PCA: Final Review

December 9, 2019

Raguvir Kunani

Dimensionality Reduction

The **dimension** of some data is the number of *independent* attributes in the data. In linear algebra, we refer to dimension as the *rank* of a matrix.

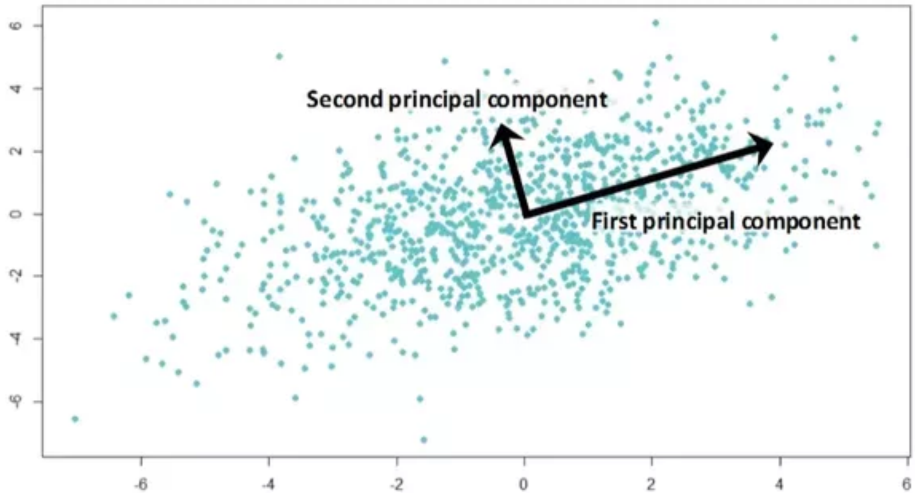
- **What does it mean to reduce the dimensionality of our data?**

  100 features $\longrightarrow$ 10 features

- **Why do we want to reduce the dimensionality of our data?**
  - features might not add info
  - easier to visualize
  - reduces storage
  - reduces model complexity
- **What are some ways to reduce the dimensionality of our data?**
  - PCA
  - manually delete columns

Principal Component Analysis (PCA): Overview

PCA takes advantage of the fact that any data inherently lies along some set of directions. Using SVD, PCA finds these directions and expresses the data as a combination of these directions.



What Are These "Directions?"

PCA uses SVD to find the "most important" directions that describe our data. What does "most important" mean? For us, the "most important" directions are the directions that our data has the most variance along.

Example: Calculating Final Course Grades

"most important" direction is the first row of V^T

this also is the "rubric" that spreads the final course grades the most

$$\left[\begin{array}{ccc} \underline{\text{MT1}} & \underline{\text{MT2}} & \underline{\text{Final Exam}} \\ & & \\ & & \end{array}\right] = X = U\Sigma V^T$$

Fortunately, SVD already calculates these directions for us, so we don't have to manually find them.

* "rubric" means coefficients for a linear combination of the exams

$X = U\Sigma V^T \Rightarrow X V = U \Sigma \underbrace{V^T V}$

$\boxed{XV = U\Sigma}$

How To Get the Directions From SVD

The SVD of a matrix X is $X = U\Sigma V^T$, where U, V are orthogonal and Σ is diagonal.

If X is the (centered) data matrix, then the rows of V^T (columns of V) contain the "most important" directions, aka **the directions of the principal components**. The columns of $U\Sigma$ contain the **principal components** themselves.

cheat sheet

Question: In our final course grade example, what is the direction of the first principal component? What is the first principal component?

first row of V^T (rubric)

the actual final course grade

The Σ matrix

Let's a take a closer look at the Σ matrix from SVD. Let's say that after doing SVD on some data matrix A, the Σ matrix looks like:

N : # of data points

$$\Sigma = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

What is the *amount* of variance explained by PC1?

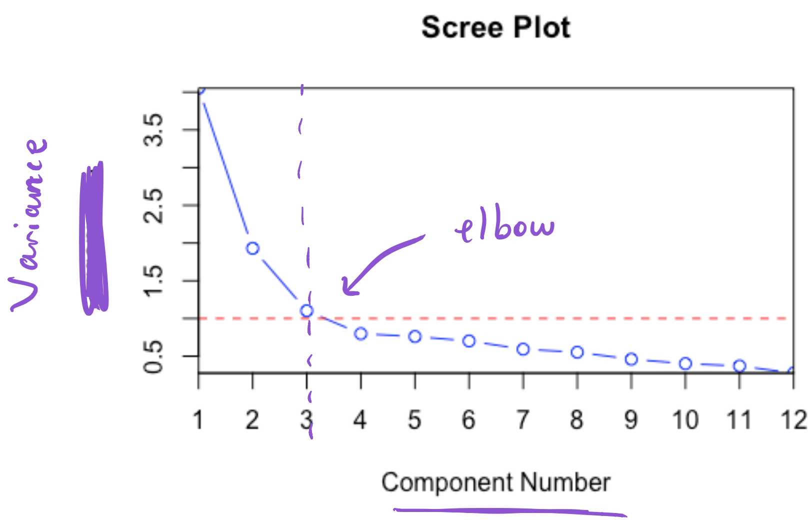
$$\frac{\sigma_1^2}{N} = \boxed{\frac{9}{N} \text{ units}}$$

What is the *fraction* of variance explained by PC2?

$$\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2} = \frac{4}{9+4+1} = \boxed{\frac{4}{14}}$$

Scree Plots

The whole point of PCA is to reduce the dimensionality of our data, so let's return to that. To reduce the dimensionality of our data, we take the k principal components that account for most of the variance in our data. To make this process easier, we use a **scree plot**:



Putting It All Together

At this point, we have chosen k principal components of our data. But what do we actually do with them? Recall that our original goal was to find a lower dimensional representation of our data.

Using our chosen PCs, we can do just that! For a data point $x_i \in \mathbb{R}^n$, we can represent it using our PCs as $(\mathrm{PC1}_i, \mathrm{PC2}_i, \ldots, \mathrm{PCk}_i)$. This means each point is now described by k attributes instead of n!

What information can we glean from this new representation?

Although the PC scores themselves are not interpretable, we can look at the actual coefficients in the rows of V^T to see which features are weighted heavily.

PCA Short Questions

- PCA still finds the principal components even when the data matrix is not centered. (True/False)

If the data matrix X is $m \times n$, what is the size of U, Σ, and V^T?

U : $m \times n$ Σ : $n \times n$ V^T : $n \times n$

If the data matrix X is $m \times n$, what is the size of PC1?

$m \times 1$

PCA Short Questions

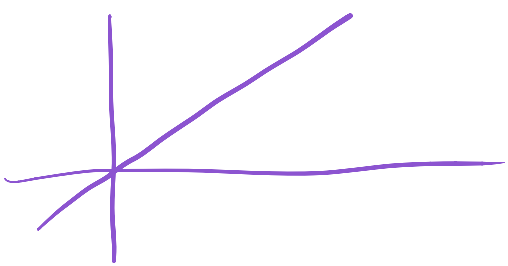
The rank of the data matrix X is the number of negative elements in the Σ matrix from the SVD of X. (True/False)

positive

What is the dot product PC1 · PC2?

anything

What does v_1 (first row of V^T) look like when plotted?



line through the origin

$||x||$: norm/magnitude/length

PCA Short Questions

What is $||u_1||$, where u_1 is the first column of U?

1

What is $u_1 \cdot u_2$?

0

What is $||v_1||$, where u_1 is the first row of V^T?

1

What is $v_1 \cdot v_2$?

0

PCA Short Questions

The columns of $U\Sigma$ contain the principal components. (T/F)

The columns of UX contain the principal components. (T/F)

The columns of XV contain the principal components. (T/F)

PCA Longer Questions

If the SVD of a matrix X is $X = \frac{1}{10}\begin{bmatrix} -6 & 8 \\ 8 & 6 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

What is the rank 1 approximation of X?

$$\frac{1}{10} \cdot 4 \cdot \begin{bmatrix} -6 \\ 8 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} = \frac{4}{10} \begin{bmatrix} 0 & -6 \\ 0 & 8 \end{bmatrix}$$

What is the rank 2 approximation of X?

X is rank 2, so the rank 2 approximation of X is X.

PCA Longer Questions

Assume the same you have a new data point $x = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$. Find the PC1 score for x using the same data matrix X from the previous slide.

For convenience, I did the matrix multiplication and the result is

$$X = \frac{1}{10}\begin{bmatrix} 16 & -24 \\ 12 & 32 \end{bmatrix}.$$

this data is not centered, so this question is invalid.

If the data was centered, then would subtract the mean of each column of X from the corresponding element of x to get a centered version x (let's call it $\bar{x}$). To get the PC1 score, compute $v_1 \cdot \bar{x}$, where v_1 is the first row of V^T.

Good Luck on the Final!

Here are my tips:

1. Choose 1-2 topics per day and just focus on those.
2. Make sure you do all the discussions and labs if you haven't already.
3. Try to study a little for all your classes every day.
4. Come to OH! All the TAs want you to succeed and we will do anything to help you make that happen.
5. Sleep consistently every day until the final.

You've all done an awesome job this semester. Congrats on making it to the end!