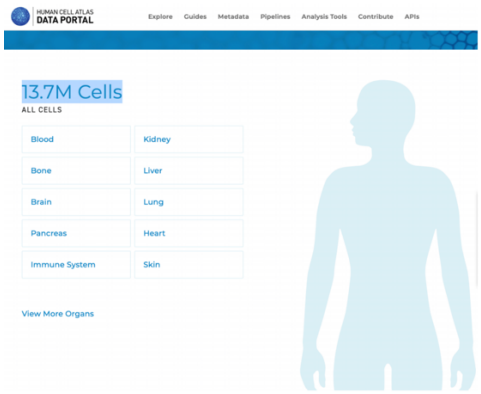
BMEG 3105 Fall 2022

Data analytics for personalized genomics and precision medicine
Feature selection & dimension reduction

Lecturer: Yu LI (李煜) from CSE, Writer: Chan Tung Man 1155142328

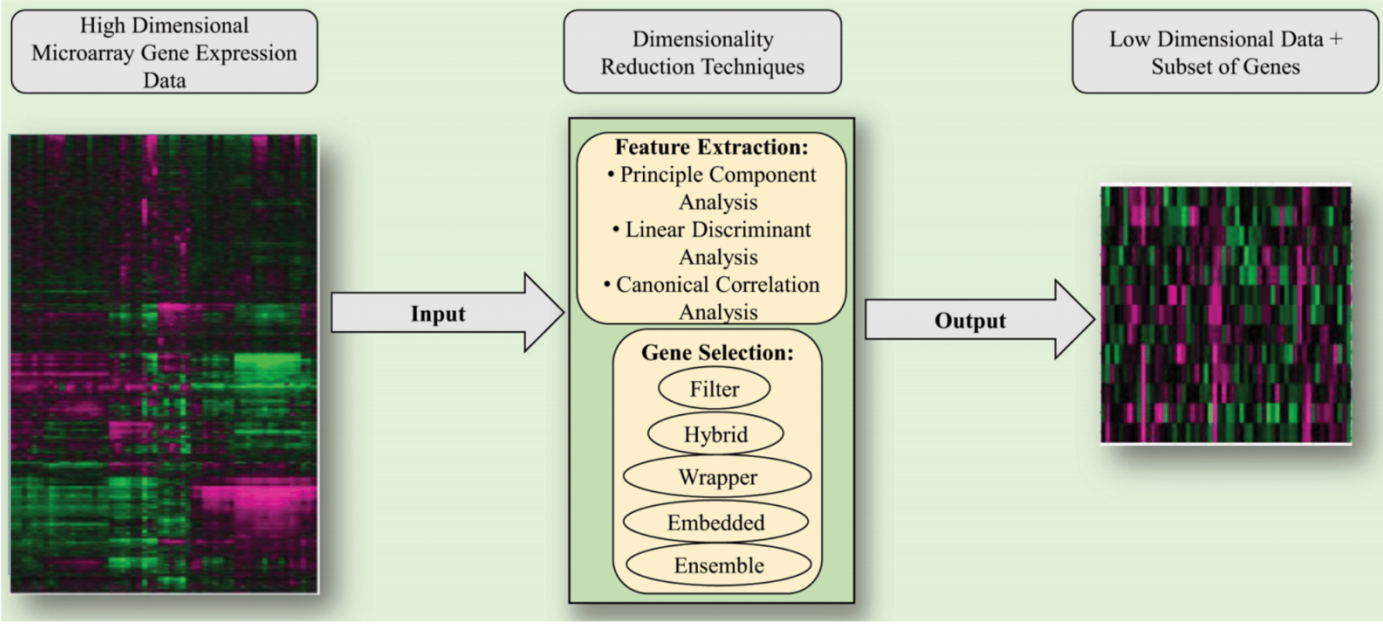
Liyu95.com, liyu@cse.cuhk.edu.hk Wednesday, 11 October

The reason of feature selection & dimension reduction

- The bio-data always very huge, require lots of storage to store the data.
  - For example:
    - Gene expression profile: 25000 gene (features) * 13.7 million cells (cells/ data point) = 1.2 TB!
    - 

  - For common: normally deal with 10000 cells with 25000 gene.
    - 25000 * 10000 = 875 MB
- We don't need to deal with all the data, some of the data is useless.
  - For example:
    - Irrelevant genes: we don't have to include them in our analysis.
    - Highly correlated genes: we don't have to include all of them.
    - Some genes are complementary: combine two values into one value may be more useful.
- Reduce the data size, more useful and friendly to use the data set.

⇨ Therefore, this procedure is called the feature (gene) selection and dimension reduction.



The Left picture can clearly see that the data matrix/ data dimension is huge as compare with the right data matrix.

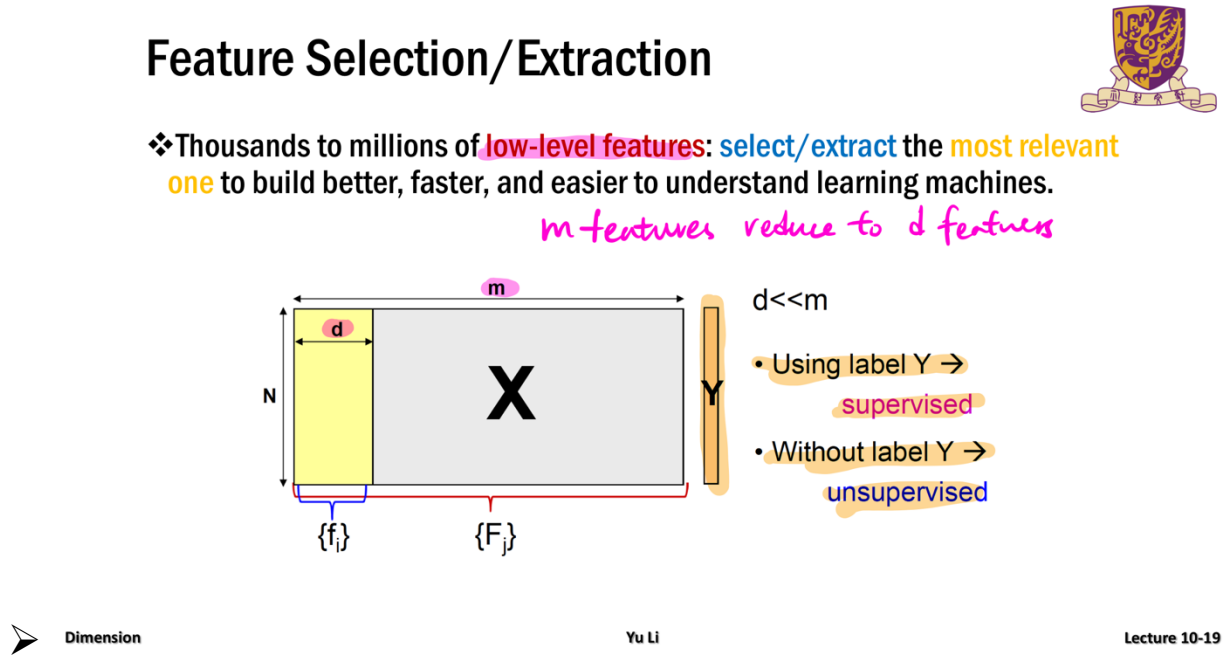
The color in right/ output matrix is more regularly with smaller dimension represent some useless data points are removed.

The middle part represents the reduction techniques and algorithm, based on this method for example PCA with different gene selection. The input data matrix can dimension into output matrix.

Benefit

- Data compression
  - ✓ Efficient storage and retrieval
    - ❖ Reduce the size of data matrix and Irrelevant data.
- Improve prediction performance.
  - ✓ Remove unrelated inputs.
    - ❖ Delete useless and irrelevant data.
- Understand the prediction results.
  - ✓ What gene are related to the cancer prediction?
- Facilitate data visualization.
  - ✓ 25000D to 2D, understand the distance between cells visually.

Feature selection



1. M numbers of features reduced to d number of features.
   - ✧ m >> d
2. using label Y -> supervised / without label Y -> unsupervised
3. N the number of cell/ data pt.

How to reduce dimensionality

- ➢ Feature selection

Choose the best subset genes from all the genes:

1. Feature ranking
2. Filter/ wrapper

- ➢ Feature extraction

Extract the feature by linear or non-linear combination.

- ✓ New gene = gene 1 + gen2

Feature ranking

- ➢ Discover the most relevant features and build a better, faster, and easier machines.
- ➢ Measurement:

1. Correlation between feature and class
   - ✓ If highly related to class, more useful feature

- ✓ Example: weight vs gender = 0.714, height vs gender 0.812
  - ❖ Height is more related to class.

2. Mutual information I(i)

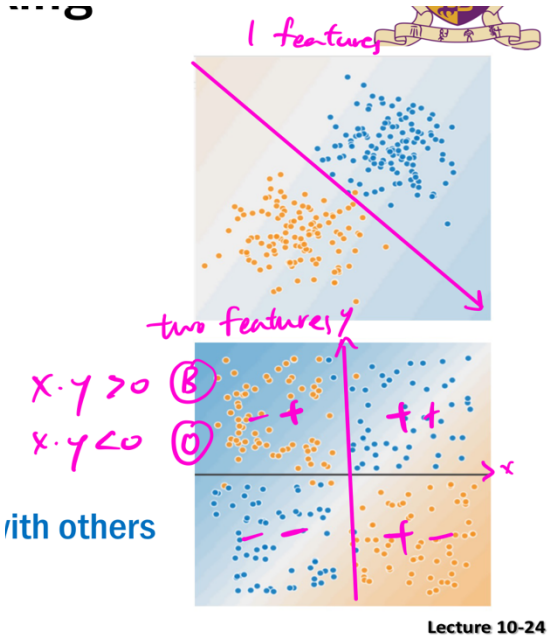
- ✓ The higher I(i), the attribute is more related to the class.

3. Fisher score F

- ✓ The higher F, the attribute is more related to the class.

Problems and issues of individual features ranking.

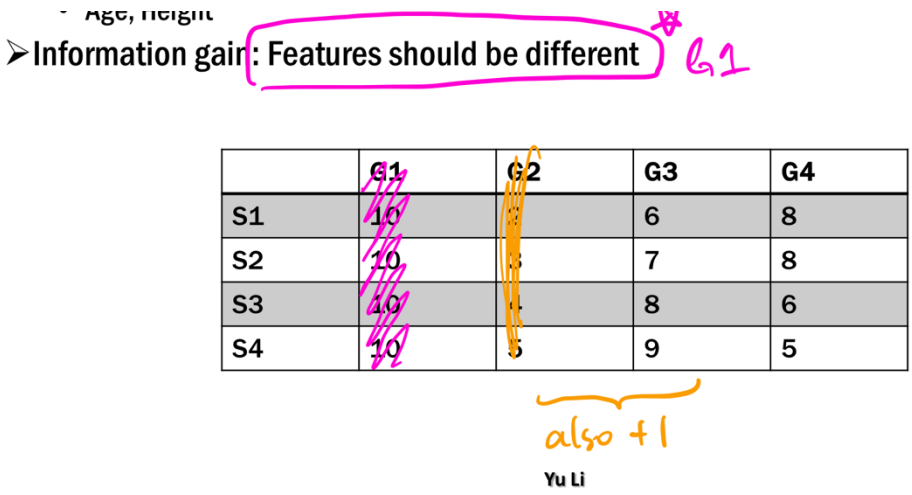
- ➢ Relevance vs usefulness:
  - ✓ They are not the same and not directly relation.
- ➢ Selection of a redundant subset
  - ✓ K best features ! = best k features
    - ❖ Best two features != the best combination of the features
  - ✓ Example: ranking Major (3) ---> weight (2) ---> Height (1)
    - ❖ Height & weight <<< height & major
- ➢ A variable that is useless by itself can be useful with others.
  - ✓ Salary: occupation +age
    - ❖ Only age is no correlate with salary but combine with occupation, it has the meaning.



- ➢
- ➢ two features have two axis to separate in 4 areas.

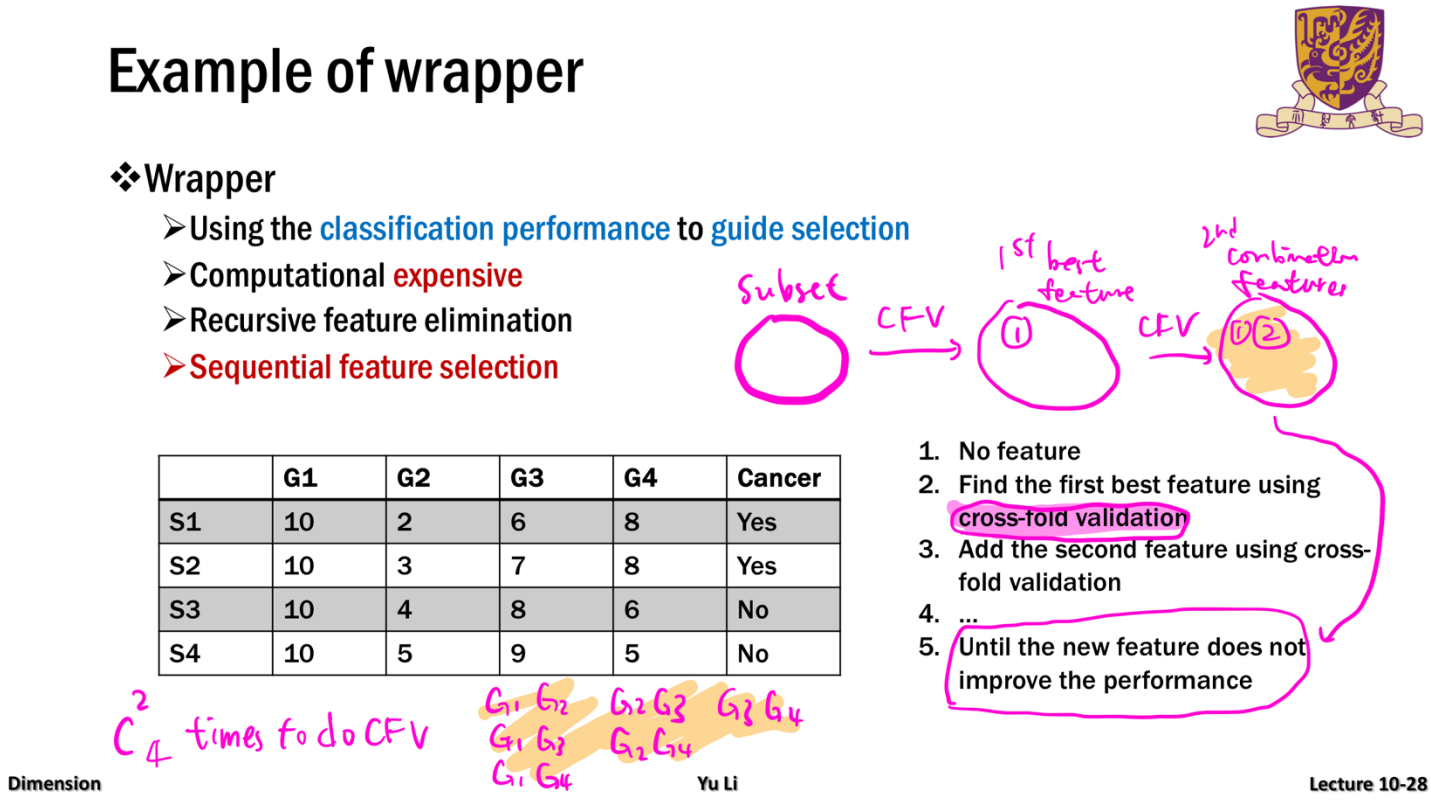
Filter

- classification performance is not involved in the selection loop.
- variance data required -> age, height, weight.
- Higher variance with more useful information, take a variance threshold to filter our data.



- Pink: delete the gene with same data pt. (information gain)
- Orange: G2 and G3 both increment is 1, choose either one is enough (variance)

Wrapper



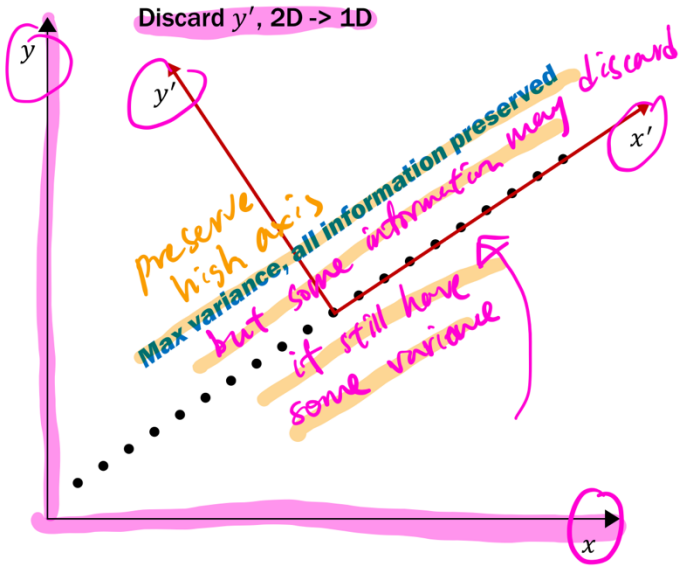
- Use cross-fold validation to update the best feature combination.

Dimension reduction

Principal Components analysis (PCA)

Principal components analysis (PCA)

high variance to low variance



A two dimensional scatter of points that show a high degree of correlation

We care about variance (information) and distance

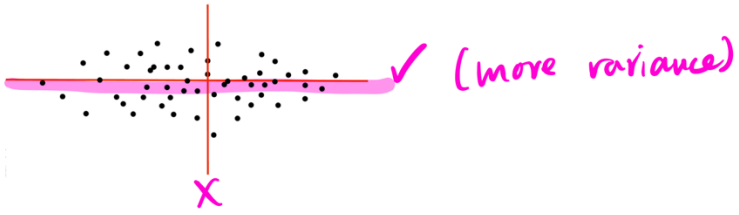
Person	Height (m)	Weight (kg)	Age	Gender
P1	1.79	75	20	M
P2	1.64	54	20	F
P3	1.70	63	20	M
P4	1.88	78	20	M
P5	1.75	70	20	??

Dimension Yu Li Lecture 10-34

- From 2D xy chart to 1D y' chart
- From high variance to low variance,
- Care about the variance and the distance between point and axis
- The scatter point are directly proportional = high correlation

Principal components analysis (PCA)

- After vector space transform, we have more "efficient" description
  - 1st dimension captures max variance
  - 2nd dimension captures the max amount of residual variance, at right angles (orthogonal) to the first
- The 1st dimension may capture so much of the information content in the original data set that we can ignore the remaining axis



Dimension Yu Li Lecture 10-36

- The 1st dimension/ axis is the horizontal one, more data point touch the axis, have more variance.

➢ The 2nd axis perpendicular to the 1st axis, less variance so we can ignore it.

How to do PCA?

How to do PCA?

❖Suppose we have a n by d data matrix, X. We first normalize each feature to make the average of each feature 0. Then, we get X'

❖Then, we calculate the covariance matrix of X'

➢$\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix

❖Find the eigenvectors and eigenvalues of Σ

❖M eigenvectors with the M largest eigenvalues

➢Principal components

❖Project the data to the M eigenvectors' direction

➢$\hat{X} = X'P$

Person	Height (m)	Weight (kg)	Age	Gender
P1	1.79	75	20	M
P2	1.64	54	20	F
P3	1.70	63	20	M
P4	1.88	78	20	M
P5	1.75	70	20	??

Dimension Yu Li Lecture 10-37

1. Normalization
2. Covariance matrix
3. Eigenvectors and eigenvalues
4. Largest eigenvalues
5. Project the data to the eigenvector of its direction

Example:

X	A	B	C
X1	1	1	1
X2	2	2	2
X3	3	3	3

1. Calculate average of each feature and do Normalization:

Average = 1 + 2 + 3 /3 =2

X'	A	B	C
X1	1-2 =-1	1-2 =-1	1-2 =-1
X2	2-2 =0	2-2 =0	2-2 =0

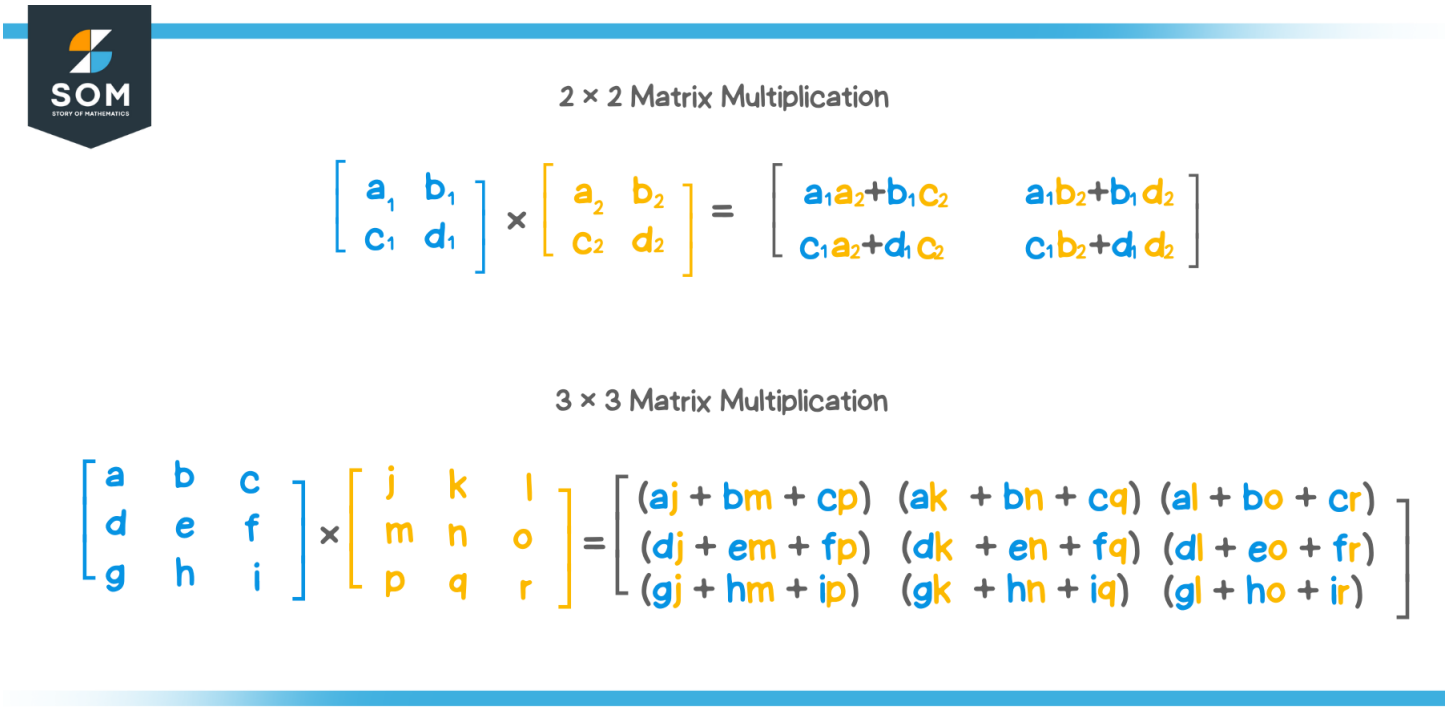
X3	3-2 =1	3-2 =1	3-2 =1

2. Covariance matrix

$$\Sigma = \frac{1}{n-1}(X'^{T})X'$$

$$X'^{T} = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

$$X' = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$



$$X'^{T}X' = \begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

$$\Sigma = \frac{1}{n-1}(X'^{T})X' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

3. Eigenvectors and eigenvalues

Example of PCA

X'

X1	-1	-1	-1
X2	0	0	0
X3	1	1	1

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- We first normalize each feature to make the average of each feature 0. Then, we get X'
- Then, we calculate the covariance matrix of X'
  - $\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix
- **Find the eigenvectors and eigenvalues of Σ**
- M eigenvectors with the M largest eigenvalues
  - Principal components
- Project the data to the M eigenvectors' direction
  - $\hat{X} = X'P$

$$\Sigma * V = \lambda * V$$

$$|\Sigma - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{vmatrix} = 0$$

$\lambda_1 = 3$
$\lambda_2 = 0$
$\lambda_3 = 0$.

$$\begin{vmatrix} x & x' & x'' \\ y & y' & y'' \\ z & z' & z'' \end{vmatrix} = xy'z'' + x'y''z + x''yz' - xy''z' - x'yz'' - x''y'z$$

$$(1-\lambda)^3 + 1 + 1 - (1-\lambda) - (1-\lambda) - (1-\lambda) = 0$$

$$\lambda = 3 \text{ or } \lambda = 0$$

eigenvalues

- $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- If eigenvalues = 0, represent No information!

Example of PCA

X'

X1	-1	-1	-1
X2	0	0	0
X3	1	1	1

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- We first normalize each feature to make the average of each feature 0. Then, we get X'
- Then, we calculate the covariance matrix of X'
  - $\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix
- Find the eigenvectors and eigenvalues of Σ
- **M eigenvectors with the M largest eigenvalues**
  - **Principal components**
- Project the data to the M eigenvectors' direction
  - $\hat{X} = X'P$

$$\lambda_1 = 3 \qquad V_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix}$$

preserve

$$\lambda_{2,3} = 0 \qquad V_{2,3} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

no information

4. Largest eigenvalues

Example of PCA

X'

X1	-1	-1	-1
X2	0	0	0
X3	1	1	1

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- We first normalize each feature to make the average of each feature 0. Then, we get X'
- Then, we calculate the covariance matrix of X'
  - $\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix
- Find the eigenvectors and eigenvalues of Σ
- M eigenvectors with the M largest eigenvalues
  - Principal components
- **Project the data to the M eigenvectors' direction**
  - $\hat{X} = X'P$

$$\lambda_1 = 3 \qquad V_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix} \qquad \lambda_{2,3} = 0 \qquad V_{2,3} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

2D $$P = \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} \qquad \hat{X} = X'P = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} = \begin{bmatrix} -\sqrt{3} & 0 \\ 0 & 0 \\ \sqrt{3} & 0 \end{bmatrix}$$

5. Project the data to the eigenvector of its direction

Example of PCA

X

X1	1	1	1
X2	2	2	2
X3	3	3	3

- We first normalize each feature to make the average of each feature 0. Then, we get X'
- Then, we calculate the covariance matrix of X'
  - $\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix
- Find the eigenvectors and eigenvalues of Σ
- M eigenvectors with the M largest eigenvalues
  - Principal components
- Project the data to the M eigenvectors' direction
  - $\hat{X} = X'P$

PCA

X1	$-\sqrt{3}$	0
X2	0	0
X3	$\sqrt{3}$	0

$\hat{X}$