CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2020

Admin

- In lab on Thursday: **project meetings** with all groups (optional if you're taking the exam)
  - Fill out questionnaire for which presentation day
  - So far: 4 groups presenting this Friday, 5 next Friday
  - Reminder that video must be turned on during your presentation
- **Second midterm** posted later today
  - Take in a 2 hour block before Dec 18 at noon
  - Create a study sheet (no other resources permitted, except for a calculator)
- Office hours **TODAY 4:30-6pm**

Outline for December 8

- Dimensionality reduction
- Principal Component Analysis (PCA)
- Midterm II review and practice problems

Outline for December 8

- Dimensionality reduction
- Principal Component Analysis (PCA)
- Midterm II review and practice problems

Principal Components 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

- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction

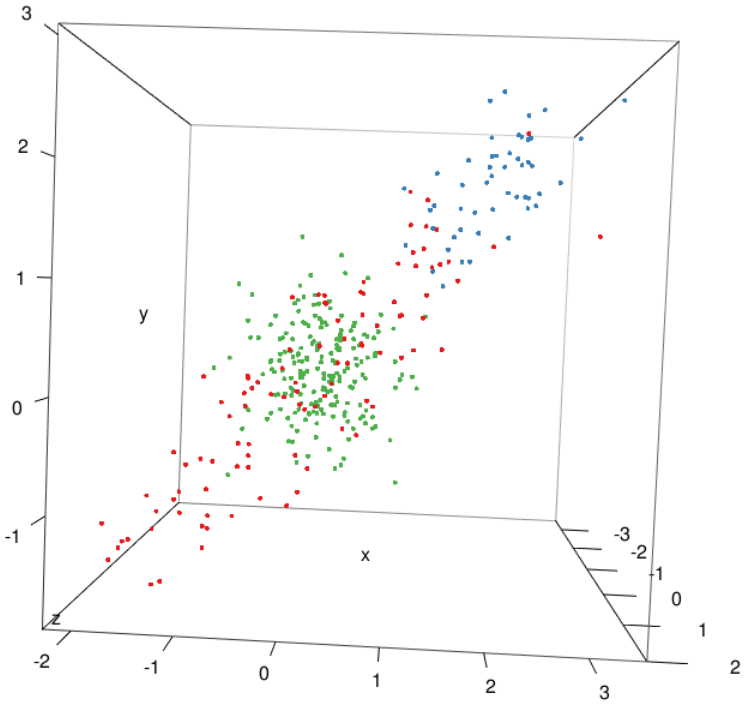
- PCA is a linear transformation

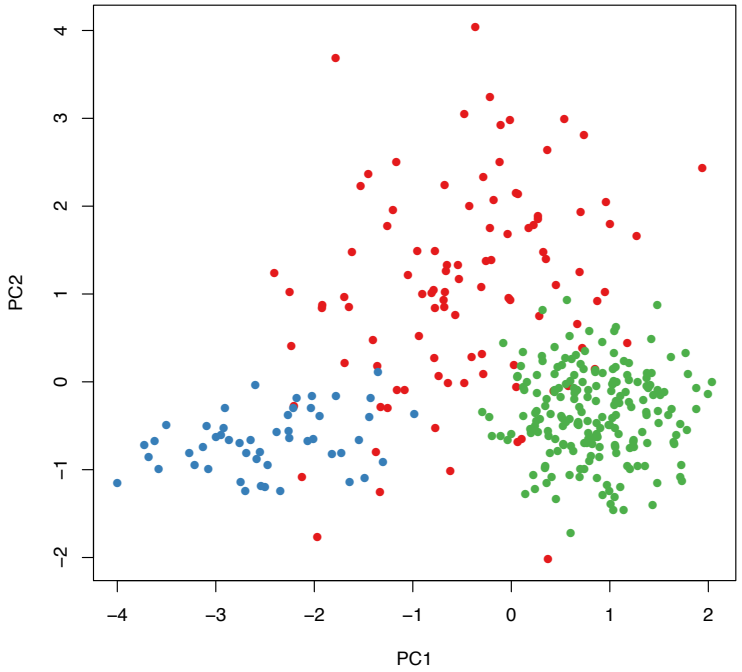
- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

Outline for December 8

- Dimensionality reduction
- Principal Component Analysis (PCA)
- Midterm II review and practice problems

Principal component analysis

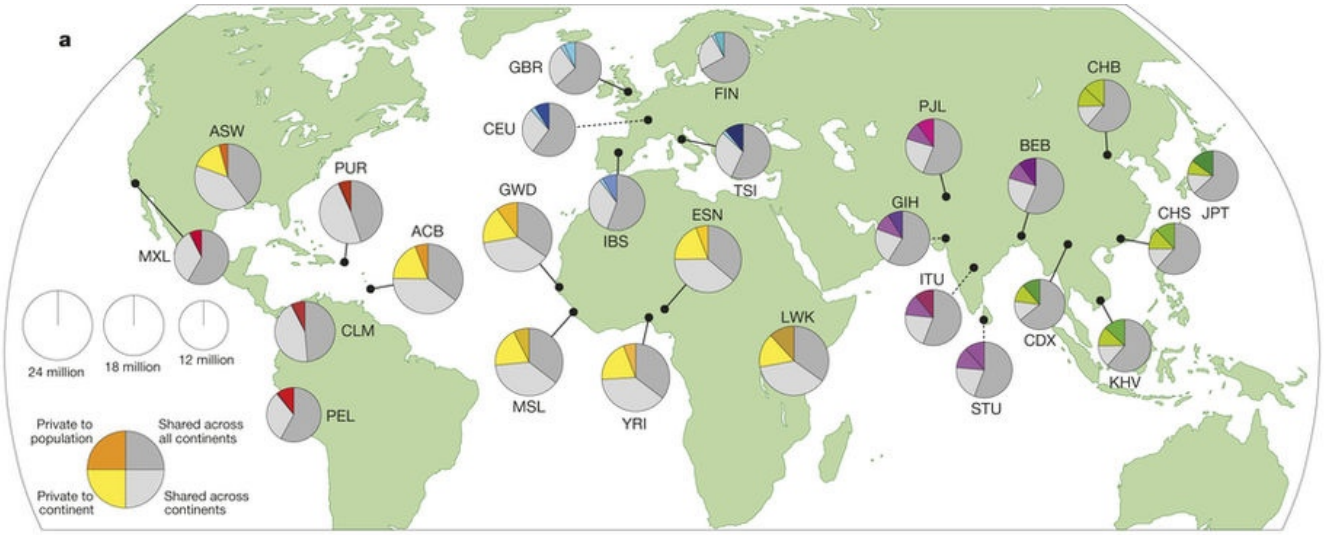




Slide: Iain Mathieson

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/

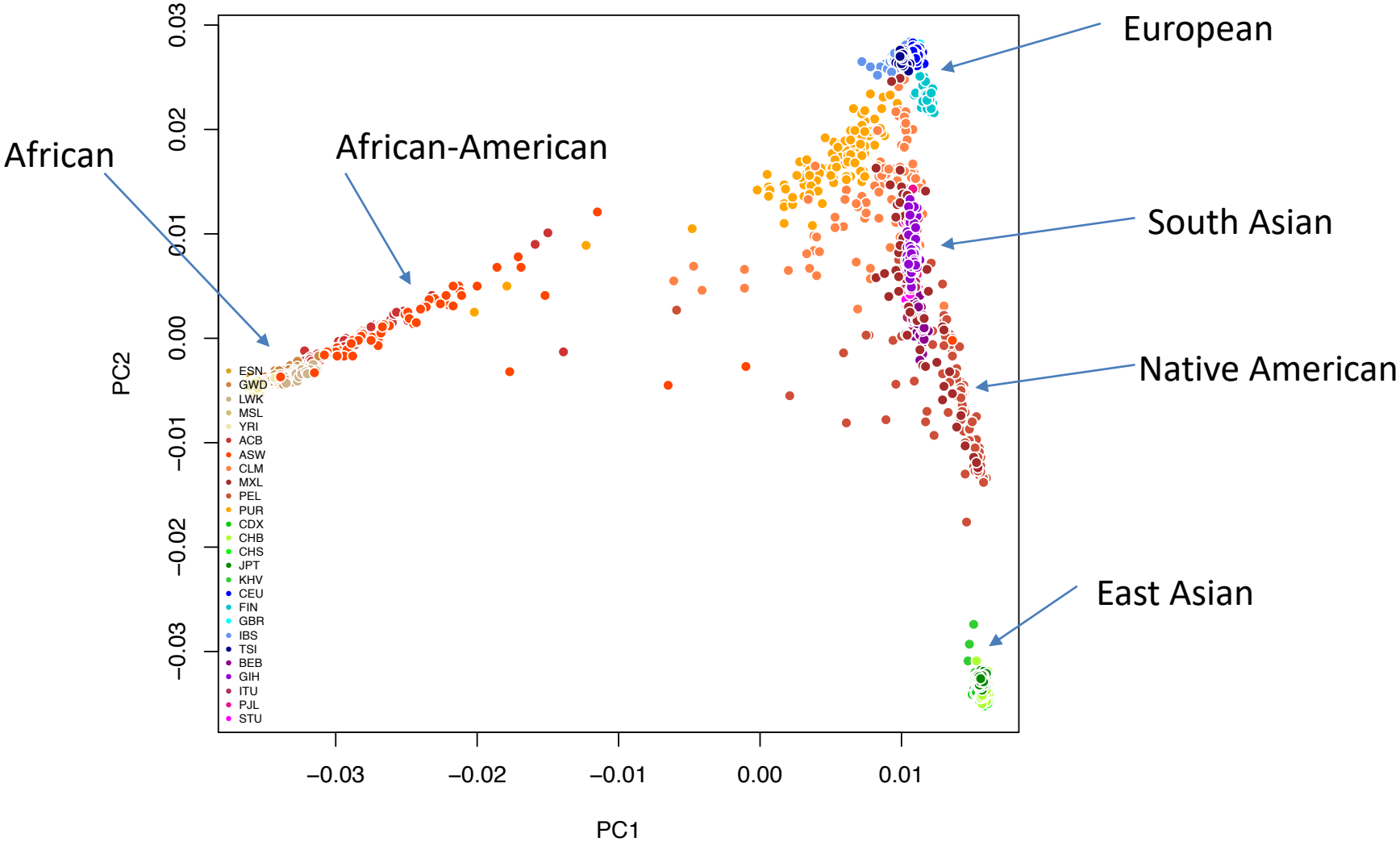


The 1000 Genomes Project Consortium, 2015; https://www.nature.com/articles/nature15393

Slide: Iain Mathieson

```
##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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     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     0|0
```

Global population structure

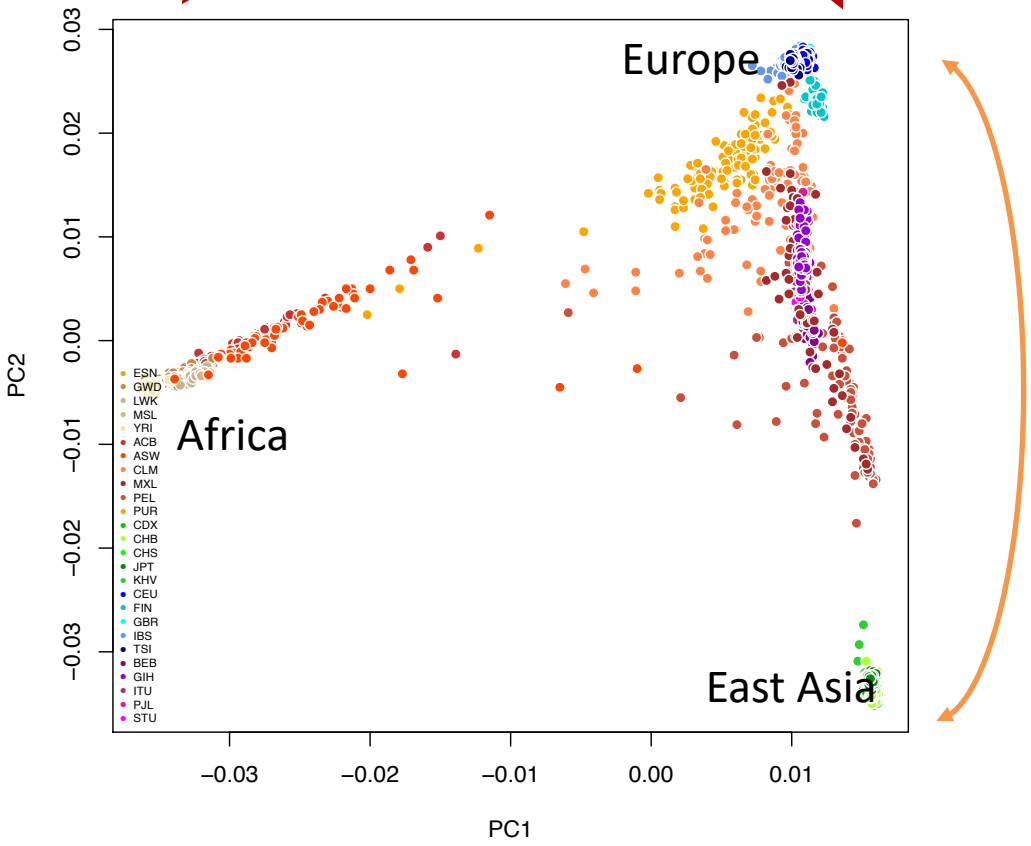


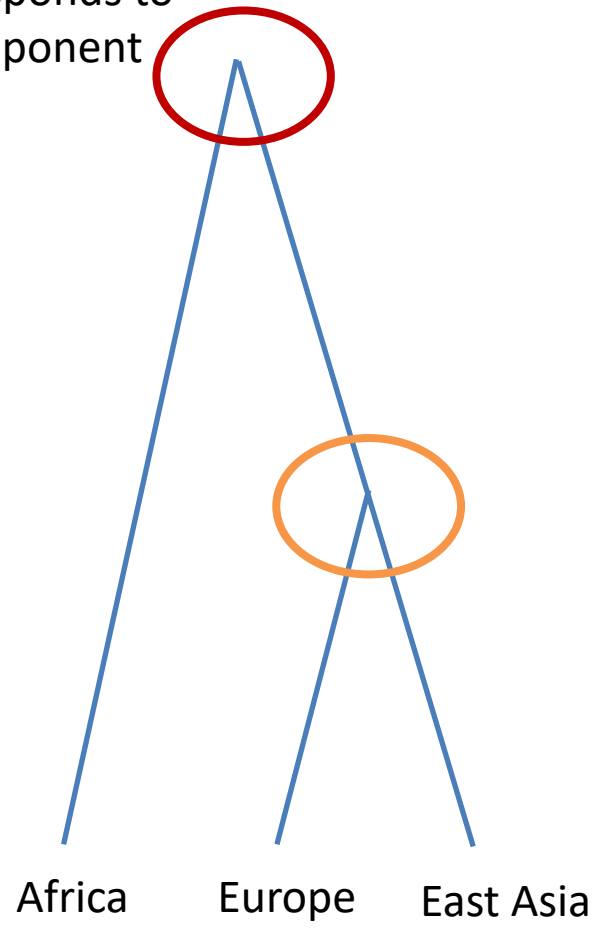
Slide: Iain Mathieson

What causes these patterns?

1. Populations splits separate populations

Oldest split corresponds to first principal component



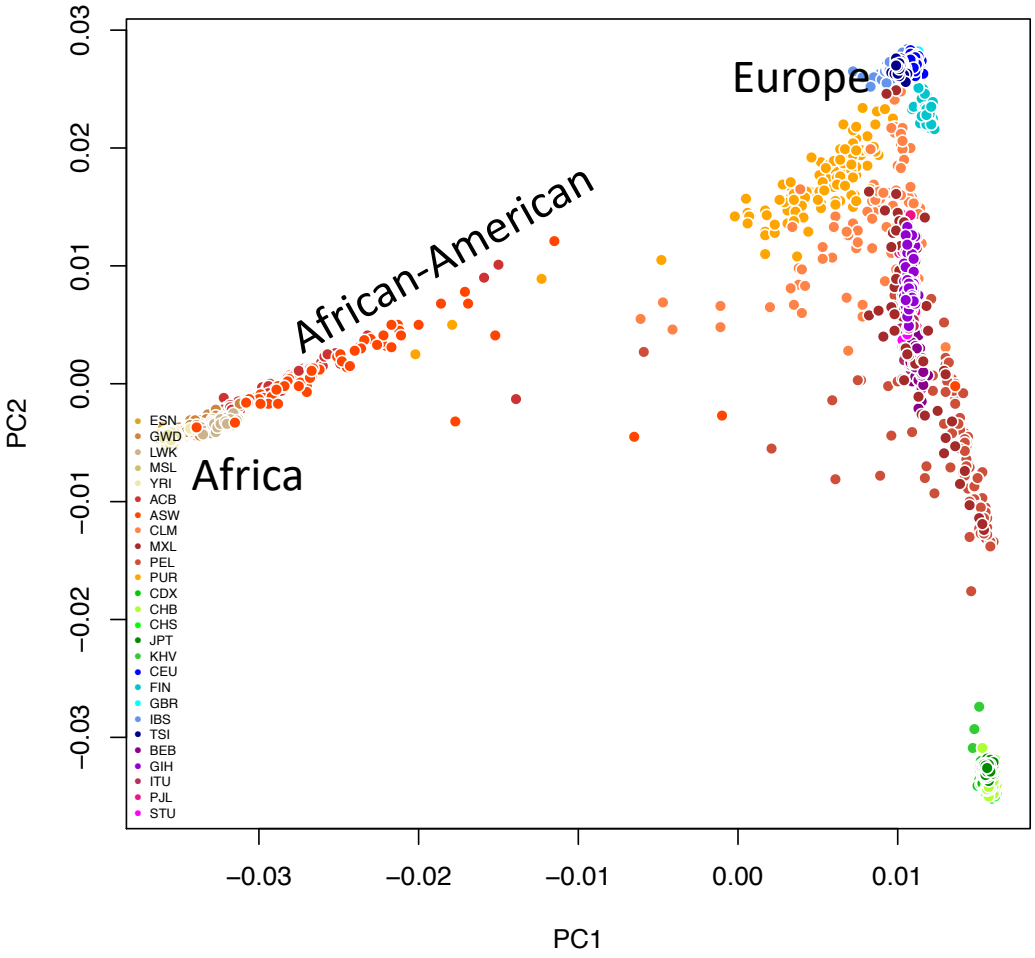


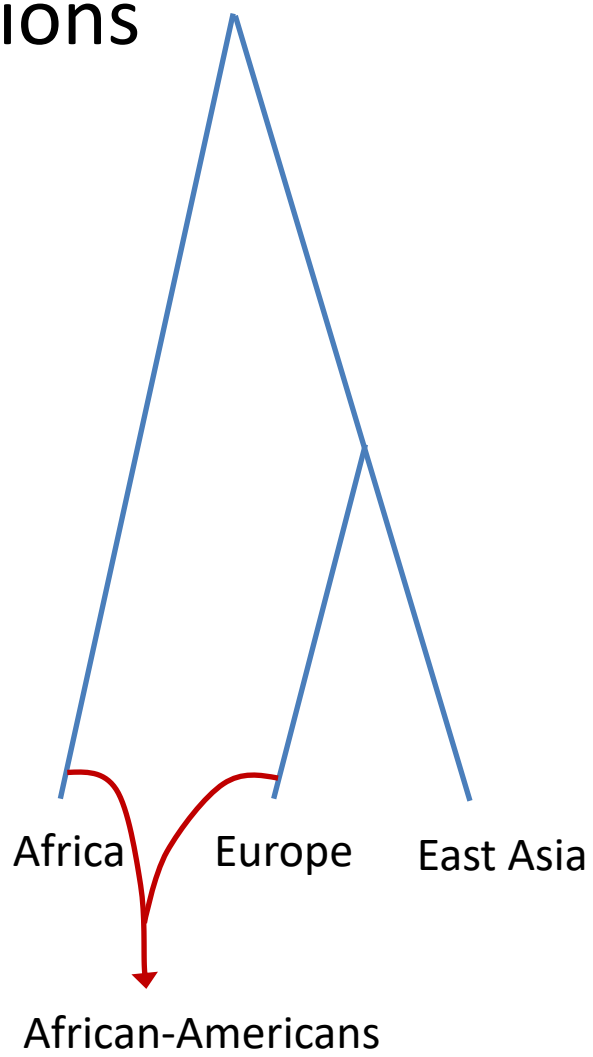
Slide: Iain Mathieson

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

What causes these patterns?

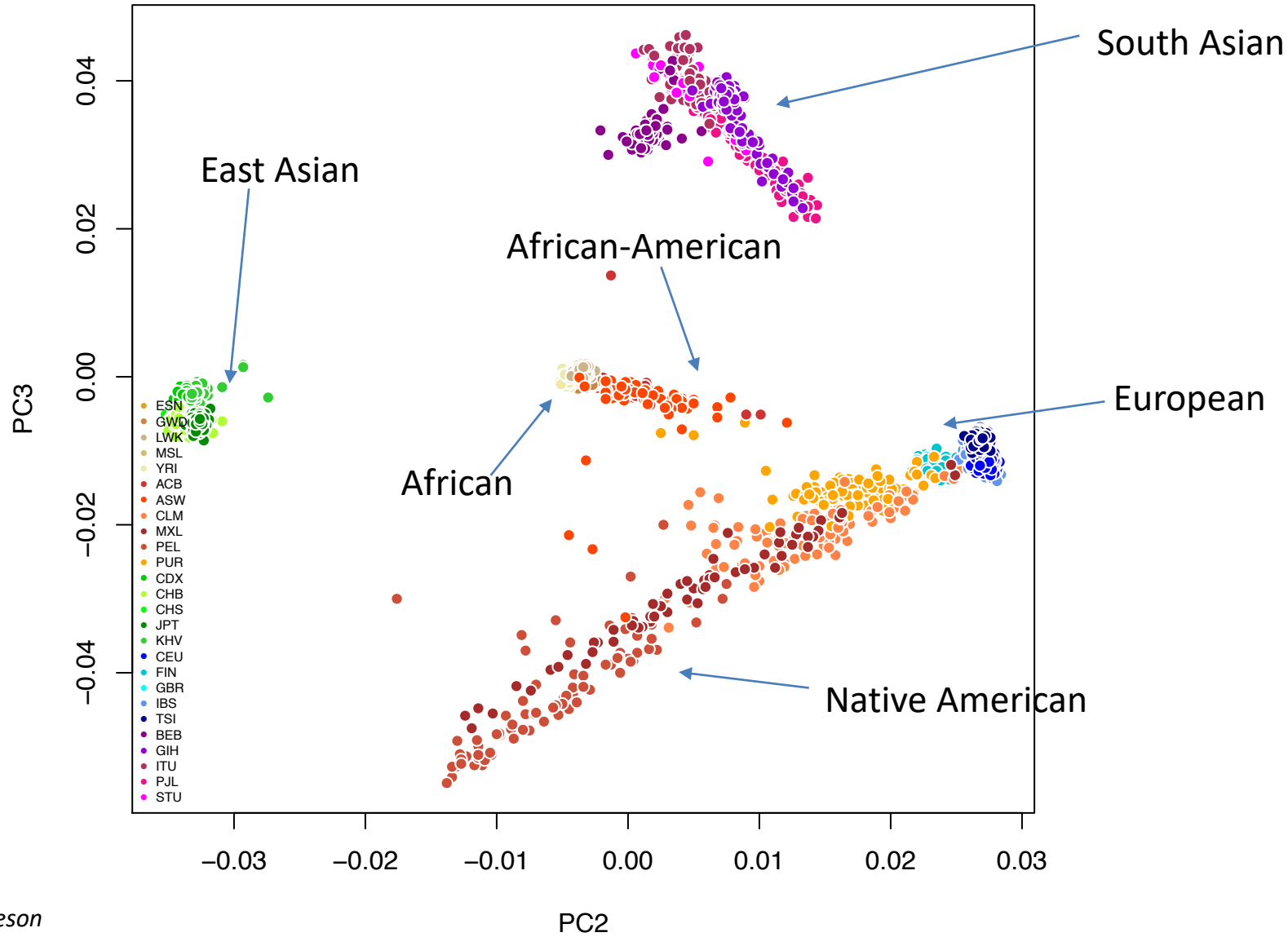
2. Admixture merges populations





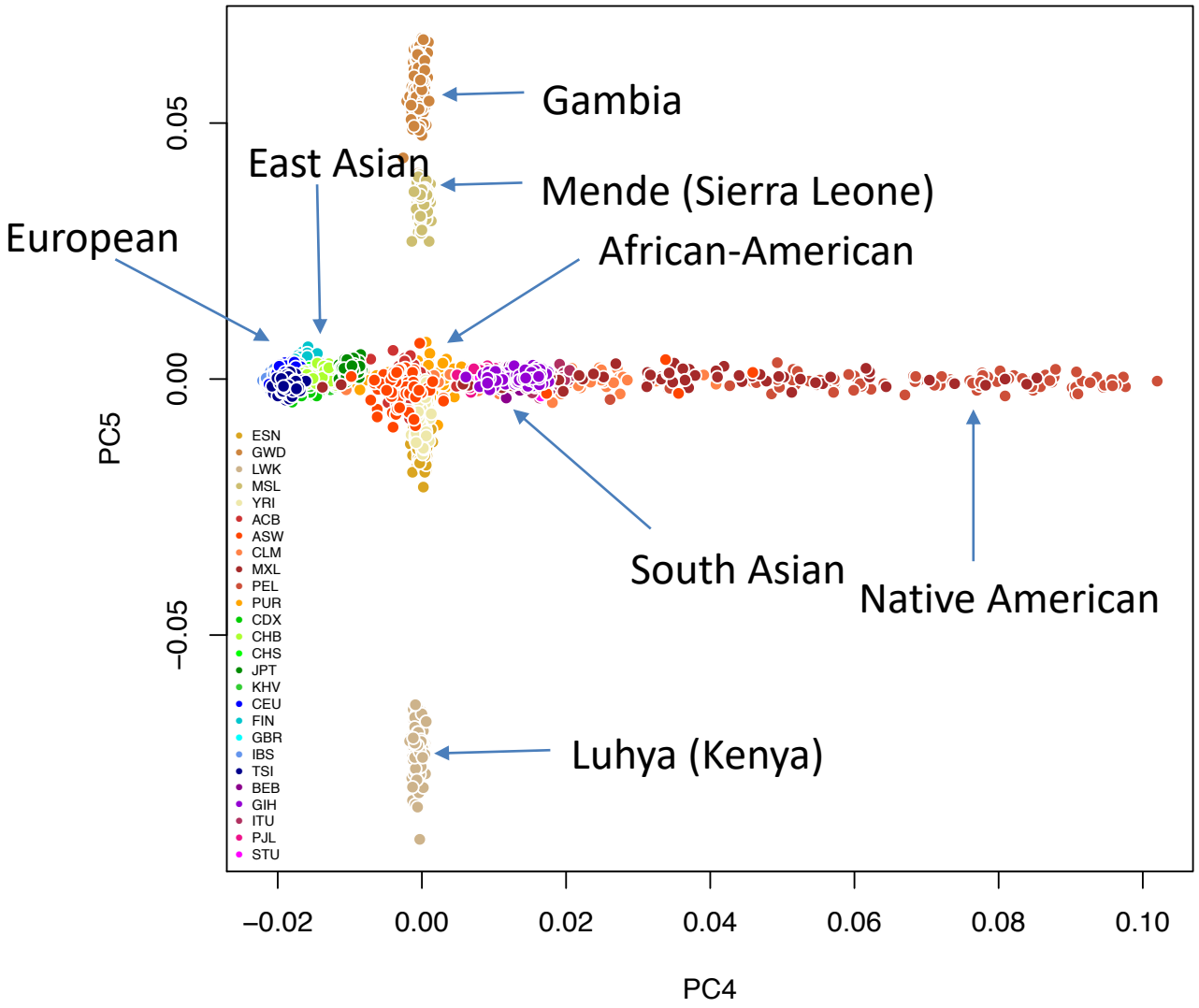
Slide: Iain Mathieson

Global population structure



Slide: Iain Mathieson

Global population structure

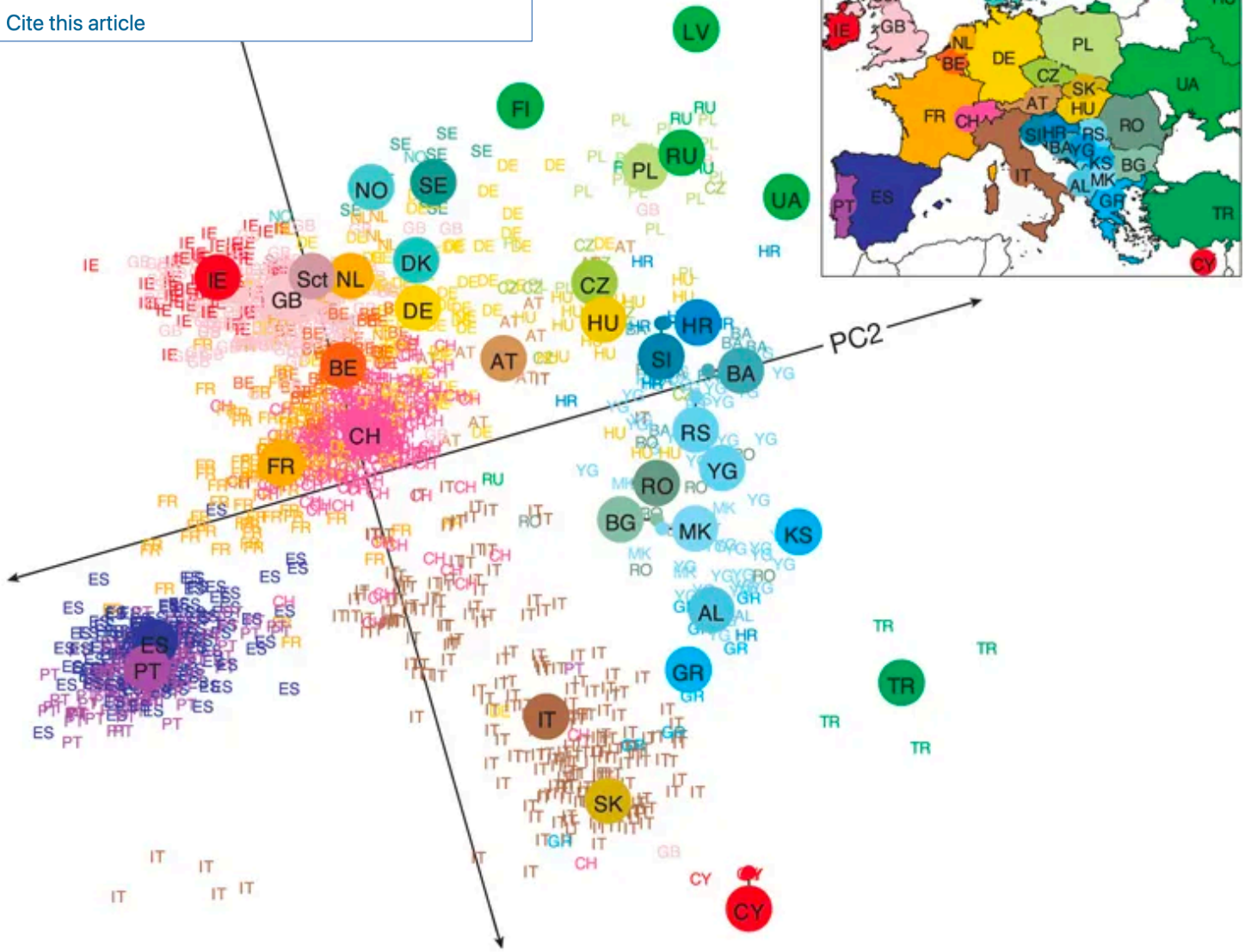


Slide: Iain Mathieson

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

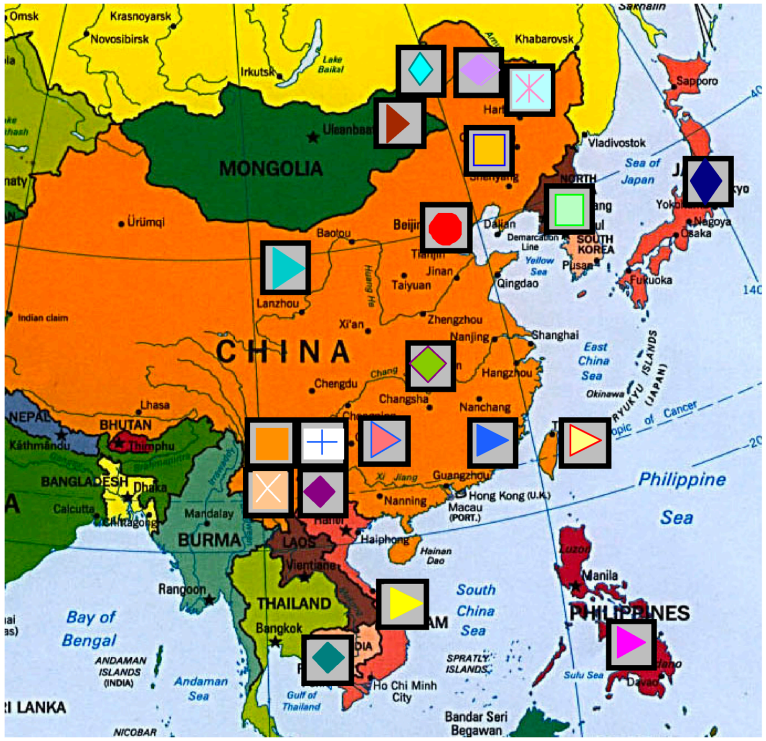


Analysis of East Asia Genetic Substructure Using Genome-Wide SNP Arrays

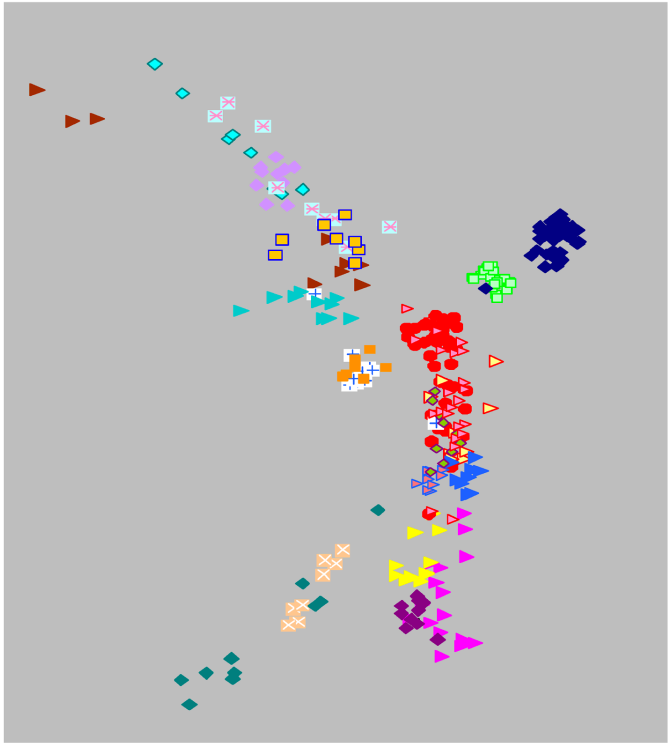
Chao Tian, Roman Kosoy, Annette Lee, Michael Ransom, John W. Belmont, Peter K. Gregersen, Michael F. Seldin

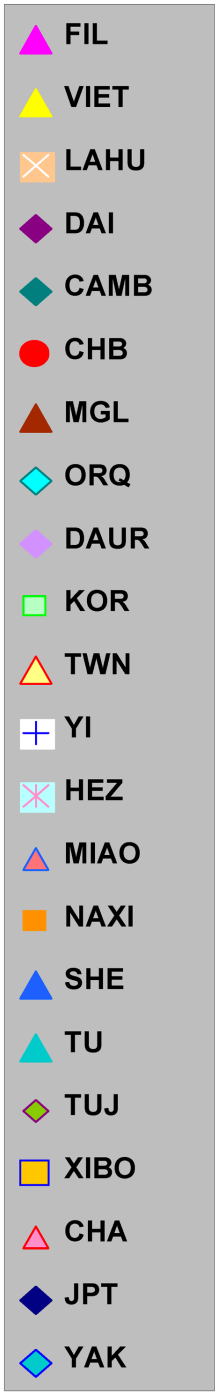
Published: December 5, 2008 • https://doi.org/10.1371/journal.pone.0003862

C



D

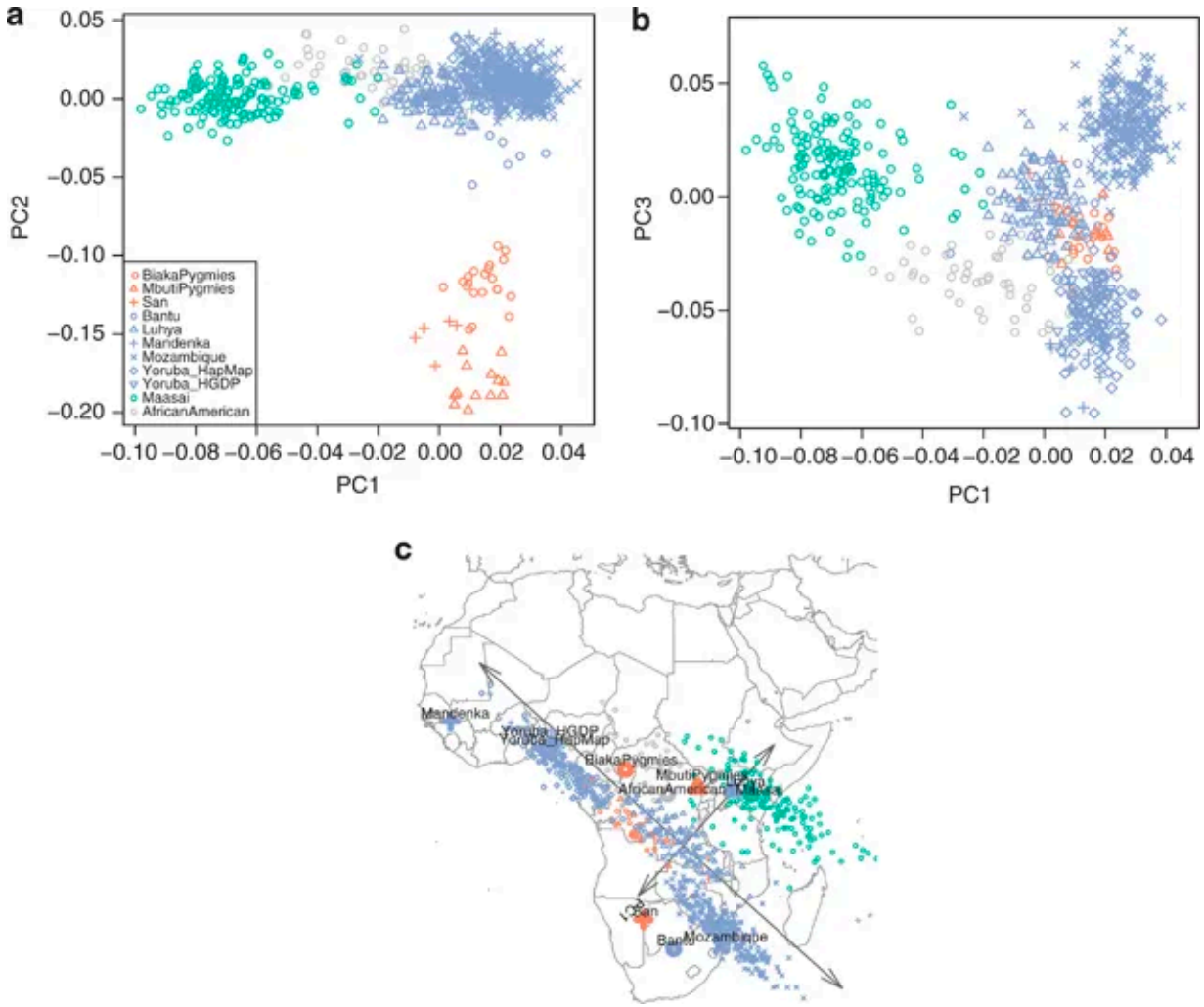




Article | Published: 25 August 2010

A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations

Martin Sikora, Hafid Laayouni, Francesc Calafell, David Comas & Jaume Bertranpetit



Handout 19

n = 1000 2 million

Step 1: Get the data. In this small example we will have $n = 6$ data points and $p = 2$ features. In reality we would have many more of each, and sometimes $p >> n$. The data matrix with n rows and p columns is called X_{orig}:

f g

$$X_{\text{orig}} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

n × p

6 × 2

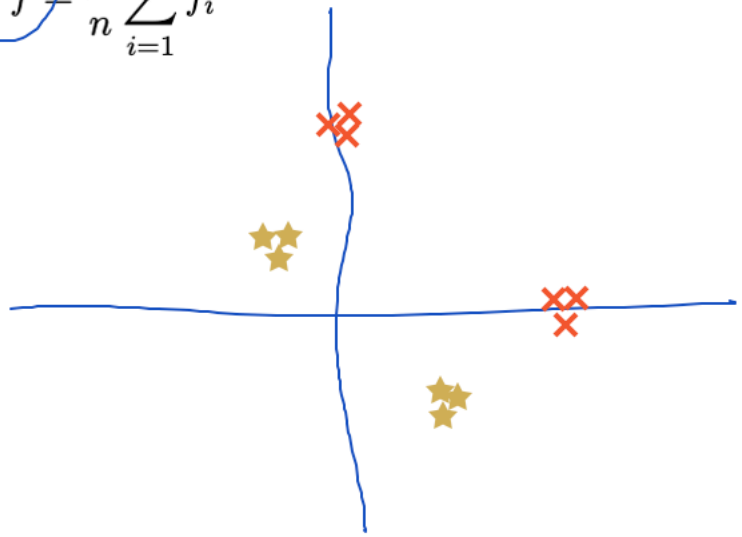
$$X = \begin{bmatrix} -1/2 & 1/2 \\ -1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix}$$

Step 2: Subtract off the column-wise mean from each column (feature) to obtain X (fill in above). The mean of column f is:

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$$

$\bar{f} = 1/2$

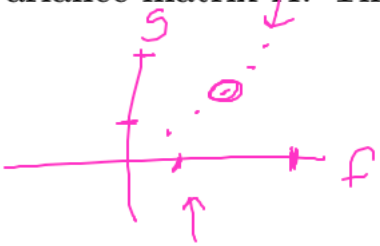
$\bar{g} = 1/2$



Handout 19

Step 3: Compute the covariance of each pair of features in X to obtain the $p \times p$ covariance matrix A. The covariance of feature f with feature g is:

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



Note that in our case, we have set all the means to be 0. Also note that variance is a special case when $f = g$:

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

Fill in A below:

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

$\text{var}(f) = \frac{1}{5}\left(6\left(\frac{1}{2}\right)^2\right) = 3/10$

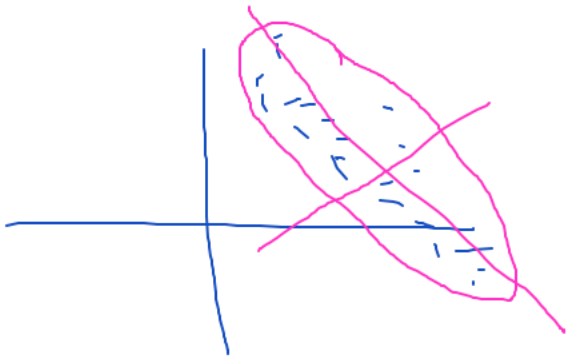
$\text{cov}(f, g) = -\frac{3}{10}$

$\text{var}(g) = 3/10$

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

Handout 19

Step 4: Compute the eigenvalues (λ_1, λ_2 for $p = 2$) and eigenvectors ($\vec{v}_1$, $\vec{v}_2$) of A. The eigenvectors (sorted by eigenvalue) will become the directions of our principal components (i.e. new coordinate system). We want our eigenvectors and eigenvalues to satisfy:



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

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

$$\vec{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

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

$$\vec{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Sort ☆

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

$$\lambda_2 = 0$$

Handout 19

Step 5: Transform the data X using the eigenvector matrix W (one eigenvector on each column, sorted by eigenvalue). The number of eigenvectors we use corresponds to the number of dimensions we retain. Say we want to retain r dimensions, then we would obtain the transformed data $T_r = XW_r$. T_r will be an $n \times r$ matrix. In our case, use $r = 2$ and compute T_r.

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

$p < 1$ $p < 2$

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

$n \times p$ $p \times r$ $n \times r$

weighted linear combination of features

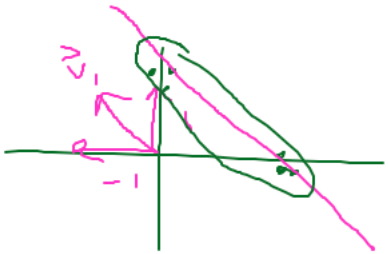
$-\frac{1}{2}(1) + \frac{1}{2}(-1)$

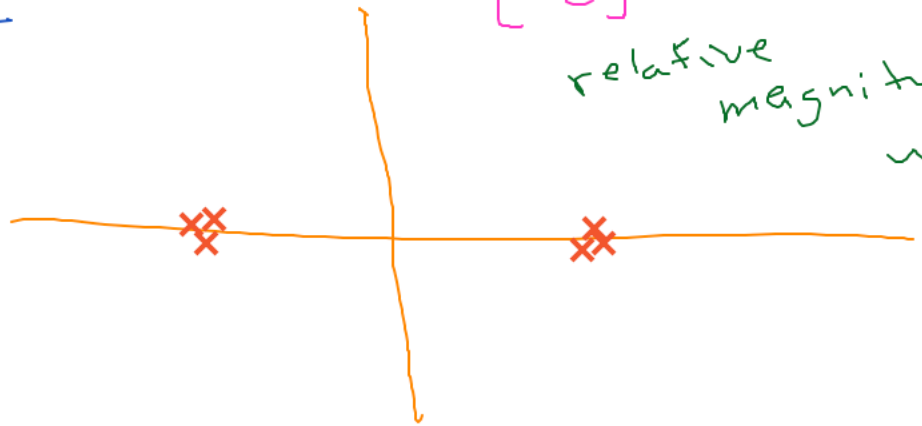
2 mill

4

$\begin{bmatrix} 2 \\ -2 \end{bmatrix}$

relative magnitude of weights





Handout 19

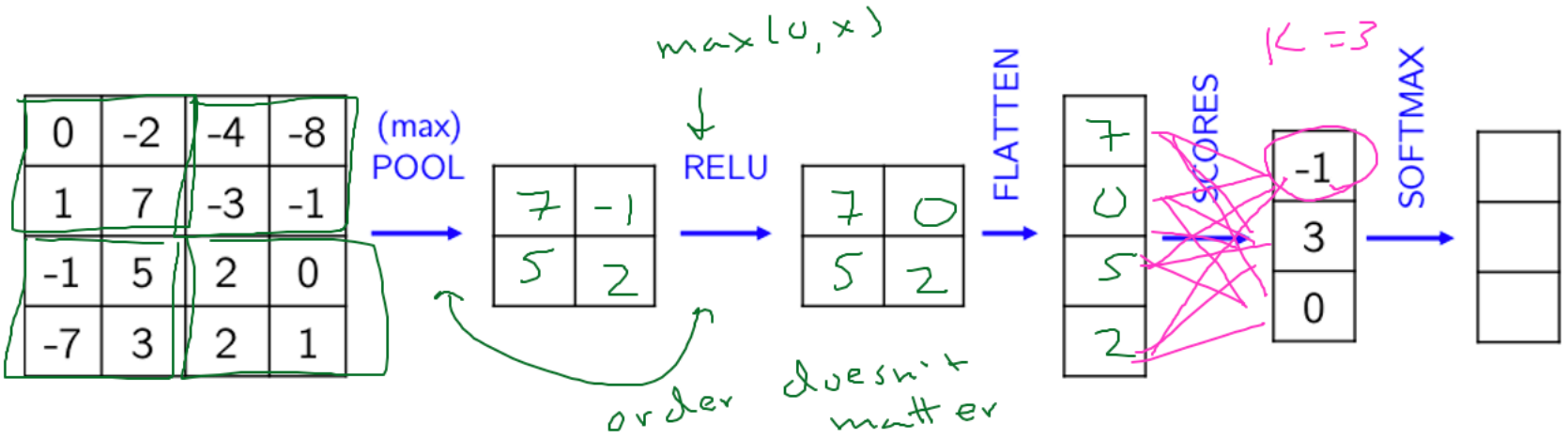
Step 6: Finally, plot the transformed data T_r with principal component 1 (PC1) on the x-axis and PC2 on the y-axis. We could plot further PCs on different coordinate systems when $p > 2$.

Outline for December 8

- Dimensionality reduction
- Principal Component Analysis (PCA)
- Midterm II review and practice problems

Handout 20, Q5

5. Say I have the following output of a CONV layer on the left. Assume no bias terms throughout.

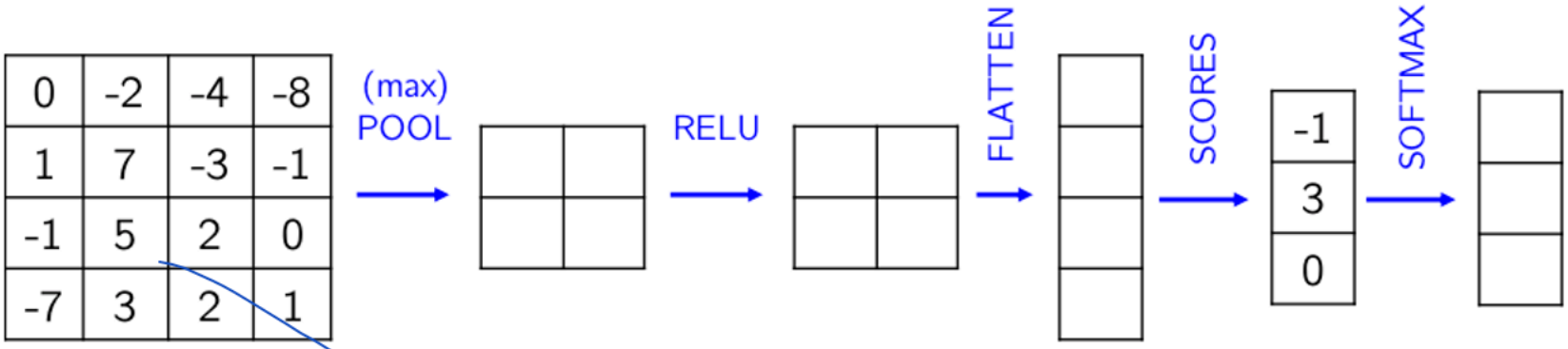


(a) If my original input was also 4×4 and I used one convolutional filter with size 3×3 (no bias), how much zero padding would I need? How many parameters would I need to learn just for this CONV layer?

(b) Fill in the steps POOL (2×2 with stride 2), RELU, FLATTEN.

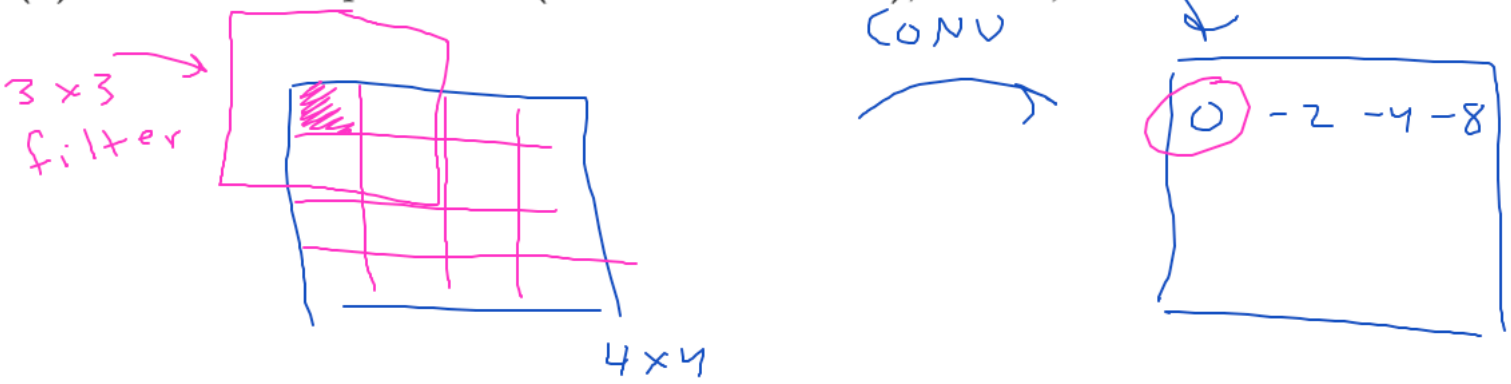
Handout 20, Q5

5. Say I have the following output of a CONV layer on the left. Assume no bias terms throughout.



(a) If my original input was also 4×4 and I used one convolutional filter with size 3×3 (no bias), how much zero padding would I need? How many parameters would I need to learn just for this CONV layer?

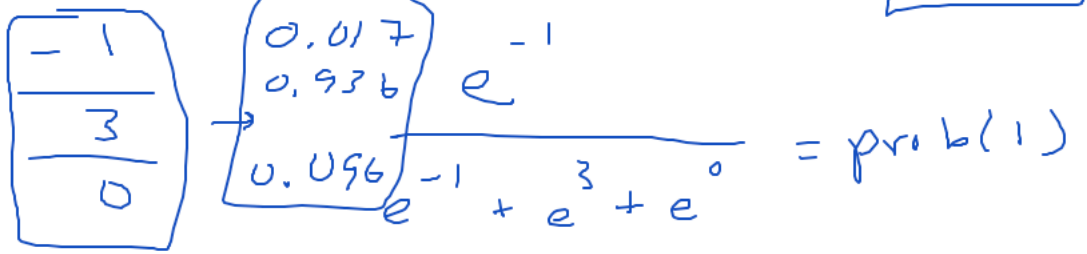
(b) Fill in the steps POOL (2×2 with stride 2), RELU, FLATTEN.

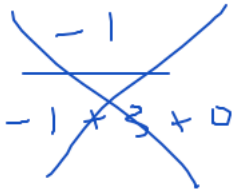


Handout 20, Q5

(c) Say the scores are as given above for three potential labels $\hat{y} \in \{1, 2, 3\}$. Compute the SOFTMAX function to obtain a probability distribution over these three classes. What would you choose for the predicted label $\hat{y}$?

$\hat{y} = 2$



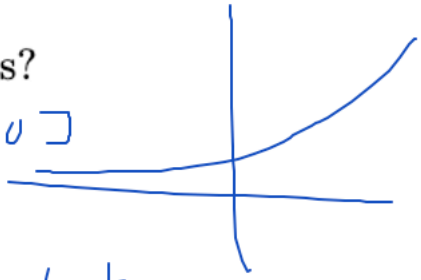


(d) If the true class was in fact $y = 2$, what is the cross-entropy loss?

$$H(y, \hat{y}) = -\sum_{k=1}^{K} y \log \hat{y}$$

$$= -1 \cdot \log(0.936)$$

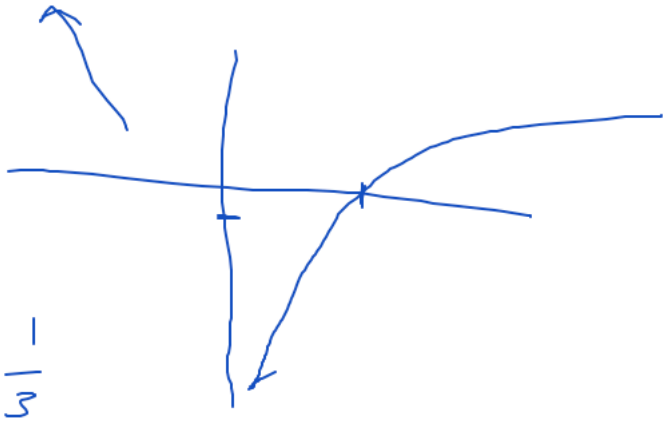
$y = [0, 1, 0]$ one hot



(e) In the input had been a matrix of zeros, what would the scores be? What would the probability distribution (output of SOFTMAX) be?

$= 0.0286$

$$\frac{e^0}{3 \cdot e^0} = \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$$



More Handout 20 solutions

Handout 20, Question 1

- First compute weighted leaf labels

$$P(+ \mid \text{sun}) = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{8} + \frac{1}{8}} = \frac{4}{7} \geq 0.5 \quad \Rightarrow +$$

Handout 20, Question 1

- First compute weighted leaf labels

$$P(+ \mid \text{sun}) = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{8} + \frac{1}{8}} = \frac{4}{7} \geq 0.5 \quad \Rightarrow +$$

$$P(+ \mid \text{rain}) = \frac{\frac{1}{12}}{\frac{1}{12} + \frac{1}{6} + \frac{1}{6}} = \frac{1}{5} < 0.5 \quad \Rightarrow -$$

Handout 20, Question 1

- Based on these labels, we can say which training points are misclassified

$$\epsilon_t = \frac{1}{8} + \frac{1}{8} + \frac{1}{12} = \frac{1}{3}$$

Handout 20, Question 1

- Based on these labels, we can say which training points are misclassified

$$\epsilon_t = \frac{1}{8} + \frac{1}{8} + \frac{1}{12} = \frac{1}{3}$$

- Note if this was > 0.5, we should have chosen different leaf labels! So this “flipping” step should happen automatically
  - (exception for pathological cases)

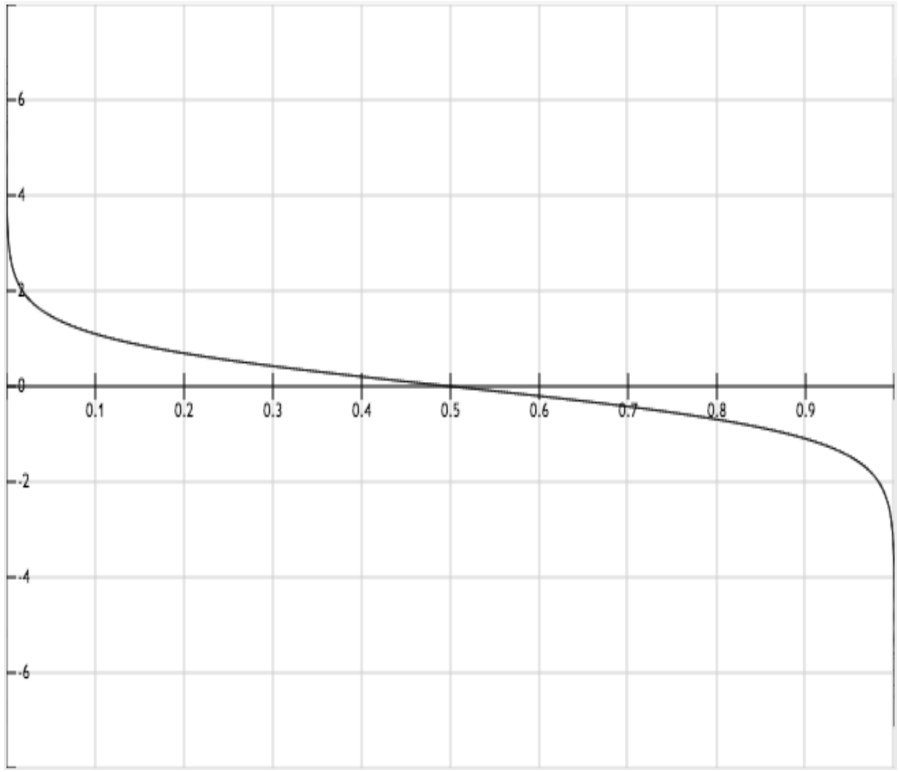
Handout 20, Question 1

- Score function:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$$

- Fraction: accuracy/error

- As error -> 0, score becomes high
- As error -> ½, score goes to 0



Handout 20, Question 2

- r = 1/3, probability of one classifier being wrong
- T = 5, number of classifiers
- R = number of votes for the wrong class
- If R=3,4,5 then we will vote for the wrong class overall

Handout 20, Question 2

- This analysis assumed classifiers were independent!

- What if they are not? How did Random Forests help us decorrelate classifiers?

Handout 20, Question 2

- This analysis assumed classifiers were independent!
- What if they are not? How did Random Forests help us decorrelate classifiers?
- Note about Bagging: choosing *n* with resampling actually does produce a very different dataset
  - As *n* increases, roughly 0.37 not chosen each time