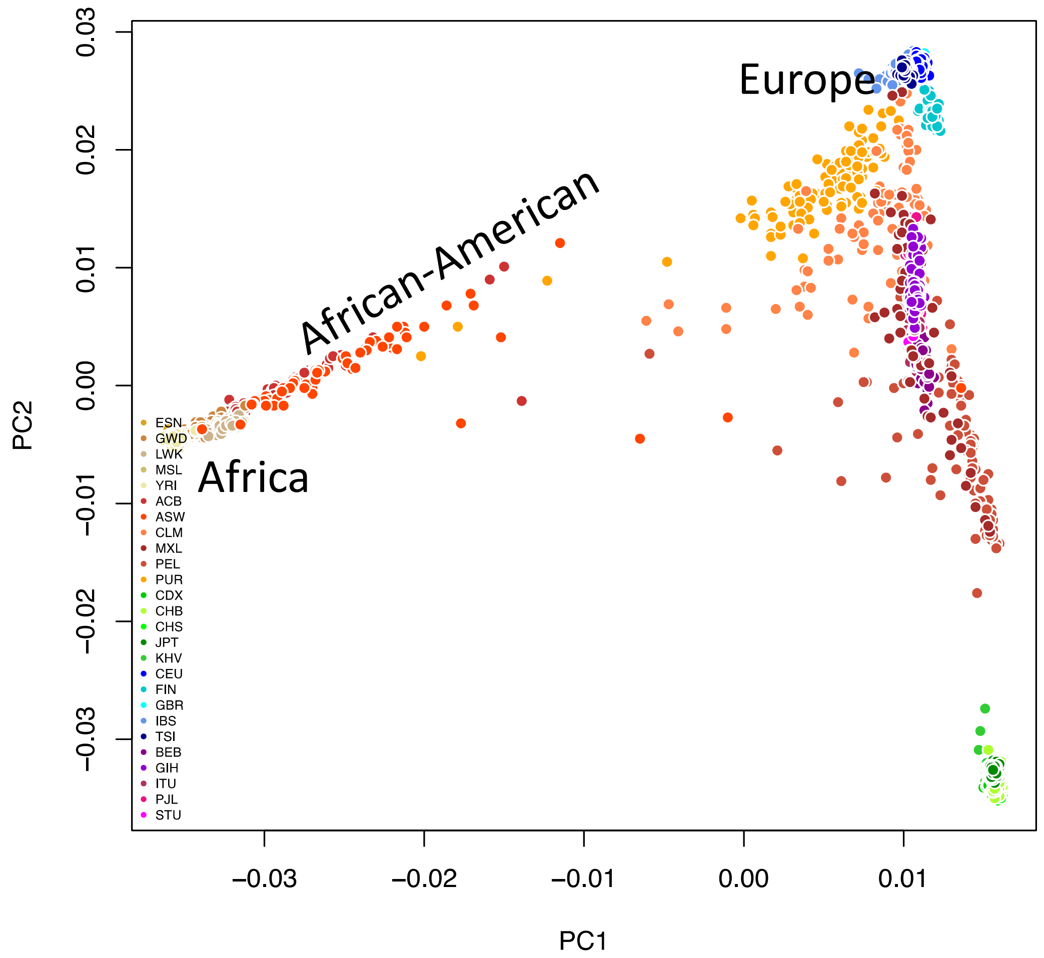


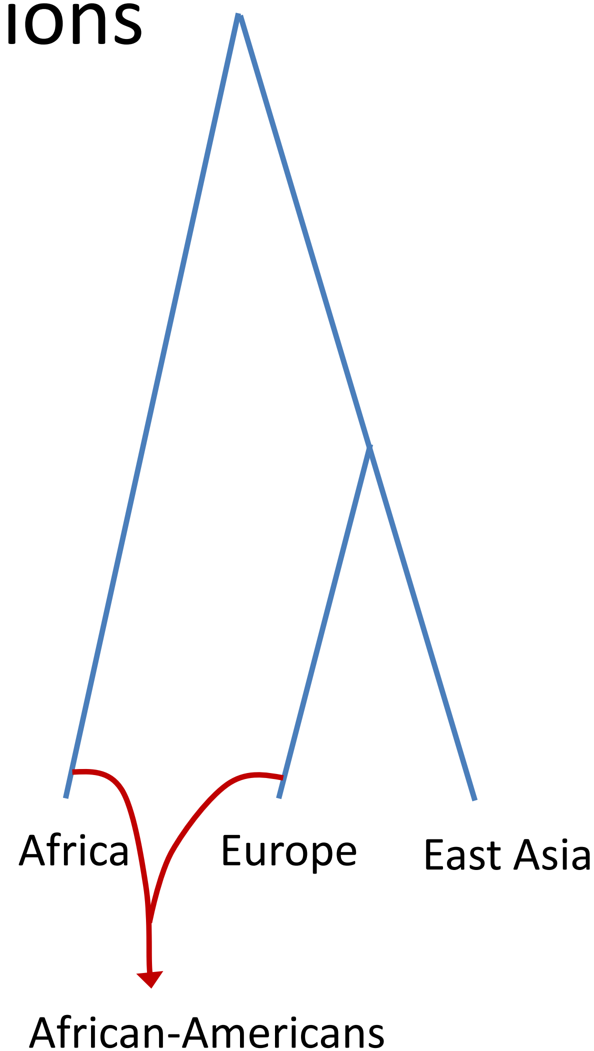
Slide: Iain Mathieson

McVean 2009: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686

What causes these patterns?

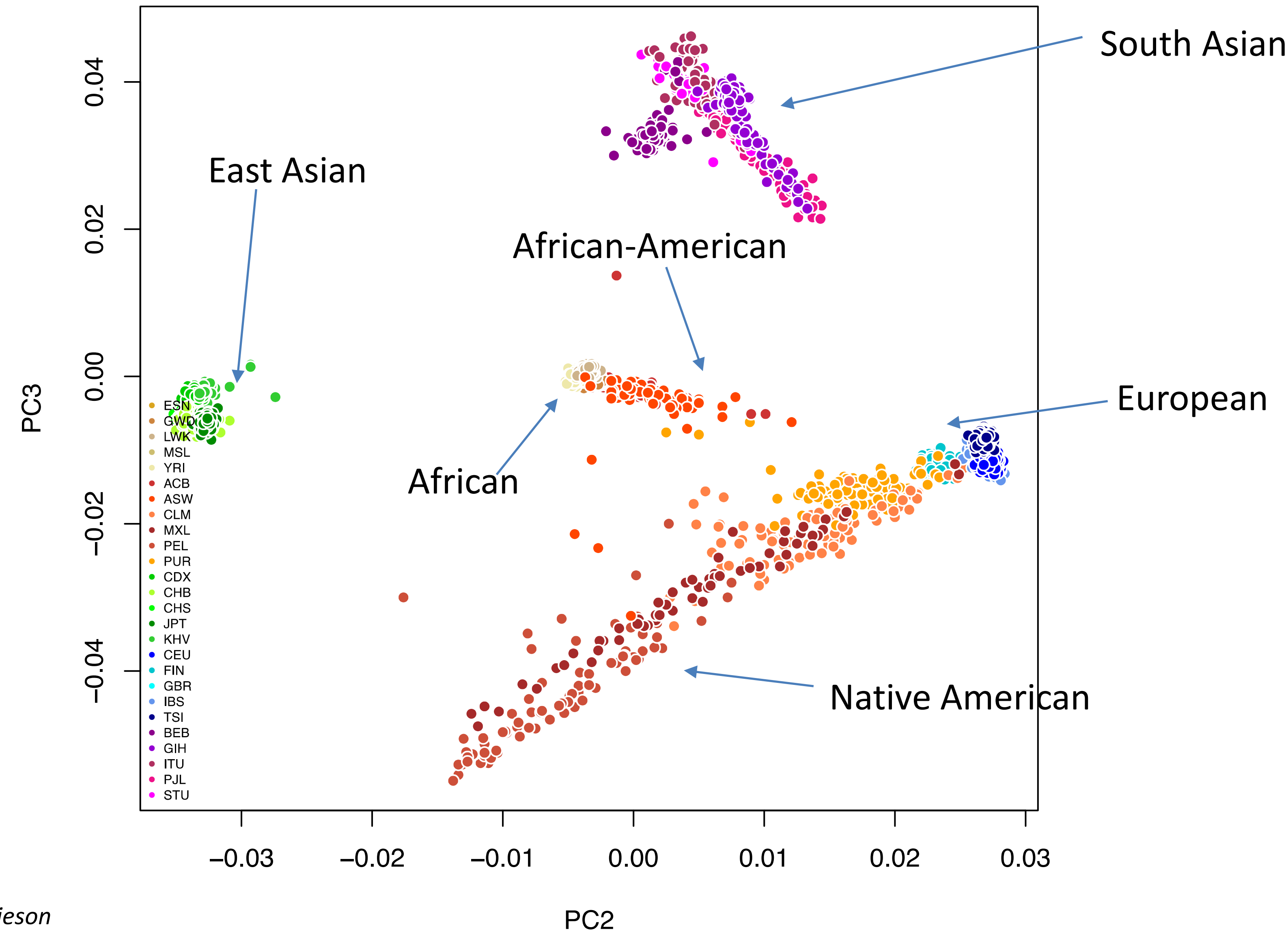
2. Admixture merges populations





Slide: Iain Mathieson

Global population structure

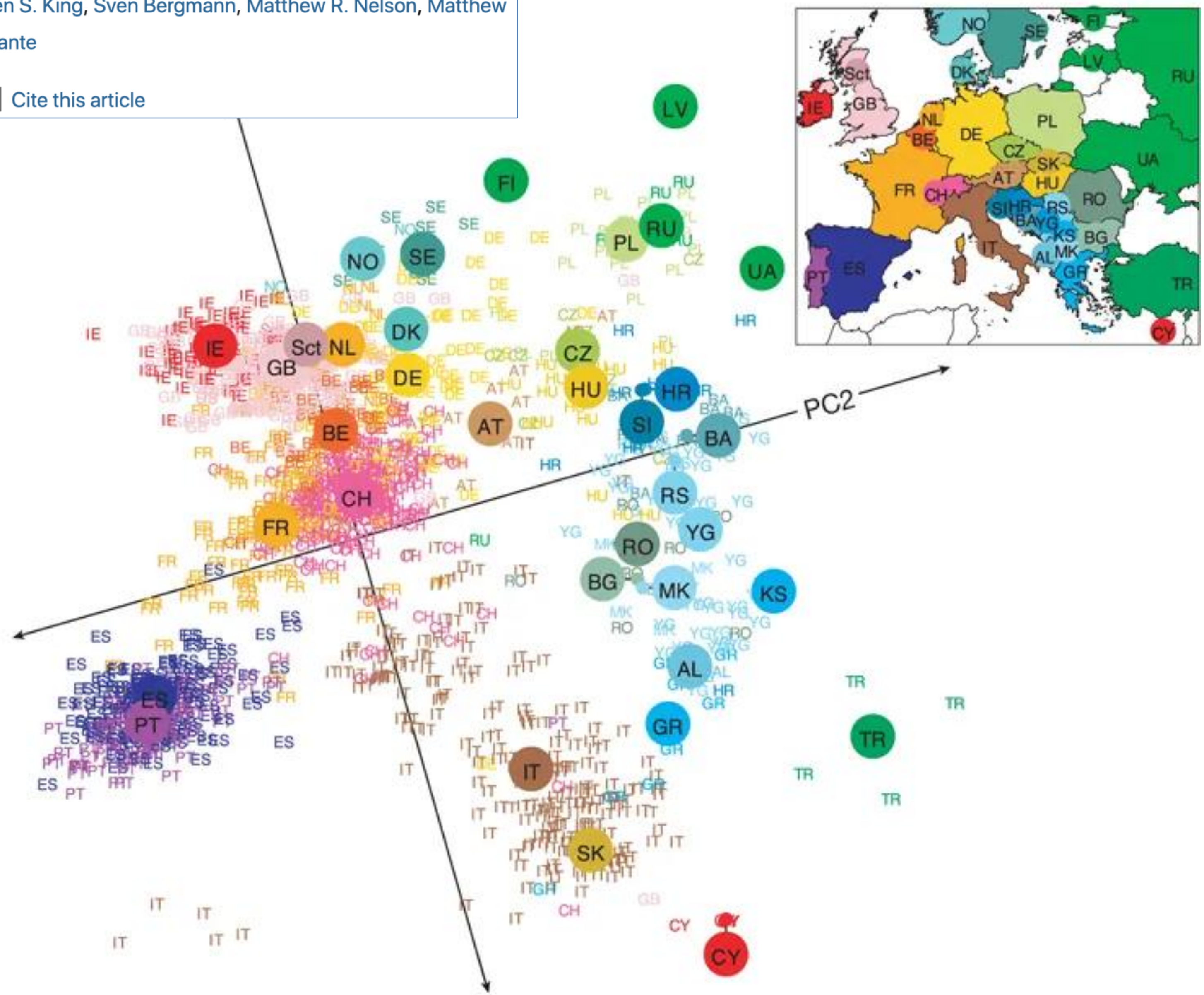


Slide: Iain Mathieson

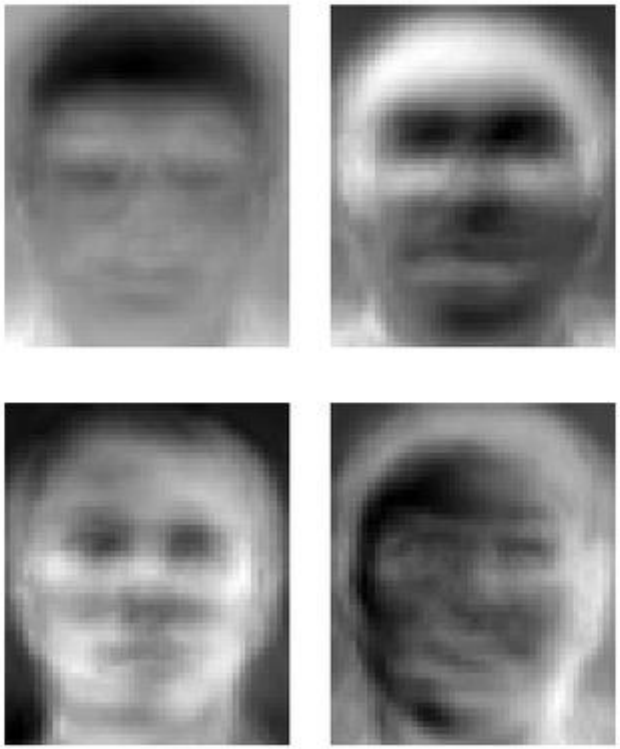
Genes mirror geography within Europe

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



PCA application: Eigenfaces



Wikipedia

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

Outline for today

- Dimensionality reduction
- PCA for data visualization

PCA Algorithm

<u>Step 1:</u>

$$X_{orig} = \left.\underbrace{\begin{bmatrix} \quad & \quad & \quad & \quad \\ \quad & \quad & \quad & \quad \\ \quad & \quad & \quad & \quad \end{bmatrix}}_{\text{p features}}\right\} n$$

p >> n

Goal: Create nx2 matrix for visualization

PCA Algorithm

<u>Step 2:</u> Subtract off column-wise mean

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

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

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

PCA Algorithm

<u>Step 3:</u> Compute covariance matrix A

$$A = \begin{bmatrix} cov(f,f) & cov(f,g) \\ cov(g,f) & cov(g,g) \end{bmatrix}$$ 2 features f, g

square & symmetric

Runtime $O(np^2)$

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

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

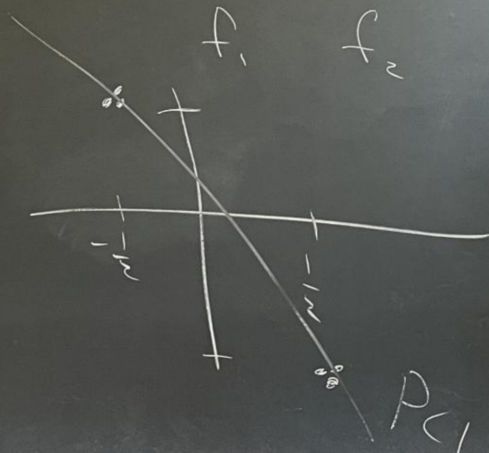
Handout 15, page 1

Handout 16

$\boxed{\text{Step 1 \& 2}}$

$$X = \begin{array}{cc} \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \\ f_1 \quad\quad f_2 \end{array}$$

orig $\begin{cases} \bar{f}_1 = \frac{1}{2} \\ \bar{f}_2 = \frac{1}{2} \end{cases}$



$\boxed{\text{Step 3}}$

$$A = \begin{bmatrix} var(f_1) & cov(f_1, f_2) \\ cov(f_2, f_1) & var(f_2) \end{bmatrix}$$

$\bar{f}_1 = 0$

$\bar{f}_2 = 0$

$$cov(f_1, f_2) = \frac{1}{6-1}\left(-\frac{1}{2}\cdot\frac{1}{2}\right)\cdot 6$$

$$= -\frac{3}{10}$$

$$\Rightarrow A = \begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix}$$

PCA Algorithm

Step 4: Compute eigenvalues and eigenvectors of A

eigenvalue

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

eigenvector

$$\det(A - \lambda I) = 0$$

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

PCA Algorithm

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

$$W = \begin{array}{c} \begin{array}{cccc} \lambda_1 & \lambda_2 & & \end{array} \\ \begin{bmatrix} \vdots & \vdots & & \vdots \\ \overrightarrow{v_1} & \overrightarrow{v_2} & \cdots & \overrightarrow{v_r} \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array} \qquad pxr$$

usually $r = 2$

first eigenvector

And compute the transformed data:

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

Handout 15, page 2

Step 4

$$\det(A - \lambda I) = 0$$

$$\left(\begin{bmatrix} 3/10 & -3/10 \\ -3/10 & 3/10 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = 0$$

6

$$\det \left(\begin{bmatrix} 3/10 - \lambda & -3/10 \\ -3/10 & 3/10 - \lambda \end{bmatrix} \right) = 0$$

$$\left(\frac{3}{10} - \lambda \right)^2 - (3/10)^2 = 0$$

$$\cancel{\left(\frac{3}{10} \right)^2} - 2 \cdot \frac{3}{10} \lambda + \lambda^2 - \cancel{\left(\frac{3}{10} \right)^2} = 0$$

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

$$\Rightarrow \boxed{\lambda^2 - \frac{3}{5} \lambda = 0}$$

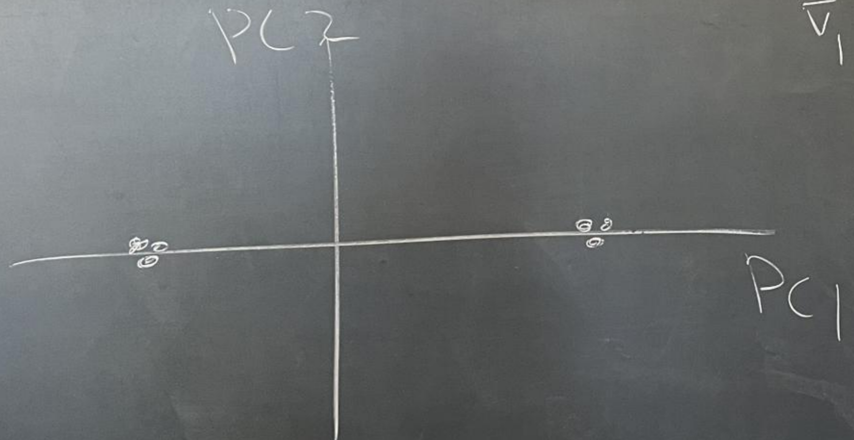
$$\lambda \left(\lambda - \frac{3}{5} \right) = 0 \Rightarrow \boxed{\begin{aligned} \lambda_1 &= \frac{3}{5} \\ \lambda_2 &= 0 \end{aligned}}$$

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

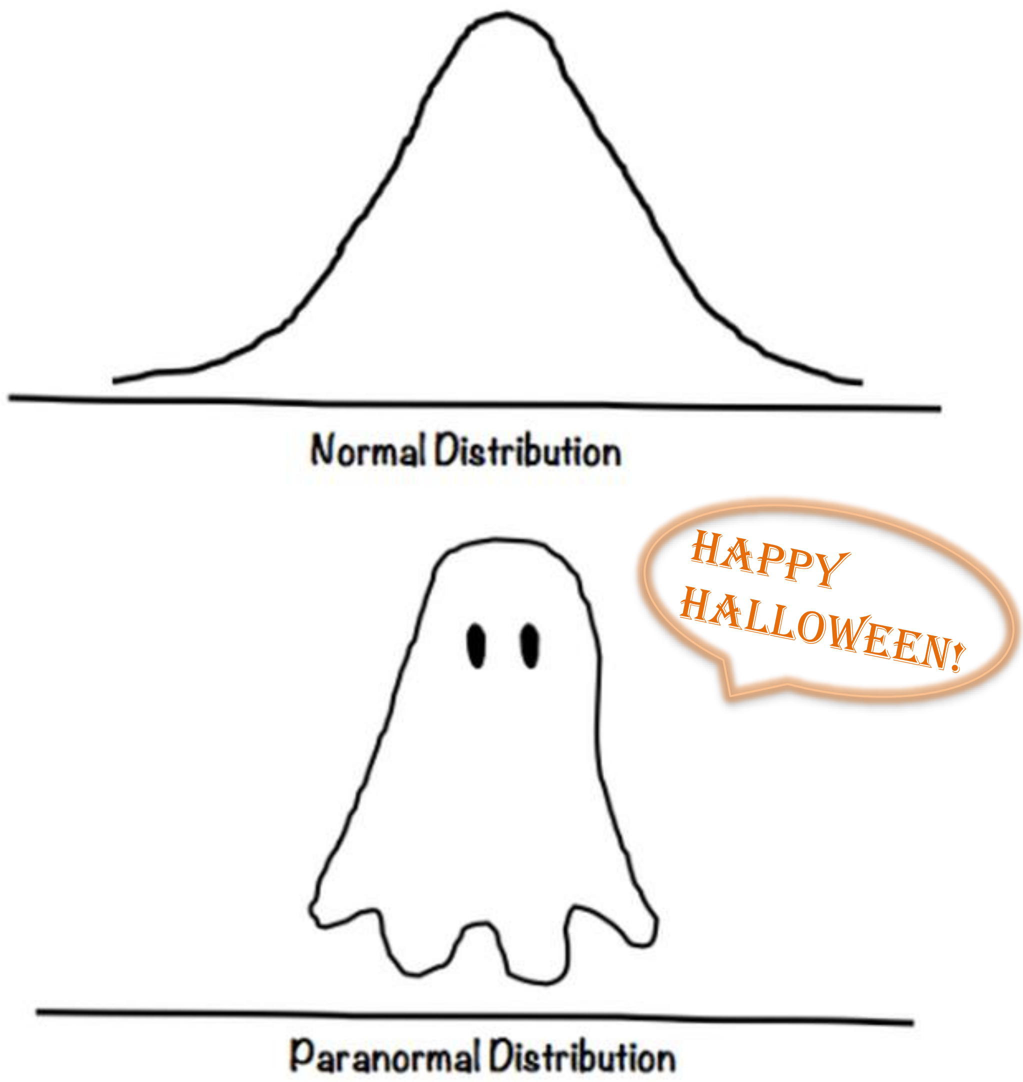
$\lambda_1 = 3/5 \qquad \lambda_2 = 0$

$$T_2 = XW_2 = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

$\vec{v}_1 \qquad \vec{v}_2$



Looking ahead: Statistics next week!



Pinterest