DATA MINING

Dimensionality Reduction

Linear Methods – PCA

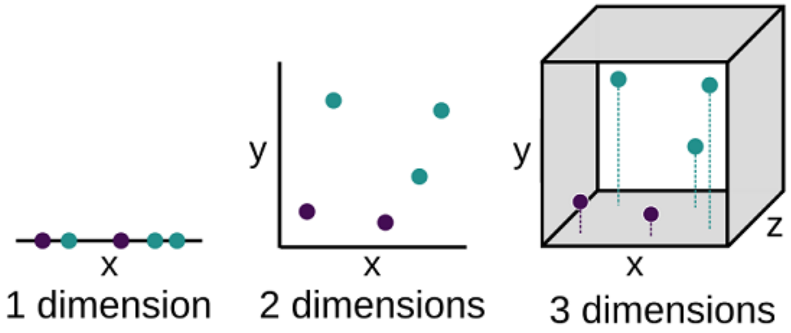
Ashish Pujari

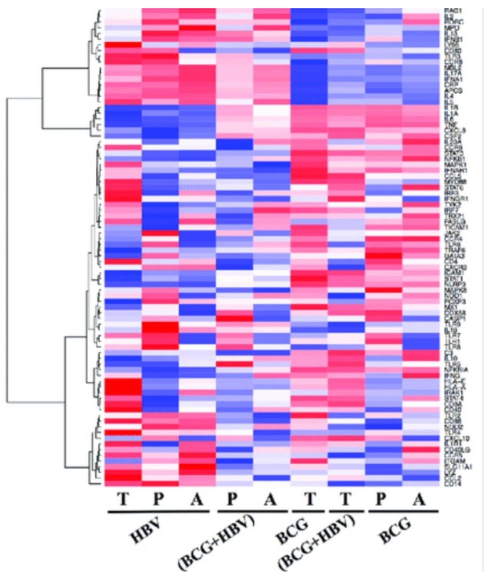
Lecture Outline

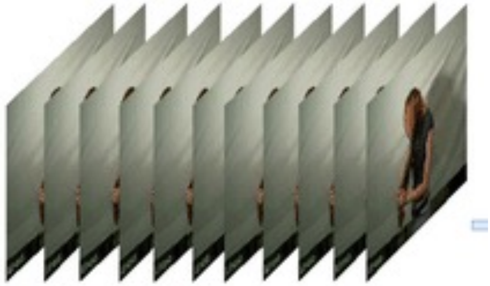
- Dimensionality Reduction
- Mathematical Foundations
- Principal Component Analysis

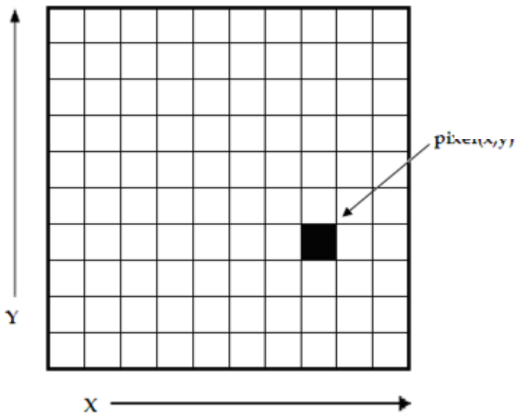
DIMENSIONALITY REDUCTION

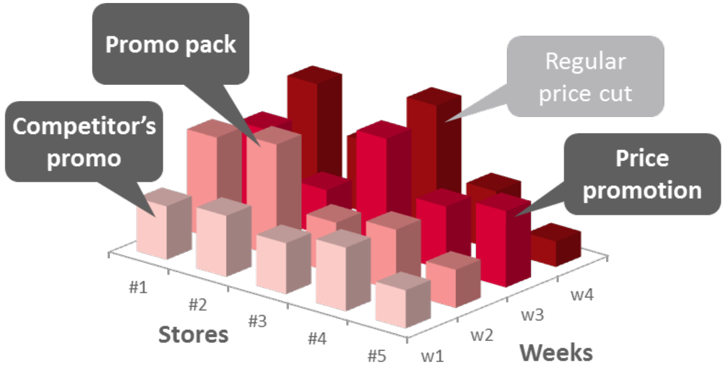
High Dimensional Data





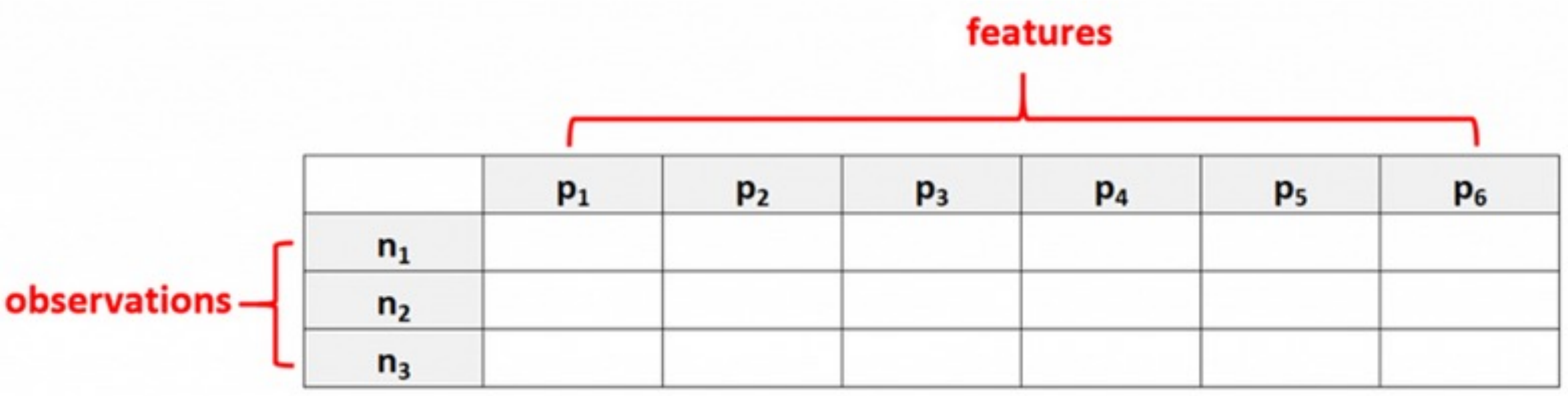




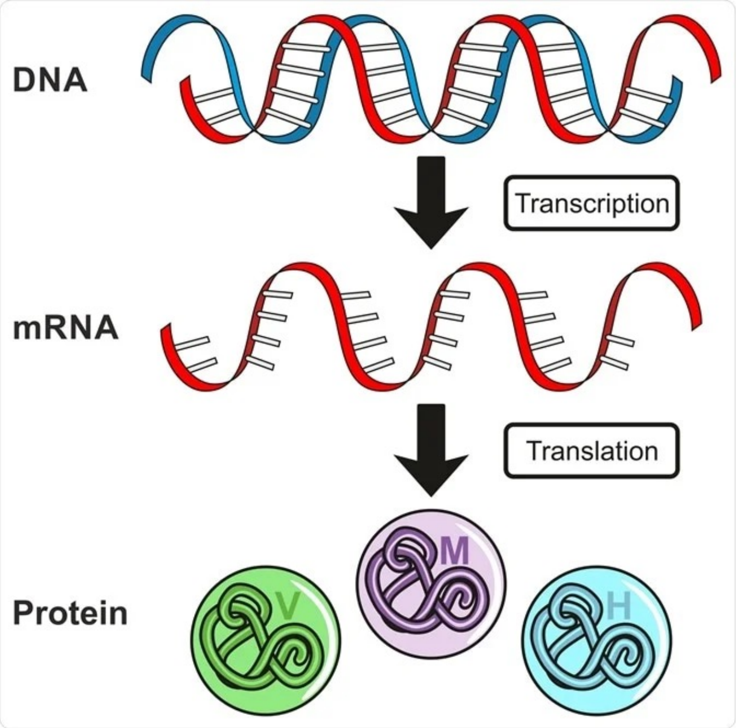


High Dimensional Data

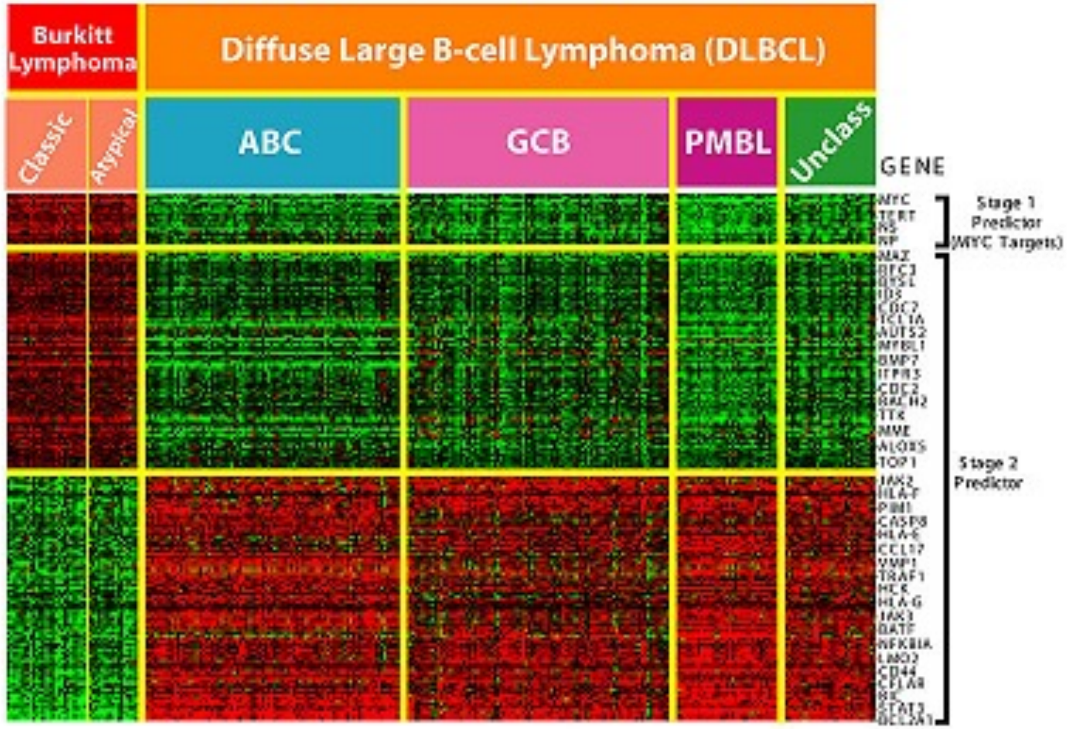
- Dataset in which the number of features p is larger than the number of observations N, often written as p >> N.



High Dimensional Data: Genomics



Source: udaix/Shutterstock.com

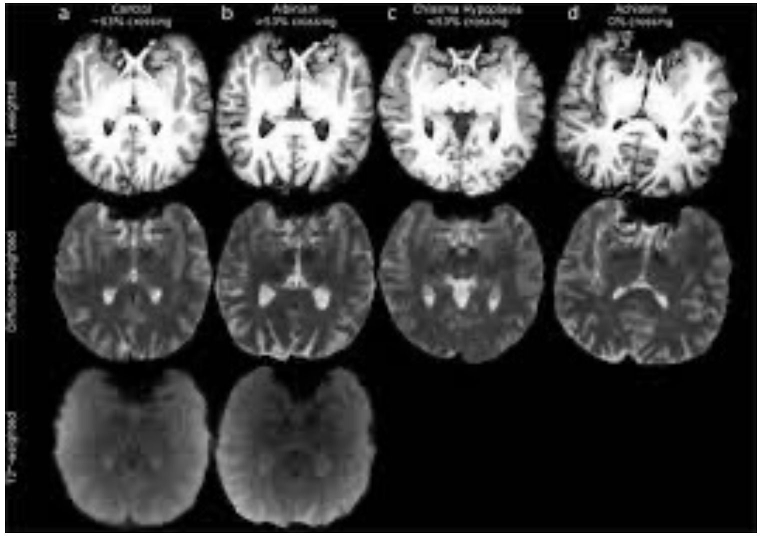


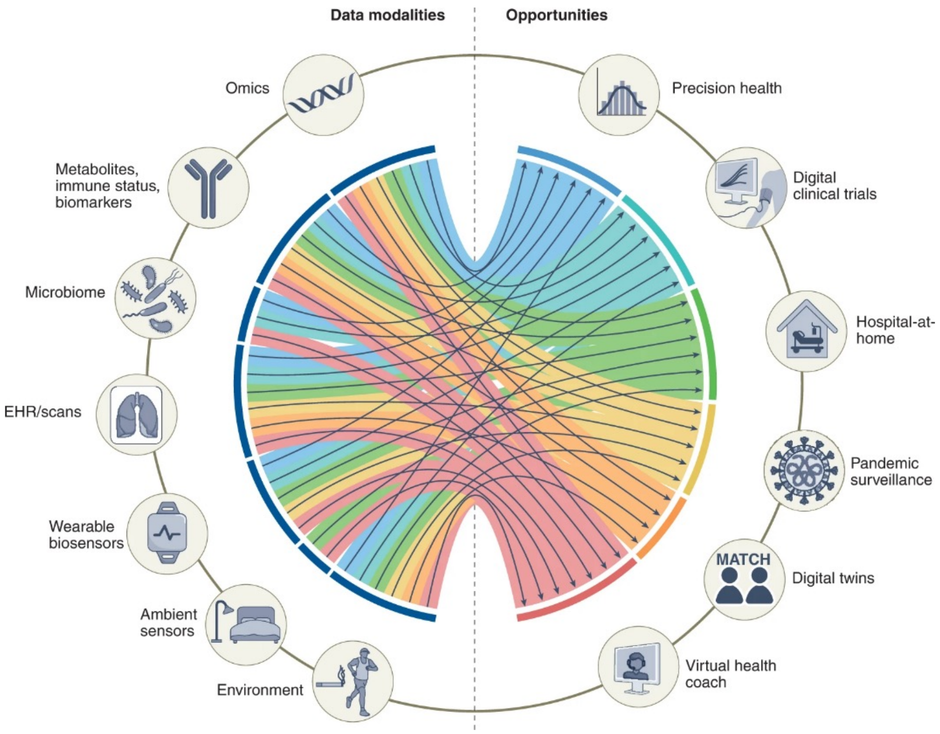
High-dimensional microarray analysis.

http://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

High Dimensional Data: Healthcare

- E.g., MRI, blood pressure, resting heart rate, immune system status, surgery history, height, weight, existing conditions, etc.





https://www.nature.com/articles/s41591-022-01981-2

High Dimensional Data: Finance

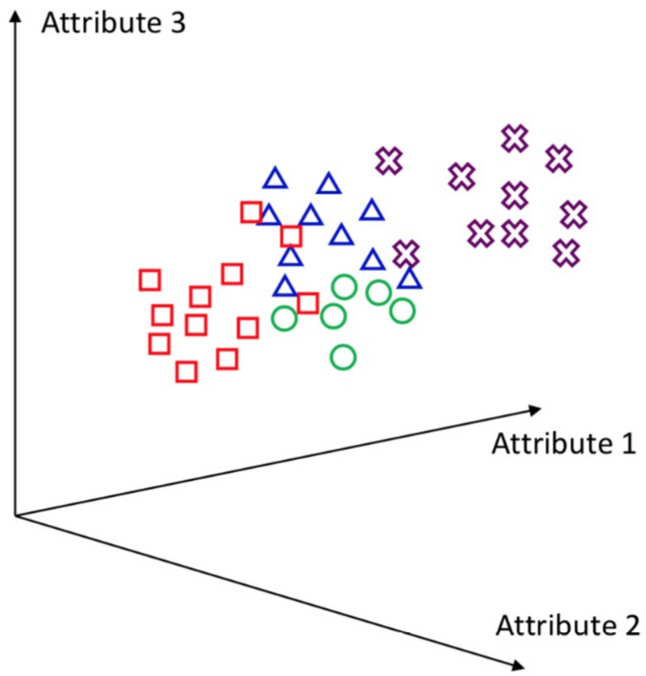
- E.g., PE Ratio, Market Cap, Trading Volume, Dividend Rate, etc.

MSFT [NASD]
Microsoft Corporation
Technology | Software - Infrastructure | USA

statements

Index	DJIA S&P500	P/E	30.35	EPS (ttm)	9.40	Insider Own	0.06%	Shs Outstand	7.50B	Perf Week	0.01%
Market Cap	2198.40B	Forward P/E	26.50	EPS next Y	10.77	Insider Trans	-17.06%	Shs Float	7.48B	Perf Month	-5.04%
Income	71.19B	PEG	1.89	EPS next Q	2.19	Inst Own	71.90%	Short Float	0.49%	Perf Quarter	-5.73%
Sales	184.90B	P/S	11.89	EPS this Y	39.70%	Inst Trans	-0.02%	Short Ratio	1.01	Perf Half Y	-7.16%
Book/sh	21.32	P/B	13.38	EPS next Y	15.41%	ROA	21.60%	Target Price	369.58	Perf Year	10.47%
Cash/sh	16.27	P/C	17.54	EPS next 5Y	16.08%	ROE	48.40%	52W Range	238.07 - 349.67	Perf YTD	-15.17%
Dividend	2.48	P/FCF	50.65	EPS past 5Y	25.70%	ROI	28.20%	52W High	-18.41%	Beta	0.93
Dividend %	0.87%	Quick Ratio	2.20	Sales past 5Y	13.00%	Gross Margin	68.80%	52W Low	19.84%	ATR	7.60
Employees	181000	Current Ratio	2.20	Sales Q/Q	20.10%	Oper. Margin	42.50%	RSI (14)	42.41	Volatility	2.75% 2.22%
Optionable	Yes	Debt/Eq	0.42	EPS Q/Q	22.30%	Profit Margin	38.50%	Rel Volume	0.61	Prev Close	285.30
Shortable	Yes	LT Debt/Eq	0.38	Earnings	Apr 26 AMC	Payout	24.30%	Avg Volume	36.34M	Price	285.30
Recom	1.60	SMA20	-4.79%	SMA50	-3.50%	SMA200	-6.47%	Volume	0	Change	0.00%

High Dimensional Data: Analysis

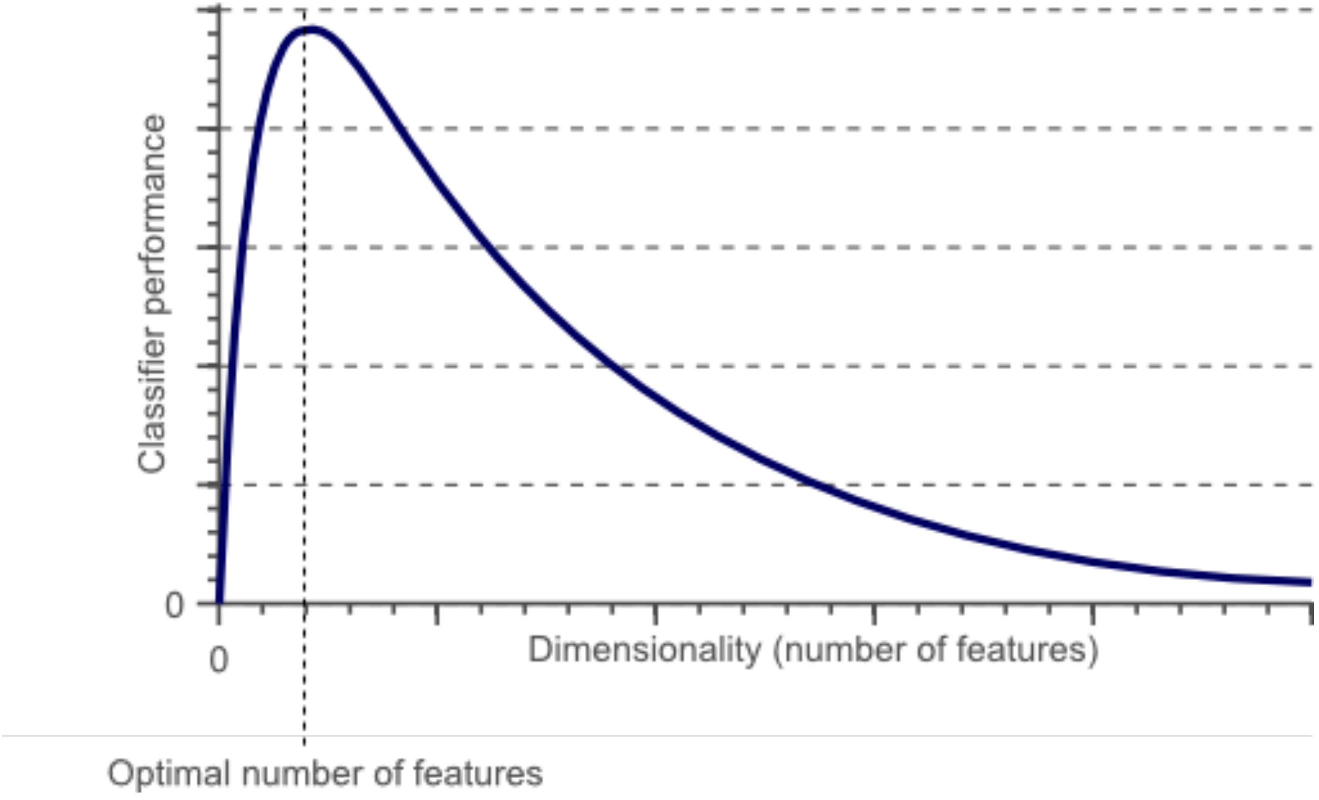


Not easy to figure out the trend in 3D

How can we find lower dimensional representation that keeps the most information about the original data?

Curse of Dimensionality

- As the number of dimensions in the data increases it impacts:
  - Sparsity
  - Sample Size
  - Metrics
  - Performance

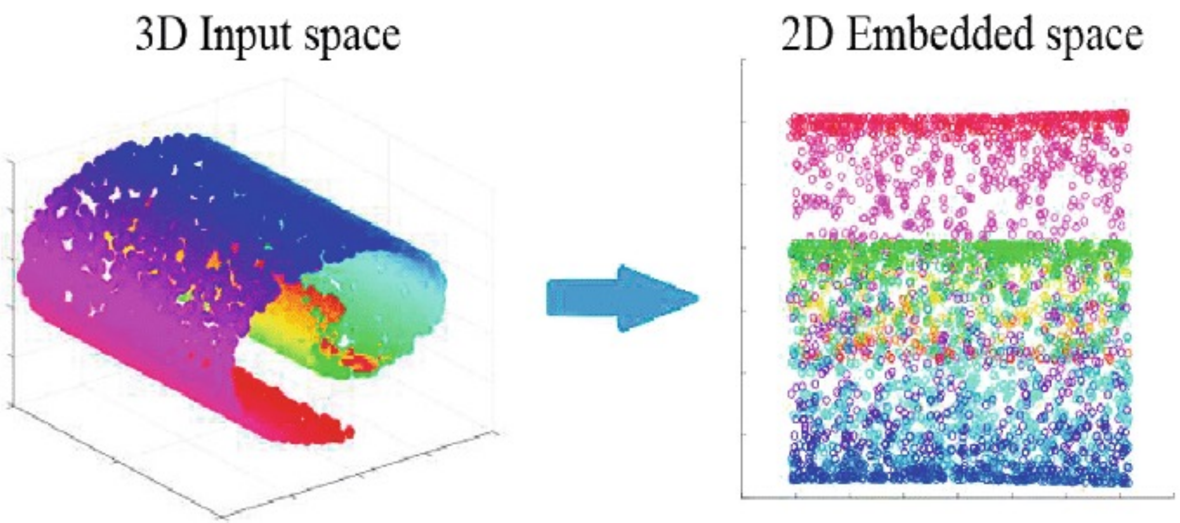


Dimensionality Reduction

- Goal
  - Simplify data understanding numerically or visually without loss of data integrity.
- Objectives
  - Reduce the number of variables
  - Examine the relationship between variables
  - Address the problem of multicollinearity
- Key Ideas
  - Exploit redundancy in the data to find a lower dimensional representation that preserves distances.

Dimensionality Reduction

$$X = \{x_{1,} x_{2,}, \ldots, x_{n,} \in \mathbb{R}^{\mathrm{D}}\} \rightarrow Y = \{y_{1,} y_{2,}, \ldots, y_{n,} \in \mathbb{R}^{\mathrm{M}}\}$$



Dimensionality Reduction: Methods

Dimensionality Reduction Methods

- **Feature Selection**
  - Missing Value Ration
  - Low Variance Filter
  - High Correlation Filter
  - Backward Feature Selection
  - Forward Feature Selection
- **Dimensionality Reduction**
  - **Factor based**
    - Factor Analysis
    - Principal Component Analysis
    - Independent Component Analysis
  - **Projection based**
    - ISOMAP
    - t-SNE
    - UMAP

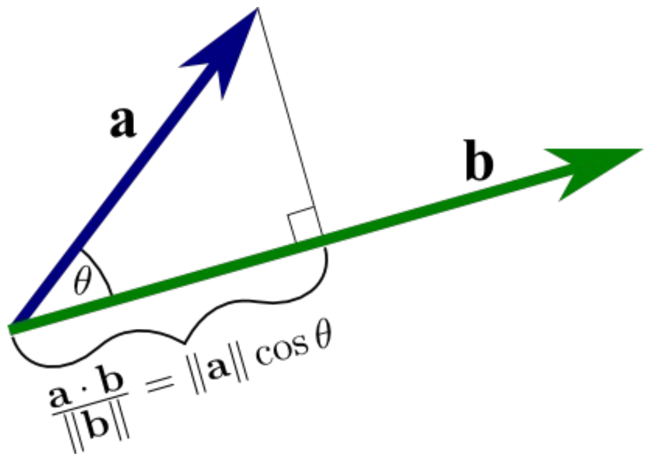
MATHEMATICAL FOUNDATIONS

Dot Products

- Dot product between two vectors is based on the projection of one vector onto another

Angle between the vectors :

is obtuse if the dot product is < 0
is acute if the dot product is > 0
is orthogonal if the dot product = 0



Orthogonality

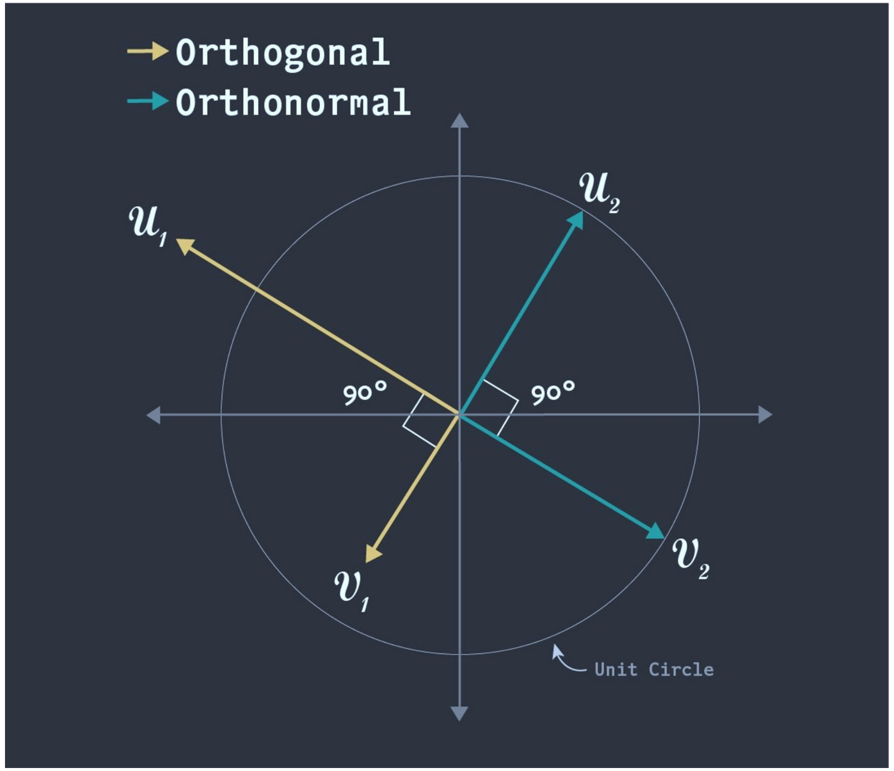
- u and v are orthogonal $u \perp v$ when the angle between them is 90°

$$\langle u, v \rangle = 0$$

- u and v are orthonormal if they are orthogonal, and each vector has unit length.

$$\langle u, u \rangle = \langle v, v \rangle = 1$$

- Orthogonality (and orthonormality) is necessary to project vectors onto subspaces



Variance, Covariance

- Variance measures the variation of a single random variable

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Covariance is a measure of how much two random variables vary together

$$\sigma(x,y) = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance Matrix

- A square symmetric matrix given by $C_{i,j} = \sigma(x_i, x_j)$ where our data set is expressed by the matrix $X \in \mathbb{R}^{n \times d}$

$$C = \frac{1}{n-1} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$$

Eigenvectors and Eigenvalues

- A vector $\boldsymbol{v}$ of dimension N is an eigenvector of a square $N \times N$ matrix $\boldsymbol{A}$ and λ is the corresponding eigenvalue if

$$\boldsymbol{Av} = \lambda \boldsymbol{v}$$

$$\boldsymbol{Av} = \lambda \boldsymbol{Iv}$$

$$\boldsymbol{Av} - \lambda \boldsymbol{Iv} = 0$$

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{v} = 0$$

$$det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

Eigen-decomposition

- Factorization of a matrix into a canonical form, whereby the matrix is represented in terms of its eigenvalues and eigenvectors
- If a square matrix A is diagonalizable, then there is a matrix P such that

$$A = P\,D\,P^{-1}$$

Original Matrix	Eigenvectors Matrix	Eigenvalues Matrix	Inverse of Eigenvectors Matrix

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

- The nondiagonal matrices P and P^{-1} are inverses of each other

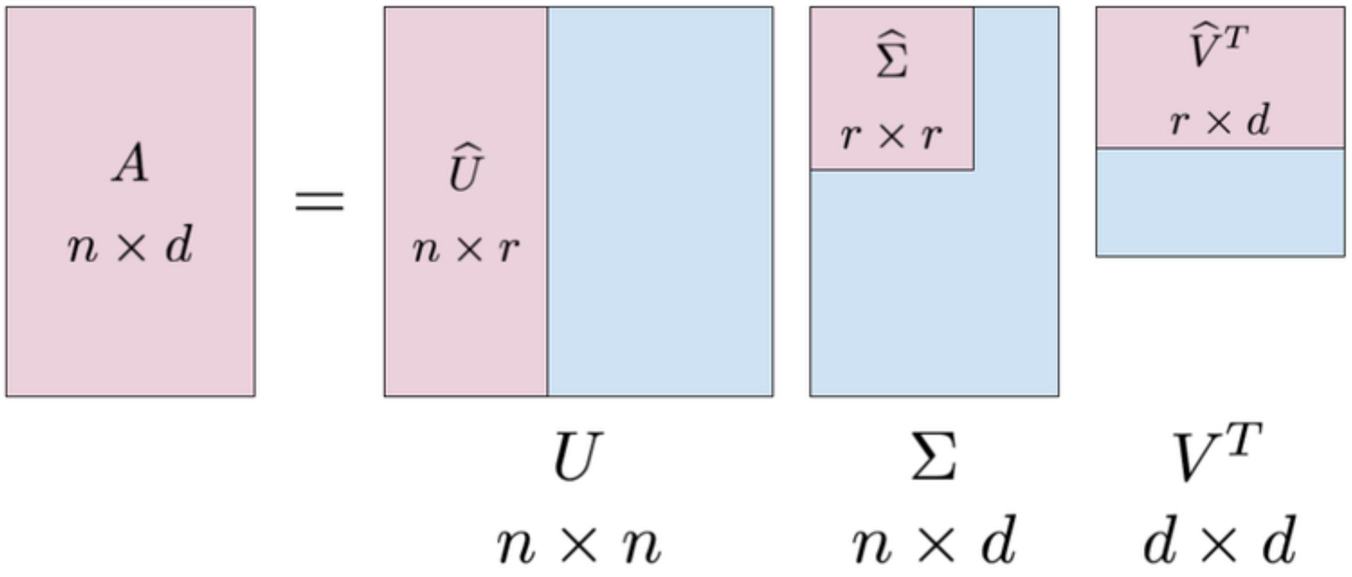
Singular Value Decomposition (SVD)

- Factorization of that matrix into three matrices given by the formula :

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

- Vectors in the matrices U and V in the SVD are orthonormal and not necessarily the inverse of one another
- SVD can be used to compute optimal low-rank approximations of arbitrary matrices.
- SVD always exists for any rectangular or square matrix

Singular Value Decomposition (SVD)

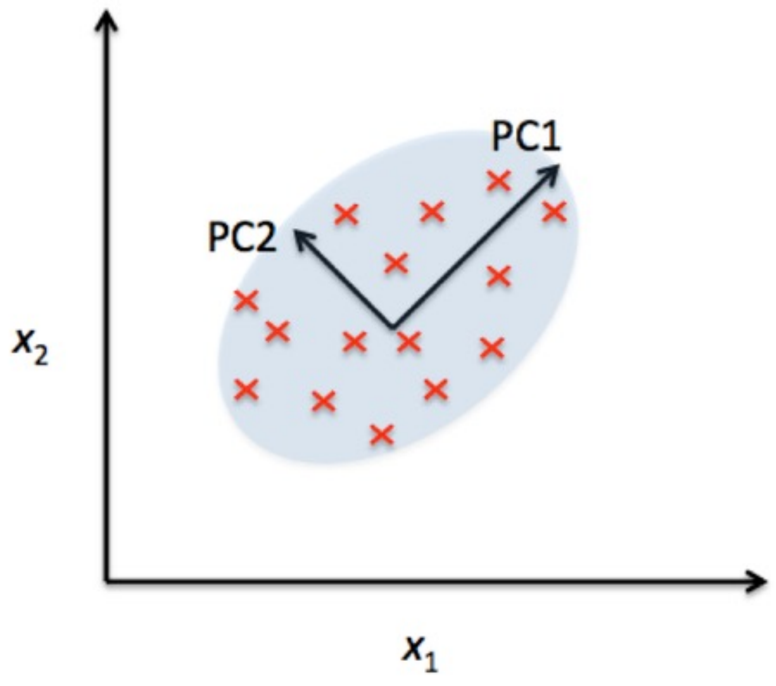


U: $n \times n$ matrix of the orthonormal eigenvectors of AA^T.
Σ: $n \times d$ diagonal matrix of the singular values of A which are the square roots of the eigenvalues of A^TA. The number of non-zero singular values is the rank of A
V^T: transpose of $d \times d$ matrix containing the orthonormal eigenvectors of A^TA

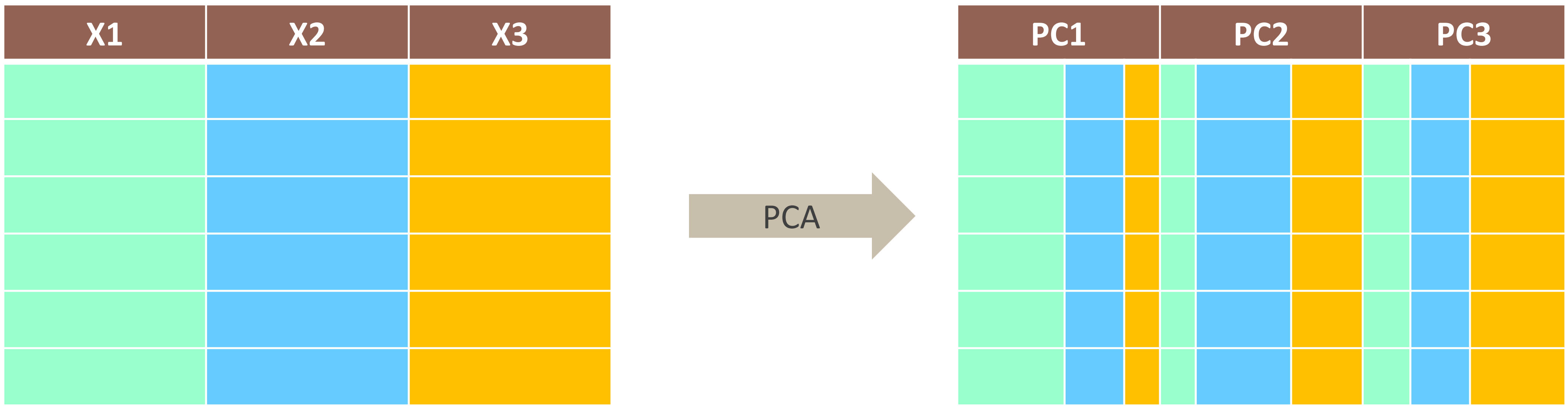
PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA)

- PCA
  - Finds a lower-dimensional representation by constructing new features - Principal Components (PCs) which are linear combinations of the original features
- Assumptions
  - Original variables should be normalized
  - Factors are independent of each other
  - There exist some underlying factors that can describe the original variables
- Approach
  - Projecting (dot product) the original data into the reduced PCA space using the eigenvectors of the covariance matrix (i.e., PCs)



PCA: Linear Method



$$PC1 = a_1x_1 + a_2x_2 + a_3\ x_3$$

$$PC2 = b_1x_1 + b_2x_2 + b_3\ x_3$$

$$PC3 = c_1x_1 + c_2x_2 + c_3\ x_3$$

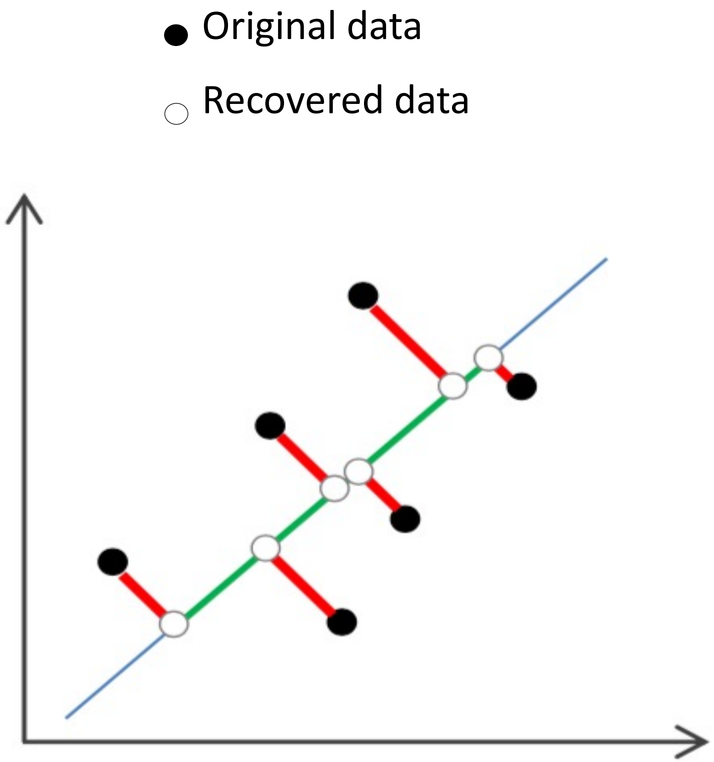
PCA: Solution Approach

- Let’s say we have an i.i.d dataset

$$X = \{x_{1,} x_{2,}, \dots, x_{n,}\} \in \mathbb{R}^{\mathrm{D}}$$

with mean value of 0

- Goal is to find projections that are as similar to the original data as possible but have lower dimensionality $(M < D)$.
- There are two approaches:
  1. Maximum variance
  2. Minimum error



PCA: Solution Approach

$$\underbrace{Variance\ of\ data}_{\text{fixed}} = \underbrace{captured\ variance}_{\text{maximize}} + \underbrace{reconstruction\ error}_{\text{minimize}}$$

- Maximum variance formulation
  - Find a low-dimensional representation which maximizes the variance of the projected data.

$$\max_{X \in \mathbb{R}^{m \times p}} \|AX\|^2 \text{ subject to } X^T X = I$$

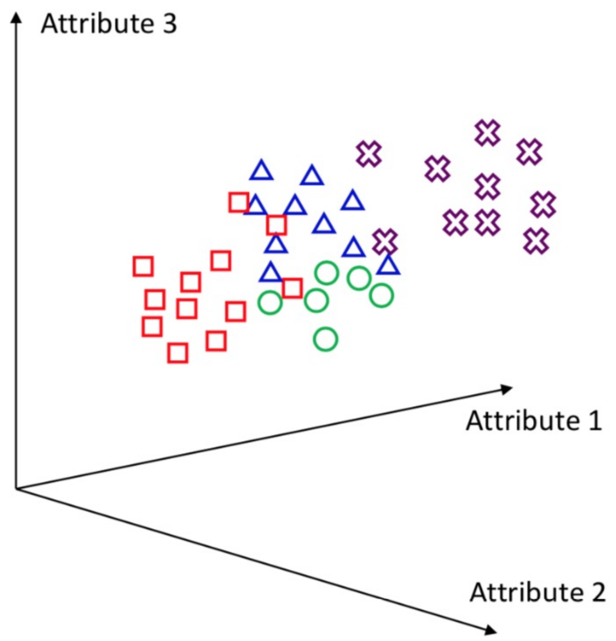
- Minimum error formulation
  - Find a low-dimensional representation which minimizes the average reconstruction error between the original data and the reconstructed data.

$$\min_{X \in \mathbb{R}^{m \times p}} \|A - AXX^T\|^2 \text{ subject to } X^T X = I$$

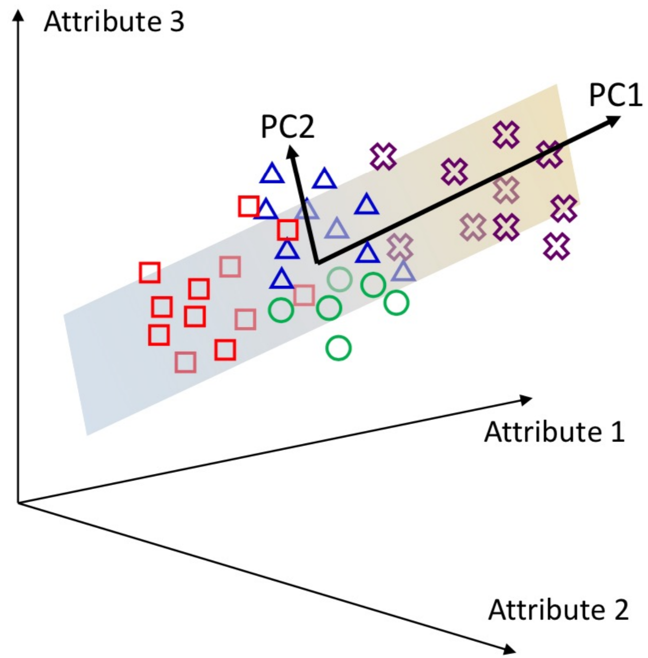
PCA: Steps

1. Standardize the data
2. Compute the Covariance Matrix C
3. Eigenvalue Decomposition
4. Sort Eigenvalues in descending order and arrange corresponding Eigenvectors
5. Select Principal Components
6. Form Principal Component Matrix
7. Transform Original Data

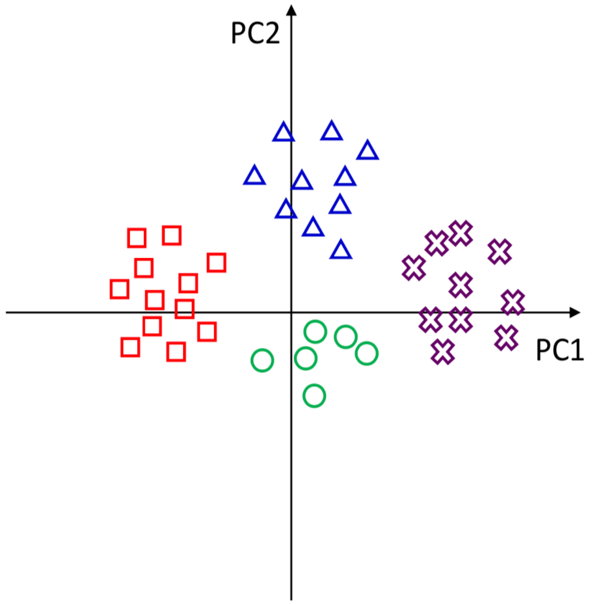
PCA: Data Visualization



1 Data in 3D



2 Principal Components



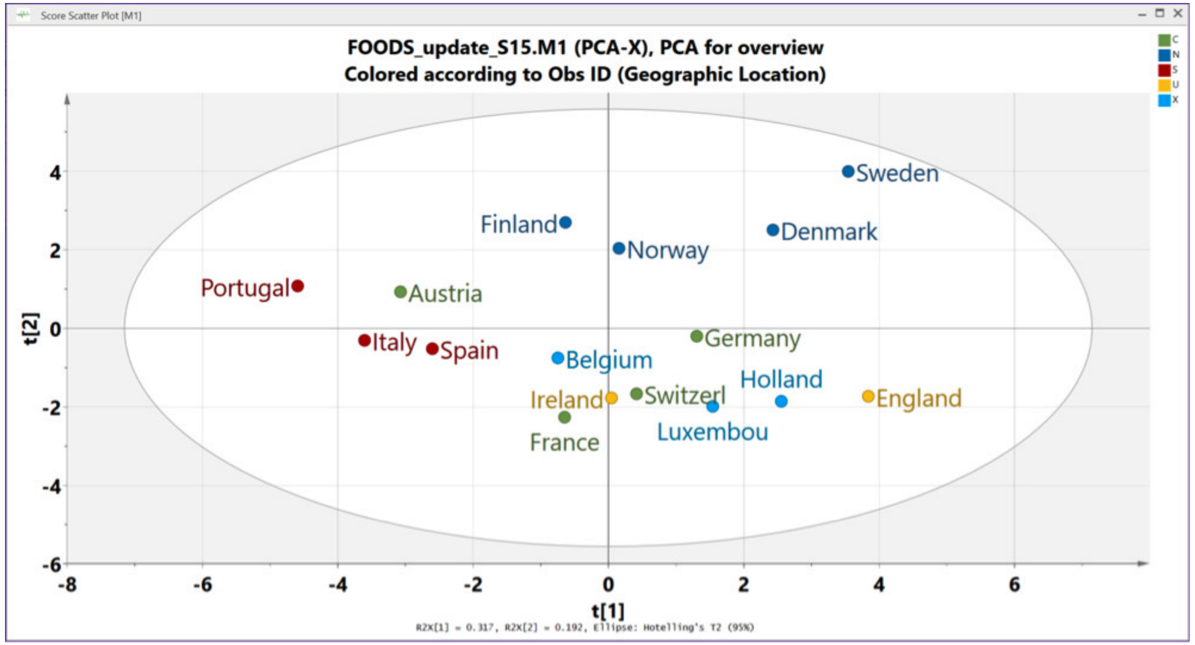
3 Data Visualization

PCA: Components

- For N original dimensions, sample covariance matrix is $N \times N$, and has up to N eigenvectors. So, N PCs.
- We can ignore the components of lesser significance
- Some information is lost, but if the eigenvalues are small, you don't lose much
  - N dimensions in original data
  - Calculate N eigenvectors and eigenvalues
  - Choose only the first D eigenvectors, based on their eigenvalues
  - Final data set has only D dimensions

PCA: Applications

- Data visualization
- Data compression (Lossy)
- Noise reduction
- Factor analysis
- Feature extraction
  - High dimensionality of the input features
  - Applied to data having multi-collinearity between the features/variables



Example 1

- Covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

- Eigenvalue-eigenvector pairs are

$$\begin{aligned} \lambda_1 &= 5.83, \quad x_1 = (0.383, -0.924, 0) \\ \lambda_2 &= 2, \quad x_2 = (0,0,1) \\ \lambda_3 &= 0.17, \quad x_3 = (0.924, 0.383, 0) \end{aligned}$$

  - $\lambda_1 > \lambda_2 > \lambda_3 \Rightarrow$ order of importance: x_1, x_2, x_3
- Dimensionality reduction
  - Pick eigenvectors (PC) with largest p eigenvalues
  - If $p = 1$, pick x_1
  - If $p = 2$, pick x_1 and x_2

Example 1 - How Many PCs?

- Comparison of recovered matrices with $p = 1,2$ and original matrix

Original data

0.8	4.4	-0.9
5.8	12.4	6.1
-3.2	-14.6	-5.9
-6.2	-15.6	-1.9
-2.2	-8.6	2.1
1.8	8.4	0.1
4.8	8.4	5.1
1.8	13.4	6.1
-3.2	-12.6	-6.9
-4.2	-10.6	-7.9
2.8	19.4	6.1
-0.2	1.4	2.1
1.8	-5.6	-3.9

Recovered data with $p = 2$

1.2	4.3	-0.9
3.4	13.0	6.0
-3.8	-14.4	-5.9
-4.4	-16.1	-1.9
-2.4	-8.6	2.1
2.3	8.2	0.1
2.4	9.1	5.0
3.4	12.9	6.1
-3.3	-12.6	-6.9
-2.8	-11.0	-7.9
5.0	18.8	6.1
0.3	1.2	2.1
-1.2	-4.8	-4.0

Recovered data with $p = 1$

0.9	3.5	1.3
3.6	13.4	5.0
-3.9	-14.6	-5.5
-4.0	-14.9	-5.6
-1.9	-7.0	-2.6
2.0	7.3	2.7
2.5	9.5	3.6
3.6	13.3	5.0
-3.5	-13.3	-5.0
-3.2	-12.1	-4.5
4.9	18.5	6.9
0.5	1.7	0.6
-1.5	-5.5	-2.0

Example 1 - How Many PCs?

- Comparison of recovered matrices with $p = 1,2$ and original matrix

Original data

0.8	4.4	-0.9
5.8	12.4	6.1
-3.2	-14.6	-5.9
-6.2	-15.6	-1.9
-2.2	-8.6	2.1
1.8	8.4	0.1
4.8	8.4	5.1
1.8	13.4	6.1
-3.2	-12.6	-6.9
-4.2	-10.6	-7.9
2.8	19.4	6.1
-0.2	1.4	2.1
1.8	-5.6	-3.9

Error matrix with $p = 2$

-0.4	0.1	0.0
2.3	-0.6	0.1
0.6	-0.2	0.0
-1.8	0.5	-0.1
0.2	-0.1	0.0
-0.5	0.1	0.0
2.4	-0.7	0.1
-1.6	0.5	0.0
0.1	0.0	0.0
-1.4	0.4	0.0
-2.3	0.6	-0.1
-0.5	0.1	0.0
3.0	-0.8	0.1

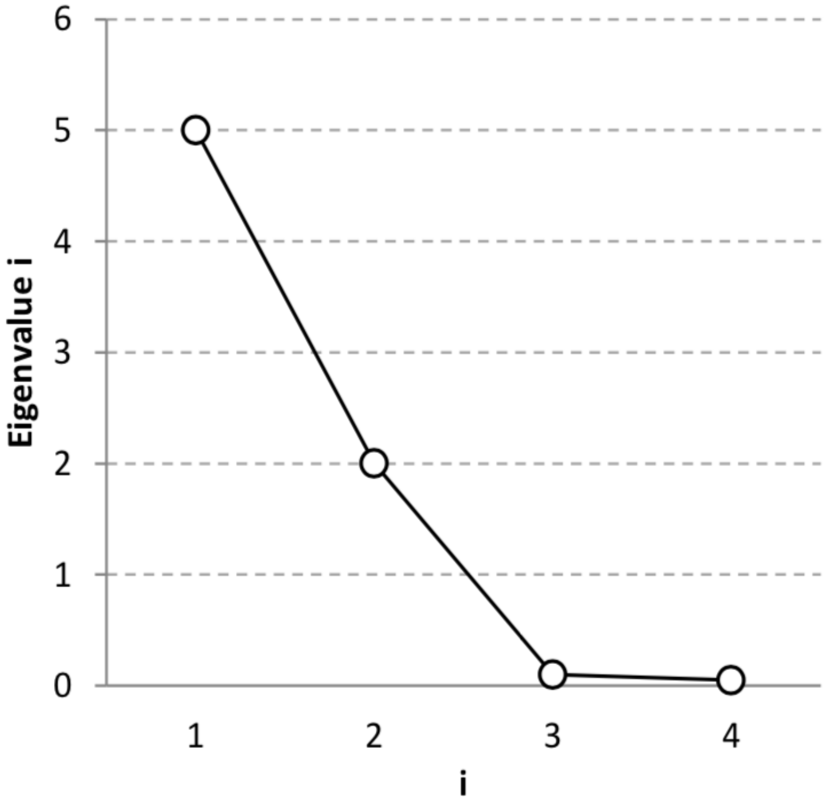
Error with $p = 1$

0.2	-0.9	2.2
-2.2	1.0	-1.1
-0.7	0.0	0.5
2.3	0.8	-3.6
0.4	1.7	-4.7
0.2	-1.0	2.7
-2.2	1.2	-1.5
1.8	-0.1	-1.1
-0.3	-0.6	2.0
1.0	-1.5	3.4
2.2	-0.9	0.8
0.7	0.3	-1.4
-3.2	0.2	1.9

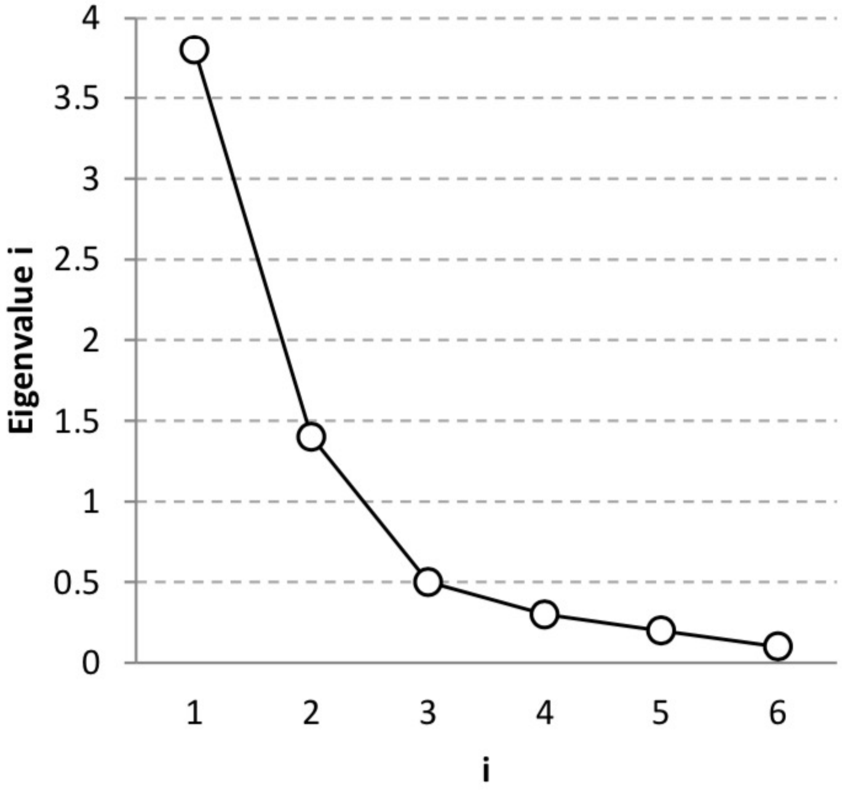
Example 1 - Number of PCs

- How many principal components?
  - There is no definitive answer
- Scree plot: a popular visual aid
  - Larger eigenvalues ⇒ more important eigenvectors
  - Small eigenvalues may be ignored without loss of important information
- Scree plot
  - Plot of λ_i versus i (sorted)
- To determine appropriate number of components (p), look for an elbow

Scree Plot



$p = 2$ *is appropriate*



$p = 2$ may be appropriate

Eigenvalues are a measure of the amount of variance accounted for by a factor

Example 2 - Iris Dataset

- Number of observations: 150
- Number of attributes: 4 numeric, predictive attributes and the class
- Attribute Information:
  - Sepal length in cm
  - Sepal width in cm
  - Petal length in cm
  - Petal width in cm

Iris Setosa

Iris Versicolour

Iris Virginica

- Classes: Iris Setosa, Iris Versicolour, Iris Virginica

Step 1: Calculate Covariance Matrix

- $A \in \mathbb{R}^{150\times 4}$ (excluding class attribute)
- $\Sigma = \frac{1}{n} A^T A$ (assuming columns of A have zero mean)

$$\begin{array}{ccccc} & Sepal.Length & Sepal.Width & Petal.Length & Petal.Width \end{array}$$

- $$\Sigma = \begin{array}{c} Sepal.Length \\ Sepal.Width \\ Petal.Length \\ Petal.Width \end{array} \begin{pmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0424 & 0.1899 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{pmatrix}$$

Step 2: Calculate Eigenvalues and Eigenvectors

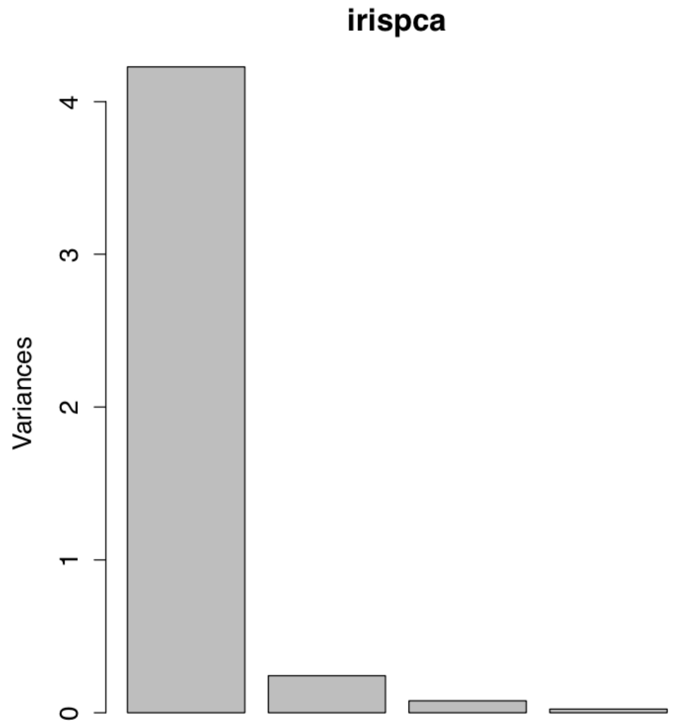
- $\text{Eigenvalues} = (4.2282, 0.2427, 0.0782, 0.0238)$

- $$\Sigma = \begin{array}{c c} & \begin{array}{cccc} PC1 & PC2 & PC3 & PC4 \end{array} \\ \begin{array}{c} Sepal.Length \\ Sepal.Width \\ Petal.Length \\ Petal.Width \end{array} & \begin{pmatrix} 0.3614 & -0.6566 & -0.5820 & 0.3155 \\ -0.0845 & -0.7302 & 0.5979 & -0.3197 \\ 0.85671 & 0.1734 & 0.0762 & -0.4798 \\ 0.3583 & 0.0755 & 0.5458 & 0.7537 \end{pmatrix} \end{array}$$

- The first principal component is the most important (largest eigenvalue), while others are not very significant.

Step 3: Scree plot

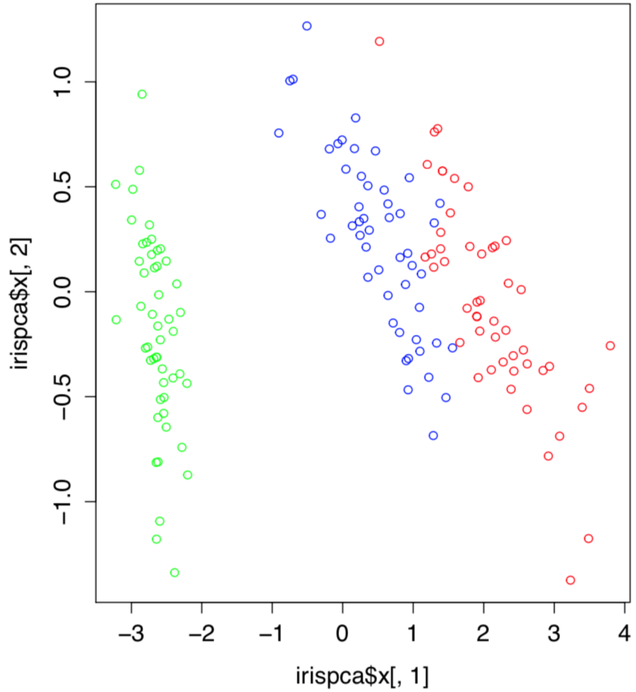
- Eigenvalues = (4.2282,0.2427,0.0782,0.0238)



- Confirm that using only one or two components is enough!

Step 4: Projection to Smaller Dimension

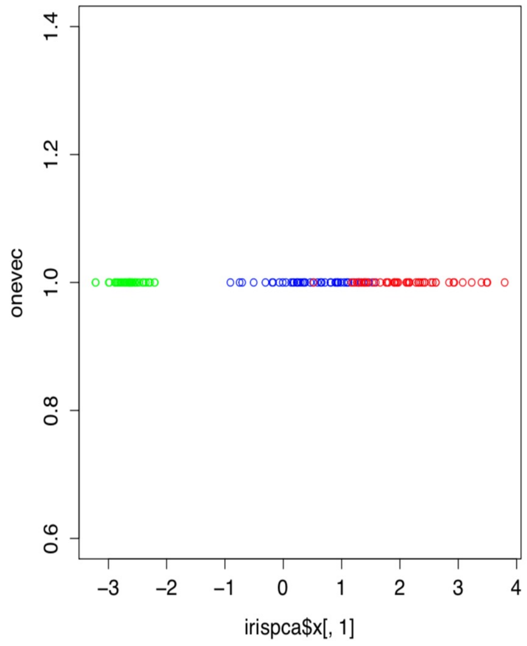
- With $p = 2$, projected data is obtained by $Y = AX \in \mathbb{R}^{150\times 2}$
- Visualize Y with class attribute (different class in different colors)



Green = setosa, blue = versicolour, red = virginica

Step 4: Projection to Smaller Dimension

- With $p = 1$, projected data is obtained by $Y = AX \in \mathbb{R}^{150\times 1}$
- Visualize Y with class attribute (different class in different colors)



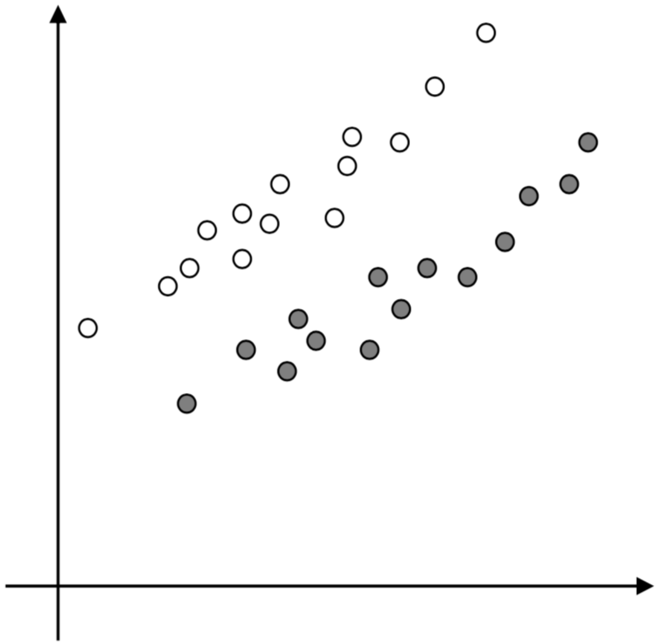
Green = setosa, blue = versicolour, red = virginica

PCA: Limitations

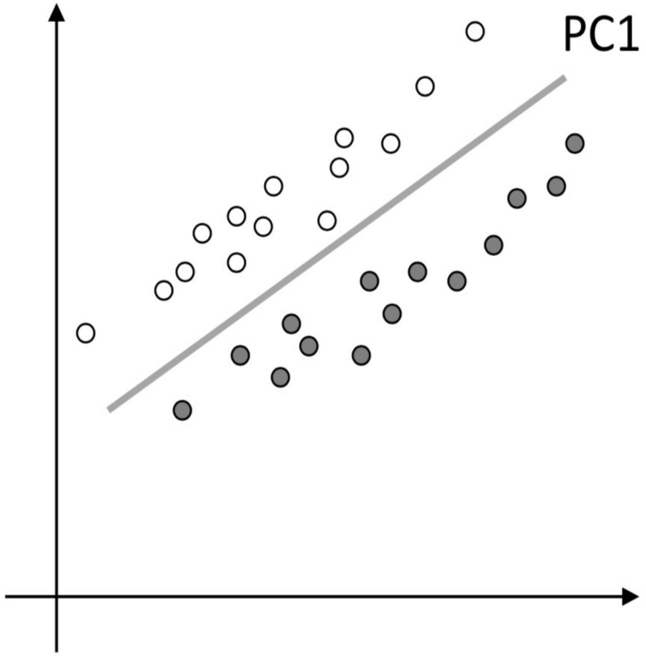
- Assumes a linear relationship between features i.e., it cannot capture non-linear structure in the data (as in many real-world applications).
- Assumes a correlation between features.
- Sensitive to the scale of the features
- Not robust against outliers
- Low interpretability of principal components.
- Trade-off between information loss and dimensionality reduction
- Technical implementations often assume no missing values

Issues: Data with Labels

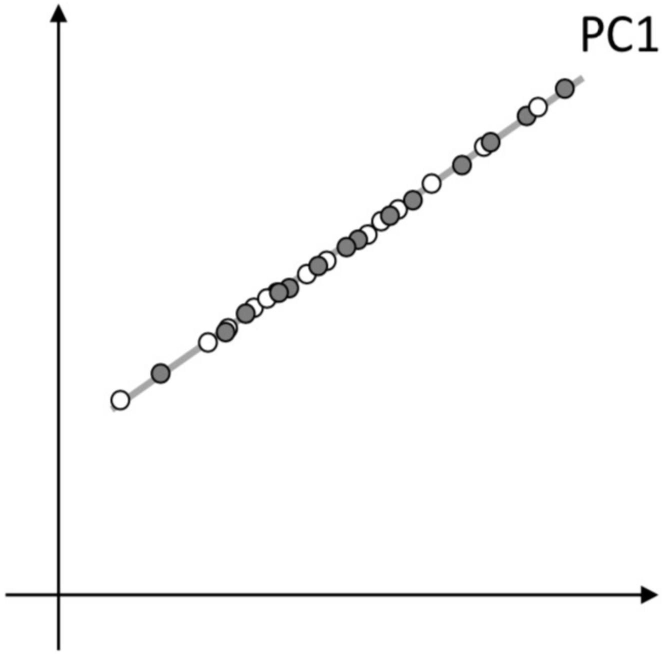
- A problematic example



Data with labels (white and gray)



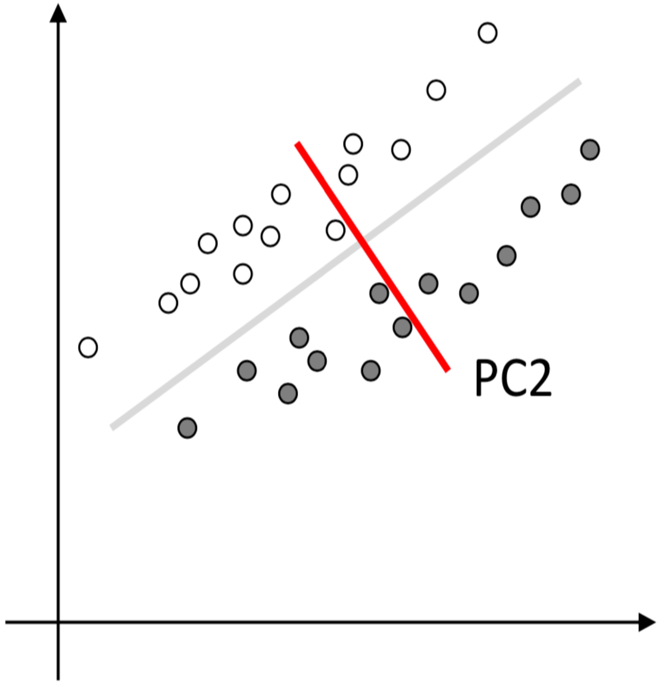
The first PC that explains most of the variance



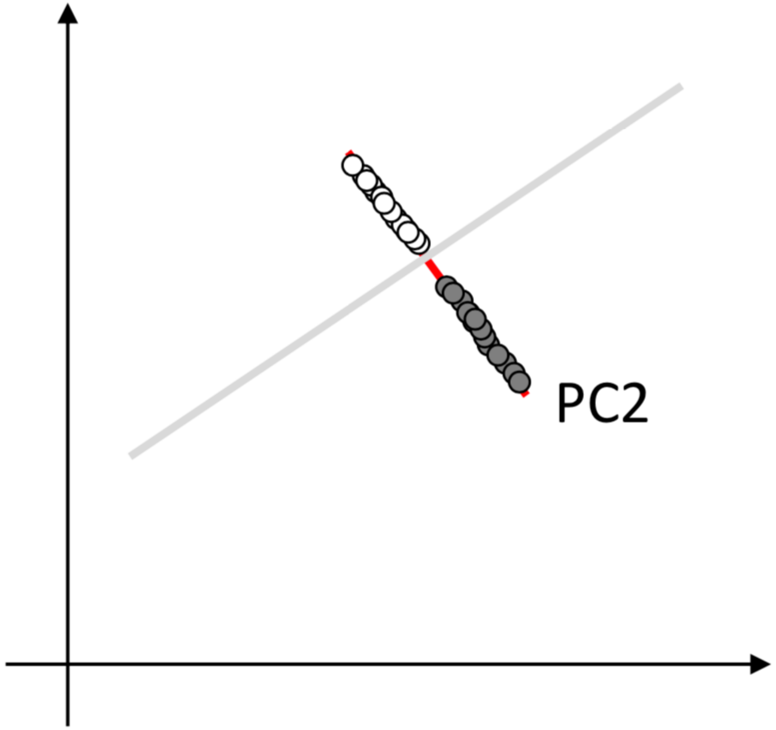
After projection, the result is useless

Issues: Data with Labels

- A problematic example



In fact, if we use the second PC, we obtain a better result



In fact, if we use the second PC, we obtain a better result