Practice#1：Canonical Correlation

目的：

Canonical Correlation Analysis 在探討複迴歸分析中有多個因變數的情形。如果去考慮每個因變數與自變數的關係，問題將變的非常麻煩。但是，如果可以將多個因變數組合成一個新的變數，問題將單純許多。Canonical Correlation Analysis 便是在探討什麼樣的組合才是最恰當的。這樣的關係推展到類別分析時，對於變數太多的情況可以得到比較簡單的判別方式，雖然犧牲了精準度，對於許多應用而言卻已經足夠了。本單元僅對 Canonical Correlation 做初步的探討，著眼點仍在如何應用 MATLAB 程式的寫作去處理一般常見的統計方法。

複迴歸分析在找出因變數與多個自變數之間的關係，其模式如

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1}$$

$\underline{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \cdots & \beta_p\end{bmatrix}$決定了自變數與因變數間的關係。一般對$\underline{\beta}$估計採最小平方法，也就是求一組最佳值$\underline{\beta}^o$，使得上述模式與因變數樣本資料得到最佳的配適。這裡的「最佳」指的是

$$\min_{\underline{\beta}} \sum_{k=1}^{N} (y_k - \beta_0 - \beta_1 x_{1k} - \beta_2 x_{2k} - \cdots - \beta_p x_{pk})^2 \tag{2}$$

最佳值$\underline{\beta}^o$滿足正規方程式(Normal Equation)

$$(X'X)\underline{\beta}^o = X'\underline{y} \tag{3}$$

其中資料矩陣 X 與 y 的結構請參考前面關於迴歸分析的單元。

另一種「最佳」的表示法：在所有自變數的線性組合中，找出一組使得其線性組合後的變數與因變數的相關係數最大。即

$$\max_{\underline{b}} \frac{\operatorname{cov}(Y, X\underline{b})}{\sqrt{\operatorname{var}(Y)\operatorname{var}(X\underline{b})}} \tag{4}$$

其中 $X=\begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$，$\underline{b}=\begin{bmatrix} b_1 & b_2 & \cdots & b_p \end{bmatrix}^T$

為方便進一步的推導，假設變數 X_i 與 Y 都已經標準化，即 $var(Y)=var(X_i)=1$。此外，也假設組合係數 $\underline{b}$ 的選擇使得 $var(X\underline{b})=1$。式(4)變為

$$\max_{\underline{b}, s.t.\, \mathrm{var}(X\underline{b})=1} \mathrm{cov}(Y, X\underline{b}) \tag{5}$$

或

$$\max_{\underline{b}, s.t.\, \frac{1}{N-1}\underline{b}^T X^T X\underline{b}=1} \frac{1}{N-1}\underline{y}^T X\underline{b} \tag{6}$$

其中 $\underline{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, X=\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{pN} \end{bmatrix}$

式(6)經 Lagrangian method 求得最佳解滿足下列條件（作業 1）

$$X'X\underline{b}^o = 2\lambda X'\underline{y} \tag{7}$$

其中 λ 即所謂的 Lagrangian multiplier。最佳的組合係數與最小平方法的解幾乎相同。這也說明最小平方法的精神就是求最大相關係數。

「找出自變數最佳的組合，使得該組合與因變數之間的相關係數最大，」可以推展到當因變數不只一個時。即「找出最佳的自變數組合與因變數的組合，使得兩個組合變數之間的相關係數最大。」假設 $\underline{u}=X\underline{b}$, $\underline{t}=Y\underline{a}$ 分別代表自變數與因變數的線性組合，也稱為 canonical variates。問題變成

$$\max_{\underline{a},\underline{b}, s.t.\, \frac{1}{N-1}\underline{a}^T Y^T Y\underline{a}=1, \frac{1}{N-1}\underline{b}^T X^T X\underline{b}=1} \frac{1}{N-1}\underline{a}^T Y^T X\underline{b} \tag{8}$$

其中目標函數為 $\underline{u}$ 及 $\underline{t}$ 間的相關係數 $r(\underline{u},\underline{t})$，又稱為 canonical correlation。(8) 的最佳解分別滿足

$$\begin{aligned} (R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY})\underline{a} &= r^2(\underline{u},\underline{t})\underline{a} \\ (R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX})\underline{b} &= r^2(\underline{u},\underline{t})\underline{b} \end{aligned} \tag{9}$$

其中

$$R_{YX}=\frac{1}{N-1}Y^TX$$
$$R_{XX}=\frac{1}{N-1}X^TX$$
$$R_{YY}=\frac{1}{N-1}Y^TY$$

式(9)是一個 eigenvector-eigenvalue 的問題。最佳解$\underline{a}$與 $\underline{b}$ 分別為 $R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$ 及 $R_{XX}^{-1}R_{XY}R_{YY}^{-1}R_{YX}$ 的 eigenvector 中 eigenvalue 最大的那一個。或是以下列的關係從一個最佳解 $\underline{b}$ 計算另一個最佳解 $\underline{a}$

$$\underline{a}=\frac{1}{r(\underline{u},\underline{t})}R_{YY}^{-1}R_{YX}\,\underline{b} \tag{10}$$

練習：

1. 從共變異矩陣(Covariance Matrix)與相關矩陣(Correlation Matrix)說起；

i、 兩變數 *x,y* 間的共變異數（covariance）及樣本共變異數（sample covariance）分別定義為：(*x(k), y(k)*代表變數 *x* 與 *y* 的第 *k* 個樣本值)

$$\mathrm{cov}(x,y)=\sigma_{xy}=\mathrm{E}[(x-\mu_x)(y-\mu_y)] \tag{11}$$

$$s_{xy}=\frac{1}{N-1}\sum_{k=1}^{N}(x(k)-\overline{x})(y(k)-\overline{y}) \tag{12}$$

ii、 當變數超過兩個（假設為 *p* 個），變數與變數間的共變異數可以透過共變異矩陣來觀察，其定義及數值計算的方式如

$$\Sigma=\mathrm{cov}(X)=\mathrm{E}[(X-\underline{\mu}_X)(X-\underline{\mu}_X)^T]=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1p}\\ \sigma_{21}&\sigma_{22}&\cdots&\sigma_{2p}\\ \vdots&\vdots&\ddots&\vdots\\ \sigma_{p1}&\sigma_{p2}&\cdots&\sigma_{pp}\end{bmatrix}$$
$$\text{其中}X=\begin{bmatrix}x_1\\x_2\\ \vdots\\x_p\end{bmatrix},\qquad \underline{\mu}_X=\mathrm{E}[X]=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{bmatrix} \tag{13}$$

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{N-1}\sum_{k=1}^{N}(x_1(k)-\mu_1)(x_1(k)-\mu_1) & \frac{1}{N-1}\sum_{k=1}^{N}(x_1(k)-\mu_1)(x_2(k)-\mu_2) & \cdots \\ \frac{1}{N-1}\sum_{k=1}^{N}(x_2(k)-\mu_2)(x_1(k)-\mu_1) & \frac{1}{N-1}\sum_{k=1}^{N}(x_2(k)-\mu_2)(x_2(k)-\mu_2) & \cdots \\ \vdots & \vdots & \ddots \\ \frac{1}{N-1}\sum_{k=1}^{N}(x_p(k)-\mu_p)(x_1(k)-\mu_1) & \frac{1}{N-1}\sum_{k=1}^{N}(x_p(k)-\mu_p)(x_2(k)-\mu_2) & \cdots \end{bmatrix} =$$

$$\frac{1}{N-1}X_c^{\ T}X_c \quad where \ \ X_c = \begin{bmatrix} x_1(1)-\mu_1 & x_2(1)-\mu_2 & \cdots & x_p(1)-\mu_p \\ x_1(2)-\mu_1 & x_2(2)-\mu_2 & \cdots & x_p(2)-\mu_p \\ \vdots & \vdots & \ddots & \vdots \\ x_1(N)-\mu_1 & x_2(N)-\mu_2 & \cdots & x_p(N)-\mu_p \end{bmatrix} \tag{14}$$

iii、 兩變數 x,y 間的相關係數(correlation coefficient)及其數值計算(sample correlation coefficient)定義為

$$\rho_{xy} = corr(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{15}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{16}$$

iv、 當變數超過兩個（假設為 p 個），變數與變數間的相關係數可以透過相關矩陣來觀察，其數值計算的方式如

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \vdots & 1 \end{bmatrix} = \frac{1}{N-1}Z^T Z \tag{17}$$

其中 $Z=\begin{bmatrix} \frac{x_1(1)-\mu_1}{\sigma_1} & \frac{x_2(1)-\mu_2}{\sigma_2} & \cdots & \frac{x_p(1)-\mu_p}{\sigma_p} \\ \frac{x_1(2)-\mu_1}{\sigma_1} & \frac{x_2(2)-\mu_2}{\sigma_2} & \cdots & \frac{x_p(2)-\mu_p}{\sigma_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_1(N)-\mu_1}{\sigma_1} & \frac{x_2(N)-\mu_2}{\sigma_2} & \cdots & \frac{x_p(N)-\mu_p}{\sigma_p} \end{bmatrix}$

v、 Matlab 亦提供共變異矩陣與相關矩陣的功能函數，分別為 *cov* 及 *corrcoef*。雖然 MATLAB 提供了這些功能，不過使用者最好能根據定義實際去計算一次，以確定 MATLAB 的函數與自己預期的一致。請自網路下載資料：canonical_data2.txt，分別利用(13)及(16)計算其 sample covariance matrix 及 sample correlation matrix。並*驗證* MATLAB 提供的指令 *cov* 及 *corrcoef*。

Hint:式(13)可以一個指令完成 *Xc=X-(ones(length(X(:,1)),1)*mean(X));*

vi、 其實當資料經過標準化後，共變異矩陣等於相關矩陣。標準化一個資料矩陣可以使用 zscore 指令。

2. 本單元探討了變數間的相關性，我們可以從已知的資料中計算彼此間的相關係數，一般稱為相關矩陣(correlation matrix)。請從網路下載一組資料檔（資料來源[1]）canoncal_data1.txt，該檔案資料包括兩個輸入變數及兩個輸出變數的資料。計算這四個變數的相關矩陣：

i、 檢查這四組資料是否為標準化資料？

ii、 觀察變數間的相關係數，即計算相關矩陣。

iii、 所謂 canonical variates 是輸入變數的組合 $u=b_1x_2+b_2x_2$ 與輸出變數的組合 $t=a_1y_1+a_2y_2$，係數的選擇使得 u 與 t 間的相關係數最大。在尚未計算最佳組合前，先就一些組合來觀察其相關性；譬如

$u_1=\frac{100}{100}x_1+\frac{0}{100}x_2$ $u_2=\frac{90}{100}x_1+\frac{10}{100}x_2$ $\vdots$ $u_{11}=\frac{0}{100}x_1+\frac{100}{100}x_2$	$t_1=\frac{100}{100}y_1+\frac{0}{100}y_2$ $t_2=\frac{90}{100}y_1+\frac{10}{100}y_2$ $\vdots$ $t_{11}=\frac{0}{100}y_1+\frac{100}{100}y_2$

建立一關係矩陣來觀察輸入變數的 11 種組合與輸出變數的 11 種組

合間的相關係數。

iv、 利用式(9) 計算最佳的組合係數及 canonical correlations.

觀察：

1. 計算式(9) 的 eigenvector 時，請留意限制條件 $\underline{a}^T R_{YY}\underline{a}=1$, $\underline{b}^T R_{XX}\underline{b}=1$ 必須被滿足。因此適當的調整 eigenvector 的 scale 是有必要的。
2. Canonical Correlation Analysis 及 principal component 在觀念上非常近似，但實際上確實不同的東西，請仔細比對這兩個分析方法的異同，才不會在應用時搞錯方向。

作業：

1. 證明最佳化問題(6)的解為(7)。
2. 證明最佳化問題(8)的解為(9)。
3. 推導(10)。

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 & Sons.
3. 黃俊英，”多變量分析<第七版>，”中國經濟企業研究所。