第十四讲（第18章） 主成分分析

(Principal Components Analysis，PCA)

简介

- 在管理上常会碰到需要以一群完整的变量來共同判断一个决策是否可行，但一方面又希望变量的计算不要过于复杂。
- 结合若干变量形成一个简单又具代表性的指标
- 变量做加权平均，如何加权？
  - 主成分分析

主成分分析

- 利用原有较多的变量，产生少数新的变量的方法
- 尽是保持原变量资料的信息
- 新的变量间需互相独立
- 变量个数适当缩减

主成成份分析的理论架构

- 几何架构
  - 以座标图表示一些观测点之间的关系
  - 所呈现的座标数最多三个
- 分析架构
  - 以数学关系来说明主成份与原有变量的关系及相关的统计量

几何架构

设X_1表入学成绩，X_2口试成绩

	X1		X2	
编号	原始成绩	$x_i - x$	原始成绩	$x_i - \bar{x}$
1	80	2	90	9
2	85	7	80	-1
3	70	-8	80	-1
4	65	-18	70	-11
5	90	12	85	4
6	75	-3	75	-6
7	80	2	85	4
8	90	12	75	-6
9	70	-8	80	-1
10	60	-18	75	-6
11	85	7	80	-1
12	80	2	95	14
13	75	-3	85	4
14	75	-3	75	-6
15	90	12	85	4
平均数	78	0	81	0
方差	85	85	43.57	43.57

两变量的总方差=128.57143
X1方差=85
X2方差=43.57

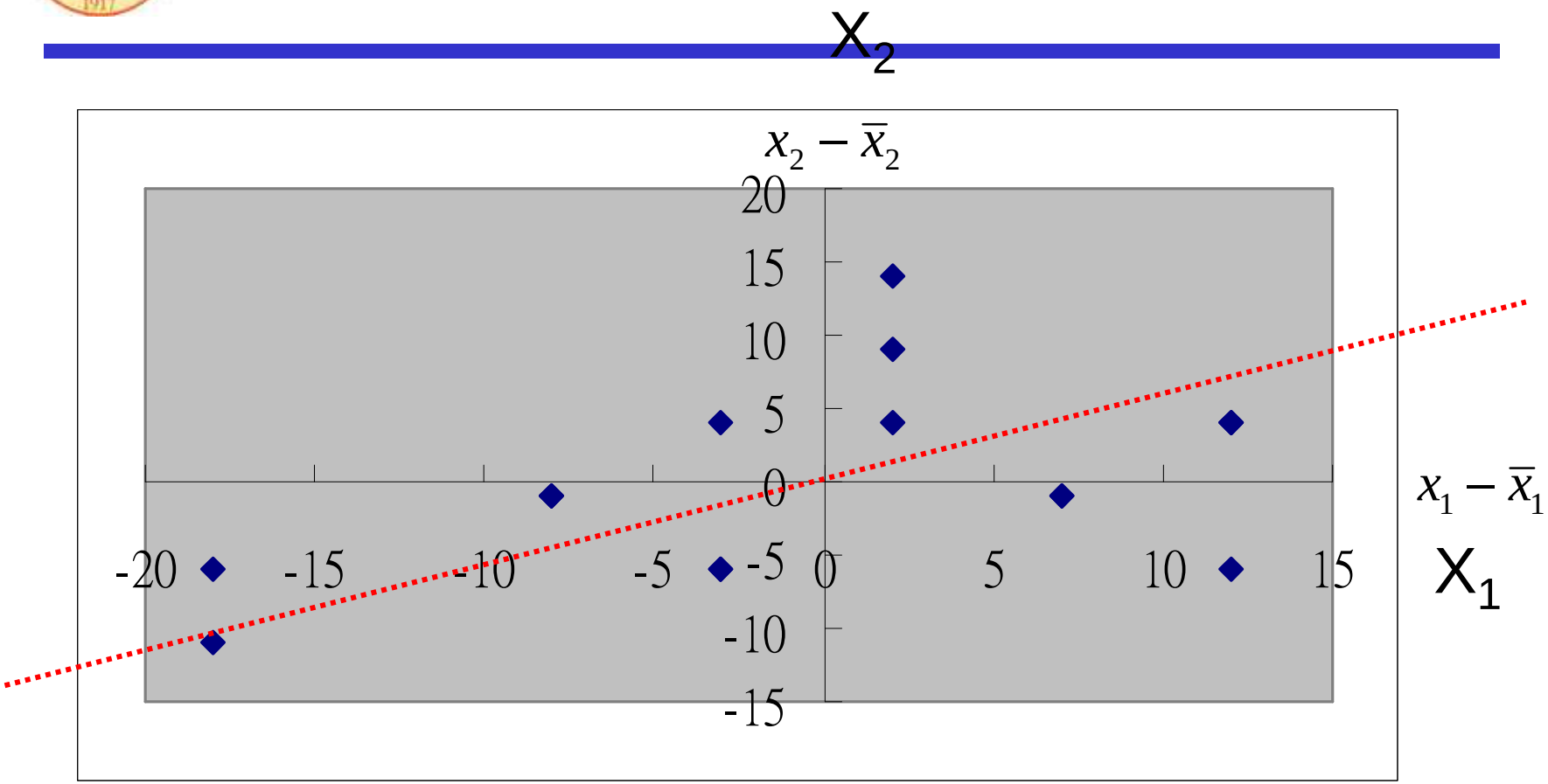
⇒X1方差占总方差之66%
X2方差占总方差之34%

方差-协方差矩阵(variance-covariance matrix)

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 85 & 25.357 \\ 25.357 & 43.57 \end{bmatrix}$$

相关系数矩阵(correlation coefficient matrix)

$$R = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.42 \\ 0.42 & 1 \end{bmatrix}$$


X_2
$x_2 - \bar{x}_2$
20
15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10
15
$x_1 - \bar{x}_1$
X_1

若将X1轴逆时针旋转θ
则对新轴的座标：

$$X_1^* = \cos\theta X_1 + \sin\theta X_2$$

$$X_2^* = -\sin\theta X_1 + \cos\theta X_2$$

X_1	X_2	X_1^*
2	9	4.96
7	-1	6.24
-8	-1	-7.86
-18	-11	-20.68
12	4	12.64
-3	-6	-4.87
2	4	3.25
12	-6	9.22
-8	-1	-7.86
-18	-6	-18.97
7	-1	6.24
2	14	6.67
-3	4	-1.45
-3	-6	-4.87
12	4	12.64
85	43.57	96.51

方差提高至75%
在新座标轴上更能将原变量的差异性突显出來

← 方差

不同旋转角度下的方差

旋转角	总方差	X_1*方差	百分比
0	128.57	85	66.11
10	128.57	92.42	71.88
20	128.57	96.45	75.02
30	128.57	96.60	75.14
40	128.57	92.85	72.22
50	128.57	85.66	66.62
60	128.57	75.89	59.02
70	128.57	64.72	50.34
80	128.57	53.49	41.61
90	128.57	43.57	33.89

取旋转角θ=25度

X1	X2	X1*	X2*
2	9	5.62	7.31
7	-1	5.92	-3.86
-8	-1	-7.67	2.47
-18	-11	-20.96	-2.36
12	4	12.57	-1.45
-3	-6	-5.25	-4.17
2	4	3.50	2.78
12	-6	8.34	-10.51
-8	-1	-7.67	2.47
-18	-6	-18.85	2.17
7	-1	5.92	-3.86
2	14	7.73	11.84
-3	4	-1.03	4.89
-3	-6	-5.25	-4.17
12	4	12.57	-1.45
85	43.75	97.3	31.55

方差

$$\Sigma^* = \begin{bmatrix} 97.03 & 0.4025 \\ 0.4025 & 31.55 \end{bmatrix}$$

$$R^* = \begin{bmatrix} 1 & 0.008 \\ 0.008 & 1 \end{bmatrix}$$

X_1^*，X_2^*非常接近互相独立的状况

⇒若角度为使 X_1^*方差最大值的旋转角
相关系数应为0

- 第一个轴可以解释最大可能比率的方差称为主成份，原样本点投影在这些轴上的值称为主成分分数(principal component scores)
- 新的变量为原变量之线性組合
- X1*与X2*方差数的总和等于X1与X2方差数的和
- X1*解释方差的比例已是尽可能最大，X2*则解释全部剩下的方差
- X1*,X2*的相关系数为0

- 若原始变量有许多个，利用主成份分析可将原有变量做适当转换，得到新的变量
- 新的变量中，第一个变量将尽可能解释原有资料的总方差，剩下的方差再依序由第2个，第3个....新变量解释
- 若少数几个新变量即可解释大部分的总方差，则原有的问题可用这几个新变量加以說明
- 新变量个数较原变量少，但整体解释能力沒有损失太多

分析架构

设 $x_1^*, x_2^*, \ldots$ 按可解释总方差的比例大小排列

即 $Var(x_1^*) > Var(x_2^*) \ldots$

$$x_1^* = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1n}x_n$$

$$x_2^* = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2n}x_n$$

$$\vdots$$

$$x_n^* = w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nn}x_n$$

考虑两个变量

$$x_1^* = w_{11}x_1 + w_{12}x_2$$

$$x_2^* = w_{21}x_1 + w_{22}x_2$$

$$\begin{aligned}
\max Var(x_1^*) &= Var(w_{11}x_1 + w_{12}x_2) \\
&= w_{11}^2 Var(x_1) + w_{12}^2 Var(x_2) + 2w_{11}w_{12}Cov(x_1, x_1) \\
&= \begin{bmatrix} w_{11} & w_{12} \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} \\
&= w_1^T \Sigma w_1
\end{aligned}$$

$\Rightarrow \quad \sigma_{11}^2, \sigma_{22}^2$ eigenvalue $\begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix}$ eigenvector

- 特征向量必为正交矩阵
- 协方差矩阵之特征值即为其所对应特征向量所组成之转换变量(主成份)的方差数
- 若变量单位不同，应将变量的资料先标准化

主成份个数的选取

- 取特征值大于全部平均
- 取特征值大于1(适用于标准化资料)
- 透过特征值排列图(陡坡图scree plot)， 选取开始变平缓的点所对应的个数

主成分分析的适用性

- 主成分分析其中的一个目的在于将一些原本互有关系的变量转换成互不相关的变量
- 原有变量相关性很低，主成分分析为多余

- 主成成份分析并非用来删除一部分原始变量
- 在变量转换过程中，每一个主成份(转换变量)都用到所有的原始变量

主成分分析与因子分析

- 相似处
  - 二者都具有将原有变量资料缩减成少数可以描述大部分原资料信息之功能
- 相异处
  - 主成成份分析主要利用原有的变量，组合成几个新的变量，最后选取的变量较原变量少，且这几个变量可尽可能解释原有资料大部分的方差
  - 因子分析在于寻找及确认可以解释原有变量间交互关系的潜在变量或建构(construct)

- 主成成份分析较偏重在分析及应用原有资料的方差
- 因子分析强调探讨原有变量间的交互影响关系
- 主成成份分析中，原有变量是用来組成新的变量，故又称为形成指标(formative indicators)
- 因子分析，原有变量是用来反映隐藏因素或建构的存在，故又称为反映指标(reflective indicators)，