



A PCA scatter plot showing the relationship between PC1 (horizontal axis) and PC2 (vertical axis) for various European countries. Countries are represented by colored circles and labeled with their ISO codes. The plot shows distinct clusters: a group of countries in the top-left (e.g., IE, GB, NL, DK, SE, NO), a group in the top-right (e.g., PL, RU, UA, HR, CZ, HU, SI, BA, RS, BG, MK, AL, GR, TR), and a group in the bottom-left (e.g., ES, PT, IT, SK, CY). An inset map in the top right corner shows the geographical locations of the countries. Arrows labeled 'PC1' and 'PC2' indicate the direction of the principal components.

S&DS 355 / 555
Introductory Machine Learning

Unsupervised Learning: PCA

Tuesday, October 1

Unsupervised Learning

Supervised learning is about being able to predict a Y using a series of predictors X_1, X_2, \dots, 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, \dots, 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:

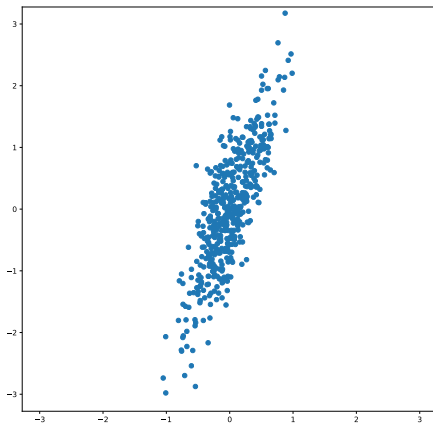
- Are there interesting ways to visualize/summarize the data?
- Are there natural subgroups in the data?

*you can think of PCA as
a kind of dimension reduction*

Principal Component Analysis (PCA)

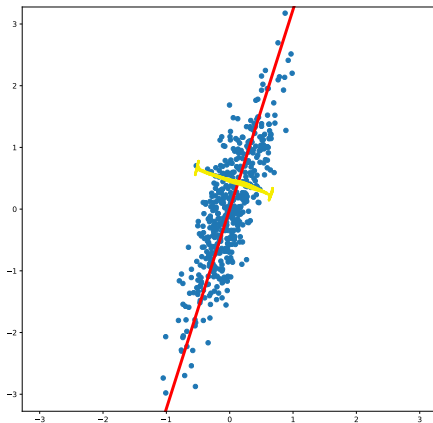
PCA finds the directions of greatest variability in the data.

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.

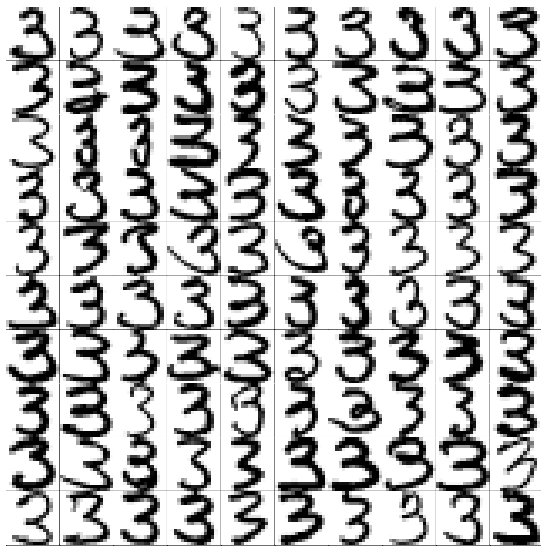


the red axis is
the first
principal direction

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

Handwritten Digits (3s)

from MNIST

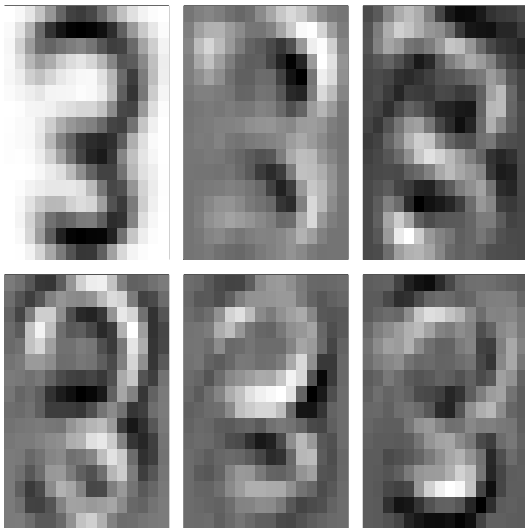


what if you average all of these?

Handwritten Digits (3s) – PCA

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

average 3 →

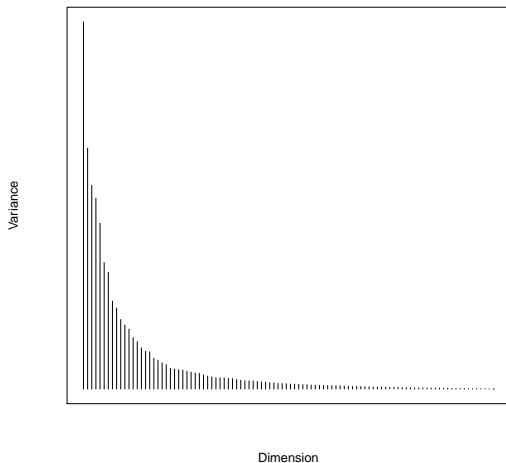


then look at
all other
data points
with the
average subtracted
away

all these 5
components are
orthogonal

multiplying
these two
together
makes them cancel

Handwritten Digits (3s) – PCA variance



Handwritten Digits (3s) – PCA reconstruction

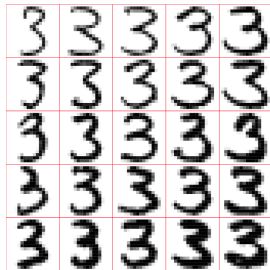
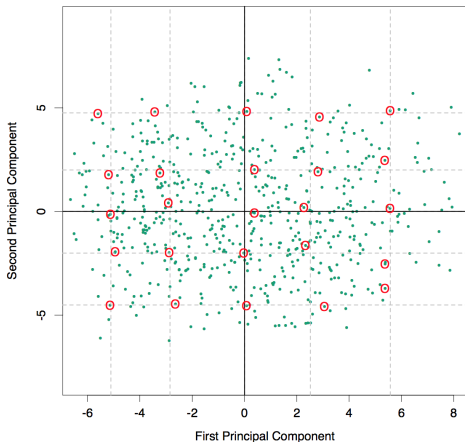


Handwritten Digits (3s)

$$\begin{aligned}\hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \boxed{\text{3}} + \lambda_1 \cdot \boxed{\text{3}} + \lambda_2 \cdot \boxed{\text{3}}.\end{aligned}$$

Handwritten Digits (3s) – Top 2 components

$$\begin{aligned}\hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \text{[Image of mean digit 3]} + \lambda_1 \cdot \text{[Image of } v_1 \text{]} + \lambda_2 \cdot \text{[Image of } v_2 \text{]}.\end{aligned}$$



Faces

test face 0



test face 1



test face 2



test face 3



test face 4



test face 5



test face 6



test face 7



test face 8



test face 9



test face 10



test face 11



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.

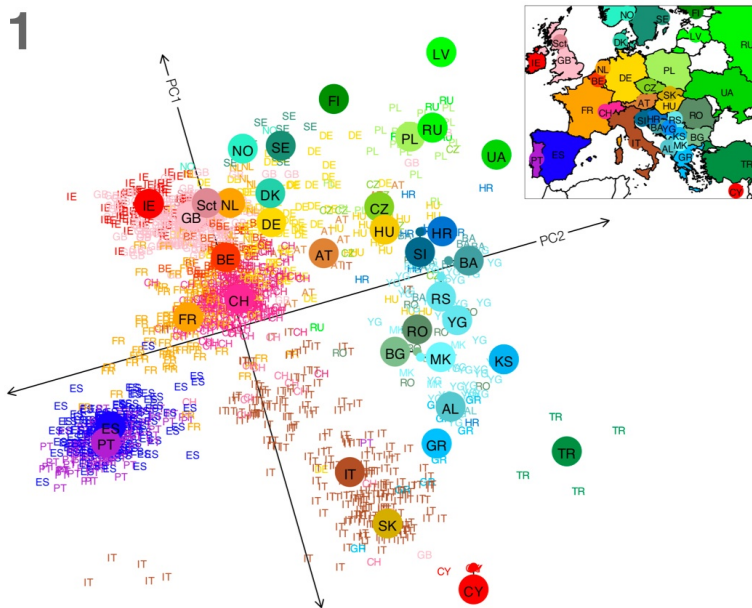
See other articles in PMC that [cite](#) the published article.

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 testing⁶; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

1



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: $x_i \mapsto x_i - \bar{x}$
- 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 $S = \frac{1}{n} \cdot \sum_{i=1}^n x_i x_i^T$. Note that

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

is the sample variance of 1-dimensional data

- 3 Find the first k eigenvectors of S ,

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

- 4 Project the data onto those k vectors:

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