UML - Data Reduction:
Principal Components Analysis

Brock Tibert
btibert@bu.edu

BU
Boston University Questrom School of Business

Outline

- Team Project Deliverable 1 - Quick Discussion
- Dimensionality Reduction Techniques
- Principal Components Analysis (PCA)
- Hands-on applications in python
- Team Data Challenge
  - Unsupervised and Supervised Techniques go well together!

Assignment 1 Discussion

- This is an **individual assignment**. This is not to be discussed with others.
- Analyze the data with the tools we have covered in class
- You will notice that there is a high level problem, so the analytics plan is up to you!
- The dataset is intentionally vague, so work through it and challenges you may face the best you can!
- No more than a 3 page PDF executive summary for your client
  - Includes your discussion of how you came to your findings/recommendations
  - Clearly state your recommendations (and why)
  - Supporting evidence as necessary (charts, tables, etc.)
- You will also include your work as a python script
  - Not a notebook, but a valid, **functioning** .py script
  - **If you are using Colab, just File > Download > Download .py is all you need to do**

Team Project Looking ahead:Deliverable 1

- A brief write up of your ideas (the problem you are looking to solve) and the dataset you are considering
- A few examples exist in the Examples folder on Q-Tools
- Restricted datasets listed in course repo
- Top Level discussion
  - Quick discussion of the project **as you see it today**
  - Identification of a dataset
  - What are you thinking about for the challenge
- You might need to build your dataset (scrape, annotate)

Google Dataset Search Beta

Search for Datasets

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

kaggle

Dimensionality Reduction

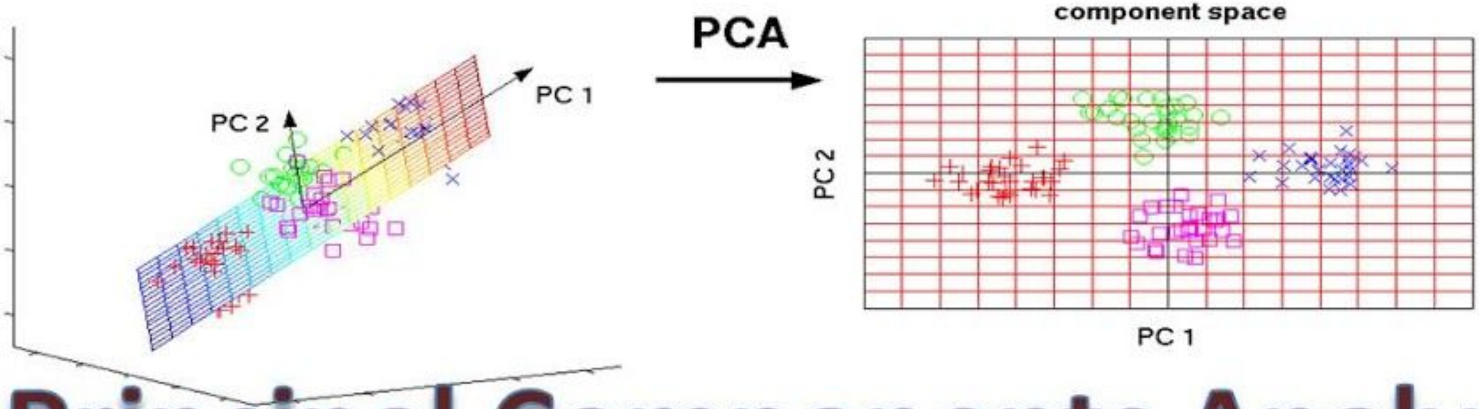
Feature Reduction

Problems with too many variables:

- Analyses can take a very long time
- Risk of correlation amongst the variables
  - Difficulty in interpreting the fit of our models
- Curse of Dimensionality!

Can be viewed as part of the Exploratory Data Analysis phase, but can also fall under unsupervised machine learning tasks

In the end, we want the included variables to contribute towards different dimensions of our problem and typically want to avoid redundant features that generate noise, not signal



Principal Components Analysis (PCA)

Principal Component Analysis (PCA)

- A very popular feature reduction procedure
- Can also be used for compressing our data or as a visualization technique
- It is useful when you have a large feature space and are not certain as to how they might be related or if they are redundant
    - Flattened Image data, computer vision, genetic and biological datasets, etc
- The variables in the dataset may be measuring some underlying constructs
    - Overly simplified example = Height and pant length

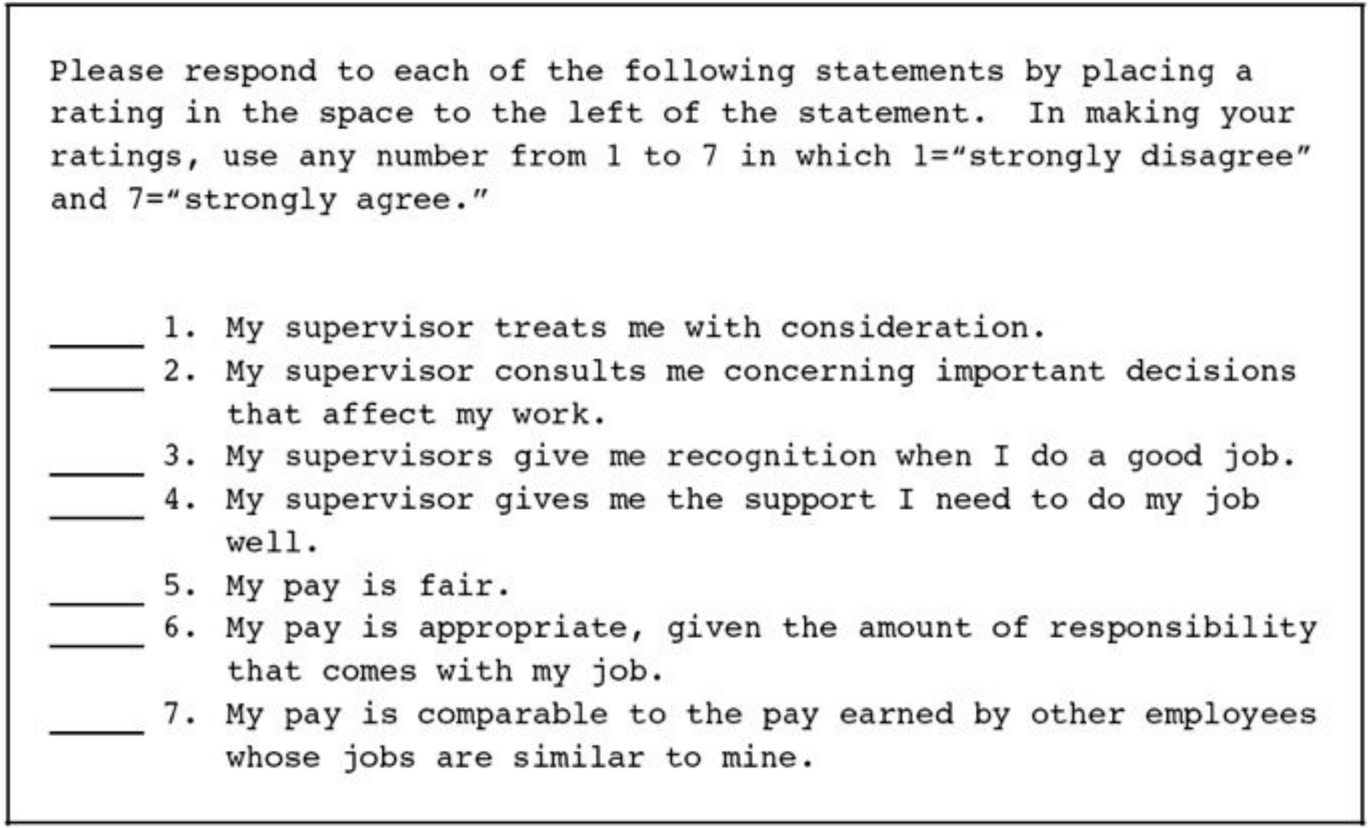
We strive to reduce the observed features into a smaller number of **principal components** (artificial variables) that account for most (if not all) of the signal in our data, and discarding noise. Once we have this space reduced appropriately, we can either visualize or compress our original data.

What is a Principal Component

A **principal component** is an artificial variable that is defined as an “optimal” linear combination of original variables.

We strive to reorganize the information (correlation) in our dataset to retain the key information while reducing redundancy across the predictors.

7 item measure of Job Satisfaction



Please respond to each of the following statements by placing a rating in the space to the left of the statement. In making your ratings, use any number from 1 to 7 in which 1="strongly disagree" and 7="strongly agree."

_____ 1. My supervisor treats me with consideration.

_____ 2. My supervisor consults me concerning important decisions that affect my work.

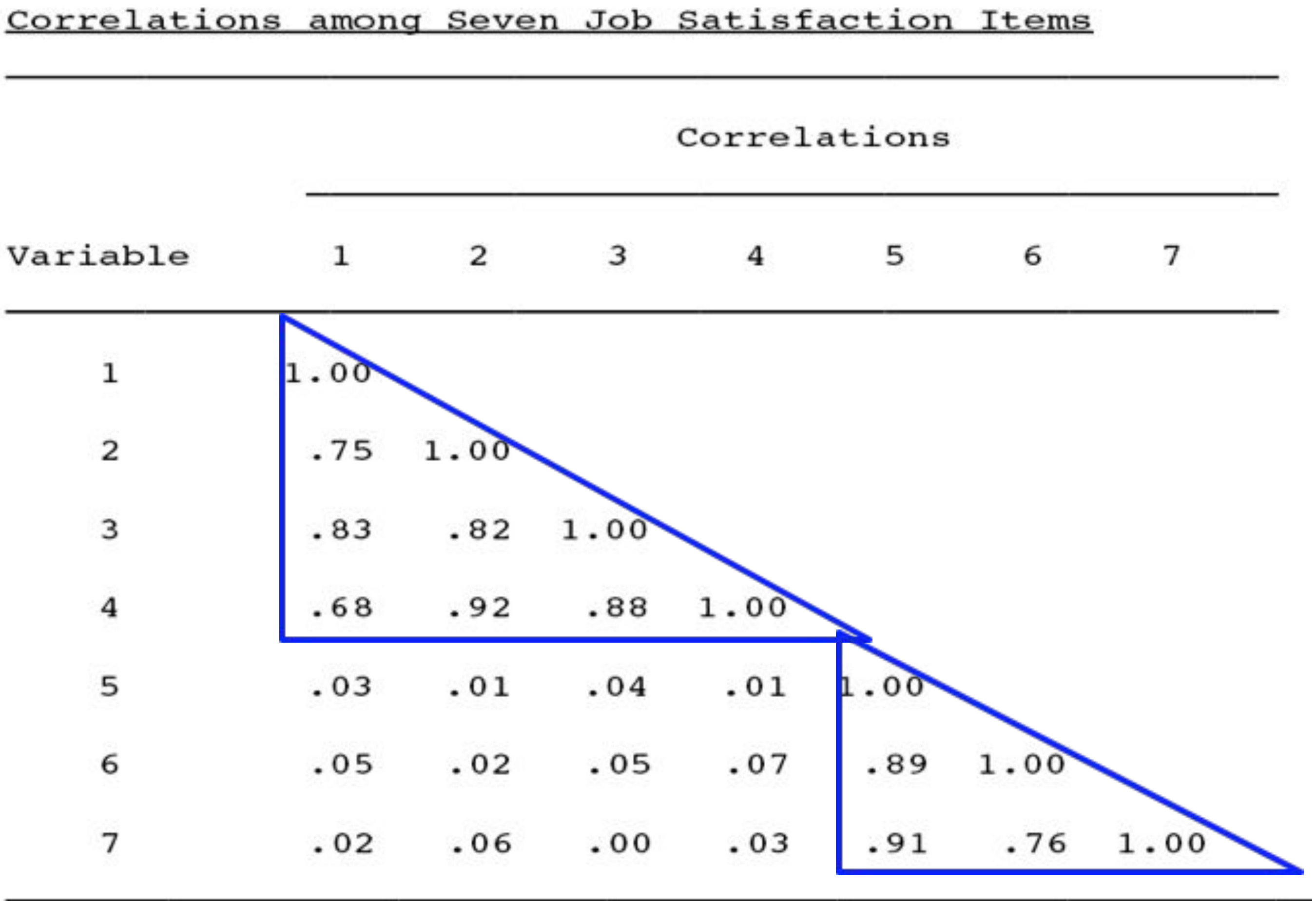
_____ 3. My supervisors give me recognition when I do a good job.

_____ 4. My supervisor gives me the support I need to do my job well.

_____ 5. My pay is fair.

_____ 6. My pay is appropriate, given the amount of responsibility that comes with my job.

_____ 7. My pay is comparable to the pay earned by other employees whose jobs are similar to mine.



Correlations among Seven Job Satisfaction Items

	Correlations						
Variable	1	2	3	4	5	6	7
1	1.00						
2	.75	1.00					
3	.83	.82	1.00				
4	.68	.92	.88	1.00			
5	.03	.01	.04	.01	1.00		
6	.05	.02	.05	.07	.89	1.00	
7	.02	.06	.00	.03	.91	.76	1.00

Reducing the Number of Variables

- Answers to questions 1-4 have high correlation
- Same for answers to questions 5-7
- It stands to reason that we should be able to reduce these related variables into a single observation

PCA: Variables that can capture most of the variance among responses and that are as uncorrelated with each other as possible.

This is a major property/benefit of PCA.

Characteristics of principal components

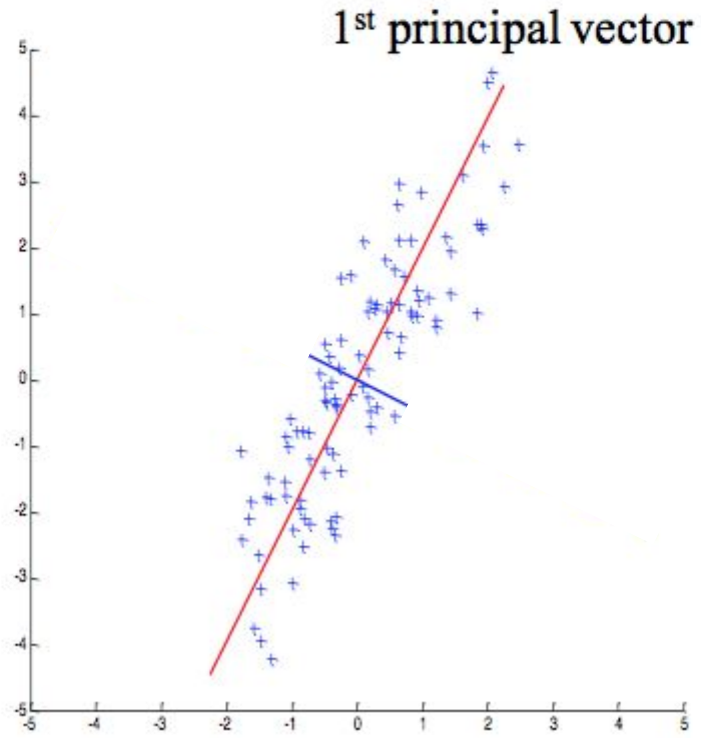
- The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables.

Principal Components in 1 dimension

1st Component:

Projection of each data point along the

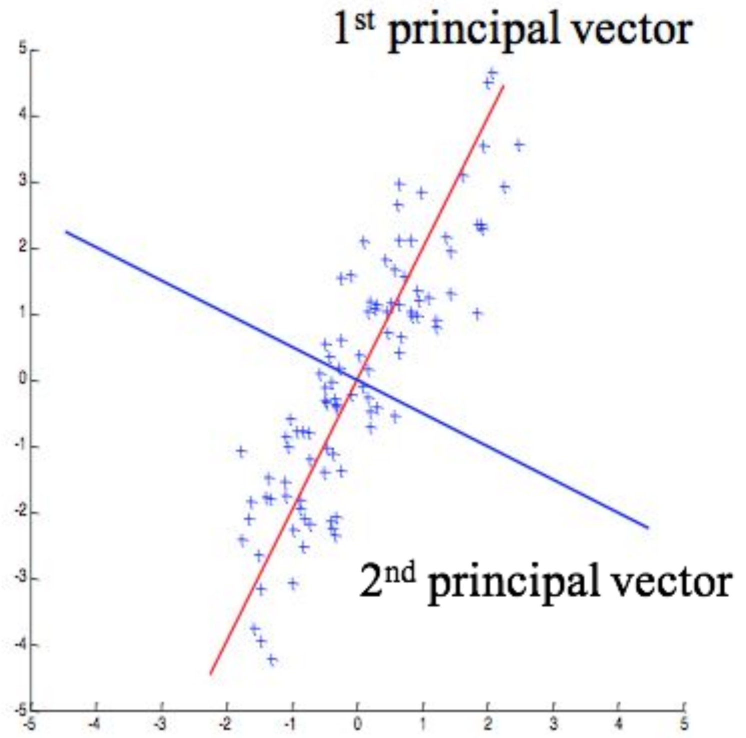
1st principal vector



Characteristics of principal components

- The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables.
- The second component extracted will have two important characteristics.
  a. First, this component will account for a maximal amount of variance in the data set that was not accounted for by the first component.
  b. The second characteristic of the second component is that it will be uncorrelated with the first component.

Principal Components in 2 dimensions



Characteristics of principal components

- The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables.
- The second component extracted will have two important characteristics.
  - First, this component will account for a maximal amount of variance in the data set that was not accounted for by the first component.
  - The second characteristic of the second component is that it will be uncorrelated with the first component.
- The remaining components that are extracted in the same way

The Extracted Components

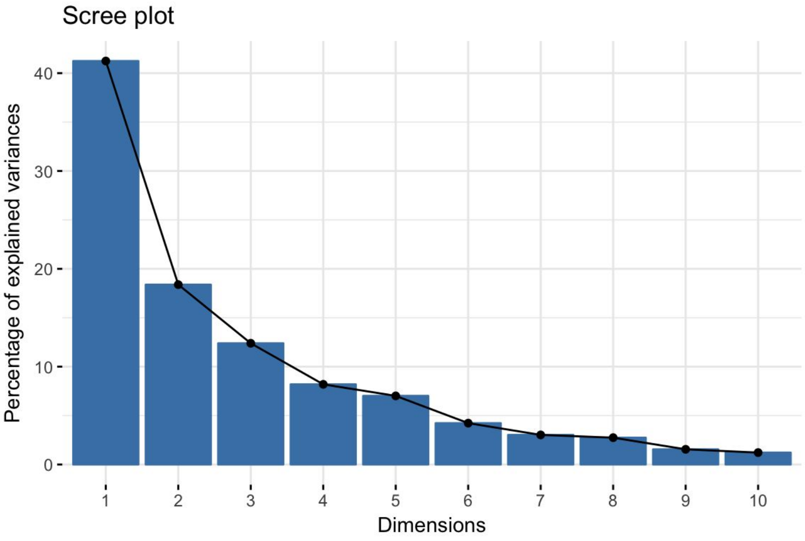
- The number of components available in a principal component analysis is equal to the number of observed variables being analyzed.

- However, in *most* analyses, only the first few components account for meaningful amounts of variance (>90%), so only these first few components are retained, interpreted, and used in subsequent analyses

How Many Components do we select?

We want to maximize variance captured but minimize the number of PC's generated

- Look for the "elbow" on the plots that look at PCs versus (cumulative) variance
- Are you capturing enough of the variance within the dataset for your problem's needs?
- Another rule of thumb approach:
  - Explained variance (eigenvalue) > 1



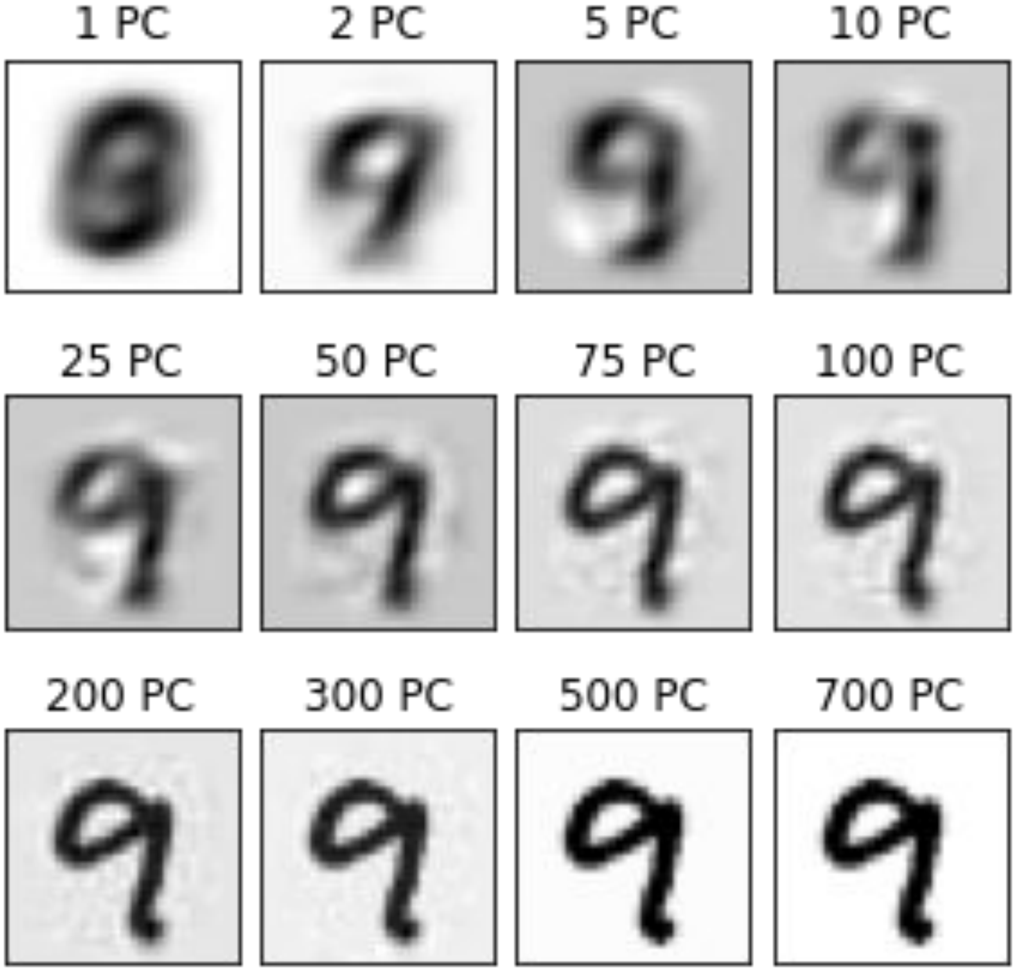
Visual Intuition

MNIST

784 features

For me, around 25-50 PCs makes this easy to understand the digit

Only 3-6% of the features required for us to assess the signal for recognition!



Considerations and Applications

Considerations

- PCA is predicated on linearity, so nonlinear structures may not perform that well
- Severe outliers can have a negative impact on the results

Applications

- Feature Engineering (the PC's become our new columns to use downstream)
- Image analysis/compression
- Biology (high-dimensional expression data, genetic data)

Data Reduction in python