看高維度資料的方法 : 矩陣視覺化(MV)

一個 31 列(單位)乘 17 行(變數)的序(rank)矩陣，利用廣義相關圖(GAP)分析三個矩陣(四個主要步驟)：

A. 矩陣圖之呈現 (此處資料矩陣的列與行隨機排列)

(1) 資料矩陣(31*17)；(2) 變數關係矩陣 : 選用相關係數矩陣(17*17)；(3) 樣本關係矩陣 : 選用歐氏距離矩陣(31*31)

B. 矩陣圖之排序

統計圖之相對性 : 將相似樣本點或變數置放在圖中靠近的位置，以呈現資料之幾何關係

階層叢聚樹排序法 : 以關係矩陣(圖 A2、A3)建構一棵叢聚樹(圖 B4、B5)，再以樹之端葉相對位置對矩陣排序，

將隨機排序的圖排成資料結構與型樣清楚呈現的圖

C. 矩陣圖之切割與摘要充分圖

依據叢聚樹將 31 個單位分成了 4 個單位群、17 個變數分成了 3 組變數群與 2 個個別變數

原來大小為(31*17, 17*17, 31*31)轉換成大小為(4*5, 5*5, 4*4)

每一區塊以該區塊之代表值(在此使用中位數)取代即得摘要充分圖 C

摘要充分圖 :1. 簡要的表現出潛藏於三矩陣中之重要結構與資料訊息

2. 呈現的是各區塊的平均趨勢，依此模式可以在各區塊中發現許多不尋常之型樣

D. 沉澱矩陣圖

排序過之圖 B : 每一列(單位)作橫向沉澱，呈現樣本沉澱圖，觀察各樣本之整體表現

: 每一行(變數)作縱向沉澱，呈現變數沉澱圖，表現各變數在所有樣本之分布狀況

: 圖 D 功能等同對 31 個單位及對 17 個變數作比鄰箱型圖(Q1，Median，Q3)

E. 條件矩陣圖

對全資料矩陣(31*17)上色，稱為矩陣條件圖

若變數間尺度差異大，大尺度變數將占用全色域而掩蓋掉小尺度變數之解析度

解決 : 將色譜之色域套用至個別變數(單位)以呈現行(列)條件圖，觀察個別變數(單位)之結構

結語 :

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 分析就可以較精準的對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之後的確認式統計分析

全矩陣式資料視覺化與資訊探索 :

(壹) EDA : 1.「看」資料獲得資料所傳達的訊息，強調的是探索式分析而非嚴謹的模式確認

2. 著重在簡單的算術與容易建構的圖、表

3. 先做初步的認知與描述，再進一步以人類的心智對訊息做全面的分析與判斷，以探索潛藏於資料中的訊息

4. 箱型圖 : 敘述性統計中最重要的工具；資料視覺化高比例研究都投注在維度縮減相關的工作上

全矩陣式資料視覺化技術在資料量不大，變數不多時，對於資料結構的探索扮演了重要的角色

(貳) 全矩陣式資料視覺化(非降維的 EDA 視覺化方法) :

1. 原始資料之呈現與關係矩陣之選擇 :

1.1 色譜與變數轉換

(1) 固定尺度的量表 : 表現其順序，任何單向漸進色階亦可勝任

(2) 變數的特性若爲雙向性 : 以二組漸進色階表示，例如微陣列、基因表現的對數資料

(3) 變數結構較複雜時 : 需要經過變數變換

(4) 變數有不同尺度 : 將變數標準化或常態化進行變數間的視覺化比較

(5) 離群值 : 對變數進行對數(或類似)轉換以淡化離群值之效果，離群值與主群體間的距離會擠壓色譜

(6) 行條件(column)變數變換 : 一般之變數變換是以變數爲主體

(7) 列條件(row)轉換 : 著重個體之間的比較時，也可以用矩陣條件轉換

1.2 關係矩陣之選擇 ： 為變數與個體選擇適當的關係矩陣，爲下一步的排序做準備，恰當的資料轉換為計算關鍵

(1) 能夠適切的表現變數問的交互作用以及個體間的關係

(2) 選擇適當色譜以呈現關係結構，例 ： 歐氏距離，故選擇單向性的灰階色譜、症狀間關係用雙向性的藍-白-紅色譜

(3) 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

(4) 計算個體間關係 ： 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

(5) 計算變數間關係 ： 離群值將嚴重扭曲變數間關係，而無法代表大部分的個體之變數間的關係

2. 關係矩陣與資料矩陣之排序 ： 將特徵一致的個體或變項放於相近的位置(統計圖的相對性)，找最佳的排序

2.1 Robinson Matrix : 注重全域性(最佳化問題計算複雜度過高)，矩陣中任二行(列)之關係皆須納入計算

(嚴格)Robinson 矩陣 ： 矩陣從主對角線往上、下、左、右 四個方向移動都是(單調)遞減

準-Robinson 矩陣 ： 矩陣若經過排序得以成爲 Robinson 矩陣

以相關係數(Pearson)矩陣收斂的特性提出翻圜排序(elliptical seriation)的方法，對於找尋 Robinson 排序有不錯效果

2.2 樹形排序 ： 區域性，階層式集群分析，其排序是以終結點(葉)之相對位置產生

排序後關係矩陣圖替代了因素分析(factor)和群集分析之功能

問題 : n-1 個節點(包含根，不包含葉)，總共有 2^(n-1)種翻轉的可能，其視覺化效果差異相當強烈。

3. 關係矩陣與資料矩陣之分割 ： 排序後下一步驟，對變數與個體進行分群

受限的群集分析問題 : p 個變數與 n 個個體都已經被排列過，群集受限爲在排序上找切割點

樹狀結構 ： 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

無樹狀結構 ： 從資料矩陣及關係矩陣之數值或圖形特徵著手

影像分析的邊緣偵測技術可以在矩陣中找尋切割點；關係矩陣必受限邊緣偵測，因爲關係矩陣具有對稱特性

4. 充分統計圖 ： 以一最精簡之圖示盡可能完整呈現並總結潛藏在原始資料矩陣與延伸之二個關係矩陣中之訊息

將每一區塊中之所有數值(原始資料或關係值)，以平均數、中位數、標準差或其它適合之統計量取代

(參) 全矩陣式資料視覺化之變通性(flexibility) : 針對不同的資料結構與應用需求，可以輕易進行改造

1. 沈澱圖 ： 類似多變量箱型圖與枝葉圖之綜合體或多個單變量之直方圖、柵欄圖(bar chart)，用以觀察每一變量分布

2. 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

(肆) 結論與可能發展方向 ：

1. 類別型資料之全矩陣視覺化 ：

二個困難 :(一)資料矩陣圖色譜之決定 ： 名目型資料需經過某種尺度化轉換再上色。二變項並列時將產生視覺上衝突。

(二)關係矩陣計算不易 ： 關係矩陣多是列聯表形式，可採尺度化轉換或對數線性模式等類別性資料統計方法

2. 多時點(相同變項)資料之全矩陣視覺化 ： 多時點的縱深式統計模型可能有所貢獻。

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。

3. 多條件(不同變項)資料之全矩陣視覺化 ： 類似多時點資料之問題。共通處是二者皆有相同的個體。

多時點資料每個時點測量相同的一套變數(量表)。

多條件資料每個條件下測量不同的變數群，正交相關類之統計理論可能可以派上用場。

4. 條件式(變項校正)全矩陣視覺化 ： 將變數之關係矩陣分解成群內與群間，

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。

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

二個困難 :(一)關係矩陣計算不易 ： 存在二個層次(群集間與群集內)，群集間之關係如何計算?是否保留群集內之結構?

(二)資料矩陣圖與關係矩陣圖如何呈現 ： 由於資料存在群集關係，最後的全矩陣視覺化仍不易表現。

6. 巨量資料之全矩陣視覺化 ： 抽樣、序貫(sequential)分析、修勻(smoothing)技術、影像處理

7. 缺失值 ： 用變數與個體二維的鄰近點進行估計；具缺失值資料之關係矩陣求算也要考慮，方法與 EM 演算法類似