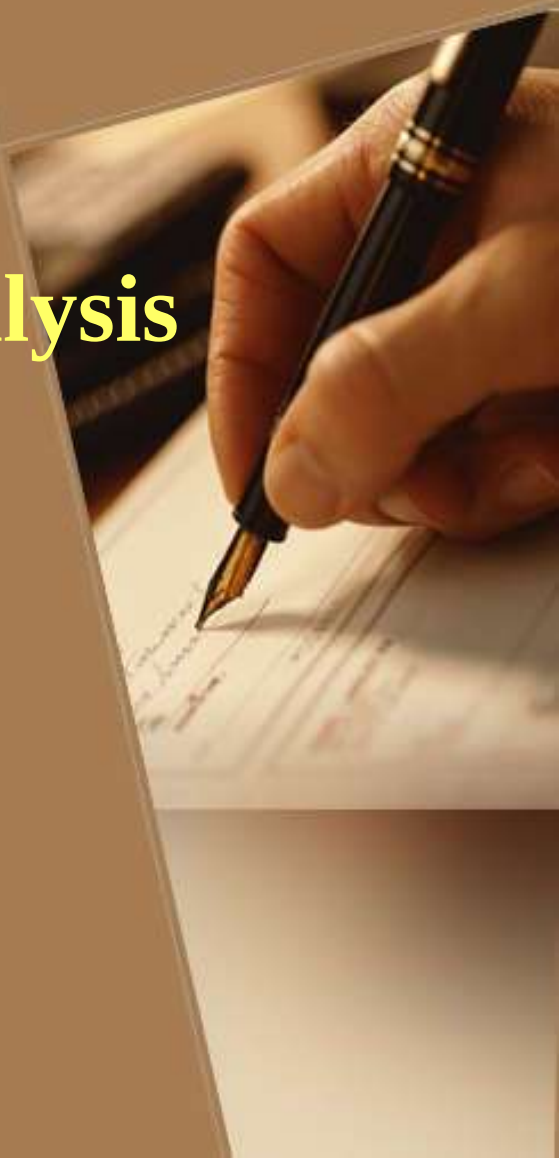
主成份分析

Principal Components Analysis



報告者：官淑蕙

2010.11.22

原　理

- 希望用較少的變數去解釋原始資料中的**最大變異**
- 期望將原本許多相關性很高的變數轉化成**互相獨立不相關的新變數**➔主成分
- p個變數，縮減到 m個主成份(m<p)
  同時儘量保留p個變數的**變異(variation)**

- 每個變數都是同等看待，並無獨立變數和應變數之分
- 對多個變數決定其權重而成加權平均，依此訂出**總指標**
- 經由**線性組合**而得的主成份，能保有原來變數最多的資訊
  - 使觀測值在主成份上，顯出**最大的個別差異**

PCA的目標：代表性、獨立性、精簡性

應　用

- 探索性資料分析的好工具
- 降低變數個數（將相關變數做簡化）
- 利用PCA可將原變數經「客觀加權」後轉化成為新的整體指標
  - 物價指數、學生成績、國家生產力等

數學架構

考慮p個變數的線性組合：

$$Y_j = a_{j1}X_1 + a_{j2}X_2 + \ldots + a_{jp}X_p \ ; \ j = 1,2,\ldots,p$$

其中 $Y_j,\ldots,Y_p$ 為P個主成分，a_{ij} 是第j個變數在第i個主成分的權重，權重要被估計，使得：

★ Y_1的變異數最大，Y_2的變異數是除了Y_1之外最大的，其餘類推。

★ $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1$，$i = 1,2\ldots,p$（使新變數的尺度固定）

★ 對所有 $i \neq j$，$a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ip}a_{jp} = 0$（使直交或不相關）

範例:
如何判定身材壯碩？

樣本編號	身高 (x_1)	體重 (x_2)
1	173	66
2	155	49
3	175	72
4	171	68
5	166	63
6	167	64
7	163	61
8	155	52
9	159	55
10	168	65
11	166	61
12	169	73
13	159	57
14	154	49
15	160	60

平均數與變異數

- 平均數(mean)

$$\overline{x}_1 = 164.0$$

$$\overline{x}_2 = 61.0$$

- 變異數(variance)

$$\left.\begin{array}{l} S_1^2 = 45.571 \\ S_2^2 = 56.429 \end{array}\right\} = 102.00$$

總變異

- 總變異= 102.00
  - ‘身高(X_1)’的變異= 45.571
  - ‘體重(X_2)’的變異= 56.429

主成份分析的運作方式 (1)

- 主成份分析的目的:
  - 將’身高(X_1)’與 ‘體重(X_2)’兩個變數重新組合在一起, 成為新的變數(Y)
  - “Y”也許是代表這個人
    “壯碩的程度”

主成份分析的運作方式 (2)

- 組合的方式是 “線性”，即:

$$Y = w_1 X_1 + w_2 X_2$$

且: $\sqrt{w_1^2 + w_2^2} = 1$ (normalized)

(所以w_1與w_2可以為負)

主成份分析的運作方式 (3)

- 例如:
  - $Y = 0.500X_1 + 0.866X_2$
  - $Y = 0.600X_1 + 0.800X_2$
  - $Y = 0.700X_1 + 0.714X_2$
  - $Y = 0.800X_1 + 0.600X_2$
  - $Y = 0.900X_1 + 0.436X_2$
  - $Y = 0.900X_1 - 0.436X_2$
  - ……

主成份分析的運作方式 (4)

- 新的變數(*Y*)必須能夠充分代表原來的變數 (身高 與 體重)

- 所謂 “充分代表” 是指新變數(*Y*)的變異數必須儘量與’身高’ 與 ‘體重’ 的總變異相當
(即儘量接近102.00)

主成份分析的運作方式 (5)

- 例如:
  - $Y = 0.500X_1 + 0.866X_2 \rightarrow V(Y) = 94.971$
  - $Y = 0.600X_1 + 0.800X_2 \rightarrow V(Y) = 98.257$
  - $Y = 0.700X_1 + 0.714X_2 \rightarrow V(Y) = 98.721$
  - $Y = 0.800X_1 + 0.600X_2 \rightarrow V(Y) = 95.217$
  - $Y = 0.900X_1 + 0.436X_2 \rightarrow V(Y) = 85.030$
  - $Y = 0.900X_1 - 0.436X_2 \rightarrow V(Y) = 10.250$
  - ……

主成份分析的運作方式 (6)

- 最後找出變異最大的組合:
  - $Y = 0.500X_1 + 0.866X_2 \rightarrow V(Y) = 94.971$
  - $Y = 0.600X_1 + 0.800X_2 \rightarrow V(Y) = 98.257$
  - $Y = 0.666X_1 + 0.746X_2 \rightarrow V(Y) = 98.951$ ✓
  - $Y = 0.700X_1 + 0.714X_2 \rightarrow V(Y) = 98.721$
  - $Y = 0.800X_1 + 0.600X_2 \rightarrow V(Y) = 95.217$
  - $Y = 0.900X_1 + 0.436X_2 \rightarrow V(Y) = 85.030$
  - $Y = 0.900X_1 - 0.436X_2 \rightarrow V(Y) = 10.250$
  - ……

主成份分析的運作方式 (7)

- 第一主成分 (1st Principle Component)
  - $Y_1=0.666X_1+0.746X_2$

- 但因$V(Y_1)=98.951$, 只解釋了總變異(102.00)的97.011% (98.951/102.00)
  另外還有3.049(2.989%)的變異有待進一步釐清

主成份分析的運作方式 (8)

- 第二主成分 (2nd Principle Component)
  - $Y_2=0.746X_1-0.666X_2$
  - $V(Y_2)=3.049$, 解釋了總變異(102.00)的 2.989%.
- $V(Y_1)+V(Y_2) = 98.951+3.049 = 102.00.$

到底可以萃取出幾個主成份?

- m個變數最多能萃取出m個主成份.
  - 例如’身高’ 與 ‘體重’ 兩個變數, 最多能萃取出2個主成份.
  - 但第2個主成份(Y_2)只解釋了總變異的2.989%, 有必要保留嗎?

特徵值

- 特徵值 (eigenvalue, λ)
  - $\lambda_1 = V(Y_1) = 98.951$
    (第一個主成份的變異數)
  - $\lambda_2 = V(Y_2) = 3.049$
    (第二個主成份的變異數)

    $(\lambda_1 \geq \lambda_2)$

使用SAS程式進行主成份分析

```
DATA    AA;
INPUT    X1    X2;
CARDS ;
173    66
155    49
175    72
171    68
166    63
167    64
163    61
155    52
159    55
168    65
166    61
169    73
159    57
154    49
160    60
;
PROC    PRINCOMP    COV    OUT = CC;
VAR    X1    X2;
PROC    CORR;    VAR    X1    X2;    WITH    PRIN1    PRIN2;
PROC    PRINT;    VAR    X1    X2    PRIN1    PRIN2;
RUN
```

Principal Component Analysis

15 Observations

2 Variables

Simple Statistics

	X1	X2
Mean	164.0000000	61.00000000
Std	6.7506613	7.51189533

Covariance Matrix

	X1	X2
X1	45.57142857	47.64285714
X2	47.64285714	56.42857143

Total Variance = 102

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	98.9511	95.9023	0.970109	0.97011
PRIN2	3.0489	.	0.029891	1.00000

特徵值

解釋量

Eigenvectors

	PRIN1	PRIN2
X1	0.665879	0.746060
X2	0.746060	-.665879

$$Y_1 = 0.666X_1 + 0.746X_2$$

CORRELATION ANALYSIS

2 'WITH' Variables: PRIN1 PRIN2

2 'VAR' Variables: X1 X2

Simple Statistics

Variable	N	Mean	Std Dev	Sum
PRIN1	15	0	9.94742	0
PRIN2	15	0	1.74610	0
X1	15	164.00000	6.75066	2460
X2	15	61.00000	7.51190	915.00000

Simple Statistics

Variable	Minimum	Maximum
PRIN1	-15.61151	15.53133
PRIN2	-4.26025	3.38514
X1	154.00000	175.00000
X2	49.00000	73.00000

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 15

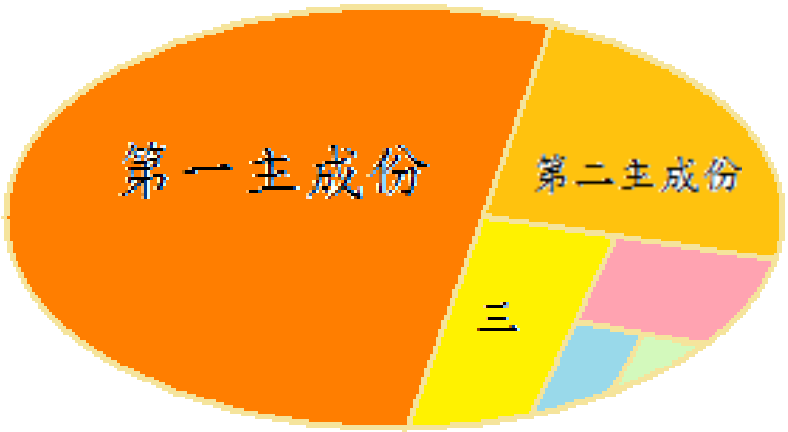
	X1	X2
PRIN1	0.98120	0.98795
	0.0001	0.0001
PRIN2	0.19297	-0.15478
	0.4908	0.5818

OBS	X1	X2	PRIN1	PRIN2
1	173	66	9.7232	3.38514
2	155	49	-14.9456	1.27601
3	175	72	15.5313	0.88199
4	171	68	9.8836	0.56127
5	166	63	2.8239	0.16036
6	167	64	4.2358	0.24054
7	163	61	-0.6659	-0.74606
8	155	52	-12.7074	-0.72163
9	159	55	-7.8058	0.26497
10	168	65	5.6478	0.32072
11	166	61	1.3318	1.49212
12	169	73	12.2821	-4.26025
13	159	57	-6.3136	-1.06678
14	154	49	-15.611	0.52995
15	160	60	-3.4096	-2.31836

I'm strong!

幾何意義

- 主成分分析就是要找一角度，使原變數的線性組合的變異數為最大。就像要對一群人照團體照，就要選一角度，使能分辨出每一個人(變異數最大、區辨力最大)。
- 主成分分析的幾何意義顯示主成份分析的目的就是要導出一組新的**直交**座標軸，使得



變異量 = $\sum \mathrm{Var}(y_1)$

第一個新變數解釋原資料最大的變異量

所得的P個新變數稱為主成份，觀察點投影至新軸的投影長即為主成份計分

第二個新變數解釋最多第一個新變數未能解釋的總變異

新變數為原變數的線性組合

第P個新變數解釋最多前P-1個新變數未能解釋的總變異

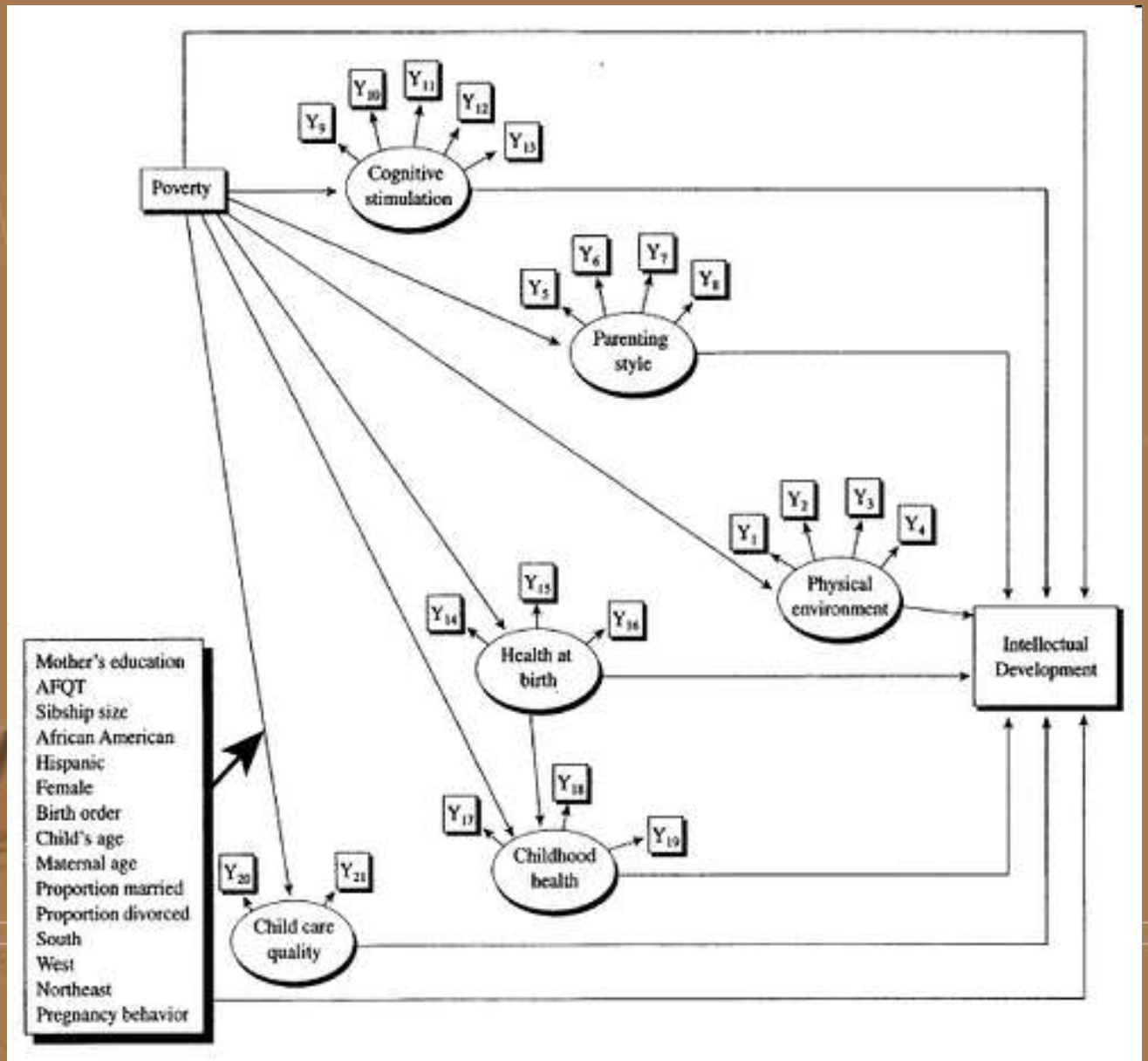
P個新變數彼此不相關

主成份 vs. 因素分析

共同因素分析與主成份分析的比較

比較內容	共同因素分析	主成份分析
I、相同部份		
1. 分析本質	兩者皆爲探索式性質	
2. 追求效果	兩者皆爲追求簡約效果	
II、相異部份		
目的本質 1. 關鍵名詞	因素	主成份
目的本質 2. 代表意涵	不可觀察的潛在構念	可觀察的線性組合變量
目的本質 3. 概念源起	Charles Spearman, 1904 年	Karl Pearson, 1901 年
目的本質 4. 理論發展	L. L. Thurstone, 1947 年	H. Hotelling, 1933 年
目的本質 5. 分析目的	因素萃取	資料縮減
目的本質 6. 著重焦點	探索觀察變項間的相關結構	解釋觀察變項的最大變異量
理論比較 1. 數學模式	分割共同因素與獨特因素	無共同成份與獨特成份之分
理論比較 2. 誤差項假定	獨立	無獨立假定
理論比較 3. 因素(成份)分數性質	爲潛在構念	爲觀察變項的迴歸預測項目
理論比較 4. 因素(成份)間關係	彼此相關	彼此獨立
理論比較 5. 共同性的估計	SMC 法、分割法、迭代法	無共同性的假定
理論比較 6. 相關矩陣的使用	縮減式相關矩陣	完整的相關矩陣
理論比較 7. 相關矩陣的複製	使用部分的共同因素	使用全部的成份
理論比較 8. 因素(成份)的唯一性	因素的唯一性未定	可決定成份的唯一性

因素分析


Poverty
Cognitive stimulation
Parenting style
Physical environment
Health at birth
Childhood health
Child care quality
Intellectual Development
Mother's education
AFQT
Sibship size
African American
Hispanic
Female
Birth order
Child's age
Maternal age
Proportion married
Proportion divorced
South
West
Northeast
Pregnancy behavior

參考資料

- 陳順宇。2005。多變量分析（四版）。華泰書局。
- 黃財尉。2003 。共同因素分析與主成份分析之比較。彰化師大輔導學報 25期：63-86。
- 主成分分析與因素分析。謝邦昌 教授。輔仁大學統計課講義。
- 主成份分析與因素分析（研究生2.0部落格）http://newgenerationresearcher.blogspot.com/2010/10/blog-post_29.html
- 因素分析（研究生2.0部落格）http://newgenerationresearcher.blogspot.com/2009/02/factor-analysis.html