Lecture #15: Principal Component Analysis

Today: How do you extract structure from high-dimensional data?

Setup: You have a matrix of datapoints

$$\underset{n \times p}{A} = \begin{bmatrix} a_1 & a_2 & \dots & a_p \\ & & & \end{bmatrix}$$

each data point a_i is an n-dimensional vector

Examples:

1. word counts in a document
2. genetic information for an individual
3. voting records
4. response patterns of a neuron to various stimuli

Q: Can we map it to low dimensions while approximately preserving its structure?

We will cast it as an optimization problem

(1) First compute the sample average:

$$\mu = \frac{1}{p} \sum_{i=1}^{p} a_i$$

and re-center the data

$$y_i = a_i - \mu$$

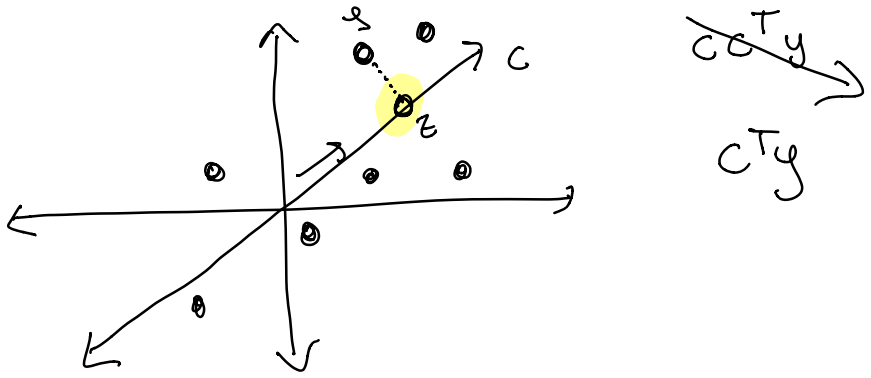
(2) Now compute the covariance

$$S = \frac{1}{p} \sum_{i=1}^{p} y_i y_i^T$$

If you haven't done statistics before this seems strange, but let's demystify it

Q2: What kind of information does S contain?

Thought Experiment: What if we project our data onto a line in some direction c



Then we might care about how "spread out" the projections are along this direction

Fact: The average is zero

$$\frac{1}{p}\sum_{i=1}^{p} z_i = \frac{1}{p}\sum_{i=1}^{p} c^T y_i = c^T\left(\frac{1}{p}\sum_{i=1}^{p} q_i - \mu\right)$$
$$= 0$$

So let's measure the spread around zero

$$\frac{1}{p}\sum_{i=1}^{p} z_i^2 = \frac{1}{p}\sum_{i=1}^{p} (c^T y_i) = \frac{1}{p}\sum_{i=1}^{p} (c^T y_i) y_i^T c$$

$$= \frac{1}{p} c^T \underbrace{\left(\sum_{i=1}^{p} y_i y_i^T \right)}_{S} c$$

This is called a quadratic form (of c on S)

When we project data onto a lower dimensional space we reduce the spread

Q3: Can we find a direction that maximizes the projected spread?

i.e. we want to solve:

$$\max_{c \text{ s.t. } \|c\|=1} \quad c^T S c$$

Let's solve it through the SVD. First write

$$A = U \Sigma V^T$$

Now we can express S through A and its SVD

$$S = \frac{1}{p} A A^T = \frac{1}{p} U \Sigma V^T V \Sigma^T U^T$$

$$= \frac{1}{p} U \Sigma \Sigma^T U^T$$

Returning to the optimization problem at hand:

$$\max_{c \text{ s.t. } \|c\|=1} \frac{1}{p} c^T U \Sigma \Sigma^T U^T c$$

Let's set $b = U^T c$ in which case we get

$$\max_{b \text{ s.t. } \|b\|=1} \frac{1}{p} b^T \Sigma \Sigma^T b \qquad (*)$$

This is the same sort of change of variables as we used in last lecture, and works b/c

$$\|b\| = \|U^T c\| = \|c\| = 1$$

Q4: What is the maximum of (*)?

i.e. $\frac{1}{p} \sum_{i=1}^{p} b_i^2 \sigma_i^2$ s.t. $\sum_{i=1}^{p} b_i^2 = 1$

It is $\frac{\sigma_1^2}{p}$ and is achieved by $b_1 = 1, b_2 = b_3 \ldots = 0$

Q5: So what is the direction c that maximizes the projected spread?

Well $b = U^T c \Rightarrow Ub = c$ so we should set

$c = u_1$ i.e. the top left singular vector of A

More generally we have:

Theorem: The k dimensional space that maximizes the spread, i.e. solves

$$\max_{\text{orthonormal } c_1,\dots c_k} \frac{1}{p} \sum_{i=1}^{p} \| C^T y_i \|^2$$

$$C = [c_1 \dots c_k]$$

is achieved by setting $c_1 = u_1, c_2 = u_2 \dots$

top k left singular vectors

The intuition is that:

best k-dimensional subspace $\supseteq$ best $k-1$ dimensional subspace ...

and once you've found the best 1-dimensional subspace $\text{span}(u_1)$, you look orthogonal to it to find the next best, etc

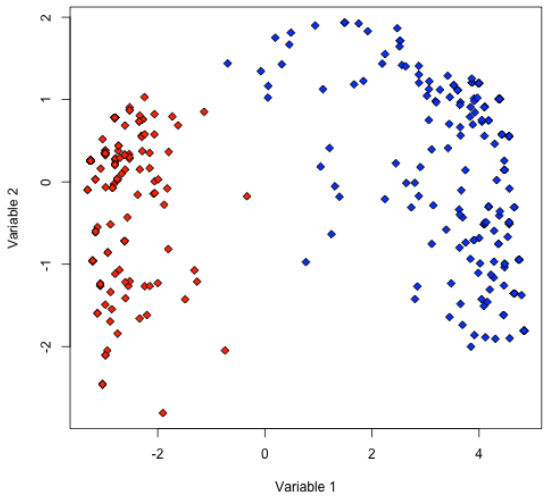
Let's circle back to our earlier examples and see what PCA discovers

Example: Each data point a_i represents a US senator and the vector is their voting record.

(104 senators voting on 172 bills)

You should ask yourself what you think will happen

Q6: what 1-dimensional projection would capture the most spread?



from Stanford STATS 202 course

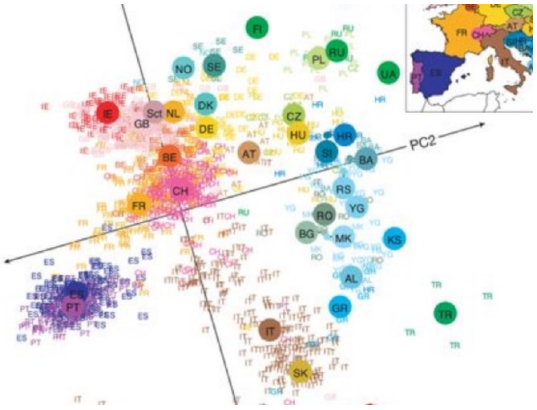
PCA is the most widely used tool for data visualization; many applications in social sciences

And the sciences too!

Example 2: Each data point a_i represents a person in an NIH study and their vector encodes their genes (SNPs, actually)

Again, what do you think will happen?

Q7: What 2-dimensional projection will capture the most genetic variation?



Before we get to other applications, let's do another interpretation of PCA

Sometimes we want to think of PCA in terms of denoising

Q8: Can we find the projection onto a k-dimensional subspace that minimizes reconstruction error? $c = [c_1 \dots c_k]$

i.e. $\min_{\text{orthonormal } c_1 \dots c_k} \frac{1}{p} \sum_{i=1}^{p} \| y_i - \hat{y}_i \|^2$

where $\hat{y}_i = cc^T y_i$

Actually this is the same question as before, in disguise:

$$\| y_i - \hat{y}_i \|^2 = \| (I - cc^T) y_i \|^2$$

$$= \| y_i \|^2 - \| cc^T y_i \|^2$$

$$= \| y_i \|^2 - \| c^T y_i \|^2$$

Hence we have:

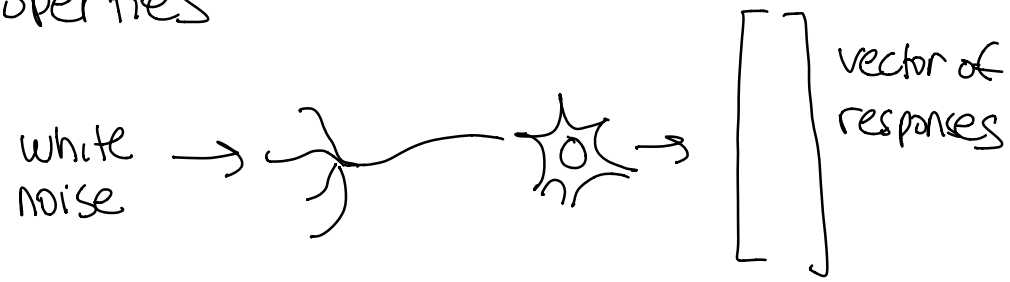
$$\min_{\substack{\text{orthonormal} \\ c_1,\dots,c_k}} \frac{1}{p}\sum_{i=1}^{p}\|y_i-\hat{y}_i\|^2 =$$

$$\frac{1}{p}\sum_{i=1}^{p}\|y_i\|^2 - \max_{\substack{\text{orthonormal} \\ c_1,\dots,c_k}} \frac{1}{p}\sum_{i=1}^{p}\|c^T y_i\|^2$$

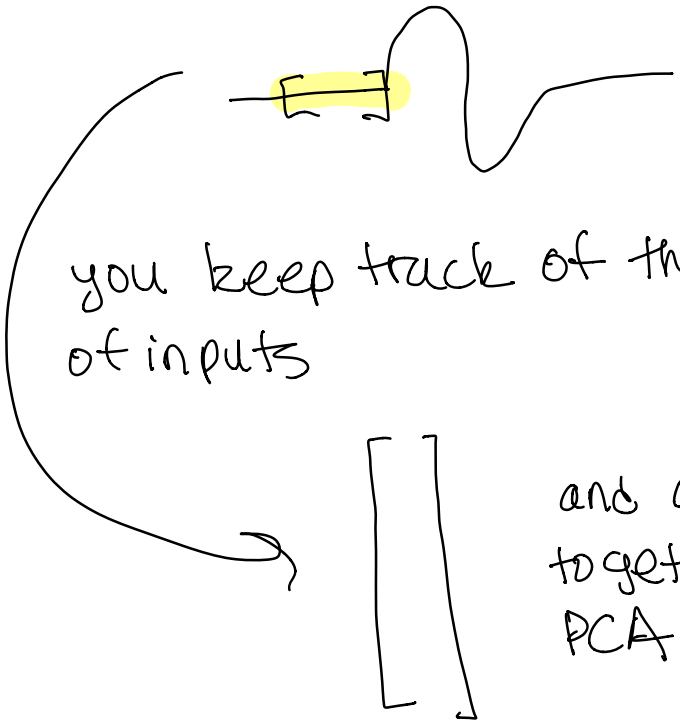
Theorem: The k-dimensional subspace that minimizes the reconstruction error is also achieved by the span of the top k left singular vectors

Let's do a last example for today:

Example 3: Given a collection of neurons, we want to understand their response properties



Spike-Triggered Covariance: When you observe a spike:



you keep track of the preceeding window of inputs

and concatenate them together and perform PCA

The largest / smallest singular vectors give you some idea of excitatory and inhibitory response patterns

That's all for today, but let me plant a question in your mind for next time:

Q9: When does PCA go wrong?