# 机器学习引论

彭玺

pengxi@scu.edu.cn

www.pengxi.me

# 提纲

一．Review

二．Canonical Correlation Analysis

三．Linear Discriminant Analysis

# 提纲

一．Review

二．Canonical Correlation Analysis

三．Linear Discriminant Analysis

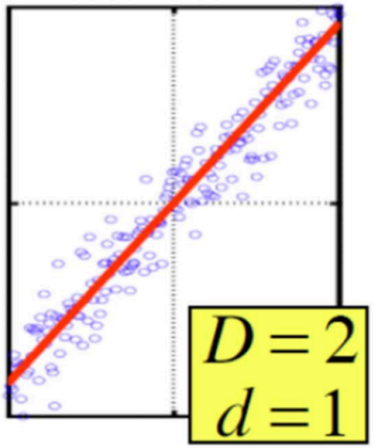
# 一、Review

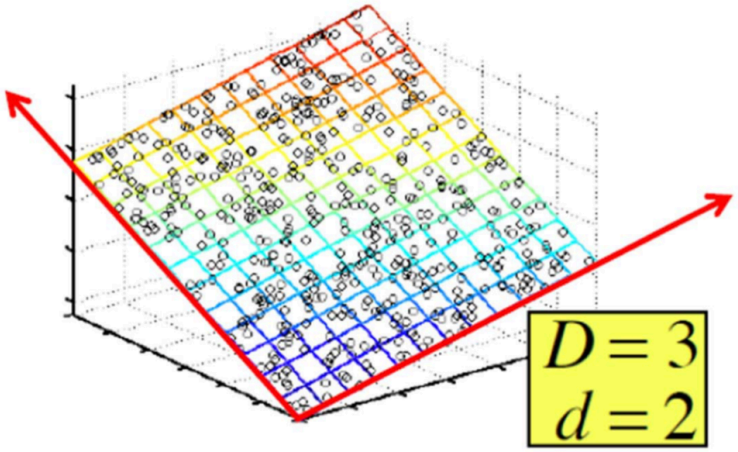


Dimensionality reduction (DR) or dimension reduction is the process of reducing the number of random variables/dimension under consideration by obtaining a set of principal variables – redundancy removal.

Basis：In mathematics, a set of elements (vectors) in a vector space V is called a basis, or a set of basis vectors, if the vectors are linearly independent and every vector in the vector space is a linear combination of this set.

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_m} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$



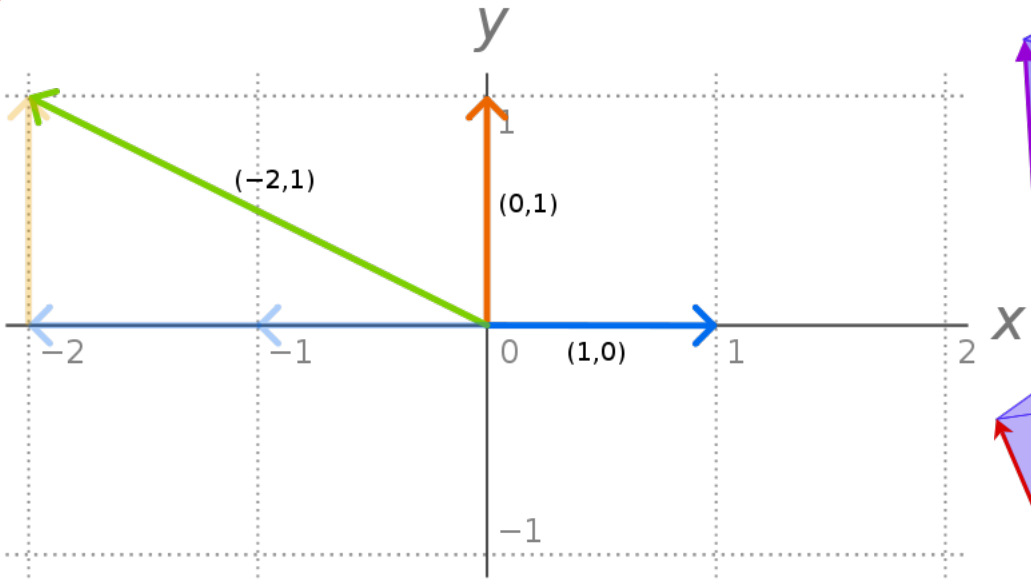
# 一、Review

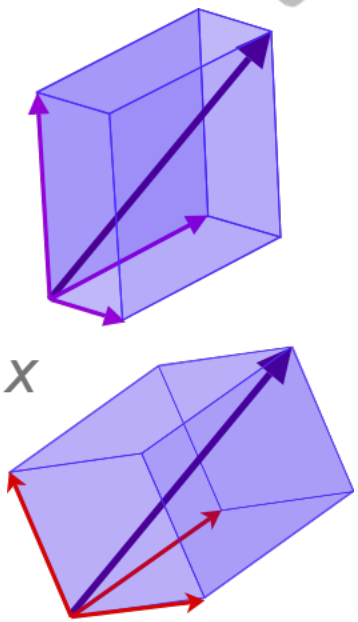


Dimensionality reduction (DR) or dimension reduction is the process of reducing the number of random variables/dimension under consideration by obtaining a set of principal variables – redundancy removal.

Basis：In mathematics, a set of elements (vectors) in a vector space V is called a basis, or a set of basis vectors, if the vectors are linearly independent and every vector in the vector space is a linear combination of this set.

**DR could be achieved by seeking the basis of a give data set!**



# 一、Review



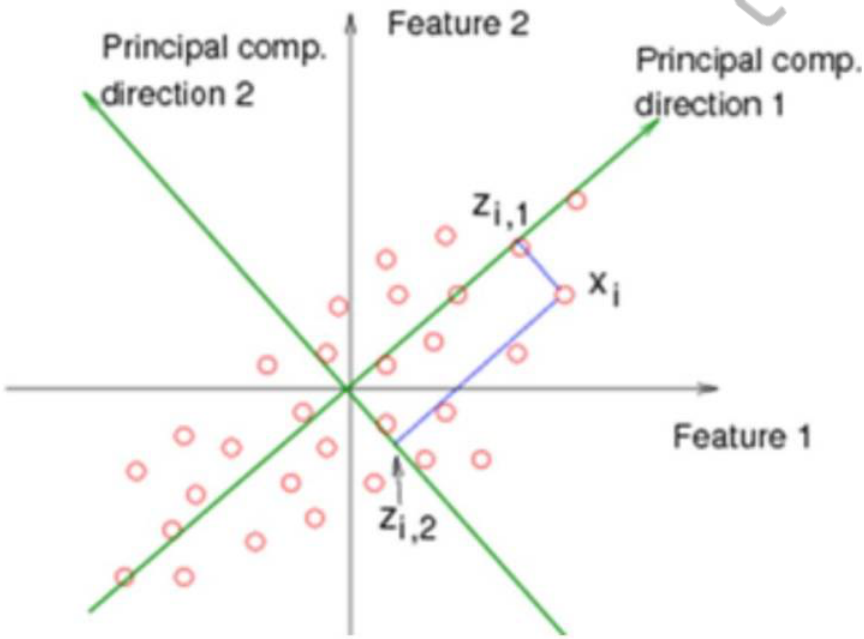
PCA aims to seek the components W from a given data set X, i.e.,

$$Y=W^TX,$$

where W is the set of components (projection matrix), and Y is the reduced representation of X.

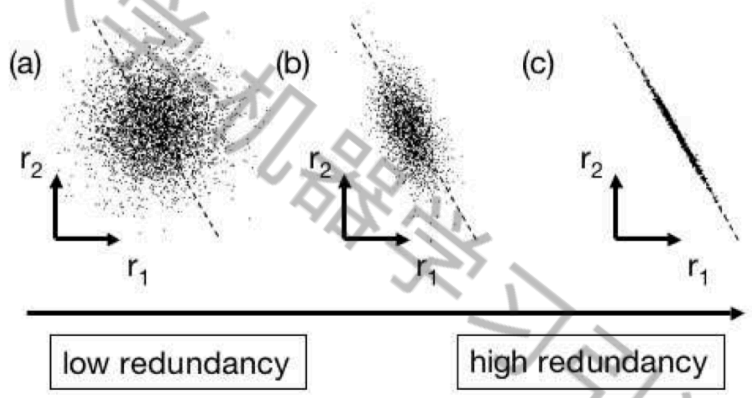
How?

The intrinsic assumption of DR is that the data contains redundancy!



# 一、Review

How to compute redundancy in mathematics?




- Covariance:

If the entries in the column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

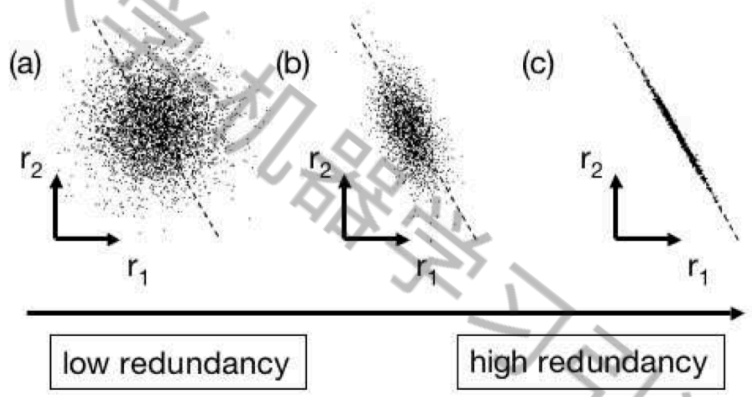
are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \operatorname{E}[(X_i - \mu_i)(X_j - \mu_j)] = \operatorname{E}[X_i X_j] - \mu_i \mu_j$$

where the operator E denotes the expected (mean) value of its argument, and

$$\mu_i = \operatorname{E}(X_i)$$

# 一、Review




How to compute redundancy in mathematics?

- Covariance:

If the entries in the column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \operatorname{E}[(X_i - \mu_i)(X_j - \mu_j)] = \operatorname{E}[X_i X_j] - \mu_i \mu_j$$

where the operator E denotes the expected (mean) value of its argument, and

$$\mu_i = \operatorname{E}(X_i)$$

- $\sum_{ij} = 0$ if and only if i and j are entirely uncorrelated.
- Otherwise, i and j are correlated.

Correlated=redundant!

Futher reading: In fact, the variances $\sum_{ij}$ also defines the signal-to-noise ratio.

# 一、Review



Let the data set X be with zero mean, then define

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}$$

i.e.,

$$\mathrm{C} = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{x^2}_1 & \cdots & \boldsymbol{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x_n x_1} & \cdots & \boldsymbol{x^2}_n \end{bmatrix}$$

Some properties of $\mathbf{C_X}$:

- $\mathbf{C_X}$ is a square symmetric
- The diagonal terms of $\mathbf{C_X}$ are the *variance* of particular measurement types.
- The off-diagonal terms of $\mathbf{C_X}$ are the *covariance* between measurement types.

$\mathbf{C_X}$ captures the correlations between all possible pairs of measurements. The correlation values reflect the noise and redundancy in our measurements.

- In the diagonal terms, by assumption, large (small) values correspond to interesting dynamics (or noise).
- In the off-diagonal terms large (small) values correspond to high (low) redundancy.

# 一、Review



As the covariance defines the redundancy, then one could remove the redundancy in low dimensional space by diagonalizing the covariance matrix.

投影：$\mathbf{w}^T \boldsymbol{x}$

方差：$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T \boldsymbol{x}_i)^2 = \mathbf{w}^T S \mathbf{w}$

$S = \frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$

拉格朗日乘数法：

$$L = \mathbf{w}^T S \mathbf{w} + \lambda(1 - \mathbf{w}^T \mathbf{w})$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2S\mathbf{w} - 2\lambda\mathbf{w}$$

$$S\mathbf{w} = \lambda\mathbf{w}$$

最大方差：

$$\max_{\mathbf{w}} \quad \mathbf{w}^T S \mathbf{w}$$
$$s.t. \quad \|\mathbf{w}\| = 1$$

方差：

$$\mathbf{w}^T S \mathbf{w} = \mathbf{w}^T \lambda \mathbf{w} = \lambda$$

Note that: $\frac{1}{n}\sum_{i=1\ldots n}\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_{\mathrm{i}}\right)^2 = \mathrm{w}^{\mathrm{T}}\mathrm{X}\mathrm{X}^{\mathrm{T}}\mathrm{w}$

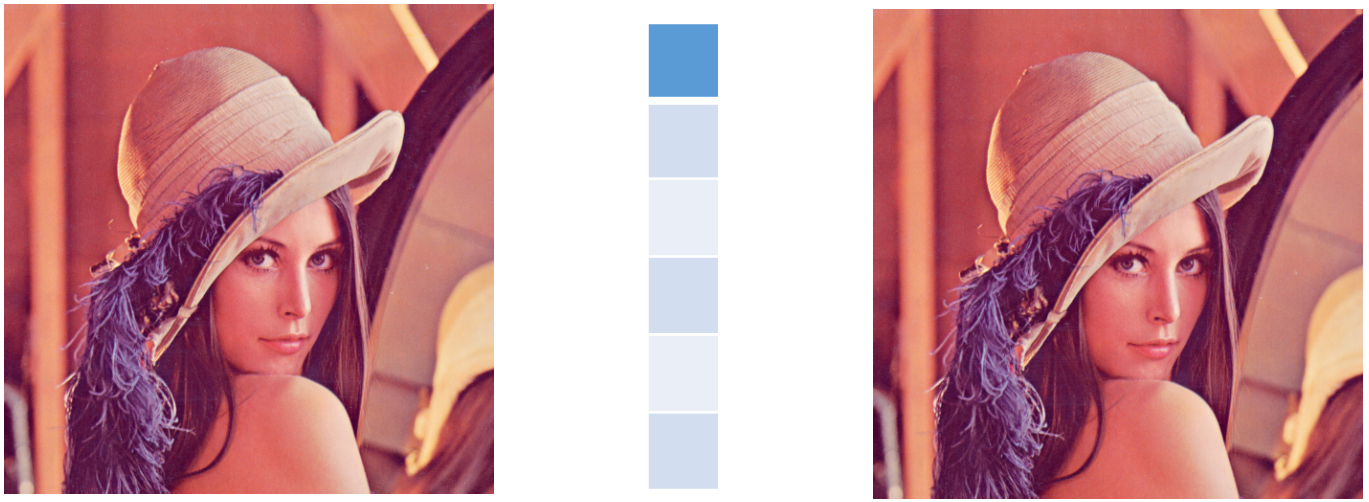
# 一、Review



View 1: Redundancy removal by removing mutual correlation

- Covariance measures redundancy
- thus DR could be achieved by diagonalizing the covariance matrix
- leading to the ED on $X^TX$

View 2: minimizing reconstruction error/description length.



Input low dimensional rep. Reconstruction

# 一、Review



View 2: minimizing reconstruction error/description length.

正交基： $\mathbf{u}_1, \cdots, \mathbf{u}_D$

原始数据： $\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \mathbf{u}_j$

基坐标： $\alpha_{ij} = \mathbf{u}_j^T \boldsymbol{x}_i$

降维重建： $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} \alpha_{ij} \mathbf{u}_j$

$$\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\sum_{j=1}^{D} \alpha_{ij}\mathbf{u}_j - \sum_{j=1}^{d} \alpha_{ij}\mathbf{u}_j\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\sum_{j=d+1}^{D} \alpha_{ij}\mathbf{u}_j\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \alpha_{ij}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \mathbf{u}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \mathbf{u}_j$$

$$= \sum_{j=d+1}^{D} \mathbf{u}_j^T S \mathbf{u}_j$$ 等价方差最小

# 一、Review



Q1: Why the variances $\sum_{ij}$ also defines the signal-to-noise ratio? And the properties of SNR w.r.t. $\sum_{ij}$.

Q2: Beside performing ED on $X^TX$, is there other way to obtain the principal components?

Q3: Can PCA handle the data drawn from multiple subspace? Why?

Q4: PCA is a unsupervised dimension reduction method, which may suffer from what problem or limitations?

Q5: What distance is adopted by PCA to measure the relation among data points? Could such a measurement solve the linear inseparable issue? If Yes/NO, why?

Multiple subspace dimension reduction

+

Supervised dimension reduction

# 提纲

一．Review

二．Canonical Correlation Analysis

三．Linear Discriminant Analysis

# 二、Canonical Correlation Analysis

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^T S \mathbf{w} \\ s.t. \quad & \|\mathbf{w}\| = 1 \end{aligned}$$

- Find the projection direction w such that the variance of projected data is maximized;
- Intuitively, find the intrinsic subspace of the original feature space (in terms of retaining the data variability).

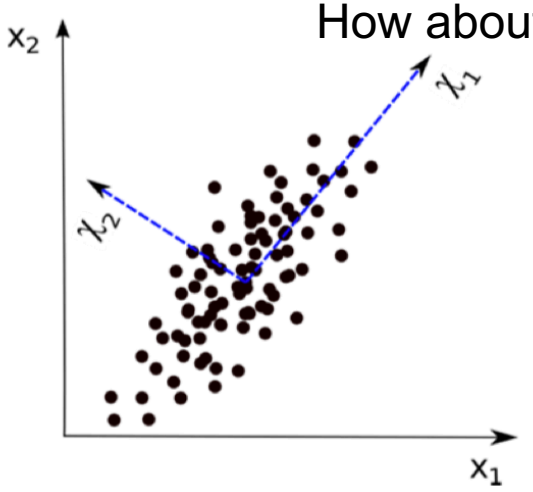
# 二、Canonical Correlation Analysis

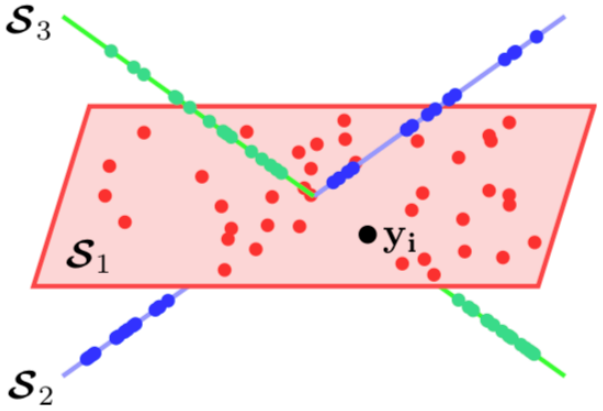


$$\max_{\mathbf{w}} \quad \mathbf{w}^T S \mathbf{w}$$
$$s.t. \quad \|\mathbf{w}\| = 1$$

- Find the projection direction w such that the variance of projected data is maximized;
- Intuitively, find the intrinsic subspace of the original feature space (in terms of retaining the data variability).

How about the performance of PCA?



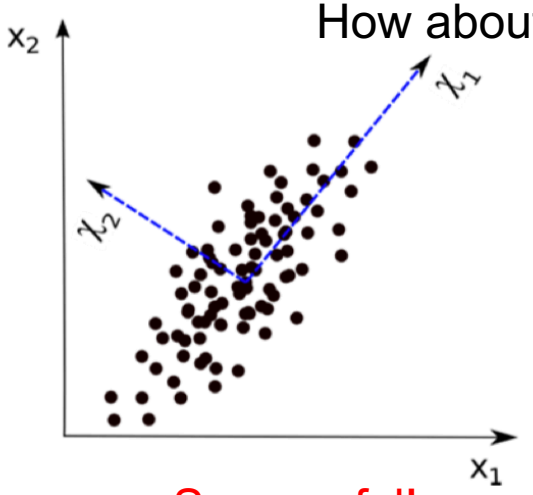
# 二、Canonical Correlation Analysis



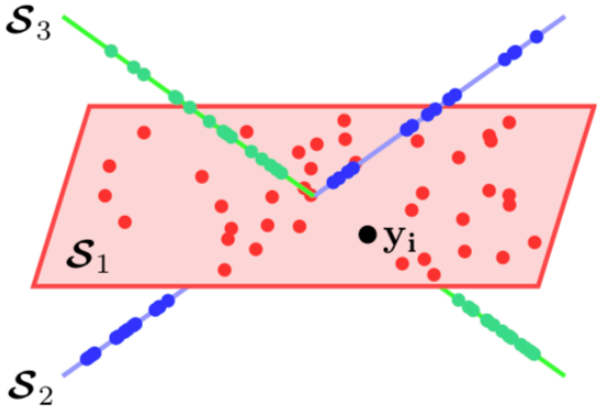
$$\max_{\mathbf{w}} \quad \mathbf{w}^T S \mathbf{w}$$
$$s.t. \quad \|\mathbf{w}\| = 1$$

- Find the projection direction w such that the variance of projected data is maximized;
- Intuitively, find the intrinsic subspace of the original feature space (in terms of retaining the data variability).

How about the performance of PCA?




Successful!

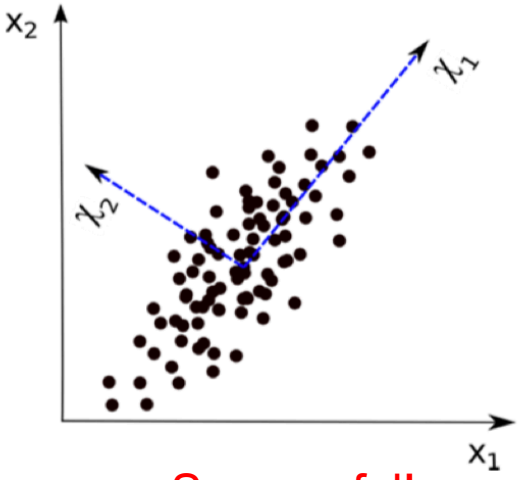



Failed!

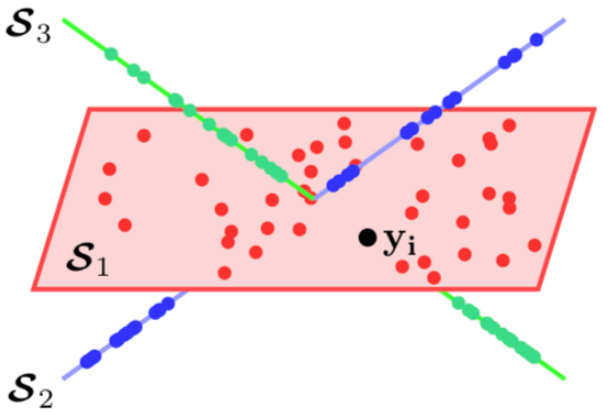
# 二、Canonical Correlation Analysis

$$\max_{\mathbf{w}} \quad \mathbf{w}^T S \mathbf{w}$$
$$s.t. \quad \|\mathbf{w}\| = 1$$

The reason is that PCA can only handle the data drawn from singe subspace!




Successful!

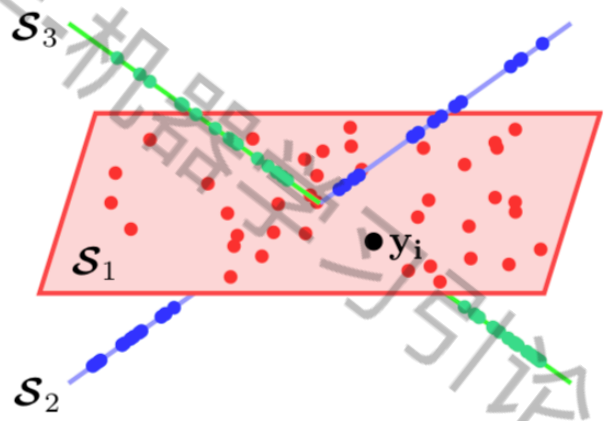



Failed!

# 二、Canonical Correlation Analysis



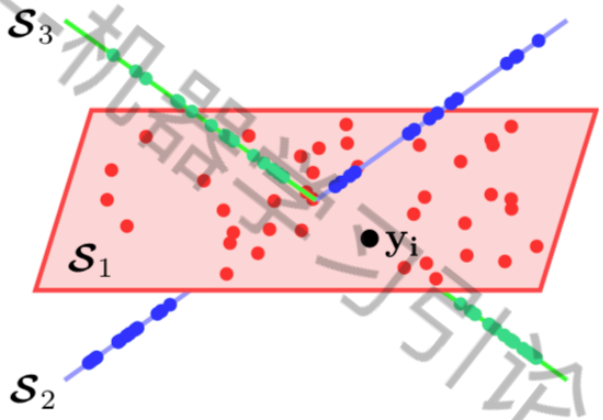
- Now consider two sets of variables x and y
  - x is a vector of p variables
  - y is a vector of q variables
  - Basically, two correlated feature spaces




- How to find the connection between two set of variables (or two feature spaces)?

# 二、Canonical Correlation Analysis






- Now consider two sets of variables x and y
  - x is a vector of p variables
  - y is a vector of q variables
  - Basically, two correlated feature spaces
- How to find the connection between two set of variables (or two feature spaces)?
  - CCA can be defined as the problem of finding two sets of basis vectors, one for **x** and the other for **y**, such that the correlations between the projections of the variables onto these basis vectors are mutually maximized.
  - Note: CCA simultaneously finds dimension reduction for two feature spaces.

# 二、Canonical Correlation Analysis

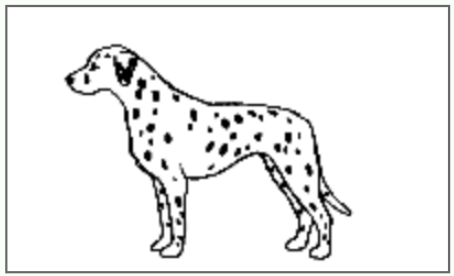


Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable, that are optimal with respect to correlations and, at the same time, it finds the corresponding correlations.

In other words, it finds the two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized.

The latent assumption of CCA is that these two variables **describe the same object in different views**, e.g., image+text,

- multi-view analysis



A Dalmatian dog. Drawing by Liisa Sarakontu.

# 二、Canonical Correlation Analysis

Let $x = \hat{\mathbf{w}}_x^T \mathbf{x}$ and $y = \hat{\mathbf{w}}_y^T \mathbf{y}$ (canonical variates), the formulation of CCA is as follows:

# 二、Canonical Correlation Analysis

Let $x = \mathbf{w}_x^T \mathbf{x}$ and $y = \mathbf{w}_y^T \mathbf{y}$ (canonical variates), the formulation of CCA is as follows:

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \mathrm{E}\left[X_i X_j\right] - \mu_i \mu_j$$

$$\begin{aligned} \rho = \quad &= \frac{\operatorname{cov}\left(\boldsymbol{w}_x^\top \boldsymbol{x}, \boldsymbol{w}_y^\top \boldsymbol{y}\right)}{\sqrt{\operatorname{var}\left(\boldsymbol{w}_x^\top \boldsymbol{x}\right) \operatorname{var}\left(\boldsymbol{w}_y^\top \boldsymbol{y}\right)}} \\ &= \frac{\boldsymbol{w}_x^\top \boldsymbol{C}_{xy} \boldsymbol{w}_y}{\sqrt{\left(\boldsymbol{w}_x^\top \boldsymbol{C}_{xx} \boldsymbol{w}_x\right)\left(\boldsymbol{w}_y^\top \boldsymbol{C}_{yy} \boldsymbol{w}_y\right)}}, \end{aligned}$$

# 二、Canonical Correlation Analysis

Let $x = \mathbf{w}_x^T \mathbf{x}$ and $y = \mathbf{w}_y^T \mathbf{y}$ (canonical variates), the formulation of CCA is as follows:

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \mathrm{E}\left[X_i X_j\right] - \mu_i \mu_j$$

$$\mathrm{argmax}_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathrm{cov}(\mathbf{w}_x^T \mathbf{x}, \mathbf{w}_y^T \mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{w}_x^T \mathbf{x}) \mathrm{var}(\mathbf{w}_y^T \mathbf{y})}}$$

$$= \mathrm{argmax}_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathbf{w}_x^T \mathbf{x} (\mathbf{w}_y^T \mathbf{y})^T}{\sqrt{(\mathbf{w}_x^T \mathbf{x})(\mathbf{w}_x^T \mathbf{x})^T (\mathbf{w}_y^T \mathbf{y})(\mathbf{w}_y^T \mathbf{y})^T}}$$

CCA is formulated as

$$\underset{\boldsymbol{w}_x, \boldsymbol{w}_y}{\arg\max} \ \boldsymbol{w}_x^\top \boldsymbol{C}_{xy} \boldsymbol{w}_y,$$

$$\text{s.t.} \quad \begin{aligned} \boldsymbol{w}_x^\top \boldsymbol{C}_{xx} \boldsymbol{w}_x &= 1, \\ \boldsymbol{w}_y^\top \boldsymbol{C}_{yy} \boldsymbol{w}_y &= 1. \end{aligned}$$

$x \in \mathcal{R}^d$

$\mathbf{x} \in \mathcal{R}^{m_x}$

$\mathbf{w}_x \in \mathcal{R}^{m_x \times d}$

$y \in \mathcal{R}^d$

$\mathbf{y} \in \mathcal{R}^{m_y}$

$\mathbf{w}_y \in \mathcal{R}^{m_y \times d}$

# 二、Canonical Correlation Analysis

Incorporating these two constraints, the Lagrangian $\mathcal{J}$ is given by

$$\mathcal{J} = \boldsymbol{w}_x^\top \boldsymbol{C}_{xy}\boldsymbol{w}_y + \lambda_x\left(1 - \boldsymbol{w}_x^\top \boldsymbol{C}_{xx}\boldsymbol{w}_x\right) + \lambda_y\left(1 - \boldsymbol{w}_y^\top \boldsymbol{C}_{yy}\boldsymbol{w}_y\right).$$

Let $\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_x} = 0$ and $\frac{\partial \mathcal{J}}{\partial \boldsymbol{w}_y} = 0$ lead to

$$\begin{aligned} \boldsymbol{C}_{xy}\boldsymbol{w}_y - 2\lambda_x \boldsymbol{C}_{xx}\boldsymbol{w}_x &= 0, \quad (1) \\ \boldsymbol{C}_{yx}\boldsymbol{w}_x - 2\lambda_y \boldsymbol{C}_{yy}\boldsymbol{w}_y &= 0. \quad (2) \end{aligned}$$

Pre-multiply (1) by $\boldsymbol{w}_x^\top$ and pre-multiply (2) $\boldsymbol{w}_y^\top$ to obtain

$$\begin{aligned} \boldsymbol{w}_x^\top \boldsymbol{C}_{xy}\boldsymbol{w}_y - 2\lambda_x \underbrace{\boldsymbol{w}_x^\top \boldsymbol{C}_{xx}\boldsymbol{w}_x}_{1} &= 0, \\ \boldsymbol{w}_y^\top \boldsymbol{C}_{yx}\boldsymbol{w}_x - 2\lambda_y \underbrace{\boldsymbol{w}_y^\top \boldsymbol{C}_{yy}\boldsymbol{w}_y}_{1} &= 0, \end{aligned}$$

# 二、Canonical Correlation Analysis

$$\begin{aligned} \boldsymbol{w}_x^\top \boldsymbol{C}_{xy}\boldsymbol{w}_y - 2\lambda_x \underbrace{\boldsymbol{w}_x^\top \boldsymbol{C}_{xx}\boldsymbol{w}_x}_{1} &= 0, \\ \boldsymbol{w}_y^\top \boldsymbol{C}_{yx}\boldsymbol{w}_x - 2\lambda_y \underbrace{\boldsymbol{w}_y^\top \boldsymbol{C}_{yy}\boldsymbol{w}_y}_{1} &= 0, \end{aligned}$$

leading to

$$\begin{aligned} \boldsymbol{w}_x^\top \boldsymbol{C}_{xy}\boldsymbol{w}_y &= 2\lambda_x, \\ \boldsymbol{w}_y^\top \boldsymbol{C}_{yx}\boldsymbol{w}_x &= 2\lambda_y. \end{aligned}$$

Since $\boldsymbol{w}_x^\top \boldsymbol{C}_{xy}\boldsymbol{w}_y = \boldsymbol{w}_y^\top \boldsymbol{C}_{yx}\boldsymbol{w}_x$, we have $\boxed{\lambda = 2\lambda_x = 2\lambda_y}$.

Then, CCA is solved by following generalized eigenvalue problem

$$\begin{bmatrix} 0 & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{yx} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{C}_{xx} & 0 \\ 0 & \boldsymbol{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix},$$

where $\lambda = 2\lambda_x = 2\lambda_y$.

# 二、Canonical Correlation Analysis



Recall the generalized eigenvalue problem:

$$\begin{bmatrix} 0 & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{yx} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{C}_{xx} & 0 \\ 0 & \boldsymbol{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix}.$$

This problem has $m_x+m_y$ eigenvalues $\{\lambda_1, -\lambda_1, \ldots, \lambda_m, -\lambda_m, 0, \ldots, 0\}$, where $m = \min(m_x, m_y)$.

The generalized eigenvalue problem can be re-written as

$$\begin{bmatrix} \boldsymbol{C}_{xx} & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{yx} & \boldsymbol{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix} = (1+\lambda) \begin{bmatrix} \boldsymbol{C}_{xx} & 0 \\ 0 & \boldsymbol{C}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_x \\ \boldsymbol{w}_y \end{bmatrix}.$$

This problem has $m_x+m_y$ eigenvalues $\{1+\lambda_1, 1-\lambda_1, \ldots, 1+\lambda_m, 1-\lambda_m, 1, \ldots, 1\}$.

# 二、Canonical Correlation Analysis



## CCA: Extension to Multiple Sets of Variables

Binary case:

$$\arg\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \ \boldsymbol{w}_x^\top \boldsymbol{C}_{xy} \boldsymbol{w}_y,$$

$$\text{s.t.} \quad \begin{aligned} \boldsymbol{w}_x^\top \boldsymbol{C}_{xx} \boldsymbol{w}_x &= 1, \\ \boldsymbol{w}_y^\top \boldsymbol{C}_{yy} \boldsymbol{w}_y &= 1. \end{aligned}$$

multiple case, how?

# 二、Canonical Correlation Analysis



## CCA: Extension to Multiple Sets of Variables

Consider $n$ multiple sets of variables, $\boldsymbol{x}_1 \in \mathbb{R}^{m_1}, \boldsymbol{x}_2 \in \mathbb{R}^{m_2}, \ldots, \boldsymbol{x}_n \in \mathbb{R}^{m_n}$.

Then, CCA is formulated as

$$\arg\max_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n} \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{C}_{ij}\boldsymbol{w}_j,$$

$$\text{s.t.} \sum_{i=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{C}_{ii}\boldsymbol{w}_j = 1.$$

Why does not separate?

# 二、Canonical Correlation Analysis



Incorporating these two constraints, the Lagrangian $\mathcal{J}$ is given by

$$\mathcal{J} = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{C}_{ij}\boldsymbol{w}_j + \lambda\left(1 - \sum_{i=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{C}_{ii}\boldsymbol{w}_j\right).$$

It follows from $\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_i} = 0$ for $i = 1, \ldots, n$ that we have

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1n} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2n} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{C}_{n1} & \boldsymbol{C}_{n2} & \cdots & \boldsymbol{C}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_n \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{C}_{11} & 0 & \cdots & 0 \\ 0 & \boldsymbol{C}_{22} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \boldsymbol{C}_{nn} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_n \end{bmatrix}.$$

The minimal generalized eigenvalue has the fixed range $[0, 1]$, whereas the maximal generalized eigenvalue has a range dependent on the dimensions of the variables. The minimal generalized eigenvalue is more convenient.

# 提纲

一．Review

二．Canonical Correlation Analysis

三．Linear Discriminant Analysis

# 二、Linear Discriminant Analysis

Q1: Curse of high dimensionality

A1: Dimension reduction by PCA based on redundancy removal

Q2: PCA will be failed if the data come from multiple subspace

A2: CCA

Q3: PCA and CCA do not utilize label information

A4: LDA

# 二、Linear Discriminant Analysis

LDA or called FDA:

- Introduced by Ronald Fisher (1936)
- One of widely-used linear discriminant analysis (LDA) methods
- FLD aims at achieving an optimal linear dimensionality reduction for classification
- Linear dimensionality reduction: PCA, ICA, FLD, MDS

# 二、Linear Discriminant Analysis



Unsupervised DR including PCA and CCA reduce the dimension of data based on the latent structure (correlation) of data without the help of label, namely,

$$\mathcal{R}^{d_1} \rightarrow \mathcal{R}^{d_2}$$

where $d_1 > d_2$

How to achieve supervised DR?

# 二、Linear Discriminant Analysis

- Given training instances (x, y)
- Learn a model/mapping f(.)
- Such that f(x) = y
- Use f(.) to predict y for new x

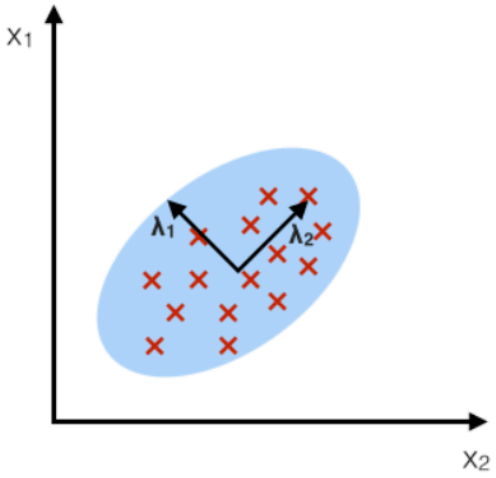
# 二、Linear Discriminant Analysis

Comparing with PCA:

- LDA could solve the multiple subspace DR in a supervised way.

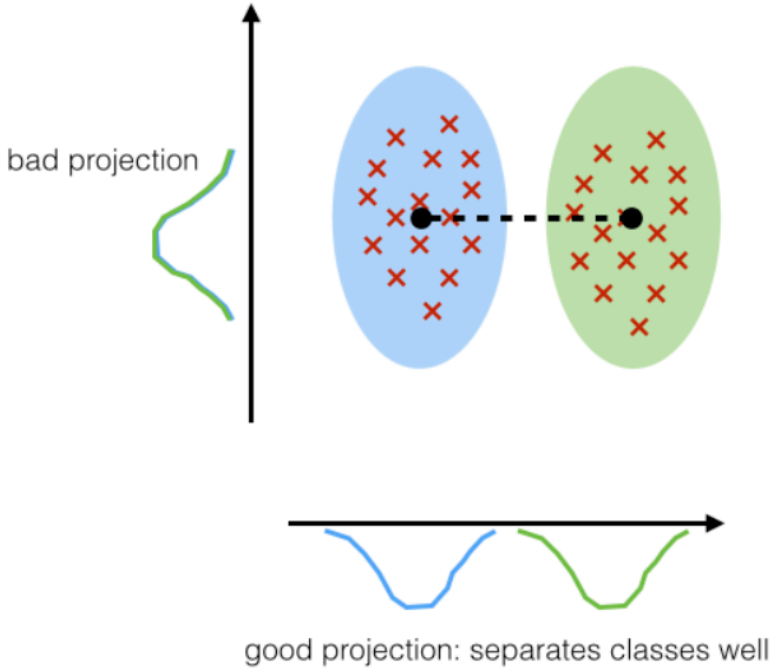
**PCA:**

component axes that maximize the variance




**LDA:**

maximizing the component axes for class-separation

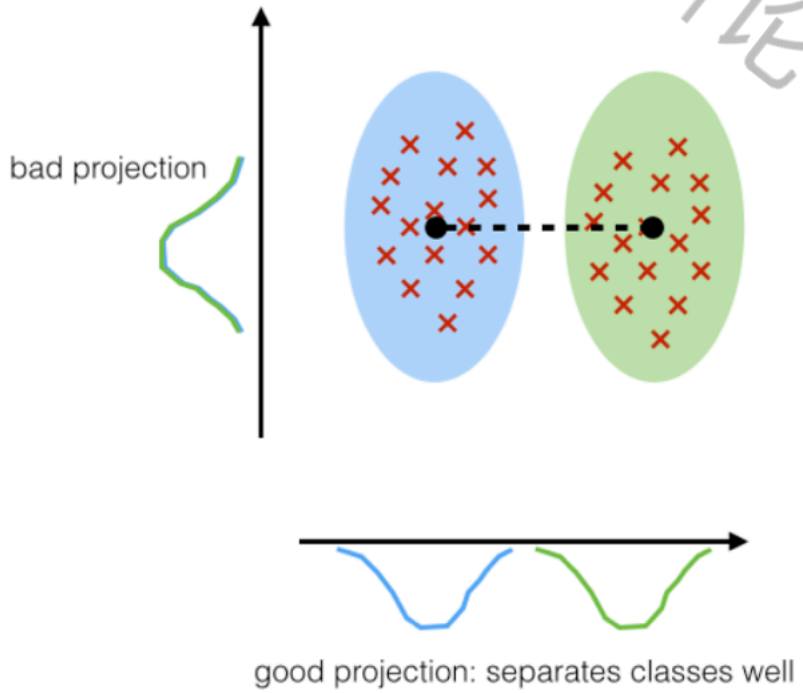


# 二、Linear Discriminant Analysis



The basic idea behind LDA:

- Learning a projection matrix W so that the within-class data points are as close as possible and between-class data points as far as possible.



# 二、Linear Discriminant Analysis

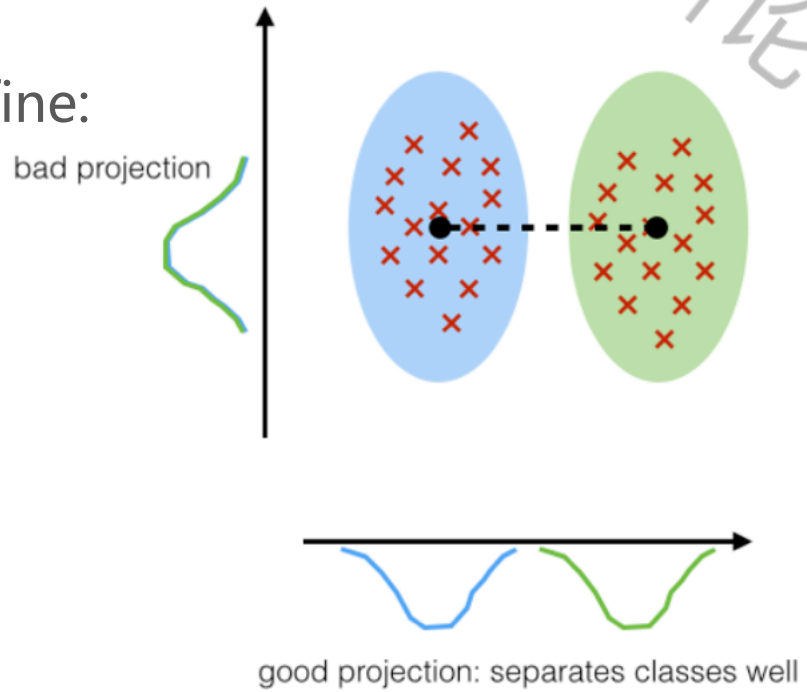


The basic idea behind LDA:

- Learning a projection matrix W so that the within-class data points are as close as possible and between-class data points as far as possible.

With the help of label, we could define:

- within-class scatter
- between-class scatter matrix



# 二、Linear Discriminant Analysis



Binary Class:

Define the within-class scatter for projected samples by $\tilde{s}_1^2 + \tilde{s}_2^2$, where

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{\mu}_i)^2 = \boldsymbol{w}^\top \underbrace{\left[ \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \right]}_{\boldsymbol{S}_i} \boldsymbol{w}.$$

# 二、Linear Discriminant Analysis

$y, \tilde{\mu}_i \in \mathcal{R}^d$
$\mathbf{x} \in \mathcal{R}^D$
$\mathbf{w} \in \mathcal{R}^{D \times d}$
$y \in \mathcal{R}^d$
$\mathbf{S}_i, \mathbf{S}_B \in \mathcal{R}^{D \times D}$



## Binary Class:

Define the within-class scatter for projected samples by $\tilde{s}_1^2 + \tilde{s}_2^2$, where

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{\mu}_i)^2 = \boldsymbol{w}^\top \underbrace{\left[ \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \right]}_{\boldsymbol{S}_i} \boldsymbol{w}.$$

and between-class scatter matrix:

$$\boldsymbol{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$$

where

$$y = \boldsymbol{w}^\top \boldsymbol{x} \quad \boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}. \quad \tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \mathcal{Y}_i} y = \frac{1}{N_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{w}^\top \boldsymbol{x} = \boldsymbol{w}^\top \boldsymbol{\mu}_i.$$

# 二、Linear Discriminant Analysis

LDA aims to find

$$\boldsymbol{w} = \arg\max_{\boldsymbol{w}} \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}},$$

# 二、Linear Discriminant Analysis



LDA aims to find

$$\boldsymbol{w} = \arg\max_{\boldsymbol{w}} \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}},$$

$$\boxed{\arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}} \Rightarrow \boxed{\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}}$$

(generalized eigenvalue problem).

Clearly, the optimal **w** consists of *K*-1 largest eigenvector of $\mathbf{S}_W^{-1}\mathbf{S}_B$

where *K* denotes the class number.

# 二、Linear Discriminant Analysis



## Extension to Multiple Classes:

Within-class scatter matrix

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top .$$

Between-class scatter matrix

$$\boldsymbol{S}_B = \sum_{i=1}^{K} \sum_{C_i} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top = \sum_{i=1}^{K} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top .$$

$$\text{Rank}(\boldsymbol{S}_B) \leq K - 1, \quad \text{Rank}(\boldsymbol{S}_W) \leq N - K, \quad \text{Rank}(\boldsymbol{S}_T) \leq N - 1.$$

# 二、Linear Discriminant Analysis

$y, \tilde{\mu}_i \in \mathcal{R}^d$

$\mathbf{x} \in \mathcal{R}^D$

$\mathbf{w} \in \mathcal{R}^{D \times d}$

$y \in \mathcal{R}^d$

$\mathbf{S}_W, \mathbf{S}_B \in \mathcal{R}^{D \times D}$

$\tilde{\mathbf{S}}_W, \tilde{\mathbf{S}}_B \in \mathcal{R}^{d \times d}$



Define

$$\widetilde{\boldsymbol{S}}_W = \sum_{i=1}^{K} \sum_{\boldsymbol{y} \in \mathcal{Y}_i} (\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}_i)^\top$$

$$\widetilde{\boldsymbol{S}}_B = \sum_{i=1}^{K} N_i (\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}})^\top .$$

One can easily show that

$$\widetilde{\boldsymbol{S}}_W = \boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W},$$

$$\widetilde{\boldsymbol{S}}_B = \boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W}.$$

# 二、Linear Discriminant Analysis



FLD seeks $K-1$ discriminant functions $\boldsymbol{W}$ such that $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$:

$$\begin{aligned}
\boldsymbol{W} &= \arg\max_{\boldsymbol{W}} \mathcal{J}_{FLD} \\
&= \arg\max_{\boldsymbol{W}} \operatorname{tr}\left\{\widetilde{\boldsymbol{S}}_W^{-1}\widetilde{\boldsymbol{S}}_B\right\} \\
&= \arg\max_{\boldsymbol{W}} \operatorname{tr}\left\{\left(\boldsymbol{W}^\top \boldsymbol{S}_W \boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^\top \boldsymbol{S}_B \boldsymbol{W}\right)\right\},
\end{aligned}$$

leading to

$$\boxed{\arg\max_{\boldsymbol{W}} \mathcal{J}_{FLD}} \Rightarrow \boxed{\boldsymbol{S}_B \boldsymbol{w}_i = \lambda_i \boldsymbol{S}_W \boldsymbol{w}_i}.$$

generalized eigenvalue problem

# 二、Linear Discriminant Analysis

Step 1: Computing the mean vectors in the input space;

Step 2: Computing the Scatter Matrices, i.e., **within** and **between-class scatter** matrix.

Step 3: compute the K-1 largest eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ as the project matrix.

Step 4: obtain the low-dimensional representation by $\mathbf{W}^{\mathrm{T}}\mathbf{x}$.

# Taking Home

Q1: The definition of Correlation in the context of CCA

Q2: Generalize ED

Q3: within-class scatter

Q4: between-class scatter

# Test Questions:



Q1: what are advantages of CCA over PCA?

Q2: what are the limitations/requirement of CCA?

Q3: the objective function of CCA, and how to solve it? (binary and multiple case)

Q4: implement CCA

Q5: what is the major difference between LDA and PCA, give two at least.

Q6: what the key idea of LDA and how LDA utilize the label to perform DR?

Q7: the objective function of LDA, and how to solve it? (binary and multiple case)

Q8: why LDA could reduce the data into a K-1 dimensional space at most?

# Q&A

THANKS!