

統計降維法

黃志勝 (Tommy Huang)

義隆電子 人工智慧研發部

國立陽明交通大學 AI學院 合聘助理教授

國立台北科技大學 電資學院合聘助理教授



分類或回歸

自變數/獨立變數
(Independent variable)

$$x_1, x_2, \dots, x_d$$

預測/
分類

$$f(x_1, x_2, \dots, x_d)$$

應變數/相依變數
(Dependent variable)

$$y$$

* 應用的問題和收集的資料

Regression:

應變數: 薪資(y) 自變數(特徵): 年資(x_1)、領域(x_2)

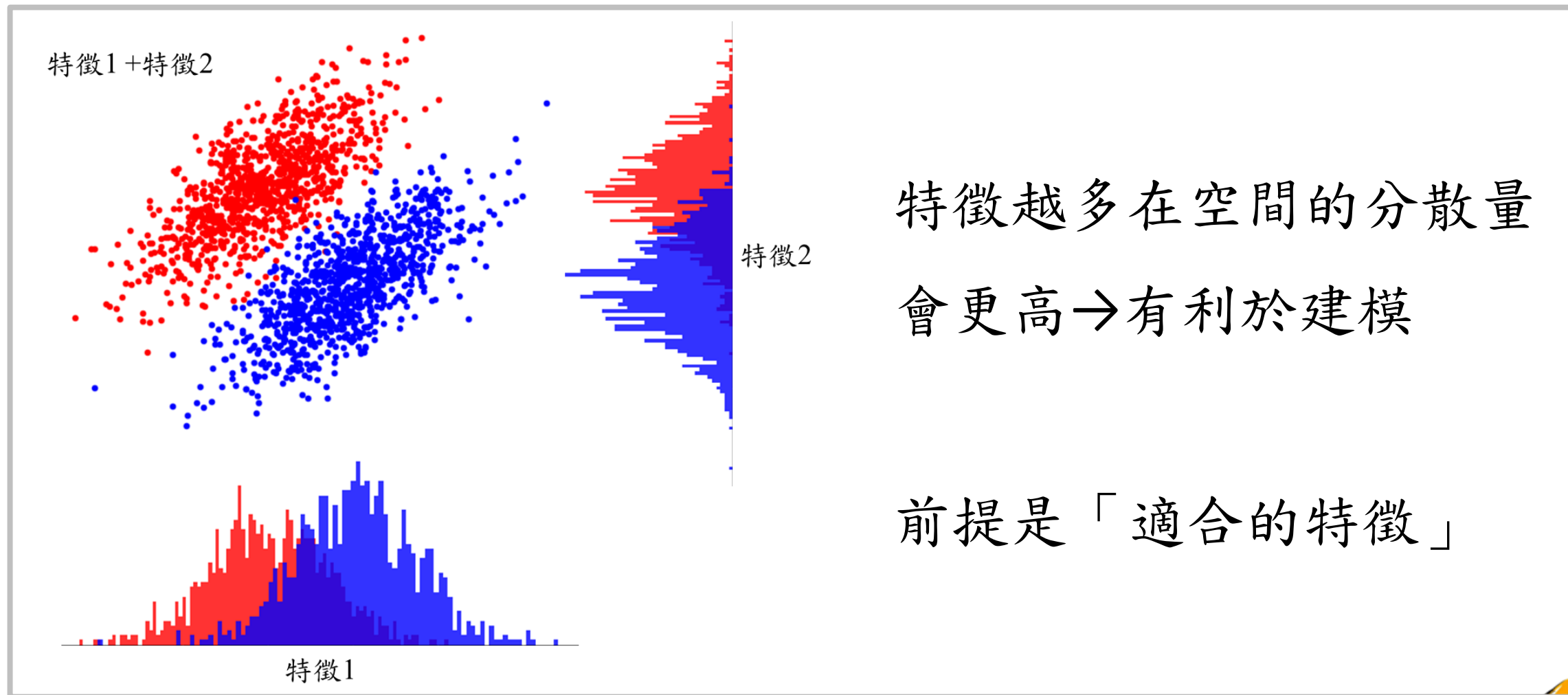
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \mathbf{x}^T \boldsymbol{\beta}$$

特徵2個 → 參數3個。

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

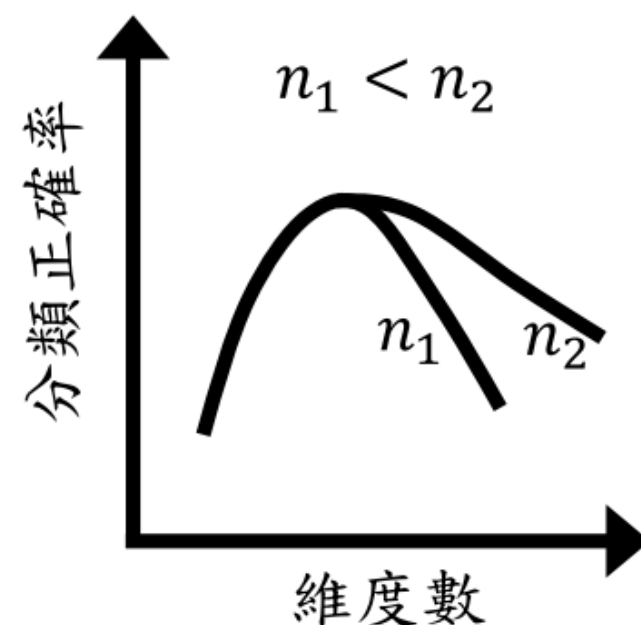
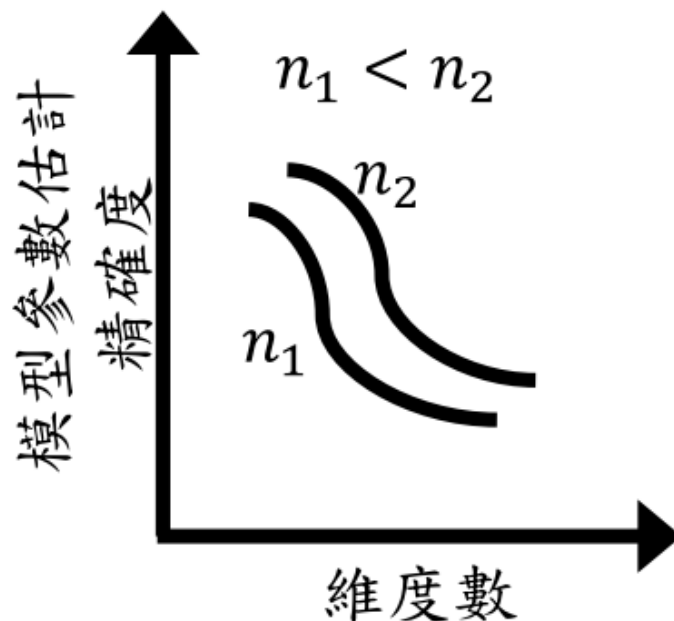
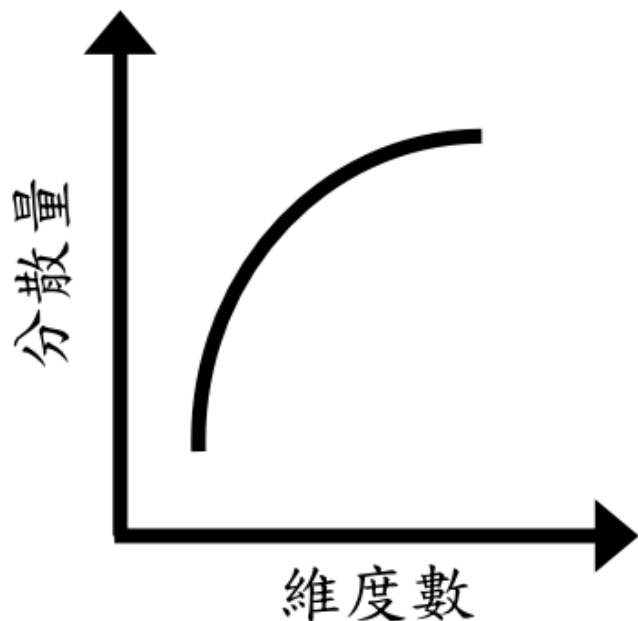


特徵數過多的問題



特徵數過多的問題

「休斯現象(Hughes phenomenon)」，也稱為「維度詛咒(Curse of dimensionality)」



特徵數過多的問題(參數>資料量)

Regression:

應變數: 薪資(y) 自變數(特徵): 年資(x_1)、領域(x_2)

$$y = \mathbf{x}^T \boldsymbol{\beta}$$

年資(x_1): 0~30年。 領域(x_2): 0 (非資訊業)、1 (資訊業)

薪資(y): 3~30萬

- 資料收集只有1筆 $(x_1, x_2, y) = (2, 1, 5)$ 。

$$\boldsymbol{\beta} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}$$

$$(\mathbf{X}\mathbf{X}^T)^{-1} = \left(\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} [1 \quad 2 \quad 1] \right)^{-1} = \left(\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \right)^{-1} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

$\det(\mathbf{X}\mathbf{X}^T) = 0 \Rightarrow \mathbf{X}\mathbf{X}^T$ 奇異(Singular)，逆矩陣不存在



特徵數過多的問題(參數>資料量)

Regression:

應變數: 薪資(y) 自變數(特徵): 年資(x_1)、領域(x_2)

$$y = \mathbf{x}^T \boldsymbol{\beta}$$

年資(x_1): 0~30年。 領域(x_2): 0 (非資訊業)、1 (資訊業)

薪資(y): 3~30萬

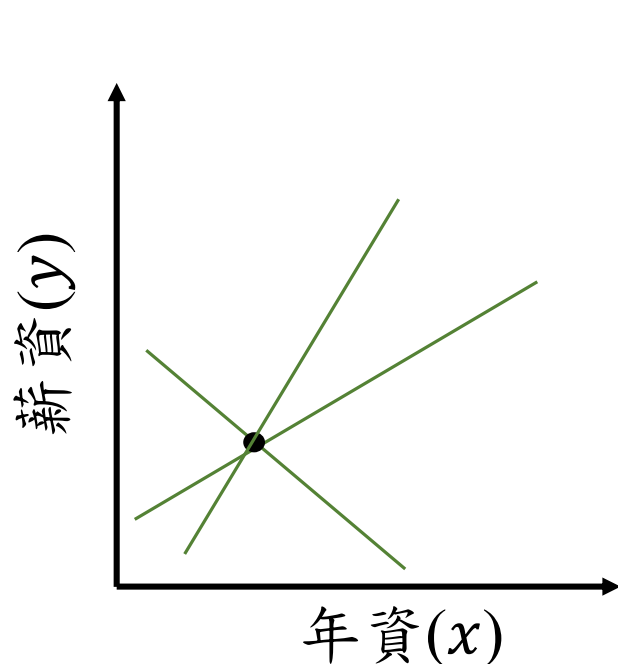
- 資料收集只有1筆 $(x_1, x_2, y) = (2, 1, 5)$ 。

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ \Rightarrow 5 &= \beta_0 + \beta_1 + 5\beta_2 \end{aligned}$$

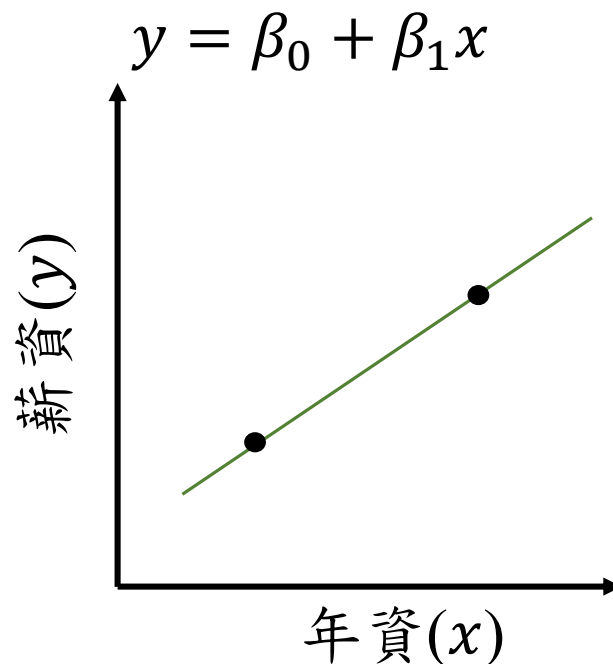
- 3元一次方程式，此方程式的解有無限多組。
- 所以從國中數學可以得知，三個參數至少要3筆資料才能找到唯一解。



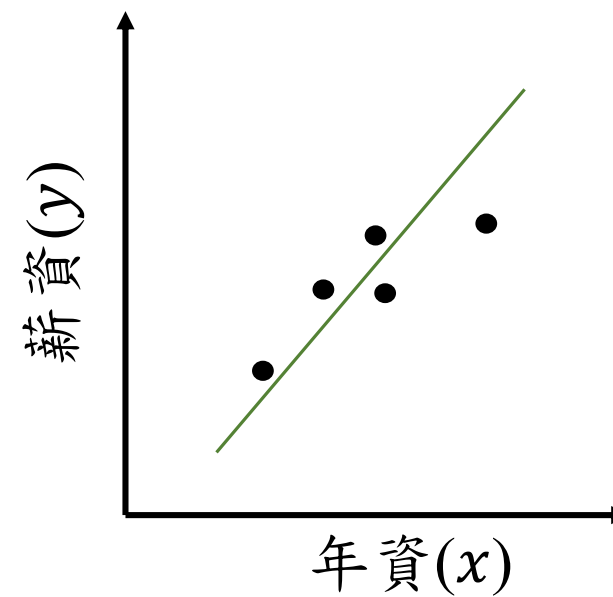
資料量(n) vs 參數(d)



參數兩個資料一筆
無窮多組解
 $n < d$



參數兩個資料兩筆
唯一解
 $n = d$



參數兩個資料五筆
近似解
 $n > d$



維度問題

Example:

Model 1: “body fat (bf)”

Model 2: “body fat (bf)”, “weight (w)”, “hair length (hl)”

$$\text{cov}(\text{model1}) = [\text{cov}(bf, bf)]$$

$$\text{cov}(\text{model2}) = \begin{bmatrix} \text{cov}(bf, bf) & \text{cov}(w, bf) & \text{cov}(hl, bf) \\ \text{cov}(bf, w) & \text{cov}(w, w) & \text{cov}(hl, w) \\ \text{cov}(bf, hl) & \text{cov}(w, hl) & \text{cov}(hl, hl) \end{bmatrix}$$



Example for single variable

If we only get one sample, and try to calculate covariance.

$$\mu = x_i$$

$$\text{cov}(\text{model } 1) = \sigma = \frac{1}{1} \sum_{i=1}^1 (x_i - \mu_x)^2 = 0$$

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Example for multi-variables

If we only get two sample, and try to calculate covariance matrix.

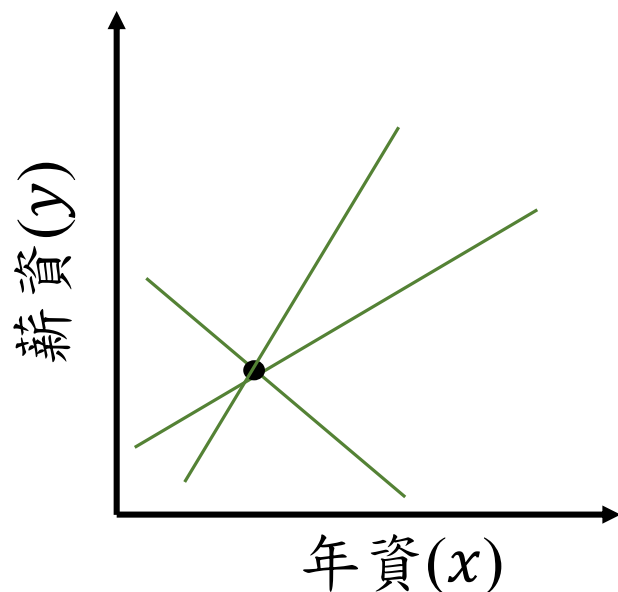
$$\Sigma = \begin{bmatrix} \text{cov}(bf, bf) & \text{cov}(w, bf) & \text{cov}(hl, bf) \\ \text{cov}(bf, w) & \text{cov}(w, w) & \text{cov}(hl, w) \\ \text{cov}(bf, hl) & \text{cov}(w, hl) & \text{cov}(hl, hl) \end{bmatrix}$$

The elements in covariance matrix are larger than 0, but the covariance matrix would be singular.

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-0.5} \exp\{-0.5(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}$$



現代AI(深度學習)



無窮多組解
 $n < d$

深度學習參數量



超過百萬個甚至到億個以上

深度學習: data-driven, 在限制的數據和假設下找到可能的近似解。

所以Learning algorithm都很容易因為資料產生模型的偏誤(bias)。



統計降維法

- 進行資料分析和統計建模分析第一件事情是希望樣本數要大於特徵數(維度數)。
- 統計資料降維：從中找到有用的資訊，進而提升建模能力。
- 統計資料降維方式可將屬性分成「特徵選取 (Feature Selection)」和「特徵萃取 (Feature extraction)」



特徵選取法

特徵

身高
體重
收入
家庭人口數
居住地
地址

哪個特徵
「有用」或「無用」



任務
分辨男生或是女生



特徵選取法

特徵		
身高	<div>1</div> <div>0</div>	2
體重	<div>1</div> <div>0</div>	×
收入	<div>1</div> <div>0</div>	2
家庭人口數	<div>1</div> <div>0</div>	×
居住地	<div>1</div> <div>0</div>	2
地址	<div>1</div> <div>0</div>	×
		2

最理想的狀況是利用窮舉法 (暴力法)
把所有組合都考慮過一輪

窮舉法: 特徵組合為 $2^d - 1$ 組。

6個特徵資料，則需要考慮到
 $2^6 - 1 = 64 - 1 = 63$ 個特徵組合

效率不佳 → 比較智慧一點的作法。

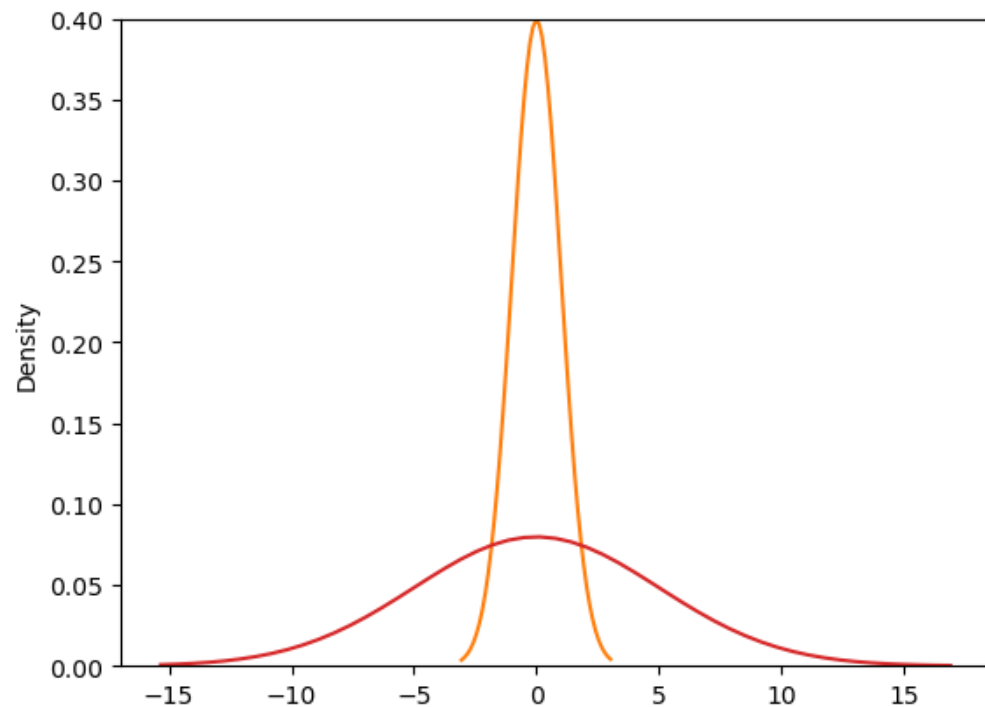


特徵選取法

- 刪除不合理的特徵: 例如變異數過低。
- 單變量特徵選擇 (Univariate feature selection):
 - 利用特徵和任務之間的統計(量)檢定進行單一變數的選擇。
 - 如果特徵彼此有關連性，此方法完全不考慮特徵的關聯性。
- 順序特徵選擇(Sequential Feature Selection):
 - 利用預先評估模型的手法，評估各種特徵組合的可能。



特徵選取法-刪除變異量最小的特徵資料



假設資料平均數都是0，變異數為0.5和5。

EX: 假設變異數是0。

這個特徵的每一筆資料數字都一樣。
→完全無用。

所以可以在這樣的假設下設定閾值濾除變異小的特徵(變數)。

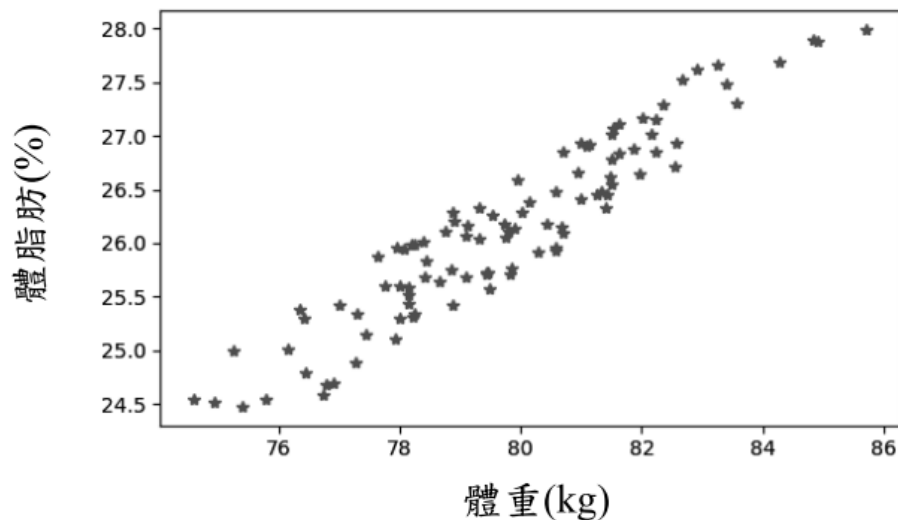
注意：容易被變數的單位影響。
例如:身高單位用公分變異量假設是10公分，但如果用成公里變異量會瞬間掉到0.0001公里。



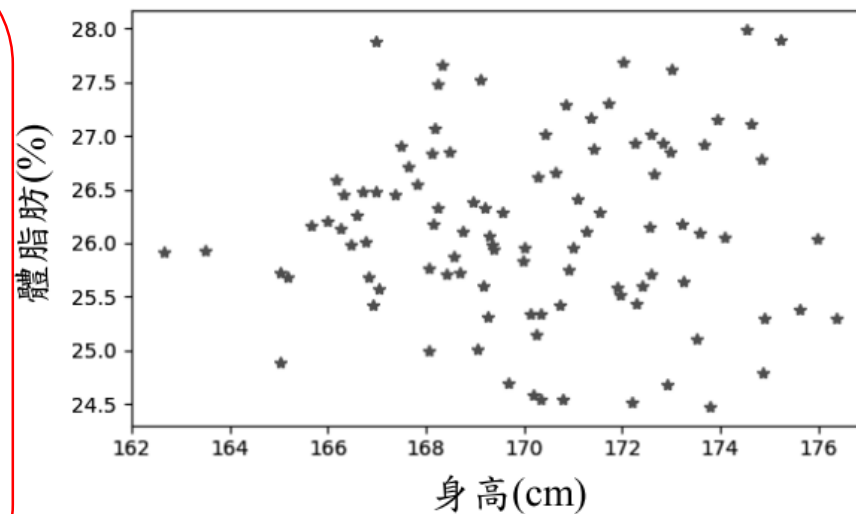
單變量特徵選擇-統計檢定-迴歸任務

任務相依方式

迴歸任務上可以採用**相關係數(r)**作為指標



$$r(\text{體重}, \text{體脂肪}) = 0.94$$



$$r(\text{身高}, \text{體脂肪}) = 0.02$$



單變量特徵選擇-統計檢定-迴歸任務

相關係數要多大才要選，有沒有個依據

相關係數的樣本分布近似自由度為 $n-2$ 的 t 分布，因此可由相關係數 r 得到 t 值為

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

有了 t 值就能計算 p 值。(見假設檢定內容)

在迴歸任務上在做的檢定就是

$$H_0: r(x, y) = 0$$

如果 $r < 0.05$ ，reject H_0 。



單變量特徵選擇-統計檢定-迴歸任務

假設經由計算得到5個特徵的 t 值與 p 值(以下相關係數的數值為示範用)：

- 體重和體脂肪的相關係數(r)為0.94 $\rightarrow t = 27.2750 \Rightarrow p \approx 0.000$
- 身高和體脂肪的相關係數(r)為0.02 $\rightarrow t = 0.1985 \Rightarrow p \approx 0.843$
- 性別和體脂肪的相關係數(r)為0.3 $\rightarrow t = 3.1132 \Rightarrow p \approx 0.002$
- 收入和體脂肪的相關係數(r)為0.1 $\rightarrow t = 0.9949 \Rightarrow p \approx 0.322$
- 年紀和體脂肪的相關係數(r)為0.4 $\rightarrow t = 4.321 \Rightarrow p \approx 0.000$

假設我們閾值設定在0.05，所以我們可以說我們在95%的信心水準下認為「體重」、「性別」、「年紀」和體脂肪預測是有相關的。



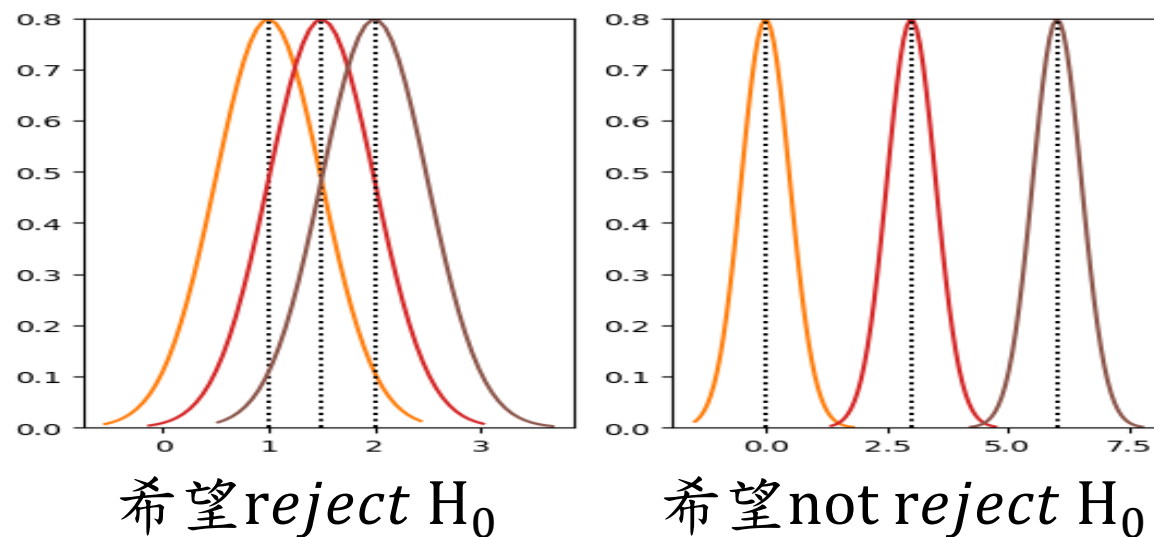
單變量特徵選擇-統計檢定-分類任務

分類任務上，可以利用卡方檢定(Chi-square)或是變異數分析(Analysis of variance, ANOVA)來檢定每個特徵不同類別的差異度。

卡方統計量可以用來量測特徵變數和分類類別的**相關性**。

$$H_0: \mu_1 = \mu_2 = \dots = \mu_c$$

$$H_1: \mu_i \neq \mu_j, i \neq j; i, j = 1, \dots, c$$



單變量特徵選擇-統計檢定-分類任務

假設經由訓練資料計算得到(以下p值為示範用)：

- 當特徵為體重時，檢定後的 $p=0.000$
- 當特徵為身高時，檢定後的 $p=0.001$
- 當特徵為收入時，檢定後的 $p=0.003$
- 當特徵為年紀時，檢定後的 $p=0.45$

$$\text{性別} = f \left(\begin{bmatrix} \text{體重} \\ \text{身高} \\ \text{收入} \end{bmatrix} \right)$$

統計上顯著的差異($p < 0.05$)，取「體重」、「身高」、「收入」來做分類即可



順序特徵選取

順序特徵選取法(Sequential feature selection)

- 向前順序特徵選取法(forward sequential feature selection)

向前特徵選取法做法是秉持「一次挑選一個特徵」進行辨識

- 向後順序特徵選取法(backward sequential feature selection)

向後特徵選取法的做法是秉持「一次刪除一個特徵」



向前順序特徵選取




6個小朋友，比賽跑100公尺老師要找出幾位小朋友，之後代表班上去跑大隊接力賽跑。但要派幾位小朋友學校還沒說，老師只好實驗找最好的組合。

老師做的是看誰行

假設要一個小朋友

 15秒  16秒



 12秒  14秒

 11秒  13秒

假設要兩個小朋友




  26秒   26秒




  24秒   24秒

  23秒

假設要三個小朋友

   40秒

   35秒

   39秒

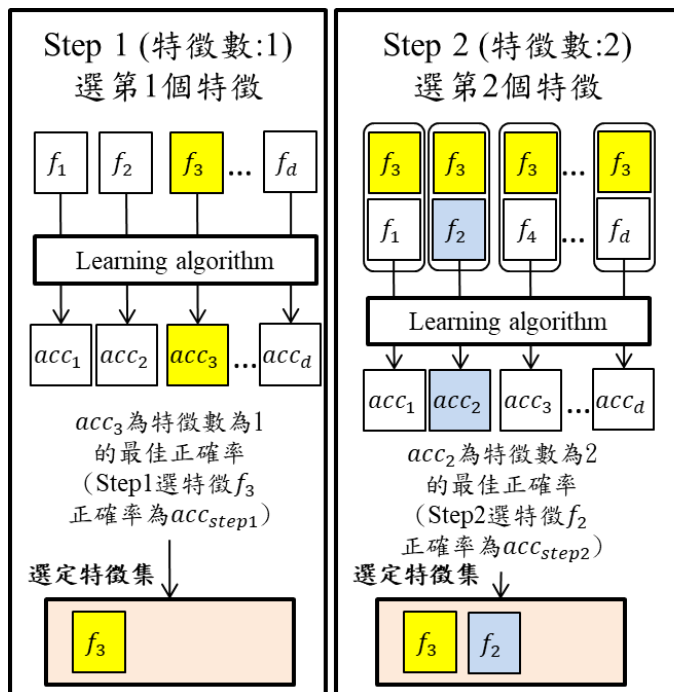
   36秒



向前順序特徵選取

- 向前順序特徵選取法(forward sequential feature selection)

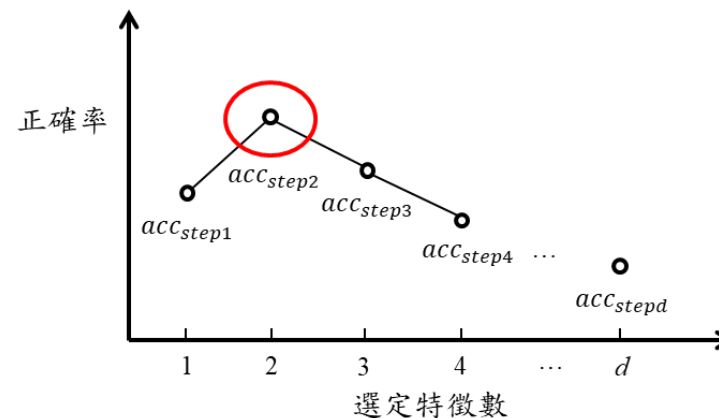
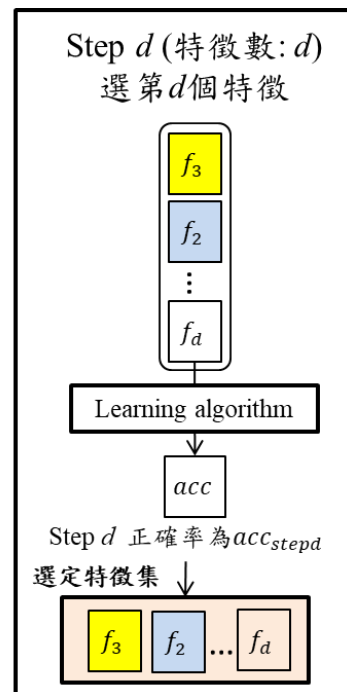
向前特徵選取法做法是秉持「一次挑選一個特徵」進行辨識



直到所有特徵都被選過

...

Step i 最佳正確率為
 acc_{stepi}



正確率最高為特徵數2個
依據向前順序特徵選取法
選取的特徵為 f_3 和 f_2







































向後順序特徵選取




























6個小朋友，比賽跑100公尺老師要找出幾位小朋友，之後代表班上去跑大隊接力賽跑。

















假設5位小朋友上場

						70秒
						69秒
						66秒
						72秒
						65秒
						60秒

假設4位小朋友上場

					50秒
					49秒
					47秒
					49秒
					48秒

假設3位小朋友上場

				40秒
				38秒
				36秒
				47秒

老師做的是看誰不行

向後順序特徵選取法(backward sequential feature selection)



特徵選取法-小結

1.刪除不合理的特徵:

刪除變異量最小的特徵資料: 把可能沒有用的資料刪除, 但需要注意變數的單位。

2.單變量特徵:

優點: 快, 因為單純看特徵和任務之間的統計量, 結果較好解釋。

缺點: 組合起來不一定最合適, 因為完全不考慮特徵之間是否有關連性。

3.順序特徵選擇:

優點: 因為會考慮特徵之間的組合, 所以結果通常比較好。

缺點: 搜尋時間會很久, 因為每搜尋一個特徵就需要建立一次分類器。



Dimension Reduction

- Dimension Reduction is proposed to overcome this issue.
 1. Feature selection
Using only “import” features.
 2. Feature extraction
Feature Fusion.

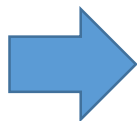


特徵萃取法

特徵 特徵選取法

身高	1	0
體重	1	0
收入	1	0
家庭人口數	1	0
居住地	1	0
地址	1	0

泛化



特徵萃取法

新特徵1 = w_1 身高 + w_2 體重 + w_3 收入 + w_4 家庭人口數 + w_5 居住地 + w_6 地址

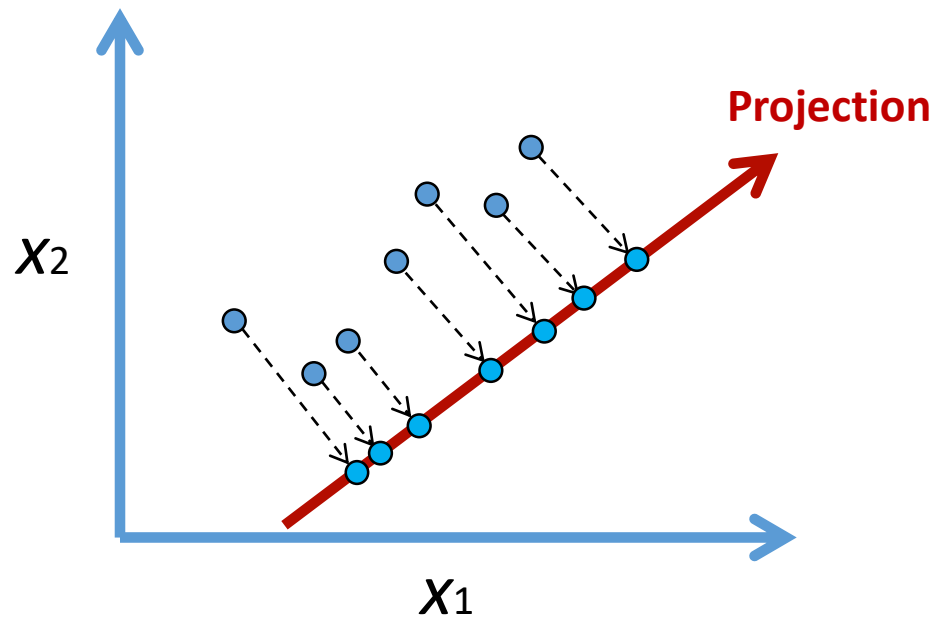
新特徵2 = w_1 , 身高 + w_2 , 體重 + w_3 , 收入 + w_4 , 家庭人口數 + w_5 , 居住地 + w_6 , 地址

...



Feature Extraction

- Feature Fusion (Projection)

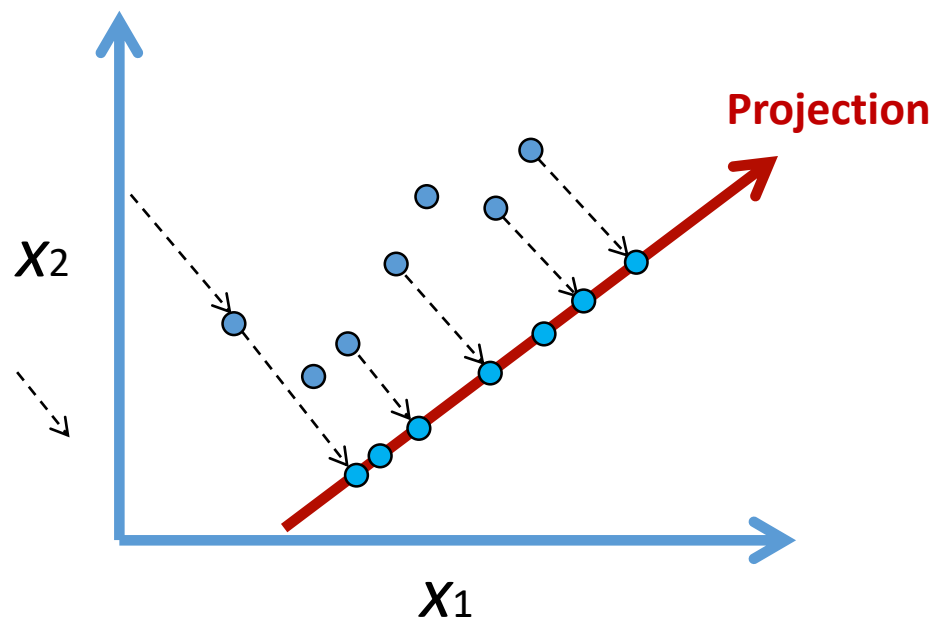


Feature extraction:
Just finding the projection vectors for input features.



特徵萃取法

- Feature Fusion (Projection)



利用新的1個特徵資料來表示兩個特徵資料。

在1-3矩陣拆解介紹中，有提到用少量的資料來還原表示大量資料。



Feature Extraction

- **Principal component analysis (PCA)**
- Independent component analysis (ICA)
- Canonical component analysis (CCA)
- Non-negative matrix factorization
- Discriminant Analysis Feature Extraction(DAFE)
- Neural Network



Principal component analysis (PCA)

Why do I introduce PCA?

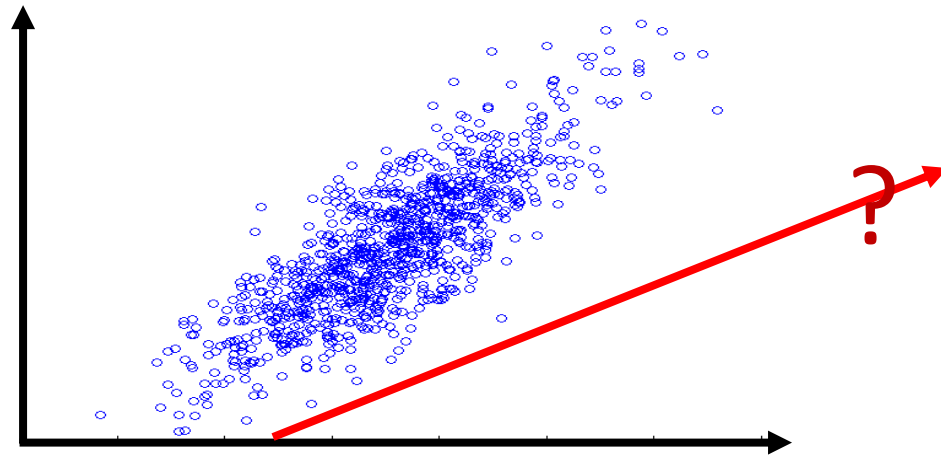
1. Stronger knowledge.
2. Unsupervised.
3. Most popular
4. Basic



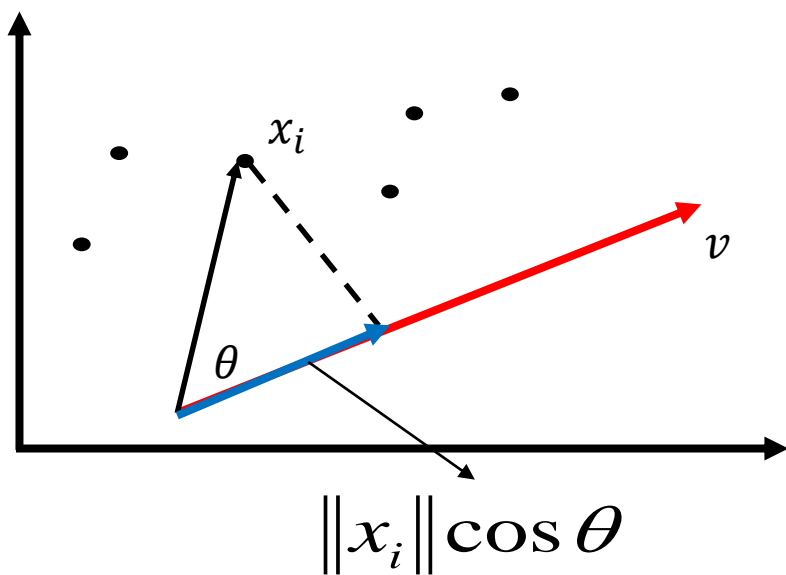
Principle Component Analysis

DL/ML/Statistics are developed by a given goal.

- PCA aims to find a set of vector containing the maximum amount of variance in the data.



Principle Component Analysis



$$\cos \theta = \frac{\langle x_i, v \rangle}{\|x_i\| \|v\|}$$

$$\begin{aligned} & \|x_i\| \cos \theta \frac{v}{\|v\|} \\ &= \|x_i\| \frac{\langle x_i, v \rangle}{\|x_i\| \|v\|} \frac{v}{\|v\|} = \frac{\langle x_i, v \rangle}{\|v\|^2} v \end{aligned}$$

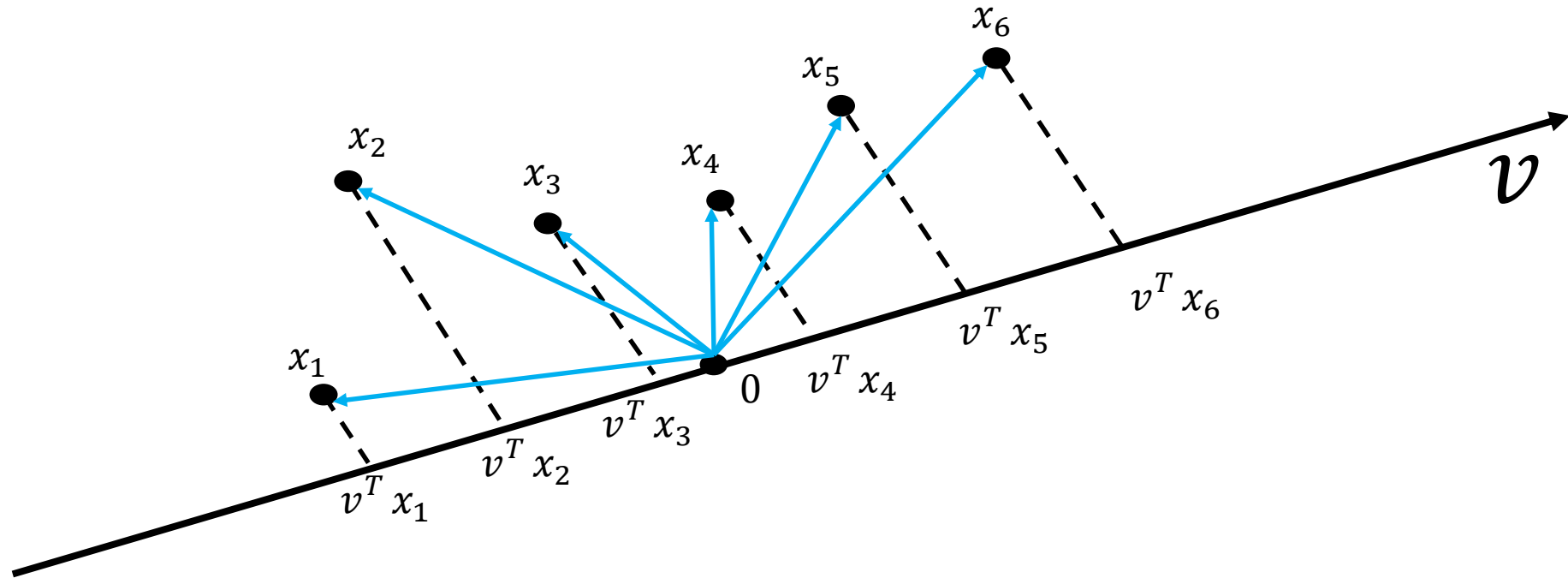
• If $\|v\|$ is unit, then $\langle x_i, v \rangle v$

$$\langle x_i, v \rangle = x_i^T v = v^T x_i$$

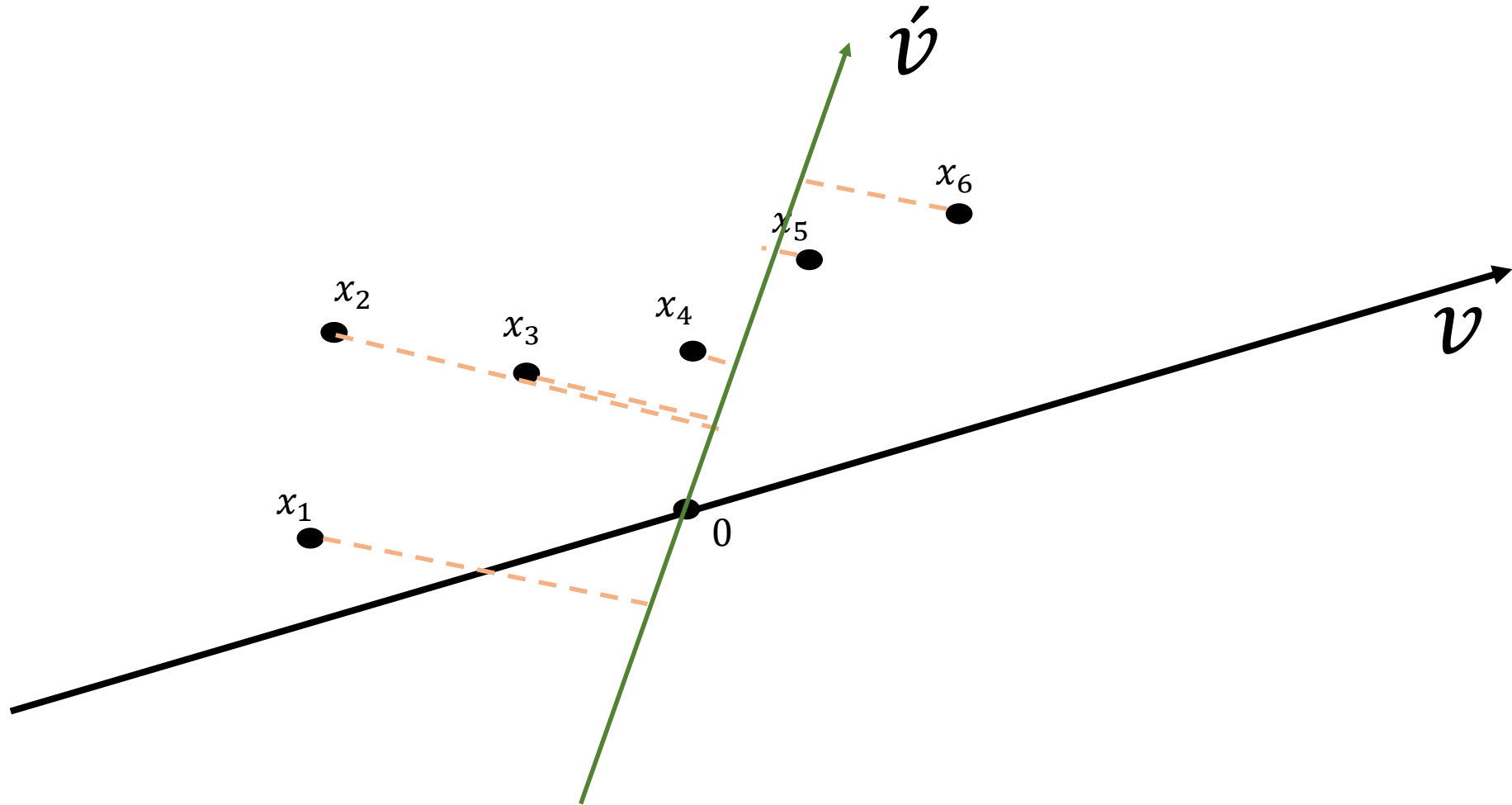
$$y_i = v^T x_i$$



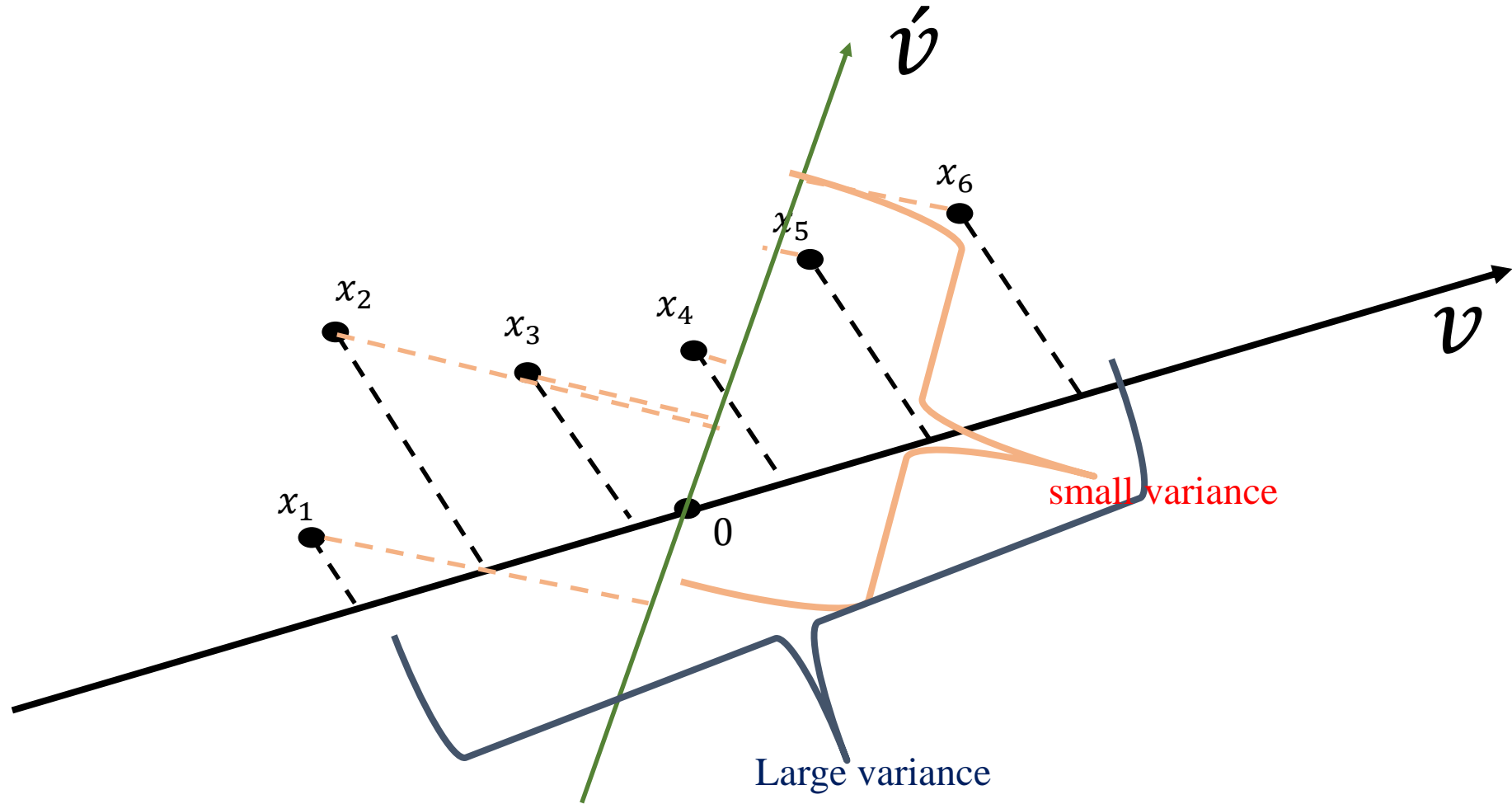
Principle Component Analysis



Principle Component Analysis



Principle Component Analysis



Principle Component Analysis

- The projections of the all points x_i into the direction v are
$$v^T x_1, v^T x_2, \dots, v^T x_N$$

The variance of the projections is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (v^T x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (v^T x_i - 0)^2 = \frac{1}{N} \sum_{i=1}^N (v^T x_i)^2$$

$$\begin{aligned} \Sigma &= \frac{1}{N} \sum_{i=1}^N (v^T x_i)(v^T x_i)^T = \frac{1}{N} \sum_{i=1}^N (v^T x_i x_i^T v) = v^T \left(\frac{1}{N} \sum_{i=1}^N x_i x_i^T \right) v \\ &= v^T C v \end{aligned}$$

C covariance matrix



Principle Component Analysis

- The first principal vector can be found by the following equation:

$$v = \arg \max_{v \in R^d, \|v\|=1} v^T C v$$



Principle Component Analysis

$$\mathbf{v} = \arg \max_{\mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{C} \mathbf{v}$$

Lagrange function:

$$f(\mathbf{v}, \lambda) = \mathbf{v}^T \mathbf{C} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)$$

$$\frac{\partial f(\mathbf{v}, \lambda)}{\partial \mathbf{v}} = 0 \Rightarrow 2\mathbf{C} \mathbf{v} - \lambda \mathbf{v} = 0 \Rightarrow \mathbf{C} \mathbf{v} = \lambda \mathbf{v}$$

$$\frac{\partial f(\mathbf{v}, \lambda)}{\partial \lambda} = 0 \Rightarrow \mathbf{v}^T \mathbf{v} - 1 = 0 \Rightarrow \mathbf{v}^T \mathbf{v} = 1$$



Principle Component Analysis

- The first principal vector can be found by the following equation:

$$v = \operatorname{argmax}_{v \in R^d, \|v\|=1} (v^T C v)$$

- This is equivalent to find the largest eigenvalue of the following eigenvalue problem:

$$Cv = \lambda v$$
$$\|v\| = 1$$

$$C = \frac{1}{N} \sum_{i=1}^N x_i x_i^T = \frac{1}{N} [x_1 \dots x_N] \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \frac{1}{N} X^T X$$



Principle Component Analysis

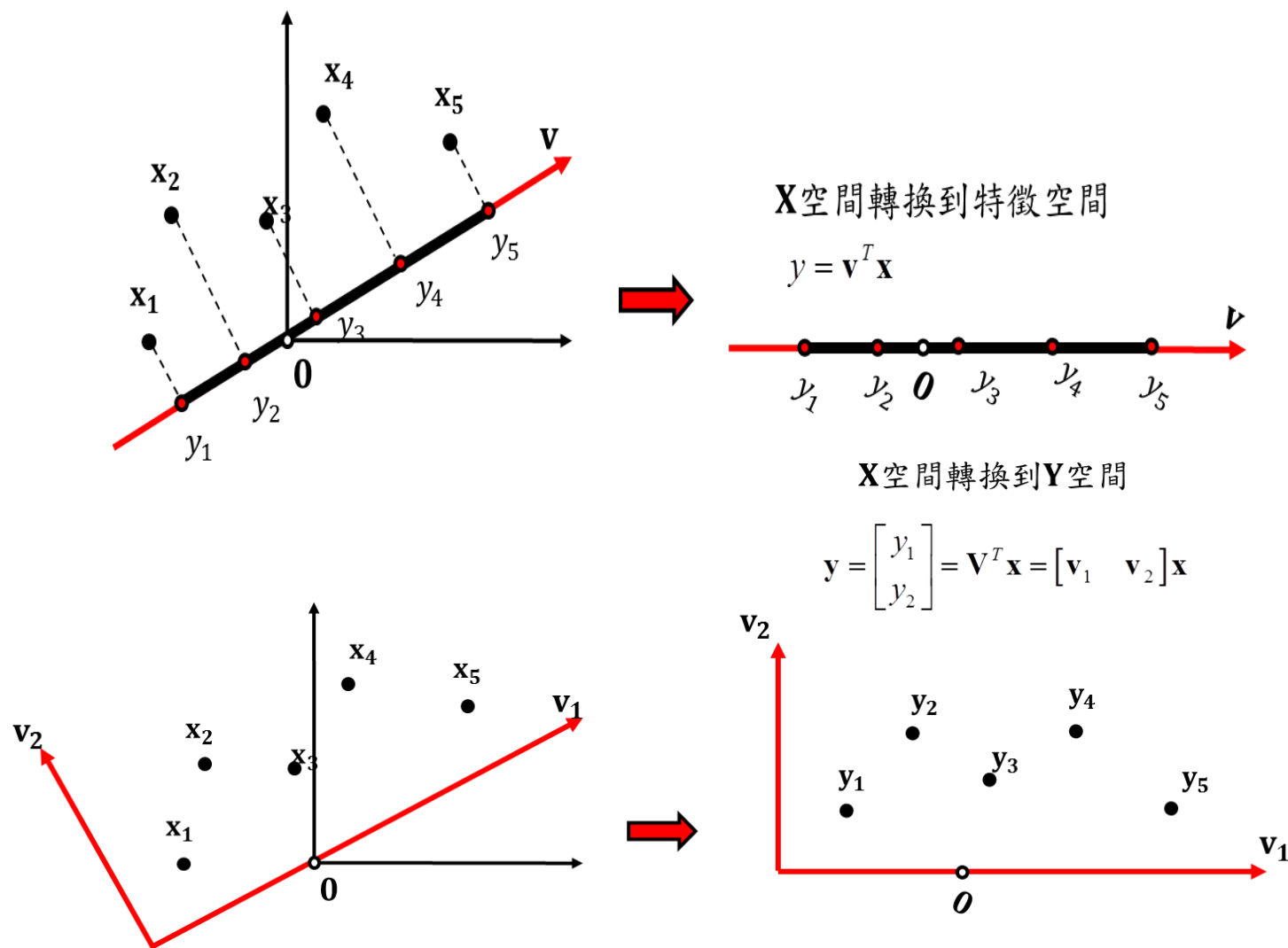
Eigenvalue vector is the corresponding variance vector.

$$\mathbf{v}^T \mathbf{C} \mathbf{v} = \sigma^2$$

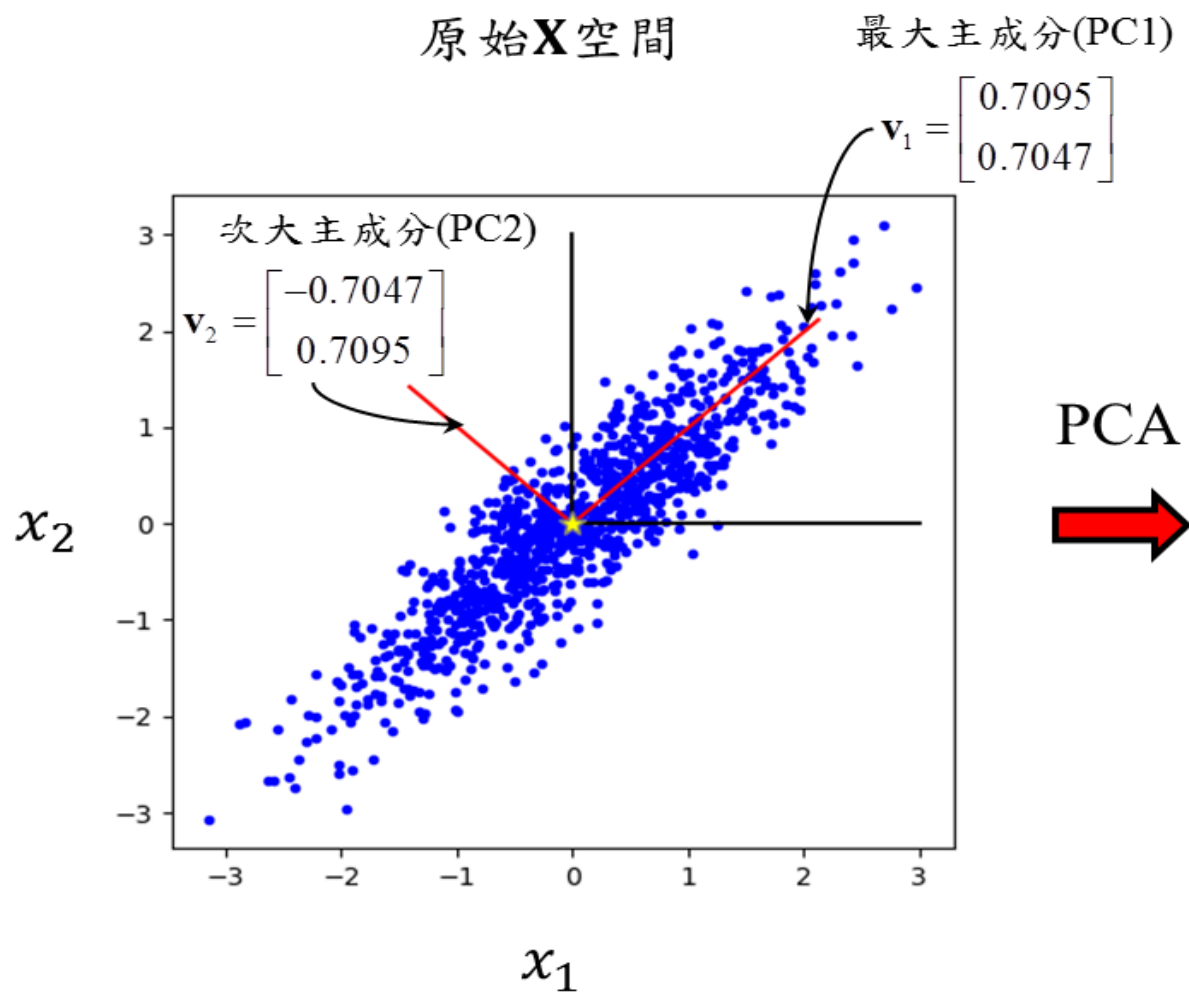
$$\Rightarrow \mathbf{v}^T \lambda \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda = \sigma^2$$



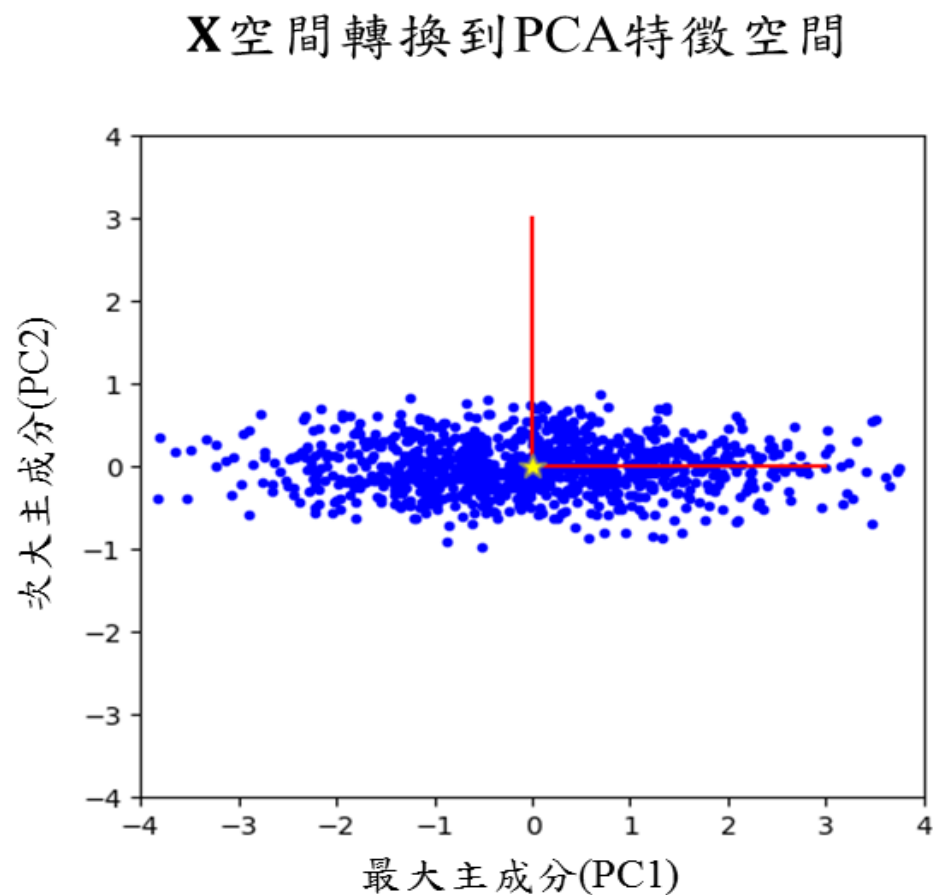
Projection



Exercise



PCA



統計學資料科學的觀點

- 統計學要怎麼決定多少主要成分出來?
- 答案是從由累積貢獻比率 (Cumulative Proportion)去決定需要取多少主要成分出來。
- 累積貢獻比率 (Cumulative Proportion)是什麼?



累積貢獻比率 (Cumulative Proportion)

- 假設有4個變數($X_1 \sim X_4$)，所以萃取出的主成份會有(PC1~PC4)，累積貢獻比率則是看前幾個主成份可以表是原始資料多少百分比的變異量。

	PC1	PC2	PC3	PC4
變異量	2.987	1.013	2.220e-15	6.955e-17
變異量百分比	74.675%	25.325%	5.55e-14%	1.73875e-15%
累積貢獻比率	74.675%	100.000%	100.000%	100%

兩個 P C 就夠代表整筆資料



特徵萃取法

- 特徵萃取法為特徵選取法的泛化演算法(Generalization)。
- 特徵萃取法在課程中介紹最常使用的主成分分析(PCA)。
- PCA就是求零平均數化的資料算出的共變異數矩陣(Σ_X)的特徵值分解。
- 找到的特徵值就是主成分對應的變異數，可以根據累積分布百分比來減少維度數進行後續的分析。

