Classification with Normal Mixtures: The E-M Algorithm

last modified September 29, 2007

群組分類 (Classification) 的方法可以分成兩種型態: 一種是 deterministic 的, 另一種是 probabilistic。後者運用了機率的假設, 善用關於資料的已知資訊 (prior knowledge) 或資料的分佈情況, 假設了母體的機率密度函數。當資料提供的資訊愈多時, 所做的「假設」會愈正確, 最後的結果也會更準確。

有別於鑑別分析 (Discriminant analysis), 事先知道每個樣本資料所屬的群組, 也不像 Cluster analysis 要將每個樣本資料的類別區分出來。有一種特殊的群組分類問題是要去估計對每個群組所假設的分配, 著名的 E-M algorithm 便是用來解決這方面問題的利器。本單元著眼於 E-M algorithm 的介紹, 看它如何巧妙的處理原本繁瑣的估計問題, 並提供必要的練習, 希望能清楚瞭解並嫻熟的運用這個簡單又有效的演算法。

本章將學到關於程式設計

較複雜演算法的程式設計概念及利用矩陣向量運算的寫作技巧。

⟨本章關於 MATLAB 的指令與語法⟩

指令:

語法:

1 背景

有些群組分析的問題常可以從資料中發現有群組重疊的現象, 並隱約展現出常態的分配。如果常態的假設正確, 對這類的問題, 其焦點往往在這些隱晦不明的常態分配的參數估計上, 而不是找到一個分界線方程式。譬如圖 1 所示的直方圖, 資料來自某種鳥類的翅膀長度量測值。明顯可以看出有兩個群組 (雄性與雌性), 其群組間的分界線也很容易「看」出來,[1] 大約在 89 左右。若單純的僅做群組的判別, 便可以將分界線定爲「數值小於 89 爲群組一 (雌性), 大於或等於 89 爲群組二 (雄性)。」不過從直方圖來看或者從生物體的結構來看, 將同性鳥類翅膀的長度假設爲常態分配應該合適。這時我們感興趣的反倒是兩個群組的翅膀平均長度、變異數與群組比例的估計。這類資料被稱爲來自「混合常態 (Normal mixtures)」的母體。

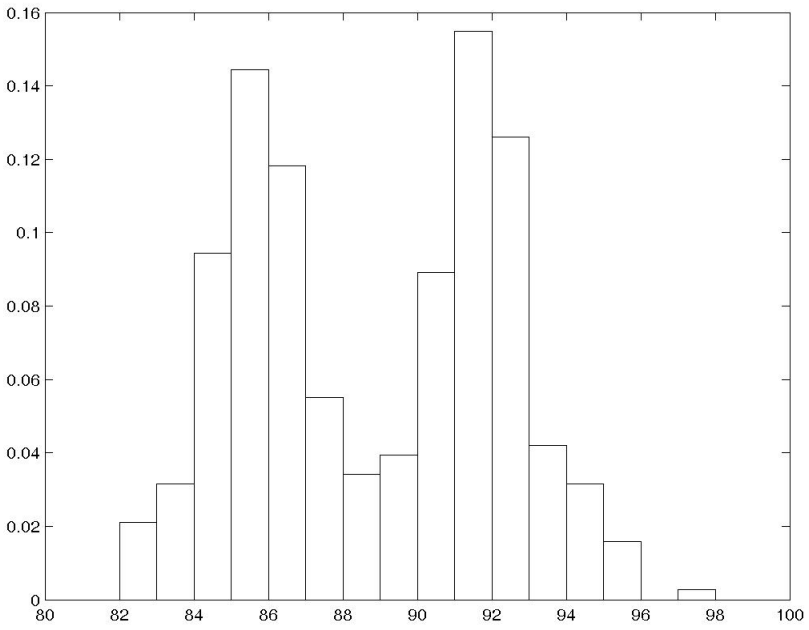

圖 1: 某種鳥類翅膀長度資料的直方圖

[1] 若要實際依數據計算出圖形中所顯示的分界線方程式, 還眞有點「小麻煩,」並非直接可得。

問題: 假設分界線已判定, 我們是否可以直接將隸屬不同群組的資料直接拿來做參數的估計呢? (參考作業 1) 以下將針對混合常態的問題進行剖析, 看看如何從群組重疊的資料中, 直接估計組出群體的參數。

從直方圖目視

1.1 General finite mixtures

在許多的應用裡, 我們想從樣本中探知某個變數 Y 的機率密度函數, 以便對未來樣本做更準確的預測及後續的推論。但往往該變數成因過於複雜或晦暗不明, 只能從一些側面的資訊去輾轉得知。譬如, 變數 Y 代表每年車禍意外的件數, 欲推知其機率密度函數

$$f_Y(y)$$

這個變數的機率密度函數常因太過複雜, 難以直接從樣本資料去推論, 比較容易的方式反而是採用條件機率的模式去迂迴取得。譬如, 假設知道某個組群 j 每年造成車禍的件數爲卜松變數 (Poisson Random Variable), 其參數設定爲 λ_j , 這樣的條件機率密度函數寫成:

$$Pr[Y = y|X = j] = f_j(y) = \frac{\lambda_j^y}{\Gamma(y+1)} e^{-\lambda_j}, y = 0, 1, 2, \cdots \tag{1}$$

另外也假設共有 k 個群組, 各群組發生的機率 (所佔比例) 爲 π_j, 即

$$\begin{aligned} &Pr[X = j] = \pi_j && j = 1, 2, \cdots, k \\ &\textstyle\sum_{j=1}^{k} \pi_j = 1 \end{aligned}$$

基於這些假設與貝式定理的應用, 機率密度函數 $f_Y(y)$ 可以寫成 [2]

$$f_Y(y) = \sum_j Pr[Y = y|X = j]Pr[X = j] = \sum_{j=1}^{k} \pi_j f_j(y) \tag{2}$$

[2]也稱爲 marginal distribution of Y

變數 Y 的機率密度函數是卜松變數的機率密度函數的線性組合, 但這樣的組合的變成非卜松變數了。變數 Y 的機率密度函數可以以這樣的方式來解讀:

在一個實驗裡面, 有 k 個可能發生的事件, 第 j 個事件的結果以密度函數 $f_j(y)$ 來表示, 而每個事件發生的機率爲 π_j。這個實驗的結果以變數 Y 表示, 其機率密度函數可以表示爲如式 (2) 的組合式密度函數。

式 (2) 中的π_j也稱爲組合權重(mixing weights) 或是組合的比例 (mixing proportions)。在群組分類 (classification) 裡也稱爲先驗機率 (prior probabilities), 但當目的是做群組分類時, 通常關心的是後驗機率 (a posteriori probability)

$$Pr[X = j|Y = y] = \frac{Pr[Y = y|X = j]Pr[X = j]}{f_Y(y)} = \frac{\pi_j f_j(y)}{f_Y(y)} \qquad (3)$$

當實驗的結果是 y 時, 上式條件機率密度函數可用來判斷 y 所屬的群組。

1.2 混合常態 (Normal mixtures)

上述藉由一個連續型 (continuous) 與一個數位型 (discrete) 分配組合成的複合型分配,[3]常見於許多的應用問題, 特別當這個連續型的分配是常態時。在式 (2) 中, 假設變數 Y 的條件機率密度函數爲常態分配時, 這類的問題稱爲「混合常態」。許多應用問題符合這個假設, 如圖 1 鳥類翅膀的長度。其中數位型的變數 X 代表「雄性」與「雌性」的群組變數, 而變數 Y 代表翅膀長度, 屬於連續型的變數。

從圖 1 翅膀長度的分配圖看來, 變數 Y 的分配很像典型的混合函數, 假設爲混合常態是合理的, 而從練習 1 模擬的範例中, 也可以看出端倪。混合常態的問題有別於一般的鑑別分析 (discriminant analysis), 主要在於它比較關心對群組所假設的機率密度函

[3]如式(2) 中的機率密度函數 $f_j(y)$ 常爲連續型函數, 而 π_j 視爲多項式分配 (k) 的分項機率。

數的參數估計, 如式 (2) 中的 π_j 與個別常態群組的參數(μ_j, σ_j^2), 不過這類的參數估計一般都很複雜, 還好簡潔有效的 E-M algorithm 正是估計混合常態參數的最典型的方式。

1.3 E-M Algorithm

在參數估計的方式中, 最大概似函數 (Maximum likelihood function) 是最常見用來估計參數的函數。它具備許多優點 [1], 經常搭配運用一階導數的最陡坡法(steepest descent) 或二階導數的牛頓法 (Newton-Raphson method) 等演算法估計參數。不過在某些應用上; 譬如有隱藏變數 (latent variables) 或缺失資料 (missing data) 時, 情況顯得比較複雜, 而 E-M algorithm 舒緩了這個複雜度, 是最大概似函數估計式中一種簡潔但也有效的估計方法, 以下以一個簡單的例子 ([1],Example 4.4.1) 介紹它出場。

範例1: 假設 X 爲一個多項分配(multinomial) 變數, 其分配寫成$M(N, \theta_1, \theta_2, \theta_3)$, 其中 $\theta_1 + \theta_2 + \theta_3 = 1$。想從下列的數據中估計這分項機率值 θ_1, θ_2。

第一次實驗: 得到 $N_1 = Z_1 + Z_2 + Z_3$, 其中 N_1 及 Z_k 分別代表樣本數及各組的數量。
第二次實驗: 得到 $N_2 = Z_1^* + Z_{23}^*$, 其中 Z_1^* 代表第一組的數量, Z_{23}^* 代表第一組以外的數量。

實驗一的概似函數爲:

$$L(\theta) = \begin{pmatrix} N_1 \\ Z_1 Z_2 Z_3 \end{pmatrix} \theta_1^{Z_1} \theta_2^{Z_2} \theta_3^{Z_3}$$

取對數後再去除常數項變爲

$$l_1(\underline{\theta}) = Z_1 \log \theta_1 + Z_2 \log \theta_2 + Z_3 \log \theta_3$$

同理, 實驗二的對數概似函數爲

$$l_2(\underline{\theta}) = Z_1^* \log \theta_1 + Z_{23}^* \log(1 - \theta_1)$$

根據實驗一與二間的獨立性, 其聯合對數概似函數寫爲

$$l^*(\underline{\theta}) = (Z_1 + Z_1^*) \log \theta_1 + Z_{23}^* \log(1 - \theta_1) + Z_2 \log(\theta_2) + Z_3 \log \theta_3 \qquad (4)$$

上式透過一次導數爲零的聯立方程式, 可以計算得到最大概似函數的估計值 (見作業3)。不過這裡我們將利用 E-M algorithm 對於缺失資料的作法, 以數值計算的方式做出相近的估計。在這個問題裡, 缺失的資料爲實驗二的第二群與第三群的數量, 若假設爲 Z_2^* 與 Z_3^*(已知$Z_2^* + Z_3^* = Z_{23}^*$)。上述的對數概似函數可以改寫爲

$$l^*(\underline{\theta}) = (Z_1 + Z_1^*) \log \theta_1 + (Z_2 + Z_2^*) \log \theta_2 + (Z_3 + Z_3^*) \log \theta_3 \qquad (5)$$

這個比較單純的對數概似函數的最大概似估計值爲:

$$\hat{\theta}_i = \frac{Z_i + Z_i^*}{N_1 + N_2} \qquad i = 1, 2, 3 \qquad (6)$$

式 (5) 稱爲完整資料的對數概似函數 (complete data log-likelihood)。明顯的, 這個估計較之原先式 (4) 不考慮缺失資料時簡單許多。而 E-M algorithm 便是利用這項特點, 避繁趨簡, 盡量在簡單的完整資料的最大概似估計上下功夫。

整理一下上面的變數, 將已知資料與缺失資料的變數分別表示出來:

$Y_1 = Z_1 + Z_1^*$	已知的第一群組的數量
$Y_2 = Z_2$	已知的第二群組數量
$Y_3 = Z_3$	已知的第三群組數量
$Y_{23} = Z_{23}^*$	已知的第二群組與第三群組混合數量
X_2	未知的第二群組數量
X_3	未知的第三群組數量

當然$X_2 + X_3 = Y_{23}$。在這些符號下, 完整資料的對數概似函數 (5) 可以改寫成

$$l(\underline{\theta}; X, Y) = Y_1 \log \theta_1 + (Y_2 + X_2) \log \theta_2 + (Y_3 + X_3) \log \theta_3 \tag{7}$$

而已知資料的對數概似函數 (4) 改寫成

$$l^*(\underline{\theta}; Y) = Y_1 \log(\theta_1) + Y_{23} \log(1 - \theta_1) + Y_2 \log(\theta_2) + Y_3 \log(\theta_3) \tag{8}$$

請記得原始的最大概似函數的問題是

$$\max_{\underline{\theta}} l^*(\underline{\theta}; Y) \tag{9}$$

E-M algorithm 並不直接針對已知資料概似函數 (8) 求最大值, 而是選擇去面對比較簡單的完整資料的對數概似函數 (7)。但因爲完整資料的對數概似函數有未知的缺

失資料 X , 無法直接計算, 需以「適當的」值來取代。較正式的說法是, 以兩個步驟迭代遞迴 (iteratively) 的方式, 克服完整資料的對數概似函數中缺失資料不能直接計算的問題。這兩個步驟描述如下: 假設在第 s 次的迭代迴圈裏

E-step(Expectation): 以缺失資料的期望值取代未知的缺失資料變數 X , 相當於以下列的函數取代 (7)

$$l_{s+1}(\underline{\theta}) = E_X \left[l(\underline{\theta}; X) | Y = y; \underline{\theta}^{(s)} \right]$$

M-step(Maximization):

$$\underline{\theta}^{(s+1)} = \max_{\underline{\theta}} l_{s+1}(\underline{\theta})$$

以式 (7) 爲例, E-step 計算
$l_{s+1}(\underline{\theta}) = E_X \left[l(\underline{\theta}; X) | Y = y; \underline{\theta}^{(s)} \right]$

$= Y_1 \log(\theta_1) + (Y_2 + E \left[X_2 | Y = y; \underline{\theta}^{(s)} \right]) \log(\theta_2) + (Y_3 + E \left[X_3 | Y = y; \underline{\theta}^{(s)} \right]) \log(\theta_3)$

換句話說E-step 是將完整資料的對數概似函數中的缺失資料換成其條件式期望值 (在已知 y 與前一次迴圈中估計出來的參數 $\underline{\theta}^{(s)}$ 的情況下)。不同的問題, 期望值的計算難易有別, 以本範例而言, 從 X_2 與 X_3 的組成可以推算兩者的條件(Y_{23} 已知) 機率分配爲,

$$X_2 \sim Binomial(Y_{23}, \frac{\theta_2}{\theta_2 + \theta_3})$$

$$X_3 \sim Binomial(Y_{23}, \frac{\theta_3}{\theta_2 + \theta_3})$$

知道機率分配後, 二項分配的期望值就是

$$E\left[X_2|Y=y;\underline{\theta}\right]=Y_{23}\frac{\theta_2}{\theta_2+\theta_3}$$
$$E\left[X_3|Y=y;\underline{\theta}\right]=Y_{23}\frac{\theta_3}{\theta_2+\theta_3}$$

Y_{23} 爲已知資料變數, 但 θ_2 與 θ_3 未知且是待估計的參數。E-M algorithm 的迭代過程中, 第一輪的參數必須採猜測或其他方式來給定, 第二輪以後的參數, 由上一輪的 M-step 參數估計值取代, 如此循環計算到收斂爲止。[4]請注意, 在 M-step 中的參數估計已經是較簡單的完整資料對數概似函數。

E-M algorithm 這樣的迭代方式直覺上很難被認同會一直朝著原始概似函數的最大值邁進, 最後收斂到一個區域的最高點 (local maximum), 不過理論上確實證明如此[2]。當然, 這個極值不保證是我們期望的極值(global maximum), 可能只是 local maximum。無論如何, 理論歸理論, 還是得實際操作看看, 才知道問題所在。練習二提供實際的數據供寫程式練習, 透過程式的撰寫將有助於對 E-M algorithm 的瞭解。之後, 才能進一步深入較複雜的混合常態的參數估計。

1.4 Normal mixture with E-M Algorithm

E-M Algorithm 是估計混合常態參數最典型的方法, 先從兩個常態的混合說起。

範例2: 假設式 (2) 中的群組變數 X(扮演隱藏變數的角色) 的代表值爲1與0, 分別代表群組一與二, 其發生的機率分別爲 π 與 $1-\pi$ 。假設變數 Y 的條件機率密度函數爲常態且分別爲 $f_1(y)$ 與 $f_2(y)$ 。此時變數 X 與 Y 的聯合機率密度函數寫爲

$$Pr(X,Y)=f_{XY}(x,y)=\left[\pi f_1(y)\right]^x\left[(1-\pi)f_2(y)\right)]^{1-x} \qquad for\ x=0\ 或\ 1$$

問題假設:

[4]收斂的準則通常以估計的參數值不再變化或變化小於某個極小的值爲判斷, 有時也可以兼併判斷目標函數(在此爲對數概似函數值) 的變化。何者爲妥, 得視每個函數在接近極值時的趨勢而定, 沒有一定的標準。

1. 共有 N 個樣本。

2. 假設 $f_1(y)$ 與 $f_2(y)$ 完全已知, 沒有任何常態分配的參數需要估計。

根據已知的 N 個樣本, 欲估計群組 1 發生的機率 π 。

當將隱藏變數納入考慮時, 完整資料的概似函數寫爲

$$\begin{aligned} L(\pi) &= \prod_{i=1}^{N}(\pi f_1(Y_i))^{X_i}((1-\pi)f_2(Y_i)))^{1-X_i} \\ &= \left\{\prod_{i=1}^{N}(f_1(Y_i))^{X_i}(f_2(Y_i))^{1-X_i}\right\}\pi^S(1-\pi)^{N-S} \end{aligned} \tag{10}$$

其中

$$S = \sum_{i=1}^{N} X_i$$

對上式取對數並去除與 π 無關的部分, 完整資料的對數概似函數變爲

$$l(\pi) = \left(\sum_{i=1}^{N} X_i\right)\log\pi + \left(N - \sum_{i=1}^{N} X_i\right)\log(1-\pi) \tag{11}$$

由上式可以很輕易的推算其最大概似函數的估計式

$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{12}$$

這個估計式其實滿符合常理的; 若事先知道每個樣本 y_i 的群組屬性 (即 X_i 已知), 其群組比例的最大概似估計就是屬於該群組的樣本數除以總數, 式 (12) 的意義在此。當樣本群組屬性已知, 其群組的參數估計只是個單純的問題。EM algorithm 解決混合常態的問題, 在於加入群組屬性的變數 X_i , 將之顯性化, 並與群組的參數一同列入待估計的行列。相反的, 若僅採取已知資料概似函數求最大值 (即只考慮變數 Y), 其結果將複雜許多 (作業6)。

利用 E-M algorithm 估計 π 在第 s 次迭代的的兩個步驟分別是
E-step:

$$
\begin{aligned}
l_{s+1}(\pi) &= E_X\left[l(\pi)|Y=y;\pi^{(s)}\right] \\
&= \left(\sum_{i=1}^{N} E\left[X_i|Y=y_i;\pi^{(s)}\right]\right)\log\pi + \left(N-\sum_{i=1}^{N} E\left[X_i|Y=y_i;\pi^{(s)}\right]\right)\log(1-\pi)
\end{aligned} \tag{13}
$$

M-step: 根據式 (12)

$$\max_{\underline{\theta}} l_{s+1}(\pi)$$

$$\pi^{(s+1)} = \frac{1}{N}\sum_{i=1}^{N} E\left[X_i|Y=y_i;\pi^{(s)}\right] \tag{14}$$

其中

$$E\left[X_i|Y=y_i;\underline{\theta}^{(s)}\right] = Pr[X_i=1|Y=y_i;\underline{\theta}^{(s)}] = \frac{\pi^{(s)}f_1(y_i)}{\pi^{(s)}f_1(y_i)+(1-\pi^{(s)})f_2(y_i)}$$

由於 X_i 非1即0, 其條件式期望值相當於群組一的後驗機率。練習 3 提供實際樣本供寫程式執行 E-M algorithm。

爲逐步的瞭解 E-M algorithm 的運作方式, 以上的討論僅假設兩個常態群組的比例爲未知。在比較瞭解 E-M algorithm 的兩個步驟後, 現可以放寬未知的限制到 k 個常態的混合, 其組合比例 π_i、均數 μ_i 與變異數 σ_i^2 都是未知, 共 $3k-1$ 個變數。

1.5 多個常態混合的問題與分析

假設 X 代表群組, 包含有 k 個 multinomial 變數

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

也就是同一個時間 (對每個樣本 $Y = y$) 只有一個 X_j 變數爲 1 其餘都是 0, 且 X_j 爲 1 的機率定義爲

$$\pi_j = Pr(X_j = 1) \qquad\qquad j = 1, 2, \cdots, k$$

若 X_j 代表其中一個事件, π_j 爲其發生的機率。根據以上的描述, 可以寫出 X 與 Y 的聯合機率密度函數

$$f_{XY}(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{k} [\pi_j f_j(\mathbf{y})]^{x_j} \qquad all\ x_j = 0\ or\ 1\ and\ \sum_{j=1}^{k} x_j = 1 \qquad (15)$$

上式看似 k 項乘積, 其實對每個 y 值只有其中的一項存在。從上式可以得出 Y 的機率密度函數

$$\begin{aligned} Pr[Y = y] = f_Y(y) &= \sum_{all\ x} f_{XY}(\mathbf{x}, \mathbf{y}) \\ &= \sum_{all\ x} \prod_{j=1}^{k} [\pi_j f_j(\mathbf{y})]^{x_j} = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y}) \end{aligned} \qquad (16)$$

一般稱爲 finite mixture density。混合常態的問題與群組分析最大的不同, 在於群組分析的每個樣本所屬的群組爲已知, 而混合常態則未知 (比較接近 clustering 的問題), 只知樣本來自一個常態混合的母體, 不知是哪一個常態, 這當然提高了參數估計的困難度。

在進入 E-M algorithm 之前, 先來看看從「已知資料的對數概似函數」求其最大值的參數估計究竟有多麻煩。對於混合常態而言, 變數不少, 先從 π_j 的估計下手, 先寫出對數概似函數

$$l(\underline{\pi}) = \log(\prod_{i=1}^{N} \left[\sum_{j=1}^{k} \pi_j f_j(y_i)\right]) = \sum_{i=1}^{N} \log \left[\sum_{j=1}^{k} \pi_j f_j(y_i)\right] \qquad (17)$$

在先不考慮其他變數的情況下, 針對 π_i 的最大概似函數的問題爲

$$\max_{\pi, \sum \pi_j = 1} l(\pi)$$

利用 Lagrange multiplier 的方式, 可以得到下列的聯立方程式 (作業 9)

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} \pi_{ji}$$

$$where\ \pi_{ji} = \frac{\pi_j f_h(y_i)}{\sum_{h=1}^{k} \pi_h f_h(y_i)} \qquad j = 1, 2, \cdots, k,\ i = 1, 2, \cdots, N \quad (18)$$

這說明第 h 群組的先驗機率, 相當於在 N 個樣本中每個樣本屬於第 h 個群組的後驗機率的平均。上式也可以寫成

$$\sum_{i=1}^{N} \frac{f_h(y_i)}{\sum_{j=1}^{k} \pi_j f_j(y_i)} = N, \qquad h = 1, 2, \cdots, k \quad (19)$$

連同$\sum_{j=1}^{k} \pi_j = 1$, 形成一組非常複雜的聯立方程式。這時通常利用如下的方式反覆求解: 假設從一組起始值開始 $\pi_1^{(0)}$,$\pi_2^{(0)}$,$\cdots$,$\pi_k^{(0)}$, 在第 s 次的迭代迴圈

1. 計算
$$\pi_{ji}^{(s)} = \frac{\pi_j^{(s)} f_j(y_i)}{f(y_i)}, \qquad i = 1, 2, \cdots, N$$
2. 計算
$$\pi_j^{(s+1)} = \frac{1}{N} \sum_{i=1}^{N} \pi_{ji}^{(s)}, \qquad j = 1, 2, \cdots, k$$

即利用前一個迴圈的先驗機率計算後驗機率, 再利用這個計算出來的後驗機率重新計算先驗機率。如此反覆計算直到兩者都不再改變爲止。

接著針對 μ_j 與 σ_j 做其最大概述函數的估計(作業10):

$$\mu_j = \frac{1}{N\pi_j}\sum_{i=1}^{N}\pi_{ji}y_i \qquad j=1,2,\cdots,k \tag{20}$$

$$\sigma_j^2 = \frac{1}{N\pi_j}\sum_{i=1}^{N}\pi_{ji}(y_i-\mu_j)^2 \qquad j=1,2,\cdots,k \tag{21}$$

以上的聯立方程式結合式 (18) 便是所謂的最大概似函數的估計值, 非常的複雜, 不但沒有 closed form solution, 即便利用牛頓法等迭代遞迴的方式, 也是囉唆至極。相對的 E-M algorithm 便顯得可愛許多, E-M algorithm 將隱藏的群組變數考慮進來成爲概似函數的一部份, 這個作法看似複雜, 卻也化解不少糾纏的參數估計麻煩, 以下是考慮隱藏變數的作法:

已知資料的對數概似函數 (17) 中缺少群組的資訊 X , 因爲並不知道每個已知資料的群組。而完整資料的對數概似函數則加入了這個變數, 使其完整, 寫成 (從聯合機率密度函數式 (15) 看起)

$$\begin{aligned} l_c(\theta;X,Y) &= \sum_{i=1}^{N}\sum_{j=1}^{k} x_{ji}\log\left[\pi_j f_j(y_i)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{k}\left[x_{ji}\log\pi_j + x_{ji}\log f_j(y_i)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{k} x_{ji}\log\pi_j + \sum_{i=1}^{N}\sum_{j=1}^{k} x_{ji}\log f_j(y_i) \end{aligned} \tag{22}$$

其中隱藏變數 $x_i = [x_{1i}\ x_{2i}\ \cdots\ x_{ki}]^T$, $\sum_{j=1}^{k} x_{ji} = 1$, 且欲估計的參數表示爲

$$\theta = [\pi_1,\cdots,\pi_{k-1},\mu_1,\cdots,\mu_k,\Psi_1,\cdots,\Psi_k]$$

即將所有待估計的參數放在一個向量中。在此也將多變量的情形考慮進來, 不再限制爲一個變數, 而是多變量混合常態, 且上式加入了 $N\times k$ 個隱藏變數 x_{ji} , 它的值非 1 即 0。再來看看這個完整資料的對數概似函數的估計值如何計算。對個別的變數取一次導數, 得到 (作業11)

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} x_{ji} = \frac{N_j}{N} \tag{23}$$

$$\mu_j = \frac{1}{N_j}\sum_{i=1}^{N} x_{ji}\mathbf{y}_i \tag{24}$$

$$\Psi_j = \frac{1}{N_j}\sum_{i=1}^{N} x_{ji}(\mathbf{y}_i - \mu_j)(\mathbf{y}_i - \mu_j)^T, \qquad j = 1, 2, \cdots, k \tag{25}$$

其中 N_j 代表所有已知資料中屬於群組 j 的數量, 當然此時因爲 x_{ji} 是變數的關係, N_j 也是變數。上式(23) $\sim$ (25) 推導的過程要比之前的簡單許多, 結果也容易解讀, 在 E-M algorithm 裡面是 M-step 的計算式。在 E-step 中需要計算隱藏變數 x_{ji} 在資料 $\mathbf{y}$ 與參數 θ 已知條件下的期望值, 即在第 s 個迭代迴圈的 E-M algorithm 的兩個步驟:

E-step: 計算

$$\begin{aligned} \pi_{ji} = E\left[X_{ji}|\mathbf{y}, \theta\right] &= Pr[X_{ji} = 1|\mathbf{y}, \theta] \\ &= \frac{\pi_j f_j(\mathbf{y}_i; \theta)}{f_Y(\mathbf{y}_i; \theta)}, j = 1, 2, \cdots, k, i = 1, 2, \cdots, N \end{aligned} \tag{26}$$

M-step: 計算 $\theta^{(s+1)}$ 的估計式

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} E\left[X_{ji}|Y, \underline{\theta}\right] = \frac{1}{N}\sum_{i=1}^{N} \pi_{ji} \tag{27}$$

$$\mu_j = \frac{1}{N\pi_j}\sum_{i=1}^{N} \pi_{ji}\mathbf{y}_i \tag{28}$$

$$\Psi_j = \frac{1}{N\pi_j}\sum_{i=1}^{N} \pi_{ji}(\mathbf{y}_i - \mu_j)(\mathbf{y}_i - \mu_j)^T, \qquad h = 1, 2, \cdots, k \tag{29}$$

計算上得按照順序 (27)(28)(29) , 因爲都需借用上個式子的計算結果。如果假設所有群組的共變異矩陣都相等, 其估計式寫爲

$$\Psi = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} \pi_{ji}(\mathbf{y}_i - \mu_j)(\mathbf{y}_i - \mu_j)^T \tag{30}$$

2 練習

1. 觀察混合常態的長相; 假設由兩個常態分配合成, 其合成密度函數爲

$$f_Y(y) = \pi_1 f_1(y; \mu_1, \sigma_1^2) + \pi_2 f_2(y; \mu_2, \sigma_2^2)$$

畫出下列混合常態的密度函數(如圖2所示), 可以利用 MATLAB subplot 的功能在一張圖上畫 6 個圖。(本題摘自參考文獻[1],p.113)

(a) $\pi_1 = 0.5, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.5, \mu_2 = 2, \sigma_2 = 1$

(b) $\pi_1 = 0.25, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.75, \mu_2 = 3, \sigma_2 = 1$

(c) $\pi_1 = 0.8, \mu_1 = 1, \sigma_1 = 1, \pi_2 = 0.2, \mu_2 = 1, \sigma_2 = 4$

(d) $\pi_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.4, \mu_2 = 2, \sigma_2 = 2$

(e) $\pi_1 = 0.9, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.1, \mu_2 = 2.5, \sigma_2 = 0.2$

(f) $\pi_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.4, \mu_2 = 2.5, \sigma_2 = 1$

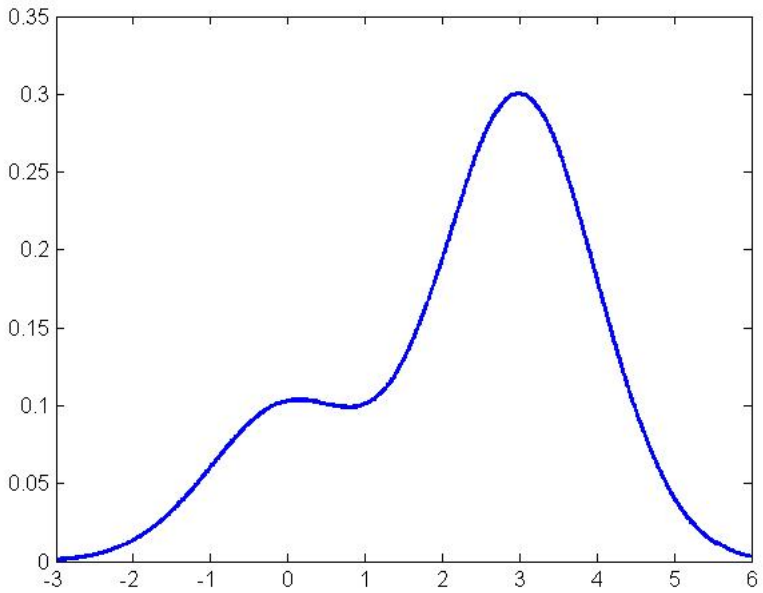

圖 2: 兩個常態混合的密度函數

2. 同範例 1 的描述, 實際的數量爲 $Y_1 = 5, Y_2 = 3, Y_3 = 2, Y_{23} = 10$ 。利用範例中描述的E-M algorithm 寫程式估計參數 θ_1 及 θ_2 , 並畫出概似函數的等高線圖及每個迭代過程中的參數估計值。如圖3所示。(本題摘自參考書目[1],Example 4.4.3) 爲證實E-M 演算法的每個迭代都會朝目標函數的區域最大值推進, 建議將每次迴圈的目標函數值列印出來, 看看是否逐步「升高」。如果不是, 可能程式寫錯了或是觀念不清楚。

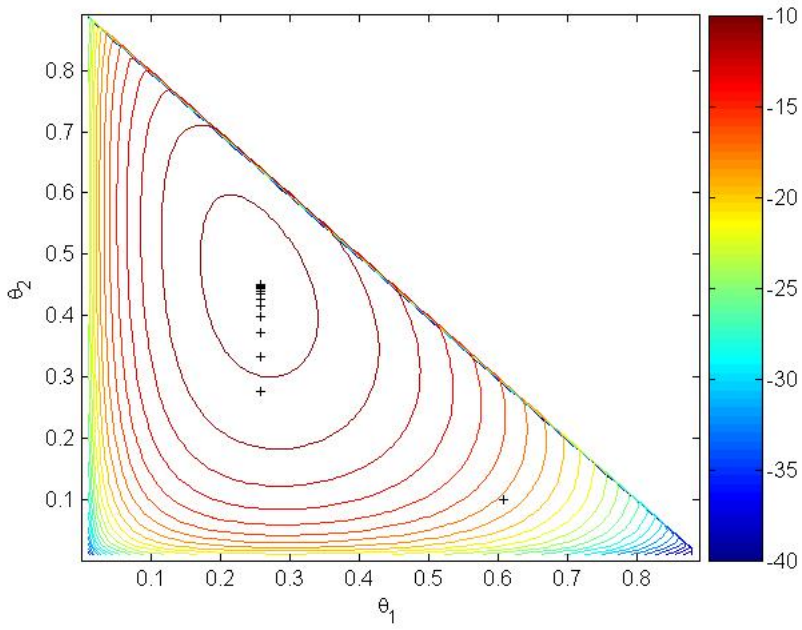

圖 3: 從概似函數的等高線圖看E-M algorithm 的迭代過程

3. 寫一支程式來執行 E-M algorithm 估計由兩個常態組成的混合常態參數。同範例 3 所述, 僅假設組合比例 π 未知。兩個常態的均值分別爲-1(for f_1) 及1(for f_2), 變異數均爲 1。根據以下樣本值估計 π:

$$Y_1 = -3, Y_2 = -2, Y_3 = -1, Y_4 = 1, Y_5 = 2, Y_6 = 3$$

畫出完整的概似函數 (11) 及估計過程中的每一個 π , 如圖4所示。從圖中的標號 (X) 看出遞迴過程中, 概似函數值只會遞增, 最後收斂在一個區域的極大值。不妨在程式的遞迴過程中, 加入概似函數的計算並列印出來, 以便觀察這個遞增的事實。

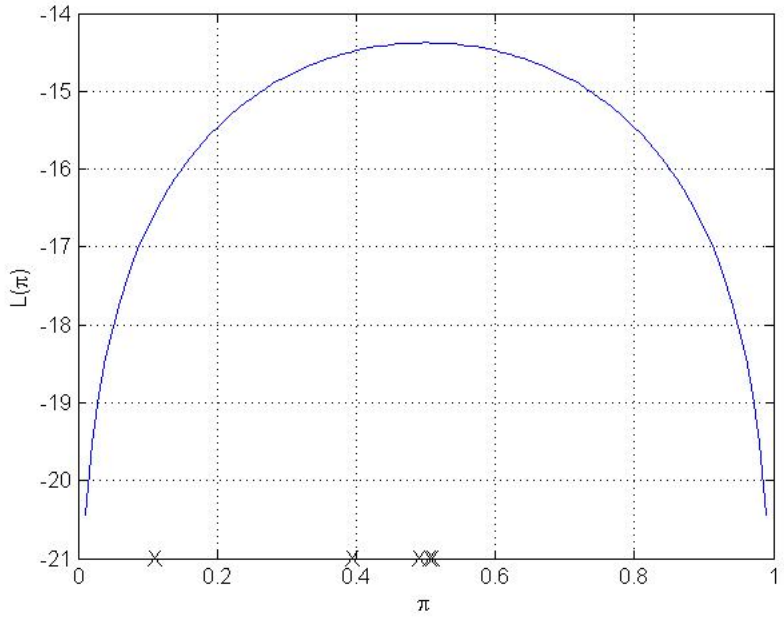

圖 4: 概似函數圖及E-M algorithm 過程中的估計值, 起始值在最左邊的 X。

4. 同上, 但是樣本值係依指定的常態分配自行產生。

3 觀察

1. 練習2,3的 E-M algorithm 都需要一個起始值做爲迭代迴圈的開始。嘗試不同的起始值, 觀察最後收斂的結果有何不同? 收斂的速度是否也有不同?

2. 對某些資料而言, 假設各群組的變異數相同反而是比較切合問題的本質, 試著在程式裡加上變異數是否相同的選項, 對每組資料分別執行變異數是否相同的情況。觀察這個假設對該資料的合理性。

3. 在畫直方圖與機率密度函數在同一張圖上時, 必須適度修正其相對的大小, 才能匹配在同一張圖上。注意, 直方圖要採用計算比例的方式, 而非計算次數 (在 MATLAB 裡可以用 bar 來配合 hist 做直方圖)。

4 作業

1. 計算出圖1資料的分界線 (方法自訂), 並依此將資料分成兩個群組, 分別計算其平均數、標準差與群組比例的估計。

2. 練習1畫出兩個常態的混合常態密度函數。試著依 (b)(e)(f) 的組合情況產生樣本, 並畫出類似其母體分配的直方圖。如圖5所示。

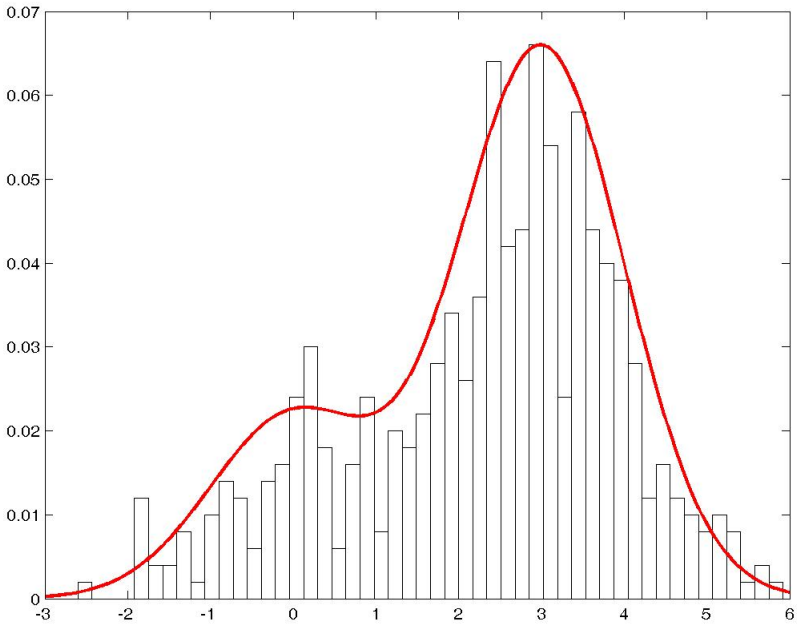

圖 5: 從混合常態密度函數產生的樣本所繪製的直方圖。

3. 證明範例 1 中的最大概似估計值爲

$$\begin{aligned}\hat{\theta_1} &= \frac{Z_1 + Z_1^*}{N_1 + N_2} \\ \hat{\theta_2} &= (1-\hat{\theta_1})\frac{Z_2}{Z_2 + Z_3} \\ \hat{\theta_3} &= (1-\hat{\theta_1})\frac{Z_3}{Z_2 + Z_3}\end{aligned}$$

4. 將 E-M algorithm 寫成一個完整的演算法, 含初始值與收斂的檢測。

5. 試著證明 EM algorithm 的迭代方式一定會收斂到一個區域的最大值, 即過程中, 對數概似函數值只會逐漸變大, 直到最高點 (請參考原論文或其他書籍的證明)。

6. 寫出範例 2 的已知資料的對數概似函數, 並求其最大值的參數估計式。

7. 有一群197隻的動物, 分屬於四個類別, 按多項分配分佈爲,$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$, 各類別機率爲

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$$

   (a) 繪製概似函數圖。

   (b) 計算最大概似函數的參數估計值 (分析解)。

   (c) 利用 EM algorithm 估計參數θ。(Hint: 將第一群分爲兩群$y_1 = (x_A, x_B)$, 機率爲$(\frac{1}{2}, \frac{\theta}{4})$。)

8. 畫出練習 3 中兩個群組參數的後驗機率。

9. 推導出式 (18)(19)。

10. 推導出式 (20)(21)。

11. 推導出式 (23)(24)(25)。

12. 寫一支程式利用 E-M algorithm 估計圖1的兩個常態群組的參數 (Example 9.3.1 [1]), 並將兩群組的個別及混合後的密度函數與直方圖畫在一起, 如圖6所示: (資料 bird.txt 請從網路上下載。)

13. 改寫上一個 E-M algorithm 程式, 使其可以同時估計多個變數的 normal mixtures。請從網路上測試資料 $em\_2d.txt$。畫出如圖7來。

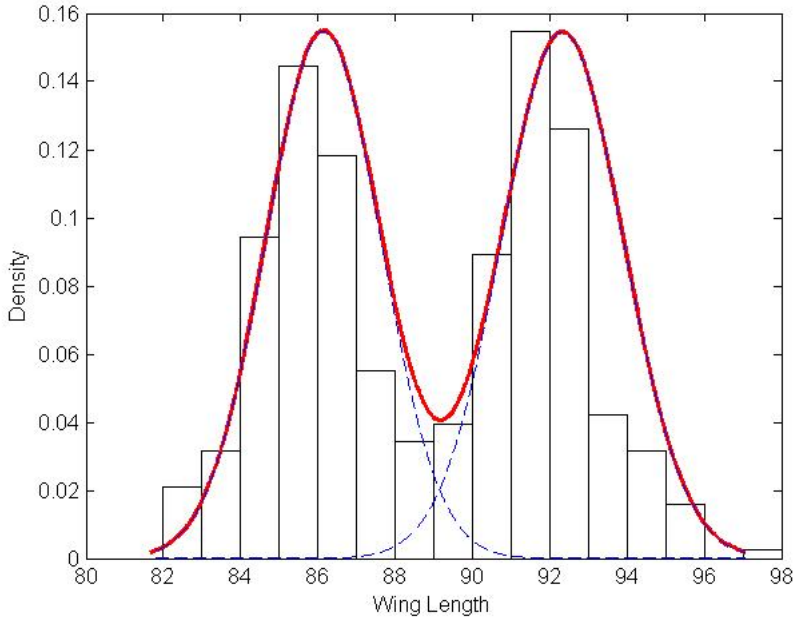

圖 6: 兩個群組的混合常態估計。

參考文獻

[1] B. Flury, "A First course in Multivariate Statistics," Springer.

[2] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning:Data Mining, Inference, and Prediction,"Springer

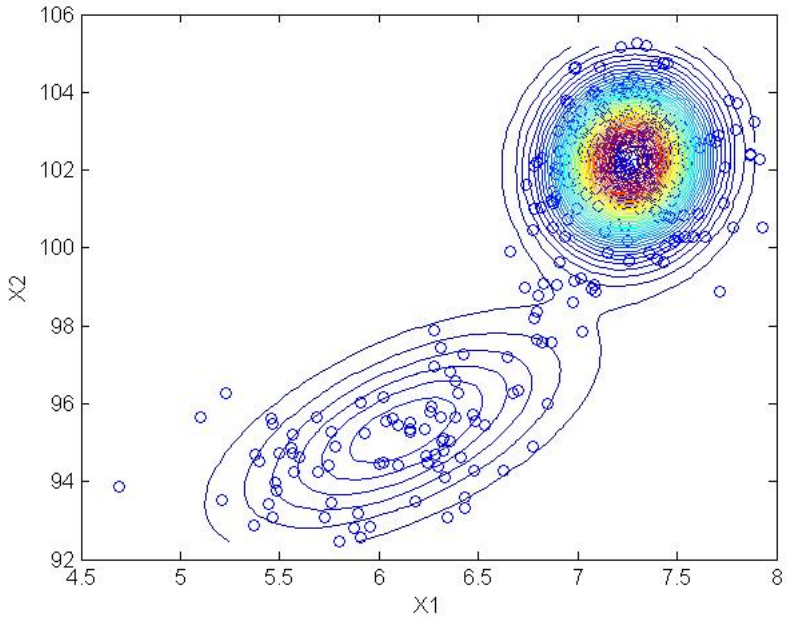

圖 7: 兩個變數、兩個群組的normal mixtures 估計。