知識天地

降維與群聚分析在高雜訊分子影像分析的應用

杜憶萍、陳素雲研究員、陳定立副研究員、謝岱霓女士（統計科學研究所）、章為皓副研究員（化學研究所）

統計科學是一門鼓勵跨界合作的科學。長久以來統計科學的合作對象跨越了士、農、工、商。統計這個英文字 statistics 的字根stat 就是和政府有關的意思。一個用心的政府在研擬重大政策時一定會蒐集相關數據加以分析解讀，作為政策依據。其他方面從我們常看到的科系或課程名稱就可以理解統計與其他領域關係之密切，例如生物統計、工業統計、商業統計。統計科學和某些領域的開創也是息息相關，例如遺傳學。孟德爾（1822-1884）分析了碗豆第一及第二子代性狀數據而提出顯性隱性基因的概念，成為遺傳學始祖。而從Ronald Fisher （1890-1962）同時是一位偉大的遺傳學家也是一位偉大的統計學家，可以看出兩個領域的發展息息相關。上述例子說明統機科學常常伴隨著一個學科領域的開創或革新，這一篇文章介紹以低溫電子顯微鏡（低溫電顯）解巨型蛋白質分子結構和統計科學的關係，就是另一個例子。

目前在蛋白質資料庫裡，有結構資訊的約90%是以X光結晶學解出來的，另外9%是以核磁共振方法，以低溫電顯解結構約佔1%。從數據來看低溫電顯市佔率遠遠被甩在後頭，為何能在這個市場內撐這麼久，沒被淘汰而且情勢看漲呢? X光結晶學以高解析度著稱，核磁共振著重在小而美（更高解析度來解較小分子），低溫電顯以解巨型分子保住其不可取代性。近年在幾個關鍵技術突破後，竟也解出解析度到3A（埃:10^{-10}公尺）的例子，所以有不可輕忽的潛力。

X光結晶學的基本原理是我們中學學過的布拉格定律，用X光通過三維晶體產生繞射圖案回解出原子間的距離而得到結構。原理簡單，真正的門檻在於長出三維晶體。分子要長成晶體需要秘笈（protocol）。秘笈的形成像古時候煉丹術，試遍配方，加溫加鹽加這個那個，更重要的可能是要加一把運氣。而醫藥科學的需求，使得科學家有興趣的分子結構越來越大而複雜，晶體的長成也就越來越艱鉅。另一方面即便晶體長成了，還是有一點缺憾。分子為了長成晶體，可能排斥了其他構型。更有甚者，可能在結晶的過程，取了一個非天然的構型。這些原因使得科學家們對於非晶體態下的結構有興趣，低溫電顯滿足了這個需求，他另有一個名稱:單分子三維結構重建分析，意即不須長晶體。

低溫電顯的做法是將純化的分子（待解結構的蛋白質分子），鑲在很薄的非晶體（玻璃態）冰層裡，再以低劑量的電子束打上去。測量到的是與分子產生交互作用的電子成像。這個影像如同對此分子的密度的投影積分再加上比例很高的雜訊。這些被鑲在冰層的分子沒有一定的方向（圖一）。理論上，只要分子數目夠多，每個方向都可以取得到，要拼出原來三維的樣貌只是時間的問題。實際上，這些蛋白質是很脆弱的，電子束劑量被限制在每平方埃低於10個電子，這使得蒐集到的影像訊噪比（signal to noise ratio）很低（圖二）。這在數據分析上，是一個很大的挑戰。

統計學裡有一個很基本的定理:中央極限定理。大意是說，假設在一個群體中取樣做某個量測（X），此量測存在一個均值（μ）。每個個別量測（X_i）和這個均值的差異稱為誤差是個隨機變量（ε_i）。而這些誤差本身均值為零（$E\varepsilon_i=0$），彼此互不干涉（independent），且對個別量測沒有偏頗（identical），則此樣本量測的平均數和原來的均值的差異可以用（誤差強度/（樣本數）$^{1/2}$）來描述。這暗示著只要在這個群體中，取足夠多樣本數來做量測，這些樣本的平均數可以很接近原來的均值，不管原來誤差強度多強。這個定理在訊噪比很低的低溫電顯影像上的應用是:只要能把影像分門別類，將相同方向（有一樣的均值）的影像取平均數，原貌就有機會呈現（參看圖二）。

取平均數的概念來降低雜訊恢復原貌，其實在X光結晶學裡也不知不覺地用了。 X光經過三維晶體的繞射圖像本身就是一個集體平均的結果。只是這個分門別類取平均值的動作不在數值上，而在樣品預備上:長晶體本身就已

經把個別分子排列整齊了。在低溫電顯中，相當於用統計分析的方法，將個別單一分子做排序， 也就是說用統計分析方法達到晶體有秩序的效果。要將訊噪比很低的影像分門別類需要兩個工具:降維與群聚分析。

降維是降低維度的縮寫。維度降低之後，後續的分析自然較簡單。降維背後有一個信念，有意義的訊號雖然以高維度方式呈現，其實是落在一個相對來說維度小的空間。從數據本身來找有效的低維度空間，一個最被廣泛使用的方法是主成份分析法。它的主要想法是找到一個主軸，使得從那個角度看（意思是把數據投影到那個軸上），這些數據最有份量（變異量最大或著說散得最開）。依此類推，找下一個正交主軸。這個方法在古典力學裡也被使用過，用來算出陀螺的穩定旋轉軸。講了半天，讀者可能會納悶影像分析和高維度空間有甚麼關係呢?

一張黑白影像是由格子般的畫素組合起來的。例如我們聽到的手機廣告強調百萬畫素的照相功能，就是指一張影像由大概是1000x1000的格子畫素組成。每個畫素呈現由黑到白的灰皆是可以被量化的。一張彩色影像則由三張負責紅、藍、綠色階所組成。在影像分析中，每個畫素都被當成一個變量，所以在這個例子裡，變量個數高達百萬。要解一組這種大小的影像的主成份分析，在數學上得解一個百萬乘百萬矩陣的本徵解，計算量非常的大。變量個數之所以如此巨大是因為單單看1000（行）乘以1000（列）的數字，而忽略了它有行與列的架構在。如果從個別的行與列（各自為1000）下手，情況就不一樣了。對於一組（多張）影像要做分析降維，我們的研究團隊說明了對行與列各自找到最佳化的降維空間，在統計上來說是比較有效率的，這個效率涵蓋計算成本及影像重建的準確度。

我們用圖三來解釋，傳統的主成份分析法與行列主成份分析法（多線性主成份分析法）對多張影像降維的差別。傳統的主成份分析法把影像上的畫素拉成一個長長的向量（每張影像成為一個向量）。這時主成份分析法再以解其共變矩陣（此矩陣之行與列的長度都和影像拉長的向量等長，在上述例子為10^6x10^6）的一群本徵向量來當作新的基底矩陣。而後續的分析是以投影到這群新基底的座標值為根據。行列主成份分析法，分別處理行空間與列空間的共變矩陣（在上述例子皆為10^3x10^3）之本徵向量解，而得到行空間與列空間的基底。每個影像左成行基底矩陣，右成列基底矩陣，即為新的投影值，供後續分析。

降維在這裡可說是群聚分析的前序作業。群聚分析的基本概念是，物以類聚。目前最廣為使用的可能算是k-means 算則了。這個算則的想法很直覺，先給定類別數k之後，第一步先隨機把全部分成k類，第二步就從這k類別中，各自找到最具代表性的插上旗幟（一般來說就是取各類別的均值）。接著再回到第一步重分k類（‘k’-step），每一個成員找最接近的旗幟歸類。再到第二步重新取均值標立旗幟（‘mean’-step）。接著不斷的在這兩步（‘k’ and ‘mean’steps）循環直到收斂。這個方法既快速又管用，只要數據夠乾淨以及類別數不要太多。可惜這兩條件低溫電顯的影像都不滿足。k-means 算則其實在第一次第一步的隨機分類就大致決定了最後類別的結構。要有好的結果，第一次隨機分類的各類別均值必須約略在不同的實際類別中。這樣的要求在類別數不大時，重複多試幾次，不難達成。但是當類別數一多時，第一次第一步的隨機分類就有致命性的影響—如果第一次隨機分類的各類均值都離某類別有些距離，這類別的成員將被迫打散分到其它類別而永遠沒機會自成一類。這個現象隨著類別數目越大，情況越嚴重。而典型的低溫電顯影像分類數至少超過100。另一個條件乾淨的數據。前面提過，低溫電顯的訊噪比很低，有些影像根本沒有訊息，不應當被歸入任何一類去做平均。我們稱這種影像為孤群。照理說一個好算則可以自動挑出孤群剔除。但在k-means 算則裡，這類孤群總可以找到最接近的類別標幟，硬闖進去，而影響到之後三維結構重建的解析度。

我們的研究團隊提出一個群聚分析算則: γ-SUP來解決這兩個問題。想像n個小朋友在操場活動時，老師宣布要分組。這時每個小朋友都會開始就近找他熟悉的朋友靠近，不到幾分鐘，人以類聚的現象就會自動產生了。在這個過程中，每個小朋友在踏出下一步時，他們的大腦裡，不斷的做計算。許多條件需要折衝，每一步都是某種加權平均的結果。有少數小朋友很可能徬徨張望但終究不入群，或兩三人組成一個極小的群體。根據這個想法，我們提出了這個 γ -SUP分類算則。我們讓空間中的n個點，各自移動而不必隨機分類，移動的走向及步伐大小透過他與周圍點的距離加權平均來決定。我們發現這個算則恰恰解決低溫電顯影像常有的孤群問題（參看圖二分類

平均結果）。

在這個例子裡，不管是降維或是群聚分析，這些工具的創新都不是從平地起，而是源自於對原來方法（主成份分析法與k-means）的理解。能確實掌握他們的特性， 知道他們的限制，才有機會突破。當然開發新的工具要有新的技術，願意學習新知識也是一重要的關鍵。到目前為止，我們的貢獻（降維與群聚分析）在整個低溫電顯影像的三維結構分析中屬於前端作業。後續的分析包括三維結構模型的建構與估計更是有挑戰性，計算一個結構動輒一兩週的計算量。我們希望，像這個經驗一樣，我們能以統計的專業，幫助合作者提供更有效率地算則分析。我們也希望在這個過程裡能回饋給統計領域，對統計方法有所提升。隨著科技的發展，每一個領域都開啟了新的機會，也各自面臨與以往不同的挑戰，解決問題所需要的工具也需要不斷地創新。某一個困難問題的解決方法（在這個例子裡是長成巨大分子的三維晶體）也許不在它原來的領域（生化領域）裡，而在友軍（統計算則）中。統計所的研究人員樂意成為你們的友軍!

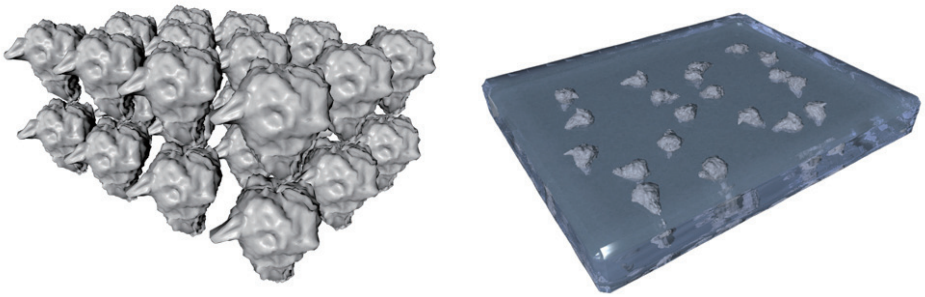
參考文獻

（1）Frank, Joachim “Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State”. (Oxford University Press, New York, 2006）.

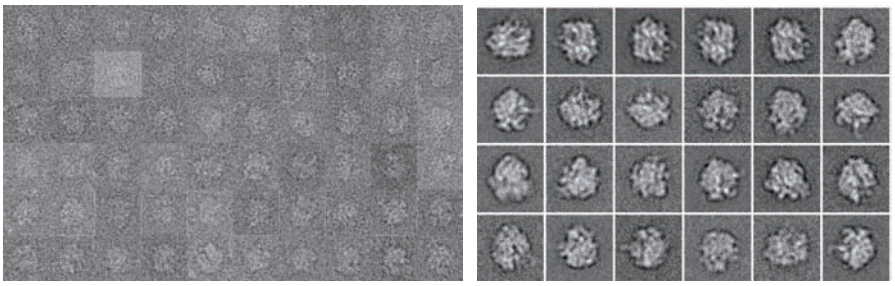
（2）Hung Hung, Pei-Hsien Wu, I-Ping Tu and Su-Yun Huang （2012）. “On Multilinear Principal Component Analysis of Order-Two Tensors”, Biometrika 99（3）: 569-583.

（3）Ting-Li Chen, Dai-Ni Hsieh, Hung Hung, I-Ping Tu, Pei-Shien Wu, Yi-MingWu, Wei-Hau Chang and Su-Yun Huang （2014）. “g-SUP: a clustering algorithm for cryo-electron microscopy images of asymmetric particles”, Annals of Applied Statistics 8（1）: 259-285.

圖一。左側為X光結晶繞射實驗所用的結晶樣品示意圖。右側為低溫電顯所使用的樣品示意圖，其投影影像，需要藉著分類的運算，把一樣方向的分子歸類，排序整齊。



圖二。左圖包含30張低溫電顯單分子（Ribosome）二維投影影像。右圖是經過分類後的平均影像，一些結構樣貌得以呈現出來。



圖三。傳統主成份分析法和行列主成份分析法對一組影像做降維的示意圖。傳統主成份分析法先把每一張影像拉成一個向量，在解基底時往往需要對一個很大矩陣做運算。