

第九部分：讀一篇關於 EM 的論文

這是一篇「集大成」型的論文。對於統計推論——尤其是古典的 parametric inference——基本上可將問題簡化為：將 likelihood function $L(\phi|\mathbf{x})$ 找出來。因為一旦找了出來，計算 MLE $\hat{\phi}$ 只是一個優化型的技術問題。有了電腦、有了軟體、只要又肯花錢，只要未知參數 ϕ 的維數不高，這工作總是可以做的。

對有些問題來說，求出 likelihood function 就難，計算 MLE 當然就更難了。若是我們將可以容易寫出 likelihood function 的數據叫做「完整數據¹ (complete data)」，那麼較難寫出 likelihood function 的數據，是不是可以經由完整數據的 likelihood function 來寫？

之所以樣的想法，多因為不完整數據有時是因為我們嘗試求得完整數據失敗而得（例如問卷中的 non-response, partial answer 等）。假如有一組數據，它可以想像成由某一組完整數據「丟掉一部分訊息」而來，那麼我們可不可以利用原來的 likelihood function 來求出相對應的、關於這組不完整數據的 MLE 來呢？

這篇文章不管求出來的 MLE 對於統計推論是好是不好，只論求 MLE 的方法。它的賣點在：對於所謂的「不完整數據」，作者有明確而廣泛的看法，而且可以舉出很多例子來——因此對應用者來說，是非常有吸引力的。同時，作者將「求不完整數據的 MLE」的基本技術，有點像「分解動作」那樣地歸納成兩個步驟，分別叫做 E 步 (E step) 和 M 步 (M step)。其中的 M 步，對於 exponential family 而言，還含有「可儘量利用關於完全數據求 MLE 的軟體」的能力。換句話說，對很多不完整數據的問題而言，你可以在現有的軟體上只增加一個計算迴路 (loop)，就可以算出想要的 MLE。

讀我的註腳需參考原始論文。見

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), *Journal of the Royal Statistical Society, Ser. B* **39**, 1-22.

以下為我的註腳。

p.1, summary. 第一句話便說明了本文的全部內容，是名家手筆。用字不多，但恰到好處。第二句話直指技術面，而 monotone 是主要的技術因素（因此收斂是必然的）。由最後一句，便引出本文的一大堆例子來。

這些例子，有很多是被前人考慮過的。有些有 EM 的想法，有些沒有（或者那些作者沒有明白說出）。但將它們都看成一類而總體都能解，並有實際的為法來解，當然是重要的結果。

這是寫得極好的論文摘要，用字不多（也沒有用任何困難的字²），但全文的內容都被

¹這是本文的簡單看法。因為要從完整才能定義不完整。世上當然有我們所直觀了解的完整數據，但又寫不出 likelihood function 的。

²一般的統計論文，除了（專有）名詞外。大多不用難字——文章好不是用字難或不難，而是用得恰當與否。

涵蓋。

p.1, introduction, line 1. 這一句和摘要是一意思，但多了 *iterative* 一詞，講法就不大一樣。有些作者喜歡將緒論的內容差不多照抄到摘要——這是壞習慣，因為明擺著不肯用心，容易引起 referee 反感。

p.1, introduction, line 4. 一般論文極少自誇。此處用 *remarkable* 已有一點算是自誇³。這類事大教授可以做，因為他們的身份和品味擺在那兒。新手以不做為宜。但 “specify, generality, wide range of examples” 是用得好的。

p.1, introduction, line 6. 在此先做 *exponential family*，是因為大家在數理統計裡一開始就會讀到，是耳熟能詳的。這是統計理論上最好用的模型：所有的好方法，在 *exponential family* 上就一定能用而且會簡化到有直觀意義。當然，若是用不上，便強烈地暗示「所提的方法有問題」。

p.1, introduction, line 9 – (1.1). 這一段定義何謂 *incomplete data*：從某一組 *complete data* 經簡化、遺漏...而來。但真的怎樣來？用嘴巴是說不清楚的，必須用數學式子，因為後面要做證明，而空口白話是無法證明的。

要做研究就得把何謂 *incomplete data* 說明白。我的意見是：公式 (1.1) 是本文的主關鍵之一：用「積分」來說明白。對 *complete data* 的 *likelihood function* 積掉那些看不到的部分，便得到 *incomplete data* 的 *likelihood function*。以 Dempster 和 Rubin 的功力 (Laird 是 75 年的博士，那時還是小教授)，在有了 (1.1) 之後，本文其它的部分，應都可手到拈來。

但能看出 (1.1) 可不簡單。當然，看出來後再舉例子就容易了。例如若有 X_1, \dots, X_{100} 算是 *complete data*，它的 *likelihood function* 是 $f(x_1, \dots, x_{100}|\phi)$ 。若我們丟掉了 x_{100} ，則不完整的數據就是 X_1, \dots, X_{99} ，而我們可以用積分的方法找到這二者者的關係：

$$f(x_1, \dots, x_{99}|\phi) = \int f(x_1, \dots, x_{99}, x_{100}|\phi) dx_{100}$$

而這就是 (1.1)。

這一段明確地交待了 *incomplete data* 和 *complete data* 間的 *likelihood functions*，應如何連接。以後的工作，只是將這個現象用證明及例子講得清楚。

p.1, introduction, line 20. 這一段只是說，要「充分利用 $f(\mathbf{x}|\phi)$ 」。求 (1.1) 的極大有兩個辦法。(1) 直接把積分算出來，再求極大，但此法的條件是「能做出這個積分」；(2) 設法另尋蹊徑，避免做 (一般幾乎做不出來的) 積分。

p.1, introduction, line 23. 說明 (1.1) 未必為 *unique*。當然，我們可在 X_1, \dots, X_{101} 中，丟掉 X_{100}, X_{101} ，然後再得到 X_1, \dots, X_{99} 。這也暗示了，在實際的問題裡，找到可用的 $f(\mathbf{x}|\phi)$ 可能並不容易。

p.1, introduction, line 25. 此處進入本文的標題，謂之「點題」。注意到起名字是一個學問⁴。如 *bootstrap*, *jackknife*, *EM* 都取得不錯——易記，就會想到去用。

³但是論文又非得說自己的東西好。所以需要你能平實地說自己好。

⁴我以前工作的公司，有一個部門叫 *Human Factor*，專門替新產品取名字。

p.2. 此一頁只講一個例子。你可以直接用 (1.2) 來做，這樣做要解一個三次方程式；也可以用 EM 的想法，此時不需要解任何方程式。即使最後到了 π^* (真正的收斂極限)，也只是解一個二次式。這一頁不難，一般的博士生應可一步一步地跟著做。注意到作者原可自己設定一個例子，但卻寧可用 Rao (1965)⁵ —— 這也是一本經典教本。

p.2, line 4. 特別提起 genetic model，雖然下文和遺傳一點關係也沒有。作者只是說這不是人造的不自然的例子，是真有實驗的。

p.2, line 20. 提「只要八步」是故意的。

p.3, table 1. 最後一個 column 中的數字，暗示的是：收斂的速度，是 exponential rate。但作者卻沒有講 —— 意在言外。

p.3, line 1 to end of section 1. 此一段敘述 EM 的歷史因緣和作者的看法。對 Hartley (1958) 這篇文章，注意到作者雖說 “many times”，但又加上 “in special circumstances” —— 故雖在給別人 credit，卻不肯給足。

p.3, line 3 after table 1. 此處又將 EM 的廣度提了一層：和 robust estimation 拉上關係。

p.3, lines 7-9. 注意到，雖一直在提別人的工作，但仍在說 “special examples”。這等於是說別人沒有進一步的理論 —— 一直都在說自己好，卻不帶火氣。

p.3, line 10. 注意到 “always increase the likelihood”，並說這是 key result：對自己的結果，定要說話。

p.3, line 14. Dempster 和 Rubin 都是 Bayesian 中的大家，當然不會忘記 Bayesian application 了。

p.3, section 2. 先只講滿足 (3.1) 的指數族 (exponential family)。這一段和第一節的例子其實是一回事。只是前面是 special example，而此處是一般的 exponential family。

p.3, section 2, lines 1-5. 這一段先自己說 “strong restriction”，免得別人說。又說 section 3 的理論遠不止此，但 exponential family 另有意義，故值得另寫一節。

若在五零年代，你可以分別做 binomial, Poisson, normal, gamma ...。但到了 1977，就不能這樣地 elementary。

p.3, section 2, line 9, line 10, line 11. 注意用詞：“regular, convex, unique”。這些都是作者小心處。表示從數理統計做問題的嚴謹。又在 line 12 to p.4 line 1 中，提出只做 natural parameter case，一直到 p.4, line 6，這些話都是為「小心」而說的，只是表現作者的數理統計唸得很扎實。最後，集合 $A_\phi = \{\mathbf{x} : f(\mathbf{x}|\phi) > 0\}$ 和 ϕ 無關，是常用的假設，此處要放上。因為若沒有它，MLE 的求法（即使對於 complete data）都可能有困難，本部的簡潔結論，就得不出了。

p.4, lines 12, 14. 注意到，提到 (2.3) 是用 “equations”，因為它可看成一組方程式。你當然可以用 “equation”，將之視作一個用 vector 來表的 equation 就是。但注意到需一

⁵此時若在一本爛書上找例子，就不如自己來造。

致。即便是 likelihood equations 在此也用複數。

p.4, line 14. “familiar form” 此詞用得精準，因為 (2.3) 並不是 exactly a likelihood equation。

p.4, line 18. \log 是不該用斜體字的。同理，在方程式裡要用 \log, \sin, \tan, \dots 而不是 \log, \sin, \tan, \dots 。但在 $\log x$ 裡的 x ，字體卻又有不同。將來在你送出的文稿中要注意這種小地方。

p.4, lines 18-22. 這一段繼續說，注意到作者一直在反覆地說明同一件事。甚至於 x 是整數，但 $E(x)$ 不是整數都交待明白——這些大教授是很小心的。

p.4, line 17. convexity 回應了 regular exponential family 的條件。主要是爲了 (2.3) 的解是唯一——前面所埋的伏筆，儘管很不重要，最好都能關照——若不是爲了要用 (2.3) 的解是唯一，何必在前面又說 convexity，又說 regular exponential family？

p.4, lines 11-28. 這一段若由國人來寫，多半是用數學式子一個一個地套下來。但統計學還有社會科學的味道，用太多公式，只是表示你只能推導式，卻不能解讀公式。所以，能用講的，就用講。在這裡最能表現你的深度。

p.4, line 29 – p.5, line 17. 這一段雖然客氣地叫做 digress to explain，但要點在 explain (若無此意，則不必 digress)。雖然全是推導，但皆是形式上的推導。最後得到的 (2.13)=0，卻正和 (2.2)=(2.3) at $\phi = \phi^\infty = \phi^*$ 一致。這是古典 EM 的關鍵。故需用種種角度，一提再提。

p.5, line 18. 注意 “in special cases by many examples”，仍然是給 credits，但不全給。作者絕口不提若無以前的 many examples，他們根本做不出來。國人的寫法太謙虛，不好，因爲別人都不謙虛。

p.5, line 20. 提到 1966 的講義。告訴讀者「我們知道最早是誰做的」。在 1966，有人明明有好結果，但並不寫論文，只發一個講義而已。

p.5, line 20 – (3.16). 這一段是引自 Sundberg (1974)。Parenthetically 指「附加說明地」。這一段只是表示作者搞懂了這篇 1974 的文章。(2.16) 在 $k = 1$ 時恰好是 (2.13)。至於在 $k = 2$ 時恰好得到 covariance matrix 一事，本文雖有提到但未深究。後來 L. Thomas 因此做了一篇不錯的文章⁶。

p.5, line 36 – p.6, line 4. 這一段其實有點多餘。Curved exponential family 並不佔太多特殊位置，此處卻用它「不能得 (2.13)」的理由，來要改用新的 M-step。但此又爲下面的一般 EM 所包括。就文論文，此一段沒有新東西且不特出，是可有可無的中間步。

p.6, line 5. 注意作者以前只用 “exponential family”，此就卻用 “exponential familys”，何以故？以前只有一個 (指 regular exponential family)，但現在已有兩個 (regular exponential family and curved exponential family)。故用複數。

p.6, line 17. 此處又回到 exponential family，這相當於 checking。當你做了更廣的一步，常需要和已知的、知道是對的結果相互驗證。如果不合，就是某一就算錯了。

⁶L. Thomas (1982). *JRSSB* 44, 226-233.

p.6, line 22 to end of section 2. 這一段話把 Bayesian method extension 就全包括了。好好地改寫, 可作一篇博士論文呢!

p.6, footnote. 注意到評審對於某一個字的字義的挑剔。英文用字, 有時比中文要精確。其實中文也可以寫得精確, 只是大家都有壞習慣了不去要求而已。國人的英文已不如洋人, 如遣詞用句再下肯用心⁷, 人家的地盤上, 文章不易被接受, 是明擺著的。慎之, 慎之。

p.6, section 3. 此一節是技術性的: 證明定理。這是國人學者的強項, 但統計學的結果不在於在數學的難度。這幾個定理都不能算難——只是說明 EM 在甚麼時候可以用而已。定理一當然不難, 因為基本上, p.7 的 (3.6) 這個式子便已足夠。若自 (3.5) 來看, GEM 的定義, 是能夠讓 $L(\phi) \uparrow$ 的算法而已 ((3.3) 是 well known 的結果)。這是「看出定理較難, 證明反而容易的定理。事實上, 直到 p.8, line 3 都沒有太難。

p.8, line 12. 此一段說明前面證明的是甚麼。文中先說 $L(\phi^{(p)}) \uparrow$, 再說 $\phi^{(p)} \rightarrow \phi^*$, 最後還得去說 $\phi^{(p)} \rightarrow MLE$ 。

這類事一般都需要條件。所謂的好期刊, 一般都會要求作者好好把條件列出來, 好好證一證, 並且希望你所用的條件不能再弱。

p.8, (3.15), (3.16). 符號 D^{10}, D^{20} 猛一看沒有定義。仔細想, 其實 (3.15), (3.16) 就是定義。

p.8, line 32. 用 “an instance of a GEM” 是用語的小心。

p.8, line 32 to end of p.9. 這一段的證明, 你看不看都無所謂。但對於「投稿一流期刊」而言, 這樣的段子就非有不可。如果你要磨鍊自己手上功夫, 不妨試作去補足所有的技術細節。這些一般在上課是不教 (因為太 detail), 但研究上你該都理清楚。這四個定理都在說甚麼? 真正所需的條件為何? 這類細節, 將來你的論文做到有初步結果時都要一一補上⁸——對某些期刊而言, 這才算 good quality work。

p.9, line 30. “can be easily verified ...” 雖是廢話, 但也要說。因為否則這些定理的條件是否有模型能滿足都不知道⁹。但作者卻小心地說 “in many instances”, 這暗示著 “not in all cases”, 也是自我保護但又不負面的寫法。

這些都是號稱為了嚴謹而玩的文字遊戲。注意到真正的要點是 GEM 基本上是 by definition, it works。這些定理都是化妝後的結果: 真正用 EM 去做問題的人, 大概都不會去查條件滿足與否——除非做出來的結果自己都不信。

p.10, lines 1-25. 這一段說話, 是爲了表現學問而多寫的 junk。完全看不出作者是否真的證明過, 還是憑經驗亂猜。但三個作者都是 Harvard 教授 (Laird 那時是 assistant professor), 應該不會弄虛做假。

p.10, lines 26-32. 這是爲了回應 table 1, column 3。也可將 table 1, column 3 看作爲伏筆, 現在才回應。

⁷將心比心, 如果你評審一篇洋人用不通的中文稿, 會讓他通過嗎?

⁸儘管可能放在附錄, 或甚至要求你刪去, 有細節而被要求刪去和無細節被認爲不夠嚴謹而受拒, 是兩回事。

⁹有時, 爲了證明定理就加條件, 結果條件太多, 以致「滿足該條件的情形」變成空集合。因此說明自己的模型的確存在。爲了嚴謹也是重要的。

p.10, line 33 to p.11, end of section 3. 這裡在儘力向深處走, 但又語焉不詳。為何用到 second derivative 時, 便會 speed up convergence? (3.19) or (3.26) 說的是甚麼? 為何與收斂速度有關? 此類言語, 不容易說。因為說得不對時, 內行人便看得出來, 便成為畫虎類犬了。

自 line 45 起的一段話, 是回頭再做 literature review。前人的類似工作, 要一一承認。但作者仍然在說別人未竟全功 (如 “unusual special cases”, “without recognition of”, “do not focus on directly” 等。極力說前人雖有建樹但都缺臨門一脚。

最後, 還是回到 Bayesian, 和以前的話輝映。這是作者寫作用心處。

p.10, line 1 to p.11, end of section 3. 此段整體而言, 仍然是在說「我們對所舉的 EM 諸性質, 都儘量做了」。這是用來堵 referee 用的 (我們已非常努力了, 別多問細節了)。因為 $Q(\phi'|\phi)$ 這個函數得要算得出來才行, 其它都是空話, 因此文字方面有點枯燥, 就不足為奇了。

p.11, section 4, lines 1-9. 整個 section 4 都是在給例子。這裡說 “either has been or can be used”, 因此挑明了這些例子都不見得是作者的發明。這一節的主要目的是, 對於所舉的 incomplete data 的問題中, 將合理的 complete data 的形式想出來 (雖不一定 unique)。並導出 E 和 M 兩步。

p. 11, section 4.1.1 to p.21, end of section 4. 以下的各小節的寫法都頗類似 (但英文用字卻都有差別): 先介紹一個情形, 再將問題改裝成 incomplete data 的形式, 並指出相對應的 E 和 M 兩步, 應如何做。

總體而言, 這是一篇蠻囉嗦的文章。它的真實例容其實只有 (1.1) 和定理一, 其它的都是佐料。能舉出那麼多例子是作者們的本事。它表面不難, 但將那麼多以前的東西整理得只用 E 和 M 兩步就弄明白了, 是不簡單的。

更要緊的, 是這個文章有市場。它的架勢是: 幾乎所有的已知關於 incomplete data 的 MLE 的問題 (或者 Bayesian 的問題), 都可以放在這個架構下討論。使用的市場一大, 哪一個期刊會拒絕它?

February 2, 2007