知識天地

矩陣視覺化一個例子：中研院有大小所嗎？

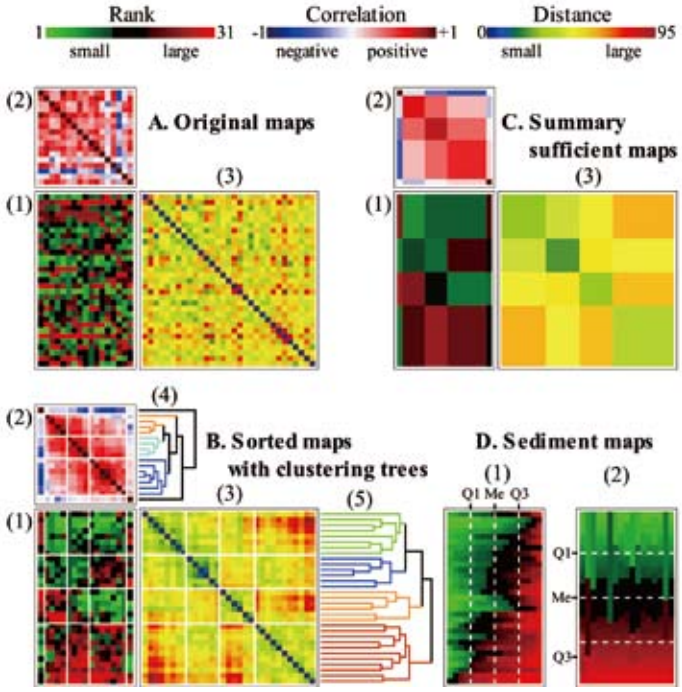
陳君厚副研究員(統計科學研究所)

中研院有大小所嗎？當然有。翻開員工名錄與預算書可看到哪些所預算數大或小、人員多或少；如果你蒐集了許多變數(variable)，則需要多變量統計方法去分析資料。筆者在此介紹一套"看"高維度資料的方法: 矩陣視覺化(matrix visualization: MV)。為了介紹MV，我們以本院31個所(處)中心為樣本蒐集20個變數(表一：17數值變數、3共變數(covariate))；資料之蒐集以公開及方便性為主。讀者對這20個變數的選擇當然有所疑慮--約聘僱人員與院外計畫等變數未納入、某些變數可能資料時間太短(如前瞻計畫)、某些變數可能應使用相對數值(如年輕著作獎)、人事變數比例是否過高等。筆者強調此資料之蒐集以方法介紹為主，非以資料分析為目的。我們將原始資料(人數、件數，千元，年分)做了排序：序1表最小值，序31為最大值，同值取平均序。進入分析的是一個31列(單位)乘17行(變數)的序(rank)矩陣，矩陣中第(*i,j*)數字表第*i*單位在第*j*變數之序，介於1~31。

我們以筆者團隊開發的廣義相關圖(generalized association plots: GAP[1-3])介紹MV基本概念； GAP分析含三個矩陣：(1)資料矩陣(31*17)；(2)變數關係矩陣(proximity matrix)--選用相關係數矩陣 (17*17)；(3)樣本關係矩陣--選用歐氏距離(Euclidean distance)矩陣(31*31)。圖一整理GAP之MV主要步驟:

A.矩陣圖之呈現

以圖一的三個色譜(1綠-紅，2藍-紅，3彩虹)，將三個數值矩陣轉換成三個矩陣圖，(圖一A)。圖一A1(31*17)在第(*i,j*)位置的一個紅(綠)點表示第*i*單位在第*j*變數之序高(低)於*j*變數之中位數(median)，明度越亮(暗)表示序越接近極端值(中位數)；圖一A2(17*17)在第(*i,j*)位置的一個紅(藍)點表示第*i*與*j*二變數為正(負)相關，顏色越深(淺)表相關程度越強(弱)。圖一A3(31*31)在第(*i,j*)位置的一個紅(藍、黃)點表示第*i*與*j*二樣本點間距離遠(近，中等)。通常資料的原排列依據研究者對資料了解與偏好而定，此處我們將資料矩陣的列與行各做了隨機排列。



圖一：GAP矩陣視覺化四個主要步驟

B.矩陣圖之排序

我們希望尋求"較佳"的排列，以呈現資料結構；筆者提出統計繪圖中與排序相關的一個概念--統計圖之相對性(relativity of a statistical graph) [1]--就是要將相似(不同)樣本點或變數置放在圖中靠近(距離遠)的位置，以呈現資料之幾何關係。常見的一種排序法為階層叢聚樹(hierarchical clustering tree)法--以關係矩陣建構一棵叢聚樹再以樹之端葉(terminal node)相對位置對矩陣排序。圖一B使用了GAP的改良樹型排序法[2]針對(圖一A2,A3)建構了兩棵叢聚樹(圖一B4、B5)，並將隨機排序的(圖一A)排成資料結構與型樣(pattern)清楚呈現的(圖一B)。

C.矩陣圖之切割與摘要充分圖

接著我們將排序後的(圖一B1,B2,B3)依據(圖一B4,B5)的樹型將31個單位分成了4個單位群(綠、藍、橘、紅)，並將17個變數分成了3組變數群(橘、青、藍)與2個個別變數。如此的切割將原來大小為(31*17, 17*17, 31*31)的三個資料點矩陣轉換成大小為(4*5, 5*5, 4*4) 的三個資料區塊矩陣。將(圖一B1,B2,B3)每一區塊以該區塊之代表值(在此使用中位數)取代即得(圖一C1,C2,C3)之摘要充分圖(summary sufficient graph)[1]，它可以簡要的表現出潛藏於三矩陣中之重要結構與資料訊息。

D.沉澱矩陣圖

若將排序過之圖一B1每一列(單位)作橫向沉澱則呈現圖一D1之樣本沉澱圖以觀察各樣本之整體表現(由上而下可看出整體序較小(綠)或大(紅)之關係)；若對圖一B1每一行(變數)作縱向沉澱則呈現圖一D2之變數沉澱圖以表現各

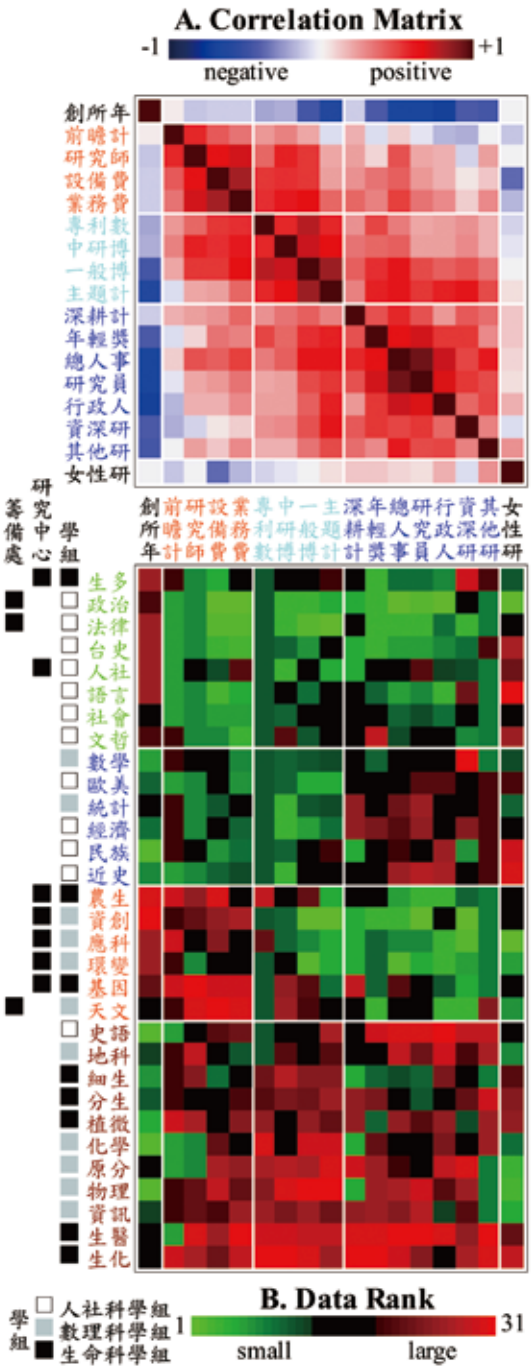
變數在所有樣本之分布狀況。圖一D1,D2之功能等同對31個單位及對17個變數作比鄰箱型圖(side-by-side box-plots；Q_1：1st quartile, Me：medion, Q_3：3rd quartile)。

E.條件矩陣圖

圖一使用(綠-紅)色譜對全資料矩陣(31*17)上色，稱為矩陣條件圖(matrix condition map)；若變數間尺度(scale)差異大，大尺度變數將占用全色域而掩蓋掉小尺度變數之解析度，此時必須將色譜之色域套用至個別變數(單位)以呈現行(列)條件圖(row/column condition map)，觀察個別變數(單位)之結構。本資料無此問題因為所有變數(序)之全域皆為(1~31)。

看甚麼與怎麼看

從圖一ABCD已經可以"看"出基本資料結構，我們將圖一B1,B2放大成(圖二B,A以便進行深入觀察；以同仁們對本院各單位與變數的熟悉度應該很容易看出端倪；筆者摘要出幾點注意項目:



圖二：中研院各所資料之矩陣視覺化

- 圖二B中31個所(處)中心分成4個單位群(綠、藍、橘、紅)，而這4個單位群在5個變數群(創所年、橘、青、藍、女研究員比)上互有消(綠)長(紅)。
- 圖二AB中橘變數群含(前瞻計、研究師、設備費、業務費)，簡稱設備群；青變數群含(專利數、中研博、一般博、主題計)，簡稱主題群；藍變數群含(深耕計、年輕獎、總人事、研究員、行政人、資深研、其他研)，簡稱人事群。主對角線上的深紅色區塊表示3個變數群群內有高度正相關；群與群間(非對角線)則呈現不同程度之正相關；創所年(歷史越久年份越小)與3個變數群(設備、主題、人事)呈現不同程度(低、中、高)之負(藍色)相關；女性研究員比例與設備費有相當程度之負相關。
- 圖二B中綠單位群含7個人文組所(處)中心加上生多中心，這些單位在3個變數群皆為序相對小之單位，相對年輕且女性研究員比例相對較高(政治所例外)；藍單位群含4個人文所加上數學與統計所，特色為歷史久、設備群及主題群低但人事群較高，女性研究員比例不低；橘單位群含自然及生醫學組共5個研究中心加上天文所籌備處，特性為年輕(皆為中心或籌備處)，設備群高，主題群中等，人事變數低(合聘之研究員歸原聘單位)，女性研究員比例亦低；紅單位群含自然及生醫學組各5個所加上史語所，特色為在3個變數群皆相對高、創所年份早、女性研究員比例在生醫(自然)學組5個所高(低)。
- 從3個共變數(學組、中心、籌備處)可以獲得以下資訊: 綠、藍單位群以人文組為主，而橘、紅單位群以自然及生醫組為主，例外的是生多、數學、統計、及史語所；綠、橘單位群各含數個中心與籌備處，而藍、紅單位群則完全為正式所；籌備處屬相對較小單位但天文所例外。
- 圖一C之摘要充分圖呈現的是各區塊的平均趨勢，另一個值得注意的現象反而是區塊中的外來值(outlier)，亦即是不合群點；例如生多中心的資深研究員比及文哲所之年輕著作獎在該區塊(綠單位群/藍人事群)中相當高；依此模式可以在各區塊中發現許多不尋常之型樣。
- 以上筆者整理出幾項重點型樣與特徵，讀者若仔細觀察圖一、二，將可再發現些有趣現象；例如圖一B3排序後之距離矩陣圖中距離最小(藍色點)兩個單位為統計與經濟所，此二所雖分屬自然與人文組，在研究性質與單位特性上卻有相當多共通處。本資料與每位讀者皆切身相關，讀者可從所屬單位立場觀察到一些現象；大部分應該與常理相符，值得深究的反而是異常者，當然有可能是資料登錄或使用有誤。立場不同讀者對變數選擇有不同需

求，對資料呈現也會有相異之解讀；筆者再次強調此資料之MV呈現是以方法介紹為主，資料分析與解釋則是另一層次的議題。

結語

中研院有大小所嗎? 針對這一份特定資料，答案是明確且肯定的；但重要的是MV可以同時呈現資料之單位群，變數組，單位群與變數組間互動關係，及其他圖法或分析不易察覺之現象。經過MV之探索式資料分析(exploratory data analysis: EDA[4])，應該對資料結構有相當了解，可以較精準的對"中研院有大小所嗎?"這一命題提出較明確合理之統計假設，再以數理統計與計算方式進行下一步的確認式統計分析(confirmatory statistical analysis)。

GAP可以在個人電腦上進行上萬個樣本點與變數之MV，我們已將GAP應用到多個生物醫學領域: 精神醫學[5-8]、癌症醫學[9-11]、流行病學[12]、與中草藥研究[13-15]。GAP在人文社會與自然科學研究上應該也有不錯的潛能，但尚未有機會開發。有興趣的同仁請至：http://gap.stat.sinica.edu.tw/Software/GAP/ 下載軟體使用。本文介紹之模組是以分析連續(continuous)資料為主，我們也開發了分析二元(binary)資料、類別(categorical/nominal)資料、以及與地圖學(cartography)資料相連結之模組，將來有機會再與大家分享。最後感謝諸位讀者之耐心，並感謝公共事務組吳春蓉，統計所賴姿秀、林芳華、張倫境、高君豪等同仁提供或整理相關資料。

表一：變數之組成

一、人事變數	**資料來源：員工名錄**（中華民國97年12月）
1. 研究員	研究員人數(含特聘、正、副、助研究員)
2. 資深研	資深(特聘及正)研究員比例
3. 女性研	女性研究員比例
4. 研究師	研究技師人數(含正、副、助研究技師)
5. 其他研	其他研究人員數
6. 行政人	行政人員數
7. 總人事	總人員數
二、計畫變數	**資料來源：學術諮詢總會網頁**（歷年）
8. 主題計	主題計畫件數
9. 深耕計	深耕計畫件數
10. 前瞻計	前瞻計畫件數
11. 中研博	中研院博士後人數
12. 一般博	一般博士後人數
三、其他變數	**資料來源：學術諮詢總會網頁及公共事務組**（歷年）
13. 年輕獎	年輕著作獎人數
14. 專利數	專利獲證件數
四、預算變數	**資料來源：會計室網頁本院98年度法定預算**
15. 業務費	業務費
16. 設備費	設備及投資
五、單位別變數	**資料來源：中央研究院2008年簡介**
17. 創所年	單位設立年分(越新的單位數字越大)
18. 學組	(數理、生命、人文): 不進入相關係數與距離之計算
19. 研究中心	是否爲研究中心: 同上
20. 籌備處	是否爲籌備處: 同上

參考文獻

1. Chen, (2002) *Statistica Sinica*, 12, 7-29.
2. Tien YJ et al. (2008) *BMC Bioinformatics*, 9:155.
3. Wu HM et al. (2010) *Computational Statistics and Data Analysis*, 54 (3), 767-778.
4. Tukey JW, (1977) *Exploratory Data Analysis*, Addison-Wesley.
5. Lin AS et al. (1998) *Psychiatry Research*, 77, 121-130.
6. Hwu HG et al. (2002) *Schizophrenia Research*, 56, 105.
7. Yeh LL et al. (2008) *Journal of the Formosan Medical Association*, 107 (8), 644-652.
8. Lin SH et al., (2009) *Genes, Brain and Behavior*, 8(8), 785-294.
9. Chen JJ et al. (2005) *Journal of Clinical Oncology*, 23, 1-12.
10. Sher YP et al. (2006) *Cancer Research*, 66, 11763-11770.
11. Chen HY et al. (2007) *The New England Journal of Medicine*, 356:11-20.
12. Lee YS et al. (2005) *BMC Genomics*, 6:132.
13. Wang CY et al. (2008) *BMC Genomics*, 9:479.
14. Chien SC et al. (2009) *Phytochemistry*, 70 (10), 1246-1254.
15. Hou CC et al. (2010) *Journal of Nutritional Biochemistry*.