多變數分析在農業科技之應用

呂秀英*
行政院農委會農業試驗所農藝組

摘要

農業科技上常會針對某研究主題同時測量一大堆不同變數(調查性狀)的資料，但我們並非針對個別變數進行統計分析，而是將所有變數合起來共同討論，針對這樣資料的統計分析，就需要用到多變數分析技術。本文試以不涉及統計理論從實用角度出發，針對農業科技研究上常用的多變數分析法，以實例分析說明其意義、應用場合、分析流程及結果解釋，期能有助於農業科技研究之參考。這些多變數分析方法包括主成份分析、對應分析、因素分析、集群分析、判別分析、路徑分析及典型相關分析。

關鍵詞：多變數統計分析、主成份分析、對應分析、因素分析、集群分析、判別分析、路徑分析、典型相關分析。

Applications of Multivariate Analysis in Agricultural Science

Hsiu-Ying Lu*

Agronomy Division, Taiwan Agricultural Research Institute, 189 Chung-Cheng Road, Wufeng, Taichung Hsien 41301, Taiwan ROC

ABSTRACT

Data containing many variables (measured characters) are often collected in agricultural science research. An interpretative analysis of multivariate data considers all the variables simultaneously is required, and entails the use of multivariate statistical analysis. This paper provides a non-statistical, practical overview of the commonly used multivariate methods in agricultural science research, including principal component analysis, correspondence analysis, factor analysis, cluster analysis, discriminant analysis, path analysis, and canonical correlation analysis. The meaning, applicability, analytical procedures, and results interpretation of these methods are presented with cited examples. Our objective is to provide the researcher an intuitive understanding of multivariate analysis and their applications in agriculture.

Key words: Multivariate statistical analysis, Principal component analysis, Correspondence analysis, Factor analysis, Cluster analysis, Discriminant analysis, Path analysis, Canonical correlation analysis.

前言

農業研究是現代科技中應用最廣、最活躍及最富挑戰性的領域之一，追根溯源，它與數學的發展，尤其是統計學的發展，具有同步性。從應用數學知識來解決農業中的實際問題，首重於試驗設計的完善以及資料分析方法的正確使用。多變數分析法(multivariate analysis)是從傳統的統計學中發展起來的一個分支，是一種綜合分析方法，它能夠在多個研究對象和多個指標相互關聯的情況下分析出它們的統計規律，非常適合農業科技研究的特點。所謂變數(variable, 又稱變量)，就是所觀測的特性，如株高、乾物重、產量、糖分含量、花色等。變數的結果，即所觀測特性的測定值，稱為觀測值(observation,又稱變值,variate)。最簡

* 通信作者, iying@wufeng.tari.gov.tw
投稿日期：2006 年 8 月 14 日
接受日期：2006 年 8 月 24 日
作物、環境與生物資訊 3:199-216 (2006)
Crop, Environment & Bioinformatics 3:199-216 (2006)
189 Chung-Cheng Rd., Wufeng, Taichung Hsien 41301, Taiwan ROC

單的觀測值為單變數，例如一群植株的株高或乾物重。大多數場合我們通常會針對某研究主題測量了一大堆不同變數的資料，雖然我們同時測量了這些變數，但我們並非針對個別變數進行統計分析，而是將所有變數合起來共同討論，針對這樣資料的統計分析，就需要用到多變數分析技術。

構成多變數分析模型的數學方法並不新穎，如與多變數有關的基本概率分布之常態分布源自 30 年代；然而，當隨機變數較多時，多變數分析的計算工作量極端繁冗，沒有電腦根本無法完成，因此直到有了電腦之後，多變數分析技術才進入實用階段並迅速發展 (Kenkel *et al*. 2002, Manley 2004)。近 20 年來，隨著電腦應用技術的發展和科技研發的迫切需要，多變數分析技術被廣泛地應用於地質、氣象、水文、醫學、農業、生物、生態、工業、心理、社會和經濟等眾多領域，已經成為處理多因素、多指標特徵問題的最有效且實用的理論和方法(Yuan and Zhou 2003, Manley 2004)。

多變數分析發展至今，已產生各種方法，各有其適用的場合，但由於其涉及較複雜的計算，對於一般農業研究者而言，往往造成怯步。而一般教科書則多著重於統計理論和公式計算，所舉列的實例也未必是農業相關的試驗資料。本文從實用角度出發，在以不涉及統計理論下，強調方法之正確選用與分析結果之正確解讀，針對一些在農業科技上常用的多變數分析法，以實例分析說明其意義、應用場合、分析流程及結果解釋，期能有助於農業科技研究之參考。這些多變數分析方法包括主成份分析 (principal component analysis) 、對應分析 (correspondence analysis)、因素分析(factor analysis)、集群分析(cluster analysis)、判別分析(discriminant analysis)、路徑分析(path analysis) 及典型相關分析 (canonical correlation analysis)，其應用場合各有不同，如 Fig. 1 所示，但最根本之目的都是用來簡化資料結構，以達到釐清變數之間的關係。

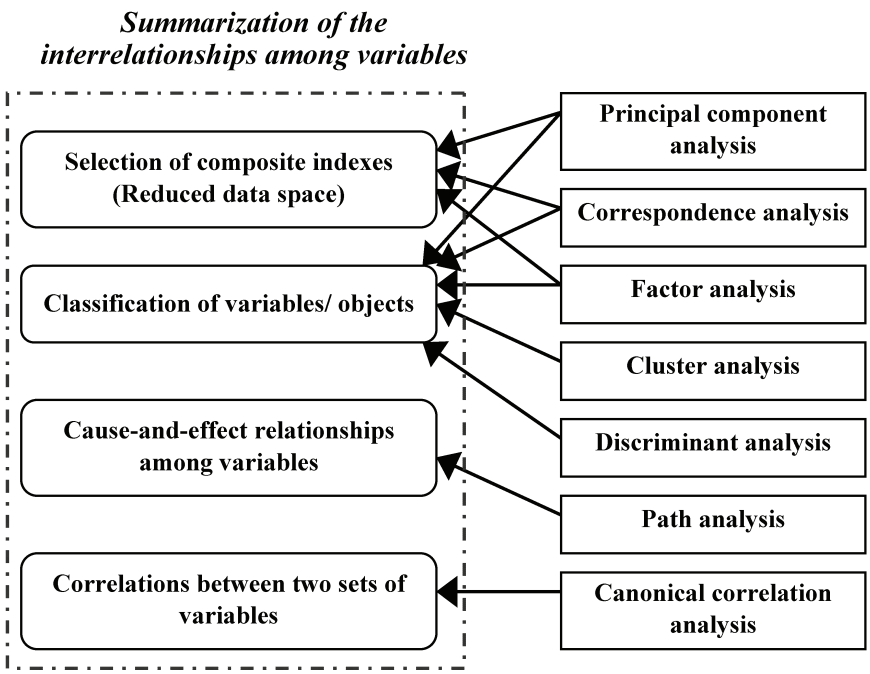

Fig. 1. The goals of multivariate analysis.

主成份分析

自然界的事物往往為多個指標的綜合結果，例如代表一個果實品質的諸種特徵包括大小、形狀、甜度、表皮外觀、纖維質含量、水分含量等性狀；反映一個昆蟲形態的諸種特徵包括體長、體寬、前翅長、觸角長、後翅鈎數等性狀；代表氣候條件中的氣象因素包括氣溫、日照時數、日照強度、風速等。以果實品質而言，可能有些品種的果實較大但甜度欠低，另有些品種則甜度和水分含量很高卻較小，其他品種又可能夠大也甜但纖維較多，因此我們不能單就一個特定性狀來決定品質的好壞，而必須將所有性狀變數一起共同考量，即應使用一個綜合性指標。

就統計上而言，數值能產生愈大變異，愈能反映彼此間之差異，但是將這些不同測量單位的性狀數值相加起來或求其平均，未必可斷言此指標就是最好的。若是對不同變數給予不同的權重，那麼加權的比重到底要設多少，也很難有一個明確的基準。因此在縮減原始變數個數以產生一個綜合性指標時，我們總希望在選擇變數(愈少愈好)與所能解釋的變異(愈多愈好)中達到平衡點，那意味著所找到新的變數不但精簡且具有代表性。在這些場合中因變數太多而不易處理時，利用變數間的相依結構，將手上許多相關性很高的變數轉化成彼此互相獨立的變數(線性組合)，能由其中選取較原始變數個數少，且能解釋大部分資料中的變異的幾個新變數，此即所謂的主成份(principal component)，而這幾個主成份也就成為我們用來解釋資料的綜合性指標。經由這種主成份分析(一般簡稱 PCA)，我們可以搜尋出主要潛在因子，捨棄次要因子，以簡化資料結構，並用來選擇變數的綜合性指標，使觀測值在這些主成份方面顯出最大的個別差異來。

一般電腦軟體進行主成份分析的主要流程，是將輸入的原始變數先求其相關矩陣(correlation matrix)後，最後計算出相關矩陣的特徵值(eigenvalue)及其貢獻量，此用來選擇主成份，進而再從計算出的特徵向量(eigenvector)來解釋各主成份的內涵(即這些主成份主要是由哪些原始變數所組成)。主成份之選擇，主要原則是依據每一主成份的特徵值(λ)大小。特徵值愈大，代表該主成份的解釋力愈強，但一般方法不只一種，例如可保留 λ >1 的主成份、保留 λ >0 的主成份、根據 λ 趨勢圖(scree plot)將徙降後曲線走勢趨於平坦之主成份捨棄不用、抽取之主成份解釋 75%變異後繼續抽取之主成份的解釋變異量少於 5%則捨棄不用等等。主成份數目之決定並沒有絕對之方法，除上述方法外，還有其他方法可用。

以 14 個木薯品種的 9 個農藝性狀進行主成份分析為例(Yang *et al*. 2006)，由資料的初步結果得知這 9 個性狀變數之間彼此相關，因此採用主成份分析找出潛在的新變數或成份，以期利用較少數而且彼此獨立的新變數，來建立木薯品種選拔的理論依據。分析結果如 Table 1 所示，從特徵值及其貢獻量，可知取前四個主成份來代表原來的 9 個變數解釋，總解釋量可達 85.69%，已經足夠代表原資料變數特性。進而從特徵向量值（Table 1）可知，第一主成份與株高、莖徑有正向相關，但其與塊根乾物質率和塊根澱粉含量呈負向相關，說明了第一主成份大的品種，其株高、莖徑表現高但塊根的乾物質和澱粉含量表現低，換言之追求過高的株高、莖徑，將會造成塊根的乾物質和澱粉含量的下降，故第一主成份適中為好；第二主成份大的品種塊根數多、塊根較長且產量高，第三主成份大的品種之塊根的鮮重、乾物質和澱粉含量都高，第四主成份大的品種塊根鮮重高且產量高。產量的構成主要在第二和第四主成份上，而塊根澱粉含量和塊根乾物質率對木薯加工很重要，主要在第三主成份上。選擇木薯時應注意產量和澱粉含量均高的品種，所以要著重對第二、三、四主成份的綜合性狀選拔。在木薯試驗中利用特徵向量值可建

構前四個主成份與原變數之間的線性關係，即可寫成以下公式：

$$PC_1 = 0.47X_1+0.37X_2-0.12X_3-0.19X_4+0.27X_5+0.28X_6-0.47X_7-0.46X_8+0.06X_9$$
$$PC_2 = 0.11X_1+0.22X_2+0.58X_3+046X_4+0.28X_5-0.13X_6+0.01X_7+0.09X_8+0.54X_9$$
$$PC_3 = 0.28X_1+0.38X_2-0.31X_3+0.20X_4+0.00X_5+0.50X_6+0.43X_7+0.45X_8-0.03X_9$$
$$PC_4 = -0.31X_1-0.43X_2-0.11X_3+0.13X_4-0.07X_5+0.65X_6-0.19X_7-0.08X_8+0.46X_9 \quad [1]$$

此即主成份分析是將主成份表示成原始變數的線性組合，換言之，主成份分析是原始變數的綜合指標。

利用主成份值可進而計算各樣本間的距離，以進行樣本的分類，不過在較多變數的研究上，主成份分析扮演的角色通常是一種手段多於目的，即它本身常不是研究的最後輸出(目的)，而是作為其他分析，如迴歸分析、因素分析、集群分析等的預備工作。

以上所述，是求變數的主成份，這種主成份稱為 R 型主成份。利用同樣的資料，當我們把樣本看成變數而獲得樣本的主成份，稱為 Q 型主成份。

對應分析

延續上述主成份的概念，對於實際問題，當我們希望同時求得變數和樣本的主成份，對變數和樣本對應進行主成份分析的方法，稱為對應分析。對應分析的優點及實質用途，是可使特徵值相同的 R 型和 Q 型主成份能用同一座標軸表示，以便在同一座標平面上可同時標示出樣本和性狀的散佈圖，以同時表達出變數和樣本兩者之間的相互關係，從而可檢視出各類樣本的主要變數為何。這種同時將變數和樣本標在一起的圖，稱為雙標圖(biplot)。但對應分析並非陳述概念之間的因果關係，而僅僅是統計學意義上的對應關係而已。

對應分析最早由 Benzécri (1973)提出，演變迄今，在很多不同時代裡該法的名稱和原理基礎略有不同(Oksanen 2004)：如最適尺度法(optimal scaling)、相互平均法(reciprocal averaging)。最適尺度法是市場、

Table 1. Results of principal component analysis performed on the 9 agronomic characters of 14 cassava varieties (adapted from Yang *et al*. 2006).

Characteristic	Principal component eigenvector			
	PC_1	PC_2	PC_3	PC_4
Plant height (X_1)	0.47	0.11	0.28	-0.31
Stem diameter (X_2)	0.37	0.22	0.38	-0.43
Root number (X_3)	-0.12	0.58	-0.31	-011
Root length (X_4)	-0.19	0.46	0.20	0.13
Root thickness (X_5)	0.27	0.28	0.00	-0.07
Root fresh weight (X_6)	0.28	-0.13	0.50	0.65
Root dry matter content rate (X_7)	-0.47	0.01	0.43	-0.19
Root starch content (X_8)	-0.46	0.09	0.45	-0.08
Yield per plot (X_9)	0.06	0.54	-0.03	0.46
Eigenvalue (λ)	3.15	2.40	1.16	1.00
Contribution (%)	35.01	26.71	12.90	11.08
Cumulative contribution (%)	35.01	61.72	74.61	85.69

社會、心理學上常用的方法，又稱雙重尺度法(dual scaling, Nishisato 1980)；相互平均法為 Hill (1974)所提出，目前名稱仍維持存在，但其算法幾乎罕見；現代統計軟體均以對應分析來稱呼，且算法採卡方測度 (χ^2 metric)的加權主成份，即分析過程如同主成份分析一樣求特徵值分析，但不同處在於以「卡方測度」取代「歐氏測度」(Euclidean metric)（ter Braak 1986, 1987）。簡言之，對應分析相當於列聯表 (contingency table)資料的加權之主成份分析，是一種用來尋求列聯表的行列兩種變數之間聯繫的低維圖示法。對應分析無需太多統計前提，它所處理的資料形式，不限連續(continous)或離散(discrete)變數，離散變數且不限計數(count)或序位(ordinal)變數，即使數字符號的虛擬變數(例如不存在=0，存在＝1)亦可。因此，對應分析非常適合於描述生物性資料，早已廣泛應用於生態學，尤其是植物生態學(Lepš and Šmilauer 1999)，現並大量應用到其他各方面之研究領域(Beh 2004)，包括基因組分析，如胺基酸組成(Tekaia *et al*. 2002)、密碼子使用偏好性(Gupta and Ghosh 2001, Liu *et al*. 2004a, b)、微陣列分析(Tan *et al*. 2004)等。

一般電腦軟體進行對應分析的主要流程，會先將變數資料進行標準常態轉換(Z 矩陣)，使 Z 對指標及樣本具有對等性，再計算出變數和樣本的主成份，然後以 Z 矩陣看作原始資料，再計算出變數和樣本的主成份值，最後將特徵值相同的各變數(R 型)和樣本(Q 型)的主成份值標在同一座標圖上表示，如此就可解釋變數、樣本以及兩者之間的關係。

以 Yuan and Zhou (2003)所列之 14 個玉米雜交種共調查 10 項性狀的實例來說明，經對應分析，各取得 R 型和 Q 型各兩個主成份後，同時將 14 個樣本和 10 個性狀變數的主成份值標在同一張圖上(Fig. 2 中空心圓點表示各品種樣本，實心圓點表示各性狀變數)，藉此可直觀地研究品種樣本和性狀點群之間的關係。由 Fig. 2 可以看出品種 13 和 14 相異於其他品種，主要是由於它們與性狀 G (千粒重)關係密切，而與其他性狀關係較遠；若試驗者特別對玉米籽實蛋白質含量 (性狀 H)感興趣，則由關係遠近可說明品種 1~12 的蛋白質含量高，而品種 13 和 14 則含量較低。

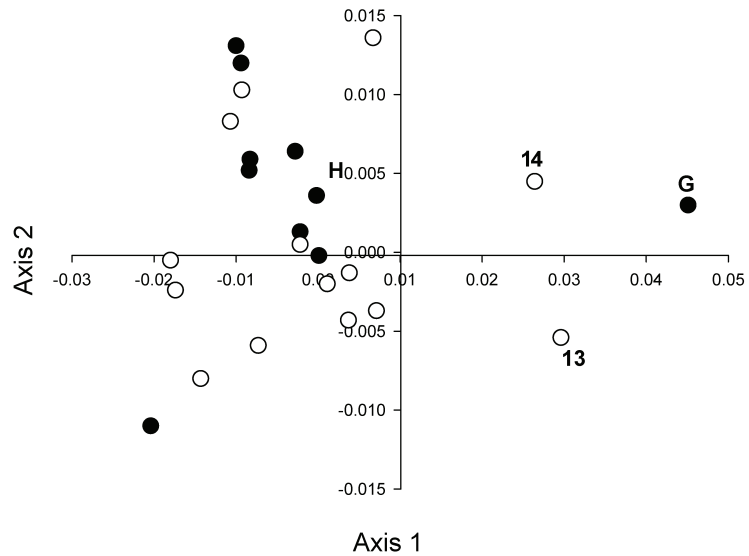

Fig. 2. Combined representation of the first two axes of correspondence analysis performed on the 10 traits of 14 maize varieties. o and • represent variety and trait, respectively. (adapted from Yuan and Zhou 2003).

因素分析

很多測定性狀之間通常彼此相關，形成的背景原因各式各樣，而其共同原因稱為共同因素(common factor)。我們希望能用這些較少的共同因素來表現原先的資料結構，以尋求基本結構、簡化觀測系統，找出資料背後隱藏的含意或潛在特徵，這是因素分析的主要目的。例如將某種紅酒與對照酒作比較，就諸多個品評項目在-4 到+4 的量表上給分，如酸味、苦味、甜味等，可利用因素分析將資料分成少數幾個共同因素，並將這些因素命名為辛辣、總品質、香醇等；將不同水稻品種穀粒形態的諸多項幾何特徵之測量值，利用因素分析將資料分成少數幾個共同因素，以從中找出最能代表水稻粒形的量化指標；將不同玉米雜交種的諸多種性狀，用因素分析將資料分成少數幾個共同因素，如第一因素反映出成熟期、分蘗、粒數與抽穗期，第二因素主要反映成熟期和粒重，第三因素反映株高。

因素分析是主成份分析的衍生方法，因此一般電腦軟體進行因素分析的主要流程，仍會先經由主成份分析程序，計算出相關矩陣的特徵值及特徵向量以決定共同因素的個數，然後從相關矩陣中抽取共同因素，計算因素負荷矩陣(factor loading matrix)及其變方，並旋轉因素以增加變數與因素之間關係的解釋，最後計算出旋轉因素的得分值(rotated factor score)，使能對變數或樣本進行分類。轉軸的方法很多，但基本原則在於使經過轉軸後的因素矩陣中每一個變數都只歸於一個或少數幾個因素上，使矩陣中零或接近於零的因素負荷量增多，以減低因素的複雜性，使因素的解釋由繁雜趨向簡單。決定因素數目的方法，與主成份分析一樣，都是依據每一因素的特徵值(λ)大小，特徵值愈大，代表該因素的解釋力愈強。

以 69 種茶葉的品質評鑑(Gao *et al*. 2003)為例，共有形狀、色澤、水色、香味 4 個變數對茶葉的品質產生影響，但這 4 個變數之間彼此可能有相關，因此，應用因素分析法來解釋資料背後隱藏的含意或潛在特徵。因素分析結果如 Table 2 所示，從特徵值及其貢獻量，可決定取兩個共同因素，因素一之特徵值為 2.1941，解釋總變異的 54.85%，因素二之特徵值為 0.3486，解釋總變異的 8.72%，故兩共同因素共同解釋變異量 63.57%。使用最大變方法(maximum variance method)進行轉軸後可將因素與因素間的差異拉開來。轉軸後因素一主要說明色澤、水色 、香味的變異情形，解釋變異量佔 48.95%；因素二主要說明形狀和色澤的變異情形，解釋變異量佔 14.62% (Table 2)。所以我們能找出轉軸後兩

Table 2. Results of factor analysis performed on the 4 quality characters of 69 tea varieties grown in spring season (adapted from Gao *et al*. 2003).

Characteristic	Factor loadings			
	Unrotated		Rotated	
	F_1	F_2	F_1	F_2
Leaf shape (X_1)	0.2698	0.4796	0.0804	0.5444
Leaf color (X_2)	0.8445	0.2195	0.7102	0.5069
Water color (X_3)	0.8495	-0.1836	0.8590	0.1323
Fragrance (X_4)	0.8286	-0.1916	0.8423	0.1174
Eigenvalue (λ)	2.1941	0.3486	1.9581	0.5846
Contribution (%)	54.85	8.72	48.95	14.62
Cumulative contribution (%)	54.85	63.57	48.95	63.76

個因素 F_1 和 F_2，以建構各變數的因素模型為

$$X_1 = 0.0804F_1 + 0.5444F_2 + e_1$$
$$X_2 = 0.7102F_1 + 0.5069F_2 + e_2$$
$$X_3 = 0.8590F_1 + 0.1323F_2 + e_3$$
$$X_4 = 0.8423F_1 + 0.1174F_2 + e_4$$
[2]

F_1、F_2 稱為這些變數的共同因素，而 e_1、e_2、e_3、e_4 誤差項各為 X_1、X_2、X_3、X_4 的獨特因素。

因素分析的用途，除了找出較少無相關的共同因素來簡化資料結構外，可進而依據因素得分值的結果，將各樣本進行排序，例如前例我們可以得到各個茶種的分數，分數愈高代表茶的品質愈好。另外，還能依據因素得分值，在因素軸所構成的空間中繪製散佈圖，由於愈接近的點表示彼此間關係愈相近，故可對樣本進行分類處理；若使用因素負荷量繪製散佈圖，則可對變數進行分類並表現變數間的關係。一個由兩個因素軸所構成的二維散佈圖，例示如 Fig. 3；在 3 個因素個數之場合，則該散佈圖會是一個三維空間，一旦因素個數超過 3 個以上，通常以兩兩成對因素間的二維散佈圖來各自展現並綜合解釋。

因素分析與主成份分析很容易被混淆不清，其間最大的不同在於：

1. 因素分析目的在建立模型，以呈現出變數和因素之間的內在聯繫關係，故才有誤差項；而主成份分析只作變數轉換，不建立模型，故無所謂誤差項。
2. 因素分析目的是要使因素比變數的數目少，且盡可能選取較少的因素，以便盡可能建構一個結構簡單的模型，重視在如何解釋變數之間的「共變異(covariance)」問題；而主成份分析目的在將一組具相關性的變數轉換成一組獨立的新變數，關鍵在「變異(variance)」問題，故主成份和原始變數的數目相等，但我們可以從中選擇貢獻量夠大的前幾個主成份來解釋資料中大部份的變異。
3. 因素分析是將原始變數表示成新因素的線性組合，即原始變數是新因素的綜合指標（如公式[2]）；而將主成份表示成原始變數的線性組合，即主成份是原始變數的綜合指標（如公式[1]）。

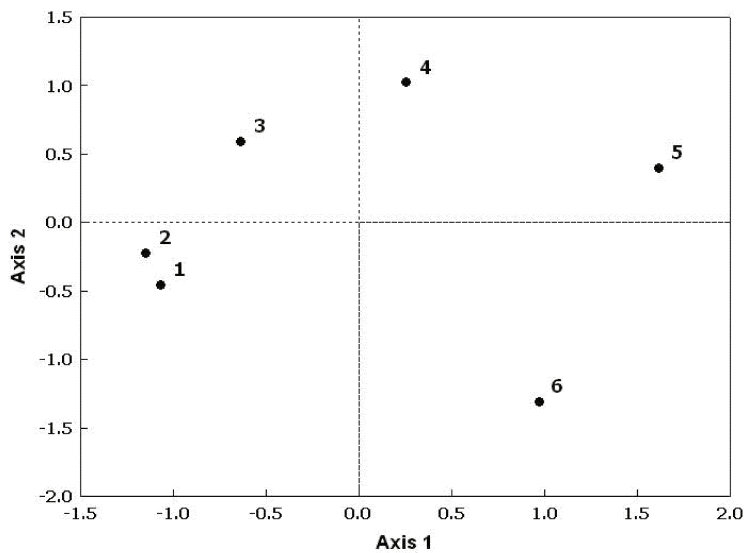

Fig. 3. Scattergram of the 6 objects in a two-dimension space. The first and second axes can be obtained from principal component analysis, factors analysis or correspondence analysis.

集群分析

集群分析是將具有多個變數的一群樣本加以區分歸類，使得性質特徵相近者納入一類。經濟、社會、人口、生物研究等領域中都存在大量分類研究、構造分類模式的問題。過去人們主要靠經驗和專家知識，作定性分類處理，很少利用數學方法，以致許多分類往往帶有主觀性和隨意性，無法揭示客觀事物內在的本質差異性和關連性，特別是對於多因素、多指標的分類問題，定性分類更難以實現準確分類。因此，我們利用集群分析，以客觀統計分析的方式，將一批樣本或變數，按照它們在性質上的親疏程度(彼此間距離或某種相似係數)，「物以類聚」地把相似的個體（或觀測值）歸於一群。集群分析在農業研究上的應用，例如調查蛋白質、碳水化合物、脂肪、卡路里、維生素等營養成份含量，對不同品牌穀類製品進行分群；調查各種血液蛋白質位點基因頻率，對不同黃牛品種進行分群；調查各種形態特徵，對某種昆蟲不同品種或棲息地進行分群等。

雖然集群分析是多變數分析方法中較簡單的一種，但其分析方法和結果判讀，一直以來爭議不斷。一個完整的集群分析流程，除了電腦執行分析程序之外，還包括對於電腦分析結果的合理決策，共包含六個主要步驟，如 Fig. 4 所示，簡單說明如下：

一、集群分析的電腦執行分析程序分為三個主要部份：

(1)變數的選擇－如各種資料轉換方法可以採資料中心化、對數轉換、資料標準化等。

(2)相似性的衡量－可採用彼此間距離或某種相似係數，距離統計量又分為歐氏距離、絕對值距離、馬氏距離等，相似係數則可以是相關係數、指數相似係數、非參數法等。

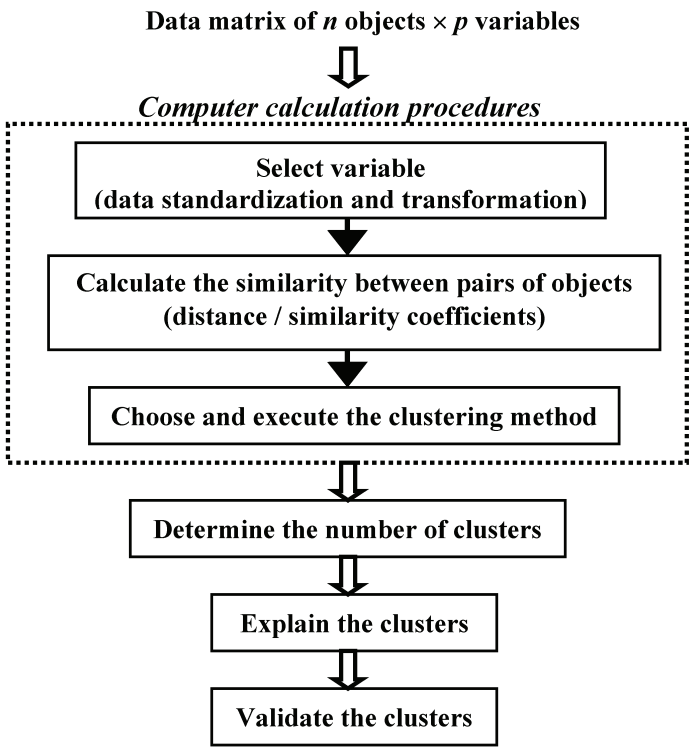

Fig. 4. Flowchart showing the six steps of clustering analysis.

(3)集群方法的選擇－可以利用系統集群法(如單連法、全連法、均連法、Ward 最小變方法等)、逐步集群法、逐步分解法、有序樣本之集群等。有關集群分析之方法，眾說紛云，各種新的方法不斷被提出，但仍沒有一種是最佳的方法且可解決所有問題，而分析出來的結果若沒有其他訊息配合，則結果究竟適不適合，也是一大考驗。因此在進行分析時，對於各種因子(變數、尺度、標準化、相似性係數、集群方法)之選擇應思慮周密，試驗者的目標、主觀性判斷、極端值之判斷皆非常重要，不同之抉擇往往造成不同的分群結果。為使集群分析結果能達到客觀性、穩定性、預測性，建議應佐以他法輔助，如圖形法等；此外專業方面的判斷和解釋非常重要，也不能忽視。單連法(single link)又稱近鄰法，以最近的個體來判斷兩者相似性，容易造成個體間的連鎖(chaining)，故所形成的樹是所有集群方法中最短的(故又稱短臂法)，此有利於發現不連續性，但在生物研究上較少使用；全連法(complete link)又稱遠鄰法，以最遠的個體來判斷相似性，最不容易造成連鎖，故形成的樹最長(故又稱長臂法)，可產生最緊湊的群，在生物學上最好解釋(Lepš and Šmilauer 1999)。均連法(average link)以群間個體的平均距離判斷，而 Ward 最小變方法則以群內變方最小為原則，這兩種方法的距離都介於單連法和全連法之間，由於是以群間相似性而非個體間距離來判斷，故較不受離群值影響；均連法適用於計算各種相異性係數或相似性係數之場合，但 Ward 最小變方法則不行。均連法的名稱並不一致，以非加權成對分群法(unweighted pair-group method using the average approach, UPGMA)為最常見版本，常應用於種原鑑定和育種材料遺傳關係研究(Warburton and Crossa 2002, Mohammadi and Prasanna 2003)。單連法、全連法、Ward 最小變方法和 UPGMA 是研究上最常用的集群方法，同一套資料集各以這些方法所形成的樹狀圖，比較如 Fig. 5。統計上，有一種共表型相關係數(cophenetic correlation coefficient, CCPC)可作為集群分析的配適性(goodness of fit)統計量，它是測量集群結果之共表型距離矩陣和原來相異(或相似)矩陣間的相關性，用來比較各種集群方法所形成的樹並衡量其失真度(Romesburg 1984)： CCPC 值愈接近 1，表示集群結果愈能準確地反映出原來資料結構；CCPC<0.8 表示配適性差，值愈低，表示樹愈是扭曲嚴重。一般而言，UPGMA 的 CCPC 值最高，適用於各種相似性係數且較不易受離群值影響，故往往被優先推薦(Romesburg 1984)。但因 UPGMA 的基本假設在實際場合很少完全符合，採用它作為主要方法之同時，最好也能考慮嘗試另外第二種方法（如單連法或全連法），檢視其結果後再作最後定奪，以免喪失某些資訊(Romesburg 1984, Mohammadi and Prasanna 2003)。根據經驗，資料轉換對分類結果的影響遠大於集群方法之選擇（Lepš and Šmilauer 1999）。進行集群分析時，直接以觀測值所求出的距離矩陣進行集群分析，與先將資料轉換成相似矩陣後再進行集群分析，兩者之間的結果往往會有很大的差異。但並非所有場合都能直接使用距離矩陣，如分子標誌的條帶資料，由於資料僅是有和無之二元型式，資料矩陣必須先標準化，即必須先求得某種相似係數後，才能進行集群分析。

二、集群分析的結果決策包含三個主要部份：

(1)集群數的決定－由於集群分析整個分析結果是以樹狀圖呈現，此結構極似生物系統樹。凡樣本間的相似性愈高時，則愈早相連，相連後可形成數小群，小群與小群間、或小群與尚未相連的樣本間再依據相似性的高低，逐步相連直到最後成為單一群體為止。一旦樹狀圖完成後，各樣本間之關係便

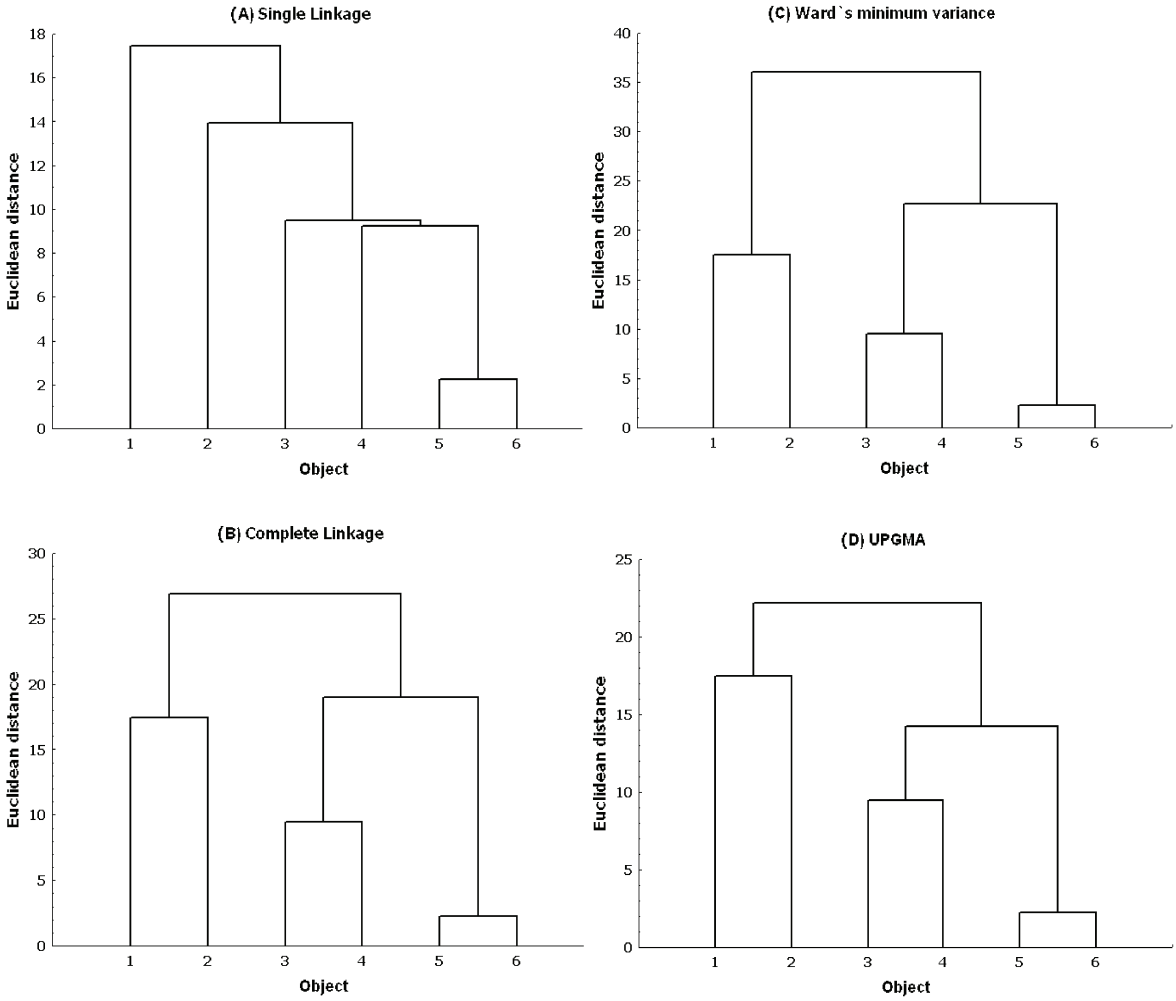

Fig. 5. Dendrograms produced from the same data matrix by using the (A) single link, (B) complete link, (C) Ward's minimum variance, and (D) UPGMA clustering methods.

一目了然。至於要區分為幾群，通常以樹狀圖的分枝狀況來決定，但由於可在樹狀圖任何層級上被切分，因此對任何型態的集群分析來說，尚無一客觀的標準程序可供遵循。實際上集群數之決定比因素數之決定更困難，倘以詳細研究集群為目的(即僅對資料做摘要分析)，而不是找出真的集群，則通常可根據集群形成過程中的各種準則統計量之綜合結果，或直接就由樹狀分群圖來約略判斷，不過在選用不同距離和集群方法時，便會得到不同的分類結果，故必須一再強調，目前雖有一些判斷統計量被提出，但基本上都只能作為選項參考而已，不能作為唯一準則，個人的專業判斷極為重要。有時候，單從集群分析的樹狀圖難以決定出適當的分群數時，同時利用其他多變數分類技術（如主成分分析、因素分析、對應分析）所繪出的散佈圖來輔助判斷(Fig. 3)，可以收到不錯的效果。

(2)集群的解釋－一旦經由集群分析而找出集群後，應設法來描述這些集群，常用的一種方法是分析各群的重心，即利用群內各個體在各變數上的平均值來描述該集群（若輸入資料是區間尺度且是在原始變數

的空間上進行集群，此方法是可行的，倘各變數是先經因素分析的運算，則須將它還原成原始變數，再來描述各集群），除此亦可計算各群的變異情形，如各群內各點間的平均距離或各點重心間之平均距離，來輔助描述該集群。

(3)集群的驗證－就集群結果對一般族群的代表性如何加以驗證，俾使集群的結果可以適用到其他事物，最直接的方法是對同一族群不同樣本進行集群分析，比較其結果並估計其一致性，但若限於時間或成本的限制，或找不到所需的事物供多次集群分析之用，可將樣本分成兩群，分別作集群分析後再比較其結果。

例如調查 14 個數量性狀對 38 種葫蘆進行集群分析(Dey *et al*. 2006) ，選用 UPGMA 法分析的樹狀圖(Fig. 6)，可看出在相似性係數 0.40 上所有品種被分為兩大群，其中第一群的 DBTG-9 與群內其他 23 個品種差異較大，而第二群中的 DBTG-201 和 DBTG-202 則與其他群內品種差異較大。若在相似性係數 0.35 上所有品種被分為三群，其中 DBTG-201 和 DBTG-202 自成一群。

判別分析

判別分析是在已知的分類之下(如草本類和木本類兩個類別)，選出具有代表性的樣本(如牽牛花、劍蘭代表草本類植物，而玫瑰代表木本類植物)，然後由這些樣本的屬性中找出一套最有效的判別函數，這個(些)函數可用來執行分類的工作，一旦遇到有新的樣本時， 可以利用此法選定一判別標準，以判定該新樣本應歸屬於哪個類群。判別分析法的用途很多，如動植物分類、醫學疾病診斷、社區種類劃分、氣象區劃分、土壤類型分類、產品等級分類、職業依能力分類、人類考古學之年代或人種分類等。集群分析與判別分析都是用來處理數值分類問題，其間的差異在於集群分析是不存在一個事前分類的情況下進行資料結構的分類，而判別分析則是已知當前研究對象的分類狀況下，建立適當的判別標準後，將某些未知個體正確地歸屬於其中某一類。集群與判別往往在一個問題上要連續運用，如先進行集群分析，再進行判別分析，就可以進行樣本的識別。

一般電腦軟體進行判別分析的主要流程，首先計算組內各變數的平均值、總平均值、離差矩陣、共變方矩陣等統計值，然後求出判別函數作為綜合判別指標，再計算各組判別係數及判別效果的檢驗統計量，以判斷待判樣本屬於何群並計算後驗機率。判別的規則，最直覺的觀念是求得各群各一個判別函數，將待判樣本帶入各群函數，以函數值最大者代表此一觀測值所被分配到的群別。但判別分析法發展至今，已產生出各種判別函數與規則，如 Fisher 線性判別、距離判別、Bayes 判別、逐步判別等(Shen 1998, Yuan and Zhou 2003)。我們可以將待判樣本的觀測值(不管新值或舊值)帶入各群所屬的判別函數：使用新的樣本觀測值，可以判斷該樣本的歸屬群別；若是使用原經驗樣本的舊值帶入各群所屬的判別函數，則可以針對判別結果進行驗證以確認其分群的正確度。分群的正確度，簡稱判別率，為判別正確之個數除以所有測試樣本個數。

例如為了區分小麥品種的兩種分蘗類型，從第一類(主莖型)取 11 個樣本，第二類(分蘗型)取 12 個樣本，作為經驗樣本，用 3 個指標(X_1, X_2, X_3)求其線性判別函數（Yuan and Zhou 2003）。若採 Fisher 線性判別分析，其判別函數為：

$$Y = 9.283X_1 - 1.020X_2 + 0.983X_3 \qquad [3]$$

將原經驗樣本帶入上述判別函數 Y，得第一類的函數平均值為 12.7358，及第二類的函數平均值為 19.0105，故得臨界值為

$$\bar{Y} = (11\times12.7358+12\times19.0105)/(11+12)$$
$$= 16.0096 \qquad [4]$$

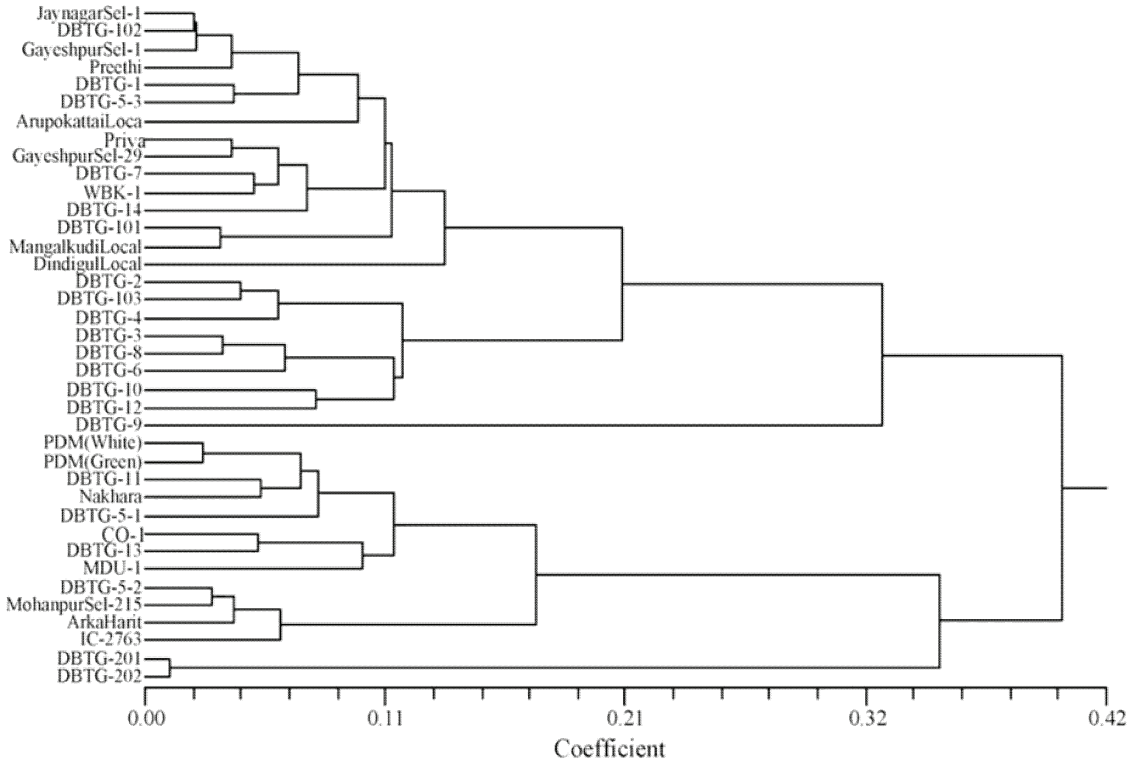

Fig. 6. Genetic relationships among the 38 genotypes of bitter gourd based on 14 quantitative traits by using UPGMA cluster analysis of the distance matrix. (adapted from Dey *et al.* 2006).

若某樣本帶入上述判別函數的值大於該臨界值 16.0096，則判別為第二類，否則屬第一類。為判斷該判別函數的判別率，將原 23 個經驗樣本分別帶入所估計的判別函數（[3]式）中，發現除了第一類第 11 個樣本被誤判為第二類外，其餘均符合，即判別率為 22/23=96%。對於一個給定的新小麥品種樣本 X_1=1, X_2=3.43, X_3=16.25，計算得到 Y =21.7582，故應歸於第二類之分蘗型。同樣資料，若採用距離判別法，則以距離判別函數值ω≧0 時將樣本歸於第一類，ω<0 為第二類，結果判別率同樣是 96%，而新樣本(X_1=1, X_2=3.43, X_3=16.25)得ω=-2.1921<0，亦判別該品種為分蘗型。

路徑分析

一連串的分析變數多半依時間順序先後發生，先發生者視為解釋變數，後發生者視為反應變數，而路徑分析就是在探討分析變數間之單向影響關係，找出變數之間的路徑係數，並畫出路徑分析圖。藉由路徑圖，研究者能清楚了解變數間之影響途徑(箭頭方向)及影響方向(正向、負向、模糊等)，利用這樣的因果模式來幫助說明假設中的因果關係。我們常以兩變數之簡單(直線)相關係數來衡量其相關程度，但此相關係數並無法說明變數間的因果關係，例如口香糖的銷售量與犯罪率之間有正相關， 但在未做路徑分析之前，不可斷言口香糖銷售量高是犯罪率高的「原因」。路徑分析法最早由遺傳學者 Swall Wright 於 1921 年所提出，主要用來解釋人類基因間的因果關係(Wright 1921)；1925 年

他將路徑分析首次應用於經濟學上，用來分析玉米及毛豬的價格(Wright 1925)。然後再被後人擴大應用至其他各領域。必須特別注意的是，路徑分析法雖屬相關關係的研究，但仍須小心下結論，除非證據十分明確，不可輕易下因果關係的結論。因果模式只是用來幫助「說明」假設中的因果關係，而非用來「證實」這種因果關係。路徑分析法的一個貢獻是鼓勵研究者在進行研究之前，作理智的預測而非毫無方向漫無目標的摸索，研究者也必須在不斷研究的過程中，不斷修正其因果模式直到能正確說明該現象為止(Li 1974, Shen 1998)。

路徑分析是迴歸模型的一種延伸，其計算流程相當簡單。但首先在執行電腦程式之前，研究者必須視其研究對象和目標，擬定可能的路徑圖架構，然後依據研究者所擬定的路徑，以變數間的相關係數作為資料，進行迴歸分析，計算所得的各路徑之迴歸係數，即路徑係數。從顯著的路徑係數估計值之正負和大小，可以判斷影響作用是正或負、以及它的影響程度。

例如 Chang *et al*. (2004)以路徑分析探討春作台南白玉米族群的生育前期植株性狀透過什麼途徑影響後期的果穗性狀和子實性狀，由於該報告所探討的性狀甚多，為有利於路徑分析的應用和結果解讀，本文僅取其中的 3 個植株性狀(株高、莖徑、葉數)、3 個果穗性狀(穗長、穗徑、穗重)和子實重來說明。首先假設各農藝性狀間可能的因果關係呈直線相加形態且為單向模式，這 7 個變數間可能的因果關係如 Fig. 7 所示：3×3 條路徑連結 3 個植株性狀與 3 個果穗性狀表現， 3 條路徑連結 3 個植株性狀與子實重表現，3 條路徑連結 3 個果穗性狀與子實重表現。根據此理論模型，以株高、莖徑、葉數之植株性狀影響穗長為例，其因果關係的線性模式為：

$$EL=b_0+b_1PH+b_2SD+b_3LN+e \qquad [5]$$

式中 EL 為穗長，PH 為株高，SD 為莖徑，LN 為葉數，e 為誤差項。其他影響路徑的線性模式同理類推。依據本例指向各個反應變數的影響路徑，由路徑係數的顯著性結果

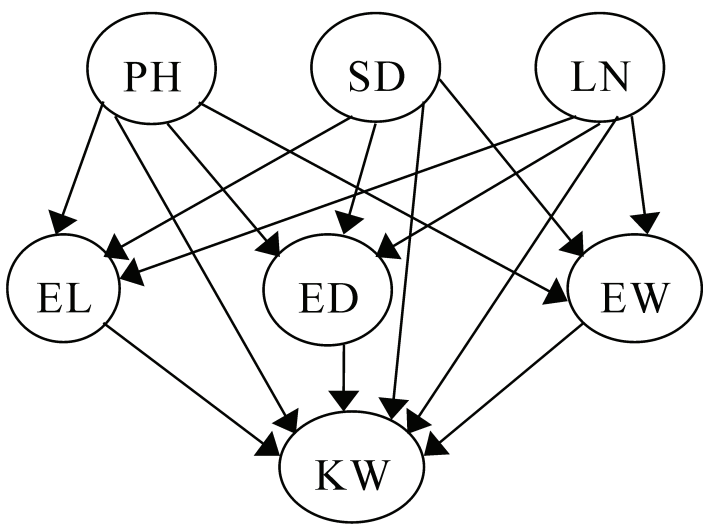

Fig. 7. Path diagram adopted in the study of the kernel and ear characters affected by the agronomic characters in the early growing stages of "Tainan-white" maize population grown in spring season. PH, plant height ; SD, stalk diameter; LN, leaf number; EL, ear length; ED, ear diameter; EW, ear weight; and KW, kernel weight. (adapted from Chang *et al.* 2004).

(Table 3)來看，指向穗長之路徑分析只有莖徑對它有影響，穗徑和穗重都受莖徑和葉數影響，子實重只受穗長和穗重影響。另一方面除了兩個植株性狀影響果穗性狀外，還有一些不知來源的誤差項也可能影響。進而可將顯著的路徑係數標示於路徑分析圖中，則更有利於因果關係的解讀，如 Fig. 8 所示，莖徑和葉數只對時間最近的果穗性狀有直接影響而對時間相隔愈遠的變數之影響力似乎愈來愈小，即植株性狀似乎沒有透過果穗性狀對子實重造成間接影響。穗長和穗重對子實重之決定性佔重要地位，尤其是穗重；株高對果穗性狀無任何影響，而穗徑對子實重影響也不大。

Table 3. Path analysis for each kernel and ear character affected by the characters in the early growing stages in "Tainan-white" maize population (adapted from Chang *et al.* 2004).

	Model			
Character[x]	EL	ED	EW	KW
PH	0.02	0.01	0.13	-0.01
SD	0.32**	0.25*	0.28**	0.003
LN	0.11	0.28**	0.30**	0.02
EL				-0.07 **
ED				-0.01
EW				1.05 **
e	0.90	0.88	0.85	0.17
F	**	**	**	**

[x] PH, plant height ; SD, stalk diameter; LN, leaf number; EL, ear length; ED, ear diameter; EW, ear weight; KW, kernel weight; e , residual factor; and F, F-test of regression model.

*, **: Significant at 5 and 1% levels, respectively.

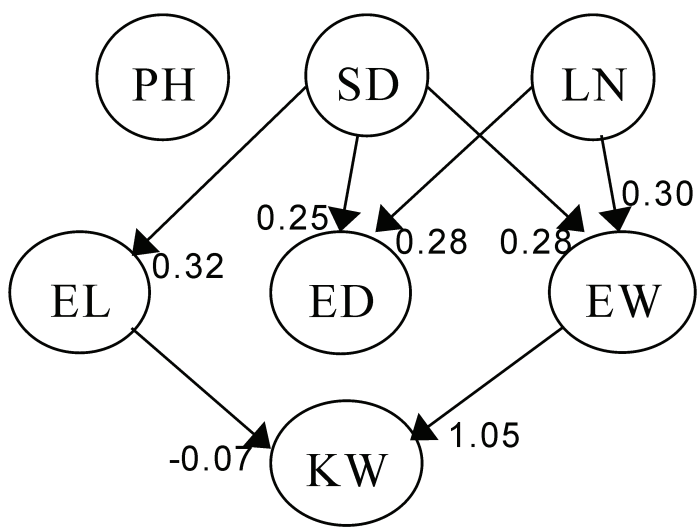

Fig. 8. Results of path analysis (only list the significant paths) on the kernel and ear characters affected by the agronomic characters in the early growing stages of "Tainan-white" maize population grown in spring season. PH, plant height ; SD, stalk diameter; LN, leaf number; EL, ear length; ED, ear diameter; EW, ear weight; and KW, kernel weight. (adapted from Chang *et al.* 2004).

典型相關

我們對一組變數綜合結果和另一組變數綜合結果間的關係感到興趣，且想從其中一組變數來預測另一組變數，例如作物一組生長特性和一組氣象因素間的關係、作物一組產量性狀和一組品質性狀間的關係、某家禽的一組生長性狀和一組生蛋性狀間的關係、農業產銷研究中一組價格指標和一組生產指標間的關係等。在農業科技研究上，我們常需要瞭解生物群與其環境間的關係、育種目標性狀與選拔性狀間的關係等，故不少實際問題可歸結為典型相關研究。為探討兩組變數(反應變數 Y 和解釋變數 X)間的關係，找出 X 的線性組合與 Y 的線性組合，以使這兩個線性組合之間具有最大的簡單相關關係。而能使這兩組變數的線性組合相關最大的權重，稱為典型相關係數。因此 Tatsuoka (1988) 將典型相關視為一種「雙管的主成份分析」。簡單相關、複相關和典型相關之間的差異，如 Fig. 9 所示。典型相關分析除了可以反映出兩組變數之間相互關係的絕大部分訊息，也能揭示兩組變數之間的內部關係。

一般電腦軟體進行典型相關的主要流程，是由變數間的相關矩陣，分別導出 X 和 Y 的兩個線性組合(此即典型變數)，使該兩個典型變數的共變方最大，以計算出典型相關係數及進行其顯著性測驗，最後計算重疊指數(redundancy index，有如複迴歸分析中的決定係數 R^2)，以衡量典型相關所能解釋的變異程度。

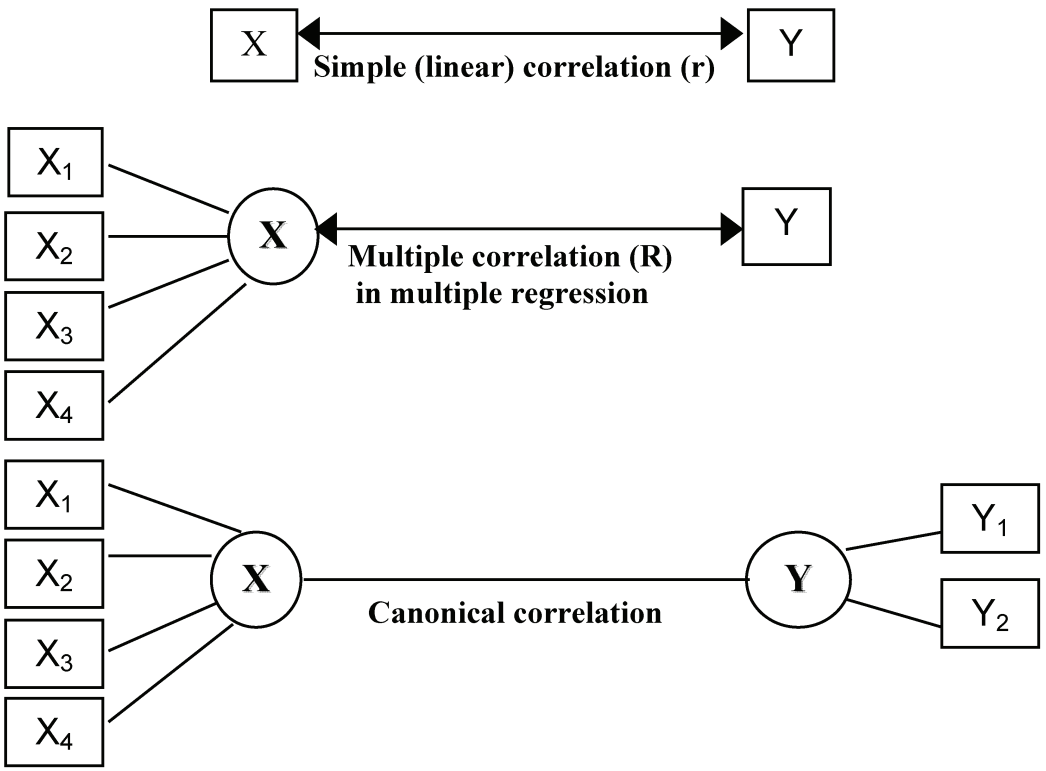

Fig. 9. Diagram showing the differences of definitions among simple (linear) correlation, multiple correlation and canonical correlation.

例如 Hao *et al*. (2006)對不同遺傳基礎的基因轉殖抗蟲棉品種(系)的產量和品質兩組性狀進行了典型相關分析，產量性狀為單株結蒴數(X_1)、單蒴重(X_2)、棉絨比例(X_3)、棉絨產量 (X_4) ，品質性狀為纖維的長度(Y_1)、強度(Y_2)、馬克隆值(micronarie, Y_3, 纖維細度和成熟的量度)、伸長率(Y_4)。分析結果顯示(Table 4)，只有第一個典型相關係數達顯著水準，其相關訊息佔兩組性狀間總相關訊息的 90.23%，表示產量性狀和品質性狀間存在一定的相關；由兩組性狀間顯著的典型相關係數所對應的各對典型相關變數的構成，可知 U_1 以 X_3 的係數最大，而 V_1 以 Y_1 和 Y_3 的係數較大且都是負值，說明了該兩組性狀間的相關顯著主要是由棉絨比例與纖維長度、馬克隆值之間密切關係所引起的。在提高基因轉殖抗蟲棉的棉絨比例同時降低細度時，應注意協調纖維長度，以防棉絨相對變短。

結論

本文純粹以實用角度出發，強調方法之正確選用與分析結果之正確解讀，配合實例說明，針對一些農業研究上常用的多變數分析技術加以介紹。若能清楚瞭解並正確掌握這些多變數分析的應用場合和結果解釋，必能有利於農業試驗資料的分析品質及資訊獲取。綜言之，主成份分析與因素分析的主要目的在縮減原始變數為少數幾個綜合指標，基本上兩者並不相同，一般人常將其混淆。對應分析是延續主成份的概念所發展出的方法，但一般多變數統計書著墨並不多，其實它對生物研究的用途甚廣，值得加以重視；對應分析的主要優點是可將各變數和樣本的主成份值標在同一座標圖（雙標圖）上，以作為最後結果的表達呈現，能用來解釋變數、樣本以及兩者之間的關係。透過主成份分析、因素分析、對應分析的散佈圖，我們可以根據資料點群的距離遠近，來對樣本或變數進行大致的分類，惟主成份分析通常扮演的角色多是一種手段多於目的，即它本身常不是研究的最後目的，而是作為其他分析的預備工作。集群分析與判別分析都是用來處理數值分類問題，其間差異在於集群分析是不存在一個事前分類的情況下以樹狀圖進行資料結構的分類，而判別分析則是在已知當前研究對象的分類狀況下建立判別函數將未知個體做分類歸屬的判斷。路徑分析是在探討分析變數間之單向影響關係，找出變數間之路徑係數並繪製路徑分析圖，以解釋變數間的因果關係。典型相關分析則可反映出兩組變數之間相互關係，以及同時揭示兩組變數之間的內部關係。至於這些方法的計算細節和理論背景，還必須再參閱相關的統計書籍。

Table 4. Results of canonical correlation analysis between yield characters and fiber quality characters[x] of transgenic cotton (adapted from Hao *et al*. 2006).

Canonical correlation coefficient	Redundancy index	Formation of the paired canonical correlation variables for significant correlation coefficient
0.8073*	0.9023	$U_1 = 0.0303X_1 - 0.2696X_2 + 0.9563X_3 - 0.1073X_4$ $V_1 = -0.8061Y_1 + 0.2223Y_2 - 0.4224Y_3 + 0.3498Y_4$
0.7441	—	
0.4728	—	
0.0166	—	

[x] X_1, boll number per plant; X_2, boll weight; X_3, lint percentage; X_4, lint yield; Y_1, fiber length; Y_2, fiber strength; Y_3, micronarie; and Y_4, fiber elongation.

*: Significant at 5% level; —: Value ignored in original paper.

引用文獻

Beh EJ (2004) A Bibliography of the Theory and Application of Correspondence Analysis. Vol. II-By Publication. School of Quantitative Methods and Mathematical Sciences, Univ. Western Sydney, Australia. 101pp.

Benzécri JP (1973) L'analyse des données. II. L'analyse des correspondances. Dunod, Paris, France. 619pp.

Chang SH, HY Lu, C Chen (2004) Path analysis of the yields affected by agronomic characters in Tainan-white maize population under different crop seasons. (in Chinese) **Crop Envir. Bioinformatics** 1:121-138.

Dey SS, AK Singh, D Chandel, TJ Behera (2006) Genetic diversity of bitter gourd (*Momordica charantia* L.) genotypes revealed by RAPD markers and agronomic traits. **Sci. Hort.** 109:21-28.

Hao D, LC He, SD Liu, JF Zhou, XYCong (2006) Canonical correlation and principal component analysis of quantity character of transgenic cotton. (in Chinese) **J. Jinling Inst. Techol.** 22(1):75-78.

Gao W, HP Wu, BC Shia, YT Cheo, CC Su (2003) Biomedical Statistics. (in Chinese) p.323-325. Tingmao Publish Company, Taipei, Taiwan.

Gupta SK, TC Ghosh (2001) Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. **Gene** 273:63-70.

Hill MO (1974) Correspondence analysis: a neglected multivariate method. **Appl. Stat.** 23:340-54.

Kenkel NC, DA Derksen, AG Thomas, PR Waston (2002) Multivariate analysis in weed science research. **Weed Sci.** 50:281-292.

Lepš J, P Šmilauer (1999) Multivariate Analysis of Ecological Data. Faculty of Biological Science. Univ. South Bohemia, České Budějovice. 110pp.

Lin CC (1974) Path Analysis – A Primer. Pacific Grove, California, USA. 346pp.

Liu Q, Y Feng, X Zhao, H Dong, Q Xue (2004a) Synonymous codon usage bias in *Oryza sativa*. **Plant Sci.** 167:101-105.

Liu Q, Y Feng, Q Xue (2004b) Analysis of factors shaping codon usage in the mitochondrion genome of *Oryza sativa*. **Mitochondrion** 4:313-320.

Manley BFJ (2004) Multivariate Statistical Methods: A Primer. 3rd. Chapman & Hall, London, UK. 214pp.

Mohammadi, SA, BM Prasanna (2003) Analysis of genetic diversity in crop plants – salient statistical tools and considerations. **Crop Sci.** 43:1235-1248.

Nishisato S (1980) Analysis of Categorical Data: Dual Scaling and Its Applications. Univ. Toronto, Toronto, Canada. 276pp.

Oksanen J (2004) Multivariate Analysis in Ecology-Lecture Notes. p.63-70. Department of Biology, Univ. Oulu, Oulu, Finnish.

Romesburg HC (1984) Cluster Analysis for Researchers. Wadsworth, Inc., California. 334pp.

Shen ML (1998) Applied Multivariate Analysis. (in Chinese) p.243-301. Jeou Chou Book Co. Ltd., Taipei, Taiwan.

Tan Q, K Brusgaard, TA Kruse, E Oakeley, *et al.* (2004) Correspondence analysis of microarray time-course data in case-control design. **J. Biomed. Informatics** 37:358-365.

Tatsuoka MM (1988) Multivariate analysis: Techniques for Educational and Psychological Research. 2nd. Macmillan Pub. Co., New York, USA. 477pp.

Tekaia F, E Yeramian, B Dujon (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends : a global picture with correspondence analysis. **Gene** 297:51-60.

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. **Ecology** 67:1167-1179.

ter Braak CJF (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. **Vegetatio**

69:69–77.

Warburton M, J Crossa (2002) Data Analysis in the CIMMYT Applied Biotechnology Center – for Fingerprinting and Genetic Diversity Studies. 2nd. International Maize and Wheat Improvement Center (CIMMYT). 29pp.

Wright S (1921) Correlation and causation. **J. Agric. Res.** 20: 557-585.

Wright S (1925) Corn and Hog Correlations. Department of Agriculture Bulletin No. 1300, US. 60pp.

Yang SZ, HZ Chen, CY Li, ZD Sun (2006) Genetic variation, correlation and principal component analyses on major agronomic characters of cassava. (in Chinese) **Chinese Agric. Sci. Bull.** 22(7):232-234.

Yuan ZF, JY Zhou (2003) Multivariate Statistical Analysis. (in Chinese) p.188-195. 2nd. Science Press, Beijing, China. 303pp.

－編輯：楊純明