Assignment 4 on PCA and logistic regression

S&DS 355 / 555

Introductory Machine Learning

Unsupervised Learning: PCA

Tuesday, October 1

Yale

Unsupervised Learning

Supervised learning is about being able to predict a Y using a series of predictors $X_1, X_2, \ldots, X_p$.

Unsupervised learning deals with data that do not have labels Y.

We are not trying to predict anything. So what else might we hope to do?

Unsupervised Learning

Supervised learning is about being able to predict a Y using a series of predictors $X_1, X_2, \ldots, X_p$.

Unsupervised learning deals with data that do not have labels Y.

We are not trying to predict anything. So what else might we hope to do?

Consider:

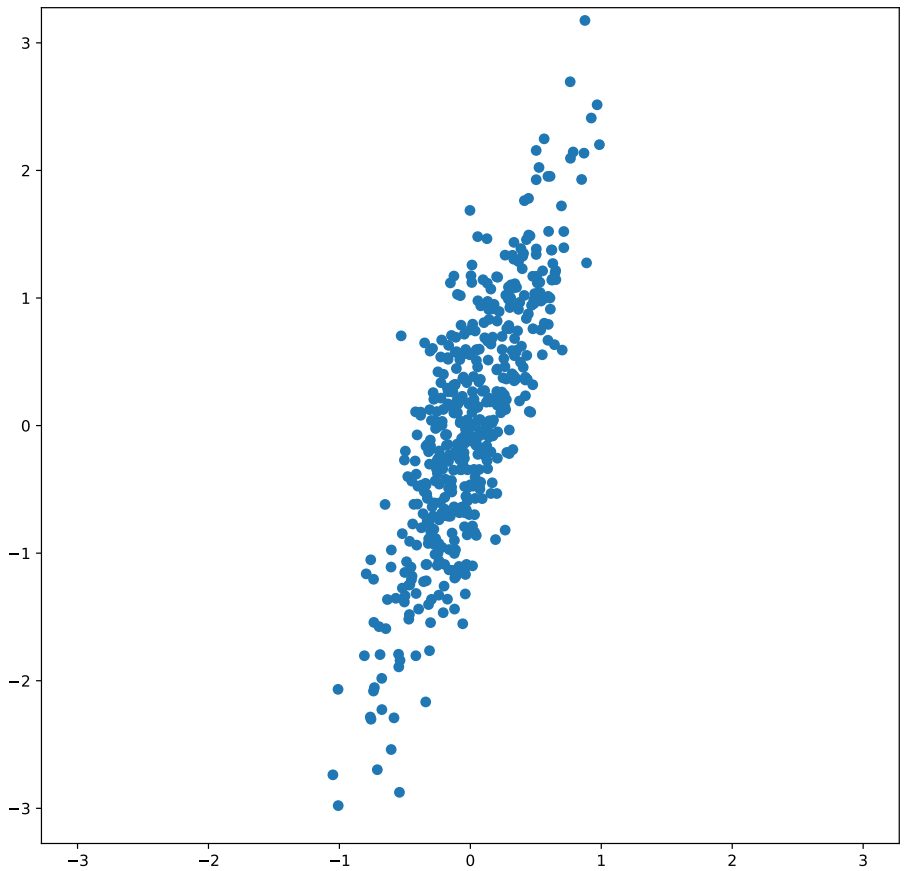
- Are there interesting ways to visualize/summarize the data?
- Are there natural subgroups in the data?

you can think of PCA as a kind of dimension reduction

Principal Component Analysis (PCA)

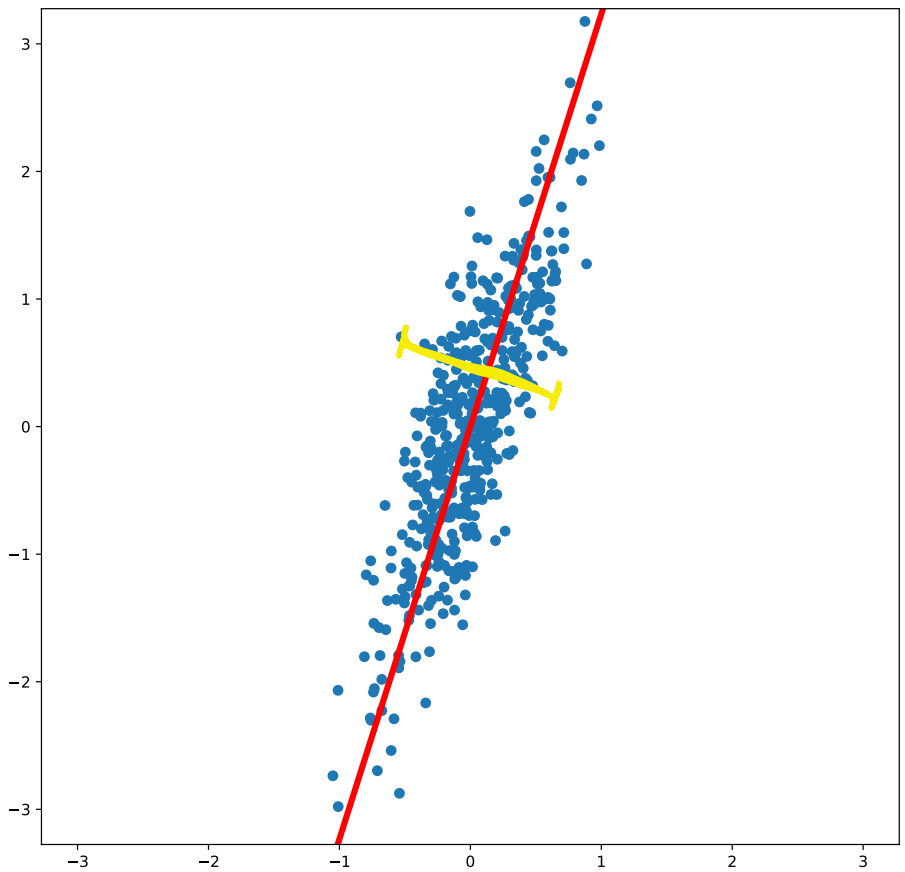
PCA finds the directions of greatest variability in the data.

the maximum variation/spread of data is along the red axis (next slide)



Principal Component Analysis (PCA)

PCA finds the directions of greatest variability in the data.

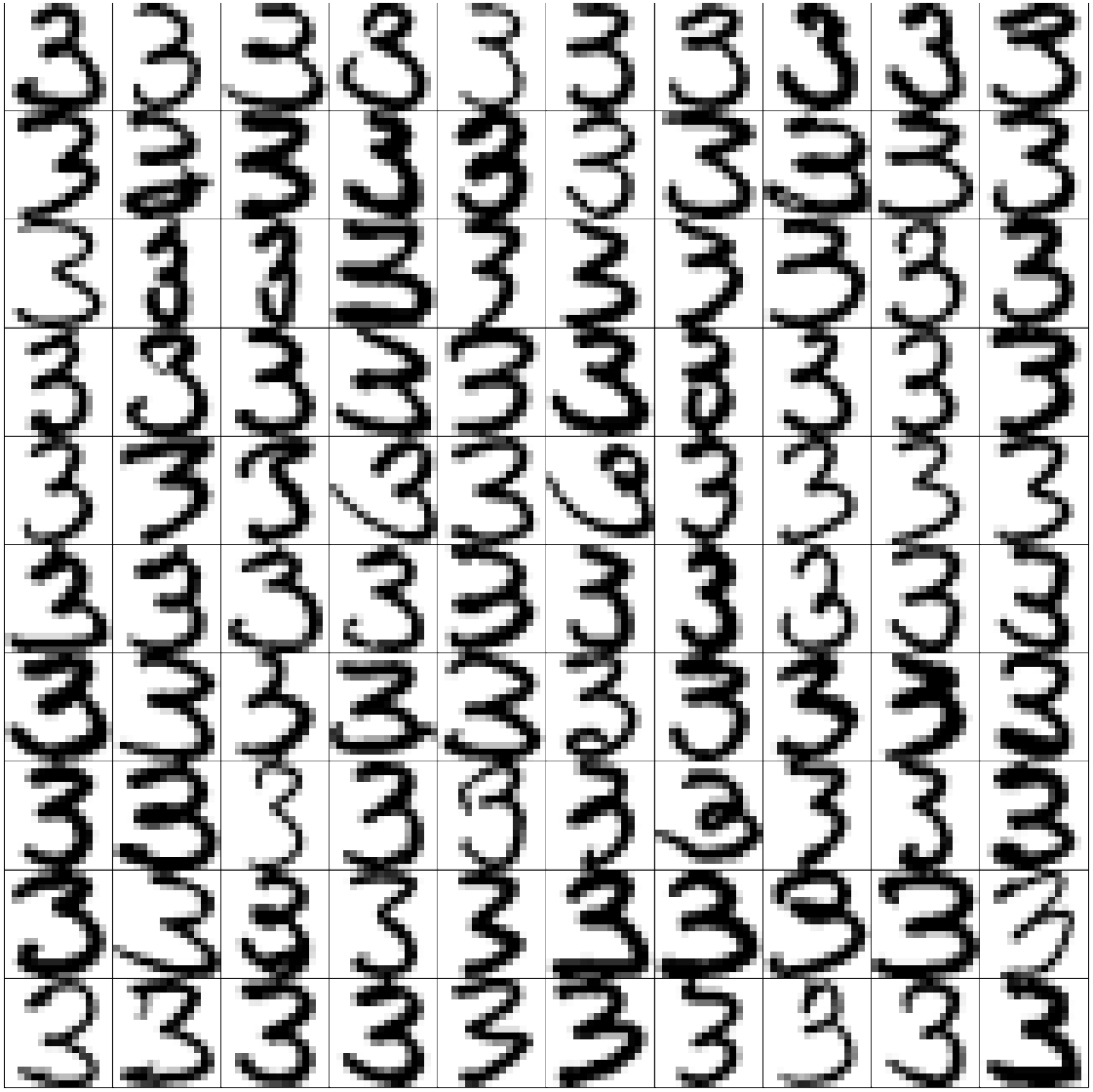


the red axis is the first principal direction

the yellow axis is the second principal direction and is orthogonal (must be orthogonal) to the first component

Handwritten Digits (3s)

from MNIST



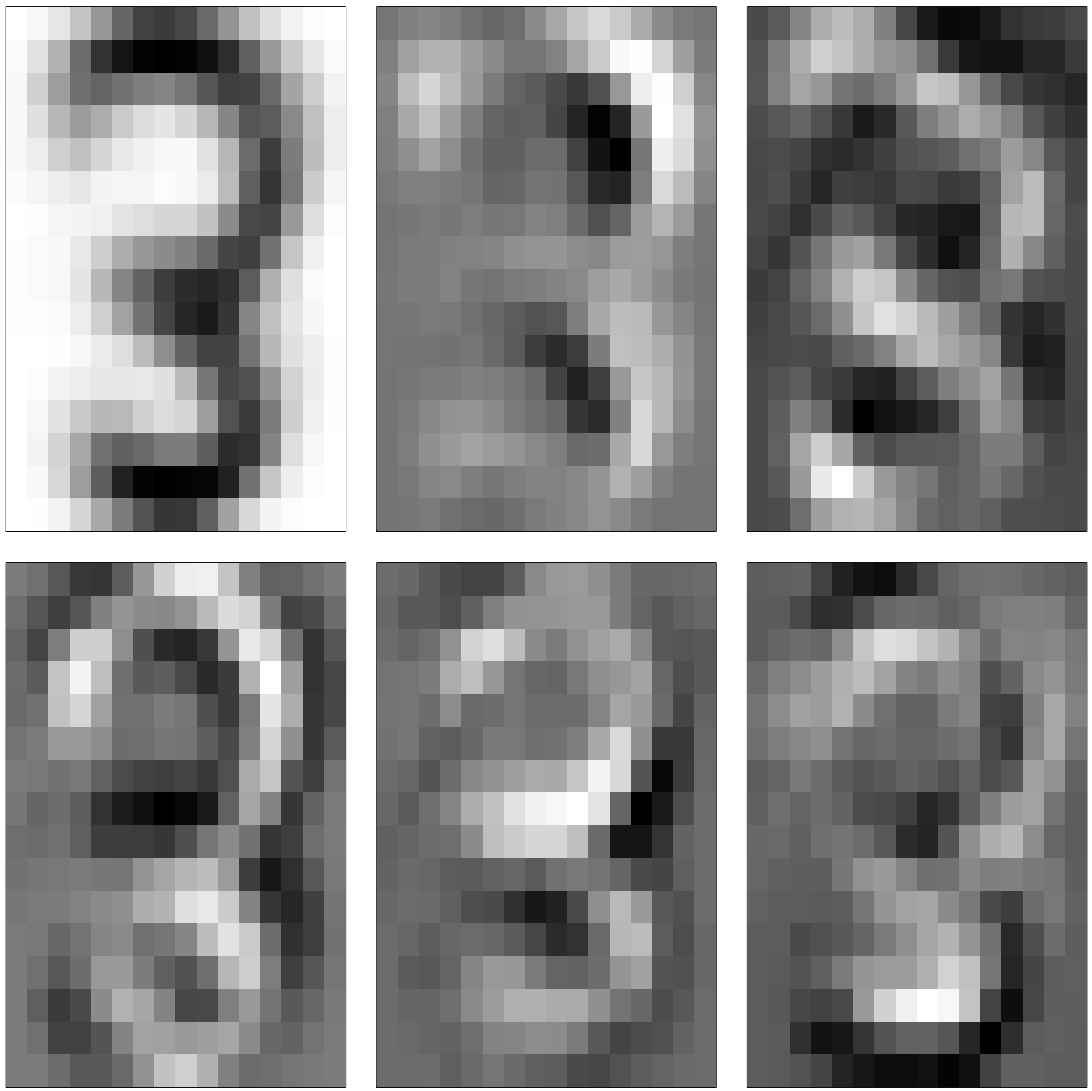
what if you average all of these?

- First, you take average of all 3s.
- Then you subtract average from all 3s to center them according to origin
- Principal components vs principal vectors?
- if you add all principal vectors, sum = = 0 (they are perpendicular to each other)

Handwritten Digits (3s) – PCA

this one modifies the average to reduce the right edges and increase the inner blacks

average 3

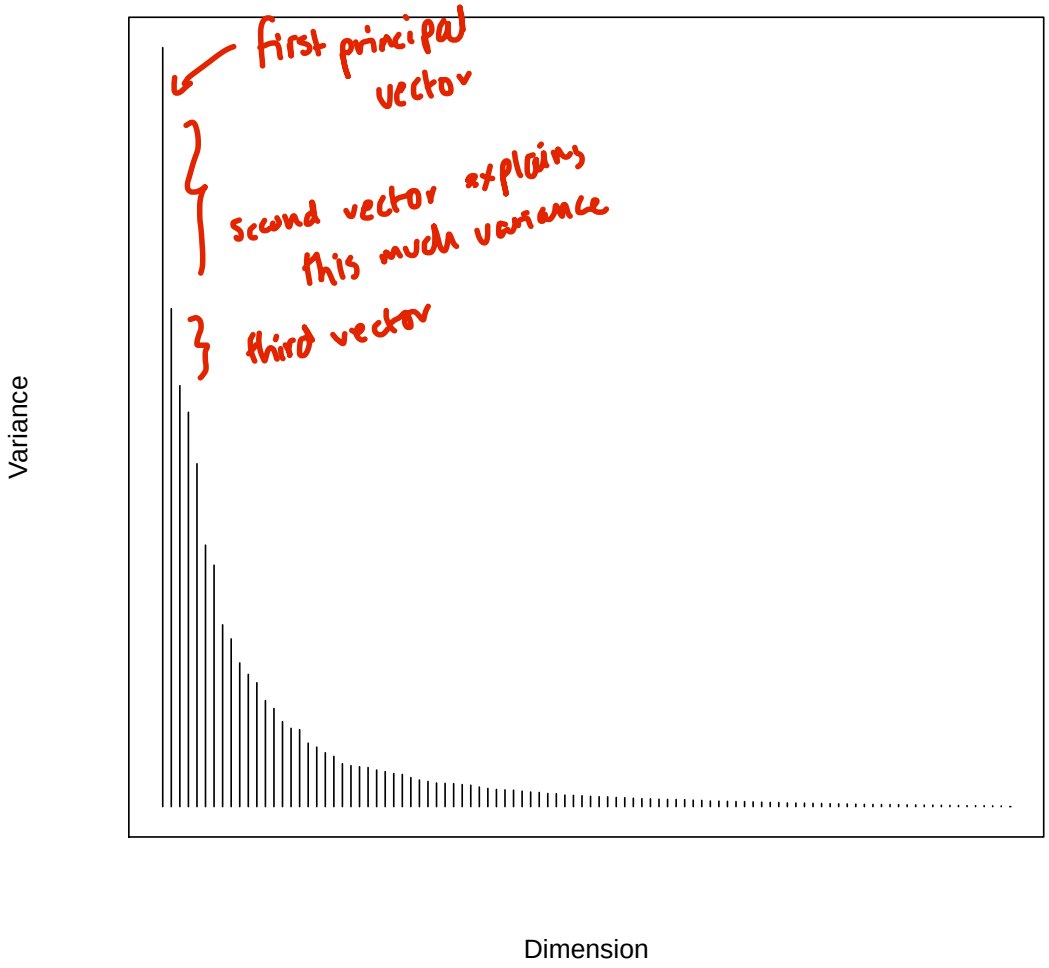


then look at all other data points with the average subtracted away

all these 5 components are orthogonal

multiplying these two together makes them cancel

Handwritten Digits (3s) – PCA variance



Handwritten Digits (3s) – PCA reconstruction

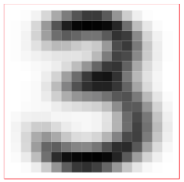
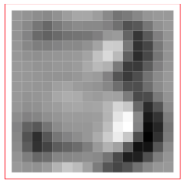
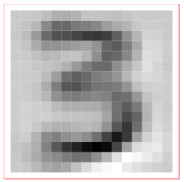
datapoints can be projected to each principal vector to give an approximation to the data point

this is the best 10-D representation of the data

Handwritten Digits (3s)

average

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

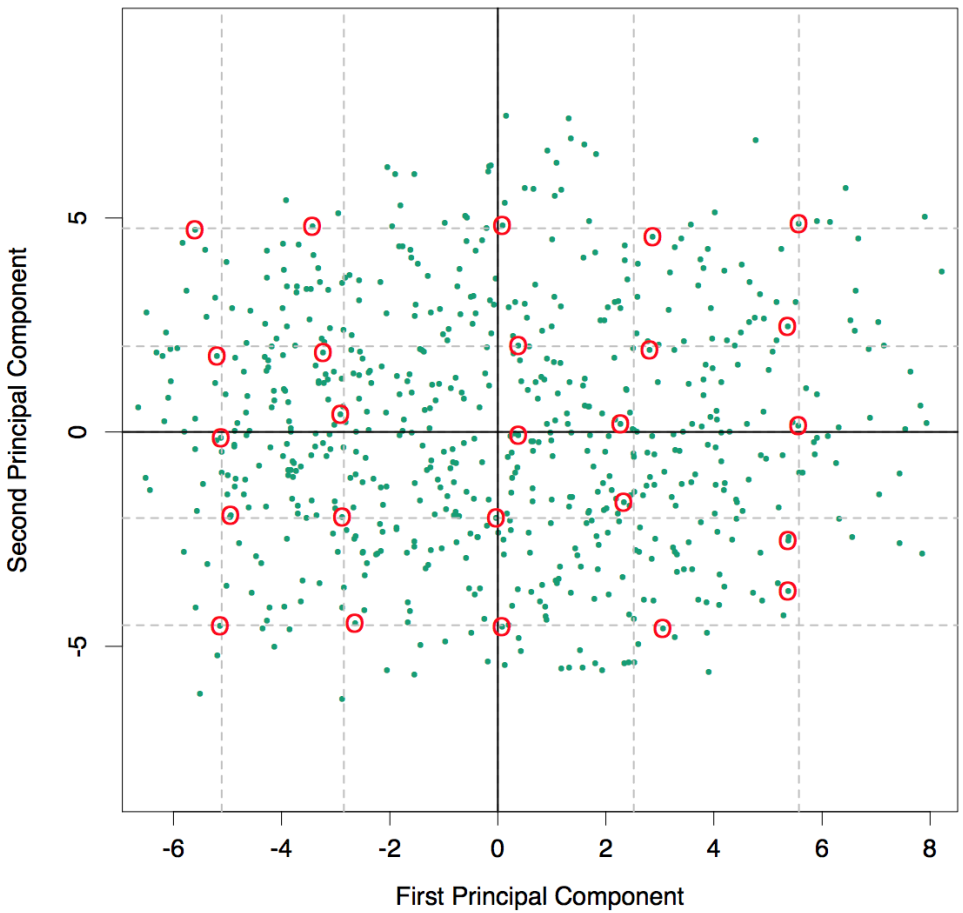
$$= \text{  } + \lambda_1 \cdot \text{  } + \lambda_2 \cdot \text{  }.$$

first two principal vectors

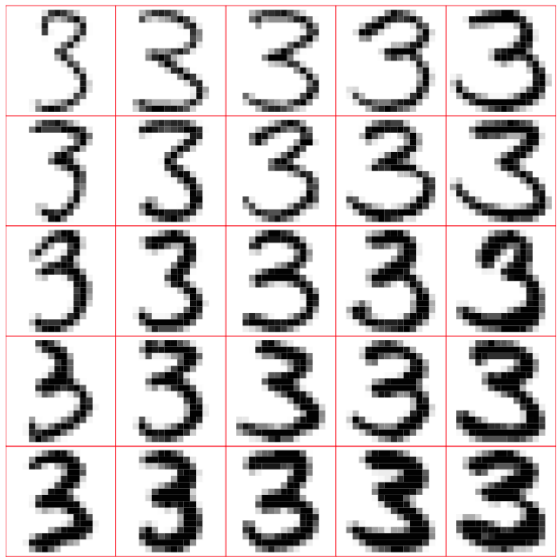
Handwritten Digits (3s) – Top 2 components

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \bar{x} + \lambda_1 \cdot v_1 + \lambda_2 \cdot v_2 .$$



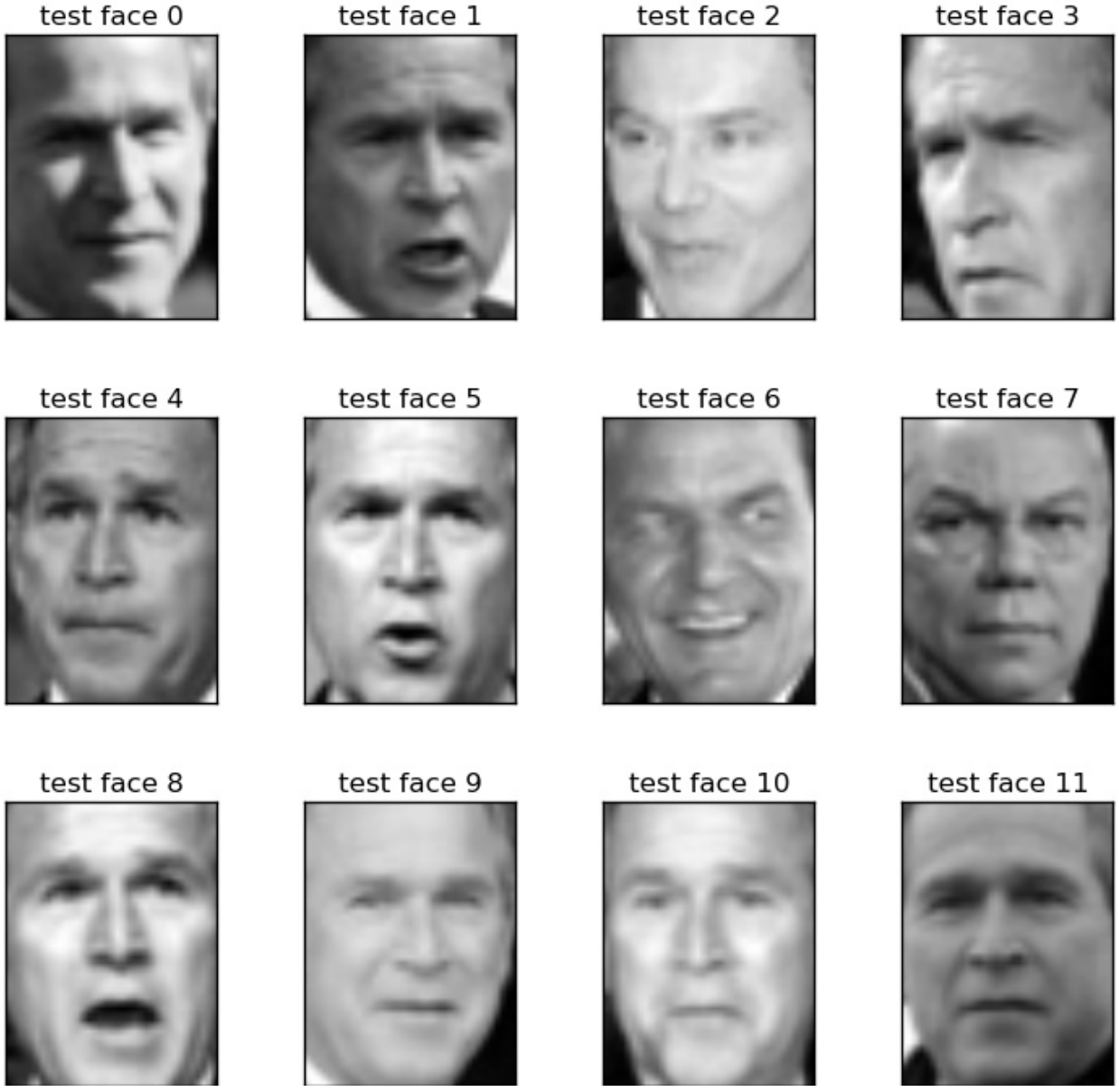
λ_2 wideness of lines



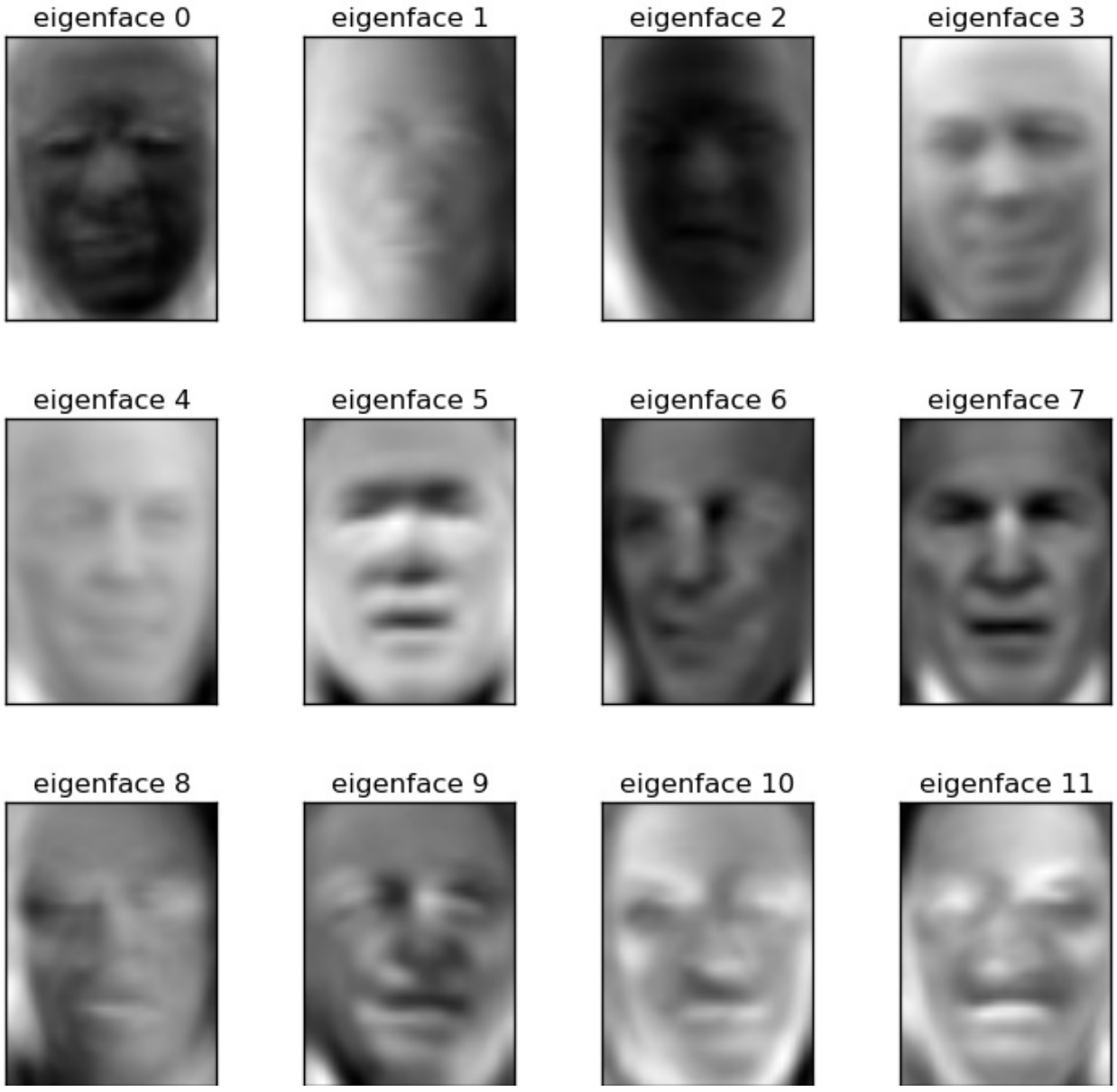
λ_1 thickness of lines

Faces

formed from principal components of ideal images.



Eigenfaces


eigenface 0
eigenface 1
eigenface 2
eigenface 3
eigenface 4
eigenface 5
eigenface 6
eigenface 7
eigenface 8
eigenface 9
eigenface 10
eigenface 11

Nature. Author manuscript; available in PMC 2009 Aug 31.
Published in final edited form as:
Nature. 2008 Nov 6; 456(7218): 98–101.
Published online 2008 Aug 31. doi: 10.1038/nature07331
PMCID: PMC2735096
NIHMSID: NIHMS132060

Genes mirror geography within Europe

John Novembre,[1,2] Toby Johnson,[4,5,6] Katarzyna Bryc,[7] Zoltán Kutalik,[4,6] Adam R. Boyko,[7] Adam Auton,[7] Amit Indap,[7] Karen S. King,[8] Sven Bergmann,[4,6] Matthew R. Nelson,[8] Matthew Stephens,[2,3] and Carlos D. Bustamante[7]

Author information ▶ Copyright and License information ▶

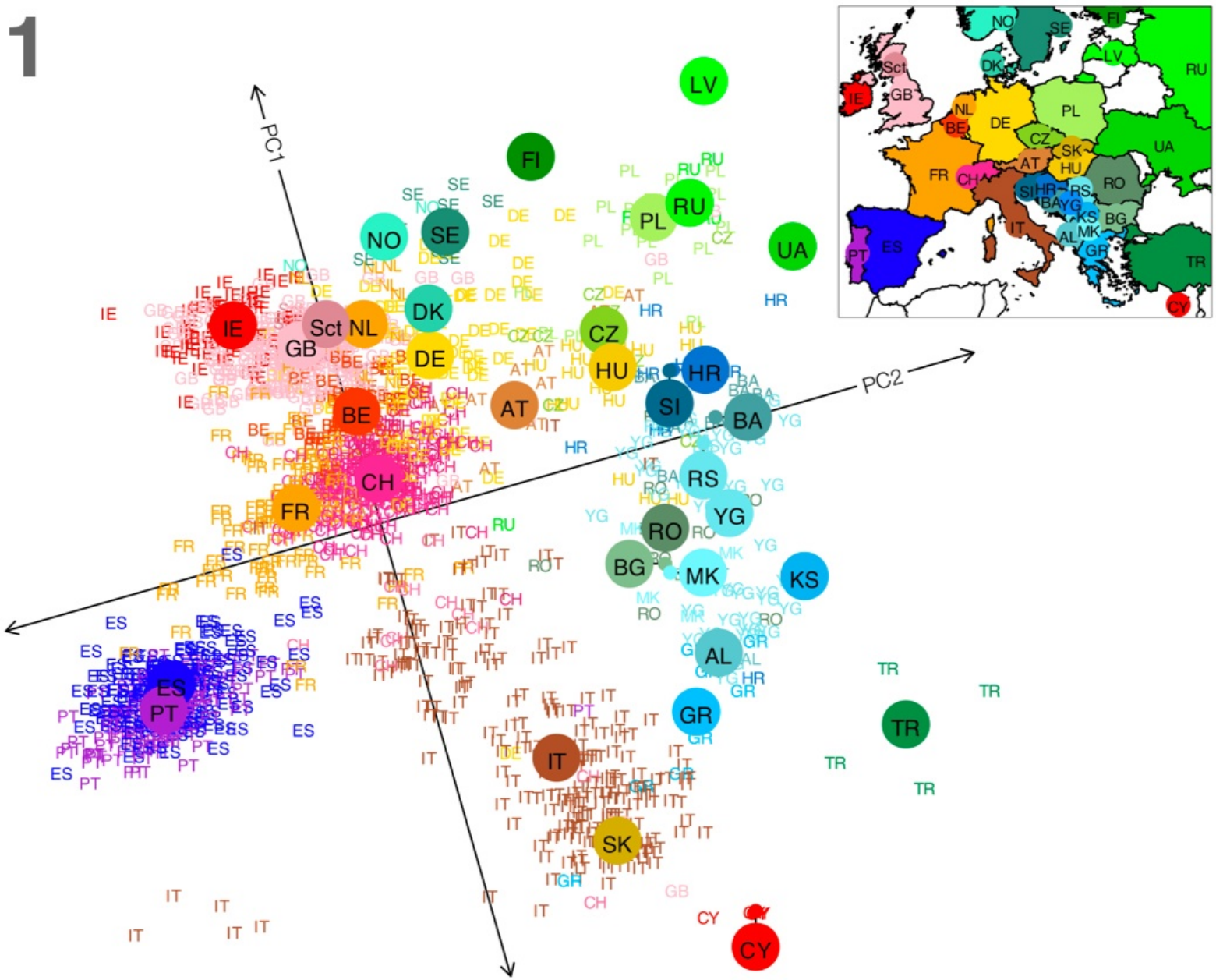
The publisher's final edited version of this article is available at Nature
This article has been corrected. See the correction in volume 456 on page 274.
See commentary "Editorial comment should accompany hot papers online." in *Nature*, volume 455 on page 861.

Abstract

Go to:

Understanding the genetic structure of human populations is of fundamental interest to medical, forensic and anthropological sciences. Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations[1–5]. Here we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing[6]; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

1


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

Robust PCA

Robust PCA (low rank plus sparse) can be used for background subtraction in video.

https://www.youtube.com/watch?v=BTrbow8u4Cw

PCA: Algorithm

1. Center the data: $x_i \mapsto x_i - \overline{x}$
2. Compute the $d \times d$ sample covariance $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$
3. Find the first k eigenvectors of S
4. Project the data onto those k vectors

PCA: Algorithm

1. Center the data: $x_i \mapsto x_i - \frac{1}{n}\sum_{j=1}^{n} x_j$

2. Compute the $d \times d$ sample covariance $S = \frac{1}{n}.\sum_{i=1}^{n} x_i x_i^T$. Note that

$$\frac{1}{n}(x_i - \overline{x})^2$$

   is the sample variance of 1-dimensional data

3. Find the first k eigenvectors of S,

$$\phi_1, \ldots, \phi_k \in \mathbb{R}^d, \qquad S\phi_j = \lambda_j \phi_j$$

4. Project the data onto those k vectors:

$$x_i \mapsto (\phi_1^T x_i)\phi_1 + \ldots + (\phi_k^T x_i)\phi_k$$