PCA for High-dimensional Data

Archer Yang

McGill University

October 4, 2024

- **Reading assignment:**
  - *Genes mirror geography within Europe*, Novembre et al (2008). Nature.
  - *Chapter 12,* Pattern Recognition and Machine Learning, Christopher Bishop.

High-dimensional PCA

PCA for high-dimensional data

- In some applications of PCA, the number of data points N is **smaller than** the dimensionality of the data space p.
- Need to deal with a $p \times p$ covariance matrix.
- Computation of eigenvalues and eigenvectors is **computationally expensive**. Complexity $O(p^3)$.

Setup

- The data matrix (**centered**) is

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix}$$

- The **sample covariance matrix** $\boldsymbol{S}$ is

$$\boldsymbol{S} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top = \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X}$$

Low-dimensional case

- In PCA, we solve the eigenvector equation

$$\boldsymbol{S}\boldsymbol{b}_m = \lambda_m \boldsymbol{b}_m, \qquad m = 1, \cdots, M$$

  where $\boldsymbol{b}_m$ is a basis vector of the principal subspace.

- **Computation complexity** $O(p^3)$.

- In low-dimensional case $N \geq p$

$$\mathrm{rank}(\boldsymbol{S}) = \text{\# of nonzero eigenvalues of } \boldsymbol{S} = \mathrm{rank}(\boldsymbol{X}) \leq p$$

High-dimensional case

- (Homework) Show that when $N < p$, a general matrix $\boldsymbol{X}$ has $\text{rank}(\boldsymbol{X}) \leq N$.
- (Homework) Show that for a **centered matrix** $\boldsymbol{X}$

$$\text{rank}(\boldsymbol{X}) \leq N - 1$$

  centered matrix: each row of $\boldsymbol{X}$ is $\boldsymbol{x}_n^\top$, which is centered by applying $\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n - \bar{\boldsymbol{x}}$ with $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$.
- What kind of $\boldsymbol{X}$ has strictly $\text{rank}(\boldsymbol{X}) < N - 1$?

High-dimensional case

- In high-dimensional case, since $N < p$ or $N \ll p$, for centered $\boldsymbol{X}$

$$\mathrm{rank}(\boldsymbol{S}) = \text{\# of nonzero eigenvalues of } \boldsymbol{S}$$

$$= \mathrm{rank}(\boldsymbol{X}) \leq N - 1$$

- Thus $\boldsymbol{S}$ as a $p \times p$ matrix has at least $p - N + 1$ eigenvalues that are zero.

- Rewrite the equation as

$$\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b}_m = \lambda_m \boldsymbol{b}_m$$

- Multiply both sides by $\boldsymbol{X}$ to give

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b}_m = \lambda_m(\boldsymbol{X}\boldsymbol{b}_m)$$

- If we now define $\boldsymbol{c}_m = \boldsymbol{X}\boldsymbol{b}_m$, we obtain

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{c}_m = \lambda_m \boldsymbol{c}_m$$

- What is the dimension of the matrix $\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top$ in the eigenvector equation?

Procedure

(Homework) If $N < p$, show that $\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top$ has the same $N-1$ eigenvalues as the original covariance matrix $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$, and also $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ has an additional $p - N + 1$ eigenvalues of value zero. In other words

- For $N \times N$ matrix $\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top$, its N eigenvalues are

$$\underbrace{\lambda'_1, \lambda'_2, \cdots, \lambda'_{N-1}}_{N-1}, 0$$

- For $p \times p$ matrix $\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$, its p eigenvalues are

$$\underbrace{\lambda_1, \lambda_2, \cdots, \lambda_{N-1}}_{N-1}, \underbrace{0, 0, \cdots, 0}_{p-N+1}$$

- We have

$$\boxed{\lambda'_1, \lambda'_2, \cdots, \lambda'_{N-1} = \lambda_1, \lambda_2, \cdots, \lambda_{N-1}}$$

Procedure

- Therefore instead we can solve

$$\frac{1}{N}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{c}_m = \lambda_m \boldsymbol{c}_m \qquad \text{for } m = 1, \cdots, N-1$$

  with computational cost $O(N^3)$ instead of $O(p^3)$.

- In order to determine the eigenvectors, multiply both sides by $\boldsymbol{X}^\top$

$$\left(\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}\right)\left(\boldsymbol{X}^\top \boldsymbol{c}_m\right) = \lambda_m \left(\boldsymbol{X}^\top \boldsymbol{c}_m\right)$$

- Therefore $\boldsymbol{X}^\top \boldsymbol{c}_m$ is an eigenvector of $\boldsymbol{S}$ with eigenvalue λ_m.

- Note: eigenvector $\boldsymbol{X}^\top \boldsymbol{c}_m$ for $m = 1, \cdots, N-1$ are not normalized, we need to re-scale

$$\boldsymbol{b}_m \propto \boldsymbol{X}^\top \boldsymbol{c}_m$$

by a constant such that $||\boldsymbol{b}_m|| = 1$.

- (Homework) Assume that $\boldsymbol{c}_m$ has been normalized to unit length $||\boldsymbol{c}_m|| = 1$, show that

$$\boldsymbol{b}_m = \frac{1}{\sqrt{N\lambda_m}} \boldsymbol{X}^\top \boldsymbol{c}_m$$

has unit length. This provide us a way to obtain a unit-lengthed $\boldsymbol{b}_m$.

Summary

- To apply PCA for high-dimensional data
  - First evaluate $\boldsymbol{X}\boldsymbol{X}^\top$.
  - Then find its eigen-vectors and eigenvalues.
  - Compute the eigenvectors in the original data space.

Application: population genetics

- **Data matrix:**
  - **Columns**: Genetic variation were genotyped at 500,568 loci using the SNP chip.
  - **Rows**: 3,192 European individuals were genotyped.
- Use the country of origin of each individual's grandparents to determine the geographic location.
- Exclude individuals with grand-parental ancestry from more than location

- Use PCA to produce a two-dimensional visual summary of the observed genetic variation.

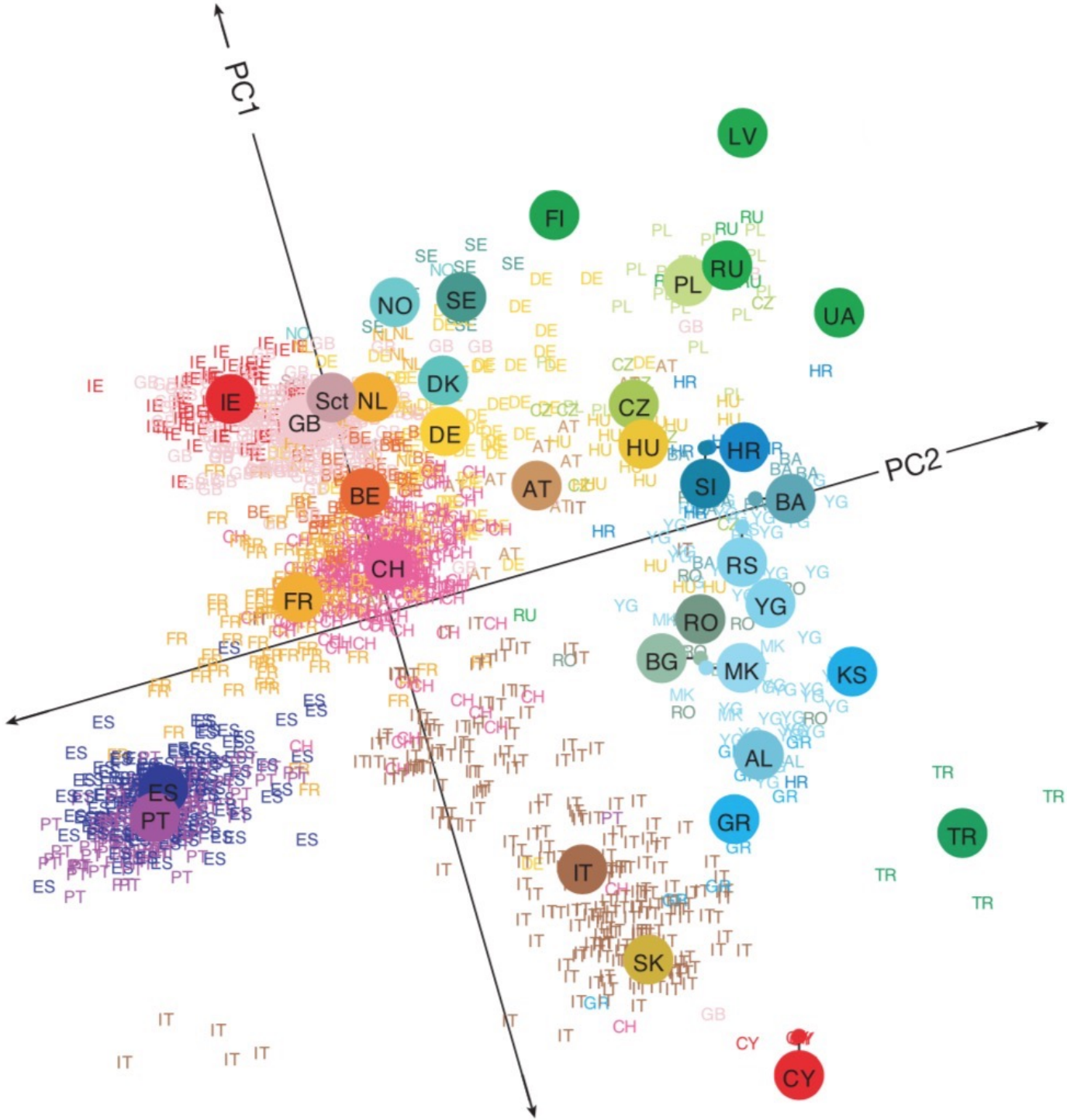
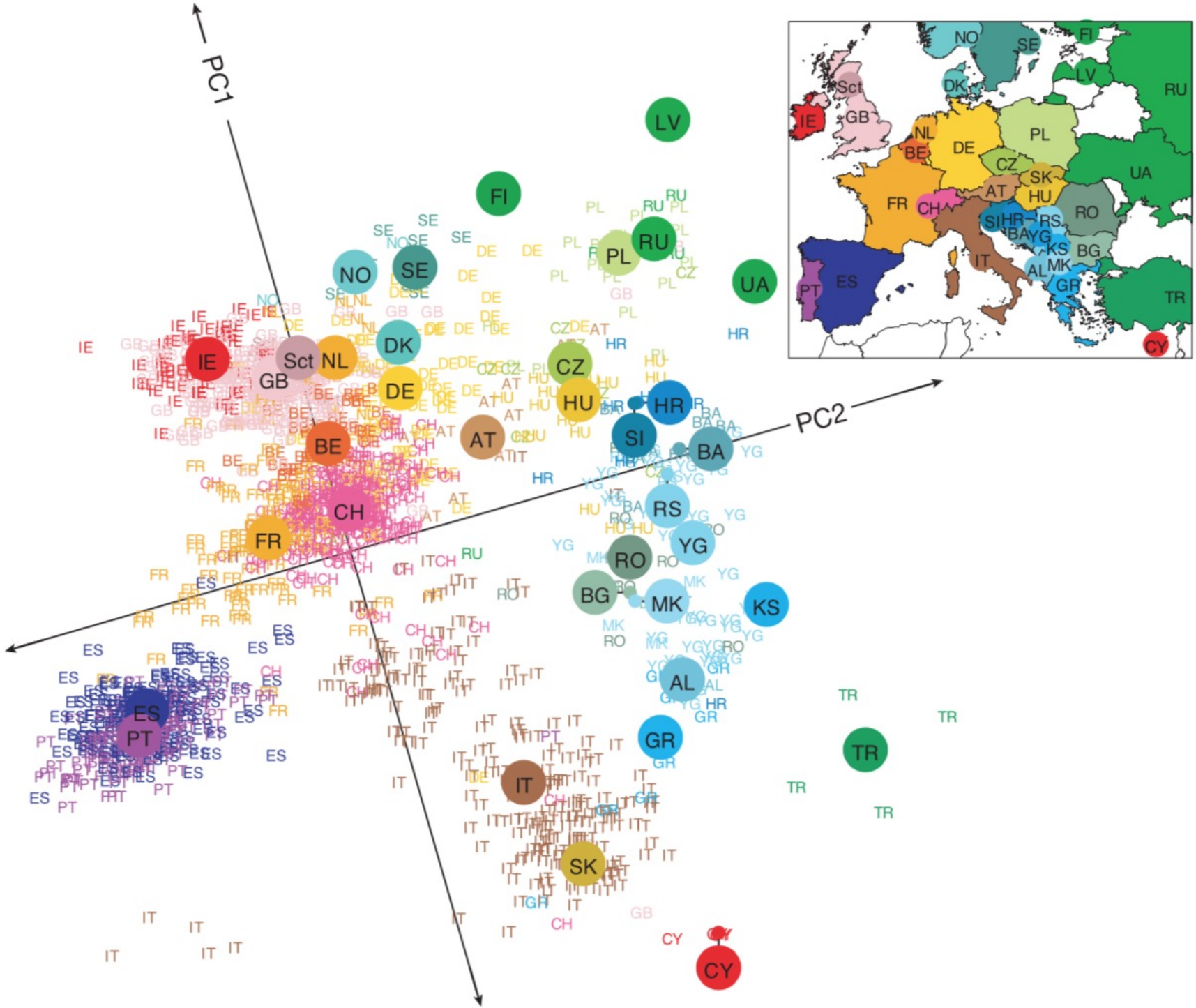
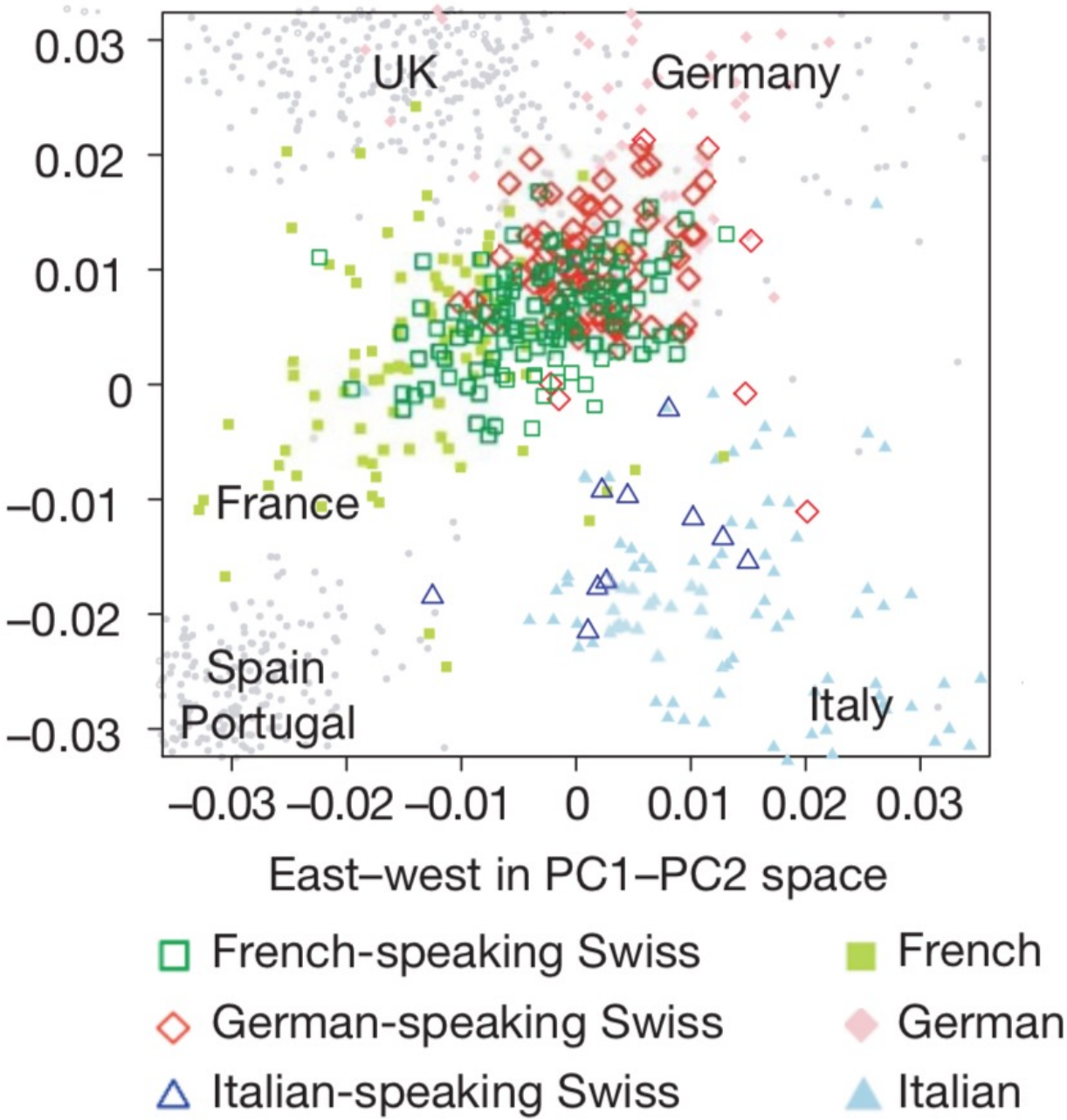

PC1
PC2
LV
FI
RU
PL
UA
NO
SE
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
GR
TR
IT
SK
CY
ES
PT

Figure 1 | Population structure within Europe. **a**, A statistical summary of genetic data from 1,387 Europeans based on principal component axis one (PC1) and axis two (PC2). Small coloured labels represent individuals and large coloured points represent median PC1 and PC2 values for each country. The inset map provides a key to the labels. The PC axes are rotated to emphasize the similarity to the geographic map of Europe. AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia. **b**, A magnification of the area around Switzerland from **a** showing differentiation within Switzerland by language. **c**, Genetic similarity versus geographic distance. Median genetic correlation between pairs of individuals as a function of geographic distance between their respective populations.


PC1
PC2
LV
FI
RU
PL
NO
SE
UA
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
ES
PT
GR
TR
IT
SK
CY


0.03
0.02
0.01
0
–0.01
–0.02
–0.03
UK
Germany
France
Spain
Portugal
Italy
–0.03 –0.02 –0.01 0 0.01 0.02 0.03
East–west in PC1–PC2 space
French-speaking Swiss
German-speaking Swiss
Italian-speaking Swiss
French
German
Italian

Results

- The resulting figure bears a notable resemblance to a geographic map of Europe:
  - Individuals from the same geographic region cluster together
  - Major populations are distinguishable.
  - A close correspondence between genetic and geographic distances.

Results

- For spatially structured data, theory predicts the top two principal components (PCs 1 and 2) to be correlated with perpendicular geographic axes.
- The direction of the PC1 axis and its relative strength may reflect a special role for this geographic axis in the demographic history of Europeans

Sparse PCA