[研究新領域報導]

全矩陣式資料視覺化與資訊探索

中央研究院統計科學研究所　陳君厚

一、EDA

John W. Tukey 在探索式資料分析 (Exploratory Data Analysis, EDA [1]) 書中開宗明義地提到：

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

學習你可以做什麼，有助於在資料分析的過程中達到事半功倍的效果。EDA 的作用在於從「看」資料獲得資料所傳達的訊息，所著重的是簡單的算術與容易建構的圖、表。透過 EDA 對於圖表中所顯露之型樣 (pattern) 做一初步的認知與描述，再進一步以人類的心智 (mind) 對所接收的訊息做全面的分析與判斷，以探索潛藏於資料中的訊息。強調的是探索式的分析而非嚴謹的模式確認。

EDA 的發展，三十年來雖有許多的學者參與研發開拓新領域，卻沒有眞正突破性的工作出現。箱型圖 (box-and-whisker plot) 仍然是敘述性統計 (descriptive statistics) 中最重要的工具。資料視覺化 (data visualization) 研究相當高的比例都投注在維度縮減 (dimension reduction) 相關的工作上 [2,3,4]。這一類的技術在資料量不大，尤其是變數不多時，對於資料結構的探索扮演了重要的角色。然而目前常見的各種統計分析資料檔，變數少則上百、多則成千；維度縮減技術，在視覺化的資料結構探索工作上已不敷使用。

二、全矩陣式資料視覺化

本文所介紹的全矩陣式資料視覺化技術，是一個非降維的 EDA 視覺化方法。其可分析資料維度大小僅受限於電腦顯示器或印表機等輸出尺寸。基本上全矩陣式資料視覺化並不是一項新的技術 [5,6]，只是新一代電腦的計算、記憶與顯示能力賦予它一個新的舞台。

全矩陣式視覺化的基本概念是很簡單的，就是將完整資料矩陣 (data matrix) 或關係矩陣 (proximity matrix) 有效的呈現於電腦顯示器或報表紙上。文中將引用 Hwu et al. [7] 分析的部分活性與負性症狀 (Positive and Negative Syndrome Scale, PANSS [8]) 精神醫學量表資料作一說明。原文中有 163 位精神分裂症患者 (schizophrenic patient)在 33 項 PANSS 症狀上的得分，本文只選用 40 位患者與 17 項症狀作爲範例。PANSS 症狀原爲 1 分 (正常) 至 7 分 (嚴重) 的順序尺度(ordinal scale)，在此假設爲連續變項以利說明。Chen [9]將全矩陣視覺化推廣爲四個主要部分，在此就以這四個部分對全矩陣視覺化作一基本介紹。

2.1 原始資料之呈現與關係矩陣之選擇

第一個工作是將數字矩陣轉化爲一個矩陣圖，基本上是以一個彩色點代表一個數字，每一個列向量 (row vector)，即一個病患的症狀組合，轉化爲一個橫向色帶，而每一個行向量 (column vector)，是一個症狀的病患分布，轉換成一個縱向色帶。整個數字矩陣的所有訊息就完全表現於一個矩陣圖內 (見封底裡圖一(a))。

2.1.1 色譜 (color spectrum) 與變數轉換 (variable transformation)

色譜的選擇是全矩陣視覺化相當重要的一個工作，例子中的 PANSS 資料是一個固定尺度 (1~7)的量表，只需找到一個色譜能表現其順序 (ordinal)特性即可。在此使用的是彩虹色譜，當然任何單向性漸進色階如灰階亦可勝任。變數的特性若爲雙向性(bi-directional)，例如微陣列基因表現(microarray gene expression)的對數

(logarithm)資料，就以二組漸進色階表示(見封底裡圖二) 。

當變數結構較複雜時，需要經過變數變換才能有效的以視覺化呈現資料結構。當變數有不同尺度(scale)時可能必須將變數標準化(standardization) 或常態化(normalization)才能夠進行變數間的視覺化比較。當資料出現離群值(outlier)時，離群值與主群體間的距離會擠壓色譜，此時視覺化呈現的僅有離群值與主群體的相對關係，主群體內的結構無法顯現，必須對變數進行對數（或類似）轉換以淡化離群值之效果。一般之變數變換是以變數(症狀)爲主體，這是所謂的行條件(column condition)變數變換。若著重個體(病患)之間的比較時，則必須進行列條件(row condition)轉換，必要時當然也可以用矩陣條件(matrix condition)轉換。

2.1.2 關係矩陣之選擇

第二個重要的工作是針對變數 (症狀) 與個體 (病患) 選擇適當的關係矩陣 (proximity matrix)，這是爲下一步的排序所做的準備動作。關係矩陣的選擇要能夠適切的表現變數間的交互作用以及個體間的關係，關係矩陣求得之後也必需選擇適當色譜以呈現關係結構。例中病患間之關係以歐氏距離 (Euclidean distance)表示，故選擇單向性的灰階色譜 (見封底裡圖一(c));症狀間之關係採用相關係數 (correlation coefficient) 並以雙向性的藍—白—紅色譜配合之(見封底裡圖一(b))。

倘若資料的分佈並不均勻或出現離群值，關係計算往往會受到與眾不同的變數或個體的嚴重影響。計算個體關係時，當某些變項與其它變項非常不同時，個體的關係很可能完全取決於這些變項。反之，計算變數間之關係時，離群值將嚴重扭曲變數間的關係，而無法代表大部分的個體之變數間的關係。因此，恰當的資料轉換是計算關係矩陣的關鍵。

2.2 關係矩陣與資料矩陣之排序

封底裡圖一(a)中的資料矩陣與關係矩陣雖以視覺化呈現卻顯得雜亂無章，原因是矩陣中變數與個體之排序 (ordering) 是隨機的。任何一張統計圖 (包含全矩陣視覺化) 要能夠表現潛藏的資料結構，必須要將特徵一致 (不同) 的個體或變項置放於相近 (相遠) 的位置。Chen [9]稱這個概念爲統計圖的相對性 (relativity of a statistical graph)，在全矩陣視覺化中就是要對二個關係矩陣找最佳的排序。

2.2.1 Robinson Matrix

想要最佳化(optimization)當然就得先定出標準(criteria)，一個矩陣是否排得好有一個常用的準則，那就是 Robinson 條件[10]。一個矩陣若從主對角線(main diagonal)往上、下、左、右四個方向移動都是(單調)遞減，則此矩陣稱爲一(嚴格)Robinson 矩陣(見封底裡圖三(a))。一個矩陣若經過排序得以成爲 Robinson 矩陣，則稱此矩陣爲準-Robinson 矩陣(pre-Robinson matrix)(見封底裡圖三(b)、(c))。一個排列過的關係矩陣若接近 Robinson 條件表示類似(不同)的變項或個體已被排在相近 (相遠) 的位置上。基本上 Robinson 的原則較注重全域性(global sense)的考量，亦即是矩陣中任二行(列)之關係皆須納入計算。事實上視覺化強調的正是全域性的原則，因爲人類視覺接受與腦部處理的都是圖中完全的資訊。然而此最佳化問題由於計算複雜度過高，並沒有實用的演算法可以套用，Chen [9]以相關係數 (Pearson correlation coefficient) 矩陣收斂的特性提出一個橢圓排序(elliptical seriation)的方法，對於找尋近-Robinson (near-Robinson) 排序有不錯之效果。

2.2.2 樹形排序 (tree seriation)

由於全域性的排序法不易開發，常見的演算法都是以區域性 (local) 最佳化爲主要訴求，其中使用最廣泛的是具有樹狀結構 (tree architecture, dendrogram) 的階層式集群分析(hierarchical cluster analysis)。其排序是以終結點（葉） (terminal node, leaf) 之相對位置產生。封底裡圖一中的變數（症狀）相關係數矩陣與個體（患者）間之距離矩陣分別產生了封底裡圖四中的變數與個體樹狀結構，二個矩陣圖也以相對應之樹狀結構重排序，最後再將資料矩陣圖經過二維（列與行）排序成封底裡圖四中之資料矩陣圖。比較封底裡圖一與圖四可以發現封底裡圖四已將相似之症狀與患者排在相鄰位置，而症狀與患者之群組與相對應之關係也

一目了然。封底裡圖四(a)之排序後資料矩陣圖相當於一個濃縮的十七個變數間一百三十六張之散布圖矩陣 (scatter-plot matrix)，二張排序後關係矩陣圖則替代了因素分析 (factor analysis) 和群集分析 (cluster analysis) 之功能。

然而，以樹狀結構對矩陣排序有一基本問題，也成爲一研究課題 [11,12]。在一個已完成的樹狀結構中，共有 n-1 個節點 (包含根 (root)，不包含葉 (leaf))，因此總共會有 2^{n-1} 種翻轉 (flip) 的可能。封底裡圖五就是以同一個樹狀結構的不同節點翻轉機制 (mechanism) 排序並呈現同一個相關係數矩陣，可以明顯看到其視覺化效果差異相當強烈。在做資料矩陣視覺化呈現時，當然也會產生不同的影響。

2.3 關係矩陣與資料矩陣之分割 (partition)

關係矩陣圖排序後的下一步驟，是直接對變數與個體進行分群的工作，這是一個受限 (constrained) 的群集分析問題。受限的原因是 p 個變數與 n 個個體都已經被排列過，群集之尋求受限爲在排序上找切割點。若是以樹狀結構對矩陣排序，則常以樹形特徵作判斷直接尋找群落或以節點高度對樹形作橫向切割自動定出群落，如封底圖六中症狀（變數）與患者（個體）樹形都以一紫色橫線切割。症狀被分成紅、綠、藍三組，患者則形成青、桃紅、黃、灰四群。

若無樹狀結構，則必須從資料矩陣（圖）及關係矩陣（圖）之數值或圖形（二者是一體兩面）特徵著手。影像分析 (image analysis) 中的邊緣偵測 (edge detection) 技術可以在矩陣圖中找尋切割點，在關係矩陣圖中進行的必須是受限 (constrained) 邊緣偵測，因爲關係矩陣（圖）具有對稱特性。

2.4 充分統計圖 (sufficient statistical graph)

封底圖六中症狀被分成三組而患者形成四群，因此症狀關係圖被切割爲九 (3x3) 塊，患者關係圖被切割爲十六 (4x4) 塊，資料矩陣圖則分割爲十二塊；關係矩陣圖之區塊包含群內與群間二種而資料矩陣圖則形成不同症狀群與患者群之組合。Chen [9] 針對此一區塊分割提出充分統計圖 (sufficient statistical graph) 之概念，目的是以一最精簡之圖示盡可能完整呈現並總結潛藏 (embedded) 在原始資料矩陣與延伸之二個關係矩陣中之訊息。其做法是將每一區塊中之所有數值（原始資料或關係值），以平均數(mean)、中位數(median)、標準差(standard deviation) 或其它適合之統計量取代。封底圖七所呈現的就是封底圖六之區塊以平均數表示之充分統計圖，圖中清楚顯示出三組症狀與四群患者之組（群）內關係強度與組（群）間關係結構，更重要的是資料矩陣之充分統計圖中扼要的總結了四群患者在三組症狀之互動關係。

三、全矩陣式資料視覺化之變通性

全矩陣視覺化具高度之變通性 (flexibility)，針對不同的資料結構與應用需求可以輕易進行改造，在此舉二例說明。

3.1 沈澱圖

一般之資料矩陣視覺化強調保留每一變數與個體之身份 (identity)，封底圖八(a)中每一圖點都是某特定患者在某症狀之分數。在資料分析之過程中，若將患者之身份忽略而在每一個症狀上由小而大排序則產生封底圖八(b)之症狀沈澱圖，此圖之作用類似多變量箱型圖與枝葉圖之綜合體或者是多個單變量之直方圖 (histogram) 或柵欄圖 (bar chart)，可用以觀察每一症狀之分數分布 (distribution)。反之，若將症狀之身份忽略則產生封底圖八(c)之患者沈澱圖，用以觀察每一患者不區分症狀之嚴重度。

3.2 分段式 (sectional) 矩陣視覺化

在排序過之關係矩陣圖方面，一個簡易卻有效的變化是分段式矩陣視覺化。其作用在於每次只呈現符合特定條件之部分關係值。圖九中每一張症狀相關係數矩陣圖僅呈現 t-檢定之 p-value 小於某特定值之相關係數（假設常態與獨立之條件成立），p-value 愈小之圖中只保留愈顯著之相關係數與愈強之症狀群組。

四、結論與可能發展方向

全矩陣式視覺化雖然不完全是一個新的研究領域，卻仍是一個尚待開發且頗具潛力的礦場，存在許多的研究課題與應用技術，值得有

興趣之學者同仁共同努力開發。當然此領域之研究工具除了常用之數理與統計方法外，增加了一項電腦繪圖之技巧（介面），不熟悉此等工具之學者可以找已進入此領域或資訊相關之研究同仁合作。一般統計分析要處理的資料都是全矩陣視覺化的可能對象，以下介紹數個較有趣且與其它統計研究領域相關之全矩陣式視覺化可能發展課題，作爲本文之總結。

4.1 類別型 (categorical) 資料之全矩陣視覺化

當資料型態爲類別性，尤其是名目(nominal)資料時，前述之全矩陣視覺化出現二個困難：(一)資料矩陣圖色譜之決定－連續型資料可以很容易使用單向或雙向漸進色階呈現資料結構，名目型資料則需經過某種尺度化 (scaling)轉換再上色。性別資料當然可以（紅）男（綠）女著色，政黨取向也可以國（藍）、民（綠）、親（橘）上色，但是將此二變項並列時，除非民進黨員等同女性，將產生視覺上之衝突。(二)關係矩陣之計算不易－名目型資料求算變數或個體之關係矩陣公式皆不多，大多是列聯表形式，可以採尺度化轉換或對數線性模式(log-linear model)等類別性資料統計方法求算。

4.2 多時點（相同變項）資料之全矩陣視覺化

範例中之 PANSS 量表資料爲患者入院時測量，資料也有可能在其它時間點取得，如出院及追蹤資料。如何將多時點資料以單張全矩陣方式呈現或多張並列以同時探索患者、症狀與時間三個因素之交互結構是一個相當具挑戰性的問題，多時點的縱深式 (longitudinal) 統計模型亦可能有所貢獻。

4.3 多條件（不同變項）資料之全矩陣視覺化

在 Hwu et al. [7] 文中不只使用 PANSS 量表，如何將多個量表的資料以全矩陣視覺化呈現患者、症狀與量表三個因素之結構是一個類似多時點資料之問題。其共通處是二者皆有相同的個體，但是多時點資料每個時點測量相同的一套變數（量表）而多條件資料每個條件下測量不同的變數群，正交相關 (canonical correlation) 類之統計理論可能可以派上用場。

4.4 條件式（變項校正）全矩陣視覺化

全矩陣視覺化與其它統計分析模型都可能有需要做變數校正 (covariate adjustment) 之場合，例如性別，年齡對模式選擇之影響。若該欲校正變數爲類別型如性別，則可能將變數之關係矩陣分解成群內 (within group) 與群間 (between group) 二關係矩陣之線性組合以探討該變數對其它變項之全矩陣視覺化影響。

4.5 相依 (dependent) 或群集 (clustered) 資料之全矩陣視覺化

當資料存在相依結構時，例如以家庭爲單位蒐集之社會調查或遺傳統計資料，資料之全矩陣視覺化亦出現二個困難：（一）關係矩陣之計算不易－此處之關係矩陣存在二個層次，即群集（家庭）間與群集內。群集間之關係如何計算？是否保留群集內之結構？都不是容易解決的問題。（二）資料矩陣圖與關係矩陣圖如何呈現－即使關係矩陣之計算解決，由於資料存在群集關係，最後之全矩陣視覺化仍不易表現。一般之相依資料統計理論亦有用武之處。

4.6 巨量資料之全矩陣視覺化

在一般的電腦顯示器上可以輕易呈現 1000 個變項與 1000 個個體之全矩陣視覺化；1000 個變項以一般的資料處理而言不算少，但是 1000 個樣本點卻不算多。處理上萬甚至百萬筆資料時，計算速度、記憶容量與圖檔顯示皆造成電腦負荷。此時抽樣 (sampling) 方法、序貫 (sequential) 分析、修勻 (smoothing) 技術與影像 (image) 處理等多種理論與方法皆可派上用場。

4.7 全矩陣視覺化之遺漏值(missing value)處理

資料出現遺漏值時可能可以用全矩陣視覺化之特性處理，排序後之資料矩陣圖已將類似的變數（症狀）與個體（患者）放在相鄰的行或列上，遺漏值可以用變數與個體二維的鄰近點進行估計。具遺漏值資料之關係矩陣求算也要考慮，其方法與 EM 演算法 [13] 精神類似。

參考文獻

[1] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley (1977).

[2] A. Buja and D. Asimov, *Proceedings of the 18th Symposium on the Interface*, American Statistical Association, 63 (1986).

[3] P. J. Huber, *The Annals of Statistics*, **13**, 435 (1985).

[4] K. C. Li, *Journal of The American Statistical Association*, **86**, 316 (1991).

[5] J. A. Hartigan, *Journal of the American Statistical Association*, **67**, 123 (1972).

[6] R. F. Ling, *Communications of the ACM*, **16**, 355 (1973).

[7] H. G. Hwu, C. H. Chen, T. J. Hwang, C. M. Liu, J. J. Cheng, S. K. Lin, S. K. Liu, C. H. Chen, Y. Y. Chi, C. W. OuYoung, H. N. Lin, and W. J. Chen, *Schizophrenia Research*, **56**, 105 (2002).

[8] S. R. Kay, A. Fiszbein and L. A. Opler, *Schizophr. Bull.*, **13**, 261 (1987).

[9] C. H. Chen, *Statistica Sinica*, **12**, 7 (2002).

[10] W. S. Robinson, *American Antiquity*, **16**, 293 (1951).

[11] B. J. Ziv, K. G. David, and S. J. Tommi, *Bioinformatics*, **17**, 22 (2001).

[12] T. Biedl, B. Brejova, E. D. Demaine, A. M. Hamel and T. Vinar, *Technical Report CS-2001-14*, Dept. of Computer Science, University of Waterloo, (2001).

[13] A. P. Dempster, N. M. Laird and D. B. Rubin, *J. R. Statist. Soc. B*, **39**, 1 (1977).