

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2026



HAVERFORD
COLLEGE

Admin

- **Lab 5** grades & feedback posted on Moodle

Lab 7 notes

- Logistic regression cost function:

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

if $h_{\vec{w}}(\vec{x}_i) = 0$ or $1 - h_{\vec{w}}(\vec{x}_i) = 0 \rightarrow$ skip $\log(0)$ or add 0

Outline for today

- Dimensionality reduction
- PCA for data visualization

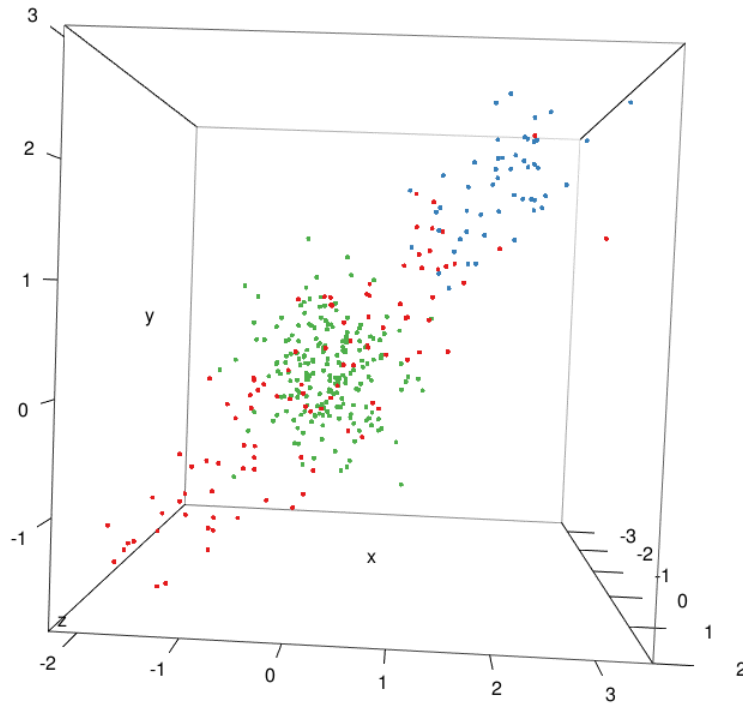
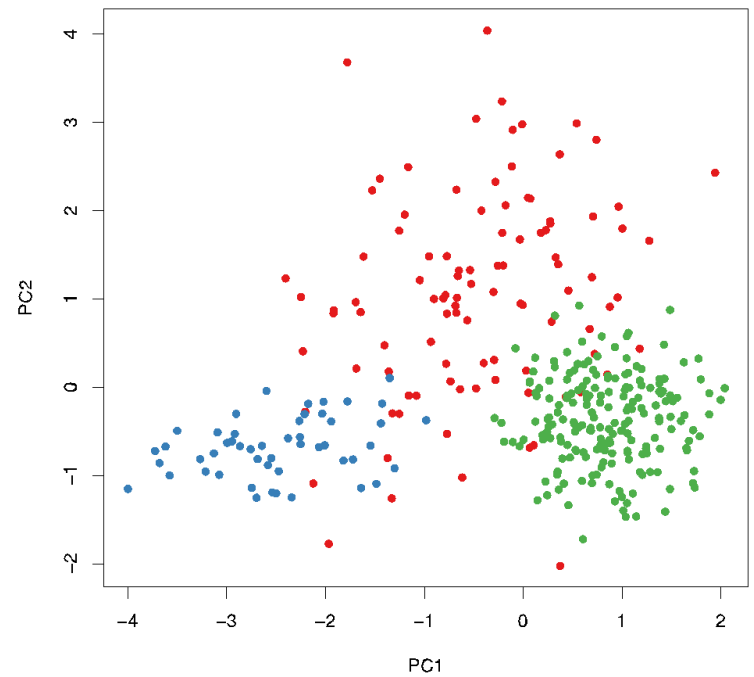
Outline for today

- Dimensionality reduction
- PCA for data visualization

Principal Component Analysis (PCA)

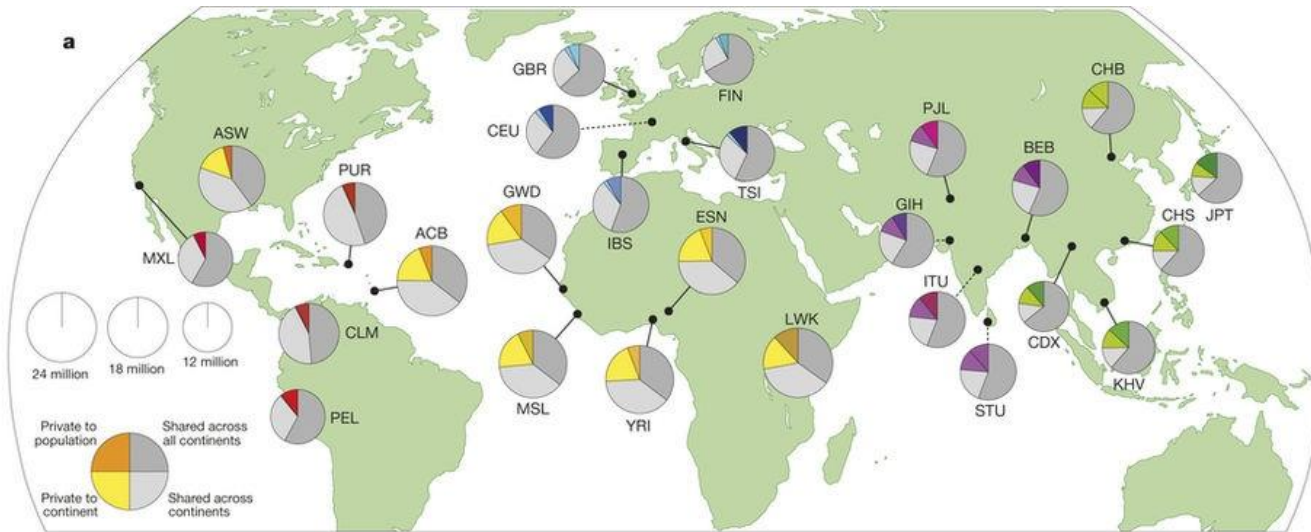
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on
- PCA is a linear transformation
- Typically, we look at the first few dimensions of the transformed data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

PCA Example

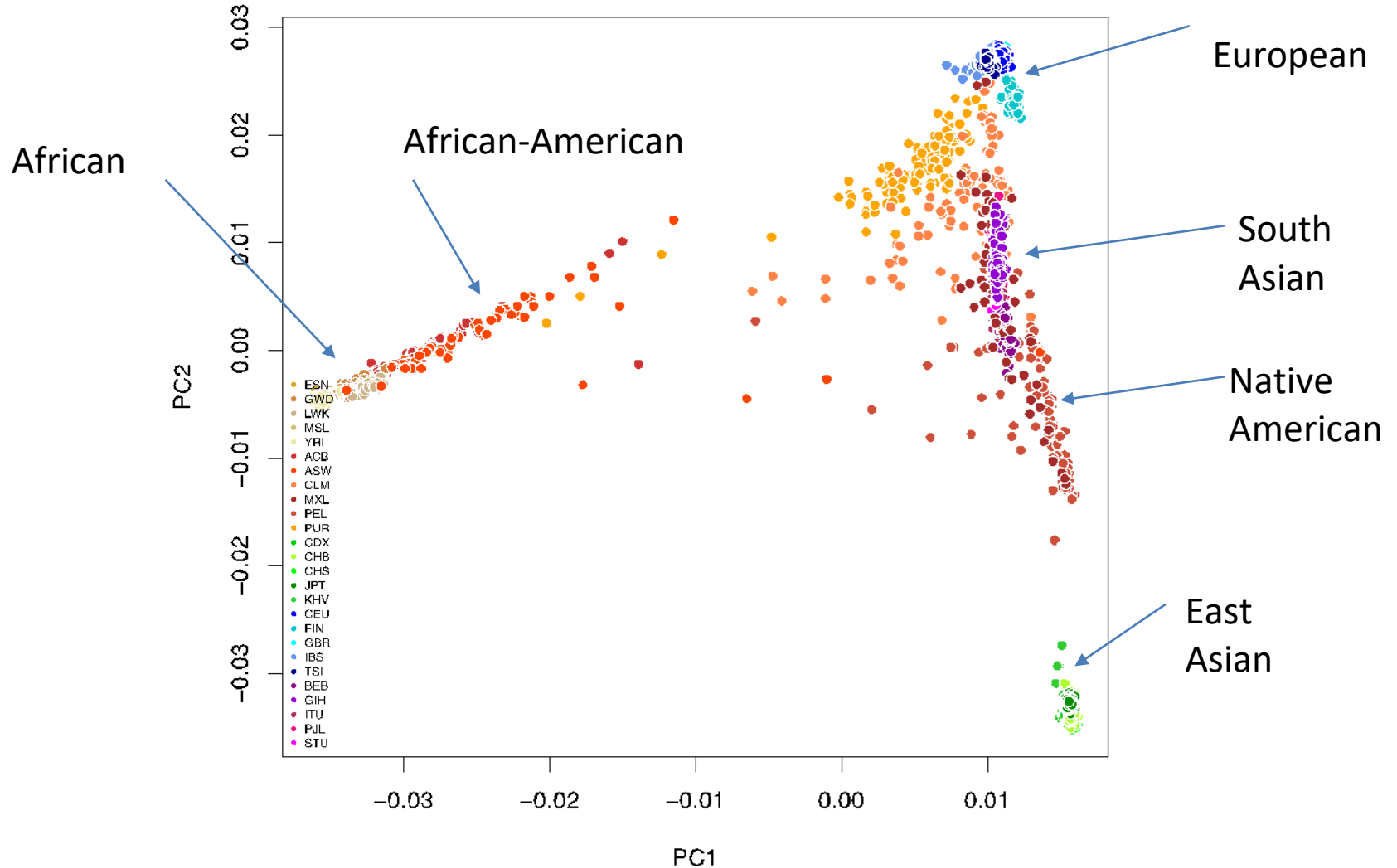


The 1000 Genomes project

- 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/>

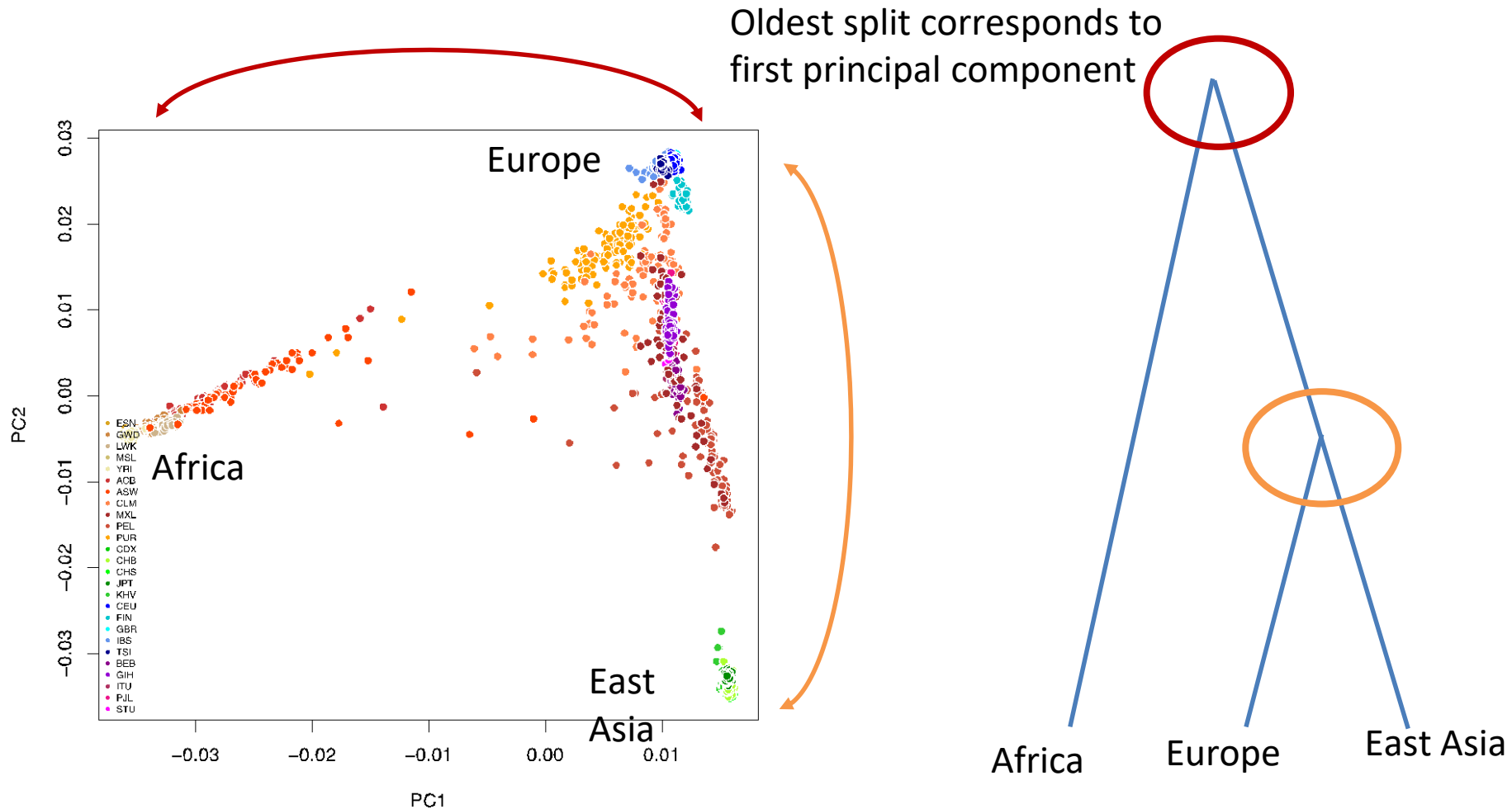


Global population structure



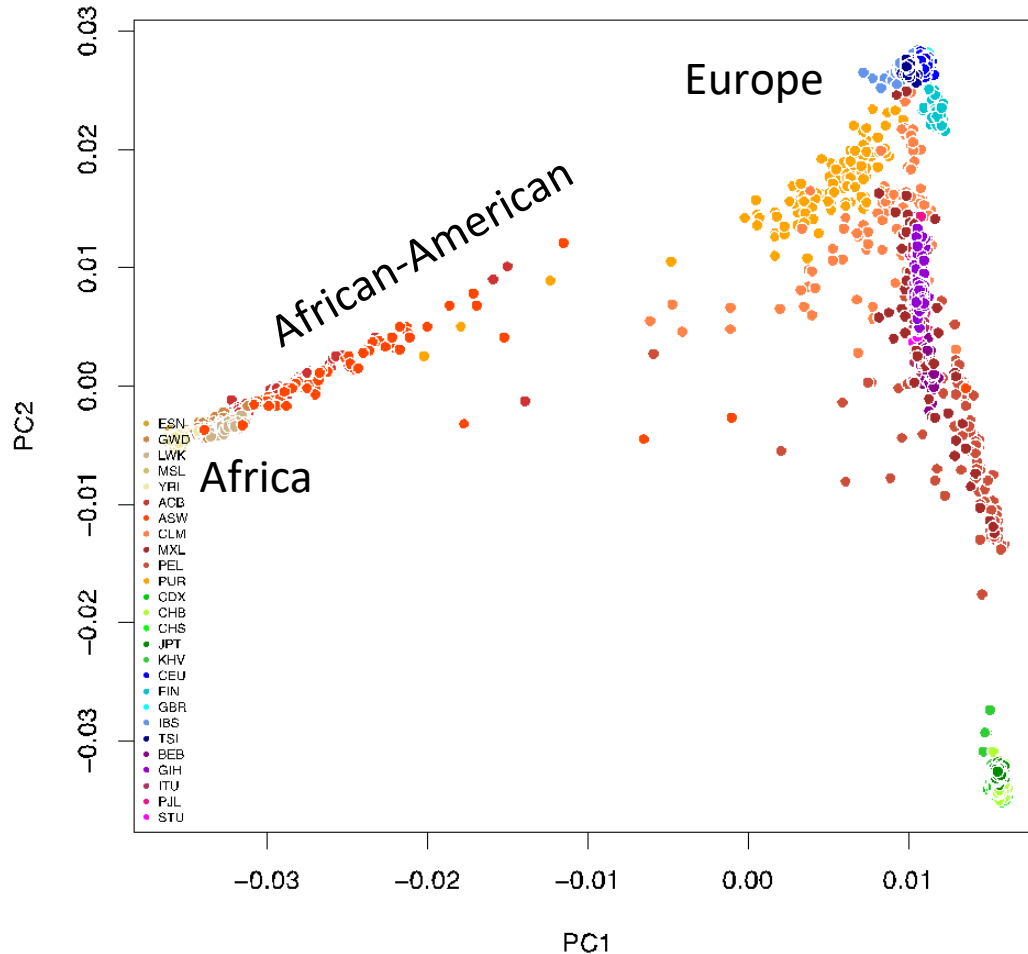
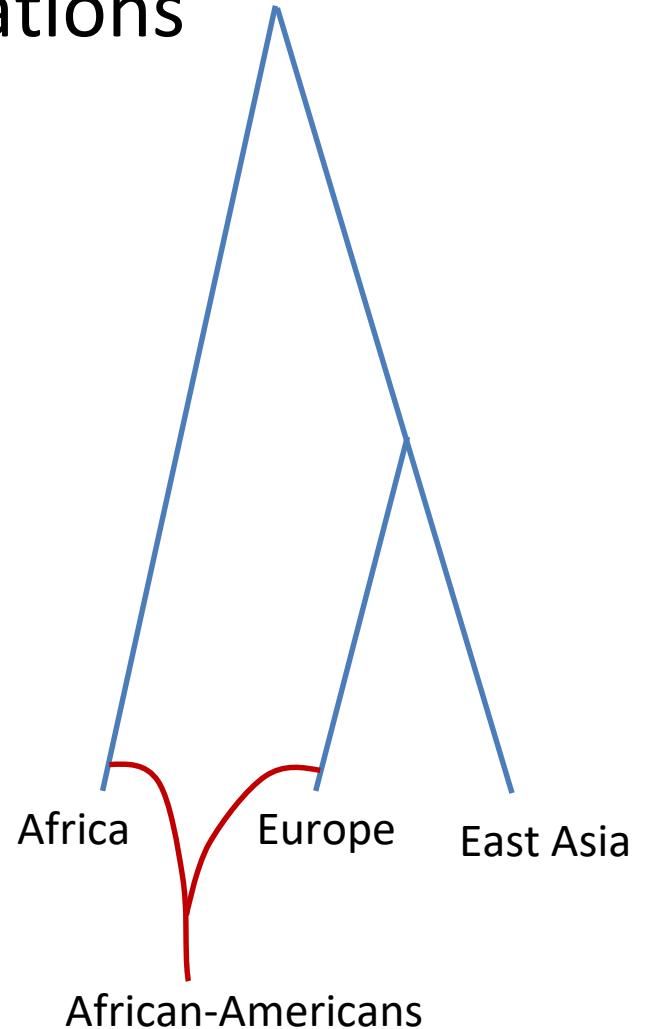
What causes these patterns?

1. Populations **splits** separate populations

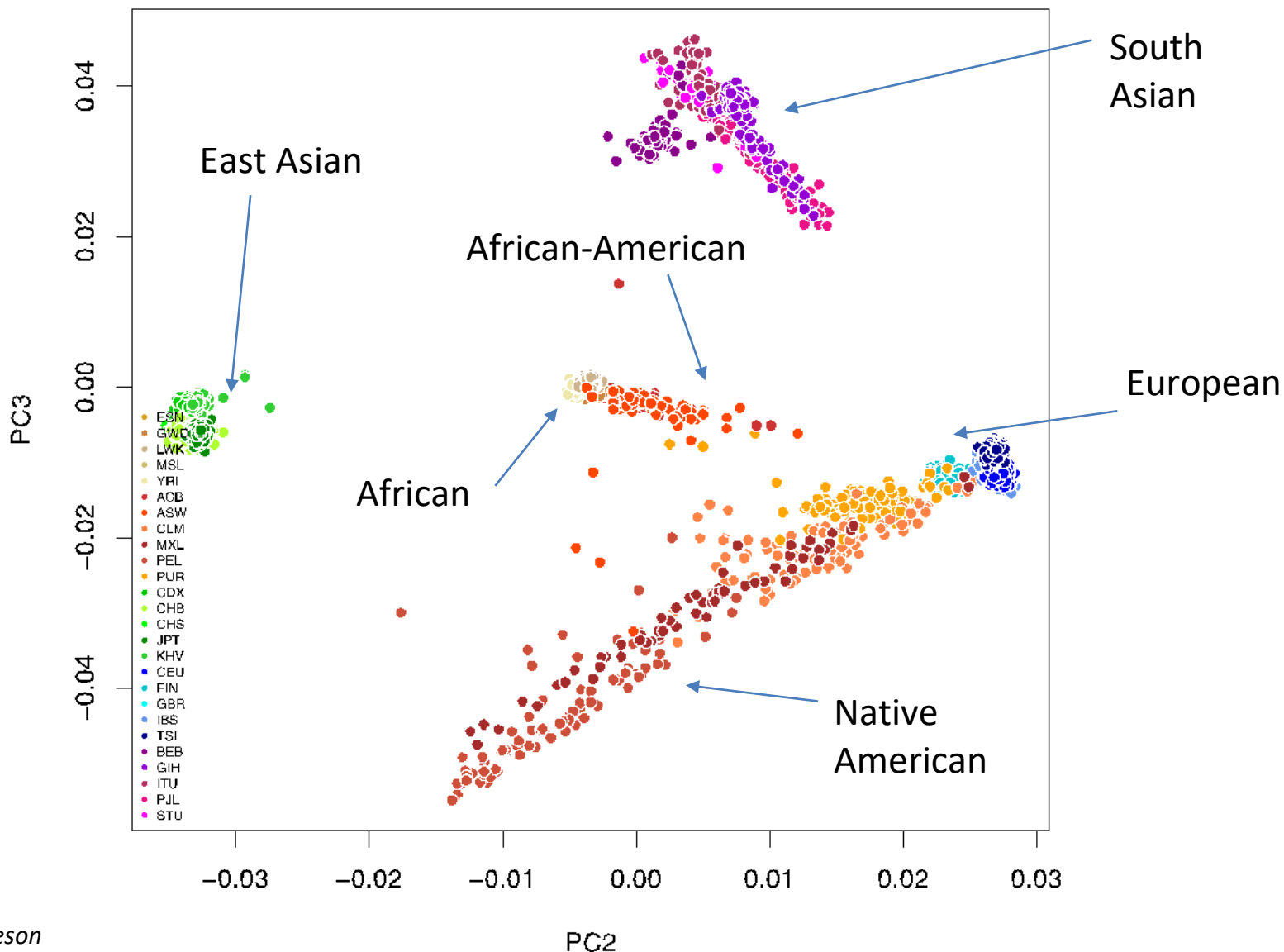


What causes these patterns?

2. **Admixture** merges populations



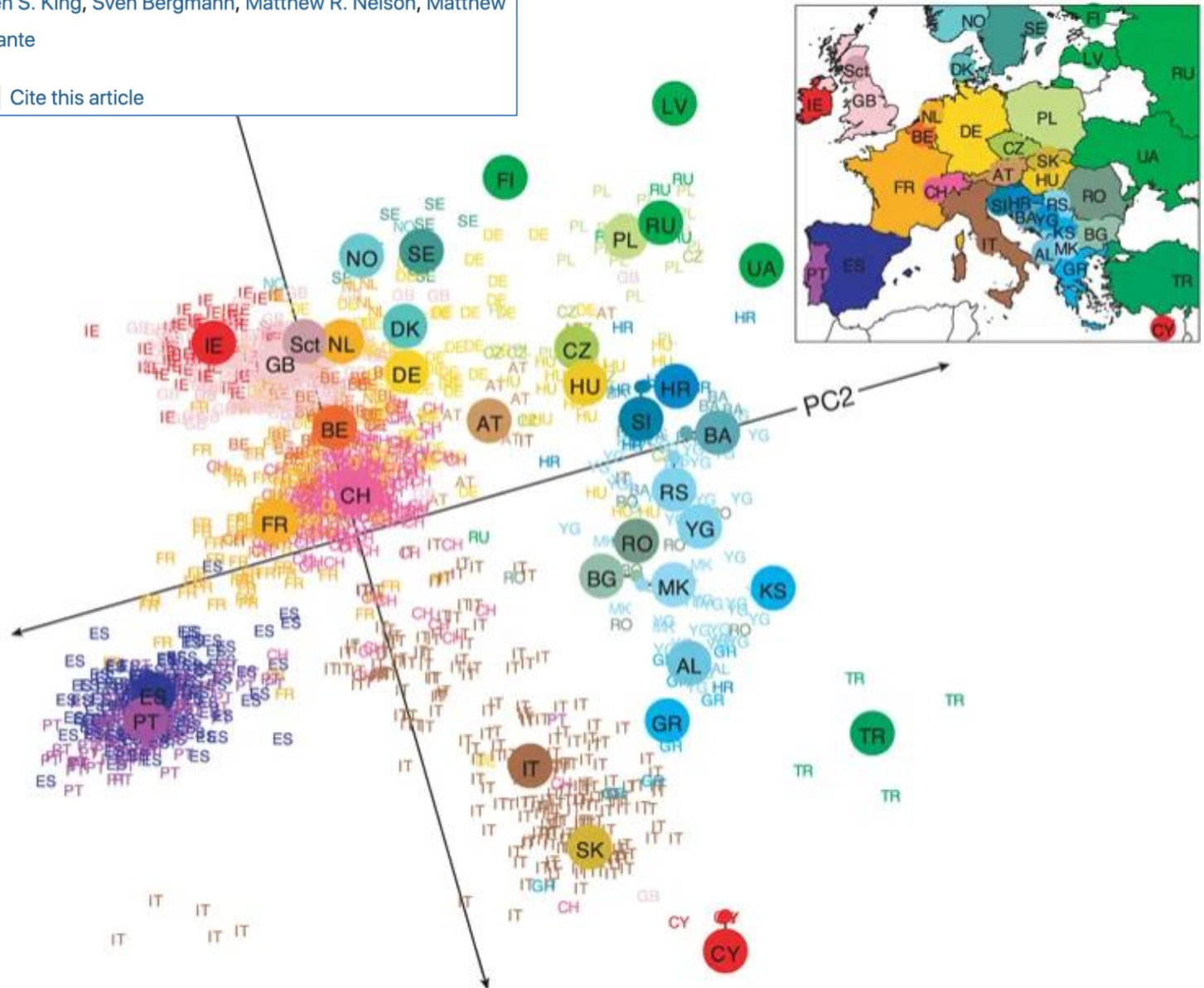
Global population structure



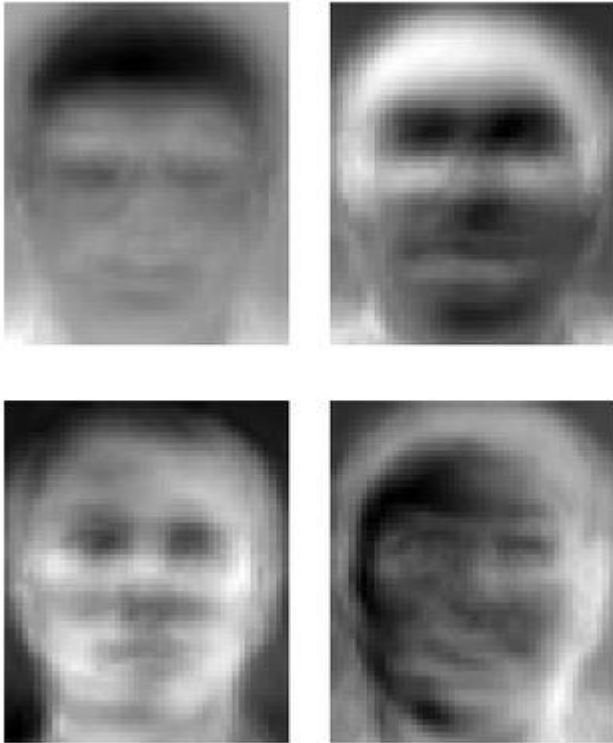
Genes mirror geography within Europe

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



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

Wikipedia

Outline for today

- Dimensionality reduction
- PCA for data visualization

PCA Algorithm

Step 1:

p 1:

$X_{orig} =$

$p \gg n$

p features

n

Goal: Create nx2 matrix for visualization

PCA Algorithm

Step 2: Subtract off column-wise mean

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



$$\bar{x}_1 = 2.5 \quad \bar{x}_2 = 2$$

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

PCA Algorithm

Step 3: Compute covariance matrix A

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

↓
square &
symmetric

Runtime $O(np^2)$

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

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

Handout 15, page 1

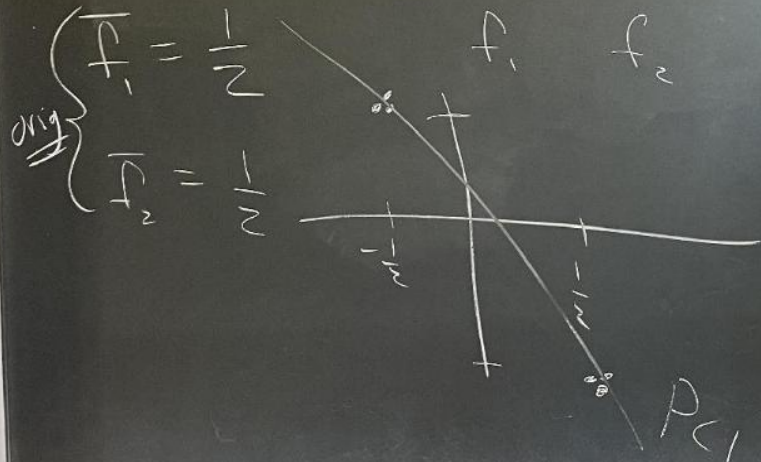
Handout 16

Step 1
4
2

X =

$$\begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix}$$

f_1 f_2



Step 3

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

$$\bar{f}_1 = 0$$

$$\bar{f}_2 = 0$$

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

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

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

PCA Algorithm

Step 4: Compute eigenvalues and eigenvectors of A

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

eigenvalue

eigenvector

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

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

PCA Algorithm

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

$$W = \begin{bmatrix} \overset{\lambda_1}{\vdots} & \overset{\lambda_2}{\vdots} & \dots & \vdots \\ \overrightarrow{v_1} & \overrightarrow{v_2} & \dots & \overrightarrow{v_r} \\ \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

 first eigenvector

$p \times r$
usually $r = 2$

And compute the transformed data:

$$T_{n \times r} = X_{n \times p} W_{p \times r}$$

Handout 15, page 2

Step 4

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

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

$$\det \begin{pmatrix} 3/10 - \lambda & -3/10 \\ -3/10 & 3/10 - \lambda \end{pmatrix} = 0$$

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

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

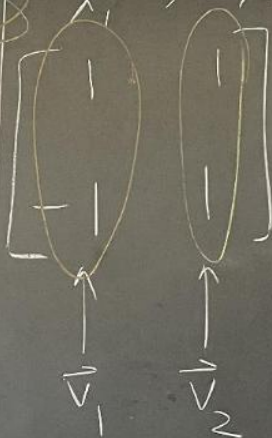
$$\lambda^2 - \frac{3}{5} \lambda = 0$$

$$\lambda \left(\lambda - \frac{3}{5} \right) = 0 \Rightarrow$$

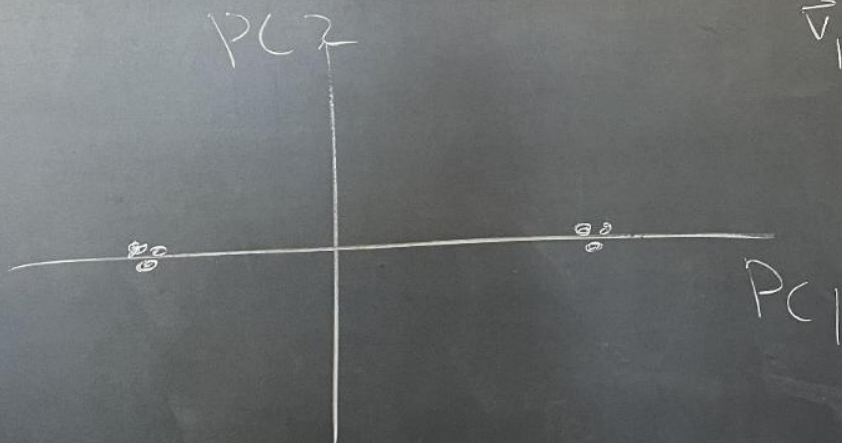
$$\lambda_1 = \frac{3}{5}$$
$$\lambda_2 = 0$$

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

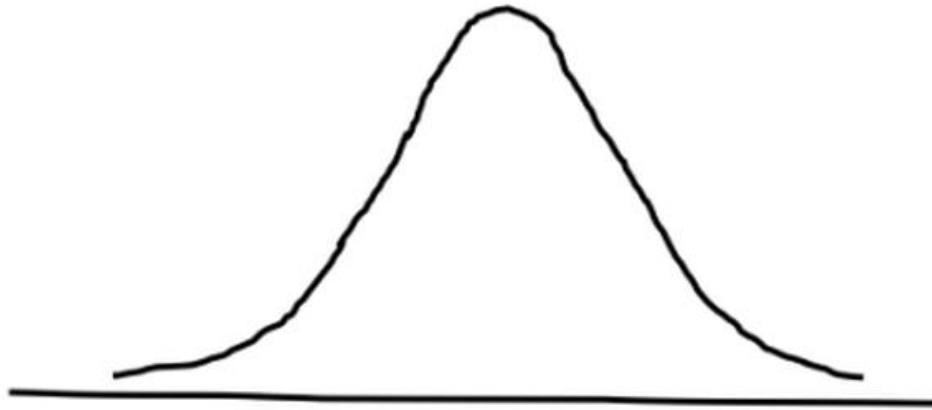
$$T_2 = XW_2 = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix} \quad \lambda_1 = -3/5 \quad \lambda_2 = 0$$



$$= \begin{bmatrix} -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$



Looking ahead: Statistics next week!



Normal Distribution



Paranormal Distribution