S&DS 265 / 565

Introductory Machine Learning

Unsupervised Learning: PCA

Tuesday, October 12

Yale

Checkpoint

- Assn 3 due Thursday; Assn 4 out
- Quiz 2 available from 12:01pm today for 24 hours
- Midterm one week from today, in class
- Practice midterm posted on Canvas
- Review sessions scheduled
- Questions?

Unsupervised Learning

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

Unsupervised learning deals with data that do not have labels Y

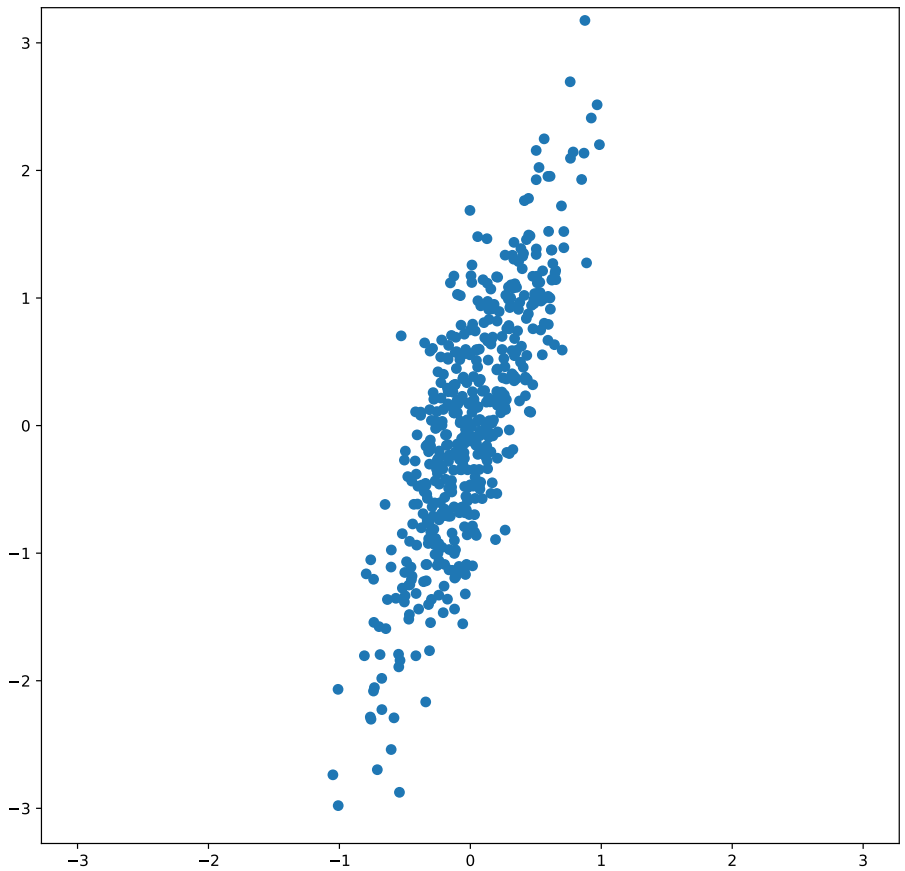
We are not trying to predict anything. So what else might we hope to do?

Consider:

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

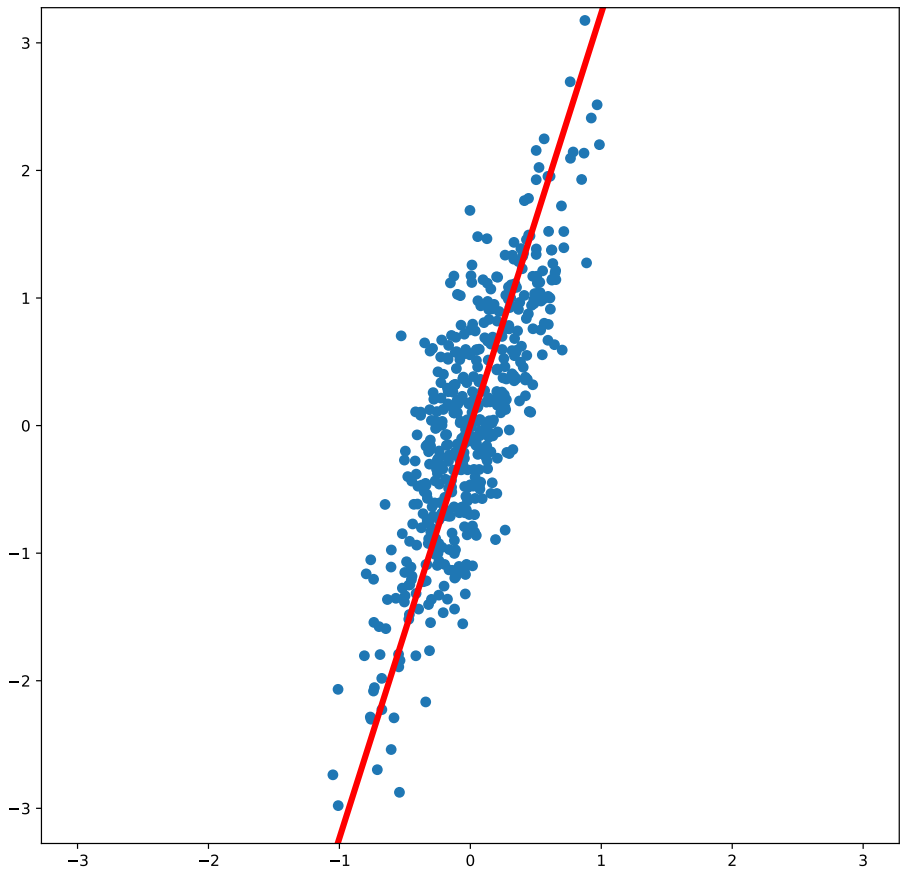
Principal Component Analysis (PCA)

PCA finds the directions of greatest variability in the data.

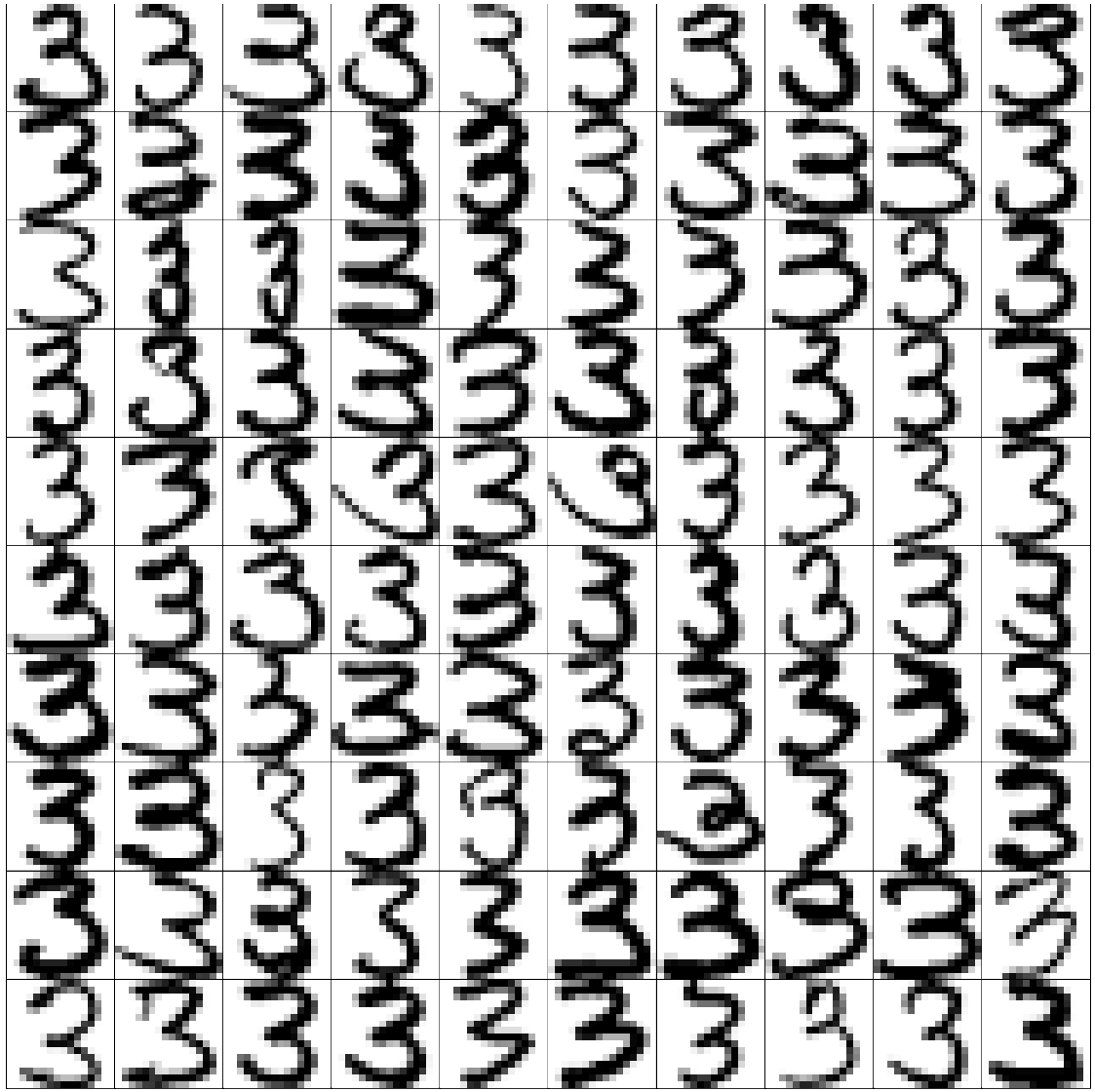


Principal Component Analysis (PCA)

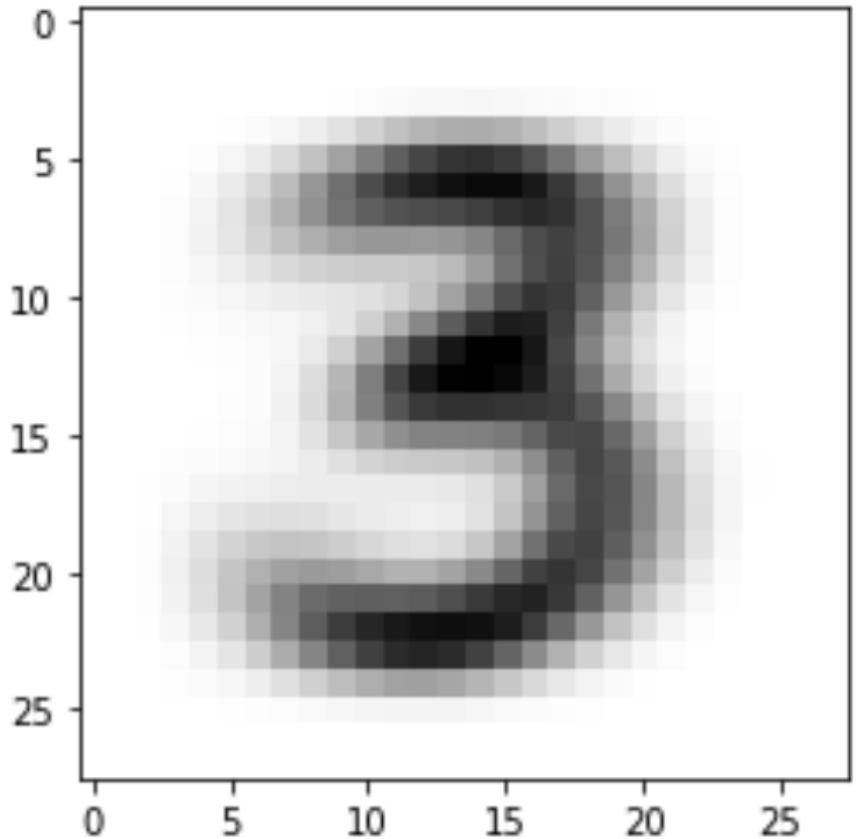
PCA finds the directions of greatest variability in the data.



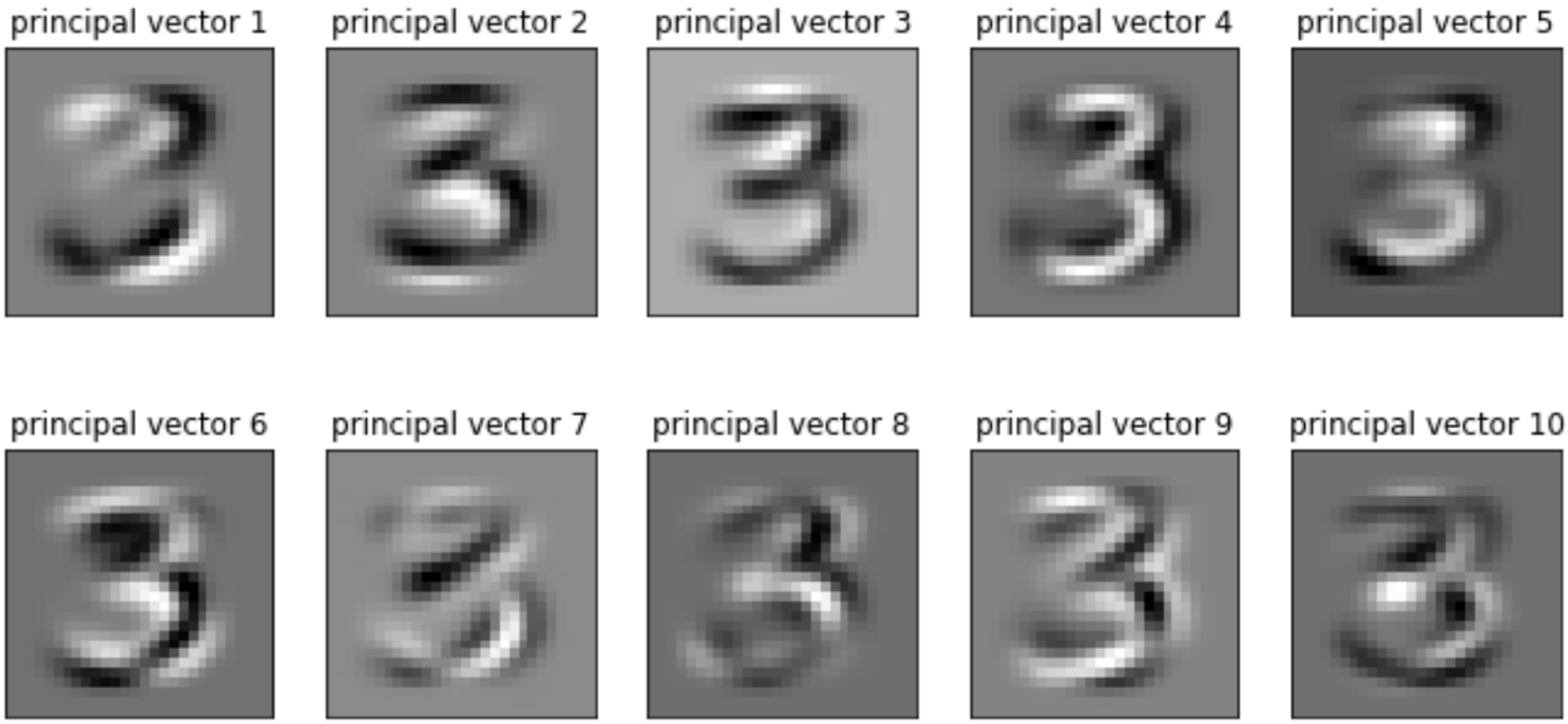
Handwritten Digits (3s)



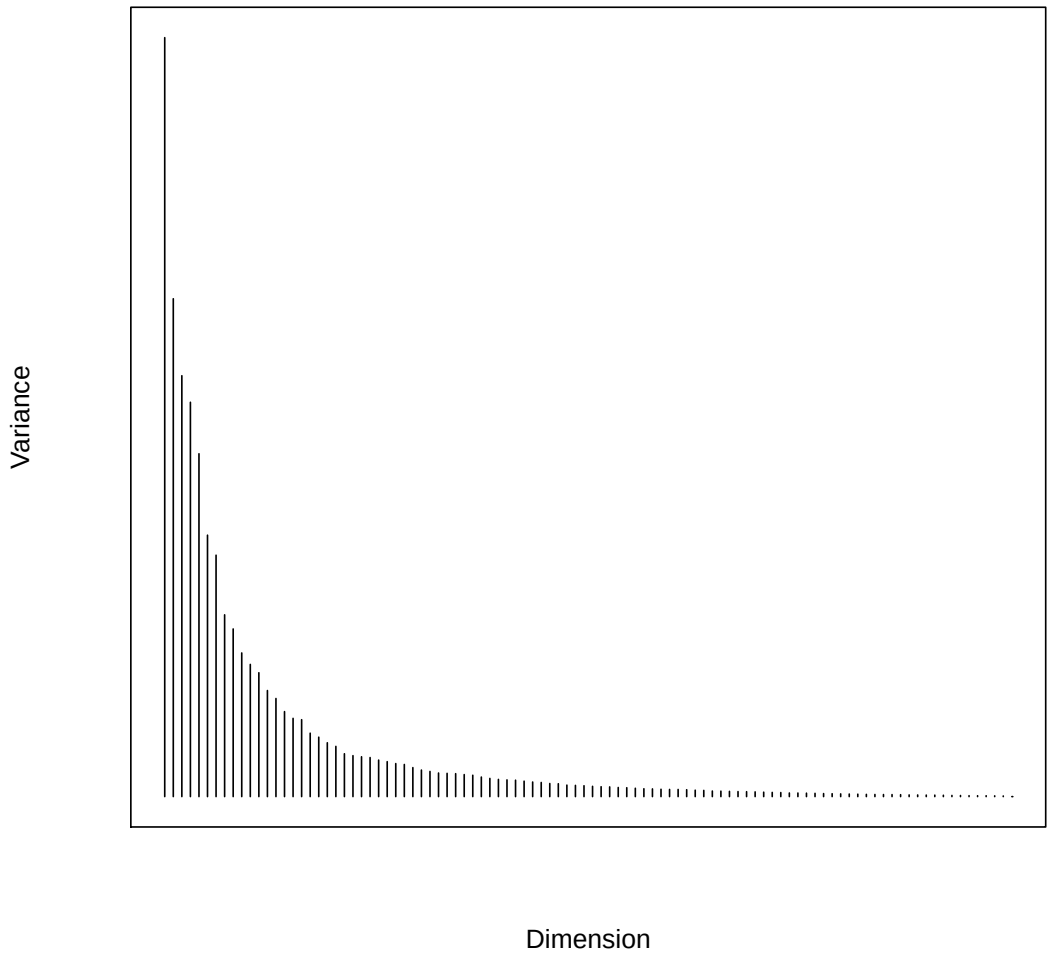
Handwritten Digits (3s) – Average



Handwritten Digits (3s) – Principal vectors



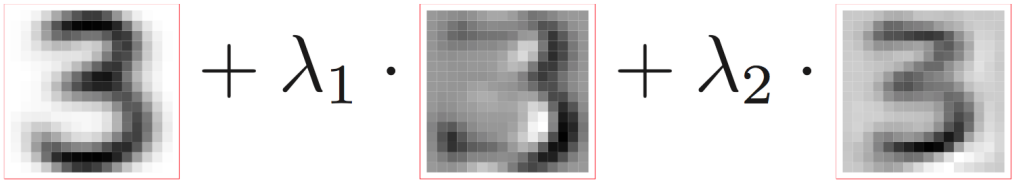
Handwritten Digits (3s) – PCA variance



Handwritten Digits (3s)

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

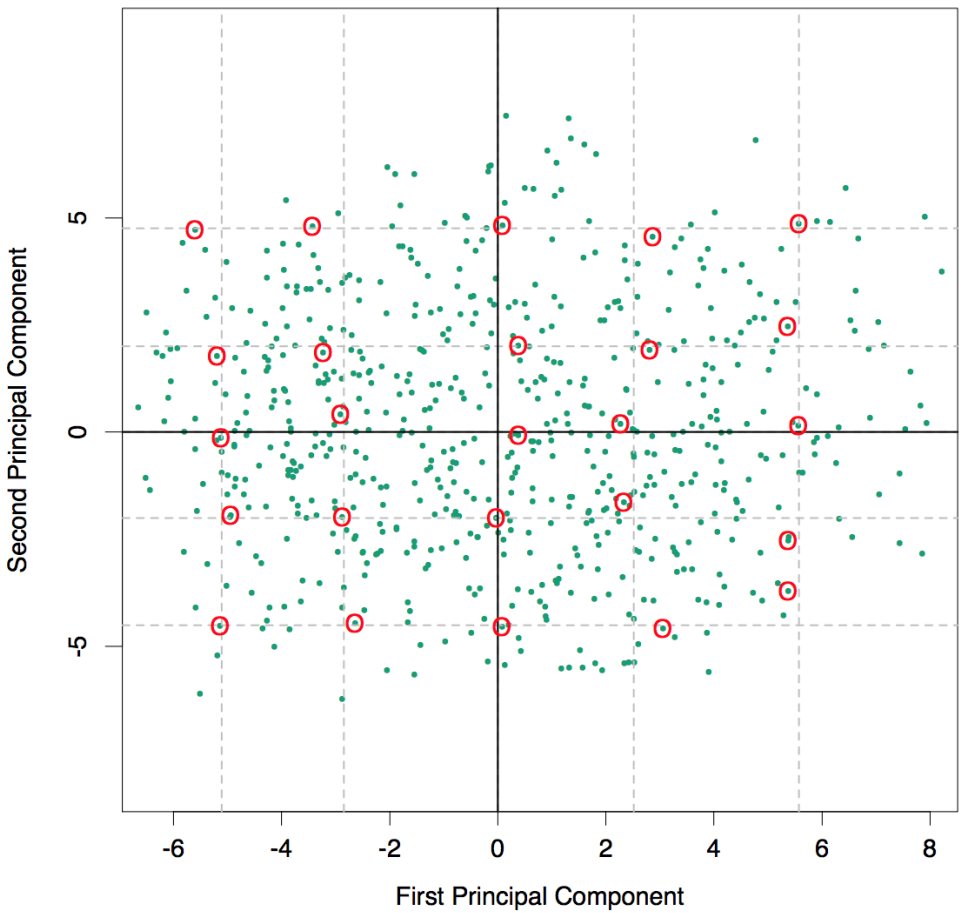
$$= \bar{x} + \lambda_1 \cdot v_1 + \lambda_2 \cdot v_2.$$

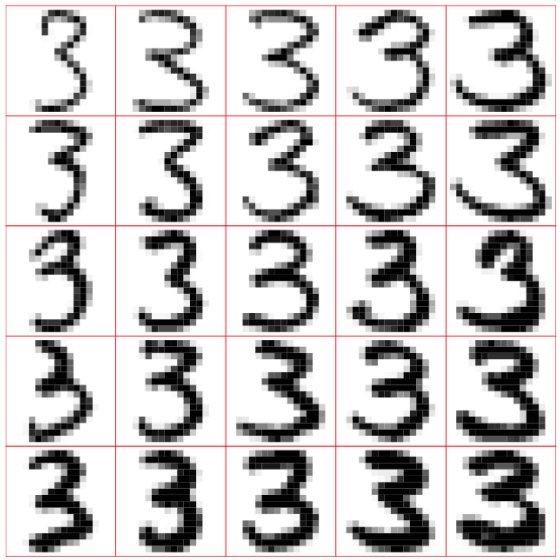


Handwritten Digits (3s) – Top 2 components

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

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



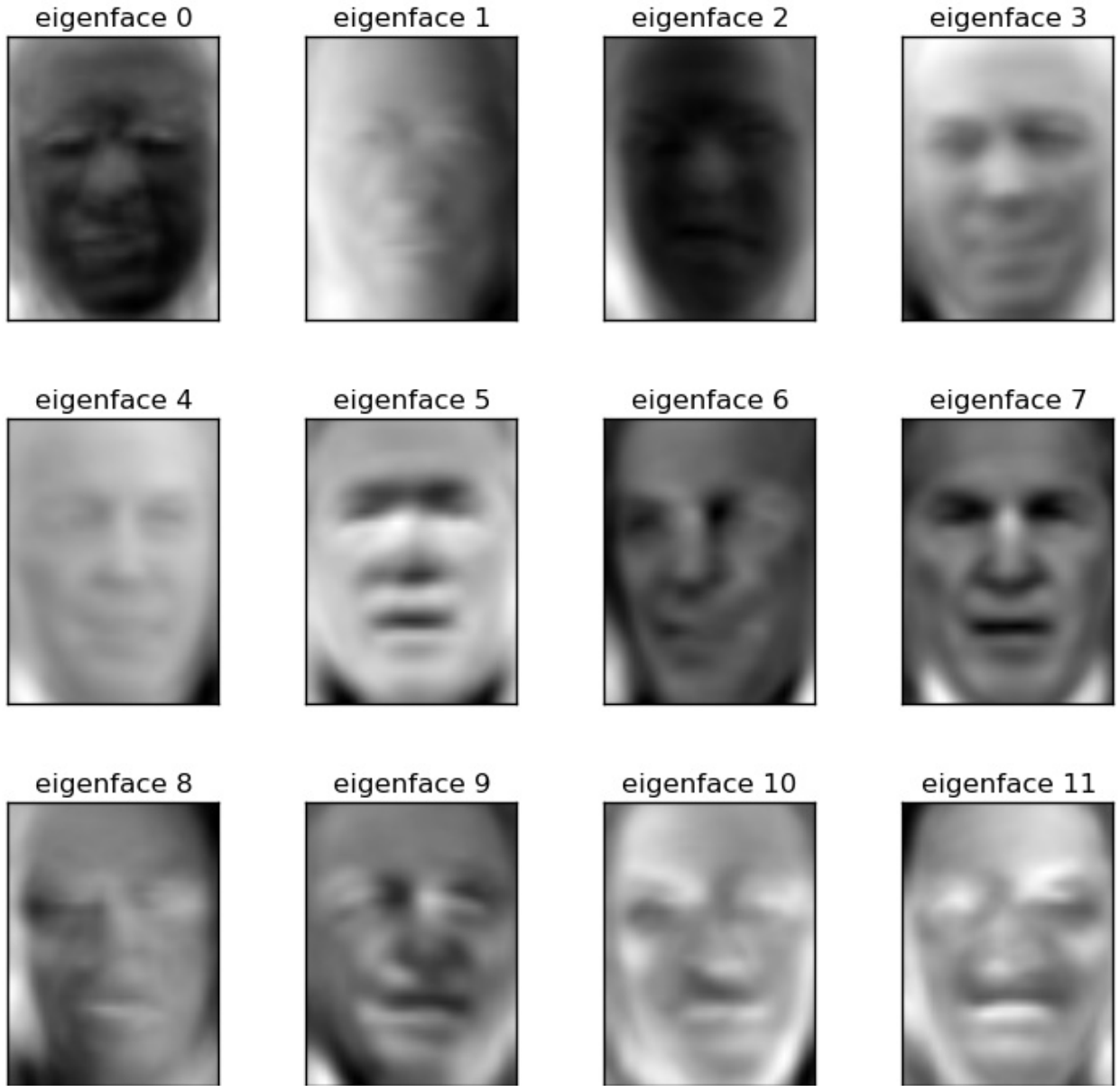


Handwritten Digits (3s) – PCA reconstruction

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 (principal vectors)



Published in final edited form as:
Nature. 2008 Nov 6; 456(7218): 98–101.
Published online 2008 Aug 31. doi: 10.1038/nature07331
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 ▶

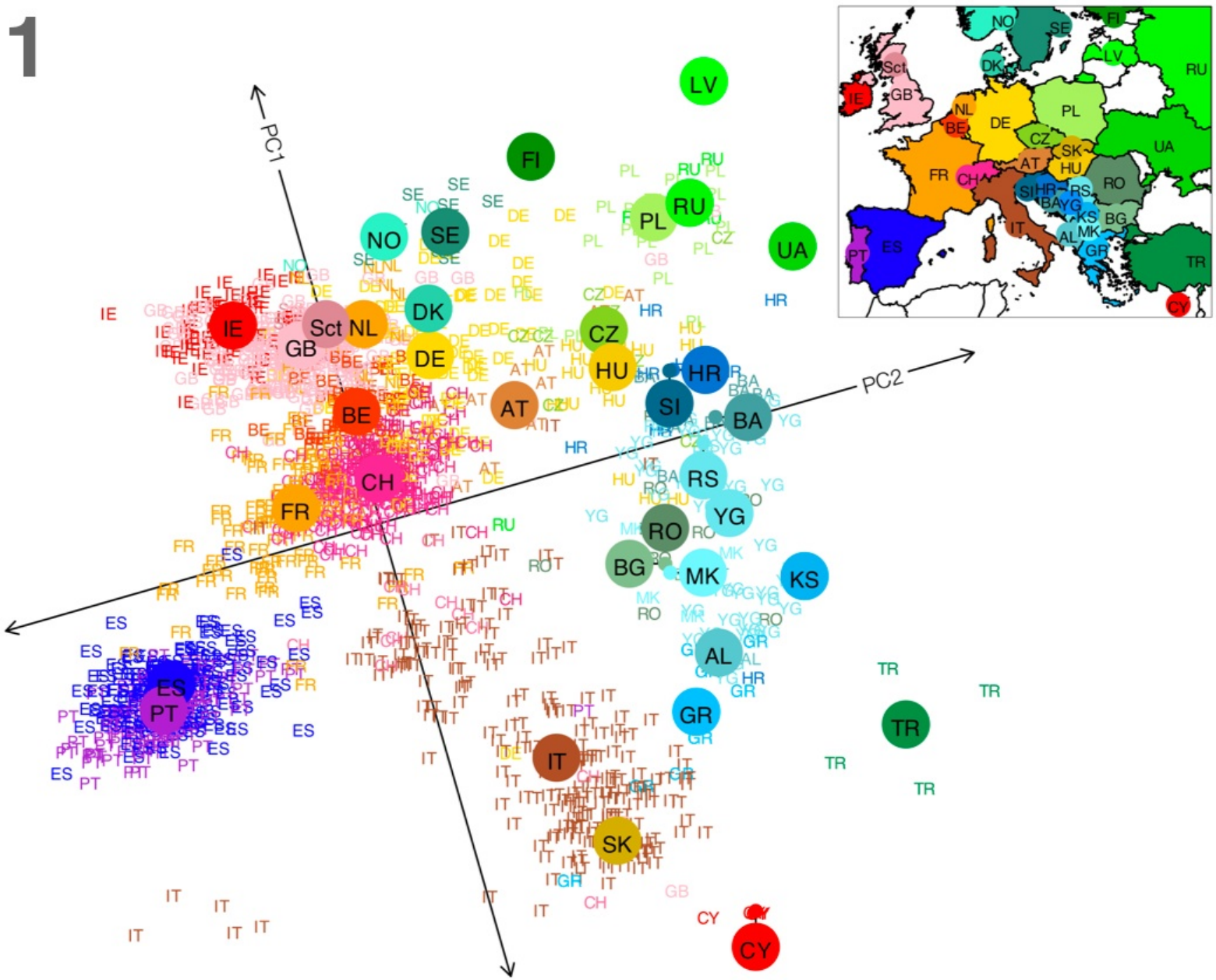
Abstract

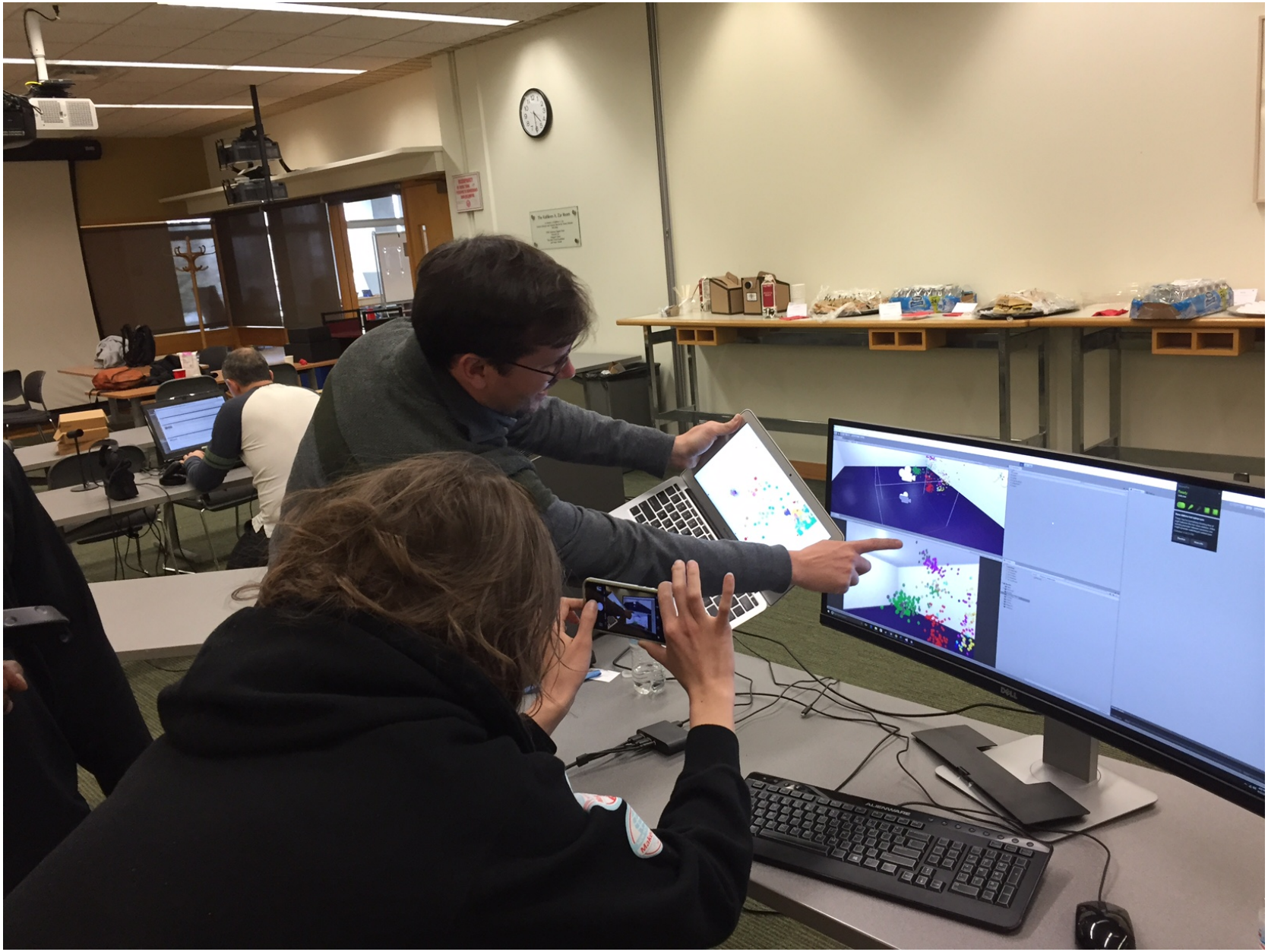
Go to:

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.

https://ghr.nlm.nih.gov/primer/genomicresearch/snp

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

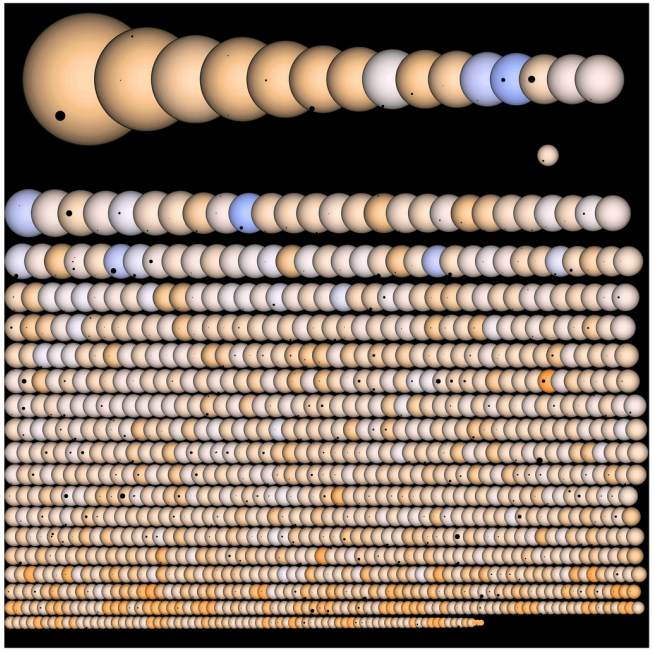

DELL
ALIENWARE

Robust PCA

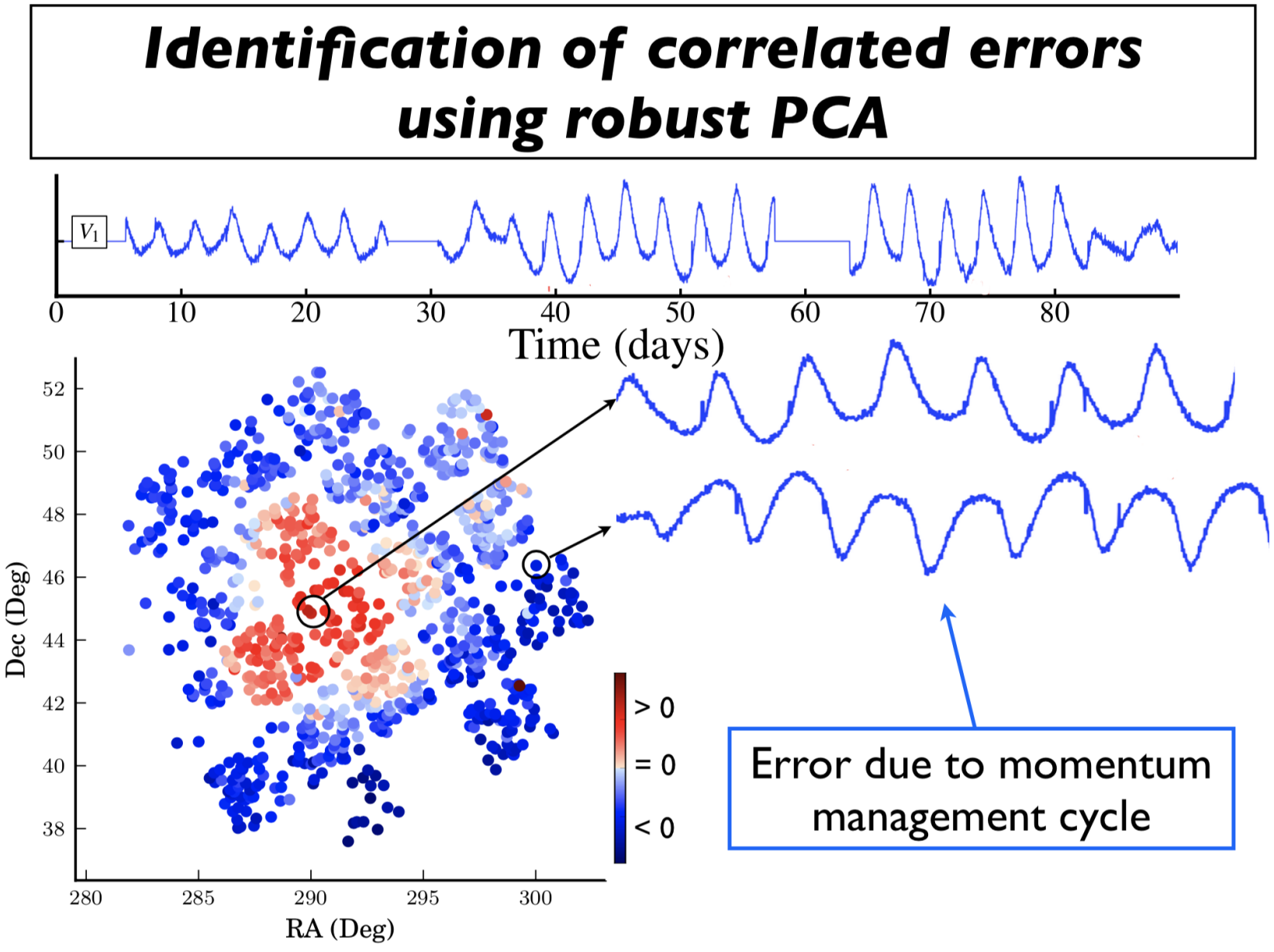
Robust PCA (low rank plus sparse) can be used for background subtraction in video.

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

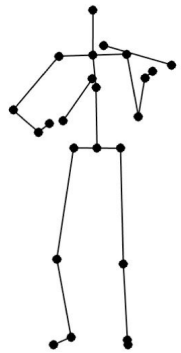
Kepler telescope



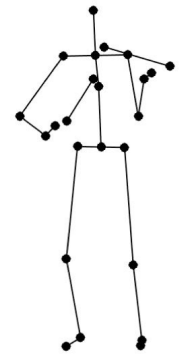
Kepler telescope



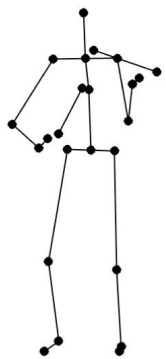
RESULTS: PC VISUALISATION, TECHNICAL VS. NON-TECHNICAL MOTION



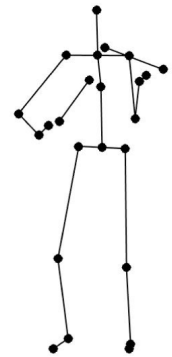
Original



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)

Aditya Chander

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: $x_i \mapsto x_i - \frac{1}{n}\sum_{j=1}^{n} x_j = x_i - \overline{x}$

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

$$\frac{1}{n}\sum_{i} \left(x_{ij} - \overline{x}\right)^2$$

   is the sample variance of jth coordinate of data.

3. Find the first k eigenvectors of S,

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

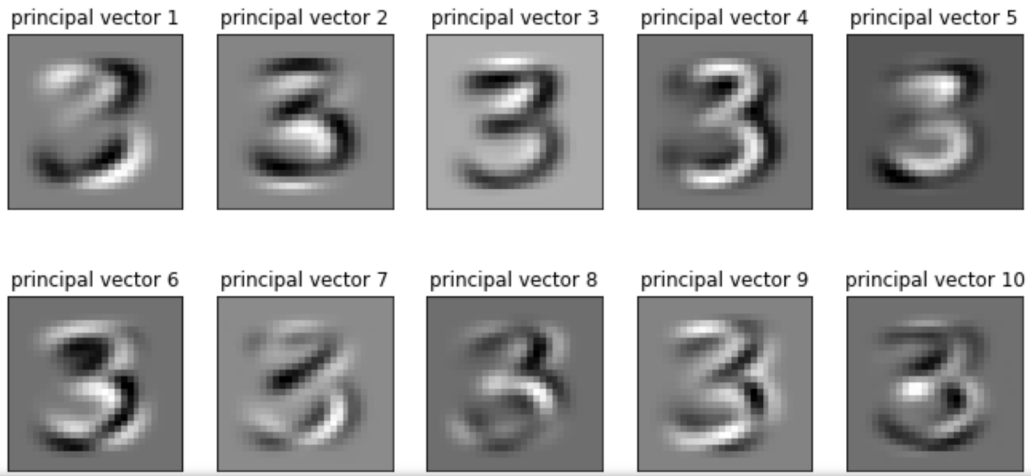
4. Project the data onto those k vectors:

$$x_i \mapsto \overline{x} + (\phi_1^T x_i)\phi_1 + \ldots + (\phi_k^T x_i)\phi_k$$

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%
```



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