交叉驗證法在淨最小平方法中選取最佳因子數目之模擬研究[1]

廖宜倫[2]、郭寶錚[3]

摘　要

利用淨最小平方法來解決資料中共線性問題是有用的，而淨最小平方法是由主成分分析改進而來，除了將解釋變數 $\boldsymbol{X}$ 矩陣轉換成直交得點矩陣 $\boldsymbol{t}$ 外，並在得點矩陣中加入反應變數 $\boldsymbol{y}$ 的訊息。淨最小平方法中，$\boldsymbol{X}$ 矩陣可分解成得點與負荷量矩陣 $\boldsymbol{t}$ 及 $\boldsymbol{p}'$，$\boldsymbol{y}$ 矩陣可分解成得點與負荷量矩陣 $\boldsymbol{u}$ 及 $\boldsymbol{q}'$，其間並經由權重向量 $\boldsymbol{w}$ 牽引著 $\boldsymbol{X}$ 與 $\boldsymbol{y}$ 之間的訊息，因此，淨最小平方法的表現要比主成分回歸法爲佳的原因。本研究利用兩個模擬研究以交叉驗證法來選擇淨最小平方回歸上最佳因子數，第一個模擬資料的產生是利用實際的NIR資料特性爲基礎來模擬產生，也就是新產生的資料中含有原始的實際資料的特性，我們只控制了模式中最佳因子數目、樣本個數、解釋變數個數與樣本變方大小，並檢查這些因素對選取最佳因子數目的影響；第二個模擬則是依照共線性程度的大小來模擬產生資料，在這個模擬中我們控制了資料的共線性大小，當共線性程度太過嚴重時，可以利用交叉驗證法來選擇保留有用訊息的因子數目，捨棄會造成共線性問題的因子。綜言之，使用淨最小平方法可以有效控制資料共線性問題，交叉驗證法對模式中最佳因子數目的估計能獲得不錯的效果，而交叉驗證法在各種情形下對模式中最佳因子數目的估計最爲準確。

關鍵字：近紅外線光譜、淨最小平方法、共線性資料、交叉驗證法、模擬研究。

前　言

在米質研究中測量水稻之直鏈澱粉含量(apparent amylase content)，利用近紅外光分析儀等光譜儀器分析其化學成分，較之在傳統實驗室中以化學分析，快速方便，且因具有非破壞性分析的特性，在近幾十年來有越來越普遍的趨勢(1)。

資料處理在化學分析方面通常分爲兩個步驟，首先先用儀器分析其化學特性，再依此化學特性估計一檢量模式 $\boldsymbol{y}=\alpha\boldsymbol{1}+\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，其中 $\boldsymbol{y}$ 爲n個樣本數的行向量，$\boldsymbol{X}$ 爲解釋變數的n×p

[1] 行政院農業委員會臺中區農業改良場研究報告第 0715 號。
[2] 行政院農業委員會臺中區農業改良場助理研究員。
[3] 國立中興大學農藝系教授。

矩陣，α 爲模式的截距，$\boldsymbol{\beta}$ 爲未知的回歸係數向量，而$\boldsymbol{\varepsilon}$ 爲逢機誤差向量，其期望值爲**0**，變方-共變方矩陣爲 $\sigma^2\boldsymbol{I}$，此步驟稱爲檢量或校正步驟(calibration step)，模式的參數稱爲回歸係數。第二個步驟爲從樣品中取出新的獨立觀測值，並將其代入步驟一所得之預測模式 $\hat{\boldsymbol{y}} = \hat{\boldsymbol{\alpha}}\boldsymbol{1} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$，以此預測反應變數，此步驟稱爲預測步驟(prediction step)[3]。

而利用近紅外光譜(near infrared spectroscopy, NIRS)分析需要用到多變數檢量(multivariate calibration)方法來處理資料，以X變數(解釋變數)代表各個波長下的光譜吸收值，Y變數(反應變數)代表化學物質的濃度或含量，來建立多重線性迴歸(multiple linear regression, MLR)。然而在解釋變數之間有著高度相關時，這種現象稱之多重共線性。目前已知利用最小平方法來估計此類型資料的回歸係數可能會得到非常差的結果，因此發展出數種解決的方法，其中包括主成分回歸(principal components regression, PCR)、脊回歸(ridge regression)、潛在根回歸(latent root regression, LRR) [9]，這些方法在很多例子中都有其良好的表現，且已有許多的研究在於比較它們之間的優缺點。

本研究中，將利用淨最小平方法(partial least squares regression, PLSR)來建立檢量模式，並利用模擬資料來驗證此檢量模式，並以交叉驗證法取得模式中最佳因子數目，模擬資料一爲利用原始資料-米質研究中以近紅外線光譜分析水稻之直鏈澱粉含量的特性來重新產生，模擬資料二則依自己所設定的共線性條件產生。

淨最小平方回歸

一般資料處理上，主要是藉由解釋變數 $\boldsymbol{X}$ 來預測反應變數 $\boldsymbol{Y}$ 的訊息，所以首先要決定的是預測模式 $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$，而在決定預測模式的方法上，有許多不同的多變數檢量方法，例如多重線性回歸、脊回歸、主成分回歸、淨最小平方法...等。一般而言，當我們面對一組資料時，會針對此資料建立檢量模式，並利用此模式對未知的觀測值做預測。

一般的資料分析中，最常使用到的檢量方法爲多重線性回歸，當解釋變數之間有著高度相關，也就是當資料中有著嚴重共線性程度時，使用多重線性回歸所估計到的回歸係數並不穩定，所以在本文中主要是討論以淨最小平方法來解決共線性問題，並利用交叉驗證法準則在PLS模式中選取最佳因子數，建立最佳的預測模式。

淨最小平方回歸是由PCR進一步發展而來，相對於PCR只利用 $\boldsymbol{X}$ 的訊息來求得負荷量(loading)矩陣 $\hat{\boldsymbol{p}}$，PLS則是同時利用 $\boldsymbol{X}$ 及 $\boldsymbol{y}$ 的訊息來找出 $\hat{\boldsymbol{p}}$，其理由是只選擇有用的 $\boldsymbol{X}$ 訊息來對 $\boldsymbol{y}$ 做精確的預測，即PLS能找出包含 $\boldsymbol{y}$ 預測值訊息的 $\hat{\boldsymbol{p}}$，而這些訊息在 $\boldsymbol{X}$ 上是沒有表現的[4,9]。以下簡單介紹PLS的概念。

第一個PLS負荷量向量 $\hat{\boldsymbol{p}}_1$ 是藉著 $\boldsymbol{x}$ 對 $\boldsymbol{y}$ 做回歸所得之標準化 $\boldsymbol{X}'\boldsymbol{y}$ 共變異矩陣而得，而利用相對於 $\hat{\boldsymbol{p}}_1$ 的得點矩陣 $\boldsymbol{t}_1$ ($\boldsymbol{t}_1 = \boldsymbol{x}\hat{\boldsymbol{p}}_1$)對 $\boldsymbol{y}$ 做回歸，所估計之回歸係數爲 $\hat{\boldsymbol{q}}_{11}$，利用所得的 $\hat{\boldsymbol{p}}_1$ 及 $\hat{\boldsymbol{q}}_{11}$ 可求得第一個殘差，其殘差分別爲 $\hat{\boldsymbol{e}}^{(1)} = (\boldsymbol{X} - \hat{\boldsymbol{p}}_1\hat{\boldsymbol{p}}_1'\boldsymbol{X})$，及 $\hat{\boldsymbol{f}}^{(1)} = (\boldsymbol{y} - \hat{\boldsymbol{q}}_{11}\hat{\boldsymbol{p}}'\boldsymbol{X})$。第二個PLS向量 $\hat{\boldsymbol{p}}_2$ 爲 $\hat{\boldsymbol{e}}^{(1)}$ 對 $\hat{\boldsymbol{f}}^{(1)}$ 做回歸所得，並將其標準化，在求得 $\hat{\boldsymbol{p}}_2$ 時可發現其爲 $\boldsymbol{X}$ 投射於 $\hat{\boldsymbol{p}}_1$ 所得，

且 $\hat{\boldsymbol{p}}_2$ 直交於 $\hat{\boldsymbol{p}}_1$，所以利用 $\boldsymbol{X}$ 在 $\hat{\boldsymbol{p}}_1$ 及 $\hat{\boldsymbol{p}}_2$ 上之投射值對 $\boldsymbol{y}$ 做回歸，得到的回歸係數爲 $\hat{\boldsymbol{q}}_{21}$ 及 $\hat{\boldsymbol{q}}_{22}$。因此，第二個殘差分別計算出：

$$\hat{\boldsymbol{e}}^{(2)} = \left(\boldsymbol{X} - \hat{\boldsymbol{p}}_1\hat{\boldsymbol{p}}_1'\boldsymbol{X} - \hat{\boldsymbol{p}}_2\hat{\boldsymbol{p}}_2'\boldsymbol{X}\right)$$

$$\hat{\boldsymbol{f}}^{(2)} = \left(\boldsymbol{y} - \hat{\boldsymbol{q}}_{21}\hat{\boldsymbol{p}}_1'\boldsymbol{X} - \hat{\boldsymbol{q}}_{22}\hat{\boldsymbol{p}}_2'\boldsymbol{X}\right)$$

如此的程序連續做下去，直到找出一個合適的因子個數A，在過程中我們得到了 $\hat{\boldsymbol{p}} = (\hat{\boldsymbol{p}}_1,\ldots,\hat{\boldsymbol{p}}_A)$，進而可求得最後的向量爲 $\hat{\boldsymbol{q}}_A = (\hat{\boldsymbol{q}}_{A1},\ldots,\hat{\boldsymbol{q}}_{AA})$，我們利用所得的 $\hat{\boldsymbol{p}}$ 可得因子 $\hat{\boldsymbol{t}}$，進而以 $\hat{\boldsymbol{t}}$ 對 $\boldsymbol{y}$ 做回歸可得PLS的最終模式。

上述爲PLS的基本概念，一般使用非線性疊代淨最小平方(nonlinear iterative partial least squares, NIPALS)算程來決定因子 $\boldsymbol{t}_a, a = 1,\ldots,A$，並利用所得到的因子來對依變數 $\boldsymbol{Y} = [\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m]$作配適。這個方法最吸引人的地方是當預測值之間有高度相關或是線性相依時，能找到一個較適當的模式，以下將介紹NIPALS算程。

NIPALS**算程**

這節將介紹標準PLS算程，在這個算程的一開始是先將 $\boldsymbol{X}$ 跟 $\boldsymbol{Y}$ 置中，即給定 $\boldsymbol{X}_0$ 跟 $\boldsymbol{Y}_0$，接下來決定 $\boldsymbol{X}$ 因子得點(scores)矩陣 $\boldsymbol{T} = \{\boldsymbol{t}_1,\ldots,\boldsymbol{t}_A\}$，及 $\boldsymbol{Y}$ 因子得點矩陣 $\boldsymbol{U} = \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_A\}$，$\boldsymbol{t}_1$ 跟 $\boldsymbol{u}_1$ 由 $\boldsymbol{X}_0$ 跟 $\boldsymbol{Y}_0$ 加權可得：$\boldsymbol{t}_1 = \boldsymbol{X}_0\boldsymbol{w}_1$ 及 $\boldsymbol{u}_1 = \boldsymbol{Y}_0\boldsymbol{q}_1$，利用NIPALS的算程的疊代程序可決定加權向量，疊代程序如下：

$$\boldsymbol{w}_1 \propto \boldsymbol{X}_0'\boldsymbol{u}_1 \tag{1}$$

$$\boldsymbol{t}_1 = \boldsymbol{X}_0\boldsymbol{w}_1 \tag{2}$$

$$\boldsymbol{q}_1 \propto \boldsymbol{Y}_0'\boldsymbol{t}_1 \tag{3}$$

$$\boldsymbol{u}_1 = \boldsymbol{Y}_0\boldsymbol{q}_1 \tag{4}$$

此處符號 $\propto$ 並不只是代表比率，且含有對向量正規化(normalization)的功能，因此加權向量 $\boldsymbol{w}_1$ 及 $\boldsymbol{q}_1$ 的長度爲1。程序開始於由 $\boldsymbol{Y}_0$ 的行向量中選擇出一向量 $\boldsymbol{u}_1$，且 $\boldsymbol{u}_1$ 有著最大的變方；結束於 $\boldsymbol{w}_1$ 或 $\boldsymbol{t}_1$ 不再改變。另外在Höskuldsson (1988)的文中提到，加權向量 $\boldsymbol{w}_1$ 及 $\boldsymbol{q}_1$ 收斂於 $\boldsymbol{X}_0'\boldsymbol{Y}_0$ 的奇異向量分解(singular value decomposition, SVD)的向量中。因爲 $\boldsymbol{w}_1\boldsymbol{X}_0'\boldsymbol{Y}_0\boldsymbol{q}_1 = \boldsymbol{t}_1'\boldsymbol{u}_1 = (n-1)\mathrm{cov}(\boldsymbol{t}_1,\boldsymbol{u}_1)$，所以 $\boldsymbol{t}_1$ 跟 $\boldsymbol{u}_1$ 向量有著最大的共變方。

一旦由原始資料矩陣產生第一個因子 $\boldsymbol{t}_1$，即可將原始資料矩陣對 $\boldsymbol{t}_1$ 做回歸並得到新的殘差矩陣：

$$\boldsymbol{X}_1 = \boldsymbol{X}_0 - \boldsymbol{t}_1\left(\boldsymbol{t}_1'\boldsymbol{X}_0\right)\Big/\boldsymbol{t}_1'\boldsymbol{t}_1 \tag{5}$$

$$\boldsymbol{Y}_1 = \boldsymbol{Y}_0 - \boldsymbol{t}_1\left(\boldsymbol{t}_1'\boldsymbol{Y}_0\right)\Big/\left(\boldsymbol{t}_1'\boldsymbol{t}_1\right) \tag{6}$$

式(5)可以寫成

$$\boldsymbol{X}_1 = \boldsymbol{X}_0 - \boldsymbol{t}_1\boldsymbol{p}_1' \tag{7}$$

$\boldsymbol{p}_1$ 爲得點向量在 $\boldsymbol{X}$ 變數上的負荷向量，

$$\boldsymbol{p}_1 = \boldsymbol{X}_0'\boldsymbol{t}_1/(\boldsymbol{t}_1'\boldsymbol{t}_1) \quad (8)$$

負荷量爲 $\boldsymbol{X}$ 變數與PLS因子 $\boldsymbol{t}_1$ 間的相關強度。同樣的式(6)可寫成

$$\boldsymbol{Y}_1 = \boldsymbol{Y}_0 - \boldsymbol{b}_1\boldsymbol{t}_1\boldsymbol{q}_1' \quad (9)$$

$\boldsymbol{b}_1$ 爲估計的內部關聯係數，即表現著經過轉換的潛在變數之間的內部關聯，

$$\hat{\boldsymbol{u}}_1 = \boldsymbol{b}_1\boldsymbol{t}_1 \quad (10)$$

$$\boldsymbol{b}_1 = \boldsymbol{u}_1'\boldsymbol{t}_1/(\boldsymbol{t}_1'\boldsymbol{t}_1) \quad (11)$$

$\boldsymbol{Y}$ 因子的權重 $\boldsymbol{q}_1$ 與 $\boldsymbol{u}_1$ 的內部關聯係數 $\boldsymbol{b}_1$ 有助於說明資料中潛在變數的結構。

NIPALS 算程以增加 1 個維度重複進行上述 (1)~(11) 式，於維度 2 時求得 $\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{w}_2, \boldsymbol{t}_2, \boldsymbol{q}_2, \boldsymbol{u}_2, \boldsymbol{b}_2, \boldsymbol{p}_2$，$\boldsymbol{t}_2$ 直交於 $\boldsymbol{t}_1$，如此重複以上步驟，直到決定出A個因子爲止，因子個數A爲預測模式所決定[6]，本文中因子數個數的決定在下一節再加以討論。

選取最佳因子數

當利用NIPALS算程以獲得PLS模式，我們會考慮預測模式中只要使用少數的成分就足以充分且適當地表現出原始資料中所攜帶的訊息。然而，一般PLS模式中，對於要保留多少精確的成分數目通常並不是很明確，因此，過去對於保留成分數目最後的決定往往是主觀的。以下將介紹選取最佳因子數目的方法與準則。

(一)預測誤差(prediction error)

選取最佳因子數目通常以預測誤差爲判斷的指標，即在多變數檢量中以降低預測誤差爲目的，並以達到最小預測誤差平方和(prediction error sum of squares, PRESS)時的因子數目視爲最佳因子數[7]。多變數檢量是以靠著將化學或物理干擾訊息模式化來降低預測誤差爲目的。而在多維的檢量模式中，以每一個獨立的干擾現象由獨立的因子代替，所以當在最小的預測誤差時，就達到一個適當的最佳檢量因子數，下面我們介紹由預測誤差爲理論基礎所建立的準則。

(二)決定最佳因子數目的準則-交叉驗證法

這裡所使用的交叉驗證法爲一次刪除一個觀測值，以剩餘的n-1個觀測值來建立檢量模式，以 $\hat{\boldsymbol{y}}_k(-1)$ 代表利用資料中的剩餘觀測值來估計含有k個因子數目的PLS模式所獲得估值，以交叉驗證法求出MSE值，其式如下：

$$MSECV = \frac{\sum(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_k(-1))^2}{n-1} \quad (12)$$

模擬研究一

(一)目的與方法

此研究主要是利用實際資料特性來模擬產生新的資料，並在模擬的步驟中利用PLS建立模式，並設定模式中最佳因子數目，再利用交叉驗證法準則來檢視模擬資料中所建立的模式之最佳因子數目與先前所設定的最佳因子數目是否一致。

本研究利用米質研究中水稻之直鏈澱粉含量，藉著水稻白米粉末的光譜與化學值資料進行分析，模式中反應變數 $\boldsymbol{Y}$ 爲水稻的化學分析值，解釋變數 $\boldsymbol{X}$ 爲光譜值，光譜資料係利用近紅外光分析儀進行掃描，掃描之光波長範圍介於1,100與2,500 nm之間，每隔4 nm掃描一次，分別共可獲得351個波長的吸光值(log(1/R)，R爲擴散反射量)，如果我們將間距增加爲28 nm，則可以將351個解釋變數減爲51個解釋變數，然後利用此51個解釋變數及依變數進行逐步回歸法選出10個解釋變數。

在實際資料中包含了10個解釋變數及1個反應變數，觀測值個數n爲351個。解釋變數平均值爲

$\boldsymbol{m}$=(7.757E-18, 9.465E-18, -9.82E-18, -3.42E-18, -6.26E-18, 3.985E-17, -6.6E-17, -1.59E-17, 3.558E-17, 5.323E-17)，反應變數的平均值爲19.677436。我們用樣本大小爲351、解釋變數數目爲51的資料進行回歸分析，可得到誤差均方MSE=6，並以此誤差均方作爲此資料的樣本變方。

我們使用k個因子來估計PLS回歸中的 $\boldsymbol{\alpha}$ 與 $\boldsymbol{\beta}$ ，其中

$$\boldsymbol{a}_k = \bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}^T \boldsymbol{b}_k \tag{13}$$

$$\boldsymbol{b}_k = \boldsymbol{H}_k \boldsymbol{s} \tag{14}$$

$$\boldsymbol{H}_k = \boldsymbol{V}_k \left(\boldsymbol{V}_k^T \boldsymbol{S} \boldsymbol{V}_k \right)^{-1} \boldsymbol{V}_k^T \tag{15}$$

$$\boldsymbol{V}_k = \left[\boldsymbol{S}, \boldsymbol{S}\boldsymbol{s}, \ldots, \boldsymbol{S}^{k-1}\boldsymbol{s} \right] \tag{16}$$

首先以原始的解釋變數 $\boldsymbol{x}$ 矩陣的乘積平方和求得 $\boldsymbol{S} = \boldsymbol{S}_{xx}$ ， $\boldsymbol{x}$ 與 $\boldsymbol{y}$ 的乘積向量和爲 $\boldsymbol{s} = \boldsymbol{S}_{xy}$ ，在事先我們已假設 $\boldsymbol{x}$ 變數及 $\boldsymbol{y}$ 變數均經過置中處理， $\boldsymbol{V}_k$ 爲 $\boldsymbol{S}$ 與 $\boldsymbol{s}$ 的疊代矩陣，當我們選取了k個因子數目時， $\boldsymbol{V}_k = \left[\boldsymbol{S}, \boldsymbol{S}\boldsymbol{s}, \ldots, \boldsymbol{S}^{k-1}\boldsymbol{s} \right]$ ，在經過式(13)~式(16)運算後，可得到k個因子所估計求得的PLS回歸係數 $\boldsymbol{a}_k$ 與 $\boldsymbol{b}_k$ [(2)]。

此模擬研究中，我們先假定原始資料最佳因子數目爲4個，所以可以得到4個因子數目的PLS回歸係數 $\boldsymbol{a}_k$ 與 $\boldsymbol{b}_k$ ，變數個數爲10的模擬資料在選取因子數目爲4個的時候，淨最小平方回歸係數:

$b_{0,10}$=24.65，$\boldsymbol{b}_{10}$ =[133.70, 40.04, 400.69, 87.44, 31.36, 92.30, -187.52, -269.96, 48.46, 50.87]。

爲了比較變異程度，另外考慮 σ^2 爲12及18。而在樣本數方面，比較樣本大小對模式中最佳因子數目是否有影響，我們將模擬產生的樣本大小設定爲30與50，我們模擬所產生的資料計考慮以下不同的組合：

解釋變數個數(m): 10

變異程度大小(σ^2)：6、12及18

樣本大小(n)：30、50

(二)模擬的步驟

本模擬試驗的目的在於產生一組樣本大小爲n，並服從原始資料的基本分布而且最佳因子數目爲4個的樣本資料，在原始資料中我們先求得實際資料的樣本平均向量與樣本變方變積

矩陣，並利用SAS程式逢機產生一個$\boldsymbol{Z}$矩陣，其中每個元素均服從標準常態分布，利用上述特性我們可產生 $\boldsymbol{x}$ 矩陣： $\boldsymbol{x} = 1\bar{\boldsymbol{x}}' + \boldsymbol{Z}\boldsymbol{S}^{1/2}$ ，最後可利用回歸估式產生一組新的反應變數數據

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\varepsilon} \quad (17)$$

其中 $\boldsymbol{\alpha}$ 與 $\boldsymbol{\beta}'$ 分別爲以先前所估計的 $\boldsymbol{a}_k$ 與 $\boldsymbol{b}_k$ 所代入，且 $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$，再利用SAS軟體，進行上述的模擬1000次，我們可以得到6種不同的組合，藉以驗證交叉驗證法準則的準確度。

(三)結果與討論

我們利用以上模擬方法所產生的資料，驗證交叉驗證法在決定最佳因子數時的表現，將所得到的結果列表一。

表一、當觀測值個數爲 30，解釋變數個數爲 10，樣品變方爲 6、12 及 18 時，在交叉驗證法準則下所得 1000 次 MSE 的平均值

Table 1. Mean value of MSE of 1000 simulations using cross validation criteria when n=30, m=10, and var=6,12,18

The number of factors	MSECV Criterion		
	var=6	var=12	var=18
1	0.356476	0.564901	0.776004
2	0.312387	0.543273	**0.773955**
3	0.295582	0.540714	0.788042
4	**0.271822**	**0.537521**	0.805161
5	0.290994	0.584529	0.879332
6	0.322608	0.647030	0.972075
7	0.352177	0.704572	1.056857
8	0.386742	0.773403	1.160291
9	0.415913	0.831790	1.247719
10	0.439529	0.879059	1.318588

表一中爲解釋變數個數爲10，觀測值個數爲30，樣品變方爲6、12及18時的模擬資料下，利用交叉驗證法所得之MSECV的平均值，在表一中可以很清楚的看出交叉驗證法在不同變方下MSE的表現， MSECV值在模式中只含1個因子時較高，並隨著模式中因子數目的增加而降低，在模式中提高至含有4個因子數時MSECV值降到最低，然後再隨著因子數的增加，MSECV值也隨著上升，所以以MSECV預測的最佳因子數目爲4個，這個結果跟模擬資料一開始時的設定，最佳因子數目爲4是吻合的。資料中樣品變方增加爲12時，模式中估計最佳因子數目仍是4個，但當樣品變方再增加到18時，估計最佳因子數目爲2個，表示隨著樣品變方的增加，利用交叉驗證法預測最佳因子數目的能力會降低。

表二是將樣本數提高至50時的模擬資料對交叉驗證法求出的MSECV值所製成。利用交叉驗證法來預測模式中最佳因子數目，因樣本變方增加，預測模式中的預測能力降低，表二我們可以很清楚的看到，這樣的現象隨著樣本數目的提高而得到修正。因此利用交叉驗證法在足夠的樣本數目時可以準確的估計模式中的最佳因子數目。

表二、當觀測值個數爲 50，解釋變數個數爲 10，樣品變方爲 6、12 及 18 時，在交叉驗證法準則下所得 1000 次 MSE 的平均值

Table 2. Mean value of MSE of 1000 simulations using cross validation criteria when n=30, m=10, and var=6,12,18

The number of factors	MSECV Criterion		
	var=6	var=12	var=18
1	0.210283	0.333516	0.458473
2	0.176568	0.306496	0.436608
3	0.163881	0.297062	0.431689
4	**0.145689**	**0.285755**	**0.427129**
5	0.148428	0.297770	0.447631
6	0.157040	0.314936	0.472589
7	0.165607	0.331079	0.496812
8	0.173502	0.347026	0.520515
9	0.179757	0.359550	0.539331
10	0.184563	0.369127	0.553690

模擬研究二

(一)目的與方法

此模擬研究主要是想了解資料中共線性程度對在PLS模式下選取最佳因子數目的影響，因此我們利用模擬方法產生一筆新的模擬資料來比較的共線性問題所造成的影響。在一般的多變量資料中，爲了比較共線性程度，會將cov(x)的變方變積矩陣進行質譜分解(spectrum decomposition)，得到一序列的的特徵值 $\boldsymbol{m}_i$ 跟特徵向量 $\boldsymbol{p}_i$ ，我們可藉由特徵值的大小來了解資料中共線性的程度。因此，在我們所模擬的資料中，可藉由控制特徵值的大小來控制資料中共線性的程度，再利用所得到的特徵值跟特徵向量來進行資料的模擬。首先我們分別將特徵值 $\boldsymbol{m}_i$ 開根號再乘以由SAS程式所逢機產生的數值，可得到此資料的所有因子 $\boldsymbol{t}$ 。利用以下式子

$$\boldsymbol{x} = \boldsymbol{t}\boldsymbol{p}' + \boldsymbol{e} \tag{18}$$

$$\boldsymbol{y} = \boldsymbol{t} + \boldsymbol{f} \tag{19}$$

可分別模擬產生 $\boldsymbol{x}$ 及 $\boldsymbol{y}$ 矩陣，式(18)跟式(19)中的 $\boldsymbol{e}$ 跟 $\boldsymbol{f}$ 分別爲逢機產生的數值，我們可針對新的模擬資料就之前所設定的五個準則來判斷模式中最佳因子數目爲何。

(二)共線性程度的控制與模擬的步驟

在設定共線性程度的方面，我們可由特徵值 $\boldsymbol{m}_i$ 的大小來控制共線性的程度，如果資料中有一個或多個解釋變數彼此間是線性相依的，那所求得的特徵值當中就會有一個或多個值會很小，也就是說當資料中所求得的某些特徵值很小時，則資料中可能會有共線性問題的存在。

那我們如何比較資料中的共線性程度呢？我們可藉由

$$k = \frac{\lambda_{\max}}{\lambda_{\min}} \qquad (20)$$

來判斷之，當k值小於100時，表示共線性問題並不嚴重，當k值介於100-1000時，表示有很強的共線性，當k值超過1000時，表示資料中有非常嚴重的共線性問題存在。因此當第1個因子與第p個因子的特徵值比值k值大於1000時，我們選取的最適因子數目爲p-1個，以後的因子將因共線性程度嚴重而被刪除[8]。

模擬研究是產生一組共線性的模擬資料如下：

利用式(18)，如模擬產生的 $\boldsymbol{x}$ 矩陣爲

$$\boldsymbol{x} = \begin{pmatrix} -5 \\ -3 \\ -1 \\ 1 \\ 3 \\ 5 \end{pmatrix} \boldsymbol{t}_1 + \begin{pmatrix} 5 \\ -1 \\ -4 \\ -4 \\ -1 \\ 5 \end{pmatrix} \boldsymbol{t}_2 + \begin{pmatrix} -5 \\ 7 \\ 4 \\ -4 \\ -7 \\ 5 \end{pmatrix} \boldsymbol{t}_3 + \begin{pmatrix} 1 \\ -3 \\ 2 \\ 2 \\ -3 \\ 1 \end{pmatrix} \boldsymbol{t}_4 + \begin{pmatrix} -1 \\ 5 \\ -10 \\ 10 \\ -5 \\ 1 \end{pmatrix} \boldsymbol{t}_5 + \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \boldsymbol{e}_3 \\ \boldsymbol{e}_4 \\ \boldsymbol{e}_5 \\ \boldsymbol{e}_6 \end{pmatrix}$$

$\boldsymbol{t}$ 及 $\boldsymbol{e}$ 分別爲獨立的常態變數，$\boldsymbol{t}$ 爲 $\boldsymbol{x}$ 矩陣中的得點矩陣，其期望值爲0，變方爲 cov$(\boldsymbol{x})$ 矩陣的特徵值 m_i ，即var(t_i)= m_i ，i=1,…,5，而 $_x$ 的五個負荷量向量 $\boldsymbol{p}_1$ 、 $\boldsymbol{p}_2$ 、 $\boldsymbol{p}_3$ 、 $\boldsymbol{p}_4$ 及 $\boldsymbol{p}_5$ 爲直交的向量。所以模擬資料的設定上，以特徵值比k=1000爲基準，假設模式中最佳因子數目爲2，3跟4，我們分別將特徵值設定爲 $\boldsymbol{m}$ = (10, 5, 10^{-8}, 10^{-8}, 10^{-8}, 10^{-8})、 $\boldsymbol{m}$ = (10, 5, 0.5, 10^{-8}, 10^{-8}, 10^{-8})、 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 10^{-8}, 10^{-8})、 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 0.01, 10^{-8})，並利用新設定的特徵值來取得資料中的得點矩陣 $\boldsymbol{t}$ ，最後由所設定的 $\boldsymbol{t}$ 跟 $\boldsymbol{p}$ 可模擬出 $\boldsymbol{x}$ 矩陣。

$\boldsymbol{y}$ 的產生是利用 $\boldsymbol{y} = \boldsymbol{tq} + \boldsymbol{f}$ ，但是因爲 $\boldsymbol{q}$ 不易求得，因此這四個模擬資料中反應變數 $\boldsymbol{y}$ 的產生方法簡略如下：

$$\boldsymbol{y} = \sum_{1}^{i} \boldsymbol{t}_i + \boldsymbol{f}$$

其中 $E(f) = 0$ ， $\text{var}(f) = 0.25$ 。

最後由式(18)跟式(19)以SAS軟體進行1000次模擬得到新的資料。這次的模擬所產生的資料爲解釋變數個數爲6，樣本個數則設定爲60，再利用MSECV準則來判斷模擬資料中的最佳因子數目。

(三)結果與討論

利用以上模擬方法所產生的資料，驗證交叉驗證法在決定最佳因子數時的表現，將所得到的結果列於表三。

表三、由模擬資料來比較不同特徵植下所得之 MSE 平均值
Table 3. Mean value of MSE for simulated data with various eigenvalue

The number of factors	MSECV			
	$\boldsymbol{m}$ =(10, 5, 10^{-8}, 10^{-8}, 10^{-8}, 10^{-8})	$\boldsymbol{m}$ =(10, 5, 0.5, 10^{-8}, 10^{-8}, 10^{-8})	$\boldsymbol{m}$ =(10, 5, 0.5, 0.1, 10^{-8}, 10^{-8})	$\boldsymbol{m}$ =(10, 5, 0.5, 0.1, 0.01, 10^{-8})
1	0.026049	0.033904	0.035624	0.035790
2	**0.006575**	0.013132	0.014821	0.014990
3	0.007120	**0.007028**	0.008130	0.008305
4	0.007340	0.007527	**0.007327**	**0.007469**
5	0.007477	0.007722	0.007733	0.007603
6	0.007616	0.007864	0.007902	0.007902

表三中的模擬資料中特徵值 $\boldsymbol{m}$ = (10, 5, 10^{-8}, 10^{-8}, 10^{-8}, 10^{-8})所模擬出的，我們將模式中最佳因子數目設定爲2，利用MSECV準則對此資料求得MSE值，由表三可很明顯看出利用MSECV準則在模式中有兩個因子時，MSE值最小，所以其估計模式中最佳因子數目是2，特徵值 $\boldsymbol{m}$ = (10, 5, 0.5, 10^{-8}, 10^{-8}, 10^{-8})所求得的模擬資料中，第1個因子與第3個因子的特徵值比的k值小於1000，而第1個因子與第4個因子的特徵值比就超過1000，所以我們判斷此資料的最佳因子數目爲3，MSECV準則估計模式中最佳因子數目跟預設的3個因子數目相同，同樣的，在以特徵值 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 10^{-8}, 10^{-8})求得的資料中，設定模式中最佳因子數目爲4，在此筆資料中MSECV準則的表現仍是好的。而在以特徵值爲 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 0.01, 10^{-8})所求得的模擬資料的，我們將第1個特徵值跟第5特徵值比k值設定剛好爲1000，藉此來比較模式中最佳因子數目爲何，在表三中我們可以很明顯的看出，MSECV在模式中有4個因子數達到最小值，也就是模式中最佳因子數目爲4個，所以我們利用特徵值比k值爲1000來當作判斷是否有嚴重共線性程度是可行的。因此 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 10^{-8}, 10^{-8})跟 $\boldsymbol{m}$ = (10, 5, 0.5, 0.1, 0.01, 10^{-8})所模擬的資料中，兩個模式中皆獲得最佳因子數目相同。

結　　語

當我們面對具有嚴重共線性的資料，使用多重線性回歸無法建立準確的回歸模式時，利用PLS方法來建立檢量模式可達到良好的效果，並可以交叉驗證法求得最佳因子數目來建立最佳PLS檢量模式。在模擬NIR米質研究資料時，建立PLS檢量模式可改善資料共線性問題，因此面對米質研究中NIR資料具有嚴重共線性時，利用PLS回歸來建立檢量模式，並以交叉驗證法取得模式中最佳因子數目，是值得推薦的。

參考文獻

1. 楊佩雅　1999　利用近紅外光分析儀檢測稻米碘呈色度與黏度特性　國立中興大學農藝系碩士論文。

2. Denham, M. C. 2000. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. J. Chemom. 14:351-361.
3. Geladi, P. and B. R. Kowalski. 1986. Partial least squares: a tutorial. Anal. Chimica. Acta. 185: 1-17.
4. Garthwaite, P. H. 1994. An interpretation of partial least squares. J. Amer. Statist. Assoc. 89: 122-127.
5. Höskuldsson, A. 1988. PLS regression methods. J. Chemom. 2: 211-228.
6. Jong, S. D. 1993. SIMPLS: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. 18: 251-263.
7. Liu, H., E. W. Robert, I. J. Robert and S. W. Neil. 1999. PRESS model selection in repeated measures data. Comput. Statist. Data. Anal. 30: 169-184.
8. Montgomery, D. C. and E. A. Peck. 1982. Multicollinearity. pp.287-346. *In* Introduction to Linear Regression Analysis.Wiley, New York.
9. Næs, T. and H. Martens.1985. Comparison of prediction methods for multicollinear data. Commun. Statist.-Simula. Computa.14(3): 545-576.

The Simulation Study on the Cross Validation for Choosing the Optimal Number of Factors on PLS[1]

Yi-Lun Liao[2] and Bo-Jein Kuo[3]

ABSTRACT

It's useful to solve the multi-collinear problem of data by partial least squares regression. The partial least squares regression is modified from principal component regression. PLS transforms the explanatory-variable matrix **X** into several orthogonal **t** scores, and it adds **y** information into scores **t** simultaneously. On PLS regression, the **X** matrix is decomposed into **t** scores and **p** loadings; **u** scores and **q** loading for **y** matrix. It introduces the relationship of **X** and **y** by using the **w** weights. This is the reason that PLS always performs better than PCR.

This study uses cross validation to choose the number of fitting factors by two simulation studies. We simulate the first data by using real NIR data character and the simulation data have the same character with real NIR data. We only control the number of fitting factors, sample size, explained variables and sample variance of the model. Finally, we check the effects for the number of fitting factors. The second simulation data is generated by the degree of collinear. On this simulation data, we control the degree of collinear. When the degree of collinear is too serious, we can use CV to keep some useful factors, and to reject some factors, which cause the collinear.

Generally speaking, controlling multi-collinearity data is effectible by using PLS. we can get good result of choosing the number of fitting factors by using MSECV. And at many conditions using the MSECV will get the accurate estimation.

Key words: NIR, partial least squares method, multi-collinearity data, cross validation, simulation study.

---

[1]Contribution No. 0715 from Taichung DARES, COA.
[2]Assistant Researcher of TDARES, COA.
[3]Professor of the Department of Agronomy, NCHU.