04_Dimension_Reduction

4.0 A Quick View

What does it do?

Dimension Reduction:

- Reduces the dimension of data.
  - Changes the data representation into a lower-dimensional one.
- It preserves the structure of the data.
- Usually unsupervised.

Why do we need DR?

- Computation Complexity
- Pre-processing stage before further learning
- Data Visualization
- Data Interpretation

4.1 Singular Value Decomposition (SVD) 奇异值分解

4.1.0 Why SVD?

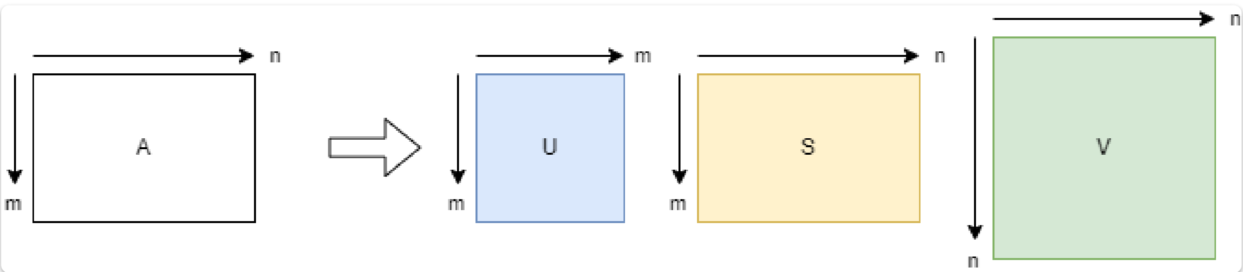
- *Redundancy* within dimensions of a single data sample. 多维间存在冗余信息
  - In a set of high-dimensional data samples, not all dimensions are useful.
  - There may be redundancies among some dimensions.
    - That is, some dimensions are highly related.
    - e.g., Suppose in a data set, for most data samples, $x_2 = 2x_1 + 3$. Therefore we only need x_1 since it could already describe x_2 with itself. This creates a redundancy.
  - SVD picks out main features, and project data into lower dimensions to remove such redundancies.
- Existence of *noise* samples. 存在噪声数据
  - Among data, smaller eigenvalues always comes with unimportant features.
  - By ignoring these data, we could reduce the noise when we are reducing data dimension.
  - That's why, during the process of SVD, we need to *sort* the eigenvalues.

4.1.1 Definition

- Suppose that matrix $A \in \mathbb{R}^{m \times n}$ contains a set of training data.
  - m: Dimensions within a data sample.
    - That is, a column vector of A represents a data sample.
  - n: The number of data samples.
  - In fact, the role of $m \times n$ could be reversed.
    - In the current version, A is a "fat" matrix; In the reversed version, A is a "tall" matrix.
- The Singular Value Decomposition process could be described as follows.

$$A_{m \times n} = U_{m \times m} S_{m \times n} V_{n \times n}^{\top}$$

where,
- A is any $m \times n$ matrix.
- U is any $m \times m$ orthogonal matrix. 酉矩阵、正交矩阵
  - $U^{\top} = U^{-1}$
  - $UU^{\top} = U^{\top}U = I$
- S is any $m \times n$ diagonal matrix. 对角矩阵
  - Singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_{\min(m,n)} > 0$ is the *main* diagonal of S.
    - $S = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m & \cdots & 0 \end{bmatrix}$
  - $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_{\min(m,n)^2}$ are the **eigenvalues** of $AA^{\top}$ and $A^{\top}A$.
- V is any $n \times n$ orthogonal matrix. 酉矩阵



Left Singular Matrix U

- When we look at the Left Singular Matrix U, we pay attention to its *Column Vectors*.
  - Since $U \in \mathbb{R}^{m \times m}$, it has m column vectors. They are the **Left Singular Vectors**.
  - These column vectors represents the *Main Directions* of the *Row Space* of matrix A.
    - Row space: The space of Row Vectors, consider the row number, i.e. the height of the matrix.
  - In other words, U denotes the relationships among the dimensions in data samples.
- How exactly?

- Each column vector of U is a unit vector, and U's column vectors are all perpendicular 垂直 to each other, with dot product of 0.
- Each column vector of U represents a *co-tendency* among all the features within a data sample.
- The more left the column vector is located, the more important it is.

4.1.2 Calculation Procedures

Problem Setup

Given

- A matrix $A = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix}$.

  #### Do
- Find U, S, and V for Singular Value Decomposition.

Basic Knowledge

$$\begin{aligned} AA^\top &= \left(USV^\top\right)\left(USV^\top\right)^\top \\ &= \left(USV^\top\right)\left(VS^\top U^\top\right) \\ &= US\left(V^\top V\right)S^\top U^\top \\ &= USS^\top U \end{aligned}$$

$$\begin{aligned} A^\top A &= \left(USV^\top\right)^\top\left(USV^\top\right) \\ &= \left(VS^\top U^\top\right)\left(USV^\top\right) \\ &= VS^\top\left(U^\top U\right)SV^\top \\ &= VS^\top SV^\top \end{aligned}$$

Step 1. Calculate $AA^\top$ and $A^\top A$

Known that:

$$A = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix}, \ A^\top = \begin{bmatrix} 2 & -1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix}$$

Therefore, we could get:

$$AA^\top = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$$

$$A^\top A = \begin{bmatrix} 2 & -1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 2 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

Step 2. Eigenvalues and S

As we obtained $AA^\top$ and $A^\top A$, we can get their common eigenvalues, and construct S matrix.

- The eigenvalues $AA^\top$ and $A^\top A$ are essentially the same, except for the zero-eigenvalue.

From the definition of Eigen Values:

$$AA^\top = \lambda I$$

where λ is the eigenvalue of $AA^\top$. Calculate the eigen values:

$$\implies |AA^\top - \lambda I| = 0$$

$$\implies \left|\begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right| = 0$$

$$\implies \begin{vmatrix} 5-\lambda & -2 \\ -2 & 5-\lambda \end{vmatrix} = 0$$

$$\implies (5-\lambda)^2 - 4 = 0$$

$$\implies \lambda^2 - 10\lambda + 21 = 0$$

$$\implies (\lambda - 3)(\lambda - 7) = 0$$

$$\implies \begin{cases} \lambda_1 = 7 \\ \lambda_2 = 3 \end{cases}, \begin{cases} \sigma_1 = \sqrt{\lambda_1} = \sqrt{7} \\ \sigma_2 = \sqrt{\lambda_2} = \sqrt{3} \end{cases}$$

Calculate Eigenvalues for $A^\top A$:

$|A^\top A - \lambda I| = 0$

$$\implies \left|\begin{pmatrix} 5 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right| = 0$$

$$\implies \begin{vmatrix} 5-\lambda & -2 & 2 \\ -2 & 4-\lambda & 0 \\ 2 & 0 & 1-\lambda \end{vmatrix} = 0$$

$$\implies (5-\lambda)\Big[(4-\lambda)(1-\lambda)\Big] - (-2)\Big[-2(1-\lambda)\Big] + 2\Big[-2(4-\lambda)\Big] = 0$$

$$\implies (5-\lambda)(\lambda^2 - 5\lambda + 4) - 4(1-\lambda) - 4(4-\lambda) = 0$$

$$\implies (5-\lambda)(\lambda^2 - 5\lambda + 4) + 8\lambda - 20 = 0$$

$$\implies (5\lambda^2 - 25\lambda + 20 - \lambda^3 + 5\lambda^2 - 4\lambda) + 8\lambda - 20 = 0$$

$$\implies (-\lambda^3 + 10\lambda^2 - 29\lambda + 20) + 8\lambda - 20 = 0$$

$$\implies -\lambda^3 + 10\lambda^2 - 21\lambda = 0$$

$$\implies (\lambda^2 - 10\lambda + 21)\lambda = 0$$

$$\implies (\lambda - 3)(\lambda - 7)(\lambda - 0) = 0$$

$$\implies \begin{cases} \lambda_1 = 7 \\ \lambda_2 = 3, \\ \lambda_3 = 0 \end{cases} \begin{cases} \sigma_1 = \sqrt{\lambda_1} = \sqrt{7} \\ \sigma_2 = \sqrt{\lambda_2} = \sqrt{3} \\ \sigma_3 = \sqrt{\lambda_3} = 0 \end{cases}$$

Therefore, the diagonal matrix S would be:

$$S = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{7} & 0 & 0 \\ 0 & \sqrt{3} & 0 \end{pmatrix}$$

Step 3. Find U

We need to find U using the eigenvalues we obtained from Step 2. Again, by the property of eigenvalues of a matrix:

$$\forall x \in \mathbb{R}^m,\ (AA^\top - \lambda I)x = 0$$

For $\lambda_1 = 7$:
$(AA^\top - \lambda I)x_1 = 0$

$$\implies \left(\begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} - 7\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)x_1 = 0$$

$$\implies \begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix}x_1 = 0$$

$$\implies \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}x_1 = 0 \text{ (row 2 - row 1)}$$

$$\implies x_1 = \begin{pmatrix} a \\ -a \end{pmatrix}$$

$$\implies u_1 = \frac{x_1}{\|x_1\|} = \frac{1}{\sqrt{x_1^\top x_1}}x_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

For $\lambda_2 = 3$:
$(AA^\top - \lambda I)x_2 = 0$

$$\implies \left(\begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix} - 3\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)x_2 = 0$$

$$\implies \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}x_2 = 0 \text{ (row 2 + row 1)}$$

$$\implies \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}x_2 = 0$$

$$\implies x_2 = \begin{pmatrix} a \\ a \end{pmatrix}$$

$$\implies u_2 = \frac{x_2}{\|x_2\|} = \frac{1}{\sqrt{x_2^\top x_2}} x_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

Construct matrix U:

$$U = \begin{pmatrix} | & | \\ u_1 & u_2 \\ | & | \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

For $\lambda_3 = 0$

Step 4. Finding V

For $\lambda_1 = 7$:
$(A^\top A - \lambda_1 I)x_3 = 0$

$$\implies \left(\begin{pmatrix} 5 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 1 \end{pmatrix} - 7 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) x_3 = 0$$

$$\implies \begin{pmatrix} -2 & -2 & 2 \\ -2 & -3 & 0 \\ 2 & 0 & -6 \end{pmatrix} x_3 = 0$$

$$\implies \begin{cases} -2x_{31} - 3x_{32} = 0 \\ 2x_{31} - 6x_{33} = 0 \end{cases}$$

$$\implies \begin{cases} x_{32} = -\frac{2}{3}x_{31} \\ x_{33} = \frac{1}{3}x_{31} \end{cases}$$

$$\implies x_3 = \begin{pmatrix} x_{31} \\ -\frac{2}{3}x_{31} \\ \frac{1}{3}x_{31} \end{pmatrix} = \begin{pmatrix} 3a \\ -2a \\ a \end{pmatrix}, ||x_3|| = a\sqrt{3^2 + (-2)^2 + 1^2} = \sqrt{14} \cdot a$$

$$\implies v_1 = \frac{x_3}{\|x_3\|} = \begin{pmatrix} \frac{3}{\sqrt{14}} \\ \frac{-2}{\sqrt{14}} \\ \frac{1}{\sqrt{14}} \end{pmatrix}$$

For $\lambda_2 = 3$:
$(A^\top A - \lambda_2 I)x_4 = 0$

$$\implies \left(\begin{pmatrix} 5 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 1 \end{pmatrix} - 3 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) x_4 = 0$$

$$\implies \begin{pmatrix} 2 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & -2 \end{pmatrix} x_4 = 0$$

$$\implies \begin{cases} -2x_{41} + x_{42} = 0 \\ 2x_{41} - 2x_{43} = 0 \end{cases}$$

$$\implies \begin{cases} x_{42} = 2x_{41} \\ x_{43} = x_{41} \end{cases}$$

$$\implies x_4 = \begin{pmatrix} x_{41} \\ 2x_{41} \\ x_{41} \end{pmatrix} = \begin{pmatrix} a \\ 2a \\ a \end{pmatrix}, ||x_4|| = a\sqrt{1^2 + 2^2 + 1^2} = \sqrt{6} \cdot a$$

$$\implies v_2 = \frac{x_4}{\|x_4\|} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}$$

For $\lambda_3 = 0$:
$(A^\top A - 0I)x_5 = 0$

$$\implies \begin{pmatrix} 5 & -2 & 2 \\ -2 & 4 & 0 \\ 2 & 0 & 1 \end{pmatrix} x_5 = 0$$

$$\implies \begin{cases} -2x_{51} + 4x_{52} = 0 \\ 2x_{51} + x_{53} = 0 \end{cases}$$

$$\implies \begin{cases} x_{52} = \frac{1}{2}x_{51} \\ x_{53} = -2x_{51} \end{cases}$$

$$\implies x_5 = \begin{pmatrix} x_{51} \\ \frac{1}{2}x_{51} \\ -2x_{51} \end{pmatrix} = \begin{pmatrix} a \\ \frac{1}{2}a \\ -2a \end{pmatrix}, ||x_5|| = a\sqrt{1^2 + (\frac{1}{2})^2 + (-2)^2} = \frac{\sqrt{21}}{2}$$

$$\implies v_3 = \frac{x_5}{\|x_5\|} = \begin{pmatrix} \frac{2}{\sqrt{21}} \\ \frac{1}{\sqrt{21}} \\ \frac{-4}{\sqrt{21}} \end{pmatrix}$$

Construct matrix V:

$$V = \begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{14}} & \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{21}} \\ \frac{-2}{\sqrt{14}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{21}} \\ \frac{1}{\sqrt{14}} & \frac{1}{\sqrt{6}} & \frac{-4}{\sqrt{21}} \end{pmatrix}$$

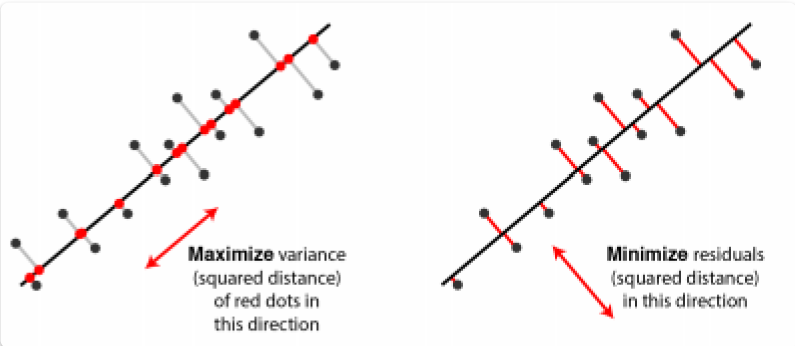
Transpose matrix V:

$$V^\top = \begin{pmatrix} - & v_1^\top & - \\ - & v_2^\top & - \\ - & v_3^\top & - \end{pmatrix} = \begin{pmatrix} \frac{3}{\sqrt{14}} & \frac{-2}{\sqrt{14}} & \frac{1}{\sqrt{14}} \\ \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{21}} & \frac{1}{\sqrt{21}} & \frac{-4}{\sqrt{21}} \end{pmatrix}$$

Step 5. Complete SVD

$$A = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \sqrt{7} & 0 & 0 \\ 0 & \sqrt{3} & 0 \end{pmatrix} \begin{pmatrix} \frac{3}{\sqrt{14}} & \frac{-2}{\sqrt{14}} & \frac{1}{\sqrt{14}} \\ \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{21}} & \frac{1}{\sqrt{21}} & \frac{-4}{\sqrt{21}} \end{pmatrix}$$

4.2 Principle Component Analysis (PCA) 主成分分析



4.2.0 Why PCA?

- Project data from higher dimension to lower dimension, while preserving a low projection error.
- Maximizes *data variance* in low-dimensional representation.
- Simple & Non-parametric method of extracting relevant information from confusing data.
- Reduce a complicate dataset to a lower dimension.

Problem Setup

Given

- An $n \times m$ training data set $X = \begin{pmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ | & | & & | \end{pmatrix}$
  - where $x^{(i)} \in \mathbb{R}^n$
  - that is, $X = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \end{pmatrix}$
  - Structural Analysis:
    - n is the *dimension* of a data sample. Each row is a feature.

- m is the *amount* of data sample. Each column is a data sample.

Do

- Reduces the dataset from n-dimensions to k-dimensions.
  - That is, to convert each feature from a n-d vector to a k-d vector;
  - Namely, to convert X from an $n \times m$ matrix into a $k \times m$ matrix.

4.2.1 Data Pre-processing: Mean Normalization

Given

- The $n \times m$ training data set $X = \begin{pmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \dots & x^{(m)} \\ | & | & & | \end{pmatrix}$

  Do

1. Calculate feature mean for all the vectors:
   - $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_m \end{pmatrix}$
   - where $\mu_j = \sum_{i=1}^{n} x_j^{(i)}$.
   - A mean of a feature with respect to all the data samples.
2. Feature scaling:
   - For each row of X, that is a set of a specific feature of each data sample,
     - Reduce each value on this row by the row mean.
     - $\begin{pmatrix} x_j^{(1)} - \mu_j & x_j^{(2)} - \mu_j & \cdots & x_j^{(m)} - \mu_j \end{pmatrix}$

What it does:

What we eventually get is:

- A scaled version of a dataset.
- Since different features may have their own range of values, which could vary, we need to normalize all features into a unified range of values.
  - E.g.: House Size is around 200 squared meters, while the price could be around 30,000.

4.2.2 Reduce Data Dimension

Given

- The normalized version of dataset X.

  Do

1. Compute the covariance matrix by:

$$\Sigma_{n\times n} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(x^{(i)})^\top = \frac{1}{m}XX^\top$$

2. Compute eigenvectors using *Singular Value Decomposition* on the covariate matrix Σ.

$$U_{n\times n}S_{n\times m}V_{m\times m} = svd(\Sigma)$$

3. Take the first k columns from U.

$$U = \begin{pmatrix} | & | & & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} & \cdots & u^{(n)} \\ | & | & & | & & | \end{pmatrix} \in \mathbb{R}^{n\times n}$$

$$\implies U_{reduce} = \begin{pmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{n\times k}$$

4. We want to reduce $x^{(i)} \in \mathbb{R}^n \to z^{(i)} \in \mathbb{R}^k$ by:

$$z^{(i)} = U_{reduce}^\top x^{(i)}$$

Namely,

$$\begin{pmatrix} - & (u^{(1)})^\top & - \\ - & (u^{(2)})^\top & - \\ & \vdots & \\ - & (u^{(k)})^\top & - \end{pmatrix}_{k\times n} \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_k^{(1)} \\ \vdots \\ x_n^{(1)} \end{pmatrix}_{n\times 1} = \begin{pmatrix} z_1^{(i)} \\ z_2^{(i)} \\ \vdots \\ z_k^{(i)} \end{pmatrix}_{k\times 1}$$

4.2.3 Choosing k

Reconstruct Original Data

After PCA, we obtain $z^{(i)} = U_{reduce}^\top x^{(i)}$. We can reconstruct the original data from $z^{(i)}$ by:

$$\tilde{x}^{(i)} = U_{reduce} z^{(i)}$$

The reconstruction comes with information loss. We will choose k based on the information loss.

Choosing k - Slow

Average Squared Projection Error:

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-\widetilde{x}^{(i)}||^2\\&=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\widetilde{x}^{(i)})^\top(x^{(i)}-\widetilde{x}^{(i)})\end{aligned}$$

Total variation of data:

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2\\&=\frac{1}{m}\sum_{i=1}^{m}{x^{(i)}}^\top x^{(i)}\end{aligned}$$

Choose the target dimension number k to be the smallest value so that:

$$\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-\widetilde{x}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2}\leq 0.01$$

i.e., 99% of the variance is retained.

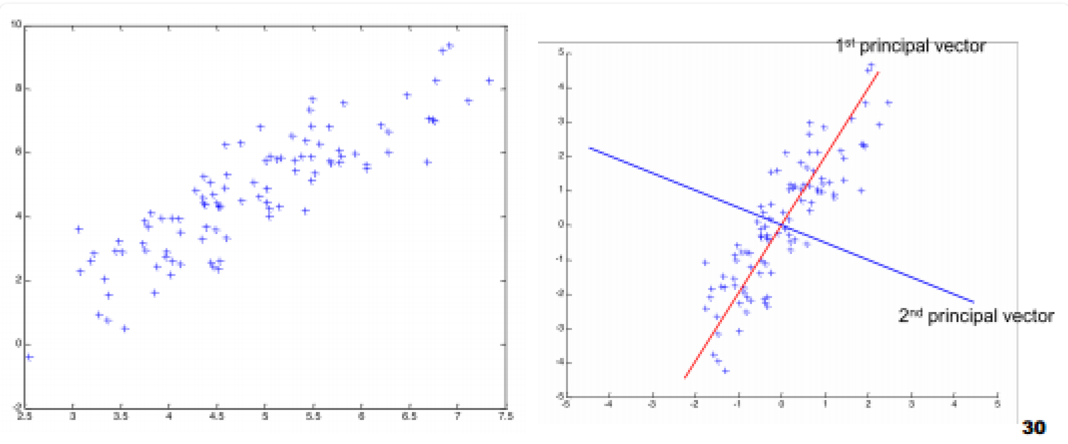
Choosing k - Fast

After performing SVD on $\Sigma=\frac{1}{m}XX^\top$, we have obtained U, S, and V. Focusing on S, we pick the smallest k for:

$$\frac{\sum_{i=1}^{k}s_{ii}}{\sum_{i=1}^{k}s_{ii}}\geq 0.99$$

i.e., 99% of the variance is retained.

4.2.4 Results



In this example, the training data X is of shape $2\times m$, thus the covariate matrix $C=\frac{1}{m}XX^\top$ is of shape 2×2. Performing SVD on C:

$$C_{2\times 2}=U_{2\times 2}S_{2\times 2}V_{2\times 2}^\top$$

There are 2 eigenvalues, with 2 principle vectors. The reduced U would be of shape 2×1.

Red Line: 1st Principal Vector

Corresponds to the *largest* eigenvalue, indicating the most significant direction the data variates.

- That is, on this line, the projected data varies the most.

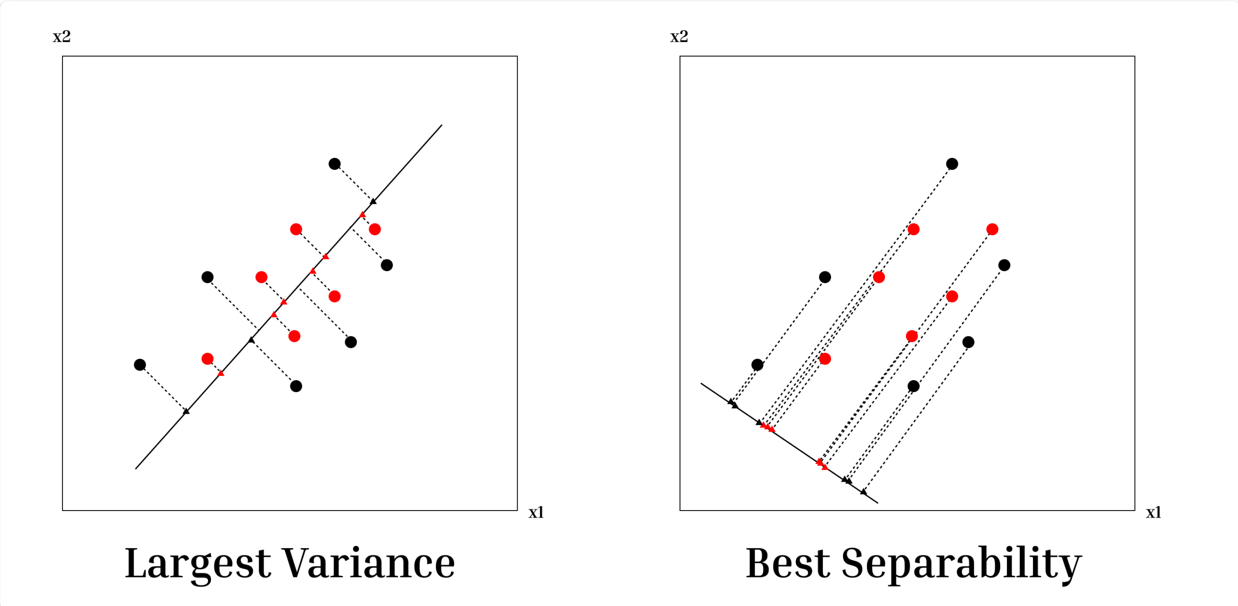
Blue Line: 2nd Principal Vector

Corresponds to the *second largest* eigenvalue, being *perpendicular* to the first one.

4.3 Linear Discriminant Analysis (LDA) 线性判别分析

4.3.0 Problems of PCA

- The *directions* of maximum variance may be useless for classification.
  - I may indeed variates, but the classes could be completely mixed together.
- LDA solves this problem by:
  - not seeking the best variance,
  - but seeking the **best separability.**
- LDA projects data to the direction *useful for classification*.



4.3.1 LDA

Given

- A set of d-dimensional samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$. From which,
  - N_1 samples belong to class ω_1.
  - N_2 samples belong to class ω_2.

  #### Do
- We seek a set of scalar $\mathbf{y} = \{y_1, y_2, \cdots, y_N\} \subset \mathbb{R}$ by projecting the N samples in x onto a line.

$$y_i = \mathbf{w}^\top \mathbf{x}_i \subset \mathbb{R}$$

- Namely,

$$y_i = \begin{pmatrix} w_1 & w_2 & \cdots & w_d \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}$$

  y_i is the projected value of $\mathbf{x}_i$ in the new space.
- LDA selects the line that maximizes the *separability* of the scalars.
- In this new space, values of y could be easily separated.

4.3.2 Measure of Separation

Supposed that we have a obtained such a line.

Sample Means of each class in x-space:

$$\mu_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{x} \in \mathbb{R}^d$$

Sample Means of each class in y-space (projected mean):

$$\begin{aligned} \widetilde{\mu}_i &= \frac{1}{N_i} \sum_{y \in \omega_i} y \\ &= \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{w}^\top \mathbf{x} \\ &= \mathbf{w}^\top \mu_i \end{aligned}$$

Distance of Means

The distance between the project mean is:

$$|\widetilde{\mu}_1 - \widetilde{\mu}_2| = |\mathbf{w}^\top (\mu_1 - \mu_2)| \in \mathbb{R}$$

Ignoring the standard deviation within classes.

Scatter

Fisher's solution is to *maximize* the difference between the means of each class.

- The means of each class is normalized by a measure of the **within-class scatter**.
- The scatter is equivalent to the *variance* of each class.

The within-class scatter of a class ω_i

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \widetilde{\mu}_i)^2$$

The total within-class scatter of all the project samples would be

$$(\tilde{s}_1^2 + \tilde{s}_1^2)$$

- ★ The criterion function would be:

$$\mathcal{J}(\mathbf{w}) = \frac{|\widetilde{\mu}_1 - \widetilde{\mu}_2|^2}{s_1^2 + s_2^2}$$

We need to find the optimal $\mathbf{w}$ that maximizes the criterion function $\mathcal{J}(\mathbf{w})$.

4.3.3 Represent $\mathcal{J}(\mathbf{w})$ with $\mathbf{w}$

We want to find the optimal $\mathbf{w}$ such that the criterion function $\mathcal{J}(\mathbf{w})$ is maximized.

- Given that $\mathcal{J}(\mathbf{w}) = \frac{|\widetilde{\mu}_1 - \widetilde{\mu}_2|^2}{s_1^2 + s_2^2}$,
- we need to use $\mathbf{w}$ to represent the scatters.

Within-Class Scatter

The scatter/variance in x-space:

$$S_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^\top \in \mathbb{R}^d \times \mathbb{R}^d$$

The within-class scatter matrix:

$$S_W = S_1 + S_2$$

To express the scatter in y-space with $\mathbf{w}$:

$$\begin{aligned}
\tilde{s}_i^2 &= \frac{1}{N_i} \sum_{y \in \omega_i} (y - \widetilde{\mu}_i)^2 \\
&= \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} (\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mu_i)^2 \\
&= \frac{1}{N_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{w}^\top (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^\top \mathbf{w} \\
&= \mathbf{w}^\top S_i \mathbf{w}
\end{aligned}$$

- ★ That is,

$$\tilde{s}_1^2 + \tilde{s}_1^2 = \mathbf{w}^\top S_W \mathbf{w}$$

Between-Class Scatter

The between-class scatter:

$$S_B = |\mu_1 - \mu_2|^2 = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$$

The difference between the projected means:

$$
\begin{aligned}
(\widetilde{\mu}_1 - \widetilde{\mu}_2) &= (\mathbf{w}^\top \boldsymbol{\mu_1} - \mathbf{w}^\top \boldsymbol{\mu_2})^2 \\
&= \mathbf{w}^\top (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{w} \\
&= \mathbf{w}^\top S_B \mathbf{w}
\end{aligned}
$$

The optimal $\mathbf{w}$

The optimal $\mathbf{w}$ will be:

$$
\begin{aligned}
\mathbf{w}^* &= \text{argmax}_{\mathbf{w}} \mathcal{J}(\mathbf{w}) \\
&= \text{argmax}_{\mathbf{w}} \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}
\end{aligned}
$$

4.3.4 Find the optimal $\mathbf{w}$

To find the optimal $\mathbf{w}$, we find that:

$$
\frac{d}{d\mathbf{w}} \mathcal{J}(\mathbf{w}) = 0
$$

$$
\implies \frac{d}{d\mathbf{w}} \left(\frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}} \right) = 0
$$

$$
\implies \frac{1}{(\mathbf{w}^\top S_W \mathbf{w})^2} \cdot \left(\mathbf{w}^\top S_W \mathbf{w} \frac{d}{d\mathbf{w}} (\mathbf{w}^\top S_B \mathbf{w}) - \mathbf{w}^\top S_B \mathbf{w} \frac{d}{d\mathbf{w}} (\mathbf{w}^\top S_W \mathbf{w}) \right) = 0
$$

- 上导下不导-上不导下导

$$
\implies \frac{1}{(\mathbf{w}^\top S_W \mathbf{w})^2} \left(\mathbf{w}^\top S_W \mathbf{w} (2 S_B \mathbf{w}) - \mathbf{w}^\top S_B \mathbf{w} (2 S_W \mathbf{w}) \right) = 0
$$

$$
\implies \mathbf{w}^\top S_W \mathbf{w} (2 S_B \mathbf{w}) - \mathbf{w}^\top S_B \mathbf{w} (2 S_W \mathbf{w}) = 0
$$

$$
\implies S_B \mathbf{w} - \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}} (S_W \mathbf{w}) = 0
$$

$$
\implies S_B \mathbf{w} - \mathcal{J}_{\max} S_W \mathbf{w} = 0
$$

Set constant $\lambda = \mathcal{J}_{\max} = \dfrac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$

$$
\implies S_B \mathbf{w} = \lambda S_W \mathbf{w}
$$

$$
\implies S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}
$$

Know that $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$ is the between-class scatter matrix.

- Therefore, $S_B \mathbf{w} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \mathbf{w} = \alpha(\boldsymbol{\mu_1} - \mu_2)$
- where $\alpha = (\mu_1 - \mu_2)^\top \mathbf{w} \in \mathbb{R}$, that is α is a scalar.
- i.e., $S_B \mathbf{w}$ points to the same direction as $\mu_1 - \mu_2$

$$
\implies S_W^{-1} (\mu_1 - \mu_2) = \lambda \mathbf{w}
$$

★ $\implies \mathbf{w} = S_W^{-1}(\mu_1 - \mu_2)$

Example:

Compute LDA projection of the following 2D dataset.

- $X_1 = \{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
- $X_2 = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$

LDA Solution:

Step 1: Data Arrangements

Arrange data into 2 separate matrices

$$X_1 = \begin{pmatrix} 4 & 2 & 2 & 3 & 4 \\ 1 & 4 & 3 & 6 & 4 \end{pmatrix}$$
$$X_2 = \begin{pmatrix} 9 & 6 & 9 & 8 & 10 \\ 10 & 8 & 5 & 7 & 8 \end{pmatrix}$$

Step 2: Class Statistics

Sample means:

$$\mu_1 = \begin{pmatrix} \frac{4+2+2+3+4}{5} \\ \frac{1+4+3+6+4}{5} \end{pmatrix} = \begin{pmatrix} 3.0 \\ 3.6 \end{pmatrix}$$
$$\mu_2 = \begin{pmatrix} \frac{9+6+9+8+10}{5} \\ \frac{10+8+5+7+8}{5} \end{pmatrix} = \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix}$$

Sample Variants:

$$
\begin{aligned}
S_1 =& \frac{1}{5}\Bigg(\begin{bmatrix} 1 \\ -2.6 \end{bmatrix}\begin{bmatrix} 1 & -2.6 \end{bmatrix} + \\
& \begin{bmatrix} -1 \\ 0.4 \end{bmatrix}\begin{bmatrix} -1 & 0.4 \end{bmatrix} + \\
& \begin{bmatrix} -1 \\ -0.6 \end{bmatrix}\begin{bmatrix} -1 & -0.6 \end{bmatrix} + \\
& \begin{bmatrix} 0 \\ 2.4 \end{bmatrix}\begin{bmatrix} 0 & 2.4 \end{bmatrix} + \\
& \begin{bmatrix} 1 \\ 0.4 \end{bmatrix}\begin{bmatrix} 1 & 0.4 \end{bmatrix}\Bigg) \\
=& \frac{1}{5}\begin{pmatrix} 1 & -1 & -1 & 0 & 1 \\ -2.6 & 0.4 & -0.6 & 2.4 & 0.4 \end{pmatrix}\begin{pmatrix} 1 & -2.6 \\ -1 & 0.4 \\ -1 & -0.6 \\ 0 & 2.4 \\ 1 & 0.4 \end{pmatrix} \\
=& \frac{1}{5}\begin{pmatrix} 4 & -2 \\ -2 & 13.2 \end{pmatrix} \\
=& \begin{pmatrix} 0.8 & -0.4 \\ -0.4 & 2.64 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
S_2 &= \frac{1}{5}\begin{pmatrix} 0.6 & -2.4 & 0.6 & -0.4 & 1.6 \\ 2.4 & 0.4 & -2.6 & -0.6 & 0.4 \end{pmatrix}\begin{pmatrix} 0.6 & 2.4 \\ -2.4 & 0.4 \\ 0.6 & -2.6 \\ -0.4 & -0.6 \\ 1.6 & 0.4 \end{pmatrix} \\
&= \frac{1}{5}\begin{pmatrix} 9.2 & -0.2 \\ -0.2 & 13.2 \end{pmatrix} \\
&= \begin{pmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{pmatrix}
\end{aligned}
$$

Step 3: Between & Within Class Scatters

Within-class scatters:

$$
\begin{aligned}
S_W &= S_1 + S_2 \\
&= \begin{pmatrix} 0.8 & -0.4 \\ -0.4 & 2.64 \end{pmatrix} + \begin{pmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{pmatrix} \\
&= \begin{pmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{pmatrix}
\end{aligned}
$$

Between-class Scatter:

$$
\begin{aligned}
S_B &= (\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top \\
&= \begin{pmatrix} -5.4 \\ -4 \end{pmatrix}\begin{pmatrix} -5.4 & -4 \end{pmatrix} \\
&= \begin{pmatrix} 29.16 & 21.6 \\ 21.6 & 16 \end{pmatrix}
\end{aligned}
$$

Step 4: Calculate LDA Projection

Inverse of the between-class scatter matrix:

$$
\begin{aligned}
S_W^{-1} &= \begin{pmatrix} 2.64 & -0.44 \\ -0.44 & 5.28 \end{pmatrix}^{-1} \\
&= \frac{1}{5.28 \times 2.64 - 0.44^2} \begin{pmatrix} 5.28 & 0.44 \\ 0.44 & 2.64 \end{pmatrix} \\
&= \begin{pmatrix} 0.3841 & 0.0320 \\ 0.0320 & 0.1921 \end{pmatrix}
\end{aligned}
$$

The LDA projection $\mathbf{w}$

$$
\begin{aligned}
\mathbf{w} &= S_W^{-1}(\mu_1 - \mu_2) \\
&= \begin{pmatrix} 0.3841 & 0.0320 \\ 0.0320 & 0.1921 \end{pmatrix} \begin{pmatrix} -5.4 \\ -4 \end{pmatrix} \\
&= \begin{pmatrix} -2.2021 \\ -0.9412 \end{pmatrix}
\end{aligned}
$$

Therefore, the LDA projection line would be:

$$
y = (-2.2021 \quad -0.9412)\mathbf{x}
$$