R 語言和商業分析 -
洞悉商業世界中的資料科學

主成份分析

總結資料資訊：上市公司財報分析

黃大維　國立台灣大學統計碩士
沈辰禧　國立台灣大學商研碩士

洞悉商業世界中的資料科學

課程大綱

主成份分析方法

生活中常見的降維分析

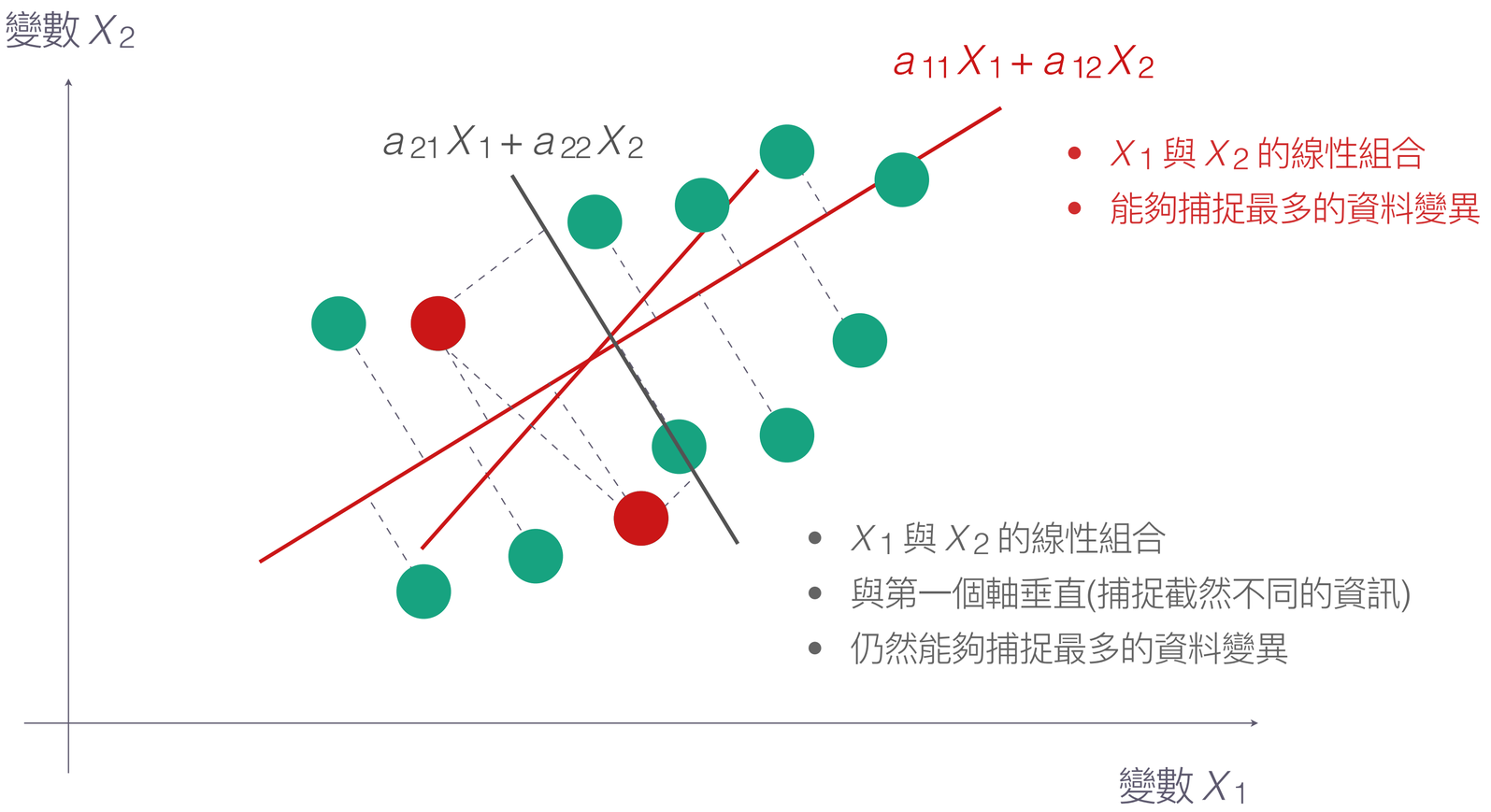
PCA 模型精神與估計

PCA 的資料前處理

PCA 分析結果詮釋

個案：上市公司財務分析

主成份分析 (PCA) 找出「有效表達資料」的新變數



- X_1 與 X_2 的線性組合
- 能夠捕捉最多的資料變異

- X_1 與 X_2 的線性組合
- 與第一個軸垂直(捕捉截然不同的資訊)
- 仍然能夠捕捉最多的資料變異

主成份分析 (PCA) 的輸入與輸出

演算法輸入 (Input)

p 個變數
(*p* variables)

- 變數間彼此相關 (correlated)
- 共蒐集 *n* 個觀察值

	變數 X_1	…	變數 X_p
個體 1	X_{11}	…	X_{1p}
個體 2	X_{21}	…	X_{2p}
…			
個體 *n*	X_{n1}	…	X_{np}

特徵值分解
(Eigen-decomposition)

演算法輸出 (Output)

p 個主成份
(*p* principal components)

- 每個主成份都是原始變數的加權平均
- 主成份彼此互不相關
- 越前面的主成份解釋越多資料變異
- 拋棄較後面的主成份進行降維

母體變數 $X_1, \ldots, X_p$ 透過「線性轉換」得到主成份 $Y_1, \ldots, Y_p$

p 個主成份

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\cdots \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

- 越前面的組成份解釋越多變異，$Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p)$
- 主成份間彼此不相關 (uncorrelated)，$Cor(Y_i, Y_j) = 0,\ i \neq j.$
- 每一個主成分的係數和為 1，$a_{i1} + a_{i2} + \cdots + a_{ip} = 1,\ \forall\ i = 1, \cdots, p.$

估計 a_{ij} 達成兩大目標：保持不相關與極大化變異數

$$Var(Y_i) = Var(a_{i1}X_1 + \cdots + a_{ip}X_p)$$

$$= a_{i1}^2 Var(X_1) + \cdots + a_{ip}^2 Var(X_p)$$

$$+2a_{i1}a_{i2}Cov(X_1, X_2) + \cdots + 2a_{ip-1}a_{ip}Cov(X_{p-1}, X_p)$$

1. 目標：極大化 Var (Y_i) （極大化上列公式）

2. 限制式：保持第 i 個主成份與前 (i - 1) 個主成份不相關。

加入其他限制式，得到更容易解釋的主成份

1. 目標：極大化 Var (Y_i) （原本的主成份分析目標）

2. 限制式：保持第 i 個主成份與前 (i - 1) 個主成份不相關。

3. 係數非負限制：(非負) 每一個係數 a_{ij} 都要大於或等於 0。

4. 係數稀疏限制：第 i 主成分係數 $a_{i1}, \ldots, a_{ip}$ 非零個數小於 k。