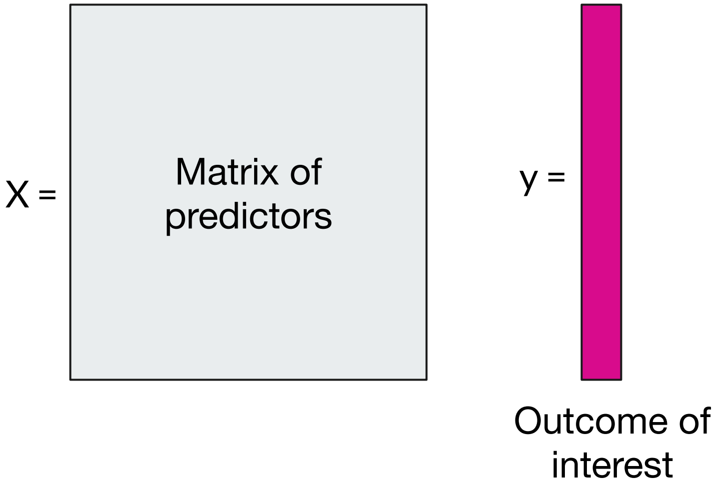
Chapter 20: Principal Components Analysis

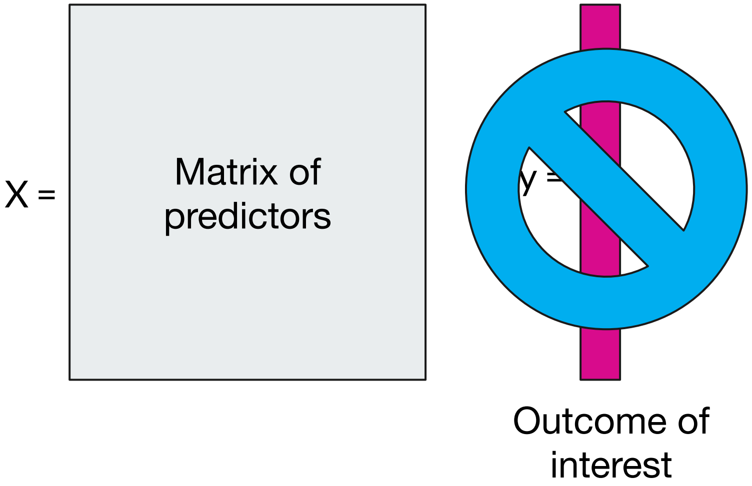
Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 30 min, max) while following along in the notes

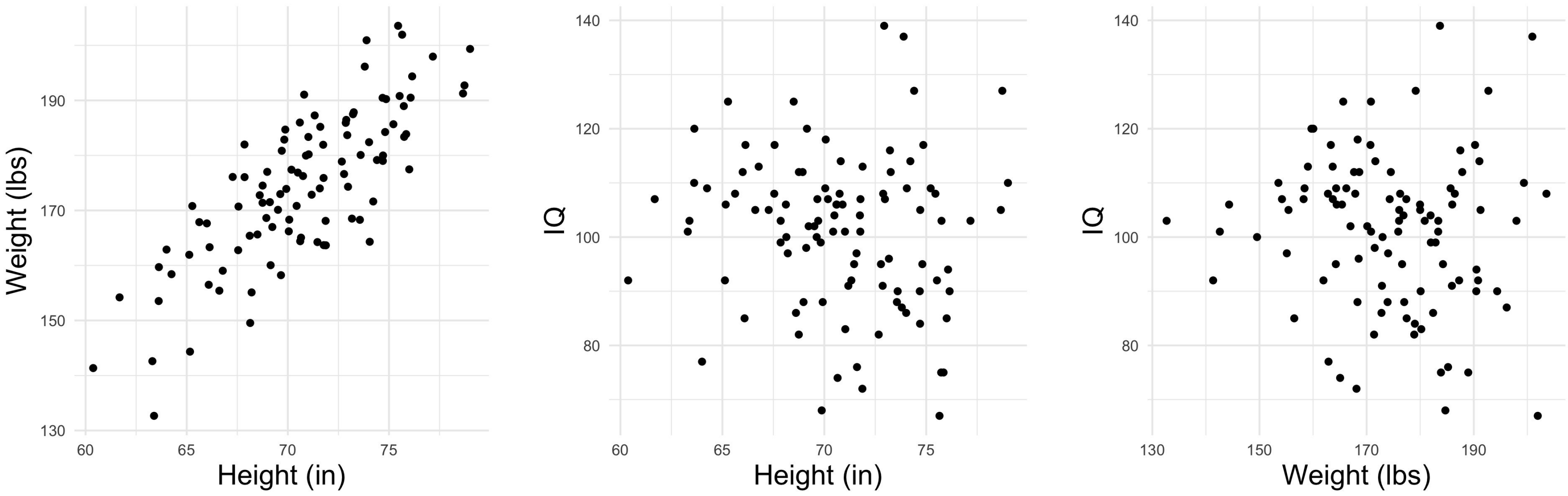


Supervised Learning

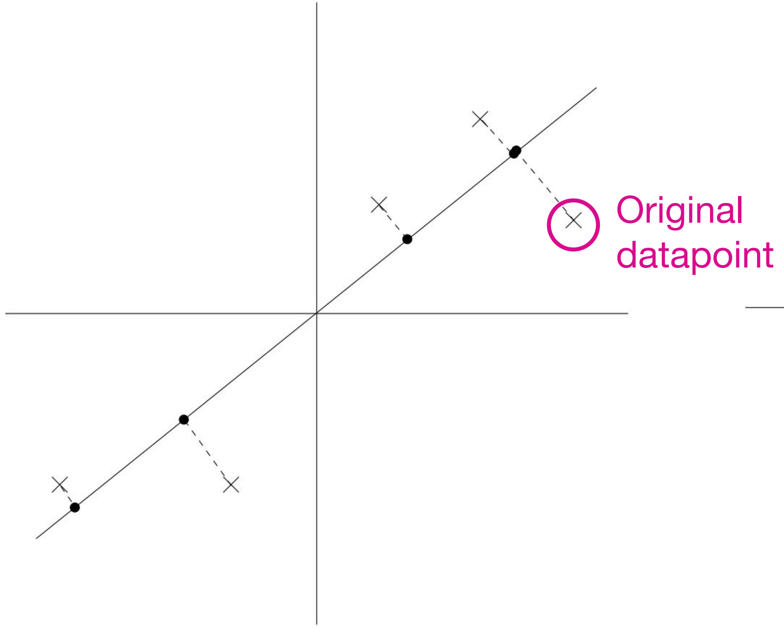


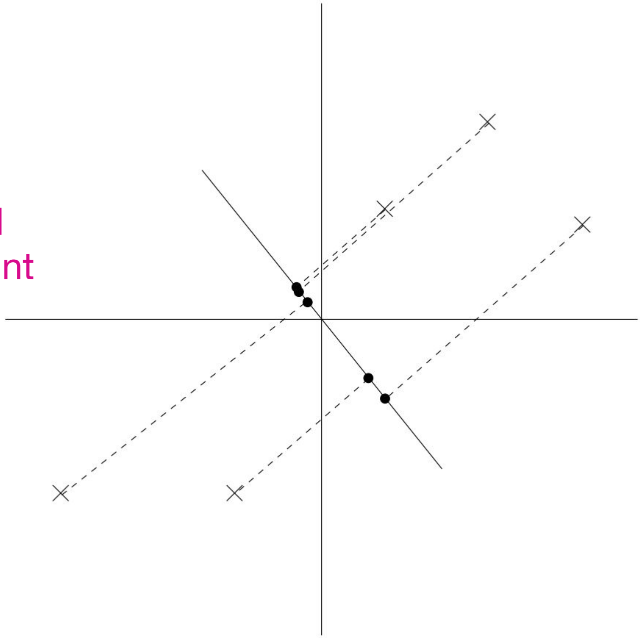
Unsupervised Learning

Example: Height, Weight, IQ



Principal Components





The first PC points along the direction of **maximum variance** in the dataset.

The first PC points along the direction of **minimum reconstruction error.**

All PCs must be **orthogonal.**

The **correlation matrix** for these data looks like this:

	height	weight	iq
height	1.000	0.790	-0.103
weight	0.790	1.000	-0.078
iq	-0.103	-0.078	1.000

Question 20.2

Looking at the correlation matrix, which features are most tightly correlated? What does this imply about the direction of the first principal component, PC1?

Question 20.3

What would happen to the principal components if you didn't center and scale the data?

Question 20.4

Why do you think the interpretation of the principal components becomes more difficult if you have features measured on lots of different scales (e.g., some categorical, some numeric/roughly normal, some numeric/highly skewed)?

```
> p$sdev
[1] 1.3454074 0.9899765 0.4580672

> p$rotation
              PC1         PC2         PC3
height  0.6996236 -0.09446119 -0.70823998
weight  0.6971414 -0.12698944  0.70559726
iq     -0.1565906 -0.98739595 -0.02299201

> p$center
   height    weight        iq
 70.74468 174.43346 101.12000

> p$scale
   height    weight        iq
 3.914989 13.853272 14.183701
```

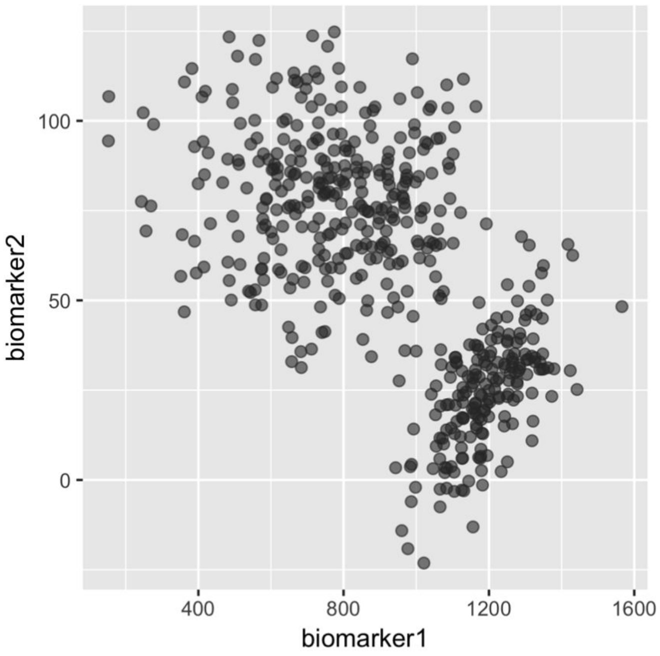
```
> p$x
              PC1          PC2          PC3
 [1,] -0.189268854  1.976180297 -0.416704298
 [2,]  1.794925031  0.710160037  0.575478783
 [3,] -0.404617344  0.283418147  0.151536071
 [4,] -0.471675025 -0.710146894  0.347006397
 [5,] -1.073877096 -0.234684815 -0.638243675
 [6,] -0.654663317  0.043481120 -0.107082277
 [7,] -1.705841441  0.368990046 -0.794047654
 [8,]  2.311353981 -0.501565781  0.033112309
 [9,] -0.006387264 -0.535422493 -1.126177513
[10,] -1.557136911  1.399023167 -0.045281059
[11,]  2.632974868 -1.046037442 -0.239934797
[12,] -0.978337630 -0.193476727  0.007814938
[13,] -0.625980903 -0.456274882 -0.305349616
(continued)
```

Question 20.5

Describe/draw the directions of the three principal component vectors, PC1, PC2, and PC3, in the coordinate system of the original predictors, height, weight, and IQ.

Question 20.6

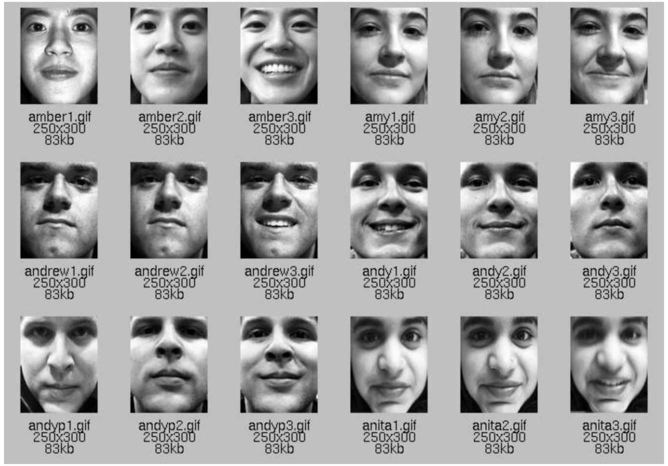
Here is a picture of the flow cytometry dataset we first encountered in Chapter 19.

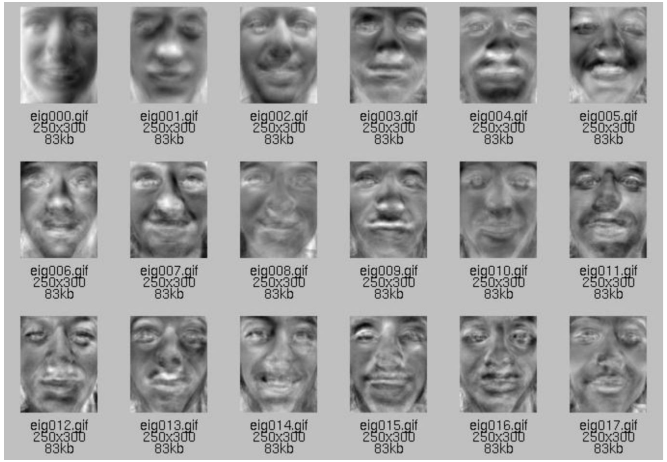


What would PC1 and PC2 look like for this dataset? (Why is there no PC3?) How could PCA help you separate the two clusters?

Application 1: Eigenfaces

One of the earliest applications of PCA to computer vision was this project, which used PCA to find characteristic “modes” by which human faces vary.





https://www.clear.rice.edu/elec301/Projects99/faces/images.html

Application 1: Eigenfaces

Question 20.7

What is X for the eigenfaces problem? What are the principal components? How could you use the principal components to match a new face to an existing database of faces?

Application 1: Eigenfaces

$$X = \begin{array}{cc} & p \\ n & \begin{bmatrix} \rightarrow \text{pixels} \\ \\ \downarrow \text{faces} \end{bmatrix} \end{array}$$

n = number of images
p = number of pixels per image

→ PCs (eigenvectors) are combinations of pixels that tend to vary together. (start fuzzy, get more specific)

→ By getting the location of a new face in PCA space, we can more easily match it to existing faces.

Application 2: Image Compression

You can also use PCA to compress an image.

Tile little boxes across the image and do PCA on matrix of these sub-images.

(a) 1 principal component
(b) 5 principal component
(c) 9 principal component
(d) 13 principal component
(e) 17 principal component
(f) 21 principal component
(g) 25 principal component
(h) 29 principal component

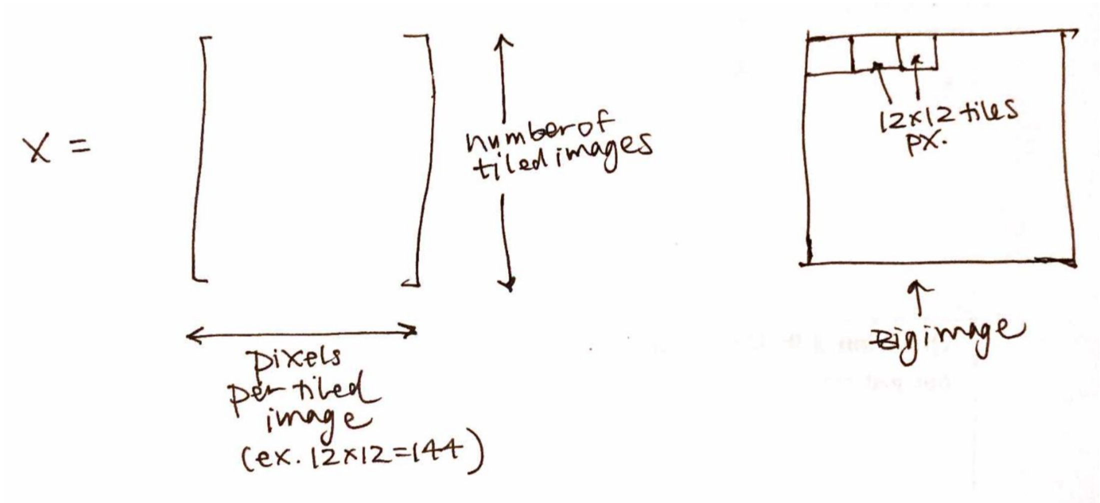
https://www.projectrhea.org/rhea/index.php/PCA_Theory_Examples

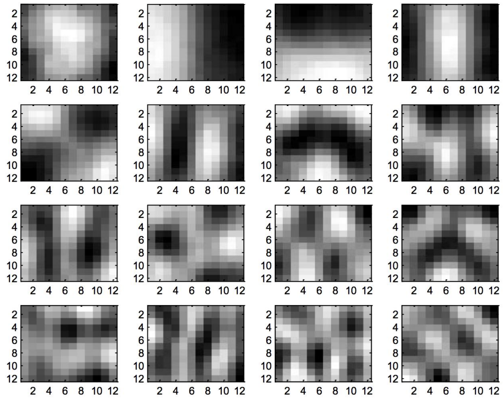
Application 2: Image Compression

Question 20.8

What is X for the image compression problem? What are the principal components? How does using PCA help compress the image?

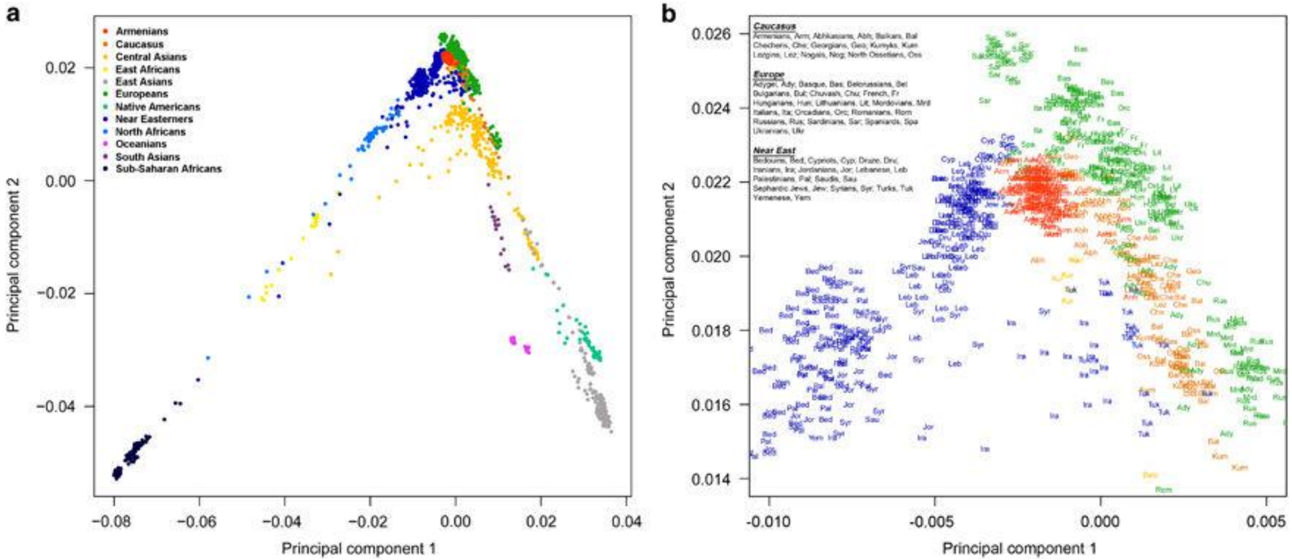
Application 2: Image Compression





http://www.cs.cmu.edu/~bapoczos/Classes/ML10715_2015Fall/slides/PCA.pdf

Application 3: Genetic Ancestry



From European Journal of Human Genetics (2016) 24, 931-936 (2016).

Application 3: Genetic Ancestry

Question 20.9

What is X for the genetic ancestry problem? What are the principal components? How were the principal components used to produce the figure above?

Application 3: Genetic Ancestry

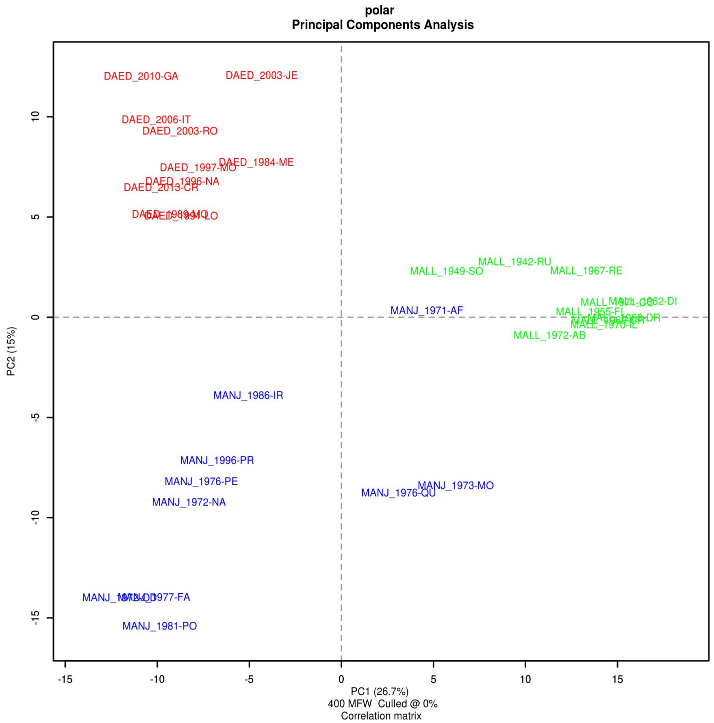
$$X = \begin{array}{cc} & \text{SNPS} \\ \text{people} & \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} \end{array}$$

Each element is 0, 1, or 2.

→ Get locations of all people in PC space (PCs 1 and 2 only).

→ Plot on a graph with color representing ethnicity.

Application 4: Topic Modeling (or) Word Similarity



https://dragonfly.hypotheses.org/530

Application 4: Topic Modeling, Word Similarity

Question 20.10

What is X for the topic modeling problem? For the latent semantic indexing problem? What do the principal components represent?

Application 4: Topic Modeling, Word Similarity

$$\underset{\text{(topics)}}{X} = \begin{array}{cc} & \\ \text{docs} & \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \\ & \text{words} \end{array}$$

PCs: combinations of words that vary together.

$$\underset{\text{(LSA)}}{X} = \begin{array}{cc} & \\ \text{words} & \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \\ & \text{docs} \end{array}$$

PCs: combinations of docs that vary together.

Question 20.11

Think of 2-3 different unsupervised learning problems from biology or medicine where PCA makes sense, conceptually at least, for modeling the data. How would you set up the data matrix in each case? What would the principal components correspond to in the data?

Deep connection to multivariate Gaussian

The eigenvectors of the covariance matrix of the Gaussian define the major and minor axes of an ellipse.