主成分分析的原理

last modified May 3, 2006

觀察兩個變數間的相關性, 可以畫散佈圖。觀察三個變數的相關性, 也可以畫出 3-D 的立體圖來觀察。但是對於三個以上的變數, 在視覺上便無從觀察起, 即便是要計算變數間的相關係數, 也顯得繁複許多。事實上, 變數一多, 就可能發生某些變數間其實存在著相依性, 或是某些變數的影響程度非常微小, 但在一般的應用上, 往往因爲人爲的直覺判斷, 造成挑選出過多的變數。多變量分析提供許多工具, 試圖化繁爲簡, 降低變數的個數, 並能抽離出眞正的核心資訊, 其中「主成分分析」極具代表性。在主成分分析的過程中, 許多統計學與線性代數的基本觀念再度被應用到, 這個單元要從這些基本觀念開始。

1 背景介紹

1.1 從一個「評量表」說起

我們常常可以讀到有關城市評比的資料。譬如, 舊金山是美國「生活品質」最好的城市、香港是亞洲生活費最高的城市。通常主辦單位會爲評比的項目做一些定義, 然後根據定義一一去評分。下列是一份針對全美三百多個城市做「生活品質」調查的 9 項評量項目,

Climate / Housing Health / Crime / Transportation / education / Atrs / Research / Econmics

這些調查項目有些只要簡單的數據即可評分, 有些或許需要經過比較嚴謹複雜的程序才能得到。可想而知這樣的調查工作所需的人力、物力及時間消耗甚鉅。但另一方面, 從調查的項目來看, 有些項目之間似乎存在『相關性』。這些相關性會讓所量測到的資

料充斥著多餘的訊息 (Redundant information)。不過, 調查項目在選擇之初通常只是表面上的認知, 或不容易發現彼此間的關係, 這些關係往往要透過相關性的分析才會出現。
『主成分分析』可以用來分析調查項目 (或稱爲變數) 間的相關性。分析後的結果或許可以因爲發現某些變數間的相關性, 而縮減調查項目 (這當然進一步節省了調查資源的使用), 或是產生另一組數量較原變數少的新變數, 這個過程即所謂的 Dimension-reduced。新變數常呈現出新的意義, 是事先分析時不易或無法察覺的, 主成分分析便是從原始變數的資料中, 找到這層關係; 不但保留大部分的「訊息」, 也有效的降低變數的數量, 對後續的統計分析, 甚至圖表的表現都很大的助益。

從以上的案例, 很清楚的可以知道, 我們不能用任一變量來代表所有變量所呈現的資訊, 這是常識。但是如果將所有變量以適度的比例組合, 成爲一新的變量, 它能代表的資訊會比單一變量來得多。主成分分析便是在這一新變量上的產生上下功夫, 試圖以最少的變數代表原始資料最大的「成分 (變量)」, 其原則如下:

- 新變數爲原變數的線性組合。
- 保留原變數間的最大變異量 (variance)。

當一個新變數不足以代表於變數間的變異, 主成分分析也會以相同的原則產生第二個、第三個... 新變數, 直到新變數間的變異能涵蓋「大部分」原變數間的變異。這裡所謂的「大部分」無法定義的非常明確, 需情況而定, 通常在 70% $\sim$ 90% 之間便能滿足需求。

1.2 理論基礎

假設將原始變數 $x_1, x_2, \cdots, x_p$ 做線性組合, 轉換爲一組新的變數 $z_1, z_2, \cdots, z_p$,

$$\begin{aligned}
z_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\
z_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\
\vdots &= \vdots \\
z_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p
\end{aligned}$$

或表示為

$$\mathbf{z} = A\mathbf{x} \tag{1}$$

其中

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

矩陣 A 從幾何的角度來看, 也稱為投射矩陣 (Projection matrix), 將資料 $\mathbf{x}$ 從原來的空間投射到另一個空間, 投射的方式與投射到的空間大小決定了矩陣 A 的組成。資料經過投射或轉置之後, 並不會損失或增加原有的「資訊」(線性的轉換不會使資料憑空增加或減少), 只是會改變資料在空間中的「長相」, 藉此提供額外的資訊, 供進一步資料處理的參考。

主成分分析的理論基礎可以從幾個面象來觀察; 分別陳述如下: (為方便分析及符號的簡潔, 原始變數均假設均數為零, 即 $E(x_i) = 0, \forall i$ 。)

1.3 從 Uncorrelated Variables 的角度

假設新變數 $z_1, z_2, \cdots, z_p$ 間彼此「不相關」(uncorrelated), 則其共變異矩陣為對角化矩陣, 即

$$\Sigma_Z = E(\mathbf{z}\mathbf{z}^T) = AE(\mathbf{x}\mathbf{x}^T)A^T = A\Sigma_X A^T = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix} \tag{2}$$

因已假設 $E(\mathbf{x}) = \mathbf{0}$, 共變異矩陣與相關矩陣相同。下面這個定理讓上式得到一個幾何上的意義:

定理 1. *A symmetric matrix Σ_X can be diagonalized by an orthogonal matrix containing normalized eigenvectors of Σ_X , and the resulting diagonal matrix contains eigenvalues of Σ_X。*

假設對稱矩陣 Σ_X 的特徵值 (eigenvalues) 及特徵向量 (eigenvectors) 分別爲 $\lambda_1 > \lambda_2 > \cdots > \lambda_p$(依大小), $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$, 根據上述定理, 新變數的共變異矩陣 (2) 可以改寫爲

$$\Sigma_Z = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}, \text{並且} A^T = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_p \end{bmatrix}$$

從 $\mathbf{z} = A\mathbf{x}$, 新變數可以寫成

$$\begin{aligned} z_1 &= \mathbf{v}_1(1)x_1 + \mathbf{v}_1(2)x_2 + \cdots + \mathbf{v}_1(p)x_p = \mathbf{v}_1^T\mathbf{x} \\ z_2 &= \mathbf{v}_2(1)x_1 + \mathbf{v}_2(2)x_2 + \cdots + \mathbf{v}_2(p)x_p = \mathbf{v}_2^T\mathbf{x} \\ \vdots &= \vdots \\ z_p &= \mathbf{v}_p(1)x_1 + \mathbf{v}_p(2)x_2 + \cdots + \mathbf{v}_p(p)x_p = \mathbf{v}_p^T\mathbf{x} \end{aligned} \tag{3}$$

其變異數分別爲 $\lambda_1, \lambda_2, \cdots, \lambda_p$。式 (2) 也可以改寫爲

$$\Sigma_X = A^T\Sigma_Z A = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^T \tag{4}$$

又稱爲原始變數共變異矩陣的頻普解構 (Spectral decomposition)。矩陣 $\mathbf{v}_k\mathbf{v}_k^T$(Rank=1) 代表組成 Σ_X 的第 k 個「元素」, 其相對的特徵值 (variance) λ_k 則表示該「元素」所貢獻的比例。當 λ_k 相對太小時, 甚至可以捨棄該「元素」, 僅以「主要成分」(λ_k 相對大的) 來近似原來的矩陣。譬如前面 $q(q < p)$ 個特徵值相對大於其餘的, 可以下列矩陣近似 Σ_X

$$\Sigma_X \approx \sum_{k=1}^{q} \lambda_k \mathbf{v}_k \mathbf{v}_k^T \tag{5}$$

1.4 從最大變異量的角度

原變數的線性組合中, 哪一種組合其變異數最大? 假設新變數爲

$$z = u_1 x_1 + u_2 x_2 + \cdots + u_p x_p = \mathbf{u}^T \mathbf{x} \tag{6}$$

問題變爲選擇一組組合係數, 讓新變數 z 的變異數最大, 即

$$\max_{\mathbf{u}} E(z^2) \equiv \max_{\mathbf{u}} \mathbf{u}^T \Sigma_X \mathbf{u} \tag{7}$$

組合係數 $\mathbf{u}$ 必須有所限制, 否則任意放大將使最大值趨近無限大而失去意義。一般假設 $\mathbf{u}^T \mathbf{u} = 1$, 問題變爲限制式最佳化問題

$$\max_{\mathbf{u}, \mathbf{u}^T \mathbf{u} = 1} \mathbf{u}^T \Sigma_X \mathbf{u} \tag{8}$$

利用 Lagrangian multiplier 的方式去除限制式, 上述問題進一步成爲

$$\max_{\mathbf{u}} \mathbf{u}^T \Sigma_X \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1) \tag{9}$$

其最佳解如下:

$$\Sigma_X \mathbf{u}^o = \lambda \mathbf{u}^o$$

這恰是原始變數的共變異矩陣的特徵結構 (eigen-structure)。此時, 新變數的變異數爲

$$var(z) = E(z^2) = \mathbf{u}^T \Sigma_X \mathbf{u} = \lambda \mathbf{u}^T \mathbf{u} = \lambda$$

換句話說, 當 λ 等於 Σ_X 最大的特徵值 時, 其相對的特徵向量 $\mathbf{v}_1$ 便是最佳的組合係數。此時的新變數稱爲第一個主成分,

$$z_1 = \mathbf{v}_1(1) x_1 + \mathbf{v}_1(2) x_2 + \cdots + \mathbf{v}_1(p) x_p \tag{10}$$

第二個主成分 $z_2 = \mathbf{v}^T \mathbf{x}$ 的推演類似上面的過程, 但多一個條件: 與第一個主成分不相關, 即

$$E(z_1 z_2) = E(z_1)E(z_2)$$

這個條件進一步爲

$$\mathbf{v}^T \Sigma_X \mathbf{v}_1 = 0 \quad \text{或是} \quad \mathbf{v}^T \mathbf{v}_1 = 0 \tag{11}$$

同樣利用 Lagrangian multiplier 的方式 (此時有兩個限制條件), 找到最佳的組合係數 $\mathbf{v}$, 求最大的變異數 $var(z_2)$ 。求解過程留待讀者親自演算, 其解爲:

$$z_2 = \mathbf{v}_2^T \mathbf{x} \tag{12}$$

其中 $\mathbf{v}_2$爲 Σ_X 第二大的特徵值相對的特徵向量。其餘的成分依此方式便可逐一呈現。以下的練習有助於瞭解主成分分析的原理及意義。

2 練習

範例1:MATLAB 提供了一組美國城市生活品質的調查資料:cities.mat。把對 329 個城市的 9 項評比資料拿出來觀察, 你可以從裡面看到什麼訊息? 如何去觀察這麼多 (9x329) 的數字資料? 要畫什麼樣的圖? 計算哪些統計量呢?

MATLAB的線上手册常有些很不錯的範例, 不僅提供資料, 也對指令的應用有詳細且完整的說明。這組美國城市生活品質的調查資料及對於主成分分析的指令 princomp, 可以從「Help Browser」裡以「princomp」爲關鍵字搜尋到。並依循範例的說明, 一步步執行相關的指令, 對於主成分分析的功能的瞭解很有幫助。藉著這個範例, 不妨可以利用本單元說明的主成分分析原理, 實際寫程式去計算所有的結果, 並與 princomp 指令執行的結果做比對, 相信對主成分分析的原理與精神更能掌握。這比起單純的執行 princomp 還要有感覺, 瞭解更深刻。

資料中的 ratings 是 329 個城市的 9 項評比資料, 針對大量資料的第一印象或是初期的瞭解可以畫 boxplot, 在 MATLAB的範例說明裡也提到並示範以下的指令

```
load cities;
boxplot(ratings,'orientation','horizontal','labels',categories)
```

結果如圖 1 所示。指令最後的變數 category 含評比項目的文字。

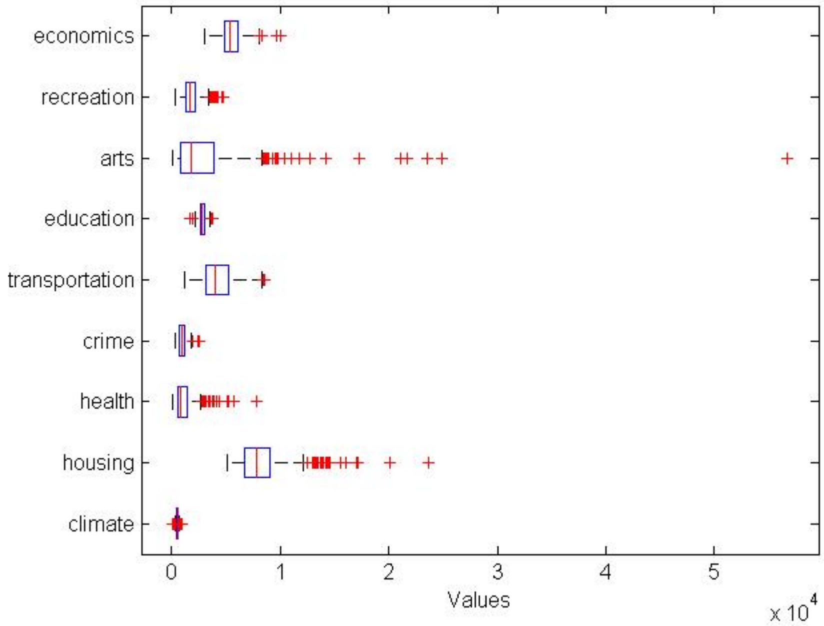

圖 1: 城市生活品質調查資料的 Box Plot

這個圖對於資料分析很重要, 可以看出資料間的差異性, 譬如大小 (scale) 的差距及資料的散佈情況, 這些都對於判斷資料是否需要做前置處理 (pre-processing) 很有幫助。從這組城市評比的資料來看, 不同項目的大小與變異相差頗大, 這對做主成分分析可能不利, 因此有必要先將這些差距以標準化的方式拉近些。譬如 MATLAB 範例建議的將每筆資料除以各項目的標準差, 指令如下

```
stdr = std(ratings);
sr = ratings./repmat(stdr,329,1);
```

變數 sr 代表標準化過的資料, 有興趣者不妨畫出它的 Box Plot 看看與之前未標準化前的差異。這裡用了一個指令 repmat(意爲: repeat matrix), 非常好用, 它將一個矩陣或向量當作一個單位, 複製成更大的矩陣。譬如上面的指令, 將 1×9 的向量 (當

作一塊磁磚) stdr 複製 (貼在) 成 329×1 個向量 (329×1 塊磁磚), 所以實際的大小是 329×9 的矩陣。如果還是不明白, 可以設定一個小矩陣來試試 repmat 的複製功能。

範例2:9項評比 (9個變數) 資料是否彼此相關? 彼此間的相關性有何差別? 如果要以畫散佈圖的方式來觀察所有變數間彼此的關係, 總共要畫幾張圖? 譬如, 畫 Health vs. Arts 及 Climate vs. Education 的散佈圖來看看他們之間的關係。

MATLAB提供指令 corrcoef 計算變數資料間的相關係數, 譬如

```
R = corrcoef(sr)
```

這是個對稱矩陣, 有了這組數據, 是否有畫散佈圖的必要? 或許見仁見智, 不過基於程式寫練習, 倒是可以一試。圖2只展示三個項目間的散佈圖, 可與相關係數做一比較。其程式片段如下:

```
k=3; %選擇多少個項目
for  i=1:k
    for j=i:k
        subplot(k,k,(i-1)*k+j),plot(sr(:,i),sr(:,j),'o')
        xlabel(categories(i,:))%從變數 categories 找出項目名稱
        ylabel(categories(j,:))
        pause(1) %停留1秒, 方便觀察
    end
end
```

範例3: 使用MATLAB 的指令 cov 及自行以公式分別計算共變異矩陣。觀察這些共變異矩陣的特徵值及特徵向量。

```
Ex = cov(sr)
lambda=eig(Ex)
```

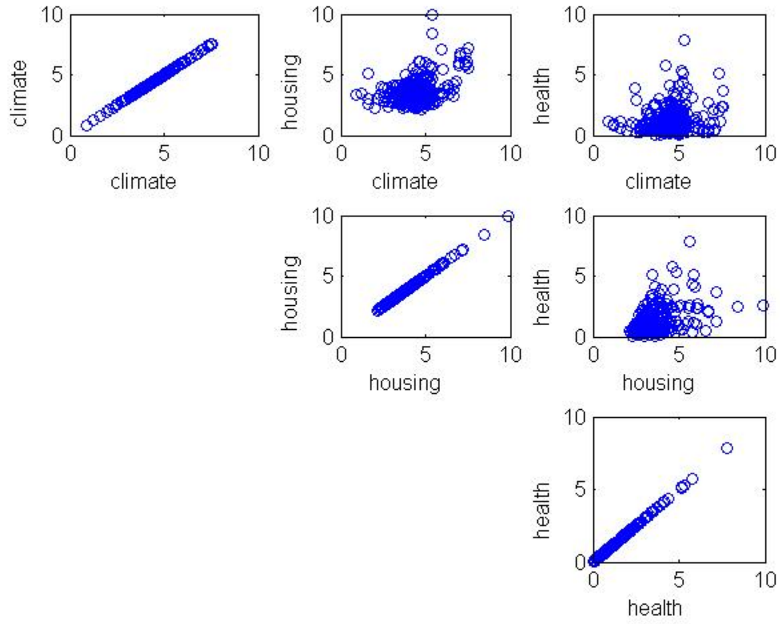

圖 2: 城市生活品質調查資料: 不同項目資料的散佈圖

式 (4) 的 Σ_X 是假設原始變數均數皆爲 0的相關矩陣, 一般情況則是使用共變異矩陣。經過指令 eig 計算得到的特徵值, 其排列並非由大而小, 使用前必須再經過排序。

範例4: 假設五個變數 x_1, x_2, x_3, x_4, x_5, 其中 x_1, x_2, x_3 爲線性獨立,$x_4 = x_1 + x_2$, $x_5 = x_2 + x_3$, 由這 5 個變數構成的共變異矩陣有幾個值爲 0 的特徵值呢? 試著去模擬這個問題。從樣本共變異矩陣中看看 5 個變數的樣本變異數與特徵值的關係。

---

x_1, x_2, x_3的樣本值可以從亂數產生器 (譬如假設爲標準常態) 產生,x_4, x_5 再從這三組資料相加取得。

範例5: 主成分分析的在幾何上的概念是「Change of basis」, 也就是座標軸的改變 (旋轉與位移)。當座標軸改變時, 原來空間中的所有點的座標也要跟著改。透過這個改變, 把資料中的主要成分抽離出來。整個過程可以透過一些簡單的二維座標轉換來展示。請按下列步驟執行: 結果如圖 3 所示

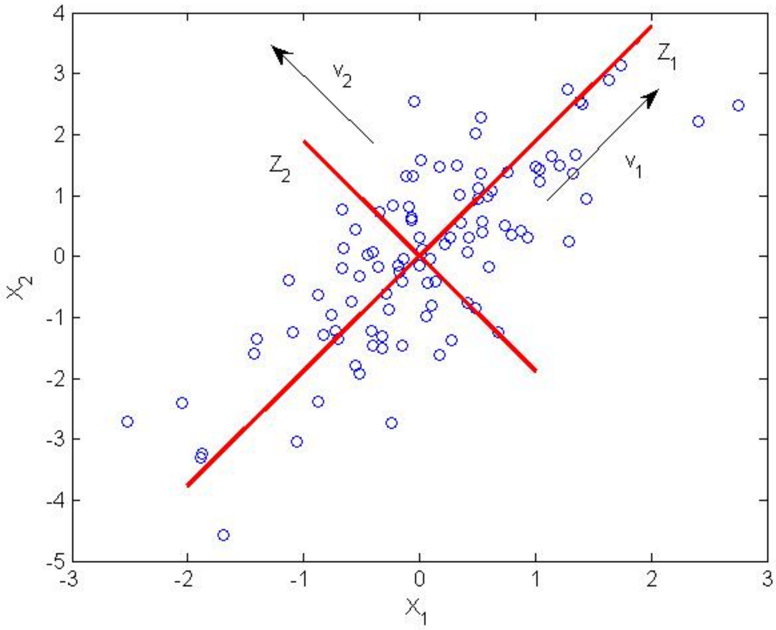

圖 3: 主成分分析的幾何意義: 座標轉換

1. 產生兩組具相依性的模擬資料, 畫出散佈圖。下列兩變數 x_1, x_2 的關係式是一個方式, 其中 c 用來調節相關性。

$$x_2 = cx_1 + \epsilon, \;\; c \in R, \;\; x_1, \epsilon \in N(\mu, \sigma^2)$$

```
x1=normrnd(0,1,100,1);
x2=1.5*x1+normrnd(0,1,100,1);
plot(x1,x2,'o')
```

2. 建立兩變數的共變異矩陣 $\Sigma_X = cov(x_1, x_2)$, 並計算其特徵值與特徵向量。

```
Ex = cov([x1 x2]);
[V, D] = eig(Ex);
[lambda, I] = sort(diag(D),'descend');%依大小排列
V = V(:, I) %特徵向量依特徵值大小重新排列
```

3. 第一個特徵向量指向新的座標軸 (以 Z_1 代表), 第二個特徵向量則指向與之垂直的另一個座標軸 (以 Z_2 表示)。畫出這兩條軸線。以下程式片段畫出第一條線。讀者試著自己畫出第二條垂直的線。

$x = [-2\ 2];$
$y = V(2,1)/V(1,1) * x;$
$plot(x, y,' LineWidth', 2,' color',' r')$

4. 建立矩陣 $A = [\mathbf{v}_1\ \mathbf{v}_2]^T$, 其中 $\mathbf{v}_1, \mathbf{v}_2$ 為共變異矩陣 Σ_X 的兩個特徵向量。計算 $\mathbf{z} = A\mathbf{x}$, 即原資料經座標轉換後的新座標。新變數 z_1 與 z_2 的關係如圖 4。請注意這張圖與圖 3 的關係, 圖 4 是將圖 3 的部分轉正來看。

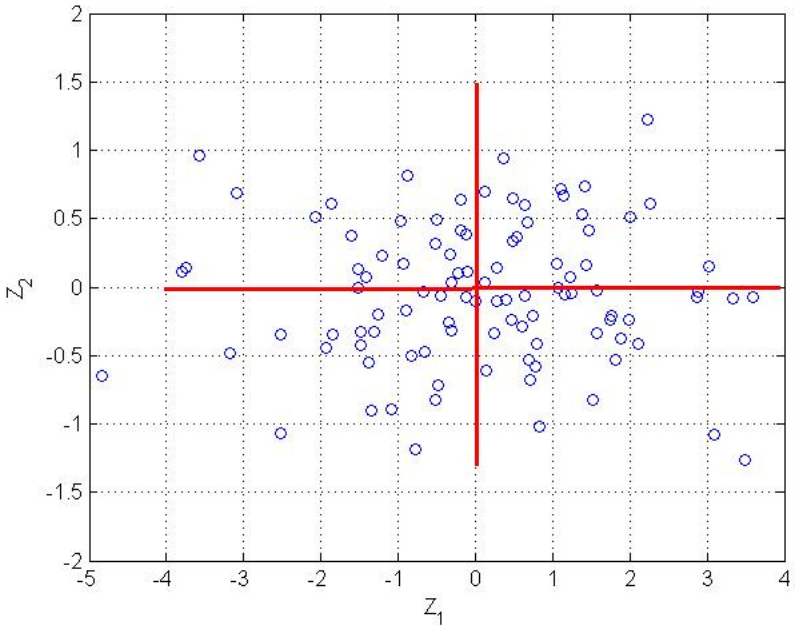

圖 4: 主成分分析的幾何意義: 座標轉換後的新座標

範例6: 主成分分析是將原變數作線性組合, 成為另一組變數, 組合的原則是保留原變

數間最大的變異, 且新變數彼此不相關。這個練習想去瞭解不同的組合的變異量與幾何意義。

假定 x_1, x_2 兩個變數, 樣本資料爲 $x_1 = [1\ 2\ 3\ 4\ 5]$, $x_2 = [2\ 1\ 4\ 5\ 4]$, 如果想要用一個新的變數 z_1 來代表這兩個變數, 在希望保留原變數最大變異 (variance) 的前提下, 下列哪一個組合最理想:

1. $z_1 = x_1$
2. $z_1 = \frac{1}{\sqrt{5}}x_1 + \frac{2}{\sqrt{5}}x_2$
3. $z_1 = \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2$
4. $z_1 = x_2$

問題:

- z_1 的樣本值來自 x_1, x_2 資料的轉換, 這相當前面練習所說的座標軸轉換, 而且都只代表轉換過後的一個座標軸。請根據上述的組合, 分別畫出這個座標軸 (含 x_1, x_2 的散佈圖) 如圖 5 所示。
- 分別計算新變數 z_1 的變異數。哪一個最大?
- 分別計算新的座標值與新座標軸垂直距離的平方和。哪一個最小?

主成分的來源是以保留原變數間最大的變異爲原則, 即式 (8) 所示。這個原則的另一面是

$$\min_P \sum_{k=1}^{n} ||(I-P)\mathbf{x}_k||^2 = \max_P \sum_{k=1}^{n} \mathbf{x}_k^T P \mathbf{x}_k \tag{13}$$

其中 P 即是所謂的 Orthogonal projection matrix。上式以樣本值爲依據, 若以變數型態則可寫成,

$$\min_P E(||(I-P)\mathbf{x}||^2) = \max_P E(\mathbf{x}^T P \mathbf{x}) \tag{14}$$

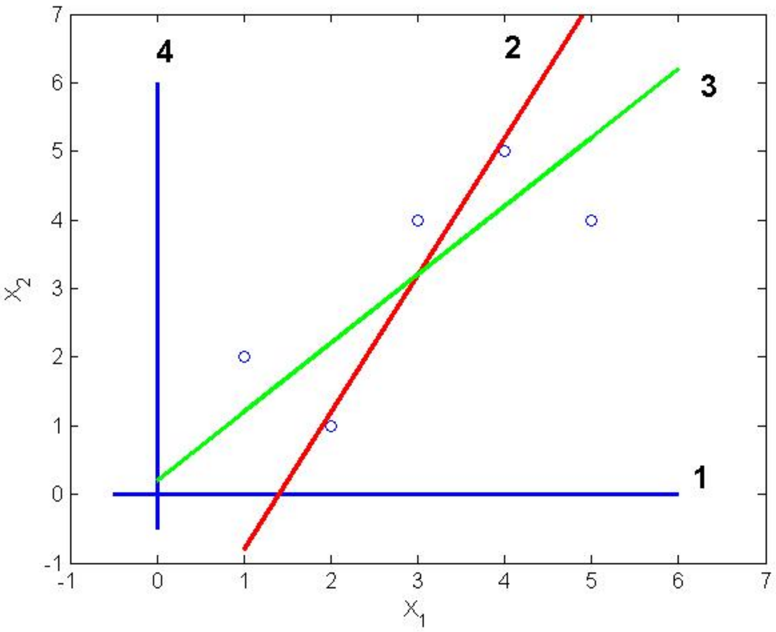

圖 5: 不同組合的幾何意義與變異量。

式 (13)(14) 是一種觀念式的表示法, 其中的 Orthogonal projection matrix P 一般並不容易直接計算取得, 通常經由別的概念切入, 推演出如 (13) 或 (14) 的表示法。譬如 $P = WW^T$。

範例7: 利用son.txt 這組資料做主成分分析:

1. 共變異矩陣 (Covariance Matrix) 是觀察兩個變數之間關係較常用的統計量。計算『頭部長度』與『頭部寬度』的樣本共變異矩陣 S(sample Covariance Matrix)。

2. 繪製兩者的散佈圖, 圖形顯示的是否與共變異矩陣呼應? 如何觀察?

3. 計算樣本共變異矩陣 S 的特徵值 及相對的特徵向量。觀察特徵值的大小分佈, 是否與兩變數間的相關程度有關? 觀察特徵向量 $\mathbf{v}_1, \mathbf{v}_2$ 的關係, 是否存在 orthogonal 的關係? 即 $\mathbf{v}_1^T \mathbf{v}_2 = 0$?

4. 假設樣本共變異矩陣 S 的特徵值爲 λ_1, λ_2, 相對的特徵向量爲 $\mathbf{v}_1, \mathbf{v}_2$。驗證

$S = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_1 \mathbf{v}_2 \mathbf{v}_2^T$

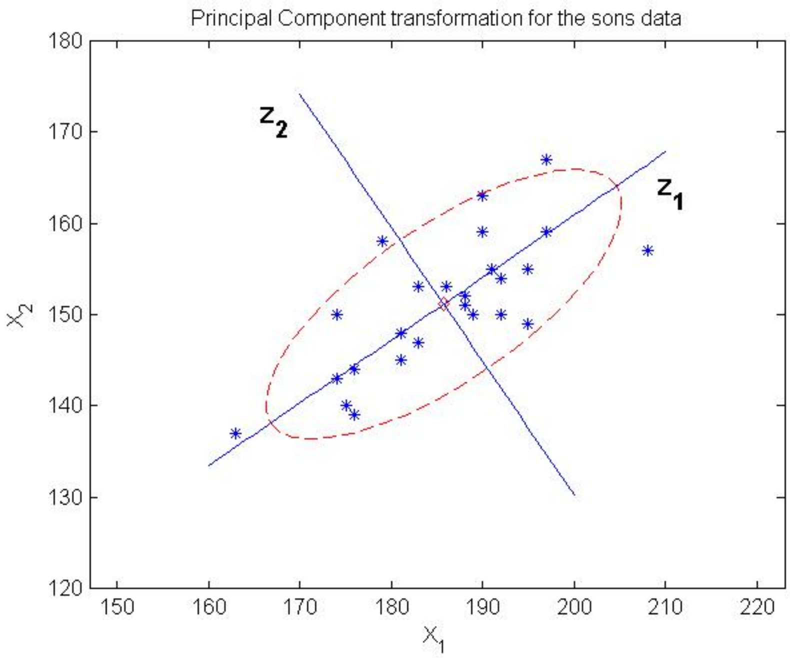

圖 6: 實際資料的主成分分析。

5. 當將資料的座標軸從 (X_1, X_2) 轉爲 (Z_1, Z_2) 時, 原資料的座標值將隨之改變。畫出如圖6的兩條垂直線。

    - 先計算中心點 (想想看這個中心點如何決定?)
    - z_1, z_2 軸就是 $\mathbf{v}_1, \mathbf{v}_2$ 的方向, 透過向量與中心點便可以畫出如圖中的新座標軸 Z_1, Z_2。

6. 從座標軸 (Z_1, Z_2) 來看這些資料, 似乎顯示出『比較散亂』的不相干關係。不過又扁向 Z_1 軸。這個『扁』的傾向或程度, 可以畫一個橢圓來表示。這牽涉到畫一個圓及橢圓的技巧。

    - 畫圓: 採 (1) $x^2 + y^2 = d, -d \le x, y \le d$ 或 (2) $sin^2\theta + cos^2\theta = d$,d 爲半徑,$-\pi \le \theta \le \pi$ 。
    - 畫橢圓: 採 (1) $rx^2 + sy^2 = d, -d \le x, y \le d$ 或 (2) $rsin^2\theta + scos^2\theta = d$,d 爲半徑,$-\pi \le \theta \le \pi$ 。

- 當圓心不在 $(0,0)$ 時, 如何畫?Hint: $(x-x_1)^2+(y-y_1)^2=d$ 或 $r(x-x_1)^2+s(y-y_1)^2=d$

7. 以新的座標軸 (Z_1, Z_2) 來看這些資料, 新的座標值如何計算? 是不是可以找到一個轉換機制 (矩陣)? 複習線性代數有關座標軸轉換的部分。

MATLAB當然也提供了關於主成分分析的指令:princomp, 習者不妨到 help 中的統計工具箱查詢並參考它的使用方式。另外還有指令 pcacov,pcares 等相關指令。以上的範例都建議讀者使用指令 priccomp 再執行一次, 對照之前的指令與結果, 相信對主成分分析會又更好的瞭解, 對指令 princomp 也會更透徹。未來需要作主成分分析時, 便可以直接採用 MATLAB 的指令。

3 觀察

1. 做主成分分析時, 由於資料來自代表不同意義的變數, 量測到的資料常有大小或變異 (variance) 差異甚大的情況, 這樣的資料有做標準化 (standardization) 的必要。在 MATLAB 的指令中, 利用 zscore 來做標準化, 譬如 $N \times p$ 的資料矩陣 X, 其中 N 代表資料樣本數,p 代表變數個數,zscore(X) 將每個變數的樣本值標準化。

2. 複迴歸分析所牽涉變數間的多重共線性, 也可以運用主成分分析的方式來解決。

3. 轉換座標軸後的第一個軸 ($\mathbf{v}_1$), 像不像一條迴歸線?

4. 主成分分析通常做爲其他資料處理方式的前置作業, 能幫助去除多餘的資料、將變數量壓低。不過並非所有的應用都適合做這樣的處理, 有些時候反而將有用的資料覆蓋或打亂 (譬如, 群組分析), 未獲其利, 先蒙其害。應用時機的選擇非常重要, 需要經驗與審慎的態度。

4 作業

1. FOOTBALL.txt 這組資料提供作爲安全帽設計與頸部傷害的研究。研究的對象是美國大學 football 與非 football 球員共60名, 並量測6種頭部相關的資

料。選擇這6種頭部相關資料是否能反映出設計的關鍵, 並不是本主題的興趣。本主題想探討這些資料彼此間是否有相關性? 也就是說: 或許更少的資料就能表達出這6種資料所能表達的意涵! 如果是這樣, 對於應用上的幫助不小, 因爲那代表需要花費的人力成本降低 (要量測的項目變少), 在分析上也比較容易 (變數少了), 結果也會比較『穩定』(獨立性強了)。這個練習要探討幾個理論與程式設計的技巧:

- 先簡單的觀察一下這6組資料的相關性, 以得到一個初淺的變數間相關的程度。建議畫出每組資料的散佈圖。
- 計算並觀察原始資料的共變異矩陣 (Covariance Matrix)。從這個關聯性值的矩陣, 能否看出初步的相依性, 或變數個別的重要性!?
- 執行主成分分析, 觀察其特徵值的分布, 並且求其比重的分布。可以畫所謂的 scree plot, 即依特徵值大小做圖。或畫特徵值 的 Pareto(柏拉圖) plot。
- 取前兩個主成分組成新的變數 z_1, z_2, 即

$$\begin{aligned} z_1 &= \mathbf{v}_1(1)x_1 + \mathbf{v}_1(2)x_2 + \cdots + \mathbf{v}_1(6)x_6 \\ z_2 &= \mathbf{v}_2(1)x_1 + \mathbf{v}_2(2)x_2 + \cdots + \mathbf{v}_2(6)x_6 \end{aligned}$$

  從由特徵向量組成的係數來看, 觀察哪些變數 x_i 的重要性比較高? 是不是可以據此說明只要這些變數即可表達所有的意義?

- 畫一張 z_1, z_2 的散佈圖, 觀察他們的相關性及分佈的情況 (像常態嗎?)。另外值得觀察的是這些資料在 Z_1 軸及 Z_2 軸的變異性 (variance), 及是否存在群聚性 (grouping)。這個問題可以自行寫程式計算, 也可以直接採用指令 princomp。

2. 證明式 (9) 與 (10) 是相同的問題。

3. 證明式 (12)。

參考文獻

[1] J. Latin, D. Carroll, P. E. Green, "Analyzing Multivariate Data,"2003, Duxbruy.

[2] A. C. Rencher, "Multivariate Statistical Inference and Applications,"1998, John Wily and Sons.