Introduction to Statistical Machine Learning
CSC/DSCC 265/465

Lecture 13: Unsupervised Learning – Part IV

Cantay Caliskan

Notes and updates

Notes and updates

- Any questions?
- Practice midterm will be posted on Wednesday (Wednesday, March 2, 2022)
- Midterm review: On Wednesday (Wednesday, March 2, 2022)
- Midterm date: Wednesday, March 16, 2022 (during class time)
- Quick clarification: Midterm will cover everything (including the lecture on Monday, March 14)

Plan for the next lectures

- *PCA*
- *SVD*

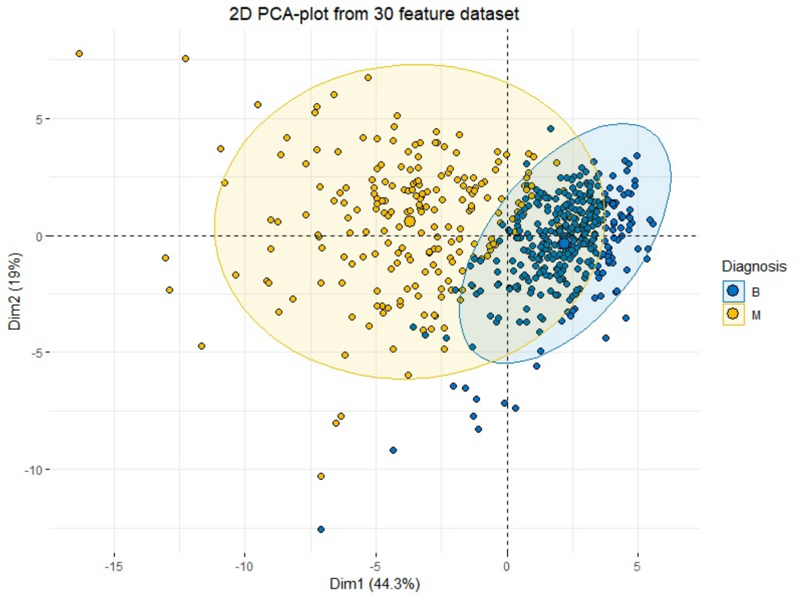
Plan for the next lectures

- ***PCA***
- ***SVD***

Dimensionality Reduction

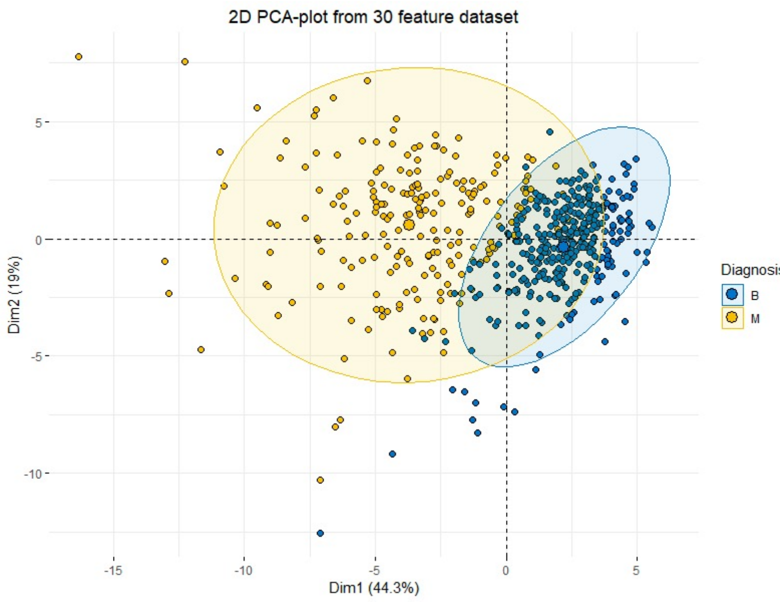
Dimensionality Reduction

- <u>Motivation for dimensionality reduction:</u>
  - Sparse data (= a lot of zeros or NA's)
  - (High/extremely high) number of features
  - Highly correlated/redundant features
  - Noisy features
  - Features that are hard to describe
  - The need to know which feature is important
  - Interpretation / visualization
  - Computational burden
  - ***Curse of dimensionality***



Dimensionality Reduction

- <u>Different techniques:</u>
  - Missing Values Ratio
  - Low Variance Filter
  - High Correlation Filter
  - Random Forests / Ensemble Trees
  - Backward Feature Elimination
  - Forward Feature Construction
  - Linear Discriminant Analysis (LDA)
  - Generalized Discriminant Analysis (GDA)
  - T-Stochastic Neighbor Embedding (t-SNE)
  - **Principal Component Analysis (PCA)**



Dimensionality Reduction

1) **Feature elimination**
2) **Feature extraction**

- ***High Dimensions = Lots of Features***

- Dimensionality reduction is *helpful*:
  - Often too many features to do a final classification
  - Higher #features -> more difficult to visualize
  - Higher #features -> more difficult to make classification

Example: Calculating the US GDP

- Let's say you want to predict the **gross domestic product (GDP)** of the United States for 2022. There is a lot of information available:
  - ***Recorded*** and ***unrecorded*** data
  - The US GDP for the first quarter of 2022
  - The US GDP for the entirety of 2021, 2020, 2019 etc.
  - Publicly available economic indicators:
    - Unemployment rate
    - Inflation rate
    - Etc.
  - US Census data from 2010 estimating how many Americans work in each industry
  - Stock price data, number of IPOs etc.

Example: Calculating the US GDP

- Question: How do I take all of the variable I have collected and focus on only a few of them?
  - You want to ***reduce dimensionality***
    - To reduce the risk of overfitting …
    - To reduce computational burden …

- So, you need:
  - Feature elimination, OR
  - Feature extraction

Feature Elimination and Extraction

Feature Elimination:

- We drop the variables that we think are not needed to come up with a good explanatory / predictive model

- <u>Disadvantage</u>: Losing the information from the variables that you drop
  - e.g. your R^2 value may drop

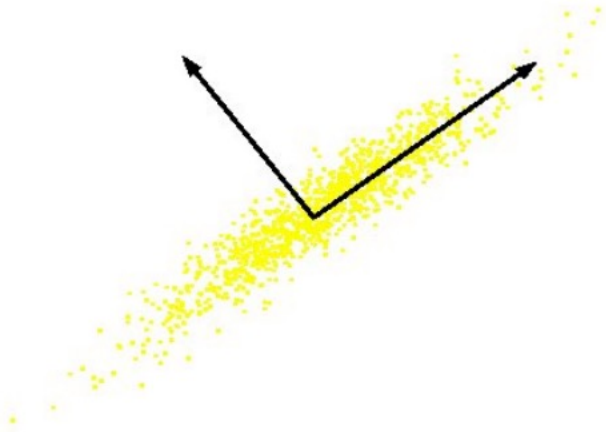
Feature Extraction:

- We take all of our **n** independent variables, and create **n** new features
- Each feature is a weighted combination of each of the "old" independent variables
- We order these new variables by how well they predict the dependent variable and we drop the ones that are <u>not</u> doing a 'good job'

Principal Components Analysis (PCA)

What is Principal Components Analysis?

- A **feature extraction** technique
- Unsupervised technique for extracting variance structure from high dimensional datasets



- An ***orthogonal*** projection or transformation of the data into a (possibly lower dimensional) subspace so that the variance of the projected data is maximized.
    - Focus is on identifying the ***correlations*** in the dataset

Math: Principal Component Analysis (PCA)

1. Standardize the data (in most cases)

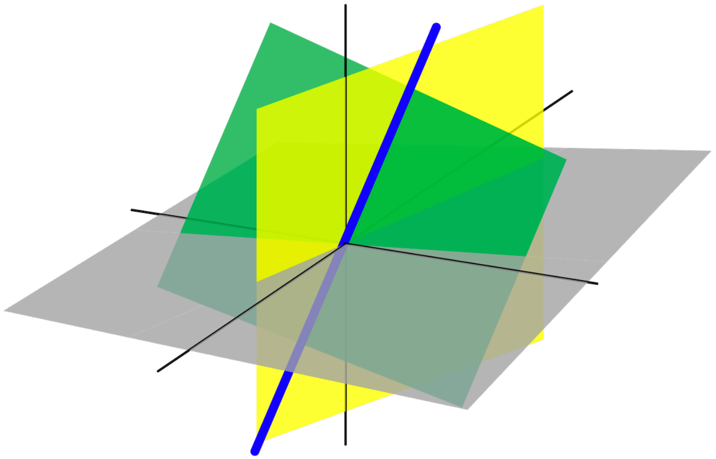
2. Calculate the covariance matrix

3. Find the eigenvalues and eigenvectors of the covariance matrix

4. Plot the eigenvectors / principal components over the scaled data

PCA: Mathematical Basics

- Linear algebra:
  - Eigenvectors
  - Eigenvalues
  - Matrix algebra



- Statistics:
  - Standard deviation
  - Variance
  - Covariance
  - Covariance matrix

Review: Standardizing the data

- **Most common types:**

  - *Standard score*: $\dfrac{X - \mu}{\sigma}$

  - Min – Max Feature scaling: $X' = \dfrac{X - X_{min}}{X_{max} - X_{min}}$

  - <u>Which one to choose when?</u>
    - Normality assumption
    - Spread of your data
      - Are you interested in the spread or controlling the spread?

Review: Eigenvalues and Eigenvectors

- Definition: If *T* is a linear transformation from a vector space *V* over a field *F* into itself and **v** is a vector in *V* that is not the *zero vector*, then **v** is an eigenvector of *T* if *T*(**v**) is a scalar multiple of **v**
  - $T(\boldsymbol{v}) = \boldsymbol{\lambda} v$
    - Where: **λ** is the eigenvalue

- If the vector space *V* is finite-dimensional, then the linear transformation T can be represented as a square matrix *A*, and the vector **v** by a column vector:
  - $A\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$

UNIVERSITY of ROCHESTER

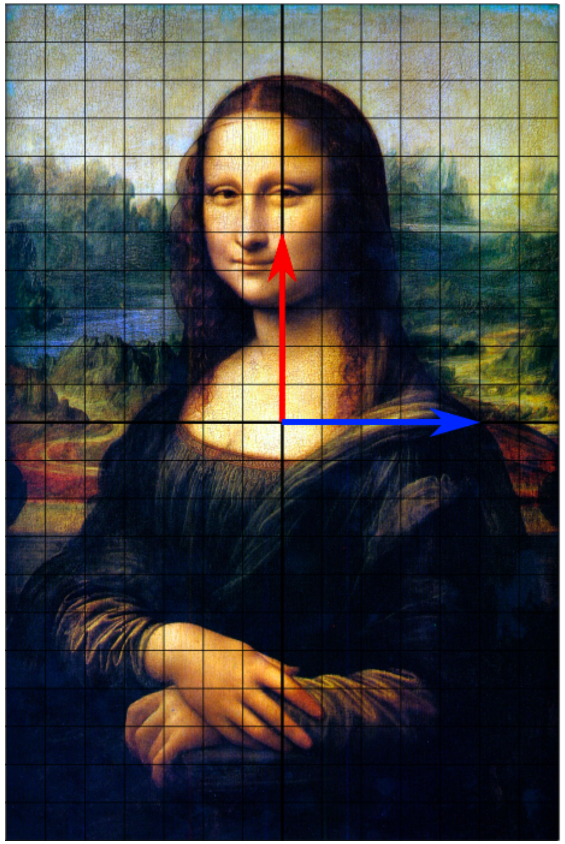
Review: Eigenvalues and Eigenvectors

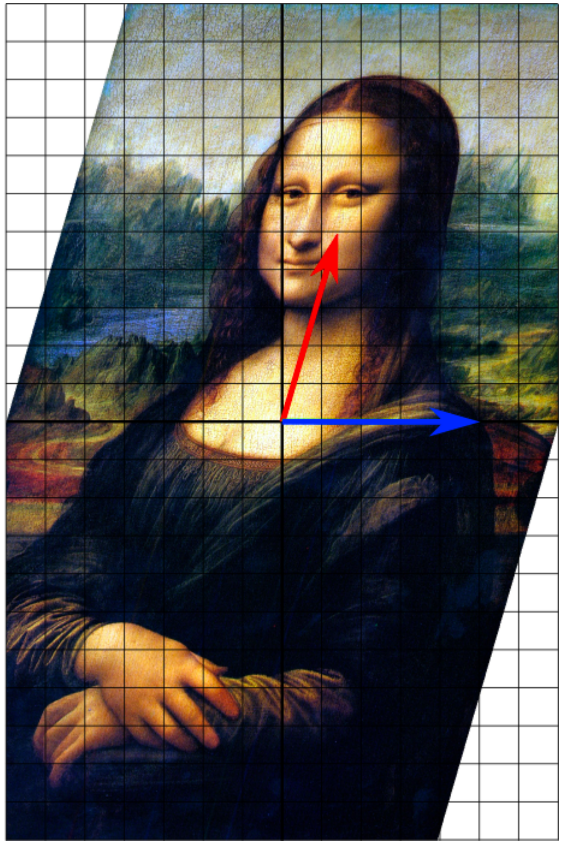
- Example: Shear mapping

- Any vector that points directly to the right or left with no vertical component is an eigenvector

- Length of vector: ***eigenvalue***

- ***Eigenvalue***: (sort of) measure of distortion
- ***Eigenvector***: (sort of) orientation of distortion





Variance – Covariance Matrix

- Let's say: You have a 3-dimensional data set, you can measure covariance between each dimension:
  - Covariance between ***x and y*** dimensions
  - Covariance between ***y and z*** dimensions
  - Covariance between ***x and z*** dimensions
- So you can construct a ***variance – covariance matrix*** that looks like:

$$\begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

The diagonal is the variances …

- You get the variance – covariance matrix by multiplying your mean-standardized matrix ($\mathbf{P}$) by its transpose ($\mathbf{P}^{\mathrm{T}}$)

Why (bother)?

- Why bother with calculating covariance when we could just plot the 2 values to see their relationship?
- In ***lower dimensions*** -> There are many options, no need for complicated analysis
- In ***higher dimensions*** -> You need to convert your data to be able to analyze it
- <u>Big</u> motivation -> ***Visualization***: Only possible in 2D or 3D (or 1D, but no one likes that).

UNIVERSITY of ROCHESTER

Principal Component

Definition:

- Direction of maximum variance in the input space
- Principal eigenvector of the covariance matrix

- But, how can we relate these two definitions?

- ***Variance***: We have:
  - A random variable fluctuating about its mean value
  - Average of the square of fluctuations

- ***Covariance***: We have:
  - Pair of random variables each fluctuating about their mean values
  - Average of product of products

Principal Components

Remember (from <u>two</u> slides ago) ☺ :

$$A = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix}$$

We have the variance – covariance matrix.

- And let's find the ***eigenvectors*** with ***k largest eigenvalues***

$$\boxed{\begin{aligned} A\,v_1 &= \boldsymbol{\lambda}\,v_1 \\ A\,v_2 &= \boldsymbol{\lambda}\,v_2 \\ &\dots \\ A\,v_k &= \boldsymbol{\lambda}\,v_k \end{aligned}}$$

$$\dots$$

$$A\,v_N = \boldsymbol{\lambda}\,v_N$$

The first ***k*** components explain most of the variance in the dataset.

And each component is <u>**independent**</u> from each other.

The first ***k*** components are called the principal components!

How do we do PCA?

Step 1:

<u>Decide on the following:</u>

- Only dimension reduction ?
  - ➢ Keep all data points
- ***First*** dimension reduction, ***then*** prediction ?
  - Separate the outcome variable ***Y*** from the rest of your dataset (***X***'s)

Step 2:

<u>Subtract the mean of each variable from itself.</u> So, each variable would have a mean zero (pre-standardization).

<u>And decide if you want to standardize:</u>

- Interested in the variance of your variables?
  - ***Don't standardize***
- Interested in central tendency?
  - ***Standardize***

How do we do PCA?

Step 3:

Calculate the variance – covariance matrix:

- Take the ("standardized / non-standardized") data you created
- Put it into matrix format ($\mathbf{P}$)
- Take matrix P and transpose it ($\mathbf{P}^{\mathrm{T}}$)
- Calculate $\mathbf{P}^{\mathrm{T}}\mathbf{P}$ and find the ***variance– covariance matrix***

Step 4:

Technical step:

- Transform $\mathbf{P}^{\mathrm{T}}\mathbf{P}$ into a $\mathbf{ZDZ}^{-1}$ format
  - $\mathbf{Z}$: matrix of eigenvectors
  - $\mathbf{D}$: diagonal matrix with eigenvalues on the diagonal
- This procedure is called ***eigendecomposition***

How do we do PCA?

Step 5:

Take a look at the eigenvalues:

- $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$
- Sort the eigenvalues from the largest to the smallest
- Find the column in $\mathbf{Z}$ that corresponds to the largest ***eigenvalue*** and place it in the first position (most important vector)
- Call this a sorted matrix $\mathbf{Z}^*$

Step 6:

Another technical step:

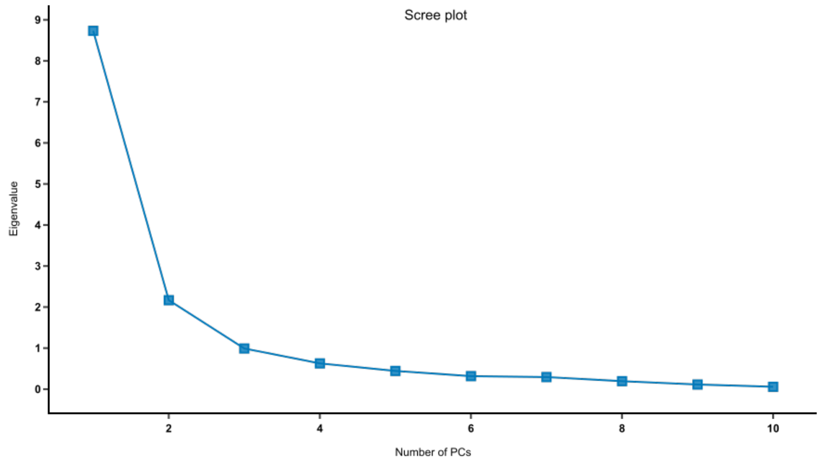
- Calculate a new matrix $\mathbf{P^*} = \mathbf{PZ^*}$
- This new matrix is a centered/standardized version of $\mathbf{X}$ and each observation is a weighted combination of original variables
- Weights are determined by eigenvectors
- And columns of $\mathbf{P^*}$ are independent

How do we do PCA?

Step 7 – Final Step:

<u>How many columns do we want to keep?</u>

- We created a new matrix **P*** where columns are independent from each other
- We have several options:
  - Arbitrarily select the number of columns we want to keep
  - Calculate the <u>proportion of variance explained</u> and pick a ***threshold***
  - Find the cumulative proportion of variance explained and set a ***threshold***
    - ***Find the elbow in the graph!***

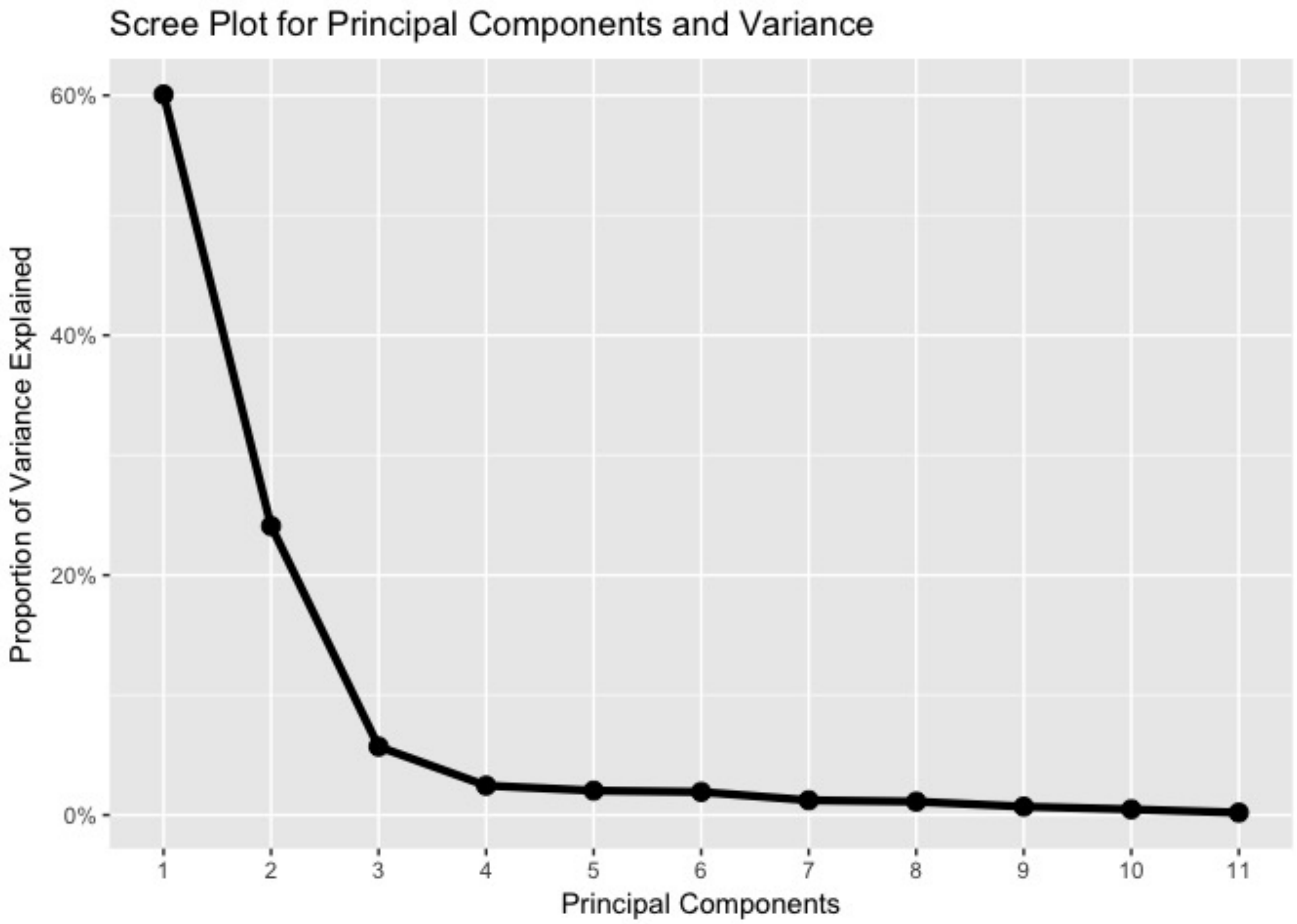


This is called a ***scree plot***

<u>Question</u>: How do we calculate the proportion of variance explained?

<u>Answer</u>: From the Var-Cov matrix (P*) of the standardized X matrix.

How do we do PCA?



<u>Scree Plot</u>

- Take a look at the cumulative proportion of variance
- Find the <u>biggest</u> slope change
- Choose the number of components ***k*** before the <u>biggest</u> slope change
    - <u>Here</u>: **k = 3**

When should you use PCA? (Methodologically)

1. Do you want to reduce the number of variables, but aren't able to identify which variables you can remove from your consideration?

2. Do you want to ensure your variables are independent of one another?

3. Are you comfortable making your independent variables less interpretable?

- If you answered ”**yes**” to all -> try PCA
- If you answered “**no**” to question 1 and/or question 2 -> you can still use PCA
- If you answered “**no**” to question 3 -> do not use PCA

UNIVERSITY of ROCHESTER

Should We Use It? (Technically)

- **Advantages:**
  - It helps in data compression, and hence reduced storage space
  - It reduces computation time
  - It helps to remove redundant features, if any
  - You can use the new variables in a linear regression setting!
- **Disadvantages:**
  - It may lead to some amount of data loss
  - PCA tends to find linear correlations between variables, which is sometimes undesirable
  - PCA fails in cases where mean and covariance are not enough to define datasets
  - We may not know how many principal components to keep- in practice, some thumb rules are applied

Kaggle Competition and Team Building

Kaggle Competition

- Goal: Classifying ***fake news*** by topic (**multi-class classification** problem)
  - Original dataset contains ~10,000 fake news
  - Plan: You will be provided with ~5,000 observations
  - Expectation: Predict the topics for the rest of the news!
  - You can work in a team of two (2) people
    - And, you are encouraged to work in a team!
  - Dataset contains:
    - ***Date***
    - ***Origin*** (Country)
    - ***Origin*** (Media Source)
    - ***Brief information***
    - ***Long information***
    - ***Topic class*** (~30 different topics)

Kaggle Competition

- ***Expectations***
  - A ***descriptive analysis***
    - <u>One section</u> for undergraduate students
    - <u>Two sections</u> for graduate students
  - A ***prediction challenge***
    - <u>Plan</u>: You will be able to choose any classification model you would like
      - Two separate lists of ranking for undergraduates and graduates
  - A ***report***
    - Summarize your findings and strategies with a final report

Team Building

- Have you met any people from the classroom?
- Would you like to get to know more people?
- Have you checked the class Facebook?

Not this Facebook! ☺

- There is a link on BlackBoard that you can use to identify your team
  - Content Menu -> Kaggle Teams
- Let’s use the remainder of the time for team building.

UNIVERSITY of ROCHESTER

Team Building Exercise

- If you are looking for a ***team partner*** for:
  - **Kaggle Competition**
  - **Final Project**

- Instructions:
  - Please come to the front (in a few minutes), and:
  - Meet with at least three (3) people
  - Submit their ***names and surnames***, ***programs and majors,*** *and a* ***'selfie'*** [as a message] on ***BlackBoard***
  - Reminder: Upload the 'selfie' as well!
  - Deadline: 3:30 PM today
  - You will receive some extra credit

UNIVERSITY of ROCHESTER