



S&DS 265 / 565
Introductory Machine Learning

Random Forests and Unsupervised Learning: PCA

October 3

Yale

Checkpoint

- Assn 2 due today (midnight)
- Assn 3 will be posted tonight
- Midterm October 15, in class
- Practice midterms (and solutions) posted on Canvas
- Review sessions will be held next week

Random Forests

- Grow many trees and average their predictions
- Trees are grown deep, to have low bias, but high variance
- To “decorrelate” the trees and reduce variance, each tree is
 - ▶ grown on a bootstrap sample of the data
 - ▶ grown with random subsets of the predictors at each split
- Tree growing can be done in parallel

Leo Breiman—“Keep it simple”



Random Forests Algorithm

- ① For $b = 1$ to B :
 - (a) Draw a bootstrap sample Z^* of size n from the training data
 - (b) Grow a random-forest tree T_b to the bootstrapped data, recursively repeating following steps, until minimum node size reached:
 - i. Select m variables at random from the p variables
 - ii. Pick the best variable/split-point among the m
 - iii. Split the node into two children nodes
- ② Output the ensemble of trees $\{T_b\}_{b=1}^B$.

Random Forests Algorithm

To make a prediction at a new point x :

Regression: Average $\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$

Classification: Majority vote of the individual trees

Out of bag (OOB) prediction

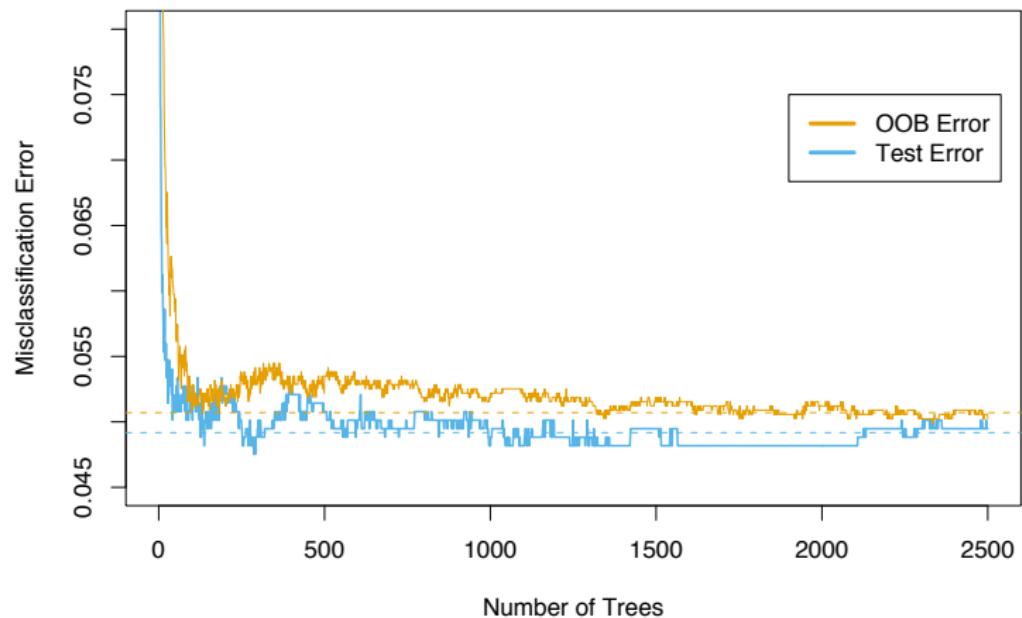
A bootstrap sample, or “bag” of data, is a set of n data points sampled with replacement from the original set of n data points. Will contain repetitions.

Each tree is grown on such a sample, which contains about $\frac{2}{3}$ of the original data (with repetitions). The remaining $\frac{1}{3}$ can be used as validation data

- For each observation $z_i = (x_i, y_i)$, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which z_i did not appear
- Thus, cross-validation can be performed “along the way”

Chance a sample x_i does not appear in a bootstrap sample is $\left(1 - \frac{1}{n}\right)^n \longrightarrow \frac{1}{e} \approx 0.37$

Out of bag (OOB) prediction



More detail: Bootstrap samples

- We have n data points
- Randomly sample exactly n of them *with replacement*
- Contains about $\frac{2}{3}$ of the data, with duplicates
- This is a “bag” of n data points

Chance a particular example does not appear in sample:

$$\left(1 - \frac{1}{n}\right)^n \longrightarrow \frac{1}{e} \approx 0.37$$

More detail: Out-of-Bag error estimation

Each bagged tree uses about 2/3 of all observations (with repeats).

The remaining data—*out-of-bag* (OOB) observations—can be put to good use.

More detail: Out-of-Bag error estimation

- Consider bags $\mathcal{B}_i \subset \{1, 2, \dots, B\}$ where (x_i, y_i) does *not* appear
- We average over those trees to form a prediction:

$$\hat{y}_i = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} T_b(x_i)$$

- OOB error estimate is then $\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2$

Performance on email spam task

Single tree: 8.7%

Random forest: 5.1%

(standard error of the estimates is $\approx 0.6\%$)

Variable importance

Random forests improve upon the predictive ability of trees, but this sacrifices the interpretability of the single tree model

Variable importance

Random forests improve upon the predictive ability of trees, but this sacrifices the interpretability of the single tree model

A good tool for interpreting a forest is **variable importance**

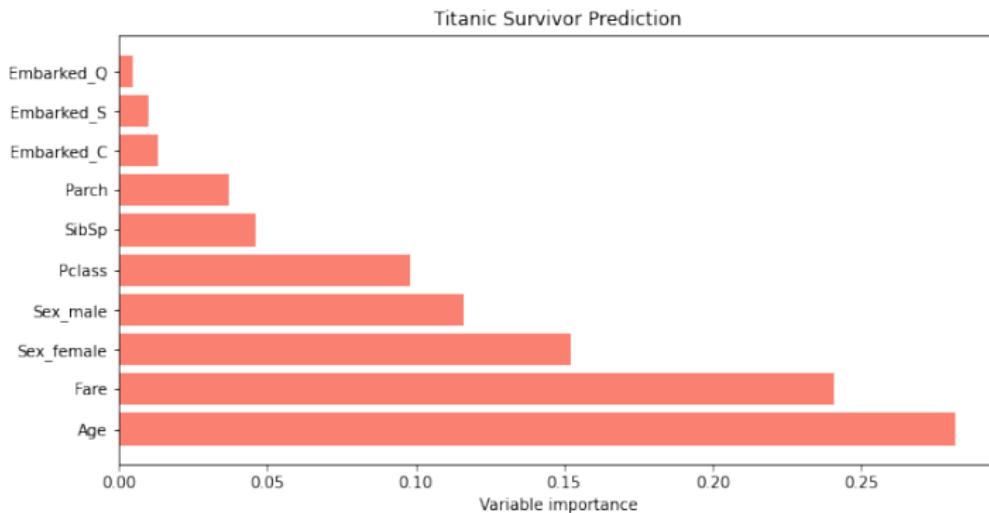
Variable importance

Random forests improve upon the predictive ability of trees, but this sacrifices the interpretability of the single tree model

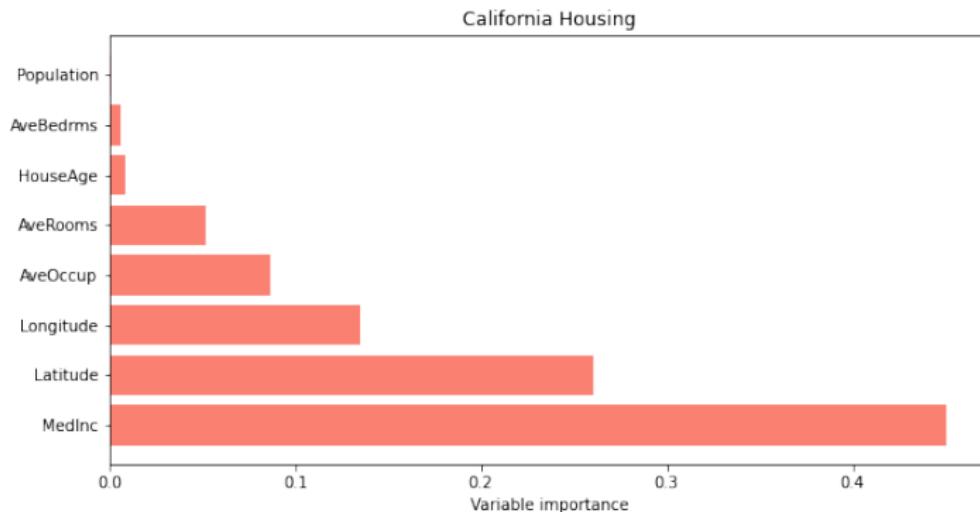
A good tool for interpreting a forest is **variable importance**

Variable importance is the amount that the RSS (or other loss) is reduced due to splits over a given predictor, averaged over all trees in the forest

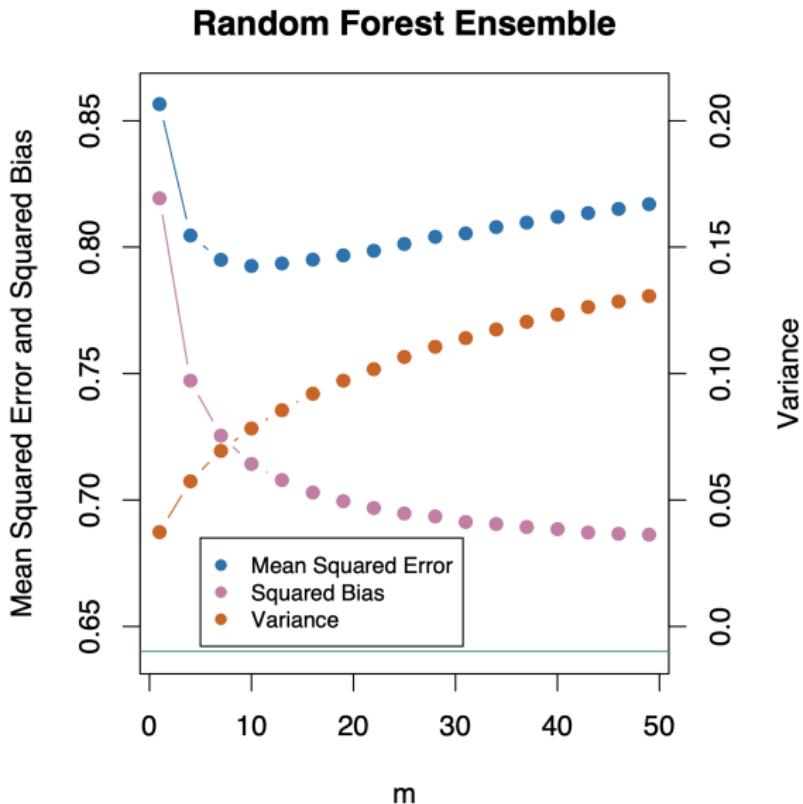
Variable importance



Variable importance



Random forest MSE



Understanding the bias-variance tradeoff for random forests

- Each tree has low (squared) bias, because it is deep
- The bootstrap sample and random subset of questions allowed at a given split result in “diversity” among of the trees
- This diversity translates to decorrelation, and tends to reduce the variance
- As we increase the number m in the random subset of questions allowed at each split, the complexity increases—variance goes up, squared bias goes down

Please step through the notebook

```
In [1]: import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

In [2]: titanic_train = pd.read_csv('https://raw.githubusercontent.com/minsuk-heo/kaggle-titanic/master/input/train.csv')
titanic_test = pd.read_csv('https://raw.githubusercontent.com/minsuk-heo/kaggle-titanic/master/input/test.csv')
titanic_train
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Allen, Mr. William Henry	male	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Montvila, Rev. Juozas	male	35.0	0	0	373450	8.0500	NaN	S
...
886	887	0	2	Graham, Miss. Margaret Edith	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Johnston, Miss. Catherine Helen "Carrie"	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Behr, Mr. Karl Howell	male	26.0	0	0	W.C. 6607	23.4500	NaN	S
889	890	1	1	Dooley, Mr. Patrick	male	32.0	0	0	111369	30.0000	C148	C
890	891	0	3						370376	7.7500	NaN	Q

891 rows × 12 columns

Unsupervised Learning

Summary: Trees and Forests

- Trees give interpretable, nonlinear prediction rules
- Deep trees have low bias, high variance
- Random forests are a way of combining trees
- Want: Different trees should capture different aspects of the data
- How: Grow each tree on a random (bootstrap) sample of the data and choosing from a random set of questions at each split

Unsupervised Learning

Supervised learning is about being able to predict out Y from 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—What are we trying to learn?

Unsupervised Learning

Supervised learning is about being able to predict out Y from predictors X_1, X_2, \dots, X_p

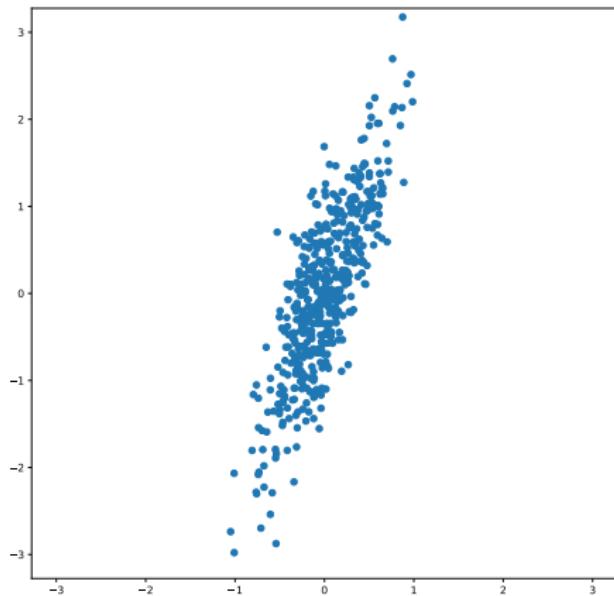
Unsupervised learning deals with data that do not have labels Y

We are not trying to predict anything—What are we trying to learn?

- Are there interesting ways to visualize/summarize the data?
- Are there natural subgroups in the data?
- What are the important types of variation in the data?

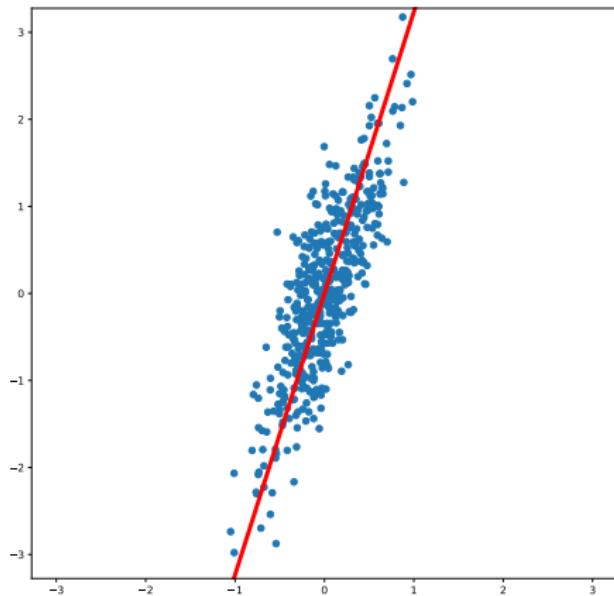
Principal Component Analysis (PCA)

PCA finds the directions of greatest variability in the data.



Principal Component Analysis (PCA)

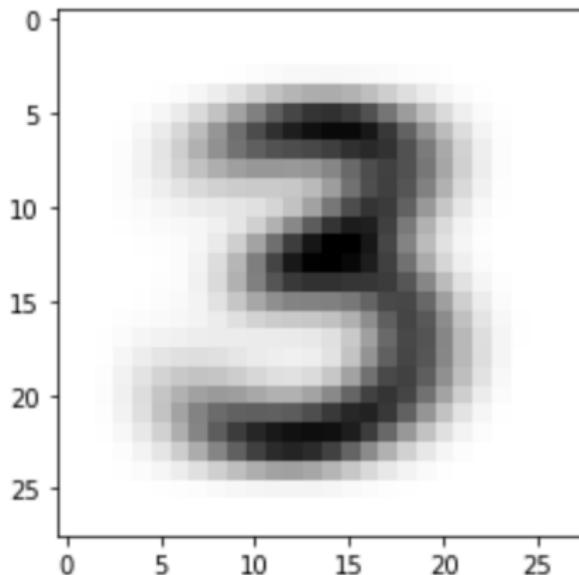
PCA finds the directions of greatest variability in the data.



Handwritten Digits (3s)

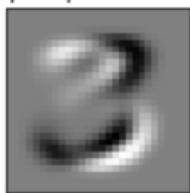
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3
3 3 3 3 3 3 3

Handwritten Digits (3s) – Average

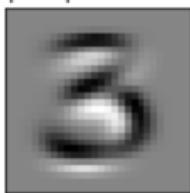


Handwritten Digits (3s) – Principal vectors

principal vector 1



principal vector 2



principal vector 3



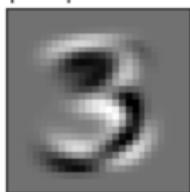
principal vector 4



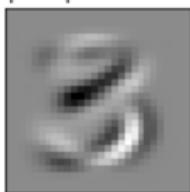
principal vector 5



principal vector 6



principal vector 7



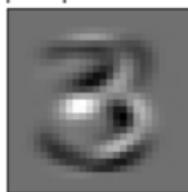
principal vector 8



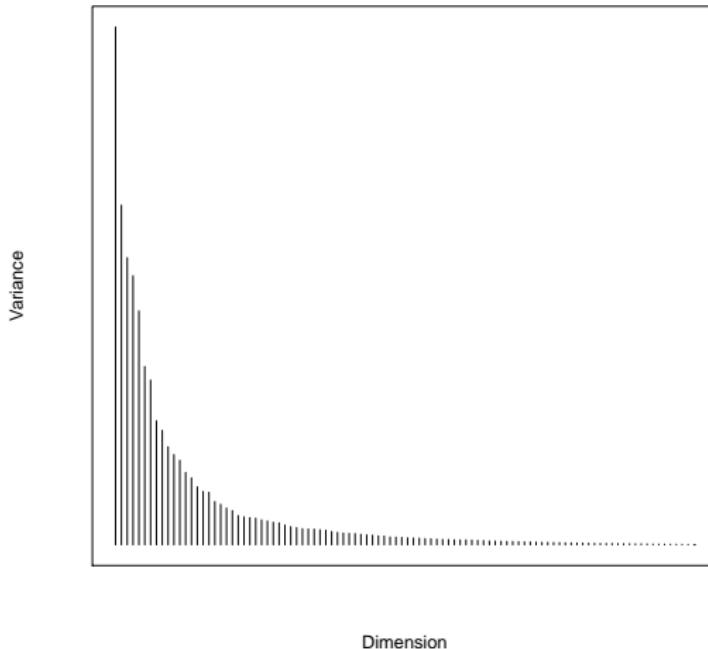
principal vector 9



principal vector 10



Handwritten Digits (3s) – PCA variance



Handwritten Digits (3s)

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

Approximation

The way to read this is:

$$\text{approximation}(x) = \bar{x} + \lambda_1(x) \cdot v_1 + \lambda_2(x) \cdot v_2$$

where

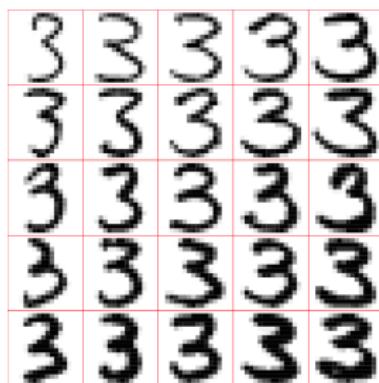
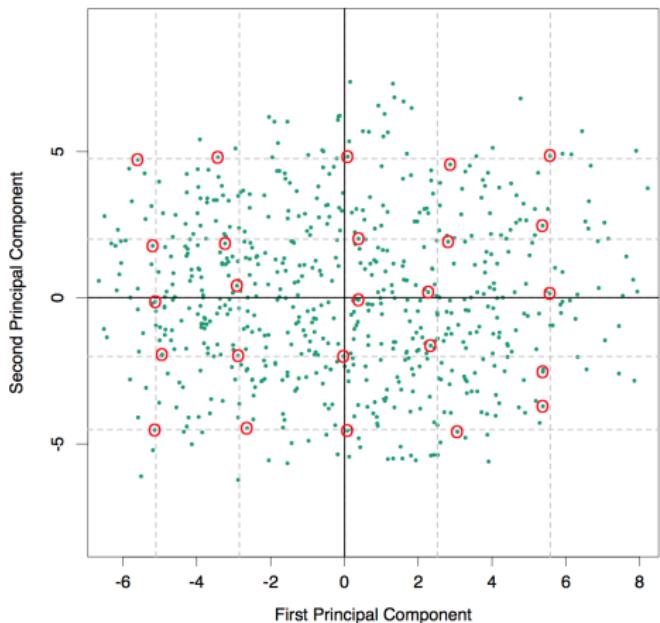
\bar{x} : average 3

v_i : the i -th principal vector

$\lambda_i(x)$: the i -th principal component for image x

Handwritten Digits (3s) – Top 2 components

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



Handwritten Digits (3s) – PCA reconstruction



Face data: LFW

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



Societal bias in image databases

Original Research | [Open Access](#) | [Published: 27 October 2021](#)

Bias, awareness, and ignorance in deep-learning-based face recognition

[Samuel Wehrli](#) , [Corinna Hertweck](#), [Mohammadreza Amirian](#), [Stefan Glüge](#) & [Thilo Stadelmann](#)

[AI and Ethics](#) 2, 509–522 (2022) | [Cite this article](#)

2187 Accesses | 3 Altmetric | [Metrics](#)

Abstract

Face Recognition (FR) is increasingly influencing our lives: we use it to unlock our phones; police uses it to identify suspects. Two main concerns are associated with this increase in facial recognition: (1) the fact that these systems are typically less accurate for marginalized groups, which can be described as “bias”, and (2) the increased surveillance through these systems.

Our paper is concerned with the first issue. Specifically, we explore an intuitive technique for reducing this bias, namely “blinding” models to sensitive features, such as gender or race, and show why this cannot be equated with reducing bias. Even when not designed for this task, facial recognition models can deduce sensitive features, such as gender or race, from pictures

Eigenfaces (principal vectors)



Let's go to the notebook

```
pca = PCA(num_components).fit(cimages)
principal_vectors = pca.components_
principal_vectors = principal_vectors.reshape((num_components, height, width))
pcs = pca.fit_transform(cimages)
capprox = pca.inverse_transform(pcs)
labels = ['principal vector %d' % (i+1) for i in np.arange(num_components)]
plot_images(principal_vectors, labels, height, width, int(num_components/5.), 5)
ratio = pca.explained_variance_ratio_.sum()
print('Variance explained by first %d principal vectors: %.2f%%' % (num_components, ratio*100))
```

Variance explained by first 25 principal vectors: 72.46%

principal vector 1 principal vector 2 principal vector 3 principal vector 4 principal vector 5



principal vector 6 principal vector 7 principal vector 8 principal vector 9 principal vector 10



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.

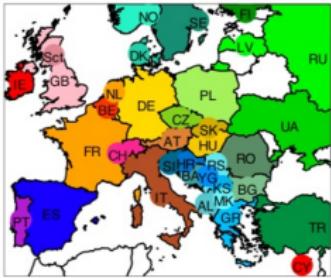
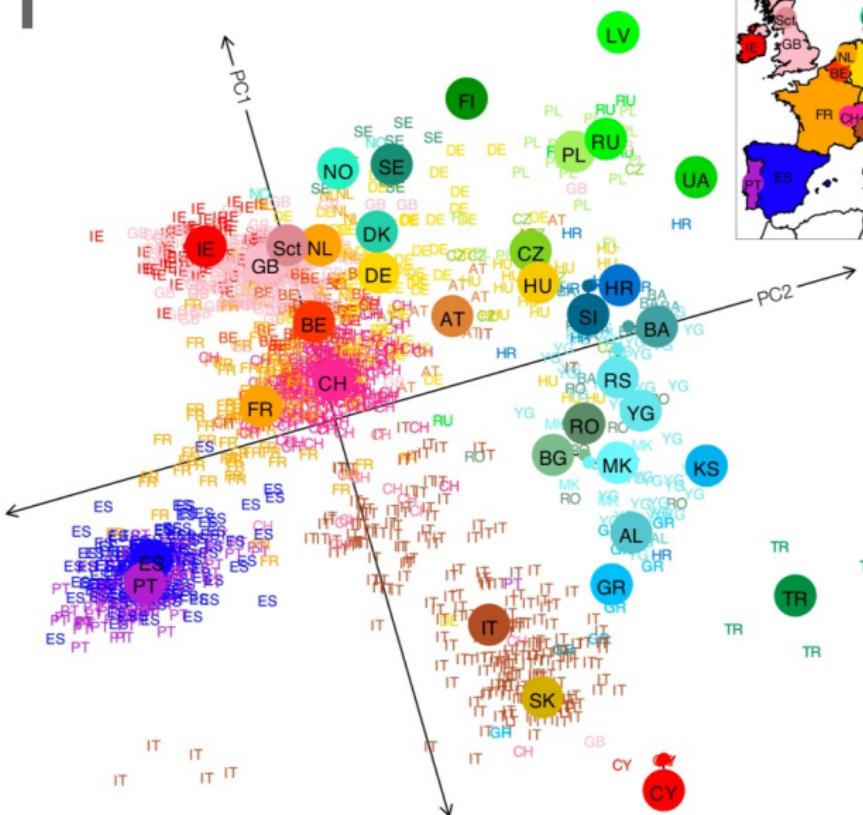
See other articles in PMC that [cite](#) the published article.

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





Examples

A couple of other interesting examples of PCA follow (without going into any detail)

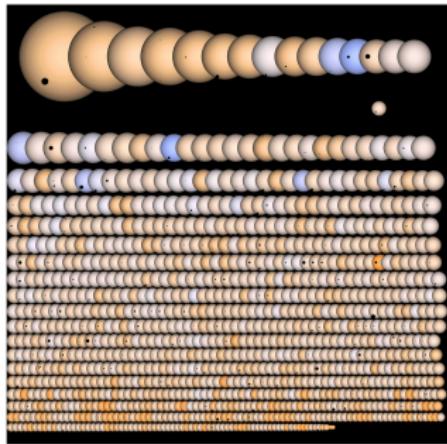
Robust PCA

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



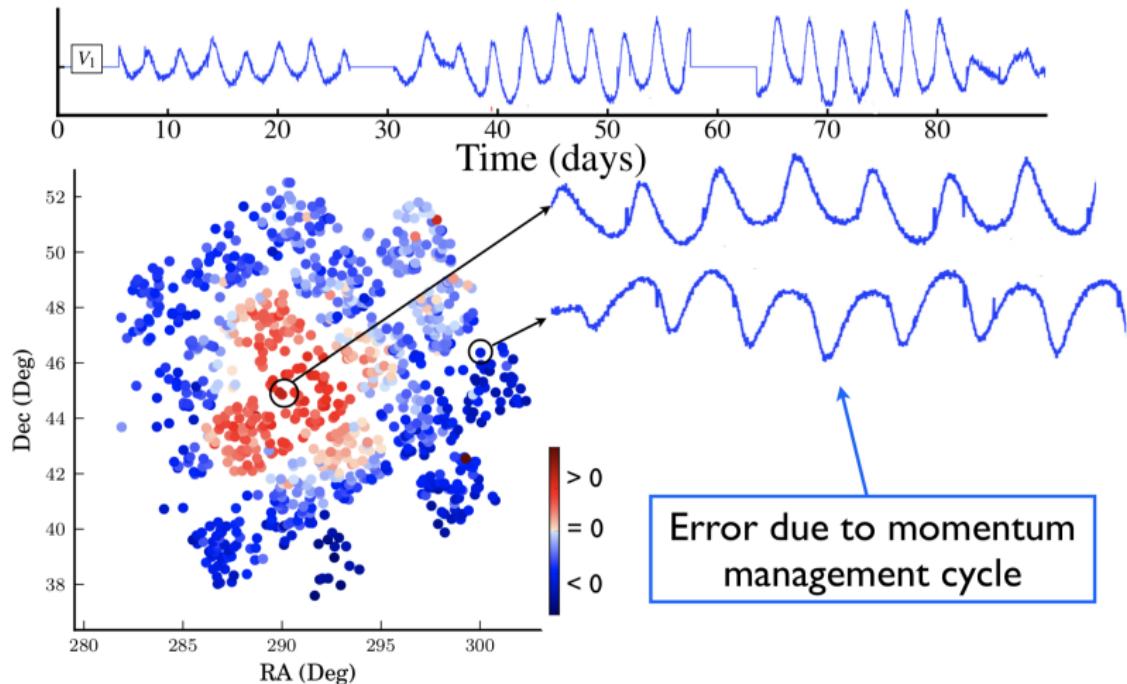
<https://www.youtube.com/watch?v=BTrbow8u4Cw>

Kepler telescope

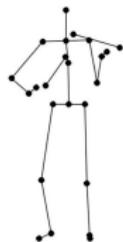


Kepler telescope

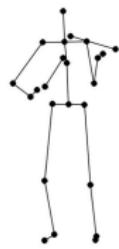
Identification of correlated errors using robust PCA



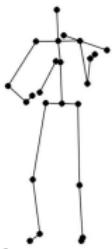
RESULTS: PC VISUALISATION, TECHNICAL VS. NON-TECHNICAL MOTION



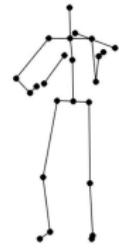
Original



PC1 (45% of variance accounted)
Side-to-side swaying (non-technical)



PC2 (25% of variance accounted)
Right arm string crossing (technical)



PC3 (15% of variance accounted)
Left-right rotation (non-technical)

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: $x_i \mapsto x_i - \frac{1}{n} \sum_{j=1}^n x_j = x_i - \bar{x}$
- ② Compute the $d \times d$ sample covariance $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$. Note that

$$\frac{1}{n} \sum_i (x_{ij} - \bar{x})^2$$

is the sample variance of j th coordinate of data.

- ③ Find the first k eigenvectors of S ,

$$v_1, \dots, v_k \in \mathbb{R}^d, \quad S v_j = \gamma_j v_j$$

- ④ Project the data onto those k vectors:

$$x_i \approx \bar{x} + (v_1^T x_i) v_1 + \dots + (v_k^T x_i) v_k$$

Let's revisit the notebook

```
pca = PCA(num_components).fit(cimages)
principal_vectors = pca.components_
principal_vectors = principal_vectors.reshape((num_components, height, width))
pcs = pca.fit_transform(cimages)
capprox = pca.inverse_transform(pcs)
labels = ['principal vector %d' % (i+1) for i in np.arange(num_components)]
plot_images(principal_vectors, labels, height, width, int(num_components/5.), 5)
ratio = pca.explained_variance_ratio_.sum()
print('Variance explained by first %d principal vectors: %.2f%%' % (num_components, ratio*100))
```

Variance explained by first 25 principal vectors: 72.46%

principal vector 1 principal vector 2 principal vector 3 principal vector 4 principal vector 5



principal vector 6 principal vector 7 principal vector 8 principal vector 9 principal vector 10



Dimension reduction—another demo

- Flower Power: PCA and classification (30 points)



In this problem you will carry out principal components analysis and classification on the iris data. The task will be to reduce the dimension from four to two using PCA, and then to train logistic regression models on the projected data.

+ Code + Text

```
[ ] import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression

import matplotlib.pyplot as plt
%matplotlib inline
```

PCA: Summary

- PCA is an unsupervised method
- Finds directions of greatest variation in the data
- The directions are called the *principal vectors*; the weightings on the vectors are called the *principal components*
- The first few vectors may be interpretable
- Orthogonality makes interpretation difficult for the higher components
- Can be used for visualization or dimensionality reduction