關於統計論文的撰寫

趙民德

中央研究院統計科學研究所

2006/9

在人民大學

前言

這本小冊主要討論的是如何撰寫統計學的論文。雖然對於一般的科學論文有所著墨, 但是以統計學爲主, 而且是以統計學的方法論爲主。

撰寫論文其實只是末節, 研究的本身才是本質。「爲學問而研究」和「爲發表而研究」是兩個層面。我退休之後有時會選一兩篇經典論文細讀, 才發現有好些以前沒有唸懂的地方。甚至可以說, 如果當年眞唸懂了, 有些我自己的研究工作, 可以少走不少冤路。道理其實很簡單: 功力不足便要上路, 總免不了跌跌撞撞。

但是人類的進步, 都不是來自四平八穩的工作。如果都要等所有的預備工作都好了才開始寫論文, 恐怕一輩子都不夠做準備的。

第一部分: 心理建設

在大學任教, 以前除了到時候去上課, 學期完了給成績, 對於是否要寫幾篇論文, 多半是自願的事。現在時代一變, 變成了「 不發表就走路 (publish or perish) 」, 因此, 發表論文, 似乎變成比好好教書更重要的工作。

這本小册, 表面上討論的是如何撰寫可以發表的論文, 但我 想深一點談關於論文寫作的種種, 因爲學術上的功力是本質, 而如何撰寫乃是末節。

沒有研究工作就沒有論文

發表論文的背後基礎是研究工作。做研究工作有如尋礦: 你沿著一條或者幾條線索, 用某一種工具一點一點地挖下去, 希望看到有價值的東面。這裡「 有價值」是比較抽象的。開礦挖到了金子、石油就算成功, 最多是蘊藏量的差別罷了。但是對研究工作而言,「有價值」指的是「新的」,「能承先啓後的」,「替別人開路的」。這裡,「新」是必要條件。新結果在理論上指的是至今人類都還沒有做過, 並且是 non-trivial 的結果。

再好的結果, 如果是幾年前被其他人先一步發表了, 除非你能夠再弄一些新見解, 那麼你就差不多等於做了虛功。你可以有一點虛榮心的滿足: 我所做的研究和某某人一樣好。落後別人之後, 如果你一再試投也許還可以在較弱的期刊發表出來: 希望他們的評審者學問不深, 看不出來。

但這不是正途。發表論文並不是沒有邪魔外道但仍勉強合乎學術倫理的辦法。但這些方法既弄不長, 也搞不大。靠這一套辦法 (如果你夠小心的話) 可以混一口飯吃, 並且可以混得不壞甚至於可以小有名氣, 但這些之上, 如果還加上「我有一點眞功夫」, 那麼這些所謂的邪魔外道的本事, 就可以變得名正言順。

這本小册, 主要還是先鼓勵大家先有些眞功夫再玩別的遊戲。沒有實力, 甚麼都是假的。

為甚麼要鼓勵學術研究?

主要的道理: 科學[1]上的實力, 是一個國家最基本的實力。論文發表得多, 表示有很多人在做科研, 表示這個國家有學問有見識的人多, 代表這些可以反映在社會和諧進步, 人民生活美好以及船堅炮利。

但這雖是一般領導的簡單想法, 卻和眞正的科學本質是有差異的。讓我們做好研究的基本力量, 雖然很大程度上包括了政府和社會的鼓勵和壓力, 但能夠讓第一流人才半夜不睡搞科研的基本力量, 並不是這些。這類第一流人才, 如果他要在政府的獎勵條件之下優化他的科研成績, 根本用不著這樣三更燈火五更雞那樣的拼命。

深入推動研究的力量

眞正的研究工作是好玩的, 是有癮的。是由興趣, 好奇心, 探險的精神這類事物後面

[1]泛指較軟性的社會人文科學和物理化學這類傳統的硬科學。

推動, 才能讓一個智商很高的人[2], 圍繞著一兩個問題的焦點, 東翻西找, 千尋萬覓地想了解眞相。好的科學研究結果, 多半來自探索未知。

我們引用一段散文:

> 在美國找一個書店是不容易的事, 找到一個有點正式的書的書店, 更難。而我信步所之, 在這麼個小街, 發現這麼多可愛的書。眼前好像有一片眩目的光芒。掏出一把在倫敦機場換來, 還不會用的錢, 讓店家挑了兩先令去, 帶回一本《新科學家》來。
>
> 三翻兩翻, 即看到很熟悉的幾張畫。仔細一瞧, 這些書全是中國的東西: 一張是一六零一年湖北的鐵塔; 一是一六二一年射的火箭; 一是一六一零年河北的拱橋; 最好玩的是一張木刻, 是用河水推磨, 用風箱吹起旺火。

這是一個叫做陳之藩的學者, 當年從美國到了英國的劍橋去訪問研究, 在那樣的環境下深思反省之後所寫的一系列散文中的一段。那本書叫「劍河倒影」, 曾經給台灣的學子深刻的影響。陳先生是中文素養極好, 而科研也卓有所成的學者。上面的一段話, 來自他讀了丹尼約瑟的《東西方的科學與社會》(此人在 1947-1964 間的論文集) 這本書之後的反思。這本書講的是中西科學的比較, 中國科學對西方的影響, 以及中國社會與中國科學的關係等。

作者發現很多科學來自中國, 而一個自然的問題是: 爲甚麼那麼多的發明, 卻沒有導出像歐洲近五百年的科學發展?

這樣的問題, 我們中國人何以沒有被認眞地問過? 何以沒有被好好回答? 我們時常說「XX 中國古已有之」, 或者搞出一些四不像的如「中學爲體, 西學爲用」的道理, 更簡單或者更浮面的做法就像李鴻章那樣的用銀子去買外國兵艦來建北洋海軍。但是, 近代中國的積弱卻是一個事實, 而這個事實在最近 20 年才開始有所改變。

丹尼約瑟的結論, 主要在中國科學在整個發展中主要是爲了「實用」, 而歐洲近五百年的科學發展, 主要是爲了「好奇」。

中國人太務實了。如果歷代將「奇技淫巧」當作科舉的測驗項目, 相信我們的科學會全然不同。但這樣做, 也許可以產生一些第一流的工程師, 但還是產生不出學術上的能引領一代風騷的大師, 就像歷代的科舉狀元, 他們也許能留下一兩首詩, 但對於學術的貢獻其實並不大一樣。

丹尼約瑟以半生的時間跑完了中國, 又淹在劍橋的書海裡, 去發掘中國的科學史, 這裡除了「好奇」, 還能說出其它原因嗎? 反過來看, 我們能不能找到一個中國的丹尼約瑟, 以半生的時間, 淹在南港的書庫裡去研究歐洲的科學史來解答「歐洲近五百年的科學發展, 主要是爲了好奇」這個假設是對還是錯[3]?

陳之藩最後說: 在這種笨人不能產生之前, 我們所謂的科學, 還是抄襲的, 短見的, 實用的。也就是說, 眞正的科學, 是不會產生的。

類似的反省, 在時下的網路小說裡也看出一部分來。這類的小說, 算是「架空歷史」

[2]我們不敢說研究工作者智商一定高, 但這一批人的確是較一般人通過較多的考試, 所以至少是一批國家菁英。
[3]這一段是陳先生的原文。

類。多半說的是某一個知青, 忽然回到過去, 如何用他們的現代知識來改變古中國的命運。但他們的著眼, 仍在科學的應用和制度的改革上, 對於「好奇」的概念, 沒有著墨。

研究的目的

好奇和實用: 這兩點我們不需再說。

明確的了解: 把原始論文說得更清楚, 以便後人更易了解。例如 Taylor 定理的原始證明是很長的, 現在任何一本微積分大概都用不到一頁。又如以前號稱世上只有七個人讀得懂相對論, 但現在任何一個物理系的畢業生都能懂。

確認: 做科研相當於人與天爭, 因此要步步爲營, 以便確認之後的結果後人敢大膽地用。

國家鼓勵研究的目的

這是使國力強大的投資。智識就是力量, 就是財富。這表現在戰爭中尤爲突出。雖然, 搞科研的人士, 一般都傾向於反戰。他們想看到的是贏在實利: 國民所得逐年增加, Gini index 逐年減少。

使國力看起來強大: 這是很多學術官員能夠做的。在全國的層面上, 他們只能做一些以論文的數量和質量爲標準的獎勵辦法, 但是上有政策下有對策, 早晚這類獎勵發表的政策和計算論文的方式都會被某些人破解。

關於 SCI/SSCI

亞洲各國的 學術官員對於 SCI/SSCI 有一種莫名其妙的信任。但 SCI/SSCI 的唯一道理, 是「它不由國內的教授控制」。但這最多只能拿來做過渡的手段來用, 不能長期倚靠。學術最終還是希望在國內有一套有聲望及傳統的學術評鑑機制。一個大國, 必須要能做到「自己說了算」。雖然這一點並不容易, 全得靠硬功夫和好的學術倫理和制度。

下面的信是 1962 年陳之藩寫給胡適的:

> 告訴您一個好消息: 我又有一篇論文, 就在 IRS 印出來了,...這篇文章很短, 但還乾淨, 題我叫「A new approach to Fourier coefficient evaluation」, 您大概數學都忘光了。這是數學上的大問題, 我提出個圖解的方法來。有個數學家說, 如果世界上的書全要燒光, 該留 Fourier series 。我的貢獻並不大, 但寫得很俐落。這是照著您在東廠胡同教我的八字秘訣:——「開門見山, 水清見底。」開門見山不難, 少廢話就行了。讓水清見底是談何容易的事。

有沒有現代的年青學者, 給他的博導寫類似的信? 這是我們傳統裡有味道的部分。如果他寫成這樣:

> 告訴您一個好消息: 我又有一篇論文, 就在 *** 印出來了,...這個是 SCI 期刊, 以 Impact factor 而論, 在同類期刊中排名第四 ...

這也許還有一些尊師重道的部分, 但是深度就差了。但這是現在的遊戲規則[4]。寫到這裡, 眞免不了有很多無奈。

心理建設

這一部分的目的, 主要是讓大家對於「寫論文」一事有一些心理建設: 有些事情是需要了解一下的。

1. 能刊載的論文未必好: 能刊載只表示通過某學術期刊的評審而已。

2. 被拒絕的論文未必糟: 有幾類論文容易被拒: 相當糟的論文, 不討喜的論文, 擋著別人路的論文。

SCI/SSCI 等只是參考: 它主要是以某 data base 上的紀錄來決定是否收入, 因此較不含個人的喜惡。觀念上我們別將它們看得那麼大: 基本上 impact factor 只是 popularity index 。也別將這些看得那麼小: 意思是說, 亞洲各國的科技官員也並非那樣的不了解狀況。對於 $(100-x)\%$ 的教授而言, 登出一篇 SCI 論文, 也並不容易 —— 寫一篇論文的基本功夫你還是要做。對政府的鼓勵和獎助, 年青時不妨積極些, 年紀大了, 就該用平常心看待。

這一切全在您把自已看成是哪一類學者。這句話說得比較有鼓勵性。但我的意思是: 做一個能讓人尊敬的學者, 其實並不是遙不可及的事。和最好的學者相比, 一個實在的博士, 在專業上也不過是相差在兩三本眞正好書和十幾篇眞正弄懂的經典論文。如果客觀環境略好一點, 有一兩年的「折節讀書」也就趕上去了。但是, 維持在第一線, 也需要用心用力。

原則上, 寫論文的技術好學, 但實質的部分則需要功力。而功力指的是學養、品味、苦工、眼界和心胸。寫論文並不是做燒餅油條, 甚至不也是製造 IC 晶片。在學術圈,「人民的眼睛是雪亮的」。長期下來, 只有眞正的學者風骨才能得到尊敬。

功力的來源: 這包括了先天的優勢, 自身的努力, 師門的教導和保護, 工作團隊的合作, 秘笈 (一個好圖書館), 諍友和好學生。

好文章

好文章有很多看法。但是有一點是公認的: 前 $x\%$ 的文章 (而 x 的值蠻小), 才叫好文章。因此, 大部分的作者, 大多數的時候, 都只是在做不怎麼好的文章。這裡並沒有太多的不滿, 因爲沒有那 $(100-x)\%$ 的普通, 怎顯得出那 $x\%$ 的不同? 因爲國家大力要求論文的發表, x 的值只會愈來愈小, 因爲分母在持續增加。可以這樣說: 論文超過某數的作者, 當然是會有好作品的, 但多數作品的也不會太好。

別好高騖遠, 儘找自已做不出, 別人也做不出的題目; 別急功近利, 儘找一定可以發表的題目; 這兩者都很難增加您的學術地位。前者是因爲多半您也做不出來; 後者則是別人也看得出來您沒有多少貢獻。對您的學術生涯規劃, 並不理想。

較佳的選擇

[4]這是以西方的價值爲價值的規則。一個大國應該有她自己的志氣。

做比自己能力略難一點的題目; 這樣才有一點挑戰性, 別人也不會覺得您總是柿子挑軟的來捏。長久下來, 會對您產生一定的敬意, 自己的滿足感也有一些。

專題的範圍, 不要太窄, 最好要有中長期的布局; 這樣, 十 年下來, 您可以有能「 成一家之言」的機會。

不要東打一拳, 西踢一脚。這種亂槍揚打鳥的方法, 投資報酬率太低。

September 4, 2006

第二部分: 好論文的特質

甚麼是好論文? 這是較難的問題, 但卻是要緊的問題。學術期刊都在挑好論文, 因此較好的論文才有較佳的機會發表, 即做是純粹爲了發表, 寫好論文也是重要的考量。

寫作是有技巧的, 但我們還是認爲: 內容第一, 包裝第二。事實上, 能做出好內容的學者, 寫作的本事也不會太差。因爲他們既然能夠在千絲萬縷的迷宮裡發現出前人未知的道理, 這些人對於邏輯推理以及清楚表達的能力, 應該都是沒有問題的。的確, 有的時候, 某些結果是經由作者經由他獨立思考方法東拐西繞之後才得到的, 如果沒有經過清楚的沉澱, 寫出的論文只沿著作者原來幽暗的思路而發展, 其他的讀者未必能懂。有時作者認爲自然的事物, 讀者未必也認爲自然。好的結果未必好懂, 但是內容, 仍然只有內容, 才是論文的本質。

有人說, 好論文如美女, 看到了就知道。錯! 美女一看便知; 好論文只有較高手的同行才知道。因爲學術論文這檔子事不是普羅文學, 而是一種精緻文化。

低段的棋手, 往往不能看懂高段棋手的布局和收官 —— 這需要足夠的背景資料的了解。學術論文絕不是通俗文化, 它是非常講究的、針對一小部分有足夠知識的讀者所作的展示。理論上, 一篇論文應包含所有能讓讀者讀懂該文的資訊。學術論文裡不見得沒有錯, 但那都該是「誠實的錯 (honest mistake) 」。因此絕不該有故意的錯。像金庸小說裡郭靖故意搞錯寫一部九陰眞經去胡弄歐陽鋒的事, 在學術論文裡就是犯了大忌的[1]。

讓我們來隨便看一看甚麼叫做學術上的好。我們說「隨便」, 因爲自己見識也是有限, 並且主觀也強, 是不可能將所有關於好論文的條件都一一列舉的。讓我們從最簡單的一個公式說起。下面的一個公式是號稱是世上最漂亮的公式:

$$e^{i\pi} = 1 \tag{1}$$

這裡面有, $e = 2.71828\ldots$ 和 $\pi = 3.14159\ldots$ 這兩個各自合乎天人之道的「 超越數」, 再加上 $i = \sqrt{-1}$ 這樣的一個虛數, 這三件東西怎麼會巧妙地連在一起?

科學家心中的審美觀是非常主觀的 。我講不出十分有說服力的道理。但是, $e = 2.71828\ldots$ 是來自

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

的極限, 而這成就了「複利」的概念: 生長的速度, 和自身的大小成正比。這正是自然界「生生不息」的描述。, 世間有那麼多「exponential growth」的模型, 並不是偶然。而另一面, $\pi = 3.14159\ldots$ 來自圓周率, 從這裡造就了多少三角函數這種周期函數。是的,「周期」是我們的關鍵字, 因此 π 和我們所知的「循環不息」密切相連。而公式 (1) 巧妙地將這兩種合乎天人之道的事物連在一起。

我不是在講《易經》。但公式 (1) 是明明白白的可以用最嚴謹的數學講清楚的事!

甚麼是「好結果」? 首先的要求是「並非無足輕重 (non-trivial) 」。大體來說, 凡是能丟給學生做, 並且你也相信他或她大概會做得出來的問題, 大概就算是 trivial 。大部分

[1]情治人員可以在某磁片上加一點或減一點去胡弄對方的間諜, 但是在學術期刊上公開發表的論文就不可以這樣做。

博導給研究生的題目, 雖然不一定算是 trivial , 但也不會相差太多。原因無他, 如果不是長期浸淫於某一專題並有相當的功力, 好題目一般並不容易找到。如果有, 博導自己爲甚麼不做?

能問好問題, 並解之

年青時常聽到一些老數學家說:「證明定理不難, 難在寫出定理的敍述」。現在年紀大了, 有時覺得寫出定理也不那麼難。玄幻小說裡常提到, 修眞的人功力一深, 便漸有預知的能力。做科研久了, 對於自已浸沉已久的專題, 甚麼東西會成立, 甚麼事情大概不會對, 多多少少是可以猜出來的。

所謂的好問題, 是我們一般猜不出答案的。但當你一步步把它梳理清楚之後, 它的答案又明白不過。這裡, 你的學術貢獻有二: (1), 問出好問題; (2) 解決它。

古典的統計方法 (例如在如 Kendall, Stuart and Ord[2] 這樣的經典教本中所列舉的方法) 多假設

$$X_1, X_2, \ldots, X_n$$

這組資料是給你分析的。基本的想法是別人已做了實驗, 取得數據, 現在請統計學者去分析。這類的架構, 在五六十年前是標準的。在數學上, 因爲

$$(X_1, X_2, \ldots, X_n) \in R^n$$

其中的樣本數 n 固定, 是一個有限維 (finite dimension) 的問題。所有的統計推論, 是根據這組數據的聯合機率密度函數

$$f(x_1, x_2, \ldots, x_n \,|\, \theta)$$

而發展的。這裡的要點是: n 是固定的。

要打破這樣的架構並不容易。在這架構之下, 有些事情是做不出來的。例如即使在

$$X_1, X_2, \ldots, X_n \sim \text{ iid } N(\mu, \sigma^2)$$

這樣的最標準的模型下, 要算出一個關於 μ 的長度爲給定值的一個信賴區間, 也做不到。比如說, 教本的做法是利用

$$\frac{\bar{X} - \mu}{s} \sim t_{n-1}$$

來造出

$$P\left[\bar{X} - st_{n-1,1-\alpha/2} < \mu < \bar{X} + st_{n-1,1-\alpha/2}\right] = 1 - \alpha$$

其中, 這信賴區間的長度是

$$L = 2st_{n-1,1-\alpha/2}$$

而因爲 $s^2 = \sum_i (X_i - \bar{X})^2/(n-1)$ 是一個隨機量, 不會等於任何給定的常數。

且不問爲甚麼我們有興趣去找一個「有固定長度的信賴區間」, 至少在幾十年前, 這是個大家都知道的問題。若是誰能解出來, 總有一點學術上的功勞[3]。這裡面有兩條路可走, 第一是設法證明這問題做不出來; 第二是換一個條件來把這個問題做出來[4]。

[2]這是三冊的 The Advanced Theory of Statistics, 初版在 1961 年。

[3]五六十年前的數理統計和數學非常掛鉤, 這類蠻數學的命題, 還是有一點市場的。

[4]這裡面有一個次序問題。先做出第一點再做出第二點是高手; 雖然多半的人是先做第二點。

證明這件事不可能是可能的。這是 C. Stein 的工作[5]。這是個好結果, 因此可以刊載在好期刊上。但是, 證明「不能做」某件事, 因爲別的數學家老早就證明了「只用直尺和圓規不能做出三等分一個任意角」, Stein 的結果, 就不能那樣的劃時代了。因爲在三等分角的問題被解決以前, 人們根本不知道「某命題不可能做出」這件事居然是可以證明的!

既然可以證明在樣本數 n 爲固定時做不出來, 那麼在 n 不預先固定的時候呢? Stein 的解法大概是這樣的: 先取 n_1 個樣本, 做一些計算, 再根據這 n_1 個樣本所算出來的結果, 來判斷是否我們還需要再取 n_2 個樣本 —— 視情況而定。在這樣的安排之下, Stein 可以造出一個關於 μ 的長度爲固定長的信賴區間。注意到, 他所用的樣本數不固定: 有時是 n_1 , 有時是 $n_1 + n_2$ 。

針對同一個問題, Stein 就至少有兩篇好論文出來。但這問題並沒有結束: Stein 的信賴區間是否可以做得短一點? 是否已做到最短? 是否可以做到 normal 之外的分布 ...。

這類的問題可以問得更深入。隨著問題的深入, 題目也愈變愈窄。解法和所需要的數學也愈來愈難 ...。這在序列分析 (sequential analysis) 裡, 最後成了一個蠻特殊的專題。

Stein 的工作, 告訴我們在樣本數不固定時, 可以做出一些樣本數預先已固定時我們做不出的事。這個結果在概念上給我們的啓發, 遠大於技術面的貢獻。而其它諸子的工作, 就多半只剩下數學上的困難, 就味道來講, 就不免要弱一些了[6]。

我們再舉一個例子。在

$$\int f(x)dG(x)$$

這樣的寫法裡, 我們基本上要要求 $G(x)$ 是一個平滑函數。雖然我們可以對 $G(x)$ 的性質降低一點標準, 例如在 Riemann 積分中, 如果 $G(x)$ 是一段一段連續的就好。而在 Lebesgue-Stieltjes 積分中, 我們只要求 $G(x)$ 可以用來定義某一測度。但這樣的條件不能降得太多。因爲 $dG(x)$ 這一個符號, 講的就是, 如果將 $G(x)$ 一小段一小段來看 (不論是對 x 軸還是對 y 軸), 因此對它的某些連續性, 總是需要的。

但是如果 $G(x) = B(x) =$ a Brownian motion, 這個積分的問題便完全不一樣。這是因爲除了它的隨機性以外, $B(x)$ 還是一個到處連續, 但無處可微分的函數。

這時, 要能夠清楚的說明白甚麼叫做

$$\int f(x)dB(x)$$

就是好結果 (K. Ito[7])。因爲要做到這一點, 你必須跳出原有的對於積分的概念那類的框框, 而重新看到並梳理出一些全新的東西。有了重新的定義, 才有所謂的「隨機積分」這回事。

抓住問題的本質並說明

衍生性商品 (derivatives) , 指的是證券的買權或者賣權那一類的未來合約。這種合約是可以交易的, 而它的交易價通常是針對某一種上市股票的價格的變動而變動。例如 X

[5]感謝張源俊教授告訴我這件事。

[6]例如, 在序列分析裡, 因爲數據是 $(X_1, X_2, \ldots) \in R^\infty$, 問題的數學結構要難得多。

[7]原始的論文大概是 1946 年。當然, 現在多半的關於 stochastic integration 的書都會詳細介紹。例如 J. M. Steele (2001). Stochastic Calculus and Financial Applicaations, Spriger , 第六章。

公司的股價現在是 30 元, 三個用後, 我是否可以用每股 35 元的價格來認購? 如果可以, 我現在要用多少錢來購買這樣的權利?

衍生性商品的市場值, 現在已大於股市。是目前財務金融市場中最要緊的商品。那麼, 它的公平價格, 應該放在哪兒?

這個問題在 1973 年[8] 才有重要的突破。甚麼叫做公平定價? 一個重要的想法是: 這個價格要能使得交易雙方都沒有完全不冒任何風險而輕鬆獲利的機會。在 Black 和 Scholes 的論文裡, 他們將公平這個觀念, 用「有效市場中無套利」這個基本條件來抓緊在一起。而在這樣的條件下用精準的數學來導出所謂的 Black-Scholes 公式。因爲他們清楚地將問題的本質抓住, 雖然 Black 和 Scholes 所做的是最基本的 European option 的定價, 但這一點突破, 卻是近代數理財務學中最關鍵的步驟之一。近十年來, 它都是最火的學科, 因爲許多其它的較複雜的衍生性商品的合理定價, 基本上也還是要靠類似的想法來推導。

我們再換一個問題, 問一下「甚麼叫做 information 」?

這個問題也不簡單, 但我們不能平空來問這樣的問題。一個有效的方法是, 先找到一個有用的架構, 然後再試著把這件事說明白。

不同的架構下, information 有不同的意義。例如在統計推論的架構下, information 一詞, 指的是 Fisher's information 。它的基本想法是, 數據是用來做統計推論的, 而未知參數的估計, 是統計推論的最基本的部分。因此, Fisher's information 所看準的問題是在問: 對於未知參數的估計而言, 我們最好能做到甚麼程度?

但對於通訊理論而言, 問題的焦點就不在於參數的估計, 而在資訊的傳達: 一連串訊號從 A 點出發, 經過一個介面, 再傳到 B 點, 關注的是: 有多少東西送進了 A, 又有多少東西由 B 傳了出來?

爲了清楚地說明這件事, 我們首先得問「在甚麼情況下, 通訊介面其實沒有傳遞任何訊息」—— 換句話說, 我們先得知道甚麼是 0 , 然後才能知道甚麼是 1 。

至於如何說明何謂沒有 information , 一個比較合理的想法是先列舉幾個「不含 information 」的合理條件出來。而在這幾個條件下導出 entropy 的觀點。

以上基本上是 Shannon (1948) 的結果[9], 這在通訊理論上是重要的一筆, 甚至於發展出一個專屬學術期刊來[10]。

客觀條件成熟時的總結

人的想和和基本解析能力, 其實是差不多的。如果某一類問題經常困擾著我們, 那麼自然便會有許多不錯的學者去鑽研它。他們的解法, 也往往有類似的想法出現。但每一個人的出發點不同, 所面臨的數據結構也不同, 因此看到的問題和解法也會有些共同點, 也有些差異。時間一長, 功力深厚的有識之士, 便會看出這類問題的眞正癥節, 從而能做

[8] F. Black and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* **81**, 637-654. 多半的關於數理財務學的教本都會提到。

[9] C. E. Shannon (1948). A mathematical theory of communication, *Bell System Technical Journal* **27**, 379-423, 623-656.

[10] IEEE Transaction of Information Theory 創辦於 1953, 是 IEEE 的招牌期刊之一。

一個清清楚楚的整理。這類的論文, 是一種「集其大成」型的論文。

例如 Dempster 等[11] (1977) 在已有很多論文討論不完全資料如何處理後, 才推出 EM-Algorithm , 將計算 MLE 的手段, 歸結到兩個步驟。因爲這個集大成的工作, 涵蓋面非常的廣, 幾乎將所有的關於不完整資料的處理方式, 一網打盡, 從此變成經典論文。

另一個在客觀條件成熟時才引起廣大迴響的論文是關於自助重抽法 (bootstrap method) 的, 我們知道, 一般的統計推論問題已基本上可以歸納到「 求出某統計量的分布」。而古典的統計方法是利用解析上的功夫 (包括大樣本理論) 來計算該統計量的分布。但解析的方法有限, 當所有的解析方法都已用盡時, 我們還有甚麼其它的方法?

當電腦計算變得容易時, 當幾乎人人都有一台 PC 時, 當統計軟體變得普遍時, 這時總結出一套「如何利用模擬以計算適當統計量的分布」這便是自助重抽法, 見 Efron (1977)[12] , 這也是一篇經典論文。

新技巧的引入

有用的新技巧的引入, 也是好論文的條件之一。例如 fast Fourier transform (FTT) , 可以將一個需要 $O(n^2)$ 次的計算, 簡化成 $O(n \log n)$ 。又如 Markov chain Monte Carlo (MCMC) 的想法, 基本上將 Bayesian 方法的罩門 —— 後驗分布多半算不出來 —— 加以解決。這類的結果, 當然會引起廣泛的重視。

當有一大類問題需要解答時

設限數據 (censored data) 是近代常見的數據, 這尤其對於描述壽命長短的問題, 經常出現。例如 Kaplan-Meier estimate , 這是對於一組設限資料如何找出合理的對 cdf 的估計量的論文。見 Kaplan and Meier (1958)[13] 。因爲這類問題實在遇到得太多, 而且這也是很多關於如何處理設限數據的基本論文, 它差不多是世界上被引用次數最多的論文之一。

比估計 cdf 更進一步, 如何對於設限數據來做迴歸分析, 那當然是更會讓人用到的結果 (Cox 1972)[14] 。如果數據中還有一部分相依, 類似的迴歸又要如何去做 (Liang and Zeger, 1986)[15]? 這類的結果, 是因爲太多組數據是這種形式而變得重要。

這類好文章的來處

這類所謂的好研究成果是如何來的? 最自然的是因爲功力到了自然形成。當你在某一個行業久了, 你自然知道甚麼是重要的問題, 甚麼是用一點苦功就能做但做出來也功勞不大的問題。因爲爲了要發表, 你需要做新的東西。如果你對當前的一般行情都摸不清楚, 你能玩這類遊戲的資格還淺。學術發表, 就是要走在別人前面。

---

[11] A. P. Dempster, N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm, *J. of Royall Statistical Soc. B*, **39**, 1-21.

[12] B. Efron (1977). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, —bf 7, 11-26.

[13] E. L. Kaplan and P. Meier (1958). Non-parametric estimation from incomplete observations. *J. of the American Statistical Association* **53**, 457-481.

[14] D. R. Cox (1972). Regression models and life-taaables. JRSSB **34**, 187-201.

[15] K. Y. Liang and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

我們常說, 文章本天成, 妙手偶得之。這一點, 對於學術論文而言, 往往不是這樣。一般來說, 妙手很難偶得。基本上, 即使你已到了作戰的第一線, 你還是不能太鬆懈。因爲和你競爭的對手都沒有鬆懈。

王國維在《人間詞話》中提出境界一說。謂

> 古今之成大事業、大學問者, 必經過三種之境界 昨夜西風凋碧樹, 獨上高樓, 望盡天涯路。此第一境也。衣帶漸寬終不悔, 爲伊消得人憔悴。此第二境也。眾裏尋他千百度, 驀然回首, 那人卻在, 燈火闌珊處。此第三境也。此等語皆非大詞人不能道。然遽以此意解釋諸詞, 恐爲晏歐諸公所不許也。

有人說做研究亦然。我比較相信前面兩個境界, 那指的是苦讀和迷惘。至於頓悟卻是少, 畢竟研究工作是靠邏輯、實驗和計算哪。

這類所謂的好文章很難教出來。畢竟, 大部分的科研工作都是苦哈哈地做出來的。我列舉這些例子, 主要是給大家看看: 甚麼叫好。如果你覺得這些離你太遠, 也不要灰心傷志, 因爲其它的大部分文章, 都不是這類水平。

我甚至覺得, 若是在所有有「項目」的學者中, 有 2% 的學者能到這種程度, 則國家獎勵學術的經費, 就沒有白花。要國家獎勵學術的經費沒有白花, 談何容易啊!

最後, 我用一小段引文來結束這一段。第一段來自《東周列國志》:

> 姬平乃歸燕都, 修理宗廟, 志復齊仇。乃卑身厚幣, 欲以招徠賢士, 謂宰相郭隗曰:『先王之恥, 孤日夜在心, 若得賢士, 可與共圖齊事者, 孤願以身事之, 唯先生爲孤擇其人。』郭隗曰:『古之人君, 有以千金使涓人求千里之馬。途遇死馬, 旁人皆環而嘆息, 侍從官問其故, 答曰:「此馬生時, 日行千里, 今死, 是以惜之。」涓人乃以五百金買其骨, 囊負而歸。人君大怒曰:「此死骨何用? 而費吾多金耶?」侍從官答曰:「所以費五百金者, 爲千里馬之骨故也。此奇事人將競傳, 必曰:『死馬且得重價, 況活馬乎?』馬且至矣。」不期年, 得千里之馬三匹。今王欲致天下賢士, 請以我爲馬骨, 況賢於我者, 誰不求價而至哉?』

國家出錢買科研成果, 是不是這樣想的呢?

September 4, 2006

第三部分: 一般論文

在功力不足時寫好論文較難, 但前面說過不能等一切都預備好了再寫論文。這樣做眞可能一輩子都只是做講師助教授。因此, 新手上路是必須的。並不是說新人就寫不出好論文, 世上大有聰明的人, 極年青時就光華燦爛, 引人注目。我們當然希望最後大家都有較大的學術貢獻, 但目前我們且把身段放軟, 將目標別定得那麼高。

如何寫論文是次要的, 如何找題目才是眞的。現代網路發達, 只要蒐尋的方向是大略正確的, 年青學者並不難找到相關的學術資訊。

別做學位論文的推廣

畢業論文裡往往有最後一節: 未來的工作或今後研究的方向。你當然會很自然地去做原論文的推廣, 不論怎樣, 這是你還算熟悉的專題, 何況你的博導說不定還會給你一點指導。

我的建議: 別走這條路; 或者, 至少你別希望靠這條路多寫幾篇論文。

理由是有一些的: 第一, 這往往是沒有多大前景的題目。學位論文是用來訓練研究生的研究方法、態度和寫作技巧用的, 而訓練的意味大於研究的意味。你的博導知道你大概做得出來, 或者至少他自己知道他一定會做得出來。這樣的題目做練習可以, 做研究就有一點不夠難。第二, 如果你仍然倚靠你的老師, 請問論文將來要不要掛上他的名字? 如果要掛的話, 是你排第一還是他排第一? 我其實並不十分反對老師和學生合寫, 畢竟老師在學生身上用了不少時間指導, 因此在學生的論文上掛一個名字也算應該。但我覺得三篇以內還算可以, 超過此數就有一點過了 —— 學生更要考慮升等, 研究項目等世俗上的事, 聯合而寫的論文, 尤其是和自己的老師合寫的, 在學術評鑑上是會被打折計算的。道理很簡單: 升等的最重要的要求是「該員有沒有獨立研究的能力」。獨立者, 不靠別人自己單幹也[1]。第三, 老師的題目, 如果他是這個專題的專家, 你多半做不過他 (和師兄師姐)。青出於藍談何容易? 別忘了這些人都是競爭者。

簡單地說: 要長大, 先斷奶。

數據和論文

統計學的論文分爲兩大類: 有好數據的和沒有數據的[2]。做沒有數據的問題比較簡單, 作者基本上只要提出, 對這個問題, A 做了甚麼, B 又做了甚麼, 但他們都沒有做出結果 C 。你只要把 C 做出來了, 當然就算是研究有成果。這在 A ,B 都是有名氣的教授時, 你的理由就更充分了。做這類問題, 最要緊的是你手上的數學計算的功夫要好。國內出身的統計學者, 在國外好期刊上經常有論文的, 幾乎沒有人數學不強的。

只要數學夠強, 你可以放心大膽地一再把條件減弱或者計算得更精細。例如某人證明了弱收斂, 你就做強收斂, 或者進一步去算它的收斂速率, 甚至大離差 (large deviation) 的結果 . . . 。

[1] 在某些單位,「沒有 single-author paper 」是研究能力的重大瑕疵。

[2] 如果你只用某些老掉牙的數據, 如 sun spot data, iris data, Nile river data 等, 不算是有數據。

下面我舉一個例子, 是關於 Kaplan-Meier estimate (KME) 的, 前面已大略提過。

KME 的故事

給一組「隨機設限 (random censored) 」的數據:

$$(\min\{X_i, C_i\}, \delta_i),\ \text{iid}\ i = 1, 2, \ldots, n$$

其中 $X_i \sim F; C_i \sim G, \delta_i = I_{[X_i \le C_i]}$ 。若 $\delta_i = 1$, 則我們只看得到 X_i 的值, 此外 $\delta_i = 0$, 我們只看到 C_i 的值。主要的目的是尋找 F 的估計量。

如果所有的 $\delta_i = 1$, 則我們能看到所有的 X_i 的值。此時, 沒有因設限而丟掉的值。那麼

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \le x]}$$

是最傳統的估計量。關於 $\hat{F}_n$ 的性質, 已經太古典了[3], 而這由二項分布和 CLT 及相關的大樣本理論就知道。簡單地看, 這是以 $\bar{X}$ 的形式出現的統計量, 能夠做的數學, 大概都已經做完了, 如果還剩下甚麼, 都難得要命。

但是對於 KME 而言, 因爲有幾個 $\delta_i = 0$, 有些 X_i 是看不見的 (實務上的說法是我們知道 $X_i > C_i$, 我們只看到 C_i)。KME 的出發點反而是風險函數 (hazard function) 。若 $X > 0 \sim F$ 是一個不會爲負的隨機變數 (例如壽命), 則

$$h(x) =: \frac{f(x)}{1 - F(x)}, \quad x > 0$$

叫做 X 的風險函數。利用微積分可求出

$$S(x) =: 1 - F(x) = e^{-\int_0^x h(u)du}$$

如果將上式中的積分用一段一段的求和來表:

$$h(x) \approx \sum_j h(u_j)\Delta$$

則有

$$S(x) \approx \prod_j e^{-h(u_i)\Delta}$$

是一個由「求連乘積」的方式來表現的。KME 的結構, 就是一個連乘積的形式

$$\hat{S}_n(x) = \prod_{t_i < t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i} \tag{1}$$

其中 $t_i = \min\{X_i, C_i\}, i = 1, 2, \ldots, n$ 。

公式 (1) 的源遠流長, 但之前的工作, 如涓涓細流, 可是到此之後, 氣勢一變。關於對設限數據求 $F(x)$ 的估計的問題, 甚至於一般地對於設限數據的統計推論問題, 都有了開朗的看法: 原來我們得從風險函數入手, 而不是從「算術平均數」的角度來看問題! 這

[3]意思是, 別在裡面找題目了, 因爲幾乎都挖光了。

就像人類一直都在用模彷鳥類的角度來設計飛機, 一直沒有成功過。直到萊特兄弟的定翼飛機機成功了, 人類才忽然懂得: 不能學鳥[4]。因此航空工程才能突飛猛進一樣。

Kaplan 和 Meier 的文章, 好像是忽然將問題的「任督二脈」打通一樣。自然後面就跟了一大堆補強、加註、推廣的文章。那是五零年代的末期, 搶做論文的風氣還不太烈。下面我舉的例子, 多多少少可以讓大家看一下怎樣去搶 (還不算太糟的) 題目。

有了 KME 這樣的公式, 甚麼時候可以大方地用它? KM 的 1958 論文中的條件是 random censoring , 這並不能一定保證, 在實用上能不能更寬些? 一個較好的結論是: 如果「設限」這件事不會影響到未來的壽命, 則 KME 的公式是可用的。用數學講, 就是

$$P(X \in [t, t+dt] | X \geq t) = P(X \in [t, t+dt] | X \geq t, C \geq t). \tag{2}$$

這樣的條件有明白的直觀意義, 比 randon censoring 要寬鬆[5], 這是蠻有意思的結果, 當然值得發表[6]。

但這樣的結果就沒有問題? 記住,「找題目來做」的技巧, 有一部分就是挑毛病。要挑毛病, 即使別人沒有毛病, 也要努力去挑。有人問: 這樣的條件 (2) 是不是可以從所給的數據裡檢驗呢? 這不是無聊的問題, 即使答案是負面的, 也有它的意義。答案眞是負面的, 當然能講清楚這一點, 也是可以發表的[7] 。這個結果當然重要, 因爲它, 我們只能利用數據的背景來討論是否可以用 KME 。因此, 取樣前的行政作業規範, 就變得重要。

你可以更進一步, 挑出 KME 估計得很差的情形來說[8]。這樣的論文, 當然需要 KME 已經被公認重要時才有價值。否則, 我隨便發明了一個估計量, 或者一個根本用不到的估計量, 好或不好都不重要, 有甚麼好討論!

較容易問的問題是它的數學性質。這裡有一大票論文[9]。最直接的是 normal approximation[10]:

$$\sqrt{n}(\hat{S}_n(\cdot) - S(\cdot)) \rightarrow \text{ Gaussian process}$$

證明某統計方法數學性質的工作是可以做的。但愈來愈難做, 愈來愈難發表, 而且學術上的影響愈來愈小。那樣難, 除了那幾個不太識人間煙火的象牙塔中人, 誰看哪[11]? 當然, 如果你眞的快手快脚, 又有幾下獨門絕招, 這遊戲還是可以玩的。

做這類學問, 需要效率。你需要眼明手快 —— 眼到、手到、心到。這是因爲有極大一批學者在幹類似的事[12], 因此只要有一點有氣味的風吹動, 這批人 (包括他們的學生) 就可能撲上來搶做。你若是慢絲條理的, 這口飯就有點難吃。

跟別人搶快的另一個條件是題目紅火的程度。一個好專題在開始的時候, 尙待開發的部分還多, 你只要夠勤快, 機會也會有。過了一陣, 這個專題中容易做的部分很快地被

[4] 另一個原因是當時的工業水平不夠, 機械效率太差。

[5] 並且, 很多其它的設限方法, 在這樣的條件下也可以用。

[6] Kalbfleisch and MacKay (1979). *Biometrika*, 87-90.

[7] Tsiatis (1975). *PNAS* **72**, 20-22. PNAS 指的是美國科學院的期刊, impact factor 是很高的。

[8] Peterson (1976). PNAS, **73**, 11-13. 抓到名人行爲不檢, 也是有市場的。

[9] 這我就不一一列舉了, 找一篇較近的關於 KME 在 Annals 上的文章, 從它引用的文章裡向後找, 總能找到一大群論文。

[10] Breslow and Crowley (1972). *Ann. Statist.*, 437-453. 到了現代這種題目已搶不到了, 因爲作者若是自己不證, 多半論文就登不出去。研究工作者的日子是愈來愈難了。

[11] 又一說法是: 很多期刊是掌握在這類學者手上的, 因此這類論文反而更有機會。

[12] 這是支持學術期刊的主流人群, 因爲要做有眞貢獻的工作不易, 做看起來有貢獻的且又挑不出毛病要容易多了。數學暗示困難和嚴謹, 這至少就是 non-trivial 。

一掃而空。也許這個問題已經被充分了解, 學術期刊的編者會覺得這個題目不再有趣並且還有更重要的熱門專題要討論。聰明的人就會考慮下車, 尋找下一個目標 —— 以上是「業績掛帥」的幹法。對學術不是沒有貢獻, 但一般沒有大貢獻。愛因斯坦說, 他最不想見的, 就是一大堆學者拼命在軟木頭上鑽洞!

當然, 即使在問題的晚期, 你也可以做出一些能發表的結果。但僅就投入/產出比來看, 是比較不合算些。有時你也可以做出相當重要的結果出來: 如果 A, B, C . . . 等作者, 在討論別的部分都很清楚, 但到了某處你想弄清楚的地方, 言詞就東閃西躲沒一句有氣力的話時, 你就知道那是一個需要突破的關鍵點: 如果你能將它搞清楚, 做得出 (包括證明做不出), 總是會有好處的。

先拿到數據再做問題

首先我得說拿到新數據並不容易。工業上的數據難拿到, 因爲廠家有業務機密要保護[13]。其他科學家辛苦由實驗而得的數據, 更是他們的寶貝[14]。而過去已發表在公共領域 (public domain) 的數據, 多半是原始的資訊, 都已榨乾。拿來驗證你的新方法可以, 但對於原來的科學問題, 已貢獻不大。因爲別人的功力也不會差, 你不見得會分析得比前人要好。

較好的數據是來自你參加別人的研究團隊。這樣, 你就在問題的早期就切入, 參與數據的規劃蒐集, 和專業人士直接對話。何況眞實數據, 即使是經由統計學家的設計蒐集, 也很難不會發生問題: 取樣的獨立性不能保證, 資料未必完整, 套不進現有的模型, 殘差的變異過大或者全然不像是常數 . . . 。只要你投入, 就會發現教本上的技術都用不上, 而這就是你發展新工具的機會。

這是一個統計學者的正途: 在別的科學研究裡找到研究的生命力。一開始要打入別人的團隊較難, 但等到你漸漸有了聲望、名氣和口碑, 上門求教者會讓你有做不完的項目。但你要注意到別做簡單的工作, 那樣你只是別人的雇用人員。已故的名生物遺傳統計學家李景均[15] 說: "I don't do chi-square", 意思就是說: 要挑嘴, 不要甚麼都吃。

最糟的事是: 辛苦做出一點結果, 投稿時評審問: 有沒有 real data? 你東翻西找, 哪裡可以找到剛好可以套進你的新模型的實數據? 這在你做純理論的工作時還好, 至少還可以模擬。在你做號稱應用的統計問題時, 這就夠你喝一壺了。初入行者最易犯此病, 切記切記!

有眞實數據的論文, 其實是比較有意思的。所謂數據, 最好有重要的學術上的或者生活上的背景。例如你如果弄得到中國衛星掃描而得的數據, 或者奧運選手的訓練數據, 就算你用較淺的統計方法來分析, 只要基本正確, 英文再清楚些, 發表不難。如果是神六數據或者關於青藏鐵路的數據, 當然就更別說了。

數據最好不要太標準。如果套進迴歸軟體就有好結果, 這樣的論文誰會要? 你需要複雜些的、有結構的、可以至少用好幾種近代模型的、最好背後還有一些有科學意味。一開頭, 模型簡單些不妨, 然後漸漸放寬條件, 做較複雜的分析 —— 有一組好數據, 可以玩好幾年呢。下面我舉一個自己的例子。

[13]一種是眞的機密, 競爭對手可能因此而得利; 另一種是假的機密, 他們其實是在用簡單的手法, 但也不想讓外人知道。

[14]實驗室間的競爭, 因爲和項目經費有關, 也很厲害。

[15]李景均 (C. C. Li, 1912-2003) , 原是北大教授, 1949 年後因爲規定要用唯物的觀點教遺傳學, 跑到美國。寫論文固然重要, 執著於學術的本質, 才是眞科學家。前人遺風, 心嚮往之。

火工件的故事

火工件指的是有火藥會爆炸的東西。爆炸通常由某物的引發, 如通電。電有大小強弱, 因此引發與否, 是和 stress 有關的。這問題是和計量響應 (quantal response) 是一類的, 只是火工件的對象是爆炸, 後者是白老鼠餵藥[16], 看幾週後腫瘤是否夠大。

對於計量響應, 有興趣的部分是分布中間的部分: 多少藥量, 可以讓 50% 的老鼠產生夠大的腫瘤[17]? 對火工件, 有興趣的部分是分布的尾部: 多大的電流, 可以讓 99.99% 的火工件都引爆?

我能拿到一組眞實數據是因爲手上有一個項目。關於 binary response 的數據, 至少有一甲子的研究了, 書都有好幾本[18]。因此在一九九幾, 理論上可做的研究不多[19]。

有了數據, 如何寫論文? 應付原來的委託單位 (client) 不難, 我們只是把近 60 年的東西補足了, 告訴他們到目前爲止最好只能做到哪兒, 再出一册中等長度的報告就好了。不論怎樣, 原委託單位只是靠一本美軍手册, 而那手册主要的理論根據是 1948 的一篇論文[20]。我們先將問題套進一個 Markov chain 中 (看起來就較深了, 因爲沒有 independence), 然後再發展理論, 用自助法 (bootstrap) 來做所要的信賴區間, 並爲了提高模疑的效率, 引入 important sampling 的觀念, 搞出一個自己叫 important re-sampling 的法子來做。經過模擬, 用傳統大樣本方法所得到的 95% 信賴區間大約只有 90% 左右, 而我們的法子眞的差不多做到 95%[21] 。

這有甚麼了不起? 學問上是沒甚麼: 論文裡很多部分的技術都是在文獻裡找得到的[22], 但我們有眞數據[23], 眞問題, 發現老方法的不足, 並合成一些較現代的方法 (bootstrap 和 important sampling) 來改進老問題, 這些「 功勞」就足夠我們覺得可以寫一篇不錯的論文了。這就是先有數據再動手研究的好處。

這只是一組小型數據的故事。但如果你參加了一個 3-5 年的計劃, 其中有 5000 份問卷, 分三年完成, 那你可以拿來玩一下的數據就夠用好一陣子了。

需讀書時要折節

關於科學意義, 如果這個科學有較大而長期的潛力, 你就該考慮用心補上一些足夠的背景知識。例如生物遺傳是近年來最紅的專題, 就有些極好的華人統計學者, 自己找生物教本來弄清楚基本生物學 —— 因爲這樣你才能面對著生物學家來談他們的問題。更重要的, 你較會知道他們的問題裡面, 生物學的意義有多少。

氣不可衰

一般論文指的是雖不見得是最好的, 但最後還是在最好或中等以的期刊上能發表的論文。因此, 請勿小看它—— 沒有這許多一般論文, 也襯托不出那少數幾篇突破性的結

[16]藥量就是 stress 。

[17]這叫 LD-50 。

[18]國內有劉寶光 (1995):《敏感性數據分析與可靠性詳定》, 國防工業出版社。這書蠻實在的。

[19]Jeff Wu 有幾篇從 DOE 的觀點來做 binary response 的文章。這也是選題的一法: 把腳硬套進鞋裡, 套得進就是結果。

[20]Dixon and Mood (1948). *JASA* **43**, 109-126. 所以, 你該有一點懂, 所謂「機密」, 有時會是甚麼事吧?

[21]Chao and Fuh (2001). *Statistica Sinica*, **11**, 1-21.

[22]其實, 論文大體都如此, 哪能全都是新的? 有一兩點新意就不錯了。

[23]指首次使用。

果。古人說千軍易得, 一將難求。像岳家軍、戚家軍這樣的精兵是可以訓練出來的, 但岳飛、戚繼光卻是天生的將才, 這種菁英自己會成長, 用不著我們去擔心。研究的結果也一樣。好結果難, 多半的結果都是一般的。但它們做了一件重要的事: 把問題週邊的枝枝節節都清掃乾淨, 因此留下未做的部分都是較難啃的部分。

因此, 即使你也在做這類「搶食大餅」的工作, 也不必輕看自己。因爲你畢竟還有些貢獻。何況這類工作並不容易, 世上的高手如雲呀。

做研究除了本身有趣之外, 就是還得撐著一口氣別放鬆了。做一些大家都做得來的工作並沒有太糟, 但你不能一輩子總是這樣幹。

你也許在想, 等我升到教授, 就改走行政路線, 系主任、所長、院長 ...一路幹上去。話是不錯, 的確有人的生涯規劃是這樣的。但是世事難料, 別忘了在大學, 教授是本職, 其它行政業務都是兼任[24]。如果你到了五十歲, 爭取某一個長失敗, 前一個長任期已滿, 不得已還是得回到本來的系, 那時候年青的學者已趕了上來, 正是「長江後浪推前浪, 前浪死在沙灘上」。這時, 若是沒有一些有功底的作品, 你在那個單位裡如何走路有風?

在學術單位裡, 學問是一輩子的事。

做較難的問題較有成就感

這話當然不必我說就懂。你可以做幾篇或甚至幾十篇搶快型的文章, 但心中最好還是有一兩個較難的、希望能突破關鍵的問題擺著。在某一個專題久了之後, 週邊的問題都搞清楚了, 偶然來一個突破還是可能的。

我一般不鼓勵年青學者做大猜測 (conjecture)[25] 或有名的未解問題 (open problem) 。小教授提出的 open problem , 學術上的份量也許不夠, 做出來功勞也不大。大教授提出來的, 往往眞正困難。他做不出[26]多半你也做不出。下面的問題是 open 的, 感謝林國棟教授給我這個題目。

給一個小題目

令 $F(x) = (1 + e^{-x})^{-1}, x \in R$ 。這是所謂的 logistic 分布的 cdf 。取 $X_1, X_2, X_3 \sim$ iid F 。再令

$$X_{(1)} < X_{(2)} < X_{(3)}$$

爲它們的次序統計量 (order statistics) 。取

$$U = X_{(2)},\ V = \frac{1}{2}(X_{(1)} + X_{(3)})$$

則可以證明 U, V 有相同的分布。但這個命題的逆定理是否成立? i.e., 若設 $X_1, X_2, X_3 \sim$ iid F , 取 U, V 如前, 如果 U, V 的分布相同, 問 F 是否恰好是 logistic ?

這題目我 不建議你去做, 道理很簡單: 林教授在這類問題上至少有二十年功力, 他若是還做不出, 肯定是難的。當然, 你若是能嚴謹證明, 發表應該是沒有問題的。

新意

[24] 起碼台灣是這樣。

[25] 好在統計學裡沒有甚麼大猜測可以讓人一夜成名。

[26] 多半他的門人弟子也做不出。

競爭的目的是汰弱留強或者變弱爲強。光是要求 SCI/SSCI 期刊, 已相當能夠汰弱了。這些期刊對於國內學者而言, 大都沒有甚麼人情可講, 如果論文沒有太多的新意, 英文講述能力再差一點, 這對於國內上千個大學裡的學者教授, 已經是較難的障礙。這裡面, 最難的部分是新意。新意有大有小。好論文有大新意, 前已述之。一般論文多半只有一點點的新意。

你怎麼知道你所做的東西是新的[27]? 主要的是: 你必須知道你現在做的研究專題的現況, 才能知道你想做的方向、你的點子 (idea) , 是不是已經被別人做過了, 是不是做得夠好。

這在大家不怎麼重視論文發表時[28], 情況還好。但現在大家一窩蜂要求發表論文, 競爭的對手太多。別人不見得笨, 壓力不見得小, 研究資源和環境不見得差, 你看得出想得到的點子, 別人也多半能看得出想得到。如果問題的難度不大, 除非你的速度比別人還快, 你的機會也不見得有多高。國人所佔的便宜是數學訓練較強。這在以前是大便宜, 但現在太多的華人學者在搞統計 (和相關學門) 的研究了, 這些人也不見得人人在前五名的期刊上發表, 因此競爭下來, 國內學者也未必在「數學較強」上佔優勢。

所謂「新意」, 因爲人人都在要求, 所以一般對於新意的要求, 其實也沒有那麼嚴: —— 你只要能清楚說明你的新意在哪裡就是。重點是你得說得明白, 而且在論文的前兩頁就說明白。

緒論的重要性

這裡說到了論文中最要緊的部分: 摘要 (abstract) 和概論 (introduction) 。摘要因爲短只要說重點就好[29], 問題不大。概論則不然, 這是最難寫的。

當學者埋頭做問題的時候, 往往是當他循著某些思路一路向前, 尤其是在有一點靈感的時候。如果你做到某一個深度, 做研究基本上還是有趣的事, 有時顧不得去找是否別人有類似或相反的想法。但這在論文的概論裡, 你就得在兩三頁的範圍裡講清楚這個題目從第一篇論文起到目前爲止的概況。概況不是目錄, 你得說出 (用較乾淨的英文) 誰做了甚麼, 結果有沒有影響, 誰貢獻最大, 誰又跟在後面但也有功勞, 這些與你目前論文的關係 —— 這一切只是要表現你是如何的專業。要用明白的話說明你的貢獻在何處, 但又不能用自己誇自己的句子。這一方面是禮貌, 一方面是規矩: 論文要用大心機寫, 但得表現出平常心的樣子。平常心, 指的是客觀和不帶感情。

這裡面學問可大著哪, 一時也說不完。當你漸漸有名而在較好的期刊上投稿時, 別忘了你面臨的評審者都是這一行的專家, 並且, 更重要地, 他們都是你的競爭對手。「行家一伸手, 便知有沒有」, 你拉出來的架勢, 已差不多決定勝敗。因爲大概由這裡已可看出你是一個青澀的菜鳥還是一個戰鬥經驗豐富不可輕侮的高手。何況這裡面還包含著學門和學派間的恩怨[30], 因此有些文章你是要捧的, 有些是你是要罵的。捧的時候你要維持做一個學者的尊嚴, 要恰到好處而不帶諛詞[31], 罵的時候要尖酸刻薄不留餘地地突顯別人的缺點但還得顯出溫柔敦厚純學術本位的樣子。你以爲只要結果好就沒有問題, 是

[27]一個方法是先查一下 Kotz and Johnson 的 *Encyclopedia of Statistical Science*, 計九冊, 是一個蠻中肯的開始點。可找到很多專題過去的關鍵論文。再去反查 *Citation Index*, 便可回溯到近年。這兩種資料都貴, 要靠一個好圖書館。至於現在別的競爭者的工作, 那就全靠你自己的本事了。

[28]如五十、六十年代。

[29]據說人大要求學位論文的摘要至少三千字, 是有點多。這樣, 誰還讀論文的本文哪?

[30]你以爲拒絕過某人或某人學生的文章, 別人眞的不知道? 學術圈子其實很小, 有些人眞的不怎麼上道。

[31]有一次某文章提到我的某論文, 說我做了 A 又做了 B 。但我確實沒有做 B , 反而因此拒絕了那篇文章: 他沒看懂我嘛!

的, 眞正的好就大概就沒有問題。一般的好, 還得看江湖上的朋友是不是肯高抬貴手。

同理, 別人文章的緒論, 也得好好讀。第一次不妨馬虎些, 讀完全文以後, 如果你覺得這文章還有點意思, 再回頭讀一次該文的緒論。問自己三件事: 第一, 這文章的主要賣點在哪兒? 第二, 作者怎麼想出來的? 第三, 現在你已經「懂了」, 請問若是由你來寫, 這緒論該如何寫?

這和研究學問的關係不大, 但和你的論文接受率有關。懂得化妝的女子, 找到愛情的機會大些。投稿相當於「求售」, 即使是和氏璧, 在未打磨包裝前都一再碰壁。文章不注意潤色就投出去, 等於穿得一身邋遢去做面試會談。—— 便縱有千種風情, 更待與何人說? 一篇論文的緒論, 和一個人的面孔一樣: 別人第一眼就要看的, 而印象好不好對下一步決定的取向, 可是有極大影響的。

要有所影響 (make an impact)

最好的論文開創新域。大如相對論, 對整個的科學投下重磅炸彈。小如前述的 information theory , 也開拓一個特殊的專題。其次的論文, 在突破一個重要的關卡之後, 前面是一馬平川。例如 Ito's lemma , 爲隨機積分舖下定石。再次是當某類問題已日漸成熟, 經過深思熟慮的梳理之後的集大成之作。

能做這類論文的學者, 根本用不著我們去擔心。但即使是一般論文, 也該極力地求新求變, 希望能對原來的學門產生影響。我們且不談眞正的科學貢獻[32], 只是爲了世俗的目的: 爲了能在較好的期刊上發表, 得到較多的引用 (citation) , 得到同儕的認可, 得到國家的獎助, 得到學術上的榮譽等等, 你就得出盡絕招, 求新求表現求有所影響。因爲「是不是有影響」才是我們對學術研究是否有意義的基本考量。

判斷一篇文章有沒有新意, 可從幾個簡單的角度來看[33]: 是否有新的資料, 新的觀點, 新的切入點, 新的分析方法, 新的發現。

判斷一篇論文是否有賣點, 也是看它是否能有所影響。因此作者要用心去強調他的論文確有影響。當然, 有沒有影響是在論文登出來之後, 別人才會去讀, 而且讀了後會去用才會漸漸地爲學界所知。因此目前誰也不能眞的確認, 也因此你就該誠實[34]地極力鼓吹[35]。

在論文裡不要謙虛。不要說自己的結果不重要, 儘管你心裡也許會覺得如此[36]。外國人常明白地告訴學生: "Be positive !" , 多講你的成功, 別說你的失敗。

懷疑論

做研究時的基本態度是誰也不要相信, 甚麼也別相信[37]: 例如一般的論文中號稱的好結果, 別人給出的 conjecture , 論文裡的定理和應用, 你的系主任, 院長甚至博導所告訴

[32]一般的論文, 本來貢獻就不可能大。貢獻大的都是好論文, 這是定義。

[33]以下幾點由管中閔教授的一篇講稿中引用。

[34]不要講你自己都不信的話, 專家還是很多, 還是要回到學術面。因此, 沒有狼, 不可以說狼來了; 但若有狼, 可以說大惡狼來了。

[35]中山先生革命時, 若是很謙虛地說滿清政府雖然腐敗但仍擁有極大的資源, 參加革命多半會死, 誰會陪他去玩命? 但若是將建國大綱, 五權憲法這些搬出來, 再說滿清是唯一的阻礙, 且已千瘡百孔, 則前景就完全不同。七十二烈士有幾人讀過三民主義的全文? 他們多是只聽到了孫先生的 introduction!

[36]這在你原來有一個大目標, 但只做出來一個次要目標時, 心中多半會覺得如此。

[37]這一段是利用林共進教授的講稿改寫的。

你的事。因爲他們可能會錯, 會不周延, 會省掉某些條件, 會條件太強, 應用範圍不廣。

爲甚麼要懷疑? 因爲如果這些人都是對的, 都已做到最好, 那麼你除了趕快換題目之外, 還能做甚麼? 因此你必須用有色的眼光來讀別人的論文, 用懷疑的態度來看科學的命題[38], 用別人沒有問過的問題來拓展自己的深度。題目要自己挑, 因爲別人給的, 可能眞是殘羹剩飯。

但天底下甚麼題目都能做, 甚麼都不能做。明明是死棋, 你也可能走活; 明明是活棋, 你也可能走進死胡同。信心和研究能力都是靠跌撞撞培養出來的。

再說懷疑論

這等於是上一段話的 corollary 。因爲別人論文的好處別人自己會說, 因此他們不提的部分 —— 也許是忘了, 也許是沒有想到, 也許是故意[39] —— 才是你該注意的。Kaplan 和 Meier 沒有說在隨機設限之外, KME 是否可用, 這就成就了 Kalbfleish 和 MacKay 的文章; 而 Kalbfleish 和 MacKay 也沒有說出他們的條件, 是否可由 KME 的數據來檢驗, 這就成就了 Tsiatis 的文章。

這些文章的題目都由懷疑而來, 這類的結果, 不是挑錯[40], 而是補強、拓展和更清楚的說明。因此即使文章交給原作者去審, 多半也不會引起反彈。

September 4, 2006

[38]例如在某演講裡聽到某問題可以做或者是 open 時, 就要多考慮: 多半是雞肋。

[39]包括他自己還留一手預備寫下一篇論文。

[40]理論上, 如果你說某人的結果是錯的, 他應該「聞過則喜」才對。可我就聽過這類文章被一票人圍堵的故事。江湖中, 固然有俠士, 也有拉幫結派的黑道。

第四部分: 動手作文

沒有研究結果就沒有論文。有了之後, 寫論文只是將你的研究結果, 清楚地告新別人而已。對學術論文而言, 你的讀者是學有專精的人士, 你的結果是實實在在的科學結論, 因此你不是寫詩, 不是寫散文, 甚至不是一般的論文[1]。學術論文要說得清楚明白, 沒有不相干的話, 基本上要以容觀的立場敍事, 而且遣詞用字以不帶感情爲宜[2]。

邏輯次序

假如你已得到 A ,B,C 三個結果。你在做研究時, 因爲你對於問題背景的熟悉, 思想方式可以跳來跳去。但是讀者雖是專業人士, 卻並不是你。因此你得將結果的邏輯順序弄清楚。

簡單地可以這樣想: 如果你要將結果講給學生聽, 你該用甚麼次序? 是不是先弄懂 B, 才能弄得懂 A? 在 C 中所用到的的幾個名詞, 在甚麼地方引入最好?「你的學生」是還不錯的模擬對象。因爲他們有點懂, 又不眞懂。至於讀者的程度, 可以假設他們有一個計量方面的博士學位。

先粗具規模

寫論文也是作文, 但目的是講故事: 將你的研究結果講給別人聽。因此, 成果第一, 故事第二。先將成果依邏輯次序列出來, 然後用文章將它們連起來。將想講的其他各要點一一列出, 試著替它們安排適當的出場點。前幾次的稿子可以疏漏, 但你得一遍一遍地潤色, 到最後打磨光亮爲止。

這樣的方法, 適合於你自己用電腦軟體來打自己的論文。因爲在軟體上修改和打印都十分容易, 所以你才能一次一次反覆地修改到滿意爲止。這和幾十年前, 你要把手稿交給打字員完全不一樣。因爲那樣一來一往, 至少一個星期。我們要儘量發揮近代科技的優點。世上有些學者的本事大, 他們第一遍的手稿就和最後的定稿差不多。如果你有這樣的本事當然最好; 如果沒有, 不妨學我。自從開始用軟體寫作以來, 一篇論文 update 十次, 是常有的事[3]。

初具規模就去動手拉開文章的架勢的好處是: 它讓你「下決心去寫」這個步驟, 變得容易[4]。

萬事起頭難。若要不難, 胡亂起一個比總不開始要好。

從第二節開始

寫論文要從第二節開始寫, 用心讀論文也是從第二節開始。何以故? 因爲一般的論文, 第一節的緒論只是誇說自己的結果多麼好, 別人是如何有缺點而已。眞功夫在後面

[1]例如報紙上的社論, 或者《古文觀止》上的那種古典論文。

[2]有時, 像「感謝某人的討論」、「慶祝某人七十生日」之類的句子是可以加上的, 但不在正文。在正文裡要一本正經討論學問。

[3]當然, A4 的紙就用了不少。可是這是你項目經費裡最該花的錢, 並且不能算貴。

[4]我做學生時, 老師的教導是: 沒有好結果就別寫 —— 寫爛文章是丟人的事, 有尊嚴的學者是不做的。但現在政府如此推動發表論文, 大家有甚麼辦法? 我現在只能主張大教授別甚麼論文都寫, 總要有點品味罷?

(有時甚至在某一個放在附錄的 lemma 裡)。緒論需要用心化裝舖排, 但化裝舖排也會掩蓋缺點[5] —— 要眞懂, 還得自己去用心讀。

最後才去寫緒論。緒論 (introduction) 指的是「介紹」—— 將你的內容介紹給讀者認識。因此先有內容, 才能介紹。

基本格式

論文的基本格式, 大體相同。雖然各期刊都會有一些要求, 投稿之前, 去找一本近期的目標期刊, 讀一下它的「 Instruction to Authors 」, 就差不多知道了。你其實還得去讀一下該刊的「編輯方針 (scope) 」, 他們多半會說明他們喜歡哪一類的文章。

論文多半以節 (section) 來分, 而不是以章 (chapter) 來分。章比較長, 在專書裡常用; 而節較短, 因爲一篇論文, 一般只針對一個專題。

節之下可再分小節 (subsection) 。有些人還要分小小節 (sub-subsection) , 我覺得是太細了[6]。好一些的論文, 在某節節尾, 常會說下一節會討論甚麼 —— 語氣是隨著本節的內容來說的, 給人的感覺是: 每一節都有 motivation , 一個忌諱是: 某節的標題後面, 沒有說明就加上小節的標題, 如

4. Estimation Procedure

4.1 Estmation equation

下面的寫法較有樣子:

4. Estimation Procedure

At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here.

4.1 Estimation equation

不少人不懂得這事, 雖然現在已不那樣講究, 但是一連兩個標題連在一起, 當中甚麼內容也沒有是蠻難看的。另外, 節或小節的內容都不要太短, 因爲印出來的版面也不好看 —— 好期刊是非常重視他們的版面的, 因爲學術論文的目標是「藏諸名山」, 如果不夠講究, 就表示他們自己都不在乎自己的內容。而學者們肯去投稿, 除了升等之類的壓力外, 另一個原因則是「好期刊是要被好圖書館收藏的」—— 這和在網站上隨便貼一下完全不同。

至於圖和表在文中該放在何處, 才較好看, 你不必去擔心, 因爲這是排版者的專業。你不妨將圖和表都放在文章的最後, 只要在正文中用一行字說

(Insert Table 3 about here)

[5]記住, 誰也別信。
[6]如果有 3.4.4 節, 多煩哪。

就可以了。

文章是給人讀的, 因此要有次序。學術論文也是要給人讀的, 但讀分兩種: 粗讀和細讀。

你的論文摘要, 要能讓讀者有興趣進一步去讀你的緒論。你的緒論, 應該能讓讀者去粗讀你的正文。至於細讀, 只是針對那些可能用到你的論文的人和你的競爭對手。刊登論文的主要目的是留不紀錄 (documentation) , 是先驗證過 (所以要學術評審) 再給人查的 (所以圖書館才收藏)。因此別太希望有人會用心細讀你的論文, 但你得希望有人會粗讀它, 抓住你的結論。因爲極少人會從頭到尾將每一個字都讀到[7]。

一個方法是以論文的主體作爲讓人粗讀用, 而將較細膩的細節, 放到附錄。如果期刊是有聲望的, 多半的讀者會覺得附錄中的東面是評審查證過的, 因此, 除非必要, 他們也不會細看。

同理, 不要把你的每一招每一式都寫出來。我有一篇文章[8] , 其中的一個公式的手稿就有 100 頁, 而我算了四遍。但寫入文章, 不過九行。我只是淡淡地說:

The computation, although tedious, can in fact be carried out. (以下爲推導的概略方法及最後公式)

沒有這個公式, 這篇文章大概不會接受。但若我寫上 20 頁, 恐怕更難被接受。

平等的精神

在科學之前, 人人平等。這表現在論文裡, 就是一般在我們提人名時, 不提他們的職位、榮銜或者尊稱。我們說 Smith (1999) , 指的是此人在 1999 年所發表的一篇論文或者專書。但我們不提此人是大教授、小教授、新出爐博士、某科學院的院士、還是某國國王所封的爵士。因此, 文章中別用 Dr. Taguchi, Professor Ito, or Sir Fisher[9] —— 要讓他們的文章著作自己說話。即使是自己的老師, 也是一樣。文章裡不作興拉關係, 說誰是我的師叔, 誰又和我在某會議中相談甚歡云云[10]。雖然, 圈子裡對大家的關係都略知一二 —— 你以爲大家去參加國際會議只是去聽別人演講?

這並不是說我們不承認好文章的權威[11]。但別人的 credit 要靠他們的論文來展示, 不是靠頭銜。「平等」的另一個意思是: 至少在表面上, 你的文章是和大家站在同一標準來評量的, 不因你是新手, 就該「讓幾個子」[12]; 也不因爲你聲名赫赫, 就不敢懷疑你的結論[13]。

嚴謹的態度

在所有的嚴謹中, 最要緊的當然是學術推理的嚴謹。當它有漏洞而被評審人員抓到

[7]現在的論文眞是太多了, 誰有那個美國時間哪?

[8]Chao (1987). *Biometrika* **74**, 426-427.

[9]這些人拿到的榮譽已經夠多, 用不著我們來湊熱鬧。

[10]但文章中可以感謝某人提供建設性的建議之類, 這類事該放在「感謝」這一小段, 不可放在正文中。且不宜感謝期刊的總編和副編, 因爲有拍馬的嫌疑。

[11]我們對好結果還是有敬意的, 文章中不妨適當表示。但別過分。

[12]有些期刊對博士論文較鬆, 這是他們的政策: 不要嚇壞了年青人。

[13]有些期刊採用「寄出評審時將作者姓名塗去」的辦法, 以免有名的學者有較高的機會通過。但這其實不易瞞過內行人。

時, 除非你能夠補救[14], 否則你只好摸摸鼻子走路。

其它的嚴謹就表現在許多小地方了。我們列舉一些。

英文要正確

英文的文法別錯[15]。該加上 s 的地方, 別忘了加上。單字別拚錯了[16], 因此自己細讀幾遍是必要的, 千萬不要以爲評審人會幫你一一挑出來, 他又不是你的博導[17]! 拚法對的字, 未必是正確的字。實在不行, 找個英文好的老外 (包括付費) 將它順一順。

國人的英文多靠上課所學。學得好的人, 寫出來可以讓老外能懂。但英文的精微細緻處, 沒有十幾年功力是不行的。比如說, 你知不知道, can not 和 cannot , 哪一個對? Disqualified 否 unqualified 有甚麼不同?

另外的要求是一致性。例如我在文章中說, 1998 年, A 做了 A1 ; 2002 年,B 做了 B1。這兩個句子, 動詞的部分, 可以用過去式 (98 和 02 都已過去); 也可以用現在式 (學術上的結果可以以眞理視之)。這兩種寫法, 現在在論文中都可以用。但是, 在同一論文裡, 不要在某一句裡用現在式, 又在另一句裡用過去式[18]。這是文法上的不嚴謹。

關於英文, 寫到此處。我推薦一本極好的小書。當年我極爲受惠, 而且至今還是經典。網路上一查便得。

William Strunk, Jr. *The Elements of Style.*

我只是將它的目錄列在下面:

Elementary rules of usage

Form the possessive singular of nouns with 's
In a series of three or more terms with a single conjunction, use a comma after each term except the last
Enclose parenthetic expressions between commas
Place a comma before and or but introducing an independent clause
Do not join independent clauses by a comma
Do not break sentences in two
A participial phrase at the beginning of a sentence must refer to the grammatical subject
Divide words at line-ends, in accordance with their formation and pronunciation

Elementary principles of composition

Make the paragraph the unit of composition: one paragraph to each topic
As a rule, begin each paragraph with a topic sentence; end it in conformity with the

[14]如重新證明, 換一個較嚴的條件, 或者證明評審錯了。

[15]我們假設你是用英文作文 —— 你要 SCI/SSCI 是嗎? 雖然, 中文的論文, 對文法的要求是類似的, 你只是沒有問題而已。

[16]現在的軟體都有 spell check, 一定要用!雖然用過之後還會有錯, 但要好得多。

[17]有一個故事是這樣的: 某生在他的論文某處加上「若老師讀到此處, 請找我, 有 XO 一瓶爲謝, 以感謝老師的用心細讀。」—— 據說 XO 還在。

[18]即使兩句是隔了 20 頁也一樣。這個毛病在作者參考了幾篇別的論文後再寫緒論時常犯。你以爲洋人都是對的。不錯, 他們分別都是對的, 但你一合起來寫就錯了。

beginning
Use the active voice
Put statements in positive form
Omit needless words
Avoid a succession of loose sentences
Express co-ordinate ideas in similar form
Keep related words together
In summaries, keep to one tense
Place the emphatic words of a sentence at the end

A few matters of form

Words and expressions commonly misused

Words commonly misspelled

其它細節

細節是常忘但忘了的結果就是 sloppy 的東西。我列舉一些於後。

一字一義

在全文中, 一個名詞指一件事。不要在第二節裡, 你的意思是 A , 到了第四節, 你又指 B 。爲了英文顯得活一點, 你在文章中可以用一些同義字, 但不要在有技術性的字彙上搞這一套。

同理, 一個符號只代表一件事。如果你用 $f(x)$ 來表示 X 的 pdf, 那麼後面的文章裡, 你就不能用同樣的 $f(x)$來表 Y 的 pdf , 除非 X, Y 是同分布的。理論上, 全文的符號都要一致。這毛病在你組合不同時期的手稿時會犯, 自己往往都不會覺得。

新詞要交待明白

所有較難一點的詞句, 在第一次出現時都要給定義[19]。不一定是非常正式的定義, 但要能說明白。

相互引證的事情要一一對應

例如正文中口提到 Wang (2002) , 那麼在最後的 References 裡, 就該將該文 (或該書) 列入。反之, 若後面的 References 中有 Wang (2002), 則正文中某處需提到以文章或書 —— 兩邊誰多一個誰少一個都是不嚴謹。這原則也用在同一文章中的圖和表上。若有圖三 (或表四), 則正文中需有某處明示「見圖三 (或表四)」。

有些較好的期刊甚至要求你只能引用有第幾卷、第幾期的論文或專書[20] —— 某某單位的技術報告的不行的, 因爲它們往往沒有經過學術評審, 份量不夠 (也許竟是錯的[21])。至於網站上的資料, 現在尚無定論。我自己的期刊, 主張用一個 footnote 來說明出處就

[19]請自行判斷何謂「較難」。例如 $e, \pi, E(X), Var(X)$ 可以不說明便逕自使用, 但 ϕ ($N(0,1)$ 的 pdf) 就可以考慮說明一下。這雖然已十分通用, 但尚未 100% 被認可。

[20]而且最好不要是不見經傳的小書店的專書。

[21]這就是前面所說的「誰也別相信」的另一種表現。

好, 暗示這類的引證, 份量要差些。並且, 網站有時會消失不見, 較沒有「立此存證」的意思。

至於文章中用到甚麼電腦或軟體, 以前是有人會提的, 但現在已不重要。因爲現在幾乎人人都電腦和軟體, 就傾向於不提了。這些都已是必需而不希奇的工具了, 和原子筆、A4 的紙一樣。誰在乎你用甚麼牌子的原子筆哪?

事事來歷明確

論文用字, 如老杜的詩: 無一字無來歷。自己發明的, 就要明白說出是自己發明的[22], 若不是你發明的, 你得告訴別人是誰做出來的 —— 別打馬虎眼。這裡若是交待得不明白, 將來說不定會有人來檢舉「 抄襲別人的研究成果」—— 這在學術界是大忌, 那時說不定教授的資格都被吊消。

來歷明白的更進一步的意思是: 你需要適當地給 credit —— 一般是引用最早的有突破性的、關鍵性的論文。這是你表現功力的時候: 有些論文是有錯的, 有時, 這些錯內行人是知道的。若是你不查或根本不知道錯而引用, 遇到內行的評審, 效果適得其反。

在你說明甚麼還沒有做出來的時候, 更該用心。文獻蒐尋 (literature search) 本是 research 的一部分, 甚至是較重要的一部分。這是初入行者的通病, 尤其在他們附近沒有一個好圖書館時, 更是吃虧。

不可跳號

論文中的公式、圖、表都有編號。通常各期刊都有自己的規定, 參考一下就好。你的編號系統也許和該期刊不同, 這一開始沒有關係, 多半在論文要排版時, 講究的期刊會要求你改 (或他們幫你改)。這些不會影響到你的論文接受與否。

但是跳號就會。例如你原有表一, 表二, 表三, 後來你自己刪掉表一, 那麼, 刪掉後原來的表二, 就該晉升爲表一, 原來的表三, 就該變爲表二。如果你不改正, 文章中只看到表二表三, 但不見表一, 這樣的跳號, 明顯地表示你不認眞用心。

拉丁文

爲了表示有學問, 有些拉丁縮寫是可以用的。例如

i.e., viz., ad hoc, et al., per se, ibid., e.g., ...

但是, 別用錯了 (尤其是標點, 甚至要用斜體字)。

不要隨便就出來一個縮寫

其它的縮寫, 應在第一次出現時給出全名。例如「maximum likelihood (ML)」,「generalized linear model (GLIM)」。

總之, 英文並非你的母語, 英文不好是不能避免的問題。下面的法子有所幫助:

[22] 而且還得大聲說, 生怕評審和讀者看不到, 漏了你的功勞。

- 寫不了長句子, 就用短而文法不錯的[23]。

- 不會玩花樣就別搞花樣。

- 文字上既然贏不了人家, 就要以內容取勝。搞清楚你的賣點在哪兒, 別人的缺點在哪兒。科學畢竟不是文字遊戲, 雖然, 文字好還是佔便宜。

- 實在不行, 先用中文寫出來再翻譯成英文再改。

- 對於技術性的論文, 其實需要用的句子不多。讀別人的論文時, 不妨將別人的句型抄下來。若有 10-20 條樣版句型, 也就差不多了 —— 畢竟我們不是寫小說、散文或者劇本!

強調緒論

我們一再強調, 一篇論文中緒論最重要。從編輯和評審的角度來看問題[24]: 第一印象大體上就決定了他是用正面的角度還是用負面的角度來處理你的文章。他是正面地幫助你呢, 還是負面地快快幹掉你的文章?

我通常在兩個小時以內就有了看法 —— 不論我懂還是不懂論文的內容。當我是正面的時候, 我才會用心地設法了解內容, 這篇文章就有了機會。一半以上的時間我是負面的。如果小錯誤一抓就一把, 你會覺得這文章會好嗎? 我覺得我還算是夠負責的讀稿人。當我挑到第十個錯 (大和小) 時, 我就會停住了 —— 這個作者已告訴我他不用心。光是 sloppy 這一個缺點, 就夠拒絕它了。

最後建議

寫稿及投稿前應去跑一跑學校的專業圖書館。親自去瞧一瞧你的目標期刊是甚麼樣子的[25], 別人的文章是怎樣的? 副編都是甚麼人[26]? 這樣, 你比較會知道你的文章是否適合這個期刊。

至於它們是不是 SCI/SSCI , 上網一查就知。大體來說, 有名氣有歷史的學會和大學所支持的期刊, 多半是有水準的。現在有很多由某書店支持的期刊, 雖也有好的, 一般就較弱。

有些期刊上, 有時也有由編委寫出的告訴大家的應注意事項[27]。我再另外附兩個 power point 的講稿, 分別來自林共進, 管中閔教授。講的都是關於論文發表的事。

September 4, 2006

[23] 其實最好用長短不一的句法, 才不單調。

[24] 別忘了這些人都是較有名的學者, 他們的一個共同特點就是沒有時間。

[25] 否則你的文章是寄給何人, 寄到何處?

[26] 主要是去查一下這些人最近的著作是否和你的文章有關。

[27] 例如 M. Schminke (2004). Raising the bamboo curtain. *Academy of Management Journal* **47**, 310-314. 就有參考價值 (感謝劉長萱教授告訴我)。

第五部分: 且先讀幾個緒論

在這一部分, 我們在較好的統計期刊上挑了幾篇論文, 將它們的摘要和緒論抄下來和大家討論一下。我在挑選時沒有先讀過論文, 我只是跑到圖書館, 將架上最近的一期的第一篇論文挑出來。期刊選了四種: *Annals, Biometrika, JASA, JRSSB* 。因此不能算是隨機挑選, 但也不能算是 100% 的立意挑選。這四種期刊都可算是「第一排 (first tier)」的國際統計期刊。意思是, 如果你有兩篇文章在這類期刊上被接受, 拿下一個研究計畫甚至升等就有點把握了。習慣上, 主編往往會將他覺得較好的論文, 排在每期的第一篇。

回應前面我們對於「緒論」的強調, 我們試著來讀它們的摘要和緒論。這些文章討論的專題我們不見得懂。但是, 就是因爲我們不懂, 才值得細讀: 看看這些寫得好的文章的緒論, 有沒有讓我們覺得「 後面的文章有內容」?

Green 和 Maridic 的文章

這一篇登在 *Biometrika* (2006) **93**, 235-254. 這個期刊在 1901 年創刊, 已有 105 年, 是一個有悠久歷史傳承的學術期刊。標題是: "Bayesian alignment using hierarchical models, applications in protein bioinformatics" 。

這裡用了幾個名詞。Bayesian 至少大家都聽過。指的是將未知參數當作 random variable 而給一個 prior 的一套想法和做法。Alignment 意指對齊, 如士兵的排成直線。Hierarchical model 則是某一類的統計模型, 這裡從字面上看不清楚, 雖然可以想像, 反正是有點複雜的模型就是了。最後說到了應用, 用了 "'protein (蛋白質)", "bioinformatics (生動資訊)" 而個非常當紅的名詞。因此, 光是標題, 就十分有科學味了。

摘要是這樣的, 其上的註, 是我加上的。

An important problem in shape analysis is to match configurations of points in space after filtering out some geometrical transformation. (註一) In this paper we introduce hierarchical models for such tasks in which the points in the configurations are either unlabelled or have at most a partial labelling constraining the matching and in which some points may only appear in one of the configurations. (註二) We derive procedures for simultaneous inference about the matching and the transformation, using a Bayesian approach. (註三) Our hierarchical model is based on a Poisson process for hidden true point locations; this leads to considerable mathematical simplification and efficiency of implementation of EM and Markov cchain Monte Carlo algorithms. We find a novel use for classical distributions from directional statistics in a conditionally conjugate specification for the case where tne gometrical tansformation includes an unknown rotation. (註五) Throughout we focus on the case of affine or rigid motion transformations. (註六) Under a broad parametric family of loss functlsan optimal Bayesian point estimate of the matching matrix can be constructed that depends only on a single parameter of the family. (註七) Our methods are illustrated by two applicaations from bioimformatics. The first problem is of matching proteib gels in two dimensions and the second consists of aligning active sites of proteins in three dimensions. In the latter case we also use information related to the grouping of the amibo acid, as an example of a more general capability of our methodology to include partial labelling information, We discuss some open problems

and suggest directions for future work. (註八)

註一: 在這裡提出 shape analysis 的問題, 當然大家不會懂。因爲畢竟這不是如 OLS 等耳熟能詳的東西, 因此要說明一下。重要的問題是 "to match configurations of points in space after filtering out some geometrical transformation" 這不一定是最重要的問題, 但他們敢說 important , 總有一些根據。此外, 在技術面上來說, "transformation" 一詞也提出來了。

註二: 此處立刻跳入問題。本文是用 hierarchical models 來架構工作的, 而數據中還有標籤 (labels) 的問題, 還有兩組要對比的點, 數目可能不相同的問題。這說明了問題有難度。

註三: 這裡面, 何謂 "inference about the matching" , 何謂 "inference about the transformation" 都沒有提。但這是只是摘要, 是被允許的。

註四: 在這裡提出模型, 其中 "hidden location" , 當然有一點像 latent 的味道, 因此用到 EM 就合理了, 至於 MCMC , 則是因爲 Bayesian 的原故。他們更指出, 因爲模型取得巧, 才能有簡化。

註五: 再於此處提出他們還用到了 directional data 中的本事, 而這又關照了前述的 transformation 有一種是轉軸 (rotation) 。

註六: 指出只做兩種 transformations, 至於 affine 和 rigid motion, 一個需要你回到高等幾何, 另一個卻是物理。暫時需要想一下, 其實是不那麼重要的。但用了較專業的名詞, 就會顯得高級一點。

註七: 這裡告訴我們, matching matrix (當然, 後面的本文一定要定義) 是本文要估計的對象。

註八: 最後歸到兩組應用。都是有科學意味的數據, 這又關照了標題中的 bioinformatic 一詞。

整個摘要告訴我幾件事: 第一, 問題是空間中兩組點是否可以相切合? 因爲這些點指的是蛋白質和胺基酸, 所以這是一個當紅的生物資訊問題, 這就有了「吸引目光」的效果。然後, 作者提出方法, 用 Poisson process 來布空間的點, 雖然太普通但尚可接受, 因爲我們也不知道還有甚麼好用的 proesss 。但他們提出了 hidden location, 就把問題的架構弄得較難一些了, 但他們還能夠簡化, 就暗示著「有內容」。第三, 他們還從空間統計裡借用工具, 並說明他們最得得到的是 "optimal Bayes point estimate of the matching matrix" 。最後, 他們還說: 他們用的是眞實數據。

摘要給人的初步印象是「有不少東西」。但我們應該看他們沒有說的。例如 "optimal Bayes point estimate of the matching matrix" 是不是科學上有意思的問題要點? 或者只是統計學家能做出的東西? 從這裡, 我們看不出來, matching 做得是好是壞, 與是否能最優地做估計是同一件事還是根本不相關? 注意到兩個作者都不是生物學家, 這是引起我一點懷疑的地方。有兩組數據給兩個統計學家玩, 總可以搞些看起來有料的東西出來。估計 matching matrix 和 matching 是同一件事嗎? 我怎麼好像覺得作者沒有交待?

現在且看一下緒論的部分。

Various new challenging problems in shape matching have been appearing from different scientific areas including bioinformatics and image analysis. In a class of problems in shape analysis one assumes that the points in two or more configurations are labelled and these configurations are to be matched after filtering out some transformation. Usually the transformation is a rigid transformation or similarity transformation. Several new problems are appearing in which either the points of configuration are not labelled or the labelling is ambiguous, and in which some points do not appear in each of the configurations. An example of ambiguous labelling arises in understanding the secondary structure of proteins, where we are given not only the three-dimensional molecular configuration but also the types of amino acid at each point. A generic problem is to match two such configurations, where the matching has to be invariant under some transformation group. Descriptions of such problems can be found in the review article by Mardia et al. (2003).

We now describe two datasets related to protein structure. One is of two-dimensional gel data, where each point is a protein itself and the transformation group is affine. In this case we have a partial matching, identified already by experts, that we can use to assess our procedures. In the second example we have a three-dimensional configuration of two active sites of two proteins which also has additional chemical information. Here the underlying transformation to be filtered out is rigid motion. In this problem, one of the main aims is to take a query active site and find matches to a given database, in some ranking order. The matches will give some idea of the functions of the unknown proteins, leading to the design of new enzymes, for example.

There are other related examples from image analysis such as the matching of buildings when one has multiple two-dimensional views of three-dimensional objects; see for example Cross & Hancock (1998). The problem here requires one to filter out the projective transformations before matching. Other examples involve matching outlines or surfaces; see for example Chui & Rangarajan (2000) and an unpublished 2002 Ph.D. thesis from the Technical University of Denmark by L. Pedersen. Here no labelling of points is involved, and we are dealing with a continuous contour or surface rather than a finite number of points. Such problems are not addressed in this paper.

The principal innovations in our approach are the fully model-based approach to alignment, the fact that the model formulation allows us to integrate out the hidden point locations, the prior specification for the rotation matrix, and the Markov chain Monte Carlo algorithm.

一般的摘要只是緒論的子集[1], 這篇文章兩者都要唸, 因爲緒論裡說得並不多。它先提出除了 bioinformatic 之外, 還有 image analysis , 但是語焉不詳, 只是要我們去看 Mardia et al. (2003) 。

然後直接跳到兩組數據。在有數據的論文裡, 數據最重要, 因此得先介紹。但其後所說的問題 "In this problem, one of the main aims is to take a query active site and find matches to a given database, in some ranking order." 似乎本文根本不曾去碰。因此看起來只是花招。

然後他們提到三篇文章: 1998, 2000 和 2002 的博士論文。但全沒有提這些文章和本

[1]但最好要文字有所不同。

文的關係, 好像它們只是 image analysis, 問題類型雖然類似, 所用的技術和本文全然不同的樣子。

最後談到本文的賣點: 幾乎全都是技術性的。得到甚麼結論呢? 似乎看不出來。緒論裡連 matching matrix 提都沒有提, 爲甚麼要估計它, 估得準是不是等於 matching 就做得好, 還是說根本只有一種 matching , 好不好我們控制不了, 最多只是量測一下 matching 的結果而已?

這是可以去讀的文章, 因爲它似乎留下很多空間讓我們去問新的問題。此外, 相關的文獻似乎不多, 這暗示問題也許還新, 下功夫搞清楚了就有機會。

關於 Zheng, Salganik and Gelman 的論文

這篇文章刊在 2006 的 *Journal of the American Statistical Association*, **101**, 409-423. 這是美國統計學會旗下的招科期刊。三位作者都來自 Columbia 大學, 其中, Salganik 是社會學的博士生, 其他兩位都是統計系的老師。

標題就十分搶眼: "How many people do you know in prison?: Using overdispersion in count data to estimate social sructure in networks", 光是憑這個標題, 很多人就會來讀一下論文摘要[2]。標題已抓全要點: 正面的使用 (using), 而不是將 oversispersion 當作數據中的缺失來小心處理。這樣的態度是值得一提的, 而標題中將它反應出來。

文章的摘要如下:

Networks —— sets of objects connected by relationships —— are important in a number of fields. The study of networks has long been central to sociology, where researchers have attempted to understand the causes and consequences of the structure of relationships in large groups of people. Using insight from previous network research, Killworth et al. and McCarty et al. have developed and evaluated a method for estimating the sizes of hard-to-count populations using network data collected from a simple random sample of Americans. In this article we show how, using a multilevel overdispersed Poisson regression model, these data also can be used to estimate aspects of social structure in the population. Our work goes beyond most previous research on networks by using variation, as well as average responses, as a source of information. We apply our method to the data of McCarty et al. and find that Americans vary greatly in their number of acquaintances. Further, Americans show great variation in propensity to form ties to people in some groups (e.g., males in prison, the homeless, and American Indians), but little variation for other groups (e.g., twins, people named Mchael or Nicole). We also explore other features of these data and consider ways in which survey data can be used to estimate network structure.

它強調幾點: (1) Social network 的研究, 有其歷史; (2) 作者用 "multilevel overdispersed Poisson regression model" 來做分析; (3) 對於估計 hard-to-find 群體大小的方法, 是參考了 Killworth et al. and McCarty et al. (注意到, 在摘要中不必將論文的年份寫出來); (4) 利用 variation 是本文技術上的創新, 和以前的論文, 皆有所不同[3]; (5) 證明美國人的社交傾向, 變異極大; (6) 甚至考慮到如何估計 network structure —— 雖然沒有說

[2]當然是因爲一般的統計論文, 不止是內容, 連標題都十分乏味。

[3]其實, 正面面對變異, 統計學者早已覺醒。如質量問題中的田口方法, 還有如 ARCH 之類的時間序列模型都是。

明如何定義一個被估計的 network structure 。

這個摘要比較平實。主要是因爲 social network 是生活的一部分, 即使不知道科學上的定義, 也可以有一些感覺。這摘要給我們的印象是, 作者也做了不少工作, 並且有些工作還是創新的。他們正面地面對 overdispersion, 並且有趣地處理了眞實數據。

緒論是這樣的 (中文的部分是我加上的):

Recently a survey was taken of Americans asking, among other things, "How many males do you know incarcerated[4] in state or federal prison?" The mean of the responses to this question was 1.0. To a reader of this journal, that number may seem shockingly high. We would guess that you probably do not know anyone in prison. In fact, we would guess that most of your friends do not know anyone in prison either. This number may seem totally incompatible with your social world.

這一段就開啓得不錯 —— 你的朋友裡有多少在坐牢? 平均數是一個人? 怎麼這樣高? 這樣豈不是一半的人都在牢裡? 美國的治安眞可怕啊!有了這樣驚人的數據, 然後再輕鬆地爲你解套: 本文的讀者大概都是教育程度好, 社經地位較高的一族, 多半沒有認識坐過牢的人。於是你心情愉快, 好奇心起, 這個「平均一人」是從哪裡出來的?

So how was the mean of the responses 1? According to the data, 70% of the respondents reported knowing 0 people in prison. However, the responses show a wide range of variation, with almost 3% reporting that they know at least 10 prisoners. Responses to some other questions of the same format, for example, "How many people do you know named Nicole?" show much less variation.

這一段明白告訴讀者數據中的 variation 極大 —— 有些人認識一大堆坐牢的人[5]。儘管, 70% 的人都不認識坐牢的人。

This difference in the variability of responses to these "How many X's do you know?" questions is the manifestation of fundamental social processes at work. Through careful examination of this pattern, as well as others in the data, we can learn about important characteristics of the social network connecting Americans, as well as the processes that create this network.

前面兩段話都沒有技術用語。但這一段開始使用了 social network 以及它的形成過程 —— 社會科學研究的味道就進來了。

This analysis also furthers our understanding of statistical models of two-way data, by treating overdispersion as a source of information, not just an issue that requires correction. More specifically, we include overdispersion as a parameter that measures the variation in the relative propensities of individuals to form ties to a given social group, and allow it to vary among social groups. Through such modeling of the variation of the relative propensities, we derive a new measure of social structure that uses only survey responses from a sample of individuals, not data on the complete network.

這一段特將 "include overdispersion as a parameter that measures the variation in the

[4]這個字指的是「被監禁」。
[5]當然不是指牢頭。

relative propensities of individuals" 加以說明, 並申明我們利用 "modeling of the variation of the relative propensities", 才導出 "a new measure of social structure that uses only survey responses".

Background

Understanding the structure of social networks, and the social processes that form them, is a central concern of sociology for both theoretical and practical reasons (Wasserman and Faust 1994; Freeman 2004). Social networks have been found to have important implications for the social mobility (Lin 1999), getting a job (Granovetter 1995), the dynamics of fads and fashion (Watts 2002), attitude formation (Lee, Farrell, and Link 2004), and the spread of infectious disease (Morris and Kretzcnmar 1995).

這一段選用一些關於 social network 的一些文獻, 注意到沒有太老的東西。這雖不是一個完整的文獻蒐尋, 但至少資料都還新。表示作者還知道最近發生了甚麼[6]。

When talking about social networks, sociologists often use the term "social structure," which, in practice, has taken on many different meanings, sometimes unclear or contradictory. In this article, as in the article by Heckathorn and Jeffri (2001), we generalize the conception put forth by Blau (1974) that social structure is the difference in affiliation patterns from what would be observed if people formed friendships entirely randomly.

用文字說明本文中所謂的 social structure 是甚麼意思。觀念是相對的: 將「隨機交友」當作 social structure 爲 0 來看問題。當然, 這裡還沒有明確的定義, 只有明確的想法[7]。

Sociologists are not the only scientists interested in the structure of networks. Methods presented here can be applied to a more generally defined network, as any set of objects (nodes) connected to each other by a set of links (edges). In addition to social networks (friendship network, collaboration networks of scientists, sexual networks), examples include technological networks (e.g., the internet backbone, the world-wide web, the power grid) and biological networks (e.g., metabolic networks, protein interaction networks, neural networks, food webs); reviews have been provided by Strogatz (2001), Newman (2003b), and Watts (2004).

這一段是可有可無的。只是說「可以有更廣的應用」, 一般人都不會去查的, 但作者應該眞的讀過所列的幾篇文章[8], 否則就是不負責任。

Overview of the Article

In this article we show how to use "How many X's do you know?" count data to learn about the social structure of the acquaintanceship network in the United States. More specifically, we can learn to what extent people vary in their number of acquaintances, to what extent people vary in their propensity to form ties to people in specific groups, and also to what extent specific subpopulations (including those that are otherwise hard to count) vary in their popularities.

[6] 引用十年以前的文章時, 最好只引用那些經典的, 才顯得你有學問有品味。
[7] 如果你有相關的數據, 可考慮做「論網路交友的社會結構」。
[8] 至少他也得用細讀過它們的緒論。

先提出數據的基本形態: 由某一特殊問題 “How many X’s do you know?” 而得, 所以是 count data 。

The data used in this article were collected by McCarty, Killworth, Bernard, Johnsen, and Shelley (2001) and consist of survey responses from 1,370 individuals on their acquaintances with groups defined by name (e.g., Michael, Christina, Nicole), occupation (e.g., postal worker, pilot, gun dealer), ethnicity (e.g.. Native American), or experience (e.g., prisoner, auto accident victim); for a complete list of the groups, see Figure 4 in Section 4.2. Our estimates come from fitting a multilevel Poisson regression with variance components corresponding to survey respondents and subpopulations and an overdispersion factor that varies by group. We fit the model using Bayesian inference and the Gibbs-Metropolis algorithm, and identify some areas in which the model fit could be improved using predictive checks. Fitting the data with a multilevel model allows separation of individual and subpopulation effects. Our analysis of the McCarty et al. data gives reasonable results and provides a useful external check on our methods. Potential areas of further work include more sophisticated interaction models, application to data collected by network sampling (Heckathorn 1997, 2002; Salganik and Heckathorn 2004), and application to count data in other fields.

這一段說明數據的來源 (McCarty, Killworth, Bernard, Johnsen, and Shelley (2001)) 和一般說明。所用到的模型 (a multilevel Poisson regression with variance components), 所用的工具 (Bayesian inference and the Gibbs-Metropolis algorithm) 及甚麼時候可以做得更好 (identify some areas in which the model fit could be improved using predictive checks)。至於這對於 social network 研究的影響, 倒是看不出來。

主要的結果似乎仍然是導出了 “a new measure of social structure that uses only survey responses” 。這對於社會學的意義是甚麼? 似乎沒有著墨, 甚至不知道這個所謂的 new measure, 是用來量測甚麼性質的。

Buhlmann 的文章

這文章發表在 2006 的 Annals of Statistics, **34**, 559-581. Annals 是國際數理統計學會的旗艦期刊。要求證明嚴謹, 是技術性要求很高的。當然, 這篇文章是蠻理論的 (否則不會投進來)。讀它的摘要和緒論, 除非你是相當夠格的專家, 根本就搞不清來他做了甚麼工作。我是一直讀到 p.565 才把此文的「成就」搞明白[9]。但還不算數學證明。

這是爲一小撮人寫的論文。標題是 “Boosting for high-dimensional linear models” 。這裡只用了一個較生疏的名詞: boosting 。這個名詞來自所謂的機械學習 (machine learning) , 原意是指將一個較弱的 (預測) 方法, 經過某類處理之後 (處理的方式叫 boosting), 變得更好[10]。

統計學家對 boosting 有興趣, 因此也有不少學者加入研究。但此文光是從標題是看不出甚麼的, 所以先讀一下它的摘要:

We prove that boosting with the squared error loss, L_2Boosting, is consistent for very

[9]因爲好幾個名詞到後面才明確定義。這對於摘要和緒論是可以的, 對正文就不行。對這類文章, 作者並沒有希望你第一次就讀懂。這種摘要和緒論, 是爲了讓專家留下深刻印象用的。

[10]我曾以爲如果我有 51% 的機率猜到明天某股票的漲或跌, 則我應可用 boosting 的辦法, 提高我的猜測率到 90% , boosting 當然沒有這樣利害。

high-dimensional linear models, where the number of predictor variables is allowed to grow essentially as fast as $O(\exp(\text{sample size}))$, assuming that the true underlying regression function is sparse in terms of the ℓ_1-norm of the regression coefficients.

這個摘要太專業了, 我要從論文的 p.559 讀到 p.565 才知道它第一句在說甚麼。基本上是這樣的。在

$$y = X\beta + \epsilon, \quad (y = n \times 1, X = n \times p, \beta = p \times 1)$$

這樣的線性模型中, 一般都要求 $n > p$, 因爲否則「自由度不夠」。但如果 $p > n$, 請問如何估計 $E(y|X)$ 這個迴歸函數? 一般當然是不行, 否則我們讀到的初等統計都錯了。但是, 如果大多數的 β 的值都很小時是不是有可能?

這裡,「大多數的 β 的值都很小」的條件, 叫做 sparse 。這篇文章所證明的是, 在 sparse 的條件下, 是可以得到 $E(y|X)$ 的「一致估計量 (consistent estimate)」的。

In the language of signal processing, this means consistency for de-noising using a strongly overcomplete dictionary if the underlying signal is sparse in terms of the ℓ_1-norm. We also propose here an AIC-based method for tuning, namely for choosing the number of boosting iterations. This makes L_2Boosting computationally attractive since it is not required to run the algorithm multiple times for cross-validation as commonly used so far.

這一段說, 在做 boosting 時, 有一個「每次疊算的次數」要交待明白。作者說他有辦法。

We demonstrate L_2Boostwg for simulated data, in particular where the predictor dimension is large in comparison to sample size, and for a difficult tumor-classification problem with gene expression microarray data.

這一段對數理統計的論文是蠻標準的, 先講方法, 再做模擬, 最後再套入眞實數據。這裡, 眞數據來自生物微晶片, 雖然不是作者自己測出的, 但這個問題仍然紅火 (事實上, 我們會去對 $n < p$ 的情形如此關注, 這類數據的大量存在是主要原因)。

這個摘要寫得算中肯 —— 反正只是寫給內行人看的 —— 它只是明白地講出所做出的主要結果。

它的緒論則較長。其實只是摘要的加強版。內容如下:

Freund and Schapire's [11] AdaBoost algorithm for classification has attracted much attention in the machine learning community (cf. [20] and the references therein) as well as in related areas in statistics [1,13], mainly because of its good empirical performance with a variety of datasets. Boosting methods were originally introduced as multiple prediction schemes, averaging estimated predictions from reweighted data. Later, Breiman [1, 2] noted that the AdaBoost algorithm can be viewed as a gradient descent optimization technique in function space. This important insight opened a new perspective, namely to use boosting methods in contexts other than classification. For example, Friedman [12] developed boosting methods for regression which are implemented as an optimization using the squared error loss function: this is what we call L_2Boosting. It is essentially the same as Mallat and Zhang's [19] matching pursuit algorithm in signal processing.

這一段是給關於 boosting 的背景資料。主要是說誰做了甚麼工作。因爲指明了 L_2Boosting 是本文的手段 (先別管 L_2Boosting 的定義, 對這類數理統計的文章, 早晚會給出明確定義的), 此段只誇獎了 [1, 2] 兩篇文章, 說 "This important insight opened a new perspective" 。至於其它的論文, 都不帶褒貶。他並提出 L_2Boosting 和 match prusit 是同一回事[11]。注意到, 他這一段只提了七篇論文, 這都是有選擇的, 這些都是關於 L_2Boosting 的重要文獻, 充分表示作者是否內行。

Recently, Efron, Hastie, Johnstone and Tibshirani [10] made a connection for linear models between forward stagewise linear regression (FSLR), which seems closely related to L_2Boosting, and the ℓ_1-penalized Lasso [22] or basis pursuit [5]. Roughly speaking: under some restrictive assumptions on the design matrix of a linear model, FSLR approximately yields the set of all Lasso solutions (when varying over the penalty parameter). This intriguing insight may be useful to get a rough picture about L_2Boosting via its relatedness to FSLR: it does variable selection and, shrinkage, similar to the Lasso. However, it should be stated clearly that the methods are not the same; an example showing a distinct difference between L_2Boosting and the Lasso is presented in Section 4.3. Moreover, we point out in Section 2.1 that FSLR and LzBoosting are different algorithms as well.

這一段又提了一些相關的論文, 說它們似乎和 L_2Boosting 做同樣的事。作者的結論反面的: 這些都不等於 L_2Boosting。他明白地說: "it should be stated clearly that the methods are not the same" 。這一段所給人的印象是: 此人將 L_2Boosting 的事情已弄得十分清楚了。也澄清了一些大家或以爲對的事情。而這些都是學問, 新的知識。

As the main result, we prove here that L_2Boosting for linear models yields consistent estimates in the very high-dimensional context, where the number of predictor variables is allowed to grow essentially as fast as 0(exp(sample size)), assuming that the true underlying regression function is sparse in terms of the ℓ_1-norm of the regression coefficients.

這裡強調本文的主要結果。

This result is, to our knowledge, the first about boosting in the presence of (fast) growing dimension of the predictor.

並且說明別人沒有做出這類型態的結果 —— 該說自己好的時候就大方地說。

Some consistency results for boosting with fixed predictor dimension include [17, 18] as well as [25]. Except for Jiang's [17] result, these authors consider versions of boosting either with ℓ_1-constraints for the boosting aggregation coefficients or, as in [25], with a relaxed version of boosting which we found very difficult to use in practice due to the nonobvious tuning of the relaxation, that is, how fast the boosting aggregation coefficients should decay. The result by Zhang and Yu [25] may be generalized without too much effort to a setting with increasing dimension of the predictor variable, but their theoretical work includes only a rigorous treatment of the classification problem (besides the above mentioned disadvantage of their relaxed boosting algorithm).

這裡說其它似乎相近的結果都不夠好。注意到作者的理由都很明確, 不是空口說白話的。需知這樣的內容最易讓同儕生氣。文字不要帶火氣。

[11]這是暗示「本人學問不錯」的手法, 要學 —— 尤其是他不僅提 [19], 還要提 "signal processing" , 這秀出了「格局不止在 machine learning」。

We believe that it is mainly for the case of high-dimensional predictors where boosting, among other methods, has a substantial advantage over more classical approaches. Some evidence for this will be given in Section 4.1, and other supporting empirical results have been reported in [3] in the different context of low- or high-dimensional additive models for comparing L_2Boosting with more traditional methods such as backfitting or MARS (restricted to additive function estimates). Notably, many real datasets nowadays are of high-dimensional nature. Besides the well-documented good empirical performance of boosting, we identify it here as a method which can consistently recover very high-dimensional, sparse functions.

此處說明高維的問題才是主要的應用, 並且因爲有 "a substantial advantage over more classical approaches" , 世上有太多這類的數據 (這一點其實大家都知道)。並且, 最主要的, 作者又將主結果再提一次。

We may also view our result as a consistency property for de-noising using L_2Boosting with a strongly overcomplete dictionary. In contrast to a complete dictionary, for example, Fourier- or wavelet-basis, the strongly overcomplete noisy case is not well understood. Our result yields at least the basic property of consistency.

這一段, 作者又將本文與其它事物拉關係。這類問題的結構大同小異, 多半在不同的架構下運作。作者說 "Our result yields at least the basic property of consistency", 但這一段話有「想當然耳」的味道 —— 這些東西是可以給學生做的題目, 有作者這樣的專家猜應該會成立, 但眞正寫出來卻是另一件事。

Besides the theoretical consistency result, we propose here a computationally efficient approach for the tuning parameter in boosting, that is, the number of boosting iterations. We give an easily computable definition of degrees of freedom for L_2Boosting, and we then propose its use in the corrected AIC criterion. Unlike cross-validation, our AIC-tuning does not require boosting to be run multiple times. This makes the AIC-type data-driven boosting computationally attractive: depending on the data, it is sometimes as fast as the very efficient LARS algorithm for the Lasso with tuning by its default ten-fold cross-validation [6, 10].

這一段說明計算的效率。作者也有做法: 用 AIC 來做。到此爲止, 論文的理論部分都已介紹完畢。此文其實只有一個主定理, 但作者洋洋灑灑地用了 25 頁的 Annals pages, 主要是靠此人確實對相關的問題知道得夠多夠深入, 所以才寫得出這些旁徵博引的話。需知說這類話最易出問題, 因爲你的看法不見得就是評審人的看法。這時深度就變得重要, 如果你眞的言之有物, 那些專家也會懂的。

We demonstrate on some simulated examples how our L_2Boosting performs for (low- and) mainly high-dimensional linear models, in comparison to the Lasso forward variable selection, ridge regression, ordinary least squares and a method which has been designed for high-dimensional regression [14]. We also consider a difficult tumor-classification problem with gene expression microarray data: the predictive accuracy of L_2Boosting is compared with four other, commonly used classifiers for microarray data, and we briefly indicate the interpretation of the L2Boosting fit along the lines of a linear model fit.

最後說明模擬和眞實數據的套用, 以及與其它類似方法的比較。

這篇論文的摘要和緒論都算是標準的寫法, 它沒有奇兵突出 (像 Zheng et al. 一樣), 但平舖直敍, 立刻進入主結果, 並大力地和現有的文獻比較, 以宣揚自己的優點, 坐實別人的缺點。這是數理統計論文的通常寫法。

Brien 和 Bailey 的論文

這篇文章登在 *Journal of the Royal Statistical Society, Series B* (2006) **68**, 571-609. 這是英國皇家統計學會的招牌刊物。

標題極短:"Multiple randomization", 只有兩個字。一般只有自認是高手者才會這樣取論文標題。這暗示本文並不針對某一特定的小範圍而討論。它的摘要是這樣的:

Multitiered experiments are characterized by invoving mutiple randomizations, in a sense that we make explcit. We compare and contrast six types of multiple randomizations, using a wide range of exampls, and discuss their use in designing experiments. We outline a system of describing the randomizations in terms of sets of objects, their associated tiers and the factor nesting, using randomization diagrams, which gave a convenient and readily assimilated summary of an experiment's randomization. We also indicate how to formuate a randomiztion-based mixed model for the analysis of data from such experiments.

這裡用了 "tier" 一詞, 這並不是一般的實驗設計 (DOE) 教本上常見的。此處就特用來和 randomization 一起配合使用 (所以後文將再說明)。摘要裡說明本文是要用很多例子來討論多重隨機化的道理的, 提到了 "randomization diagrams", 和如何處理 "randomization-based mixed model" 。這個摘要沒有甚麼不對, 只是看不太懂而已。DOE 已是非常精微細緻的學問, 細緻到非專家就不想去碰的程度。作者之一的 Bailey 是此中的有名的專家 (如果不是權威), 但這樣的摘要, 我讀起來還是有「閑人免進」的感覺。

它的緒論如下:

Experiments are distinguished from observational studies and happenstance data by the purposive application of treatments to observational units. Nelder (1965a, b). White (1975), Bailey (1981,1991) and Heiberger (1989) formulated methods for the analysis of experiments that take this distinction into account by classifying factors in the experiment as either 'block' or 'treatment' factors. Here we interpret block and treatment factors to mean the sets of unrandomized and randomized factors respectively. Many researchers have advocated the use of such a distinction (Fisher (1935), Wilk and Kempthorne (1957), Cox (1958), section 6.3, Yates (1975), Mead and Curnow (1983), section 14.1, Brien (1983) and Piepho et al. (2003)). The distinction enables the direct construction of analysis-of-variance tables exhibiting all the confounding, as well as the inclusion, in a mixed model, of all terms that are warranted by the randomization.

這一段是名家手筆。把區集 (block) 和 (treatment) 用「不隨機化」和「隨機化」來分辨。等閑人寫不出來。

Of course, it is possible to formulate an analysis without making this distinction. For example, in the analysis of a randomized complete-block design as a 'two-way analysis of variance' based on a two-factor, no-interaction model, no distinction is made between the treatments that are randomly assigned and the blocks that result from inherent features of the observational units. As Kempthorne (1955) noted, this results in the same analy-

sis for the randomized complete-block design and the two-way factorial experiment, even though they differ markedly in the randomization that is employed. This is because the confounding that is inherent in each experimental situation has not been recognized (Brien and Payne, 1999). Similarly, the common description of the classic split-plot experiment in terms of only the three factors blocks, mainplot treatments and subplot treatments also fails to distinguish between randomized treatments and inherent features of the observational units.

讀過一點 DOE 的人都知道 "this results in the same analysis for the randomized complete-block design and the two-way factorial experiment, even though they differ markedly in the randomization that is employed", 但讀過 Brien and Payne (1999) 並肯去深入了解的人就少了。若是能夠了解到 "the confounding that is inherent in each experimental situation has not been recognized", 是否實驗數據的分析方法就會不同?

Brien (1983) identified experiments that involve more than the single randomization of treatments to plots and concluded that two sets of factors are not enough to describe the randomization in"such experiments. The general term 'tiers' was introduced for the sets of factors in an experiment which result from the classification of the factors according to their status in the randomization: see also Cullis et al. (2003). Experiments that involve multiple randomizations, and hence more than two tiers, were labeled 'multitiered'. Multitiered experiments include two- phase, some superimposed and some single-stage experiments, and some multistage experiments using the same units at each stage; they do not represent a collection of new designs, but are a class of designs made up of several existing design types.

這一段說明, 在 DOE 中,"experiments that involve more than the single randomization of treatments" 的情形, 20 年前就已發現。並由此引入 "tier" 一詞, 並讓它和「隨機化」相連。照顧到摘要開始時的句子。但像 "Multitiered experiments include two-phase, some superimposed and some single-stage experiments, and some multistage experiments using the same units at each stage" 的寫法, 就太專業了, 我讀了一點感覺都沒有。也許可以 impress 專家罷。

Two-phase experiments were introduced by Mclntyre (1955) and were discussed by Cox (1958), although Cox used the term 'stage' rather than 'phase'. They are characterized by the following:

(a) a complete experiment in the first phase, although not necessarily with the measurement of response variables;
(b) the randomization of the units from the first phase to the units in the second phase.

到此處, 過去的和本文有關的文獻回顧就差不多結束了。這幾段算是較長的, 但是, 懂就懂; 不懂還是不懂。因爲搞懂也需要功力。至於文字, 是寫得有專家的味道。

A common situation in which they occur is where the produce from a field trial has to be processed in a laboratory or an evaluation phase (Brien, 1983; Brien et al., 1987; Wood et al., 1988; Brien and Payne, 1999; Cullis et al., 2003; Kerr, 2003). All of the examples of two-phase experiments that have been published in the literature employ only one of the simplest type of multiple randomization.

In this paper we distinguish between stages and phases in experiments, the former including the latter as a special type. We note that usage of stages and phases in experiments is unrelated to that in sampling. Multistage experiments are conducted in several distinct time intervals with randomization of factors for each interval. They include two-phase, superimposed and change-over experiments (but not longitudinal studies), as well as multistage reprocessing experiments, in which the response is measured only after several stages of processing the same units (Miller (1997), Mee and Bates (1998) and Box and Jones (1992), section 5). Not all multistage experiments are multitiered as they do not involve multiple randomizations —— the usual repeated measurements experiments that are discussed in text-books are not. Single-stage experiments that involve multiple randomizations include some grazing and some plant experiments. For example, Brien and Demetrio (1998) discussed multiple randomizations in the context of continuous grazing trials and Preece (1991) described a multitiered horticultural experiment.

由此段開始介紹本文所要做的工作。先說明本文中所用到 "stage" 和 "phase" 意味上的不同。

Although the experiments do not involve new designs, there has been uncertainty and difficulty in their randomization and analysis. The purpose of this paper is to describe and compare six different types of multiple randomization and how they are employed in a range of multitiered experiments. We concentrate on equireplicate designs, because they demonstrate all the relevant issues but without extra complications. We hope to increase understanding of the way in which multiple randomizations can be employed in designing experiments. We also develop randomization diagrams for depicting the randomization in an easily absorbed form. In Section 2, we introduce and define the terms that are required and illustrate their use in the context of two-tiered experiments. Section 3 describes a simple three-tiered experiment. The six types of multiple randomizations are then described: Section 4 discusses and gives examples of multiple randomizations in which the two constituent randomizations can be performed in either order, whereas Section 5 does the same for the inclusive multiple randomizations. Section 6 gives examples involving more than one type of multiple randomizations, and hence three or more randomizations. A strategy for formulating a mixed model for analyzing multitiered experiments is outlined in Section 7. Section 8 addresses some general issues concerning terminology and multiple randomizations. Further examples involving multiple randomizations are available in Brien (1992) and from the multitiered experiment Web site at http://chris.brien.name/multitier/.

這一段說本文所做的事情, 是比摘要要說得多的完整版。所做的事有二:"to describe and compare six different types of multiple randomization and how they are employed in a range of multitiered experiments" 和 "develop randomization diagrams for depicting the randomization in an easily absorbed form" 。最後, 再用「目錄式」的列舉, 告訴讀者在哪一節做哪一件事。這是緒論的習慣, 雖然, 我每次唸到此就, 就必然會跳過。

大體說來, 這仍然是一篇中規中矩的論文緒論 —— 不能因爲我對 DOE 知識淺薄而小看它。因爲它的目的, 不是讓我覺得這文章好, 而是讓 DOE 的專家覺得這文章內容。

這是一篇所謂的 discussion paper —— 原則上應該是較有份量的文章。統計學者常以「被邀往 RSS 宣讀 (並討論) 論文」爲榮。

題外的感覺

一門學問發展到最後就是愈來愈細, 到最後, 只有高手才會有興趣。高手加入表示論文困難, 論文困難就讀者少。精妙設計而來的實驗配置及規範, 像藝術品。因爲他們設計得得精密, 最後只有專家才懂他們講究的要點是在何處。以至於可能去眞正去用的人 (實驗者), 可能因背景不足而卻步。

精妙的設計如美女 ——增之一分則太長, 減之一分則太短。君不見全眞七子的北斗七星陣, 在七子都健在的時候, 可以抵擋東邪西毒這樣的大高手。但將譚處端換上尹志平以後, 便威力大減。何以故? 計算太精, 所以不能容許小錯。

學問到了一定程度就「閑人免進」了。從另一個角度來說, 當成吉斯汗用萬人隊沖鋒陷陣的時候, 北斗七星陣有個屁用!

Goorge Box 曾說過一個故事。他說, 有一個他系上的學生, 因爲拓樸學 (topology) 不夠好, 所以沒能通過博士的資格考 (qualify) 。他於是跑到一家公司去工作。在公司裡, 他僅僅靠著一個 2^3 實驗, 一路立功升到公司的副總經理!

本段練習

我們附上四篇論文在他們投稿時的摘要和緒論。這些論文最後都接受了, 也刊登出來了。所以也不能算太差。這類文章都在 可拒絕可接受之間 —— 看遇到的評審人 (referee) 想當好人, 還是想當壞人。

現在請你用評審人的角度來讀它。先不要管主要部分 (以後再給), 現在只讀閉要和緒論。請問:

1. 我看得出作者想做甚麼嗎?

2. 他們的做法或思路有道嗎?

3. 作者的學問夠不夠大?

4. 你的直覺是傾向於推薦刊登還是建議拒絕? 理由?

第一篇

Testing for Activation in Data from FMRI Experiments

Abstract

The traditional method for processing functional magnetic resonance imaging (FMRI) data is based on a voxel-wise, general linear model. For experiments conducted using a block design, where periods of activation are interspersed with periods of rest, a haemodynamic response function (HRF) is convolved with the design function and, for each voxel, the convolution is regressed on prewhitened data. An initial analysis of the data often involves computing voxel-wise two-sample t-tests, which avoids a direct specification of the HRF. Assuming only the length of the haemodynamic delay is known, scans acquired in transition periods between activation and rest are omitted, and the two-sample t-test is used to compare mean levels during activation versus mean levels during rest. However, the validity of the two-sample t-test is based on the assumption that the data are Gaussian with equal variances. In this article, we consider the Wilcoxon rank test as well as modified versions of the classical t-test that correct for departures from these assumptions.

The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

Introduction

Functional Magnetic Resonance Imaging (FMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to FMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully and the data follow the Gaussian distribution. In this article, we consider the analysis of FMRI data collected in one of two states, called "activation" and "rest",i based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the activation data and the rest data are Gaussian with equal variances. In this article, we propose use of a modified two-sample test for FMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937) which is a modification of two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for FMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in FMRI and the most common methods used to process FMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for FMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of FMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

第二篇

An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population

Abstract

Behavioral risk factors for cancer tend to cluster within individuals, which can compound risk beyond that associated with the individual risk factors alone. There has been increasing attention paid to the prevalence of multiple risk factors (MRF) for cancer, and to the importance of designing interventions that help individuals reduce their risks across multiple behaviors simultaneously. The purpose of this paper is to develop methodology to identify an optimal linear combination of multiple risk factors (score function) which would facilitate evaluation of cancer interventions.

Introduction

Despite the considerable biomedical advances of the last half-century, facilitating improvement in lifestyle behaviors remains the most efficacious population-level strategy for reducing cancer risk. Estimates vary, but suggest that over fifty percent of new cancer cases and up to one-third of cancer mortality could be prevented though improvements in health behavior practices (American Cancer Society, 2004; Doll and Peto, 1981). A 19 percent decline in the rate at which new cancer cases occur, and a 29 percent decline in the rate of cancer deaths, could potentially be achieved by 2015, if prevention efforts were heightened and behavior change sustained. This would translate to the prevention of approximately 100,000 cancer cases and 60,000 cancer deaths each year, by the year 2015 (National Cancer Policy Board and Institute of Medicine, 2003). There is ample epidemiological evidence for the consideration of red meat consumption, physical activity, and folic acid intake in cancer prevention efforts. Regular physical activity lowers the risk of cancers of the colon, breast, and possibly prostate (Colditz, Cannuscio, and Frazier, 1997; Friedenreich and Rohan, (1995)). An additional 30 percent of cancer deaths can be attributed to adult diet (Anonymous, 1996); higher intake of red meat has been associated with increased risk of colon (Sandhu, White and McPherson, 2001) and prostate cancers (Michaud, Augustsson, Rimm, Stampfer, Willett, and Giovannucci 2001). Associated with both physical inactivity and diet is obesity, which may account for between 25-30 percent of cancers of the colon, breast (postmenopausal), endometrium, kidney, and esophagus (Vainio and Bianchini, 2002). Folic acid is protective against colon cancer (Giovannucci *et al.* 1998); long-term multi-vitamin use, in particular has been found to reduce risk for colon cancer, likely because of its folic acid content (Giovannucci *et al.* 1998). The risk for many diseases, including colon cancer, is associated with multiple behavioral risk factors (MRF); these behaviors are highly interrelated and tend to cluster within individuals. For example, those who eat high-fat diets are also more likely to be sedentary, suggesting that the behaviors may be mutually reinforcing (see e.g., Emmons, Marcus, Linnan, Rossi, and Abrams, 1999). Change in one behavioral risk factor thus may serve as a stimulus or gateway for change in the other health behaviors (see e.g., Emmons *et al.* 1999), and there are overarching behavioral principles and intervention frameworks that guide behavior change efforts across risk factors. Consequently, to facilitate population-level reductions in cancer risk, it may be inefficient to target discrete behavioral risk factors, when similar principles might be applied simultaneously to multiple behaviors (Institute of Medicine, 2000). The literature provides little consensus as to the most appropriate analytic strategy for evaluating the efficacy of MRF interventions; most studies have analyzed the various outcomes independently or by creating a simplistic sum (e.g., 1 RF + 1 RF = 2RFs) (see e.g., Prochaska and Sallis 2004; Campbell, James, Hudson, Carr, Jackson, Oakes, Demissie, Farrell, and Tessaro, 2004). This could be problematic, because the use of separate analytic strategies may result in improper inferences regarding the effect of an MRF intervention because of correlation among the factors. Such strategies may overlook the clustering effect brought about by the agglomeration of multiple behavioral risk factors and have been criticized as being too simplistic. The purpose of this paper is to develop a methodology to identify an optimal linear combination of multiple behavioral risk factors (MRF score function) for cancer that would best facilitate evaluation of an MRF cancer intervention.

第三篇

Reducing Subjectivity in the Likelihood

Abstract

Some scientists prefer to exercise substantial judgment in formulating a likelihood function for their data. Others prefer to try to get the data to tell them which likelihood is most appropriate. We suggest here that one way to reduce the judgment component of the likelihood function is to adopt a mixture of potential likelihoods and let the data determine the weights on each likelihood. We distinguish several different types of subjectivity in the likelihood function and show with examples how these subjective elements may be given more equitable treatment.

Introduction

We propose methods for modeling the likelihood function that will require fewer subjective judgments. We first discuss the nature of the problem of subjectivity in the likelihood function; then we review some related research; and finally, we define a mixture likelihood function and suggest estimation procedures that reduce the effects of subjective views imposed on the observed data.

1.1 Statement of the Problem

It is sometimes desirable that beliefs of experimenters should be brought into a scientific analysis in ways that minimally distort the measured data (see, for example, Hogarth, 1980; Kyberg and Smokler, 1980; Lad, 1996). But that having been said, scientists observing data sometimes interpret the data points subjectively, according to what they want the data to show, and according to how precisely they believe the data points were measured. The latter procedure is of course quite common. This subjective interpretation of observed data may be totally at the unconscious level, or it may be purposeful (with the purposeful interpretation, the analysis may become fraudulent; see for example, Grayson, 1995, 1997; Howson and Urbach, 1990; and Press and Tanur, 2001).

The subjective interpretation of empirical data in medicine was discussed by Kaptchuk (2003). He stated (page 1, op. cit.):

"Doctors are being encouraged to improve their critical appraisal skills to make better use of medical research. But when using these skills, it is important to remember that interpretation of data is inevitably subjective and can itself result in bias. Facts do not accumulate on the blank slates of researchers' minds, and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists' convictions or theoretical apparatus".

Statistical analysis of a data set most often proceeds by summarizing the distribution of the data in terms of its likelihood function. In order to specify the form of the likelihood function, various assumptions are made about the data, such as mutual independence, identical distributions, unimodality, etc. After the likelihood function has been specified, additional assumptions are sometimes made (significance levels thought to be appropriate are specified, a prior distribution about the underlying unobservable quantities may be

brought in, etc.). Analysis of the data generally proceeds by trying to keep the likelihood function treatment of the data as simple as possible, so that the scientist or analyst will introduce minimal distortion of the data. The analyst tries not to discard data, and tries to maximize the chance of understanding what nature is trying to tell us through the revealed data about the underlying phenomenon. In this way, when the analysis of the data has been completed, the claim can reasonably be made that the conclusions drawn from the analysis approximate, if not precisely reflect, the laws of nature, rather than the possible misinterpretations and misunderstandings of the laws of nature by human beings. It will be useful to first briefly define what we mean by objectivity and subjectivity, in this context.

According to Mandik, 2001,

"The word objectivity refers to the view that the truth of a thing is independent from the observing subject. The notion of objectivity entails that certain things exist independently from the mind, or that they are at least in an external sphere. Objective truths are independent of human wishes and beliefs. The notion of objectivity is especially relevant to the status of our various ideas, and the question is to what extent objectivity is possible for thought, and to what extent it is necessary".

This is but one of many definitions that have been suggested. The elusive quest for objectivity in science has been, and remains, an important topic of discussion among historians and philosophers of science (for extensive additional discussions of the meaning of "objectivity", see for example, Bower, 1998; Porter, 1995, 1996; and Daston and Galison 1992). For some, scientific objectivity involves the search for certainty in knowledge about one of nature's well-kept secrets, independent of what human beings believe; but in many cases, we find that what we earlier thought to be true about nature, turns out later to be questionable.

In an interesting example from physics, Folger, 2003, pointed out that:

"Pioneer 10, launched in 1972, is now some 8 billion miles from home. But is has been slowing down, as if the gravitational pull on it from the sun is growing progressively stronger the farther away it gets. Milgrom proposed (see the MOND pages-MOdified Newtonian Dynamics) that Newton's laws might change at these accelerations. If Milgrom is right, Newton's and Einstein's laws will be in for some major tweaking."

Sometimes the scientist has such deep understanding and insight into the phenomenon he/she is studying that the scientist's own predictions of what should be found from the analysis are far superior to what the data analysis seems to indicate. In some cases the beliefs of the scientist or analyst are so strong, even before actually taking any data that bear on the phenomenon, that the data are interpreted or manipulated so that they will reflect these preconceived views of the scientist. Any preconceived personal views (views held before taking any data), weak or strong, are what we refer to in this context as subjectivity.

1.2 Related Research

One approach to reducing the effects of differing assumptions about likelihoods may

be found in a line of research that involves use of the empirical likelihood function. In this approach, most useful in large samples, a discretized, binned, version of the empirical cdf, instead of a specific likelihood function, is used. Inference is then made from a multinomial distribution. An unfortunate feature of this approach is the additional unknown parameters that are concomitantly introduced into the model. See: Owen, 1988, 2001. For typically small and moderate size samples this could be a problem, but for the massive data sets typical of data mining applications (see, for example: Berry & Linoff (1997); and Hastie, Tibshirani, and Friedman, 2001) such an approach could be a helpful alternative.

We show in the next section how we might understand and account for some types of subjectivity that sometimes enters the likelihood function, and might not be desired. We will use the definition and form of the likelihood function in which for absolutely continuous random variables, up to a proportionality constant, it is the joint probability density function of the observables given the unobservables.

第四篇

Application of One Sided t-tests and a Generalized Experiment Wise Error Rate, GFWER(k), to High-Density Oligonucleotide Microarray Experiments: An Example Using Arabidopsis

Abstract

Motivation: A formidable challenge in the analysis of microarray data is the identification of those genes that exhibit differential expression. The objectives of this research were to examine the utility of simple ANOVA, one sided t tests, natural log transformation, and a generalized experiment wise error rate methodology for analysis of such experiments. As a test case, we analyzed a Affymetrix GeneChip microarray experiment designed to test for the effect of a CHD3 chromatin remodeling factor, PICKLE (PKL), and an inhibitor of the plant hormone gibberellin (GA), on the expression of 8256 Arabidopsis thaliana genes.

Results: The GFWER(k) is defined as the probability of rejecting k or more true null hypothesis at a given p level, and ignores the probability of making less than k Type I errors. The method was shown to be simple to apply and greatly increases power. A k value as small as 2 or 3 was concluded to be adequate for large or small experiments respectively. A one sided t-test along with GFWER(2)=.05 identified 43 genes as exhibiting PKL-dependent expression. Expression of all 43 genes was re-examined by qRT-PCR, of which 36 (83.7%) were confirmed to exhibit PKL-dependent expression.

Introduction

The advent of inexpensive microarray technology has enabled individual laboratories to easily obtain a global perspective on the expression pattern of thousands of genes. This powerful technology has allowed investigators to diagnose early cancers (Kim, J.W. and Wang, X.W., 2003; Zhang, H. et al., 2003), discover genes that contribute to quantitative traits (Gu, C.C. et al., 2002), and detect coordinated gene regulation during pivotal developmental events such as embryogenesis and sexual maturation (Girke, T. et al., 2000; Lo, J. et al., 2003; Ruuska, S. A. et al., 2002).

The first generation microarrays were generally based on two dye methodologies. These cDNA microarray experiments involve hybridizing two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000-20,000 cDNA probes (Yang and Speed, 2002). In contrast, more recent high-density oligonucleotide microarrays, such as those offered by Affymetrix , provide direct information about the expression levels in an mRNA sample and can have a much higher density (Yang and Speed, 2002).

The majority of methodologies for microarray analysis have been developed for two dye spotted arrays (Kerr et al. 2000; Kerr and Churchill, 2001; for review see Quackenbush, 2001 and Yang and Speed, 2002). Unfortunately these two-dye spotted arrays also pose other statistical issues, such as normalization to correct for dye bias. Furthermore if more than 2 treatments are used, it is not possible to compare all treatments on the same chip thus necessitating an Incomplete Block Design (IBD) type design (Kerr and Churchill, 2001). As such, special experimental designs, such as the reference and rotational design are needed for correct analysis (Kerr and Churchill, 2001; Quackenbush, 2001 and Yang and Speed, 2002).

In contrast, oligonucleotide microarrays use a single dye technology and pose some advantages, including a greatly increased density of genes and simplified experimental design because treatment effects are tested independently on each chip, eliminating the need for IBD designs. Nevertheless, statistical issues remain, such as normality of residuals, homogeneity of residual variance, correlation of errors within an array, and correlation of biological samples across arrays.

Mixed model methods for analysis of microarray experiment, proposed by Wolfinger et al. (2001), solves most of these issues (see Craig et al., 2003 for review). However, the complexity of analysis dramatically increases with these advanced methods. Unfortunately, many of the current practitioners of microarray technology do not possess the mathematical expertise necessary to meaningfully employ these methods. On the other hand ANOVA is a tool that is easy to implement with methods common to most researchers. Kerr and Churchill (2001) conclude that "The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors".

The first objective of this research was to examine the utility of simple ANOVA for analysis of replicated oligonucleotide microarrays experiments. The motivation was given eloquently by Kerr and Churchill (2001) who stated "An advantage of model based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering why it is inadequate can help the analyst identify sources of variation and bias." A secondary objective of this study was to show how using a one sided t-test can be used to increase power. The final objective was to introduce an alternative method to increase power by accepting a base number of false positive with high probability.

The ANOVA is particularly suited to analyzing data from microarray experiments that employ a replicated factorial arrangement of t reatments. An example of such an experimental design is one in which the investigator looks at gene expression in wild-type and mutant plants in the presence or absence of an added chemical. Many microarray studies incorporate this type of experimental design, e.g. the response of genes in non-

tumorigenic and tumorigenic tissues to different concentrations of toxic or therapeutic drugs (Lundquist, H. et al., 2002; Martinez, J.M. et al., 2002) or the response of genes from different tissues to estrogen or other hormones (Abe, H. et al., 2003; Faccioli, P. et al., 2002; Fujita, N. et al., 2003; Goda, H. et al., 2002). This design easily extends into any number of genotypes (or tissues) by any number of developmental time points (or biochemical exposures).

The primary biological objective of this research was to understand how a CHD3-chromatin remodeling factor, PICKLE (PKL), and a plant growth regulator, gibberellin (GA), regulate gene expression during germination of Arabidopsis seeds (Rider, S.D. et al., 2003). PKL is necessary for repression of embryonic traits in Arabidopsis (Ogas, J. et al., 1997). Expression of the embryonic state in pkl seedlings is inhibited by the plant growth regulator gibberellin (GA) and is enhanced by application of uniconazole-P, an inhibitor of GA biosynthesis (Izumi, K. et al., 1985; Ogas, J. et al., 1997). Specifically, gene expression was examined in wild-type and pkl seeds grown in the absence and presence of 10-8 M uniconazole-P. Thus the genotypes were 'wild type' vs. the pkl mutant, and biochemical exposure was to either 10-8 M uniconazole-P or no uniconazole-P during seed germination.

Our working hypothesis was that PKL functions during germination to repress genes that promote embryonic identity. In support of such a hypothesis, the transcript levels of two positive regulators of embryogenesis, LEAFY COTYLEDON1 (LEC1) and LEAFY COTYLEDON2 (LEC2) (Lotan, T. et al., 1998; Stone, S.L. et al., 2001), are elevated during germination of pkl seedlings (Ogas, J. et al., 1999; Rider, S.D. et al., 2003). Our interest was to find new genes that exhibited PKL-dependent expression, i.e. were up regulated. As such, we had a natural one sided test.

September 4, 2006

第六部分: 讀一篇要大改的文章

附上的論文是某作者投稿用的第一稿。他肯投出去, 大概以爲沒有問題了。我拿掉了原作者的名字和機關, 爲省空間, 也拿掉了所附的圖和表, 因爲我的目的不是要讓大家搞懂這論文。

這是典型的國人用英文寫的論文 —— 大、小毛病一大堆。最後被接受的形式, 和下面所顯示的頗不相同。

請挑出所有你能挑出的毛病 (包括大小寫的不當和標點的小錯)。用紅筆在錯的地方註上就好。

The impact of FDI on regional innovation capability:

a case in China

Abstract: FDI have been traditionally considered as an important channel in the diffusion of advanced technology. Whether it can promote technology progress for the host country or not is a focused problem in recently decades. This paper analyzes the relationship between FDI and regional innovation capability (RIC). We find that the spillover effects of FDI are not significant as we usually thought. The impact of FDI on RIC is weak; the entry of FDI has no use for enhancing indigenous innovation capability. The research manifests, increasing domestic R&D inputs, strengthening the innovation capabilities and technology absorbency in domestic enterprises are determinants to improve RIC.

Key words: FDI, less developed countries, regional innovation capability, spillover effect, the level of entrepreneurship

1. Introduction

Multinational companies (MNCs) play an important role in the process of global economic integration. Through capital outward, MNCs maintain tight connection with the international economy. The foreign direct investment (FDI) concomitant with MNCs has already become vital source for many less developed countries (LDCs) to obtain international capitals and advanced technology increasingly. Chinese government always attaches lots of importance to attracting foreign investment since the reformation and the "open-door" policy was carried. They implemented the "market for technology" policy, and tried to facilitate the technological progress through attracting inward FDI. China has already

received the total amount of foreign investments over 562.1 billions dollars until the end of 2004, and since 1993 China has been the largest FDI recipient in the LDCs. In 2003, despite of the breakout of the epidemic SARS, the amount of inward FDI reached 53.505 billions dollars which surpassed America and made China become the most inward FDI country all over the world. In 2004, the amount of inward FDI reached 60.63 billions dollars. Now the ratio of FDI to GDP has surpassed 40 percent. [Gong, 2005]

FDI is the important driving force to boost economic development in China. The impact of FDI has penetrated into many aspects of the national economy with the increase of total amount. The negative function appeared gradually such as the homogeneous expansion of Chinese manufacturing and the international trade dissension because of the excessive reliance of FDI and FDI technology. How about the strategic effect of the "market for technology" policy? Does FDI facilitate the technology progress in domestic enterprises? More and more Chinese scholars begin to think about the FDI strategy retrospectively. Many international scholars plunged into the vehement dispute.

As the main way of international capitals fluxion, the impact of FDI on the economy and technology to the host country has been paid more and more attention by international researchers. Many researches showed that FDI was important to facilitate economic development and technological progress for the host country, especially for the developing countries [Kokko, 1994; Kokko et al., 1996; Sjoholm, 1999; Borenztein et al., 1998]. However, the empirical research on the spillover effects of FDI didn't support standpoints above. Recently some empirical researches indicated that the spillover effects of MNCs were weak, and the positive effects of FDI on economy development should have some certain conditions [Young, 1992; Haddad and Harrison, 1993; Kokko, 1994; DeMello, 1996, 1997]. For many LDCs, the correlation between FDI and technological progress or productivity growth was not significant, except for those export-oriented countries [Balasubramanyam, Salisu, and Sapsford, 1996] and those countries having high-level of human capitals [Borensztein, DeGregorio, and Lee, 1998).

Whether FDI can bring the technology progress in China? Chinese scholars also have different opinions. Two scholars in Chinese Academy of Social Sciences Xiaojuan Jiang and Chunfa Wang raised some researches on this problem in 2001 and 2003 respectively. They both investigated the enterprises invested by MNCs and after analyzing the collected data, they obtained contrary conclusions. According to the view of Xiaojuan Jiang, FDI can boost technology progress. There need not any precondition. MNCs will definitely bring their advanced technology, machine and equipments to share Chinese market and consequently enlarge their proportions. While Chunfa Wang considered that the technology spillover effects of FDI were not distinct. Advanced machine and equipments are not equal to technological capabilities. On the contrary, FDI enterprises would decrease and squeeze out R&D activities in domestic enterprises.

Foreign direct investments are carried through by MNCs primarily. MNCs have obvious advantages relatively compared to domestic firms, and hold the most vigorous parts in the world economy. The contribution of MNCs through direct investments on technology transfer is obvious in theory. MNCs combine their own predominance including capital, technology, management skills, marketing channel, R&D and so forth with the advantages in the host country including nature resources, human resource and market scale and so forth to realize the advantage being repaired with each other in the whole world. MNCs

pay more great attention on the R&D of new products. The direct investments of MNCs may bring precious resources including capitals, technology, management skills, R&D capabilities, and the network of international trade for the host country. LDCs attract FDI; bring technology spillover effects through demonstration, imitation, reverse engineering, individual contact, diffusion of management skills, and the exploitation of international market. This is beneficial to shrink the gap in high-technology with developed countries, to upgrade the industrial technology in acceleration, and to fetch up the technology indentations and lag in the course of development. Furthermore, MNCs have stronger technology consciousness and skills; they make use of research institutions, universities and other service organizations more positively, so MNCs can facilitate the construction of national innovation system for the host country to some extent.

FDI may also bring negative spillover effects. The technology spillover effects of MNCs are likely to very feeble. It is one kind of crucial means that MNCs invest in LDCs in order to keep their own technology advantages and at the same time to share the cost advantages with LDCs. MNCs mostly invested the capital-intensive and technology-intensive industries which were laggard in China under the consideration of the global strategy and they aimed to occupy Chinese market in the long term. MNCs have obvious competitive advantages compared to domestic firms. They always bring more advanced technology and equipments through FDI. However, the spillover effects can't come into being automatically. Imbriani and Reganati (1997) represented that the spillover effects from FDI enterprises were in inverse proportion to the size of technology gap between foreign and domestic firms from the Italian evidence. Kokko (1994), Kokko, Tansini and Zejan (1996) also found that a positive and statistically significant spillover effect only in plants with a moderate technology gap. There almost were not any spillover effects when the technology in foreign firms was much higher than domestic firms according to the research of Mexico and Uruguay. Generally speaking the introduced technology must be suitable with the factor endowments in the host country. If the technology gap were too large, although there may be many opportunities for domestic firms to imitate and learn, domestic firms may have not enough technological capabilities to absorb and imitate advanced technology from FDI enterprises and can't make the advanced technology be endogenetic. Thus it results in little spillover effects. Because of the stickiness of information, most technology and knowledge are tacit knowledge. Only through practice can they be mastered. Just as the study of Borensztein (1998) showed that the introduction of more advanced technology and the requirement of absorptive capability in the host country were twinborn factors of economic growth. FDI was more productive than domestic investment only when the host country had a minimum threshold stock of human capital. DeMello (1999) discovered that FDI had positive effects to the countries with high technology and had negative effects on the followers. So the contribution of FDI not only depends on the technology level brought by foreign investment, but also depends to a large extent on the ability of absorption and assimilation of advanced technology of domestic firms. To acquire the spillover effects of FDI, the host country must be in possession of enough human capital that has received good education and trainings, and proceeds to reform and innovate constantly after utilizing the acquired technology effectively.

Because MNCs possess more advanced technology and management skills compared with domestic firms, their entries monopolize the original competition market in the host country, squeeze out the domestic firms in the industry and lower the market share of domestic enterprises. Kokko (1994) found that no evidence would show the positive effects of FDI on domestic productivity growth when the market share of MNCs was big and the

technology gap was too large. Demonstration and competition effects are really effective mechanism of spillover effects, but this must be based on corresponding technology capabilities in domestic firms. MNCs invest, incorporate and purchase the potential rivals in the host country, which may cause the innovation activities in the purchased enterprise be decreased, be transferred or be closed; furthermore decrease the regional R&D activities, make us excessively rely on foreign countries about key technology and bring the negative influence.

The majorities of MNCs invest in our country and engage in only the production and operation of the final product at present. Those key intermediate products were usually supplied by MNCs internally. So the linkage effects were limited. MNCs integrate with the Chinese base of manufacture very slowly to protect their own technology and the manufacturing expertise. Also, they seldom have relation with Chinese universities and research institutes. MNCs usually take the global production strategy and arrange every kind of functional behaviors in the whole world to excavate the competitive advantages on each step. Chinese enterprises are just one of its manufactured chains in the global production. This is manufacture transfer but not technology transfer. Technology circulation happened only within the MNCs. This is one kind of floating economic with bad groundwork. The advanced technology and equipment can't be transferred to domestic firms automatically. The technology stream can't be transformed into endogenetic technology capabilities easily.

The purpose of this paper is to analyze quantitatively the relationship between FDI and regional innovation capabilities (RIC) and to find the determinants of RIC. We will empirically do research on the correlation between FDI and RIC using data of each province in China; verify whether the more of inward FDI in a province will lead to a higher level of innovation capability. The paper is organized as follows. Section 2 presents our research method. Section 3 gives the concept of regional innovation capabilities. Section 4 researches empirically the relationship between FDI and RIC using multivariate statistics analysis. Section 5 constructs empirical model to analyze the impact of FDI on RIC further. Section 6 analyzes the effect of FDI to the level of entrepreneurship. Section 7 concludes.

2. Research design

In this paper, we focus on the relationship between FDI and RIC and highlight the role of FDI in the RIC in China. What is the regional innovation capability? This is the principal issue we should conduct on. So we make certain the concept of RIC firstly. In fact, the connotation of RIC is very abundant. Inward FDI is just one of the main factors of RIC; it means the ability of technology introducing and transfer. On the basis of the comprehension on the meaning of the concept, we establish a series of indictors to reflect RIC completely and objectively. Then we construct synthetic evaluating function. The score of the constructed function can represent RIC accurately.

We use principal component analysis of multivariate statistics to obtain synthetic evaluating function. We adopt the first principal component (PCR1) as the synthetic evaluating indictor to appraise RIC and calculate the correlation between FDI and RIC. Because PCR1 is the weighted sum of each original variable, and its variance is the biggest, it can

reflect majority of the information of all original data. The factor loadings of each original variable on PCR1 are the correlation of each original variable and PCR1. Through this analysis we can also obtain the determinants of RIC. We will elaborate on the characteristics of some typical regions. We conduct on detailed compare and analysis to summarize the rules of development of RIC.

We establish the empirical model with regression analysis to compare the importance of each factor that be related to RIC using the number of patent applications to measure RIC. Then we use correlation analysis through scatter diagram and the determinate coefficients of regression model to manifest the impact of FDI on the level of entrepreneurship. All analyses are dealt with SAS statistical software.

All data used in this research are obtained from China Statistical Yearbook, China High-tech Industry Statistical Yearbook, and China Science and Technology Statistical Yearbook. These three yearbooks are compiled by National Bureau of Statistic of China. Some of the variables can't be directly obtained and we proceed to some simple computation. All data are the most up-to-date currently available.

3. What is the regional innovation capability?

Technological innovation is a concept of economic category; it means the economic-technological activities including R&D, production and commercial applications of new technology (include new product and new craftwork). Technological innovation is one kind of commercial activities which creates new economic value by means of new technology (entirely new or modification). It is the first commercial application of new technology, and realizes the combination of economy and technology. Technological innovation has three particular characteristics. First, it emphasizes that the degree of market realization and acquisition of the business benefits are the ultimate standard to verify whether the innovation is successful or not. Second, it emphasizes that it is a systematic engineering from the research and development of new technology to the first commercial application. Third, it emphasizes that the business enterprises are the subjects of technological innovations.

Technological innovation has two types of realization modes, namely the independent innovation mode based on R&D, and the second-innovation mode based on introducing in and assimilating advanced technologies. The investments on independent innovation of new technology and new products are considerable, and the risk is also stupendous. It needs very strong capability of R&D and needs large amounts of funds for support. Because of the large technology gap between developed countries and China, the second-innovation mode is usually taken in China. With the fast development in technology and science, the scope of technological innovations become more and more wide, the competence of making use of the exterior knowledge has become an important part of technological innovation capability.

The regional innovation vitalities are enslaved to innovation capabilities. The regional innovation capabilities are the potentialities of producing streams of innovations related to commerce in a region. They refer to the capabilities of converting knowledge into new product, new craftwork, and new service. The regional innovation capabilities are not

merely the capabilities of science and technology, nor just the technological competition ability, for they pay more attention to the economic applications of new technology. The strong technological competition ability of a region doesn't means that the innovation capabilities are strong too. The regional innovation capabilities are made up of those factors as follows: Technology human resource, the creators of knowledge and employees who grasp the subjects of craftsmanship; the ability of knowledge fluxion, that is the ability of making use of all kinds of resources in the world constantly, and the ability of knowledge fluxion among each innovation units; the capabilities of technological innovation in enterprises; innovation environments and the economic performances of innovations, that is the output of innovations.

The innovation capabilities will decide the long-term economic competitiveness in a region. The innovation capabilities are the most important factor to explain the differences of the degree of economic prosperity among high-income nations. For LDCs to say, obtaining and developing the technology has become an essential driving force to improve the competitiveness [The Global Competitiveness Report 2002-2003]. For LDCs to say, introducing in advanced technology and assimilating, realizing the second-innovation is the shortcut to improve the regional innovation capabilities and international competitiveness.

4. The correlation analysis between FDI and RIC

4.1. To establish the series of indictors of RIC

To evaluate the correlation between FDI and RIC accurately, we need to describe RIC exactly first, so we must establish a series of indictors to evaluate RIC. The establishment of the series of indictors of RIC is the further step of comprehension on the essence and the meaning of the concept of innovation, and is also the development of innovation theory. The connotation of RIC is very abundant; the determinants include education, science and technology resources, innovation capability in enterprises, regional synthetic strength and information condition; and also include the regional policy and management skill. The series of indictors of RIC must reflect the present conditions and the utilization efficiency of regional knowledge and technology objectively and try to evaluate the RIC completely and objectively. The indictors being chosen should be brief and terse. On the basis of reflecting RIC accurately, try to select those synthetic indictors which have the commonness and are maneuverable. Therefore, we select 22 indictors altogether to measure RIC as listed in table 1.

Table 1 about here

4.2 The principle component analysis

We use the principal component analysis to establish the synthetic variable evaluating RIC of each province. Its excellence is that the weights are based on the inherent configura-

tion relationships of the variables from the data analysis, there doesn't have any influence of subjectivity, and is objective perfectly. This is beneficial to analysis and appraise synthetically. We analyzed the data of 31 provinces in China about above 22 variables with principal component analysis, using SAS software to deal with. The biggest eigenvalue of the correlation matrix values 11.1167 and the first principal component reflects 50.53

Table 2 about here

By the eigenvector of the biggest eigenvalue in table2, we can get the expression of the first principal component (PCR1) as following:

$$\begin{aligned} PCR1 = {} & 0.1237RFUND + 0.1338FDI + 0.2793TVMARKET - 0.0814EXPPT \\ & + 0.1971EXPIT + 0.1107RR\&D + 0.0818RS\&T + 0.2875UTILITY \\ & + 0.2380DESIGN + 0.0727OVEQUIP - 0.0136EXPTR \\ & + 0.2426NEWSALES + 0.2494NEWRATE + 0.2589COLLEGE \\ & + 0.2781HIGHEDU + 0.2611BOOKEXP + 0.2685NS\&E \\ & + 0.2237NPTE + 0.2303NHTE + 0.2597NNE \\ & + 0.2871PGDP + 0.1447LPRODUCT \end{aligned}$$

The expression shows, PCR1 is the weighted sum of each original variable, most of the coefficients are positive excluding two indictors of EXPPT and EXPTR. Indictor EXPPT means the average expenditure on purchase of domestic technology of each enterprise. Generally speaking, the higher of technological capabilities, the less domestic technology enterprises would purchase. So the negative value of the coefficient is reasonable. Indictor EXPTR means the average expenditure on technology renovation of each enterprise. The coefficient is negative, but the number values only -0.0136, so its influences is not big. We can neglect it. Thus PCR1 can reflect the innovation capability of each province synthetically. So we use PCR1 as the synthetic evaluating indictor to appraise RIC.

4.3 The impact of FDI on RIC

The loadings in table 2 are the factor loadings of each original variable on PCR1. Factor loadings are the correlations of each original variable and the principal component in fact. The correlation coefficient between FDI and PCR1 is only 0.4462, there has no significant correlation between FDI and RIC. By table 2 we find that the variables having high relationship with RIC include: TVMARKET (respects the ability of technology transfer), UTILITY and DESIGN (respect the ability of design in enterprises), NEWSALES and NEWRATE (respect the ability of innovation output in enterprises), COLLEGE, HIGHEDU, BOOKEXP (respect the caliber of employees0, NS&E (respects the human resource in S&T), NNE (respect the level of entrepreneurship) and PGDP (respects the performance of innovation). Therefore, for a nation or a region to say, the determinants of innovation capability include: paying attention to the investments of S&T, the ability of upgrading the human resource in S&T and making full use of all kinds of resources

of S&T; the technological innovation capability in enterprises; having good environments beneficial to innovations, the mechanism and system that is open and can make full use of the local special resources and the global S&T resources; the outstanding performance of local economic, which become the powerful pull force of innovation. We analyze the correlation between FDI and RIC further using the synthetic variable obtained from the principal component analysis, drawing the scatter diagram, establishing the regression model and obtaining the regression curve. See as figure 1.

Figure 1 about here

We can observe the relationship between FDI and regional innovation capability clearly by picture 1. From the scatter of points that represent each region in picture 1, we can discover that there have many points far away with the regression curve, such as Beijing, Shanghai, Tianjin, Guangdong and Jiangsu. The determinate coefficient R2 is only 0.1991, it indicates that FDI can explain 19.91

By picture 1, we can also find that there have a lot of points almost gathered together. These points represent the provinces in which both FDI and RIC are all lagged behind. These provinces locate in the centered or western region of China, are less developed in economy because of the lagged notion, policy and geographic location. The behindhand innovation capabilities are mainly induced by the lagged economic conditions. To eliminate the influence of economic development, we calculate the correlation again canceling the regions in which both FDI and RIC are all lagged behind and obtain the correlation is minus 0.3626. This indicates that FDI has the negative impact on RIC.

FDI in Guangdong and Jiangsu are the biggest. The amount of FDI in Guangdong is 11.334 billions dollars in 2002, and 10.19 billions dollars in Jiangsu. The amount preponderated over other regions greatly. But their regional innovation capabilities lie in the fourth and the fifth respectively, and just bigger than the latter regions a little. Then we compare each original variable carefully and discover that on three variables that respect the caliber of employees, which are number of college and higher level per 10000 populations (COLLEGE), number of graduates from institutions of higher education per 10000 populations (HIGHEDU), the books sales per capita (BOOKEXP), Guangdong located in 12 and 13 and 21respectively in the whole country. Variable that respects the human resource in S&T, number of scientists & engineers per 10000 populations (NS&E), Guangdong located in No.7. And Jiangsu located in 20, 6, 4and 5 on above variables respectively. These show that the level of human capital doesn't match the amount of FDI in Guangdong and Jiangsu. Whether the stock of human capital is high or low will decide the spillover effects directly, and influence a nation's independent innovation capabilities and potentiality directly. Only when the volume of FDI match the stock of human capital and technological capabilities, can the RIC be developed and enhanced.

The region with the strongest RIC is Shanghai, followed by Beijing, while FDI in these two regions listed in the fourth and the eighth respectively. What drive the innovation capability in Beijing and Shanghai? Obviously we can't get the reasonable explain only from the FDI. The superiority in Beijing include the abundant S&T resources, strong abil-

ity of creating new knowledge, attaching importance to the investments for S&T, having the most excellent employees, and the good environments for innovation. While in Shanghai, there have a good foundation for industrial innovation; the ability of technological innovation of enterprises in Shanghai keep ahead in the whole country; and there have the strong financial strength and powerful capital advantage. Zhejiang is the province in which private economy developed very well, the capability of innovation in private enterprise is extraordinary. Zhejiang developed quickly because of the dual superiorities of S&T and the system. People have great enthusiasm to start-up an enterprise. There have the perfect system for regional scientific and technological innovation, investment and financing and supporting for the talented person in Zhejiang. Zhejiang is the apotheosis of making full use of local resources to improve the innovation capability.

Therefore, being open to external, attracting FDI is not the unique means to increase RIC, and it isn't the important means either. The speed of reformation and the degree of marketization in economy will decide the level of innovation, the notion and system will influence the innovation capability. The determinants of the innovation capability include the original drive force of innovation, the technological innovation capability in enterprises, the ability of making full use of the local special resources and all kinds of S&T resources, having the good environments beneficial to innovation, perfect mechanism and system for regional scientific and technological innovation. Besides above factors, the local economic condition is the strong pull force for innovation. On the other hand, the level of GDP will be high in the regions that have strong innovation capability, and the international competitiveness will be strong too.

5. The empirical model of the effect of FDI to the innovation capability

We measure the innovation capability with various different variables in the above analysis. But scholars often use the patent data to measure the technological innovation, because when an inventor or creator apply patent to the patent censorship, this is usually the potential sign of economic value of innovation and representation of the innovation capability. Furthermore the patent data are complete, accurate and can be obtained easily.

In less than 20 years, China has made tremendous progress in establishing a legal system for the protection of innovation. China's first patent law was enacted in 1984 and came into effect in 1985. Since then, the law has been amended twice. Since the passage of the 1984 patent law, the central government has issued over 20 regulations and guidelines so as to promote innovation activity in China. Today's patent law in China is pretty much in line with the international standard [Kui-yin Cheung, 2004].

The patent law of China divides patents into three categories: invention, utility model, and external design. The invention patent refers to the new techno-project put forward on the product, method or its modification, which can form the products having independent intelligent property. The utility model patent means the new practical techno-project put forward on the shape, structure of the products or their combinations. The external design patent is the new design which is full of pleasant impression and suited for the application in industry about the shape, pattern, color, or their combinations of the products. The innovative level is relative low of the utility model patent and the external design patent. The term of protection is 20 years for invention patents and 10 years for utility model and

external design patents. Among three types of patents the invention patents are regarded as major innovations, they can represent the innovation level mostly.

The number of patent authorized was influenced by the abilities of the patent censorship, so in this text we use the number of patent applications in our country to measure the regional innovation capabilities.

We use the following model to estimate the spillover effects of FDI on innovation capabilities in China.

$$Patent = f(FDI, PGDP, S\&TEXP)$$

We use the number of patent applications (Patent) as a measure of innovation capability. FDI refers to the realized values of FDI lagged one period considering that FDI inflow to China impacts on domestic innovations within a short period of time. As measures of input to R&D activity, we use expenditures on science and technology development (S&TEXP). Finally, considering the fact that different provinces are at different stage of economic development so that their innovation capabilities should also differ, we include the level of per capita GDP (PGDP) in our estimation.

The model can also be expressed as: $Patent = A \cdot FDI^{\alpha} PGDP^{\beta} S\&TEXP^{\gamma}$, then we take logarithm in both sides, and the model can be written as following:

$$LnPatent = C + \alpha \ln FDI + \beta \ln PGDP + \gamma \ln S\&TEXP + u$$

Where C is the estimated intercept of the equation, u is the statistic error, the coefficient α, β and γ is the elasticity of the increase of FDI, PGDP, S&TEXP to the patent increase respectively. For convenience, the equation was written as follows:

$$Patent = \beta_0 + \beta_1 FDI + \beta_2 PGDP + \beta_3 S\&TEXP$$

The coefficients $\beta_1, \beta_2, \beta_3$ measure magnitude of the influence of FDI, PGDP and S&TEXP respectively. The data are taken from China Statistical Yearbook and China Technology Statistical Yearbook, covering 30 provinces. Tibet is excluded in our analysis because most of the relevant data for it is either not available or zero during the time period examined.

We used SAS software to estimate the equation with the ordinary least squares (OLS). And we estimated the equation for each type of patent (invention, utility model, and design) as well as the total patent applications.

Table 3 about here

By table3, the determinate coefficients R^2 of four models are all above 0.77, and the values of F-statistic are significant at the level of 0.0001. This shows that four models all have great statistic significance and they can explain the variations of all types of patent

applications more than 77 percents. For the amount of invention patent, the coefficient of FDI is positive, but it is statistically insignificant even under the level of 0.10. So FDI has no significant effect to the amount of invention patent. There has positive effect for FDI to the amount of utility patent, and has significant effect to the amount of external design patent and total patent. S&TEXP has significant effect to the amount of all three types of patent and the total patent under the significant level of 0.0001, except the external design patent is under the significant level of 0.05.

The results manifest that S&TEXP is the most important factor to increase the innovation capabilities. FDI has positive spillover effects to some extent, but has no significant effect to increase creative inventions and the inventions having independent intelligent property rights.

6. The effect to the level of entrepreneurship

We use the number of new registered enterprises (NNE) and private technology enterprises (NPTE) and high technology enterprises (NHTE) as a measure of entrepreneurial level. FDI refers to the realized values of FDI lagged one period considering that FDI inflow to China impacts on innovation activities within a short period of time.

The effect of FDI to the entrepreneurial level can be showed as the following figures.

Figures 2, 3, 4 about here

We observe three scatter plots in which each point represent each province, and find that there have many points far away with the red regression curve. Three correlation coefficients between FDI and the number of all kinds of new enterprises and the determinate coefficients of regression equations are all very low. So we can make conclusions that FDI has no direct significant effect to the level of entrepreneurship. The more FDI will not bring the higher entrepreneurial level.

This result may be induced by the crowing-out effect of FDI. Because of the high risk to start-up an enterprise, a successful entrepreneur must be adventurous pioneer daring to take the tremendous risk and failure. The presence of foreign enterprises may provide higher wage for employees and increase the competition in the market, this increase the difficulty of successfully starting an enterprise. So it is likely that many people having certain entrepreneur spirits would rather select the secure and steady job in foreign enterprises than undertake the risk to carve out.

7. Conclusions and policy implications

As far as LDCs are concerned, introducing in advanced technology is the shortcut to facilitate the technology progress. FDI has been regarded as an important channel for technology diffusion. However, the results in this paper demonstrate as follows.

First, the correlation between FDI and RIC is insignificant statistically. The impact of FDI on the regional innovation capability is inappreciable. The regions which attract more FDI haven't the higher regional innovation capability. Only when the volume of FDI match the stock of human capital and technological capabilities, can the RIC be developed and enhanced. This finding is the same to Borensztein (1998). Attracting FDI is not the unique means to increase RIC, and it isn't the important means either. The determinants of RIC include the original driving force of innovation, the technological innovation capability in enterprises, the ability of making full use of the local special resources and all kinds of S&T resources, having the good environments beneficial to innovation, and the local economic conditions.

Second, the results of the statistical model indicate that investment in R&D activities is the most important factor to enhance the innovation capabilities. FDI has positive spillover effects to some extent, but has no significant effect to increase creative inventions and indigenous innovation capabilities.

Third, through three scatter diagrams and correlation coefficients between FDI and the number of new enterprises and the determinate coefficients of regression equations, we find that FDI has no direct significant effect to the level of entrepreneurship.

To sum up above results, MNCs have the technological advantages to some certain; the inward FDI will facilitate the technology progress and improve the regional innovation capabilities in some degree. But we can't think blindly that we introduce in FDI the more the better. The advanced technology and equipments of FDI can't be transferred to domestic firms automatically. There have no significant correlation between FDI and RIC, the more of FDI will not necessarily bring the stronger innovation capabilities and spirits. So we can barely rely on FDI to improve RIC. The determinants to improve RIC include such factors as following: increasing domestic R&D investments, enhancing the stock of human capitals, improving the technological innovation capabilities and technology absorbency in domestic enterprises, having good environments for innovations.

The findings in this paper may provide some insights for both the host countries and foreign investors. Based on above analysis we put forward some policy implications as follows, which may also be effective to other LDCs. Firstly, it is urgent for China to improve the stock of human capital and indigenous R&D capabilities. Whether the stock of human capital is high or low will decide the spillover effects directly, and influence a nation's indigenous innovation capabilities and potentiality directly. It is necessary for China to enlarge investments in fostering human capitals, especially the investments for fostering R&D personnels. It is necessary for China to strengthen cultivating the independent R&D capabilities, particularly supporting for the R&D activities of techno-intensive industries. These are important to shrink the technology gap between domestic enterprises and foreign capital enterprises. Only when the host country has certain technological talented persons, can MNCs arrange R&D projects in the host country and train the native high technological talented persons, to reduce the cost of human resource and make products locally better. Therefore making native innovative systems perfect, cultivating R&D capabilities of domestic enterprises positively, promoting and stimulating independent innovations in domestic enterprises are the basis for improving technological progress from making use of the spillover effects of FDI. Secondly, Chinese government should focus on the quality of inward FDI and insist on the sticking point to advance independent innovation capa-

bilities. Try to urge and guide FDI to be on the trajectory which is beneficial to improve our independent innovation capabilities. Thirdly, China should try to create a fair competition environment for domestic enterprises to compete with foreign enterprises, and keep sufficient rivalrousness of the market. Under the intense competition environments it may force MNCs to increase the degree of technology transfer to our country.

References are omitted, end of paper for correction.

September 4, 2006

第七部分: 讀一篇關於自助法的論文

在這一部分, 我們選出 B. Efron 的一篇在 1979 年的論文來細讀 —— 主要在探討他的寫法, 論文寫得好的境界有三: 初級爲易售, 讓評審覺得可以刊登; 再進一步是眞有內容, 此時, 即使寫得略差, 也不會受太大的影響。最糟是其實沒有實質內容但寫得花團錦簇[1]。

此文爲 1977 年的 Rietz 的獎座論文。這是國際數理統計學會 (Institute of Mathematical Statistics, IMS) 在它的年會中所邀請的主要演講。這演講會在會場的最大廳內舉行, 且同時沒有其他的論文在宣讀。能受到邀請的學者皆非等閑之輩 —— 事實上此文在 Annals 上根本不會被拒絕, 因爲否則 IMS 豈不是自己打自己的嘴巴?

Rietz 講座因爲要面對較大的聽衆 (雖然 IMS 的會員都是以數理統計爲主, 但也不是人人皆是高手), 因此這樣的論文也不會寫得太深。但更不會太淺, 因爲否則又豈不弱了被邀演講者的名頭? 對這類的文章, 作者有較大的期望, 希望能 (對統計科學) 發生影響。因此他會寫得不僅能懂而且能用, 且同時讓讀者覺得這文章有學問。

這篇文章的全文, 見 *Annals of Statistics* **7**, 1-26 。下面的註脚, 要配合原文來讀, 可以當作此文的「導讀[2]」來用。我們的目的不是教你自助法, 而是介紹你這論文的寫作。

論文的標題是 "Bootstrap methods: Another look at the jackknife". 用了兩個名詞: bootstrap 和 jackknife 。這兩件事都是「重抽法 (resampling method) 」。在演講的時候, jackknife 已廣爲人知, 但 bootstrap 卻是新的東西。以下爲我的註脚。

---

p.1, abstract, line 1[3]. random sample 指的是 $X_1, X_2, \ldots, X_n \sim$ iid F 。此處直接引入 $R(\mathbf{X}, F)$ 這個統計量, 這個統計量是本文今後要討論的主弦律。注意到它有未知的分布 F 在內。這和我們熟知的統計量如 $\bar{X}, s^2$ 有異 —— 它們不包含未知參數[4] 。但目前你可將它想像爲 $\bar{X} - E(X)$ 或者 $s^2/Var(X)$ 之類的統計量。

p.1, abstract, line 4. 在此直接提 jackknife , 並和本文目的做一個區隔。本文可看成爲 jackknife 的推廣, 但作者並不這樣看, 這樣看就把本文看小了。他將 bootstrap 看成爲一種自然的功法, 而將 jackknife 看成爲一種類似的, 但較低層次 (因此反而較複雜、較不清楚明白) 的工具。當然, 在寫此文的時候, 多數人知道 jackknife , 而 bootstrap 則正要推出, 知者甚少[5]。所以他用 jackknife 作爲比較也是合理的。此時, 他對 jackknife 既不能捧也不能罵, 因此說 "is a linear approximation of jackknife" —— 但此話可不能空口說, 後文得要有本事證明。

p.1, introduction, line 5. 用 "attempt to explain" 是恰當的用語。這回答了「何謂

[1]「花團錦簇」是舊小說中形容八股文寫得好的用語。

[2]這裡當然充滿了我的看法 (如果不是偏見), 也不盡然是作者的原意。此文我每讀一次, 都有些新的領悟, 因此這一次不到之處, 必然也是有的。

[3]這指從 abstract 起第一行, 下同。

[4]在無母數問題中,F 可視作未知參數, 但它是「無窮維 (infinite dimension)」的。

[5]事實上眞在這一行中的人應該早已知道。開玩笑, Stanford 的統計系正預備推出的一套方法, 如果你還想在這個行業競爭, 豈有沒聽過一點風聲之理? 在第一線作戰的戰士, 耳朵當然是豎起來的。

jackknife」這樣的較深層次的問題。

p.1, introduction, line 8. "more dependable" 在後面可是要用例子說明的。故下一句立刻說: 對於 sample median 而言, bootstrap 就比 jackknife 有用 —— 因爲已知 jackknife "is known to fail".

p.1, introduction, line 10. 立刻又說, 對於 linear discrimation 中, bootstrap 比 cross validation (另一種已知的重抽法, 常在 linear discrimation 中使用) 要好。

p.1, introduction, line 13. 此一段又提 jackknife , 說可以用 delta method 來看出 jackknife 可以視作爲 bootstrap 的一次微分。並說因爲可以這樣看, 不止對 jackknife 可以提出理論基礎, 並因此可以回答一些原先 jackknife 需要回答但一直說不清楚的問題。此段的末尾還說, 若有額外條件 —— 如已知 F 爲對稱或者 smooth —— 該如何做 bootstrap 。最後, 更進一步提到, 對於迴歸, bootstrap 又該如何做[6]?

p.1, introduction, line 21. 此一段說本文以舉例爲主, 對於一般性的 bootstrap 理論著墨不多[7]。

p.2, line 3. 此一段承認以前各作者的功勞。如 [5, 6] 爲 "ideas closely related" 。注意他不承認 100% 相同, 因此否則自已就沒有貢獻了。又說 [13] 中的方法可以導出本文的公式 (3.4) , 又謊 [10] 中的 infinitesimal jackknife 直接和本文的 method 3 相關。這一段是預告, 將後文的 method $i, i = 1, 2, 3$ 都一一交待。

注意到 [10] 並非一般論文, 而是 Bell Labs 的 technical report 。一般在重要論文中是不引用技術報告的, 但 [10] 卻是有名的一篇文章, 內行人都知道。所以可以在本文中大方引用。

本文最後提出 parametric bootstrap, 以印證本文所討論的 non-parametric bootstrap 爲主, 但 parametric bootstrap 的道理是類似的, 且和 R. A. Fisher 的 information bound 一致。注意到 Fisher 基本上已解決了所有傳統型的 parametric inference 的問題, 可由此顯示出 bootstrap 的不同凡響 —— 可以大小古今通吃。

p.2, section 2, line 6. 此處 "plays no role" 一句, 是爲了後面的例子 (如相關係數) 做伏筆。在此也等於是說, 此例雖然 F 爲一維, 但 bootstrap 的應用則不限於此。

p.2, section 2, line 12. 此一段用簡單的一段文字介紹 jackknife, 並暗示 jackknife 的不足: (1) 只有興趣在 bias 和 variance; (2) jackknife 只有興趣處於 (2.3) 或 (2.4) 這種形式的統計量[8]。

p.3, lines 3-16. 這一段直接介紹 bootstrap 的運作方式 (但 line 10 中所提的 "drawing samples of size $n - 1$ without replacement" 是一直沒有講清楚的)。

p.3, lines 17-23. 這幾行不那樣易懂, 但卻是 bootstrap 的基本道理。"Fisher consistency" 並不那樣 well-known , 但打出 Fisher 的大名, 不懂的人只好承認自己的學問不

[6]這裡提一下, 一般的統計方法, 若無法推展到迴歸型的數據, 就不能算是完整。所以他提「能做迴歸」, 是有「本研究基本已完備」的暗示。

[7]不要以爲 Stanford 的人不做 bootstrap 的理論。相信此時作者必有些學生在努力。甚或已有一些成果。要比這些人快, 你需要更快, 並且要消息靈通。去國外某學術單位訪問, 這是一個目的。

[8]其實本文中 bootstrap 也多在對付這類統計量, 但因在架構上 $R(\mathbf{X}, F)$ 是廣泛得多, 所以看起來較高級。

夠。bootstrap 的精神是: 用 R^* 的分布來近似 R 的分布。爲何這竟是對的? 我曾有較數學的表示法如下。$R(\mathbf{X}, F)$ 的分布可以用它的 characteristic function (ch.f)

$$Ee^{itR} = \int e^{itR(\mathbf{x},F)} dF(\mathbf{x}) \tag{1}$$

來表現。同理, $R^* = R(\mathbf{X}^*, \hat{F})$ 的分布, 則用類似的 ch.f

$$Ee^{itR^*} = \int e^{itR(\mathbf{x}^*,\hat{F})} d\hat{F}(\mathbf{x}^*) \tag{2}$$

來看。在 (1) 和 (2) 之間, 有兩點相異: $\mathbf{x}$ vs. $\mathbf{x}^*$, F vs. $\hat{F}$ 。但仔細看一下, (2) 中的 $\mathbf{x}^*$, 其實是一個用來指示對誰積分的 dummy variable 。若在 (2) 式裡的積分中令 $\mathbf{x} = \mathbf{x}^*$, 則 (2) 變成

$$\int e^{itR(\mathbf{x}^*,\hat{F})} d\hat{F}(\mathbf{x}^*) = \int e^{itR(\mathbf{x},\hat{F})} d\hat{F}(\mathbf{x}). \tag{3}$$

比較 (1) 和 (3) 則知道若是 $F = \hat{F}$, 則 (1) 和 (3) 是完全一樣的。因此二者的分布在此時是一樣的。

p.3. lines 24-29. 這一段的好句子是 "let the varying degree of success of the examples speak for themselves" —— 這又暗示了本文大量舉例的原因。

p.3, line 30 及以下。這是做研究的基本方法之一: 用最簡單的 non-trivial case 來驗證理論。此處 $X_i \sim$ iid $b(1; p)$ 是幾乎不能再簡單的情形, 用前述的 bootstrap 方式來做, 是否會得到理想的結論? 如果有, 可以向下面繼續做; 如果沒有, 就得去找原因。

p.4, line 4. 用 "universally familier" 一語帶過多少教本上的結果,

p.4, lines 5-16. 這一段是可以硬算的。此處舉的是一個較複雜的例子。注意到因爲 $\bar{X}$ 的情形等於已做過, 雖然那是最簡單的情形, 此就再做就沒有意思了。因此此處用 sceond moment 未討論, 以示和 binomial case 有基本的不同。至於引用 [3] , 也是寫作技巧之一。這 Cramer (1946) 可是經典的古典教本。你若是隨便引用一本 Statistics in Business, 看在評審人的眼裡, 你的份量就全不是那回事了。

p.4, line 19 及以後。此處提出三種方法來算 R^* 的分布。最後當然大多數人會用 method 2, 但方法一是用來印證用的: 能做解析計算時還是要做解析計算。Simulation 終究不如 exact 的計算結果。方法三其實是無用的, 在本文中有用, 是因爲我們要靠它建立 bootstrap 和 jackknife 間的關係。

但此文若沒有和 jackknife 的一段姻緣, 文章的支撐就少一些。所以這一文字是可加上的。此外, 用三個角度來看 bootstrap , 也表現作者上功夫。

p.4, line 29 及以後。介紹本文的其它內容。主要指的是 2-sample 問題的處理。雖然指的是 section 4 和 section 6 (沒有連貫), 但二者都是 2-sample problems, 故可放在同一段落。

p.5, section 3. 本節以硬算爲主, 要點在 p.6, lines 8-13. 這幾行用明確的事實說明 "bootstrap works for sample median, while jackknife does not" 。注意到 [11] 也是經典教本 (且引用時直指 p.237 —— 這是應該的。直接引用某書的時候, 應結出較細的資訊。如 [3], Section 27.4 (p.4, line 11).)。這表示兩件事: (1) 作者對讀者負責, 不希望你在

一本 500pp 的書中去找他所引用的某件事, (2) 表示作者眞的讀過, 不是鬼混的。至於指 [14], p.8 有錯一事, 一般是有風險的。但若你眞的挑得出錯, 也可以明說。此處 [14] 的作者 Miller 是 Efron 的博導, 二人關係良好, 是無所謂的。

p.6, line 14 – p.8 end of Table 1. 這兩頁若是由一個小教授來寫, 多半會被要求刪去。但 Eforn 是有名的大教授, 所以可以多用篇幅。重點在 table 1 後面所說的 "the most notiable feature ..." 。有這一句才能 justify 這 兩頁的篇幅。並同時強調了最簡單的 bootstrap , 其實非常不錯。至於極力地將問題中的額外條件 (如對稱、平滑) 用進 bootstrap 的努力, 其實是沒有甚麼大道理的。

如果在 1979 年, 如果你數學不錯, 一個立刻可以做的題目是: 用數學來證明或反證這兩件事。此文只用了 $n = 13$ 而已。那麼, 當 $n \to \infty$ 時, 情形會如何? 對數學好的學者, 這不會太難。而在 1979, 這類文章多半可以刊載。我的直覺是: (1) 你必需手脚要快, (2) 二者的差別應在 $O(n^{-1/2})$ 的項以外, (3) 這種文章即使能做也沒甚麼了不起, 你只是證明你技術不錯。

但現在是 2006 , 你必須好好查一下才會知道此事有沒有已經發表。而 p.8 的最後一段是作者的感覺, 雖然這不會是無的放矢, 但也不會有眞正的大意義。明顯地, Efron 目前根本不想追究這些現象。

這一節只有兩個要點: (1) 對 sample median 而言 (因爲已知標準答案), bootstrap 較 jackknife 要好: 前者做對, 後者做錯。(2) 簡單的 bootstrap 不比複雜的 bootstrap 要差。這裡面, (1) 是數學證明, (2) 則只是模擬的結果。但作者巧妙地安排 bootstrap 的出場場景, 並用 jackknife 來陪葬, 雖然不是自己的計算 (作者大方承認, 本節的主要技術來自 [13])。這是重要的, 不要掠美 —— 這是學術論文的規矩。

p.9, section 4.

本節又舉了一個例子: 對於 discriminant analysis 中的 error rate 的估計量, bootstrap 要比 cross validation 要強。

全節都是在用 simulation 來做 (道理當然是算不出解析的結果)。這個例子和前面的 sample median 不同, 否則就沒有甚麼意義。它的不同點在: (1) 這是 2-sample 問題; (2) cross validation 和 jackknife 是不同的事; (3) 我們在此算不出 exact 的結果。因此 method 2 是唯一的方法。並且, 到目前爲止, method 2 尙沒有舉例。

主要的結論在 p.10 的 Table 2 中: 差別是 0.09 vs. 0.03, 有三倍的不同。

本節沒有理論, 但模擬的步驟交待得十分清楚 —— 我可以根據作者的描述, 寫出一個程式同樣做出作者的結果。這也是科學論文的要點之一: 所描述的實驗[9] 資料, 應豐富到別人也能據此做出同樣的實驗的程度[10]。這是「可重覆性 (repeatability)」的要求。

此節除了選題正確 (需要一個 well-known 的問題, 它還得有一個 well-known 的解, 這個解還得比不上 bootstrap), 且本例的各種條件都尙未被討論過, 因此除了寫得仔細之外, 並不算難。可提出的是, 在 p.10 的 (4.7) 中, 要用 $[X_i \in B^*]$ 而非 $[X_i \in B]$; 在 p.11, line 19 處, 要指出該用 m 而非 $m - 1$ 。這些都是應該有理由的, 想來眞正的專家都知道。但我們只講文章的寫法, 而非導出結果的數學。我們就不提了。

[9] 模擬當然是一種實驗, 儀器是電腦。

[10] 道理是: 別人有辦法查證你的實驗是否做對了。

要注意的是 p.9, (4.1) 中的「符號成對」時的寫法。在字母裡, 先 X 再 Y, 先有 i 才有 j , 先有 m 才有 n 。因此

$$X \quad \text{用足標} \quad i \text{ 計有 } m \text{ 個}$$
$$Y \quad \text{用足標} \quad j \text{ 計有 } n \text{ 個}$$

其中隱含的次序, 數學上一點都不重要。但如此寫將出來, 會讓讀者在唸這一段時少用很多心力。並且你自己做計算時也會多費心力。例如 (4.2) 中的分母自然是 m , 不會是 n 。如果我們將 (4.1) 寫成

$$X_i = x_i, \quad X_i \sim_{\text{ind}} F, \quad i = 1, 2, \ldots, m$$
$$Y_i = y_i, \quad X_i \sim_{\text{ind}} F, \quad i = 1, 2, \ldots, n$$

雖然數學上正確, 但就不夠漂亮了。即使我們寫成

$$X_i = x_i, \quad X_i \sim_{\text{ind}} F, \quad i = 1, 2, \ldots, m$$
$$Y_k = y_k, \quad X_k \sim_{\text{ind}} F, \quad k = 1, 2, \ldots, n$$

也不夠漂亮。這類的小小化妝, 也是要學的。

p.12, section 5, line 7. 這一節目的是建立 bootstrap 和 jackknife 的關係, 是和論文標題有關的主要結論。在 line 7 處, 就讚揚 [10] 是 excellent paper 。此事世間已有公論, 所以誇講一下也無所謂[11]。何況 bootstrap 還要用 jackknife 來做墊背用!墊背夠好, 被墊的文章當然會更好。

這一段的數學技巧是將 X^* 看成爲多項分布。這是 standard trick , 因爲原來的 $X_i \sim F$ 用到 X_i^* , 就變成了 P^* , 而後者的分布是可以寫出 exact form 的。技術上, 等於是由一個 non-parametric 的架構, 轉換到 parametric 的架構上去[12]。

另一個要點是 R 在這一段中完全沒有限制。因此此段說的是: 一般的 bootstrap 和一般的 jackknife 間的關係。當然, (5.11) 所得到的結果, 和 [10] 中所得一絲不差 —— 這才是本文中較硬的結果。此事到了 p.14 的第一段結束, 才算眞正回到傳統的 jackknife 。

這一段要唸得眞懂, 得找張紙找枝筆跟著做。此外, 需得去讀 [10][13] 或者 [14] 。

這一節以 2006 年的角度來看, 並不重要 —— 因爲 bootstrap 已可以獨立站起。但本文的標題是 "Bootstrap methods: Another look at the jackknife", 因此,「借用 jackknife 以成就本文」的味道很濃。而這一節就變得有意義。任何情形下, jackknife 和 bootstrap 都是重抽法 (resampling method) , 因此, 在第一篇介紹 bootstrap 的文章中,「和 jackknife 比一比」,「和 jackknife 拉上關係」都是有意思的。這一節雖然對後來使用 bootstrap 的人無用, 但對於 referee 的影響, 卻是有一點用的 —— 至少, 這一節讓我們對於 bootstrap 和 jackknife 都有較深的了解, 而了解卻是研究的目的。

本節的最後一段有點多餘, 只是回應前面的對稱情形, 再多晃一招而已[14]。最後, (5.15) 中最後一行中 $\bar{y}$ 的足標是多餘的。

p.14, section 6. 這一節又舉了一個例子, 例子性質又與以前的例子不同: (1) 它做的是標準的 2-sample case , 因此可望由此看出一般的迴歸該怎麼做[15]。(2) 對於 2-sample

[11]但對不怎麼樣的論文, 就別亂捧: 可能有人要吃味, 也可能讓評審人覺得你品味不足。
[12]嚴格地講, 因爲還有 $x_1, x_2, \ldots, x_n$ 的值可視爲有 n 個未知參數的 pamametric model 。
[13]現在大概是找不到它了。因爲 Bell Labs 已變成 Lucent , 而後者差不多要垮了。
[14]我不敢說一定是虛晃一招。
[15]結果迴歸未明顯地由此看出。

問題而言, 尚有不知如何處理 jackknife 的問題存在。而經由 bootstrap 的角度以觀之, 可望能解決 jackknife 這個問題。(3) 我們可算出一些 exact 的結果 [公式 (6.7)], 亦可經由 method 2 (delta method) 算出近似的結果 [公式 (6.13)] , 而這兩者相差極少 —— 從而說明 delta method 用在 bootstrap 上也是有道理的。

在 p.16, (6.14) 之後, 明白說明如何用 jackknife 來做一個 2-sample 的問題, 這解快了一個未決的問題。

p.17, section 7. 對這一節我的感覺不好。主要的是其技術內容, 和前面的 2-sample case 似無交集, 也不搭調。也許從 Wilcoxon test 直接到到 regression 是太快了, 我看不出之間的關連。(2.7) 固然解釋了「做法正確」, (7.8) 又用 jackknife 來墊背 —— 說 [15] 和 [9] 都做錯了。但這一節因爲和前面連不上, 寫來是有一點弱的。

當然, 文章到了尾巴這裡, 文氣已盡。讀者到此也都累了, 一般的想法是: 前面的似乎都沒有錯, 最後一節大概也錯不了, 因此,「深究」大概是不會去做的了。

p.19, section 8. Remark 指的是文章中較有深度的註脚。此文有 6pp 的 remarks, 是長了些。但這是 Efron 的風格。

總的來說, 這文章其實只有一個主要的 idea: 「若我們想知道 $R(\mathbf{X}, F)$ 的分布, 用 bootstrap 就好」。除了幾個有解析解的地方外[16], 全文其實沒有多少技術面的東西。但這卻是極好的統計論文: 見林又見樹, 而見林的部分, 遠大於見樹, 因爲作者選了一個廣大的角度來看 bootstrap 的問題。數學計算太多的統計論文, 基本上都是令人卻步, 只有一小撮人可能有興趣的。

這文章沒有講甚麼? 想一想。有沒有模糊的地方需要弄清楚的? 哪些問題他該問卻沒有問?...

這是「找題目來做」的方法之一。當然, 這是老文章了, 容易補充的地方, 大概都被補齊了。去好好查一下 1979 後出現的關於 bootstrap 的文章, 我想, 只是讀它們的緒論, 你就該學會怎樣找題目了。

September 4, 2006

[16] 而且都不算難。

第八部分: 評審幾篇論文

本部分是個習題。附上四篇論文, 請你用 referee 的角度來讀它。你的態度是「儘量找麻煩」, 因爲你預備建議主編「這文章不能要」。因此你要挑出所有的錯來。

爲省空間, 我們拿掉所有的圖。

第一篇 (此文的摘要和緒論曾用作習題)

Testing for Activation in Data from FMRI Experiments

Abstract: The traditional method for processing functional magnetic resonance imaging (FMRI) data is based on a voxel-wise, general linear model. For experiments conducted using a block design, where periods of activation are interspersed with periods of rest, a haemodynamic response function (HRF) is convolved with the design function and, for each voxel, the convolution is regressed on prewhitened data. An initial analysis of the data often involves computing voxel-wise two-sample t-tests, which avoids a direct specification of the HRF. Assuming only the length of the haemodynamic delay is known, scans acquired in transition periods between activation and rest are omitted, and the two-sample t-test is used to compare mean levels during activation versus mean levels during rest. However, the validity of the two-sample t-test is based on the assumption that the data are Gaussian with equal variances. In this article, we consider the Wilcoxon rank test as well as modified versions of the classical t-test that correct for departures from these assumptions. The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

Key words: Excess kurtosis, haemodynamic response function, Shapiro-Wilk test, skewness, two-sample t-test, Welch test, Wilcoxon Rank test.

1. Introduction

Functional Magnetic Resonance Imaging (FMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to FMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully *and* the data follow the Gaussian distribution. In this article, we consider the analysis of FMRI data collected in one of two states, called "activation" and "rest," based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the activation data and the rest data are Gaussian with equal variances. In this article, we propose

use of a modified two-sample test for FMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937), which is a modification of two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for FMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in FMRI and the most common methods used to process FMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for FMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of FMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

2. FMRI Experiments

2.1 Some physiology

All neuronal activation is linked to an increase in oxygen consumption, causing a local increase in the blood flow. The body's response is to supply more oxygen than is required for the neuronal activity. Due to the different magnetic properties of oxygenated and de-oxygenated blood, the excess oxygenated blood that circulates during neuronal activation alters the magnetic properties of the venous blood, resulting in the so-called *blood oxygenation level dependent* (BOLD) signal. FMRI produces a sequence of brain images that is sensitive to changes in the BOLD signal.

In a classical FMRI experiment, the subject is scanned every few seconds to obtain an image of the brain; the subject is exposed to an experimental stimulus in some time periods, and is in a rest state during the remaining time periods. The stimulus can either be applied for brief periods in rapid, possibly random succession (an "event-related" experimental design, Josephs *et al.*, 1997), or for longer periods with interspersed rest periods (a "block" experimental design, Frackowiak *et al.*, 1997). In this paper, we focus on FMRI experiments conducted using a block experimental design.

Even though neuronal activation occurs immediately after exposure to the experimental stimulus, the vascular response evolves more slowly, resulting in the BOLD signal. The temporal relationship between neuronal activation and the observed BOLD signal is called the haemodynamic response. To model the haemodynamic response, it is common to convolve the experimental design with a so-called haemodynamic response function (HRF). Poisson, gamma, and Gaussian distributions are used widely as HRFs (Friston *et al.*, 1994).

The region of the brain where there is neural activation is found by regressing the

observed FMRI data on the expected BOLD signal, obtained as a convolution of the experimental design with the HRF. Of course, this depends on a well-specified HRF.

2.2 FMRI data

Observed FMRI data are four-dimensional, in space and time. At each time point, a three-dimensional image of the brain is acquired, called a volume. Each volume consists of voxels, and each voxel has an associated one-dimensional time series of observed signal intensities.

The most common approach to the analysis of FMRI data is to consider the voxels independently. A widely-used approach assumes a general linear model (GLM) for the voxel-wise time series (Friston *et al.*, 1995). For example, after various preprocessing steps, including prewhitening to achieve approximately independent errors, a two-sample test statistic is computed for each voxel where the two samples correspond to activation data and rest data. A voxel is declared to be significant if the test statistic exceeds some threshold. The distribution theory associated with this approach is based on the assumption of Gaussianity of the observed data and the proper specification of the HRF leading to the expected BOLD signal.

For initial data analysis, it is enough for us to know the length of the haemodynamic *delay* between neural activation and changes in the BOLD signal (Banddettini *et al.*, 1993). This knowledge is used to omit scans acquired in transition periods between possibly "activated" BOLD signals and "resting" BOLD signals. The delay between the neural activation and changes in the BOLD signal depends on many different factors; the type of stimuli, the duration of each stimulus, and the brain activation regions can all effect the length of the delay. Empirical studies have proposed methods for estimating HRFs that can adapt to different experimental designs. By using the block designs described in Section 2.1 and deleting transition data in our preliminary analysis, we have a sample of data acquired under activation and a second sample of data acquired under rest. In the next section, we describe four possible two-sample tests that might be used to test for the presence of activation at each voxel.

2.3 Two-sample tests

The null hypothesis of no difference between the means of two populations can be investigated with appropriate two-sample tests. In what follows, we summarize the four tests to be compared where, under activation the voxel data are $\mathbf{F}Y_a = \{Y_i\}_{i\in\mathbf{F}A}$ and, under rest the voxel data are $\mathbf{F}Y_r = \{Y_j\}_{j\in\mathbf{F}R}$; here $\mathbf{F}A$ and $\mathbf{F}R$ denote the activation and rest acquisition times, respectively.

The classical two-sample t-test

The classical two-sample t-test assumes:

(A1) Observations $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$ are uncorrelated.

(A2) The observations within each of $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$ have identical Gaussian distributions; that is,

$$\mathbf{F}Y_f \quad \sim \quad Gau(\mu_f * \mathbf{F}1,\ \sigma_f^2 * \mathbf{F}I); \qquad f \in \{a, r\}.$$

(A3) $\sigma_a^2 = \sigma_r^2$.

To test the hypothesis:

$$H_0 :\ \mu_a \leq \mu_r \ \text{ versus } \ H_1 :\ \mu_a > \mu_r, \tag{2.1}$$

the classical two-sample t-test uses test statistic,

$$T \equiv \frac{\overline{Y}_a - \overline{Y}_r}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_r}\right)\left(\frac{(n_a-1)s_a^2+(n_r-1)s_r^2}{n_a+n_r-2}\right)}}, \tag{2.2}$$

with

$$\overline{Y}_f = \frac{1}{n_f}\sum_{i\in\mathbf{F}} Y_i \quad \text{and} \quad s_f^2 = \frac{\sum_{i\in\mathbf{F}}\left(Y_i - \overline{Y}_f\right)^2}{n_f - 1}; \qquad f \in \{a, r\},$$

where $\mathbf{F}$ is the set of activation times $\mathbf{F}A$ (rest times $\mathbf{F}R$) if $f = a$ ($f = r$), and n_a (n_r) is the number of the observations in the sample $\mathbf{F}Y_a$ ($\mathbf{F}Y_r$).

If Assumptions (A1), (A2), and (A3) are satisfied, the classical two-sample t-test with significance level α is:

Accept H_0 if $T < t_d(1-\alpha)$
Accept H_1 otherwise,

where $t_d(1-\alpha)$ is the $100(1-\alpha)$ percentile of the t distribution on $d = n_a + n_r - 2$ degrees of freedom.

The Welch test

The Welch test (Welch, 1937) is used to test the same hypotheses (2.1), but it assumes only (A1) and (A2); that is, it is possible that $\sigma_a^2 \neq \sigma_r^2$. Welch (1937) has shown that under the null hypothesis H_0, the test statistic

$$T^* \equiv \frac{\overline{Y}_a - \overline{Y}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}} \tag{2.3}$$

has approximately a t distribution with

$$e \equiv \frac{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)}{\left(\frac{\sigma_a^4}{n_a^2(n_a-1)} + \frac{\sigma_r^4}{n_r^2(n_r-1)}\right)} \tag{2.4}$$

degrees of freedom. In practice, the population variances σ_a^2, σ_r^2 in (2.4) are estimated from data using sample variances s_a^2, s_r^2. The Welch test with significance level α is:

Accept H_0 if $T^* < t_e(1-\alpha)$

Accept H_1 otherwise,

where the cut-off value $t_e(1-\alpha)$ is based on fractional degrees of freedom and is obtained by interpolation of the $t_d(1-\alpha)$ cut-off levels based on the nearest integers d to e.

The CW test

The CW test (Cressie and Whitford, 1986) also tests hypotheses (2.1), but makes only Assumption (A1); that is, it is possible that the data are non-Gaussian with unequal variances. To account for this, we use the same statistic T^* given by (2.3) as Welch, but modify its null distribution according to the skewnesses α_{3a}, α_{3r} and the excess kurtoses α_{4a}, α_{4r} of the non-Gaussian activation and rest distributions, respectively.

By calculating the Cornish-Fisher expansion of T^*, Cressie and Whitford (1986) show that under Assumption (A1) and H_0, the distribution of T^* is approximately that of the random variable,

$$V = U + \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{6\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^{3/2}}(U^2-1) - \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{2\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^{3/2}}U^2 - \frac{1}{2}gUZ, \tag{2.5}$$

where U, Z are i.i.d. $N(0,1)$ and

$$g \equiv \left\{ \frac{\frac{\sigma_a^4}{n_a^3}(\alpha_{4a}+2) + \frac{\sigma_r^4}{n_r^3}(\alpha_{4r}+2)}{\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^2} - \frac{\left(\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}\right)^2}{\left(\frac{\sigma_a^2}{n_a}+\frac{\sigma_r^2}{n_r}\right)^3} \right\}^{1/2}. \tag{2.6}$$

The CW test with significance level α is

Accept H_0 if $T^* < v(1-\alpha)$

Accept H_1 otherwise,

where $v(1-\alpha)$ is the $100(1-\alpha)$ percentile of the distribution of V, obtained by simulation. As for the Welch test, the population moments in (2.5) and (2.6) are estimated from data using sample versions; see Section 3.3.

The Wilcoxon Rank (WR) Test

The WR test (Wilcoxon, 1945) makes only assumption (A1), as does the CW test. In addition, it assumes that the distribution function $F(y)$ of the observations $\mathbf{F}Y_r$ is continuous and the distribution function of the observations $\mathbf{F}Y_a$ is $F(y-\delta)$, for $\delta \in \mathbb{R}$. Then the WR statistic tests the hypotheses,

$$H_0: \ \delta \le 0 \ \text{ versus } \ H_1: \ \delta > 0. \tag{2.7}$$

In order to test (2.7), the WR test sums the ranks of each of the $\mathbf{F}Y_a$ values in the combined sample of $N = n_a + n_r$ data consisting of the $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$ values ordered from smallest to largest. Let R_i denote the rank of Y_i; $i \in \mathbf{F}A$. The test statistic for the WR test is

$$W = \sum_{i \in \mathbf{F}A} R_i.$$

An exact p-value is then computed based on the null distribution ($\delta = 0$) of W, which is obtained by considering all possible $N!$ permutations of ranks of the $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$.

However, this is computationally demanding for large n_a and n_r. For large n_a and n_r, we approximate the distribution of the centered and scaled version of W,

$$W^* = \frac{W - .5 - n_a(n_a + n_r + 1)/2}{\sqrt{n_a n_r (n_a + n_r + 1)/12}},$$

with a standard normal (Hollander and Wolfe, 1999). Hence the WR test with significance level α is:

$$\text{Accept } H_0 \text{ if } \; W^* < z(1-\alpha)$$
$$\text{Accept } H_1 \text{ otherwise,}$$

where $z(1-\alpha)$ is the $100(1-\alpha)$ percentile of the Gaussian distribution with zero mean and unit standard deviation.

3. Methods of Analysis and Comparisons

In this section, we continue to consider inference based on a single generic voxel. Simultaneous inference involving all voxels is considered in Section 4.

3.1 Application of Two-Sample Tests to FMRI Data

Let $\mathbf{T}$ be the set of acquisition times of the observed intensities associated with the given voxel. Assuming the subject was exposed to only one type of neural activation, $\mathbf{T}$ can be divided into three groups: the time points $\mathbf{F}A$ where activation of the BOLD signal is expected, the time points $\mathbf{F}R$ during which the BOLD signal is expected to be in a rest state, and the time points $\mathbf{B}$ corresponding to the transition periods between the activation and the rest times. An example of such a division of time points is illustrated in Figure **??**. In the two-sample tests considered in this article, one sample corresponds to $\mathbf{F}A$ and other sample corresponds to $\mathbf{F}R$; intensities corresponding to $\mathbf{B}$ are omitted from further analysis.

Figure 1 about here

Consider the two-sample tests of H_0 versus H_1 given in Section 2. For a given voxel and a given test, accepting the alternative hypothesis H_1 means that the associated voxel is declared to be activated by the experimental stimulus.

3.2 Simulated FMRI data

Six datasets were obtained from 3 healthy volunteers (1 female, 2 males) using a 1.5T Signa scanner. The data were collected under rest conditions; that is, the subjects were not exposed to any stimulus during the experiment and they were instructed to relax in the scanner with their eyes closed. One such rest dataset was obtained from the first male subject (30 years old), two rest datasets were obtained from the second male subject (27

years old), and three rest datasets were obtained from the female (30 years old). Each dataset consisted of 200 volumes, every observed volume contained 28 slices, and each slice had 64x64 voxels. These datasets were preprocessed for motion correction and prewhitened to make the time series uncorrelated (using the software FEAT, which is part of the FSL package; see Smith *et al.*, 2001).

We created activation datasets by essentially adding a signal having *known magnitude and location* of the activation to each preprocessed rest dataset. The signal component was calibrated against an image acquired from a previous unrelated visual-activation FMRI experiment; see Figure **??** for an example. By applying the signal in the locations acquired from a previous visual experiment, we avoided the possibility of applying the signal near so-called default regions (regions which show decreased neuronal activity during the activation of the stimulus) and their confounding effects on the simulated signal. The activation datasets alternated blocks of 10 time points of rest with 10 time points of activation. The average peak-signal change, defined as a ratio between the average of the intensities under the activation and the average of the intensities measured during the rest periods for the most activated voxel, was set to be 3%. Each dataset contains 200 time points; the three sets of time points $\mathbf{F}A$, $\mathbf{F}R$, and $\mathbf{B}$ were obtained assuming a haemodynamic delay of 3 time periods, resulting in $n_a = 70$ and $n_r = 73$.

Figure 2 about here

3.3 Violations of equal variances and Gaussianity assumptions

Several methods were used to assess the degree of departure of the activation datasets from (A2) and (A3). Consider a generic voxel and recall from Section 1 that $\mathbf{F}Y_a = \{Y_i\}_{i\in\mathbf{F}A}$ make up the so-called "activated" sample and $\mathbf{F}Y_r = \{Y_j\}_{j\in\mathbf{F}R}$ make up the "rest" sample.

To investigate the violation of Assumption (A3) given in Section 2, thereby allowing $\sigma_r^2 \neq \sigma_a^2$, we computed the sample variances for $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$ for each voxel in each activation dataset. The pairs of sample variances of active and rest samples for all voxels that are located in subject's brain (out of all $64 \times 64 \times 28 = 114,688$ voxels, only $2,2340$ of them were located in subject's brain) are plotted in Figure **??**; the 45-degree line corresponding to equal variances is superimposed. In all panels, and especially in 3(c), we see some points far from the diagonal, which suggests that the assumption of homogeneity is violated for three voxels. A formal F-test ($\alpha = 0.05$) of equal variances detected $1,225$ out of $22,340$ (5.5%) brain voxels to have significantly different sample variances, and visual inspection of these voxels indicated no spatial pattern. This indicates that, overall, unequal variances may not be a serious problem for these FMRI data.

Figure 3 about here

To investigate departures from Gaussianity, Assumption (A2), we computed the sample skewness and sample excess kurtosis for $\mathbf{F}Y_a$ and $\mathbf{F}Y_r$, for all six activation datasets. For

the activation sample these are:

$$\widehat{\alpha}_{3a} = \frac{\sqrt{n_a}\sum_{i\in\mathbf{F}A}(Y_i-\overline{Y}_a)^3}{\{\sum_{i\in\mathbf{F}A}(Y_i-\overline{Y}_a)^2\}^{3/2}},$$
$$\widehat{\alpha}_{4a} = \frac{n_a\sum_{i\in\mathbf{F}A}(Y_i-\overline{Y}_a)^4}{\{\sum_{i\in\mathbf{F}A}(Y_i-\overline{Y}_a)^2\}^{2}} - 3,$$

and likewise we computed $\widehat{\alpha}_{3r}$ and $\widehat{\alpha}_{4r}$ for the rest sample.

To illustrate graphically the relationship between skewness and kurtosis, we chose one activation dataset. The pairs $(\widehat{\alpha}_{3a}, \widehat{\alpha}_{4a})$ for the 22,340 brain voxels from one activation dataset are plotted on the left panel of Figure **??**, and the pairs $(\widehat{\alpha}_{3r}, \widehat{\alpha}_{4r})$ are plotted on the right panel. For Gaussian data, the plotted pairs should be very close to the origin. In Figure **??**, we observe strong departures from zero skewness and zero excess kurtosis in both panels. Thus, we might expect an improvement in hypotheses testing for activation using the CW test or the WR test over the classical two-sample t-test or the Welch test.

Figure 4 about here

More formally, we calculated the Shapiro-Wilk test (e.g., Royston, 1982) for normality ($\alpha = .05$) for each voxel and rest/activation combination. For the dataset used in Figure **??**, Table **??** summarizes the number (out of 22,430) of brain voxels that were significantly non-Gaussian. About 12% of activated samples and about 11% of rest samples were declared significant by the Shapiro-Wilk test; if the samples were Gaussian, we would expect only 5% to be declared significant. More than 20% of voxels were declared significant in at least one of the activated or rest samples.

Table 1: Brain-voxels declared significant using Shapiro-Wilk test ($\alpha = .05$), based on one of the six datasets.

		Activated samples		
		Significant	Not significant	Total
Rest samples	Significant	647	2095	2742 (12.3%)
	Not significant	1774	17824	19598
Total		2421 (10.8%)	19919	22340

The spatial distribution of the voxels declared significant is shown in Figure **??**; while they are distributed fairly homogeneously between regions of the brain, there is some indication that, within a region, they can clump together.

Figure 5 about here

4. Results

All four two-sample tests were used to test for activation in each voxel. We obtained p-values as in Section 2 where the p-value for the CW test was obtained from simulation of the random variable given by (2.5) and that for the WR test was obtained from the standard normal approximation to W^*.

Because of the multiple hypotheses being tested (one for each brain voxel), the voxels declared as active were obtained by comparing the p-values with $\alpha^* \equiv \alpha/\{\#$ of brain voxels$\}$ with $\alpha = .05$. This is the voxelwise Bonferroni-adjusted level of significance based on an overall level of significance of $\alpha = .05$. Voxels with p-values less than or equal to α^* were pronounced active. Because the activation pattern of each dataset was known, we can estimate and compare the sizes and powers of the two-sample tests.

Let A denote the set of voxels to which an activation signal has been added and R the set of voxels with no added activation. Let $\mathcal{A}_{\text{right}}$ denote the voxels in A declared to be active, and let $\mathcal{A}_{\text{wrong}}$ denote the voxels in A not declared active. All voxels from category R can be similarly divided into $\mathcal{R}_{\text{right}}$, those non-activated voxels not declared active, and $\mathcal{R}_{\text{wrong}}$, those non-activated voxels which were declared active.

The achieved size of each test was estimated by

$$\widehat{\alpha} \equiv (|\mathcal{R}_{\text{wrong}}| \,/\, |\mathcal{R}|),$$

where $|\mathcal{C}| \equiv \#$ voxels in the region $\mathcal{C}$ of the brain. The quantity $\widehat{\alpha}$ is also called the false-positive rate and should be comparable to the desired familywise level of significance α $(= .05)$. If $\widehat{\alpha} < \alpha$, the test is conservative. The power of each test was estimated by

$$\widehat{\pi} \equiv (|\mathcal{A}_{\text{right}}| \,/\, |\mathcal{A}|)\,,$$

which is the true-positive rate.

Table **??** lists the estimated sizes and powers of each test for all six simulated FMRI datasets. All four tests were consistently very conservative, with the Wilcoxon test being the most conservative. The classical t-test and Welch test had equivalent power, which was consistently greater than that of the Wilcoxon test. The CW test was the most powerful test, uniformly over the six datasets.

Table 2: Estimated size and power of the four two-sample tests for the six datatsets.

Dataset	TEST							
	Classical t-test		Welch		CW		WR	
	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$	$\hat{\alpha}$	$\widehat{\pi}$
1	.496E-4	.289	.496E-4	.288	.992E-4	.305	0	.277
2	0	.208	7.466E-4	.208	19.927E-4	.224	4.479E-4	.200
3	0	.221	0	217	0	.235	0	.202
4	.583E-4	.233	.583E-4	.233	1.750E-4	.253	.583E-4	.224
5	0	.205	0	.205	0	.223	0	.188
6	0	.239	0	.237	27.527E-4	.251	1.101E-4	.227

Table **??** gives a more detailed comparison of the classical t-test and the CW test for one of the datasets. While 626 out of 2,173 activated brain voxels were correctly detected as significant by both tests, 37 additional activation voxels were correctly detected by the CW test that were not identified by the classical t-test. Only one activation voxel was identified by the classical t-test that was missed by the CW test.

Table 3: Comparison of the performance of the CW test and the classical t-test, based on one of the six datasets

			CW test			
			Voxels from $\mathcal{A}$		Voxels from $\mathcal{R}$	
			$\mathcal{A}_{\text{right}}$	$\mathcal{A}_{\text{wrong}}$	$\mathcal{R}_{\text{right}}$	$\mathcal{R}_{\text{wrong}}$
Classical t-test	Voxels from $\mathcal{A}$	$\mathcal{A}_{\text{right}}$	626	1	·	·
		$\mathcal{A}_{\text{wrong}}$	37	1509	·	·
	Voxels from $\mathcal{R}$	$\mathcal{R}_{\text{right}}$	·	·	20165	1
		$\mathcal{R}_{\text{wrong}}$	·	·	0	1

5. Discussion and Conclusions

While the results were obtained from only one type of scanner, the 1.5T Signa GE, and with FMRI data for three subjects, they show that FMRI data can exhibit both unequal variances and non-Gaussianity. Using the Shapiro-Wilk test, more than 20% of voxels in the dataset were declared significant in one or both of the rest or activated samples. We believe that more powerful scanners will lead to data that are even more non-Gaussian, since their finer spatial resolution involves less averaging of the response.

The Welch test is valid for unequal variances but when non-Gaussianity is suspected, the CW test accounts for both. The WR test is a nonparametric analog of the classical t-test. In the six datasets studied in Section 3, non-Gaussianity was a bigger problem than unequal variances. The results in Section 4 showed that the CW test performed better than the other three tests. These results suggest that the CW test should replace any standard use of the classical parametric or nonparamentric two-sample tests based on FMRI data.

Acknowledgement

This research was supported by the Office of Naval Research under grants N00014-99-1-0214 and N00014-02-1-0052 and by the National Science Foundation Grant DMS-0406026. The authors would like to thank Antonio Algaze and Petra Schmalbrock for providing the FMRI data and the members of FMRIB, Oxford UK for initial consultation about simulating activation FMRI datasets. Perceptive comments by the referees led to improvements in the exposition and strengthening of our conclusions.

References

Bandettini, P. A., Jesmanowicz A., Wong, E. C. and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine* **30**, 161-173.

Cressie, N. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal* **28**, 131-148.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. and Mazziotta, J. C. (1997). *Human Brain Function*. Academic Press.

Friston, K. J.,. Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189-210.

Friston, K. J., Jezzard, P. and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* **2**, 69-78.

Hollander and Wolfe, (***** add initials ****) (1999). *Nonparametric Statistical Methods, 2nd edn*. John Wiley and Sons.

Josephs, O., Turner, R. and Friston, K. J. (1997). Event-related fMRI. *Human Brain Mapping* **5**, 243-248.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* **31**, 115-124.

Smith, S. M., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, J., Leibovici, D., Ripley, B., Woolrich, M. and Zhang, Y. (2001). FSL: New tools for functional and structural brain image analysis. *NeuroImage* **13**, S249.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

第二篇 (此文的摘要和緒論曾用作習題)

An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population

Abstract: Behavioral risk factors for cancer tend to cluster within individuals, which can compound risk beyond that associated with the individual risk factors alone. There has been increasing attention paid to the prevalence of multiple risk factors (MRF) for cancer, and to the importance of designing interventions that help individuals reduce their risks across multiple behaviors simultaneously. The purpose of this paper is to develop methodology to identify an optimal linear combination of multiple risk factors (score function) which would facilitate evaluation of cancer interventions.

Key words: Community based research, conditional logistic regression, multiple risk factors, random effects.

1. Introduction

Despite the considerable biomedical advances of the last half-century, facilitating improvement in lifestyle behaviors remains the most efficacious population-level strategy for reducing cancer risk. Estimates vary, but suggest that over fifty percent of new cancer cases and up to one-third of cancer mortality could be prevented though improvements in health behavior practices (American Cancer Society, 2004; Doll and Peto, 1981). A 19 percent decline in the rate at which new cancer cases occur, and a 29 percent decline in the rate of cancer deaths, could potentially be achieved by 2015, if prevention efforts were heightened and behavior change sustained. This would translate to the prevention of approximately 100,000 cancer cases and 60,000 cancer deaths each year, by the year 2015 (National Cancer Policy Board and Institute of Medicine, 2003).

There is ample epidemiological evidence for the consideration of red meat consumption, physical activity, and folic acid intake in cancer prevention efforts. Regular physical activity lowers the risk of cancers of the colon, breast, and possibly prostate (Colditz, Cannuscio, and Frazier, 1997; Friedenreich and Rohan, (1995).). An additional 30 percent of cancer deaths can be attributed to adult diet (Anonymous, 1996); higher intake of red meat has been associated with increased risk of colon (Sandhu, White and McPherson, 2001) and prostate cancers (Michaud, Augustsson, Rimm, Stampfer, Willett, and Giovannucci 2001). Associated with both physical inactivity and diet is obesity, which may account for between 25-30 percent of cancers of the colon, breast (postmenopausal), endometrium, kidney, and esophagus (Vainio and Bianchini, 2002). Folic acid is protective against colon cancer (Giovannucci, Stampfer, Colditz, Hunter, Fuchs, Rosner, Speizer, and Willett, 1998); long-term multi-vitamin use, in particular has been found to reduce risk for colon cancer, likely because of its folic acid content (Giovannucci *et al.* 1998).

The risk for many diseases, including colon cancer, is associated with multiple behavioral risk factors (MRF); these behaviors are highly interrelated and tend to cluster within individuals. For example, those who eat high-fat diets are also more likely to be sedentary, suggesting that the behaviors may be mutually reinforcing (see e.g., Emmons, Marcus, Linnan, Rossi, and Abrams, 1999). Change in one behavioral risk factor thus may serve as a stimulus or gateway for change in the other health behaviors (see e.g., Emmons *et al.* 1999), and there are overarching behavioral principles and intervention frameworks that guide behavior change efforts across risk factors.

Consequently, to facilitate population-level reductions in cancer risk, it may be inefficient to target discrete behavioral risk factors, when similar principles might be applied simultaneously to multiple behaviors (Institute of Medicine, 2000).

The literature provides little consensus as to the most appropriate analytic strategy for evaluating the efficacy of MRF interventions; most studies have analyzed the various outcomes independently or by creating a simplistic sum (e.g., 1 RF + 1 RF = 2RFs) (see e.g., Prochaska and Sallis 2004; Campbell, James, Hudson, Carr, Jackson, Oakes, Demissie, Farrell, and Tessaro, 2004). This could be problematic, because the use of separate analytic strategies may result in improper inferences regarding the effect of an MRF intervention because of correlation among the factors. Such strategies may overlook the clustering effect brought about by the agglomeration of multiple behavioral risk factors and have been criticized as being too simplistic. The purpose of this paper is to develop a methodology to identify an optimal linear combination of multiple behavioral risk factors (MRF score function) for cancer that would best facilitate evaluation of an MRF cancer intervention.

2. Methods

2.1 Study design

The data analyzed in this paper are from the Harvard Cancer Prevention Program Project (HCPPP) Healthy Directions, which is composed of two randomized controlled trials, one in health centers (HC) (Emmons, Stoddard, Gutheil, Suarez, Lobb, and Fletcher 2003), and another in small businesses (SB) (Hunt, Stoddard, Barbeau, Wallace, and Sorensen 2003). The overarching goal of the HCPPP was to create a new generation of cancer prevention interventions that would be effective among working class, multi-ethnic populations. Together, the two arms of the trial were successful in enrolling a sub-population of the multi-ethnic working class population in eastern Massachusetts. The study aims and sampling strategies are published in greater detail elsewhere (Emmons *et al.*, 2003; Hunt *et al.*, 2003).

2.2 Health centers

Healthy Directions-HC (Emmons *et al.*) was a randomized controlled trial conducted in collaboration with a large health care delivery system, comprised of 14 multi-specialty medical group practices that serve over 270,000 patients. Ten of the fourteen health centers were invited to participate in this study, and all agreed. Health center served as the unit of randomization and intervention. Briefly, patients who resided in low income, multi-ethnic neighborhoods (defined using census block-groups that were predominantly working class, impoverished, or with low levels of education) were identified and approached for participation through their health center. Individuals identified through geocoding to be residents in the target neighborhoods were deemed eligible if they met the following criteria: (1) being 18-75 years old, (2) having a well-care or follow-up visit scheduled with a participating provider, (3) being able to speak and read either English or Spanish, (4) not having cancer at the time of enrollment, (5) not being employed by the participating health centers, (6) not being employed by a worksite participating in the companion small business study, and (7) providing consent to participate in the randomized study. All providers practicing in the Internal Medicine Departments of the health centers were approached for permission to recruit from among their patient pools. Provider participation averaged 83% across sites (range 50%-100%; 97 clinicians). Patients scheduled for appointments with the participating providers and in the eligible age range were identified through the automated central appointment system. Study staff attempted to recruit 8,963 potentially eligible candidates; 2,547 (28%) individuals were

unreachable. Among the 6,414 potential subjects reached, 867 (14%) were ineligible, 3,330 (52%) refused, and 2,219 (35%) were enrolled. Assuming that 14% of those not reached were also ineligible, the response rate is 29% of those assumed eligible. The cohort recruited at baseline was contacted by telephone after the intervention period to complete a follow-up survey. Of the 2,219 who completed the baseline survey (n=1088 intervention condition; n=1131 control condition), 1,954 (88%) completed the follow-up survey. The follow-up response rate was equivalent across conditions.

2.3 Small business

The Healthy Directions-SB study (Hunt *et al.*, 2003) was a randomized controlled trial in which the worksite was the unit of randomization and intervention. Worksites were identified using the Dun and Bradstreet database to locate small businesses with Standard Industrial Classification (SIC) codes 20-39 (manufacturing industries) and employing between 50-150 employees. Additional inclusion criteria included: (1) employing a multi-ethnic population (defined as 25% of workers being first-or second-generation immigrants or people of color), (2) having a turnover rate of less than 20% in the previous year, (3) being autonomous in decision-making power to participate in a study, and (4) agreeing to be randomly assigned to the intervention condition. One hundred thirty-three (133) companies met the eligibility criteria, and of these, 26 agreed to participate (Barbeau, Wallace, Lederman, Lightman, Stoddard and Sorensen 2004).

Data were collected using interviewer-administered surveys among individuals who were permanent employees and worked 20 hours or more per week. On site interviews were administered on company time in the language (either English, Spanish, Portuguese, or Vietnamese) preferred by respondents. Two cross-sectional samples were collected, one at baseline in which 1,740 participants from 26 worksites completed the survey (response rate 84%). The second sample was collected at follow-up 1,408 participants in 24 worksites (during the course of the intervention two worksites dropped out, one intervention and one control) with a response rate of 77%. 974 participants (518 in control worksites and 456 in intervention worksites) completed both the baseline and follow-up surveys forming the embedded cohort used in this analysis.

2.4 Data and analysis

The goals of the intervention were to: (1) increase fruit and vegetable intake, (2) decrease red meat consumption, (3) increase physical activity levels, and (4) increase daily multivitamin usage. The following variables assess the individual risk factors measured on a continuous scale: number of servings of fruit and vegetables per day, number of servings of red meat consumed per week (RM), and hours of moderate or vigorous physical activity per week (PA). The fourth measure is a binary variable indicating use of a multi-vitamin on 6 or 7 days per week (MV). In order to keep all variables on an equivalent time scale, we created a new variable for fruit and vegetable consumption that calculated the amount of fruits and vegetables consumed in one week (FV) by multiplying the current measure of fruit and vegetable intake by seven. The continuous variables (FV, RM, PA) were standardized using the formula in Equation 2.1;

$$STV = \frac{V - P_{05}}{P_{95} - P_{05}}, \tag{2.1}$$

where V are the original values for the continuous variables (FV, RM, PA), P_{05} and P_{95} are the fifth and ninety-fifth percentile values respectively for a given variable and STV are the new standardized variables (STFV, STRM, and STPA respectively). Standardization was implemented for consistency (to make a one unit change in one variable similar to a one unit change in another) and interpretability. The 5^{th} and 95^{th} percentiles were used to minimize the influence of outliers.

For the purposes of identifying an optimal linear combination that would show an intervention effect we restricted our sample to only those subjects who received the intervention, responded to both the baseline and follow-up surveys, and have complete data for the four risk behaviors. As opposed to the usual situation of observing how the covariate vector or a linear combination of the covariate vector will change because of treatment, the idea here is to determine how the covariate vector or the linear combination will predict the intervention status. This is similar in spirit to a matched case-control analysis.

A popular method for the analysis of longitudinal data with a dichotomous outcome is a mixed effects logistic regression model. A mixed effects logistic regression model with a logit link will have the form:

$$\log\left[\frac{pr(Y_{ij}=1)}{1-pr(Y_{ij}=1)}\right] = a_i + \beta' X_{ij}. \tag{2.2}$$

Here Y_{ij}, $i = 1, ..., n$, $j = 1, 2$, denotes the indicator of intervention time (i.e. pre-intervention $Y_{i1} = 0$ and post-intervention $Y_{i2} = 1$), X_{ij} is the covariate vector, and a_i is a random cluster effect. The subscript i is an indicator for individual and the subscript j is an indicator for time. Each individual subject i is a cluster of two sets of observations, pre-intervention and post-intervention. The random effect variable a_i can be thought of as measuring an individual's demographic characteristics (i.e., age, gender, race). In our analysis, we want to control for an individual's specific demographic characteristics, therefore, we treat the random effect variable a_i as a nuisance parameter and condition it out of the model. We can condition them out by using the conditional likelihood based on the fact that $Y_{i1} + Y_{i2} = 1$. We are left with a conditional logistic regression model. These types of models are often used to analyze matched case-control studies, where the outcome of interest is whether a subject is a case or control.

In this framework we intend to model

$$\text{logit}(Pr(Y_{ij}=1|a_i)) = \beta' X_{ij}, \tag{2.3}$$

where an optimal linear combination, or the best score, will be $\hat{\beta}' X$.

We set up our data as if it came from a 1:1 matched case-control study; each individual is a cluster of two observations, one "case" and one "control". One observation is pre-intervention ("control"/baseline) and the second observation is post-intervention ("case"/follow-up). At each time point (pre and post-intervention) each subject has a vector (containing STFV, STRM, STPA, and MV) of covariates.

For matched case-control studies with one case per matched set, the likelihood function for the conditional logistic regression reduces to the partial likelihood of the Cox model for the continuous time scale (Hosmer and Lemeshow 1998). We created dummy survival times so that all cases have the same event time and the corresponding controls are censored at a later time. We used Proc PHREG in SAS[1] to fit the conditional logistic regression model by forming a stratum for each matched set (individual id number). This allowed us to obtain estimates for $\hat{\beta}$.

3. Results

Using the combined Health Center and Small Business data from the Healthy Directions baseline and follow-up surveys on the 1,209 study participants that received the intervention, we found an optimal score function for the four risk factors:

$$score = 1.05 * STFV + 1.70 * MV + 0.25 * STPA - 1.35 * STRM. \tag{3.1}$$

The score is a summary measure of the health behaviors of a subject based on these four factors. From this score, we can see that increasing the number of fruits and vegetables consumed per week,

taking a multivitamin six or more days a week, increasing the amount of physical activity done in a week, and/or decreasing the amount of red meat consumed in a week will increase the score for a subject which in turn means an overall improvement in health behaviors. The dynamics of the score are consistent with the goals of the intervention. A participant can increase their health behavior score by changing one risk factor, or combinations of the four risk factors in a manner consistent with the goals of the intervention.

We believe that these factors not only have individual effects, but that some factors may also have compounding effects. This belief is based on previous evidence of the interrelationships seen in modifying behavioral risk factors (see e.g., Emmons *et al.*, 2004; Butterfield *et al.*, 2004). Therefore, we looked for significant interactions between the four variables. Table **??** shows the analysis of maximum likelihood estimates for our final model. In our final score function (see Equation 3.2), we multiply the effects (parameter estimates) by 100 to increase the range of the scores as well as to simplify interpretation.

Table 1: Analysis of Maximum Likelihood Estimates

Standardized Variable	Parameter Estimate	Standard Error	P-Value
STFV	0.576	0.303	0.0570
MV	2.008	0.2078	<.0001
STPA	0.232	0.193	0.2294
STRM	-1.515	0.343	<.0001
STFV*STRM	1.229	0.565	0.0296
MV*STRM	-0.707	0.343	0.0392

$$\begin{aligned} \text{score} &= 57.6 * STFV + 200.8 * MV + 23.2 * STPA - 151.5 * STRM \\ &\quad + 122.9 * STFV * STRM - 70.7 * MV * STRM \end{aligned} \tag{3.2}$$

There was a significant interaction between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, sugessting that changing both behaviors simultaneously is better than changing either behavior alone, but the effect of changing both behaviors is not equal to the sum of the individual changes on the MRF score. There was also a significant interaction between multivitamin usage more than six times a week and the amount of red meat consumed per week, suggesting that changing either behavior alone is good, but changing both behaviors simultaneously will result in an even larger increase on the MRF score.

Table 2: Examples of changes in individual risk factor measures and resulting MRF score

	FV	MV	PA	RM	MRF Score
Baseline values	20	0	4	5	−30.23
Case 1: Optimal values at final					
Final values	35	1	10	1	247.46
change	+15	+1	+6	−4	+277.69
Case 2: Improves only FV					
Final values	35	0	4	5	17.39
change	+15	0	0	0	+47.62
Case 3: Improves only MV					
Final values	20	1	4	5	135.22
change	0	+1	0	0	+165.45
Case 4: Improves only PA					
Final values	20	0	10	5	−19.09
change	0	0	+6	0	+11.14
Case 5: Improves Only RM					
Final values	20	0	4	1	14.64
change	0	0	0	−4	+44.87
Case 6: Improves FV and RM					
Final values	35	0	4	1	72.82
change	+15	0	0	−4	+103.05

Table **??** displays a few examples of how a change in an individual risk factor from the baseline case to the optimal case will change the score. If we consider the first row of Table **??** to be a baseline value in which a subject consumes 20 servings of fruits and vegetables per week, does not take a multivitamin six or more days a week, has four hours of physical activity per week, and consumes five servings of red meat per week (the average values for study subjects at baseline, meeting only the recommend level of physical activity), the standardized values would be 0.32,0,0.32, and 0.5 respectively. Therefore the score for a subject at baseline would be

$$score = 57.6*0.32+200.8*0+23.3*0.32-151.5*0.5+122.9*0.32*0.5-70.7*0*0.5 = -30.2. \quad (3.3)$$

We can consider an arbitrary optimal case as a subject who consumes 35 servings of fruits and vegetables per week (or five a day), takes a multivitamin 6 or more days a week, engages in 10 hours of physical activity per week, and eats one serving of red meat per week (meeting and/or exceeding all of the recommended levels). Table **??** shows the effects of these changes on the score from the baseline case to the optimal case for each variable alone and the effects of combinations of two and three variables. Figure **??** compares our final model (MRF, Equation 3.2) with a main effects model (a model without interactions) showing that the main effects model can both overestimate and underestimate scores predicted from the MRF model due to the absence of the two significant interactions.

Table 3: Score changes with one, two, and three variable changes

Variables Changed	Score Change
FV	47.62
MV	165.45
PA	11.14
RM	44.87
FV + MV	213.07
FV + PA	58.76
FV + RM	72.82
MV + PA	176.59
MV + RM	238.60
PA + RM	56.00
FV + MV + PA	224.21
FV + MV + RM	266.55
FV + PA + RM	83.96
MV + PA + RM	249.73

Although we used only those subjects that received the intervention to develop the score, the score is generalizable to the entire study population. It was created, and is most useful for, the purpose of comparing the subjects that received the intervention to those that received usual care, because it provides a summary measure of the health behaviors of a subject on all intervention risk factors pre and post-intervention. There were 1,297 subjects that received usual care and took both the baseline and follow-up surveys. These subjects can be considered controls for the effect of the intervention. Figure **??** shows box plots of score comparing baseline and follow-up for subjects that received the intervention compared to those that received usual care. In the intervention group, the mean score at baseline was 48.1, while the mean score at follow up was 104.3. In the usual care group the mean score at baseline was 40.4, and the mean score at follow-up was 53.2. The mean change in score for the usual care group was 12.8, while the mean change in score for the intervention group was 56.2. There was a statistically significant difference in the mean change in score from baseline to follow-up when comparing the usual care group to the intervention group ($p<0.001$). The intervention group showed greater improvements in score at follow-up proving the intervention quite successful.

4. Discussion

Increasing attention has been paid to multiple risk factor interventions, across a range of disease outcomes, both because adverse behavioral risk factors tend to cluster within individuals and because of recognition of the utility of facilitating change across multiple risk behaviors. However, most MRF studies to date have used individual risk factor methods to analyze intervention effects (see e.g., Prochaska and Sallis (2004); Campbell *et al.*, 2004). As shown in Figure **??**, the main effects model both over-estimates (e.g., FV & PA & RM) and under-estimates (e.g., MV & PA & RM) the scores predicted from the MRF model, depending on the combination of variables and the degree of change for a given participant in the intervention. Thus, such analytic models may compromise determinations of the efficacy of a MRF intervention. We were successful in modeling a linear combination of behavioral risk factors including interactions between risk factors, an effort that represents an advance over the existing methods for analyzing MRF intervention efficacy.

To illustrate, note that in our final model there are two interaction terms. One between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, and another between multivitamin usage more than six times a week and the amount of red

meat consumed per week. Looking at Table **??**, we can see that with all the other variables held constant, a change in fruit and vegetable consumption alone from 20 to 35 servings per week will increase the score by 47.62, and a decrease in red meat alone from 5 to 1 servings per week will increase the score by 44.87. However, because of the interaction term, if both variables are changed by the amounts indicated the score would increase by 72.82, which because of the interaction is a smaller than 92.49, the sum of the individual changes. Similarly, if a subject begins to take a multivitamin daily the score will increase by 165.45, and if they decrease red meat from 5 to 1 serving per week the score will increase by 44.87. However, if a participant begins to take a multivitamin daily and decreases red meat consumption by 4 servings per week the score will increase by 238.60, a larger increase than 210.32 that you would get by adding 165.45 from taking a multivitamin daily and 44.87 by decreasing red meat consumption. Cluster effects are not captured by main effects models and are an advantage of this method.

There are some limitations to the method proposed here, namely that the score function depends on the efficacy of the intervention to determine variable weighting. For example, if the intervention was most effective at increasing multivitamin use, the weight (coefficient) for the multivitamin use variable would be largest in magnitude, whereas if the intervention was least effective in changing the participants' physical activity patterns, the weight (coefficient) for the physical activity variable would be the smallest in magnitude. In some cases then, the weights may be a proxy for the amount of participant effort necessary to change the health behavior. For example, in this study we saw that multivitamin usage had the largest weight and thus the most influence on the score.

There are at least two potential explanations for this finding. First, the promotion of multivitamin usage may require less participant burden when compared to the other health behaviors (e.g., physical activity). Thus, it may be easier for participants to modify their multivitamin use; this supposition appears to be supported by the finding of an almost thirty percent increase from baseline to follow-up of the number of subjects taking a multivitamin daily. However, it is important not to undermine the significance of a change in multivitamin usage which is strongly related to the prevention of disease outcomes. Sustained use of multivitamins containing folic acid have been associated with the reduction in risks for numerous conditions including colorectal cancer and cardiovascular disease (Ggiovannucci *et al.*, (2002); Fairfield and Fletcher, 2002). Physical activity on the other hand, is among the most challenging health behaviors addressed in the study to intervene upon. In this population, 66 percent of the subjects were getting the recommended level of physical activity at baseline, and 69 percent at follow-up. Of those subjects that were not at or above the target level of physical activity at baseline, almost 9.5 percent were at or above the target level at follow-up. Another factor to consider is that multivitamin usage was treated as a binary variable in our models. That is, many potential changes are captured in the categorization of either taking a multivitamin 6 or more times a week or not doing so. Relative to increasing one serving of fruits and vegetables a week, decreasing one serving of red meat in a week, or increasing an hour of physical activity a week, this is a substantial change.

Although the purpose of our method was to develop a health behavior score (composite variable), there are some limitations to using this type of variable. The purpose of such a variable is to allow for easy comparisons of the four factors with one number. When there are changes in the score, however, a composite variable does not provide any insight into which individual risk factor(s) have contributed to the change.

Another potential limitation of applying this method to the HCPPP data is the merging of the two cohorts, small businesses and health centers. Our method develops a score function that is independent of the population but not independent of the intervention. By combining the two data sets, we have made the assumption that the interventions given to these two populations are the same. In reality, although the two interventions were quite similar, they were not identical. We decided, however, to combine the two cohorts in order to increase power, and to create a universal score that could be applied to both cohorts. This not only allowed us to make comparisons within

a cohort, but between cohorts. Taking these limitations into account, our methodology remains preferable compared to existing techniques that do not accord weights to the risk factors or adjust for cluster effects.

In summary, we have developed a score that effectively integrates multiple behavioral cancer risk factors into one measure, irrespective of individual demographic factors. We believe that the methods are generalizable to other working class multi-ethnic populations, and future research should be done to evaluate the effectiveness of these methods in other groups. The primary strength of the methodology used to develop the score is that it can be easily implemented to develop scores for other populations, for other combinations of behavioral risk factors, or for other disease outcomes (e.g., cardiovascular disease). Given the increasing attention being paid to the development of MRF interventions, we believe the described method to be the preferred means of analysis in comparison to previously used strategies. Ultimately, we believe that analytic focus on examining clusters of behavioral risk factors will enhance the design of multiple risk factor intervention approaches.

Figure 1 about here

Figure 2 about here

Acknowledgements

The research of Melody Goodman was supported by National Institute of Child Health and Human Development grant 5 F31 HD043695. The research of Dr. Li was supported by National Institute of Health grant R01CA95747. The research of Dr. Bennett, Dr. Stoddard, and Dr. Emmons was supported by grant 5 P01 CA75308 from the National Institutes of Health, and support to the Dana-Farber Cancer Institute by Liberty Mutual, National Grid, and the Patterson Fellowship Fund.

References

American Cancer Society (2004). Cancer Facts and Figures. Technical Report, American Cancer Society.

Anonymous (1996). Harvard Report on Cancer prevention Volume: 1 Causes of Human Cancer, *Cancer Causes and Control* **7**, S3-S9.

Barbeau, E. M., Wallace, L., Lederman, R., Lightman, N., Stoddard, A. and Sorensen, G. (2004). Recruiting small manufacturing worksites that employ multi-ethnic, low-wage workers to a cancer prevention research trial. *Preventing Chronic Disease* **1** 1-9.

Butterfield, R. M., Park, E. R., Puleo, E., Mertens, A., Gritz, E. R., Li, F. P., and Emmons, K. (2004). Multiple risk behaviors among smokers in the childhood cancer survivors cohort. *Psychooncology* **13**, 619-629.

Campbell, M. K., James, A., Hudson, M. A., Carr, C., Jackson, E., Oakes, V., Demissie, S., Farrell, D. and Tessaro, I. (2004). Improving multiple behaviors for colorectal cancer prevention among african american church members, *Health Psychol* **23**, 492-502.

Colditz, G. A., Cannuscio, C., and Frazier, A. (1997). Physical activity and reduced risk of colon cancer: Implications for prevention. *Cancer Causes and Control* **8**, 649-667.

Doll, R. and R. Peto (1981). The causes of cancer : quantitative estimates of avoidable risks of cancer in the United States today, *J. Natl. Cancer Inst.* **66**, 1191-1308.

Emmons, K. M., Marcus, B. H., Linnan, L. A., Rossi, J. S. and Abrams, D. B. (1999) Mechanisms in multiple risk factor interventions: Smoking, physical activity, and dietary fat intake among manufacturing workers. *Preventive Medicine* **23**, 481-489.

Emmons, K. M., Stoddard, A. M., Gutheil, C., Suarez, E. C., Lobb, R. and Fletcher, R. (2003). Cancer prevention for working class, multi-ethnic populations through health centers: The healthy directions study. *Cancer Causes and Control* **14**, 727-737.

Fairfield, K. M and Fletcher, R. H. (2002). Vitamins for chronic disease prevention in adults: scientific review. *Journal of the American Medical Association* **287**, 3116-3126.

Friedenreich, C. M., and Rohan, T. E. (1995). Physical activity and risk of breast cancer. *European Journal of Cancer Prevention* **4**, 145-151.

Giovannucci, E., Stampfer, M. J., Colditz, G., Hunter, D., Fuchs ,C., Rosner, B., Speizer, F. and Willett, W. (1998). Multivitamin use, folate, and colon cancer in women in the Nurse's Health Study. *Annals of Internal Medicine* **129**, 517-524.

Hosmer, D. W., and Lemeshow, S. (1998). *Encyclopeida of Biostatistics*, 2327-2333. John Wiley.

Hunt, M. K., Stoddard, A., Barbeau, E. M., Wallace, L. and Sorensen, G. (2003). Cancer prevention for working class, multiethnic populations through small businesses: The Healthy Directions Study. *Cancer Causes & Control* **14**, 749-760.

Institute of Medicine (2000). *Promoting Health: Intervention Strategies from Social and Behavioral Research.* National Academy Press.

Michaud, D. S., Augustsson, K., Rimm, E. B., Stampfer, M. J., Willett, W. C. and Giovannucci, E. (2001). A prospective study on intake of animal products and risk of prostate cancer. *Cancer Causes & Control* **12**, 557-567.

National Cancer Policy Board and Institute of Medicine (2003). Fulfilling the potential of cancer prevention and early detection, Technical Report, Washington D. C.

Prochaska, J. J. and Sallis, J.F. (2004). A randomized controlled trial of single versus multiple health behavior change: promoting physical activity and nutrition among adolescents. *Health Psychology* **23**, 314-318.

Sandhu, M. S. and White, I. R. and McPherson, K. (2001). Systematic review of the prospective cohort studies on meat consumption and colorectal cancer risk: A meta-analytical approach. *Cancer Epidemiology, Biomarkers & Prevention* **10**, 439-446.

Vainio, H. and Bianchini, F. (2002). *IARC Handbooks of Cancer Prevention. Volume 6: Weight Control and Physical Activity.* IARC Press.

第三篇 (此文的摘要和緒論曾用作習題)

Reducing Subjectivity in the Likelihood

Abstract: Some scientists prefer to exercise substantial judgment in formulating a likelihood function for their data. Others prefer to try to get the data to tell them which likelihood is most appropriate. We suggest here that one way to reduce the judgment component of the likelihood function is to adopt a mixture of potential likelihoods and let the data determine the weights on each likelihood. We distinguish several different types of subjectivity in the likelihood function and show with examples how these subjective elements may be given more equitable treatment.

Key words: Mixture likelihood, model averaging, subjectivity.

1. Introduction

We propose methods for modeling the likelihood function that will require fewer subjective judgments. We first discuss the nature of the problem of subjectivity in the likelihood function; then we review some related research; and finally, we define a mixture likelihood function and suggest estimation procedures that reduce the effects of subjective views imposed on the observed data.

1.1 Statement of the problem

It is sometimes desirable that beliefs of experimenters should be brought into a scientific analysis in ways that minimally distort the measured data (see, for example, Hogarth, 1980; Kyberg and Smokler, 1980; Lad, 1996). But that having been said, scientists observing data sometimes interpret the data points subjectively, according to what they want the data to show, and according to how precisely they believe the data points were measured. The latter procedure is of course quite common. This subjective interpretation of observed data may be totally at the unconscious level, or it may be purposeful (with the purposeful interpretation, the analysis may become fraudulent; see for example, Grayson, 1995, 1997; Howson and Urbach, 1990; and Press and Tanur, 2001).

The subjective interpretation of empirical data in medicine was discussed by Kaptchuk (2003). He stated (page 1, *op. cit.*):

> Doctors are being encouraged to improve their critical appraisal skills to make better use of medical research. But when using these skills, it is important to remember that interpretation of data is inevitably subjective and can itself result in bias. Facts do not accumulate on the blank slates of researchers' minds, and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists' convictions or theoretical apparatus.

Statistical analysis of a data set most often proceeds by summarizing the distribution of the data in terms of its likelihood function. In order to specify the form of the likelihood function, various assumptions are made about the data, such as mutual independence, identical distributions, unimodality, etc. After the likelihood function has been specified, additional assumptions are sometimes made (significance levels thought to be appropriate are specified, a prior distribution about the underlying unobservable quantities may be brought in, etc.). Analysis of the data generally proceeds by trying to keep the likelihood function treatment of the data as simple as possible, so that the scientist or analyst will introduce minimal distortion of the data. The analyst tries not to discard data, and tries to maximize the chance of understanding what nature is trying to tell us through the revealed data about the underlying phenomenon. In this way, when the analysis of the data has been completed, the claim can reasonably be made that the conclusions drawn from the analysis approximate, if not precisely reflect, the laws of nature, rather than the possible misinterpretations and misunderstandings of the laws of nature by human beings. It will be useful to first briefly define what we mean by objectivity and subjectivity, in this context.

According to Mandik (2001)[1],

> The word *objectivity* refers to the view that the truth of a thing is independent from the observing subject. The notion of objectivity entails that certain things exist independently from the mind, or that they are at least in an external sphere. Objective truths are independent of human wishes and beliefs. The notion of objectivity is especially relevant to the status of our various ideas, and the question is to what extent objectivity is possible for thought, and to what extent it is necessary.

This is but one of many definitions that have been suggested. The elusive quest for objectivity in science has been, and remains, an important topic of discussion among historians and philosophers of science (for extensive additional discussions of the meaning of "objectivity", see for example, Bower, 1998; Porter, 1995, 1996; and Daston and Galison 1992). For some, scientific objectivity involves the search for *certainty* in knowledge about one of nature's well-kept secrets, independent of what human beings believe; but in many cases, we find that what we earlier thought to be true about nature, turns out later to be questionable.

In an interesting example from physics, Folger, 2003, pointed out that:

> Pioneer 10, launched in 1972, is now some 8 billion miles from home. But is has been slowing down, as if the gravitational pull on it from the sun is growing progressively stronger the farther away it gets. Milgrom proposed (see the MOND pages-MOdified Newtonian Dynamics) [2] that Newton's laws might change at these accelerations. If Milgrom is right, Newton's and Einstein's laws will be in for some major tweaking.

Sometimes the scientist has such deep understanding and insight into the phenomenon he/she is studying that the scientist's own predictions of what should be found from the analysis are far superior to what the data analysis seems to indicate. In some cases the beliefs of the scientist or analyst are so strong, even before actually taking any data that bear on the phenomenon, that the data are interpreted or manipulated so that they will reflect these preconceived views of the scientist. Any preconceived personal views (views held before taking any data), weak or strong, are what we refer to in this context as *subjectivity.*

[1]Mandik, P. (2001). *The Internet Encyclopedia of Philosophy*, http://www.utm.edu/research/iep/o/objectiv.htm.

[2]MOND pages — http://www.astro.umd.edu/ ssm/mond/

1.2 Related Research

One approach to reducing the effects of differing assumptions about likelihoods may be found in a line of research that involves use of the *empirical likelihood function*. In this approach, most useful in large samples, a discretized, binned, version of the empirical cdf, instead of a specific likelihood function, is used. Inference is then made from a multinomial distribution. An unfortunate feature of this approach is the additional unknown parameters that are concomitantly introduced into the model. See: Owen, 1988, 2001. For typically small and moderate size samples this could be a problem, but for the massive data sets typical of data mining applications (see, for example: Berry and Linoff (1997); and Hastie, Tibshirani, and Friedman, 2001) such an approach could be a helpful alternative.

We show in the next section how we might understand and account for some types of subjectivity that sometimes enters the *likelihood function*, and might not be desired. We will use the definition and form of the likelihood function in which for absolutely continuous random variables, up to a proportionality constant, it is the joint probability density function of the observables given the unobservables.

2. Types of Subjectivity in the Likelihood Function

We distinguish three of the types of likelihood subjectivity problems that may occur:

(a) how to determine the distributional form of the likelihood function in a way that is largely objective, but permits the data themselves to guide the modeling as to whether the data are Normally-distributed, or Gamma-distributed, or possibly follow some other convenient distribution. We call this problem, "*distributional subjectivity*";

(b) how to treat observed data that have possibly been weighted subjectively so that some data points are valued more heavily than others, and some are even ignored; we call this problem, "*weighted-data subjectivity*";

(c) how to account for the nature of the experiment used to obtain the data that may have favored one type of response over another; we call this problem, "*experiment subjectivity*".

We treat each of these types of subjectivity in Section 3.

3. Reducing Likelihood Subjectivity

3.1 The mixture likelihood

We use a convex mixture of various likelihoods for the data; the usual likelihood function results as a special case.

Suppose an experiment is repeated n times with the resulting one-dimensional data outcomes: $x_1, x_2, \ldots, x_n$. We suppose that there are J models for the data that potentially we might reasonably entertain. For simplicity, merely to suggest a general type of approach, we consider problems

involving only one unknown parameter, namely, the means of the J distributions, ω.

In some situations, the parameters may be quite different from one another but they can generally be related functionally. For example, the case of distinguishing between the means of normal and log-normal distributions, where the mean parameter has different meanings in the two cases is sometimes particularly interesting. In such cases, functional relationships among the parameters are required.

Suppose, in the one-parameter problem, we can assume these data to be mutually independent and identically distributed, and we agree to adopt the likelihood function for Model m_j:

$$\ell_j(x_1, \ldots, x_n \,|\, m_j, \omega) \equiv \ell_j(\underline{x} \,|\, m_j, \omega).$$

Define a "mixture likelihood function",

$$L_M(x_1, \ldots, x_n \,|\, \omega) \equiv L_M(\underline{x} \,|\, \omega),$$

such that:

$$L_M(\underline{x} \,|\, \omega) = E\{ \text{ likelihood } \} = E_{Model}[\ell(\underline{x}) \,|\, \omega]] = \sum_{j=1}^{J} \ell_j(\underline{x} \,|\, \omega) P(m_j \,|\, \omega) \tag{3.1}$$

where $\ell_j(\underline{x}|m_j, \omega)$ denotes the usual likelihood function of the data under model m_j, $\ell(\underline{x}|\omega)$ denotes a model-independent likelihood function, and $P(m_j|\omega)$ denotes the prior probability of model m_j. The mixture likelihood function is of course a likelihood function itself. If there were only one model ($J = 1$), L_M reduces to the ordinary likelihood. The mixture likelihood function explicitly assumes that we should combine different models in a linear way. Other possibilities exist of course, and perhaps in certain cases, they are even more desirable. But because for a wide variety of cases, the linear assumption seems appropriate, we will retain this assumption throughout. We next address the issue of how to reduce model subjectivity (how to choose the weights).

3.2 Reducing "model subjectivity"

In some instances, the scientist has very strong, theory-based, beliefs about how the data were generated, and how the corresponding likelihood function should behave. In such instances, especially in small samples, the analyst should surely use that information to permit the desired likelihood function to emerge. In other situations where the scientist/analyst wants the data to speak as loudly as possible relative to the scientist's pre-conceived beliefs, there is no unique way to accomplish this objective. The approach suggested here is to take equal weights in the mixture. Accordingly, take all $P(m_j|\omega)$ in eqn. (3.1) to be equal (discrete uniform distribution). This interpretation of equal treatment for the different models is:

(1) in keeping with the approach frequently used for weighting in mixture models to express indifference or ignorance among the various components in the mixture;

(2) it is the procedure suggested by Laplace when he adopted his Principle of Insufficient Reason (Laplace, 1812, 1814);

(3) it is consistent with a basic result of information theory that the distribution that corresponds to maximum entropy, or minimum information, is the uniform distribution.

This gives the mixture likelihood function (equally-weighted average likelihood):

$$L_M(\underline{x}\,|\,\omega) = \frac{1}{J}\sum_{j=1}^{J} \ell_j(\underline{x}\,|\,m_j,\omega). \tag{3.2}$$

For example, suppose there are just two potential models ($J = 2$) that might reasonably represent the data: $N(\omega, 1)$ and a Student t-distribution centered at ω, with 3 degrees of freedom (a fat-tail distribution that has a population mean). Then, the mixture likelihood function becomes:

$$L_M(\underline{x}\,|\,\omega) = \frac{1}{2}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x_i-\omega)^2\} + \prod_{i=1}^{n}\frac{m^{m/2}/B(1/2,m/2)}{[m+(x_i-\omega)^2]^{(m+1)/2}}\right\}. \tag{3.3}$$

Clearly each term in equation (3.3) is non-negative and integrates to one (with respect to $\underline{x}$), so $L_M\{\underline{x}|\omega\}$ is a bone fide likelihood function for the data (as would be the case whichever models we choose). In some situations, one scientist might favor the normal distribution for representing the distribution of the data, while another might favor the Student t-distribution. By using $L_M\{\underline{x}|\omega\}$ to represent the likelihood function for all inferences, the analyst reduces the model subjectivity in the description of the data distribution. Maximum likelihood estimation of ω is now more complicated numerically than it would be with use of either the normal or the Student t distributions separately, but the numerical problem is straightforward (see numerical example below) and easily generalizes to more than two possible ordinary likelihoods.

We next numerically illustrate the example suggested in this section of how to reduce model subjectivity when the models under consideration are the $N(\omega, 1)$ and the Student t_3 centered at ω. We randomly generated a total of 20 observations, 10 observations from t_3, a Student t-distribution with 3 degrees of freedom centered at $x = 10$, and 10 observations from $N(10, 1)$. The resulting data are shown in columns 2 and 3 of Table 1a. Then, using the Newton-Raphson method, we calculated the mixed MLE. It is given at the bottom of Column 2 as: $\hat{\omega} = 9.9168$. To illustrate variability, there are four replications of this entire process shown in Table 1a; the four resulting mixed maximum likelihood estimates (mixed MLE's) are also shown in Table 1a.

Table 1a: Four replications of model subjectivity

	t	normal	t	normal	t	normal	t	normal
1	11.1861	9.6734	8.1266	11.182	8.9931	8.331	11.476	9.8325
2	9.9749	11.542	11.092	10.175	10.641	10.131	10.922	11.051
3	8.5632	10.259	11.374	11.720	9.8717	7.8108	10.270	10.642
4	5.5471	9.4442	9.3422	10.757	8.8824	8.3177	8.3906	9.0293
5	9.5188	10.779	13.189	9.8871	9.7579	9.4354	11.899	8.8359
6	10.994	9.3448	9.6007	9.715	10.385	10.092	9.5234	10.566
7	9.1875	9.9779	9.9069	9.7106	13.659	9.4326	14.559	9.495
8	11.283	9.2274	8.3785	9.8394	9.5543	10.361	10.177	10.247
9	8.7574	10.724	10.073	11.637	9.2285	9.0399	10.155	9.0938
10	7.9804	11.263	10.379	9.4837	11.274	9.7291	9.2795	10.366
Mixed MLE	9.9168		10.240		9.6237		10.116	

For comparison purposes, we also computed the separate ordinary MLE's assuming all 20 observations were generated from a normal, and then, that all 20 observations were generated from a t_3 distribution. Results are given in Table 1b.

Table 1b: Separate MLE's For normal and Student data

Normal MLE	9.761	10.278	9.7463	10.291
t_3-MLE	9.924	10.187	9.6104	10.111

Thus, it may be seen that in the first instance, while the mixed MLE is 9.9168, the MLE assuming all 20 observations came from a normal is 9.7614, whereas the MLE assuming all 20 observations came from a t_3 is 9.9240. Results for the other 3 cases are shown in Tables 1a and 1b as well. Depending upon the assumptions made for the modeling, results for the mixture MLE obtained from the model averaging may differ substantially from those of the separate models, or not.

3.3 Reducing "weighted-data subjectivity"

We examine two distinct cases of weighted data subjectivity and model the two cases separately below.

Case 1 — Several Observers (Scientists) Rate the Same Data Points Differently

In this case, different observers (scientists) might interpret the same points differently. Some observers might view certain points as mistakes (outliers that were generated from different distributions from the other points), and therefore delete them from the analysis; and others might, according to their own beliefs, weight certain points more heavily than others (perhaps difficult-to-measure points might be weighted less heavily because the error associated with the measurement might be greater than with most of the other points; perhaps certain points obtained were measured under censored conditions; etc.).

For simplicity, assume the data points are mutually independent. We define the likelihood function for Observer O_k as:

$$\ell(\underline{x}\,|\,\omega) = \prod_{j=1}^{n} [f(\delta_{jk}x_j\,|\,O_k,\omega]^{p_k(\delta_{jk}x_j|O_k)}, \tag{3.4}$$

where: $p_k(\delta_{jk}x_j|O_k) = 1$, if Observer O_k includes the data point x_j in the analysis, and $p_k(\delta_{jk}x_j|O_k) = 0$ if not; δ_{jk} denotes the weight that Observer k places on observation x_j, $f(x_j|\omega)$ denotes the pdf (probability density function) of X_j, conditional on ω. The mixure likelihood function may be defined as:

$$\begin{aligned} L_M(\underline{x}\,|\,\omega) &= E\{\text{likelihood}\} = E_{data}\{\ell_k(\underline{x}\,|\,\omega)\} \\ &= \sum_{k=1}^{K} \ell_k(\underline{x}\,|\,\omega)P_k(O_k), \end{aligned} \tag{3.5}$$

where $P_k(O_k)$ denotes the prior probability that the data analyst places on the model that has been developed by Observer k. To be objective (or indifferent among the choices), in the sense we have been discussing, we take $P_k(O_k) = 1/K$, for all k. Then,

$$L_M(\underline{x}\,|\,\omega) = \frac{1}{K}\ell_k(\underline{x}\,|\,\omega) \tag{3.6}$$

As a simple example, suppose that all K observers adopt the same distribution for the data, say, $N(\omega, 1)$ (in Section 3.2 the analyst adopted two different possible distributions for the data),

and assume that they weight the points in the same way, so that $\delta_{jk} = 1$ for all k, for all points they include in their analyses, but they may include different points. Then, since the n observations are independent,

$$\ell_k(\underline{x}\,|\,\omega) = \prod_{j=1}^{n}\left[\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(x_j-\omega)^2\}\right]^{p_k(x_j|O_k)}. \tag{3.7}$$

To be specific, suppose that $n = 102$, and that there are two observers, O_1 and O_2. Suppose further that O_1 believes x_{102} is an outlier, and O_2 believes that both x_{101} and x_{102} are outliers, but they agree that the first 100 points $(x_1, \ldots, x_{100})$ should be included in their analyses. Then,

$$\begin{aligned} p_1(x_j \| O_1) &= 1 \quad \text{for } j = 1, 2, \ldots, 101, \\ &= 0, \quad \text{for } j = 102, \end{aligned}$$

also,

$$\begin{aligned} p_1(x_j \| O_2) &= 1 \quad \text{for } j = 1, 2, \ldots, 100, \\ &= 0, \quad \text{for } j = 101, 102, \end{aligned}$$

Then,

$$\ell_1(\underline{x}\,|\,\omega) \equiv \ell_1 = \left(\frac{1}{\sqrt{2\pi}}\right)^{101}\exp\{-\frac{1}{2}\sum_{j=1}^{101}(x_j-\omega)^2\}, \tag{3.8}$$

and

$$\ell_2(\underline{x}\,|\,\omega) \equiv \ell_2 = \left(\frac{1}{\sqrt{2\pi}}\right)^{100}\exp\{-\frac{1}{2}\sum_{j=1}^{100}(x_j-\omega)^2\}, \tag{3.9}$$

Then,

$$\begin{aligned} L_M(\underline{x}\,|\,\omega) &= \frac{1}{2}\left\{\left(\frac{1}{\sqrt{2\pi}}\right)^{101}\exp\{-\frac{1}{2}\sum_{j=1}^{101}(x_j-\omega)^2\}\right. \\ &\qquad \left. + \left(\frac{1}{\sqrt{2\pi}}\right)^{100}\exp\{-\frac{1}{2}\sum_{j=1}^{100}(x_j-\omega)^2\}\right\} \end{aligned} \tag{3.10}$$

We may now estimate ω by maximizing $L_M(\underline{x}|\omega)$ with respect to ω. Note first that if we let n_1, n_2 be the numbers of data points used in the respective analyses of Observers O_1 and O_2, they are also the numbers of terms in the two summations, and in this example, $n_1 = 101$ and $n_2 = 100$. We may readily find by ordinary differentiation, the mixture maximum likelihood estimator (mixture MLE) to be:

$$\omega = \alpha(\omega)\bar{x}_1 + [1-\alpha(\omega)]\bar{x}_2, \quad 0 \le \alpha(\omega) \le 1, \tag{3.11}$$

where:

$$\alpha(\omega) \equiv \frac{n_1\ell_1(\omega)}{n_1\ell_1(\omega) + n_2\ell_2(\omega)} \tag{3.12}$$

$$\bar{x}_1 \equiv \frac{1}{101}\sum_{j=1}^{101}x_j, \quad \bar{x}_2 \equiv \frac{1}{100}\sum_{j=1}^{100}x_j. \tag{3.13}$$

That is, we find the interesting result that $(\hat{\omega}|\ell,\ell_2)$ is a weighted average (actually a convex combination) of the separate MLE's that the two observers might adopt separately, and the weights

are their respective proportions of their ordinary likelihoods, an intuitively sensible result. But note that because $\alpha(\omega)$ depends upon ω, equations.(3.11) and (3.12) must be jointly solved numerically for $\hat{\omega}$.

While in large samples, the (continuous) data will generally ultimately swamp any prior distribution weights placed on the data points (see Le Cam, 1956), in small or moderate size samples, certain very influential points that may have been deleted from an analysis can have substantial effects on the interpretation of the experiment outcomes.

We next illustrate this example numerically. We randomly generated 18 points from $N(0,1)$. We then ordered the points, and added 2 larger outliers. We assumed the first observer dropped the largest point as an outlier, and the second observer dropped the two largest points as outliers. We then calculated the mixture MLE numerically from equation. (3.11) using the Newton-Raphson method. We replicated the procedure four times to examine variation. Data are shown in Table 2a.

Table 2a: Four replications of weighted-data subjectivity

Observation	$N(10,1)$	$N(10,1)$	$N(10,1)$	$N(10,1)$
1	8.3959	7.6748	8.1260	7.7977
2	8.4063	7.8796	8.5249	8.8122
3	8.559	7.9954	9.6225	8.9922
4	8.7975	8.7684	9.6490	9.0079
5	9.3082	8.9002	9.7041	9.0501
6	9.6001	8.9819	9.7444	9.1783
7	9.8433	9.2957	9.7660	9.2580
8	9.9802	9.3553	10.0400	9.3645
9	10.2570	9.3687	10.1180	9.4404
10	10.5710	9.5069	10.3150	9.7344
11	10.6690	9.6790	10.4280	9.8685
12	10.6900	9.8179	10.5690	10.0880
13	10.7120	9.8868	10.5780	10.2120
14	10.7140	10.0860	10.6230	10.2380
15	10.8160	10.3790	10.6770	10.3900
16	10.8580	10.4620	10.7310	10.4440
17	11.1910	10.5510	10.7990	10.5690
18	11.2540	10.9440	10.8960	10.7810
19	12.0000	12.0000	12.0000	12.0000
20	13.0000	13.0000	13.0000	13.0000

Calculations of MLE's for the data in Table 2a are given in Table 2b:

Table 2b: MLE's for data with outliers

Mixture MLE	10.0406	9.4205	10.0568	9.6267
$\bar{x}_{18}$	10.0346	9.4185	10.0506	9.6237
$\bar{x}_{19}$	10.1380	9.5544	10.1532	9.7487

We see that for the data in column 2 of Table 2b, for example, the mixture MLE was 10.0406. Had the observers carried out separate MLE's, with Observer 1 dropping only the last observation, he would have found his MLE to be 10.1380, while Observer 2 who dropped both of the last 2 observations would have found her MLE to be 10.0346. While the differences are not large they are intended to be illustrative.

Case 2 — One Observer (Scientist) Rates Each Data Point Differently

The second case of weighted-data subjectivity involves a single scientist weighting the importance of the data points differently from one another. Here we envision a single scientist who has carried out an experiment many times, but sometimes, for one reason or another, the scientist carried out the experiment with extremely small error, whereas on some other occasions, the scientist associated the experimental outcomes with considerably more error. Thus, which observed results had small associated error, and which had large associated error might differ from one replication of the experiment to the next.

In this context there is just one scientist who rates his/her experimental data differentially, according to how "well" the data point was measured, or what he/she thought should have occurred, or whatever. This is the more typical situation, compared with the first case. The mixture likelihood function is obtained from equations (3.4) and (3.5), for $K = 1$, as:

$$L_M(\underline{x}\,|\,\omega) = \ell_1(\underline{x}\,|\,\omega)P_1\{O_1\} = \prod_{j=1}^{n}[f(\delta_{j1}x_j\,|\,O_1,\omega)]^{p_1(\delta_{j1}|O_1)} \quad . \tag{3.14}$$

To follow the paradigm suggested here we should take $\delta_{j1} = 1$ for every j. Of course the individual scientist would often argue that he/she knows better than anyone else that certain points were really not as good as others, and should therefore be down-weighted.

A now-classical example of this type of subjectivity of special historical interest has been documented with real data. It involves the data collected by R. A. Millikan (1868-1953). Dr. Millikan was an American physicist who successfully measured the charge on a single electron, winning a Nobel Prize in 1923 for this famous oil-drop experiment (as well as other prizes). Holton (1978) scrutinized Millikan's laboratory notebooks and found that Millikan had repeated his oil-drop experiment 39 times, obtaining outcomes: $x_1, \ldots, x_{39}$ for the charge on the electron. Holton reported that Millikan had given each of his original sets of observations a personal quality-of-measurement rating: "best", "very good", "good", "fair", and no rating at all for discarded measurements (we interpret his weights to represent his prior probabilities for these measurements). The distribution of his rating results is summarized in the Table 3.

Table 3: Millikan's measurements

rating descriptions	effective raating	δ_{j1} = Weight	number of measurements
best	4	4/10	2
very good	3	3/10	7
good	2	2/10	10
fair	1	1/10	13
discard	no rating	——	7

For Millikan, $p_1(\cdot) = 1$, for 32 data points and $p_1(\cdot) = 0$ for the discarded 7 points. We order the measurements according to their effective ratings, from "best" to "fair", and form the weighted average. The estimated value of the charge on the electron is then given by the weighted average:

$$\hat{e} = \frac{4}{10}\sum_{j=1}^{2} x_j + \frac{3}{10}\sum_{j=3}^{9} x_j + \frac{2}{10}\sum_{10}^{19} + \frac{1}{10}\sum_{j=20}^{32} x_j.$$

Millikan formed the weighted average of his measurements and accordingly estimated the charge on the electron as 4.85×10^{-10} esu (electrostatic units). The ordinary equally weighted average would have been 4.70×10^{-10} esu. In his reported value he also averaged in the values obtained by other researchers. By contrast, the accepted value for "e", the charge on the electron, today, is 4.77×10^{-10} esu. But the impressive closeness of Millikan's values with today's accepted value is deceptive; it occurred only because his values were based upon, "a faulty value for the viscosity of air, which when corrected, increases the discrepancy with the modern value by over 40%" (Mathews, 1998).

3.4 Reducing "experiment subjectivity"

Suppose there are two experiments that might be performed: E_g ("g" for "good"), and $E_{\bar{g}}$ ("$\bar{g}$" for "not good"). In E_g the scientist knows that the experiment will contain one or more variables that might produce effects that will be confounded with the effect of fundamental interest. In E_g, there are likely to be fewer such confounding variables, so the scientist believes that he/she is more likely to be able to distinguish the effect he/she is seeking. Concomitantly, it may be that by carrying out E_g, the scientist is missing the important variables that suggest that the effect sought is really artifactual, and the seeming effect is explainable in other ways. Because the scientist is so convinced that the effect sought is real and not artifactual, he/she reasons that E_g is a "cleaner" and more promising experiment. The scientist might even argue, in a moment of enthusiastic zeal, that E_g is cheaper and/or less subject to error.

In both experiments, for simplicity of interpretation, we assume the data are normally distributed with variance equal to 1. Suppose that the scientist referred to above, call him/her Scientist A, would like to show that the population mean for the underlying phenomenon of interest is positive. If Scientist A carries out E_g, it is more likely that the sample mean $\bar{x}$ will be positive than if Scientist A carries out $E_{\bar{g}}$ wherein the sample mean $\bar{y}$ will imply the alternative hypothesis $H_{\bar{g}}$: that the population mean is not positive. If $E_{\bar{g}}$ is performed the scientist believes results are either unlikely to be supportive of the theory, or they are likely to be sufficiently marginal so that the theory will be in doubt. A priori, the experimenter adjudges the chances for concluding H_g: the population mean is positive, when performingE_g as greater than the chances for concluding that the population mean is positive when performing $E_{\bar{g}}$. Consequently, Scientist A decides to perform E_g.

Suppose some other scientist, say Scientist B, performs $E_{\bar{g}}$, and subsequently observes $\bar{y}$ (using the same sample size, n). Let θ denote an indexing parameter such that $\theta = 1$ if the hypothesis H_g is true, and $\theta = 0$ if the hypothesis H_g is false.

$$L_M\{\text{data} \mid \theta\} = E\{\text{lokelihood}\} = E_{experiment}[\ell(\text{data} \mid \theta)]$$
$$= \ell(\bar{x} \mid E_g, \theta)P\{E_g\} + \ell(\bar{y} \mid E_{\bar{g}}, \theta)P\{E_{\bar{g}}\}$$

The mixture likelihood function becomes:

$$L_M\{\text{data} \mid \theta\} = P\{E_g\}\frac{\sqrt{n}}{\sqrt{2\pi}}\exp\{-\frac{n}{2}(\bar{x}-\theta)^2\} + P\{E_{\bar{g}}\}\frac{\sqrt{n}}{\sqrt{2\pi}}\exp\{-\frac{n}{2}(\bar{y}-\theta)^2\}.$$

An investigator cognizant of both experiments has both and available. In the same spirit of a desire for equity of treatment in the likelihood function, the investigator takes $P\{E_g\} = P\{E_{\bar{g}}\} = 0.5$. Then,

$$L_M\{\bar{x}, \bar{y} \mid \theta\} = \frac{1}{2}\left[\frac{\sqrt{n}}{\sqrt{2\pi}}\exp\{-\frac{n}{2}(\bar{x}-\theta)^2\} + \frac{\sqrt{n}}{\sqrt{2\pi}}\exp\{-\frac{n}{2}(\bar{y}-\theta)^2\}\right].$$

Define $z = (\bar{x} + \bar{y})/2$. Then, combining terms shows that:

$$L_M\{z \mid \theta\} = \frac{1}{2}\frac{\sqrt{n}}{\sqrt{2\pi}}\exp\{-n/4\}\exp\{-n(z-\theta)^2\}.$$

Thus, the MLE for θ is clearly: $\hat{\theta} = z = (\bar{x} + \bar{y})/2$. If Scientist A were correct in his/her a priori assessments of what was likely to happen in the experiment, $\hat{\theta}$ is likely to be closer to zero than $\bar{x}$ (or even negative), a result that would tend to vitiate Scientist A's conclusions.

For example, for Scientist A's experiment, E_g, we generated 100 observations from $N(1, 1)$ and found $\bar{x} = 1.0598$. Then, for Scientist B's experiment, $E_{\bar{g}}$, we generated 100 observations from $N(-1, 1)$ and found $\bar{y} = -.9531$. So the generalized MLE, $\hat{\theta}$, is 0.053, a sample value just barely positive, which might not be convincing in many contexts for asserting that the population mean is really positive.

4. Conclusions

We have been concerned with how to reduce the effects of a scientist's pre-conceived beliefs in the analysis of his/her supposedly objectively-observed data. We have found that we can reduce the effect of some of those subjective interpretations by using a mixture likelihood function, and then choosing the mixture weights that weigh the various interpretations of the data equally.

References

Berry, M. J., and Linoff, G. (1997). *Data Mining Techniques.* John Wiley.

Bower, B. (1998). Objective visions: Historians track the rise and times of scientific objectivity. *Science News* **154**, 360-362.

Daston, L. J. and Galison, P. (1992). The image of objectivity. *Representations* **40**, 81-128.

Folger, T. (2003). Nailing down gravity: New ideas about the most mysterious power in the universe. *Discover Magazine*, Oct., 2003, 34-41.

Grayson, L. (1995). *Scientific Deception*. The British Library.

Grayson, L. (1997). *Scientific Deception – An Update.*: The British Library.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

Hogarth, R. (1980). *Judgment and Choice*. John Wiley.

Holton, G. (1978). Sub-electrons, presuppositions, and the Millikan-Ehrenhaft dispute. In *The Scientific Imagination: Case Studies* (Edited by Gerald Holton), 25-83. The Cambridge University Press.

Howson, C. and Urbach, P. (1990). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing Co.

Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal* **326**, 1453-1455.

Kyberg, H. E. Jr., and Smokler, H. E., Editors (1980). *Studies in Subjective Probability*. Robert E. Krieger Publishing Co.

Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. John Wiley.

Laplace, P. S. (1812). *Theorie Analytique des Probabilities*.[3] Paris: Courcier.

Laplace, P. S. (1814). *Essai Philosophique sur les Probabilities*. [4] Paris.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symposium on Math. Statist. And Prob.* **1**, 128-156. University of California Press. p. 308. ****** Please check where the p.308 come from? ***

Mathews, Robert A. J. (1998). Facts versus factions: The use and abuse of subjectivity in scientific research. Cambridge, England: The European Science and Environment Forum, Working Paper 2/98, September, 1998.[5]

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall.

Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.

Porter, T. (1996). Statistics, social science, and the culture of objectivity. *Oesterreichische Zeitschrift fur Geschichtswissenschaften* **7**, 177-191.

Press, S. J., and Tanur, J. M. (2001). *The Subjectivity of Scientists and the Bayesian Approach*. John Wiley and Sons.

[3]The second, third, and fourth editions appeared in 1814, 1818, and 1820, respectively. It is reprinted in 0euvres Completes de Laplace, Vol. VII, 1847, Paris: Gauthier Villars.

[4]This book went through five editions (the fifth was in 1825) revised by Laplace. The sixth edition appeared in English translation by Dover Publications, New York, in 1951. While this philosophical essay appeared separately in 1814, it also appeared as a preface to his earlier work, Theorie Analytique des Probabilities.

[5]See http://www.esef.org., p.5.

第四篇 (此文的摘要和緒論曾用作習題)

Application of One Sided t-tests and a Generalized Experiment Wise Error Rate to High-Density Oligonucleotide Microarray Experiments: An Example Using Arabidopsis

Abstract: **Motivation**: A formidable challenge in the analysis of microarray data is the identification of those genes that exhibit differential expression. The objectives of this research were to examine the utility of simple ANOVA, one sided t tests, natural log transformation, and a generalized experiment wise error rate methodology for analysis of such experiments. As a test case, we analyzed a Affymetrix GeneChip microarray experiment designed to test for the effect of a CHD3 chromatin remodeling factor, PICKLE, and an inhibitor of the plant hormone gibberellin (GA), on the expression of 8256 *Arabidopsis thaliana genes.*
Results: The GFWER(k) is defined as the probability of rejecting k or more true null hypothesis at a given p level. Computing probabilities by GFWER(k) was shown to be simple to apply and, depending on the value of k, can greatly increase power. A k value as small as 2 or 3 was concluded to be adequate for large or small experiments respectively. A one sided t-test along with GFWER(2)=.05 identified 43 genes as exhibiting PICKLE-dependent expression. Expression of all 43 genes was re-examined by qRT-PCR, of which 36 (83.7%) were confirmed to exhibit PICKLE-dependent expression.

Key words: ********** Please add keywords ***

1. Introduction

The advent of inexpensive microarray technology has enabled individual laboratories to easily obtain a global perspective on the expression pattern of thousands of genes. This powerful technology has allowed investigators to diagnose early cancers (Kim, J. W. and Wang, X. W., 2003; Zhang *et al.*, 2003), discover genes that contribute to quantitative traits (Gu *et al.*, 2002), and detect coordinated gene regulation during pivotal developmental events such as embryogenesis and sexual maturation (Girke *et al.*, 2000; Lo *et al.*, 2003; Ruuska *et al.*, 2002).

The first generation microarrays were generally based on two dye methodologies. These cDNA microarray experiments involve hybridizing two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000-20,000 cDNA probes. In contrast, more recent high-density oligonucleotide microarrays, such as those offered by Affymetrix , provide direct information about the expression levels in an mRNA sample and can have a much higher density (Yang and Speed, 2002).

The majority of methodologies for microarray analysis have been developed for two dye spotted arrays (Kerr *et al.*, 2000; Kerr and Churchill, 2001; Lee *et al.*, 2003, Nguyan *et al.*, 2004, for review see Quackenbush, 2001 and Yang and Speed, 2002). Unfortunately these two-dye spotted arrays also pose other statistical issues, such as normalization to correct for dye bias. Furthermore if more than 2 treatments are used, it is not possible to compare all treatments on the same chip thus necessitating an Incomplete Block Design (IBD) type design (Kerr and Churchill, 2001). As such, special experimental designs, such as the reference and rotational design are needed for correct analysis (Kerr and Churchill, 2001; Quackenbush, 2001 and Yang and Speed, 2002).

In contrast, oligonucleotide microarrays use a single dye technology and pose some advantages, including a greatly increased density of genes and simplified experimental design because treatment effects are tested independently on each chip, eliminating the need for IBD designs. Nevertheless, statistical issues remain, such as normality of residuals, homogeneity of residual variance, correlation of errors within an array, and correlation of biological samples across arrays.

Mixed model methods for analysis of microarray experiment, proposed by Wolfinger *et al.* (2001), solves most of these issues (see Craig *et al.*, 2003 for review). However, the complexity of analysis dramatically increases with these advanced methods. Unfortunately, many of the current practitioners of microarray technology do not possess the mathematical expertise necessary to meaningfully employ these methods. On the other hand ANOVA is a tool that is easy to implement with methods common to most researchers. Kerr and Churchill (2001) conclude that "The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors".

The first objective of this research was to examine the utility of simple ANOVA for analysis of replicated oligonucleotide microarrays experiments. The motivation was given eloquently by Kerr and Churchill (2001) who stated "An advantage of model based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering *why* it is inadequate can help the analyst identify sources of variation and bias." A secondary objective of this study was to show how using a one sided t-test can be used to increase power. The final objective was to introduce an alternative method to increase power by accepting a base number of false positive with high probability.

The ANOVA is particularly suited to analyzing data from microarray experiments that employ a replicated factorial arrangement of treatments. An example of such an experimental design is one in which the investigator looks at gene expression in wild-type and mutant plants in the presence or absence of an added chemical. Many microarray studies incorporate this type of experimental design, e.g. the response of genes in nontumorigenic and tumorigenic tissues to different concentrations of toxic or therapeutic drugs (Lundquist *et al.*, 2002; Martinez *et al.*, 2002) or the response of genes from different tissues to estrogen or other hormones (Abe *et al.*, 2003; Faccioli *et al.*, 2002; Fujita *et al.*, 2003; Goda *et al.*, 2002). This design easily extends into any number of genotypes (or tissues) by any number of developmental time points (or biochemical exposures).

The primary biological objective of this research was to understand how a CHD3-chromatin remodeling factor, PICKLE, and a plant growth regulator, gibberellin (GA), regulate gene expression during germination of Arabidopsis seeds (Rider *et al.*, 2003). PICKLE is necessary for repression of embryonic traits in Arabidopsis (Ogas *et al.*, 1997). Expression of the embryonic state in pickle seedlings is inhibited by the plant growth regulator gibberellin (GA) and is enhanced by application of uniconazole-P, an inhibitor of GA biosynthesis (Izumi *et al.*, 1985; Ogas *et al.*, 1997). Specifically, gene expression was examined in wild-type and *pickle* seeds grown in the absence and presence of 10^{-8} M uniconazole-P. Thus the genotypes were 'wild type' vs. the *pickle* mutant, and biochemical exposure was to either 10^{-8} M uniconazole-P or no uniconazole-P during seed germination.

Our working hypothesis was that PICKLE functions during germination to repress genes that promote embryonic identity. In support of such a hypothesis, the transcript levels of two positive regulators of embryogenesis, LEAFY COTYLEDON1 (LEC1) and LEAFY COTYLEDON2 (LEC2) (Lotan *et al.*, 1998; Stone *et al.*, 2001), are elevated during germination of pickle seedlings (Ogas *et al.*, 1999; Rider *et al.*, 2003). Our interest was to find new genes that exhibited PICKLE-dependent expression, i.e. were up regulated. As such, we had a natural one sided test.

2. 21 Biological Methods

Seeds and tissues from the Arabidopsis *pickle-1* mutant (in a Columbia ecotype background) and wild-type Columbia were used for all investigations. Plants were grown as described previously (Ogas *et al.*, 1997; Rider *et al.*, 2003).

The Affymetrix GeneChip Arabidopsis Genome Array[6] contained 8256 sets of oligos representing approximately 30% of the *Arabidopsis thaliana* transcriptome. A 2×2 factorial arrangement of treatments were examined. The first treatment was genotype (*pickle* mutant vs. wild type), the second treatment was uniconazole (applied vs. control), the treatment combinations were designated pkl, Upkl, wt and Uwt (*pickle* mutant untreated, *pickle* mutant treated with uniconazole-P, wild type untreated, and wild type treated with uniconazole-P) were each represented by four biological replicates ($n = 4$) for a total of 16 chips (Rider *et al.*, 2003).

3. StatisticalL Methods

3.1 The ANOVA, partitions, and transformations

The model for the completely randomized design (CRD) associated with the k-th spot (or gene) is $Y_{ij}^k = \mu + \tau_i^k + \epsilon_{(i)j}^k$ where Y_{ij}^k is the expression (or log transform) for the k-th gene, in the j-th replicate of the i-th treatment; μ^k is the overall mean; τ_i^k is the effect of the i-th treatment on that gene, and $\epsilon_{(i)j}^k$ is random residual. For maximum information treatment effects are further partitioned into main effects and interactions. The partitioning should be reduced to single degree of freedom tests by use of orthogonal contrasts. Because the ANOVA must be completed for each spot on the array, methods to automate the test are needed. To accomplish this goal, we use the well-known result that any single degree of F tests can equivalently be constructed as a t-test (Gill, 1978). A simple t-test for any contrast can be computed with the means procedure in SAS or in any standard spreadsheet, such as Excel. The t-test also offers the advantage of being able to test for a one sided alternative. In some experiments, as in this one, the researchers may only be interested in genes that are either up or down regulated, as a result, the power to detect those genes will be greatly increased.

For expression type data, the variance is usually correlated with the mean, violating a critical assumption for the ANOVA. For such data, transforming to logs will usually correct this problem. Interpretation of log transformed data also better meets the interest of the biologist as significant differences are interpreted as being significant ratios on a non-transformed basis, i.e. the difference between logs of numbers is the same as the log of a ratio. A log base 2 is interpreted as fold change, while base 10 is interpreted as orders of magnitude difference. Natural logs have not been widely used for array data but perhaps represents the most valid biological interpretation due to kinetics. A common rate equation in chemistry is where the rate of change in product (∂Y) per unit of time (∂t) is proportional (c) to the product (Y), thus $\partial Y = cY\partial t$. The solution to this differential equation is $Y = ce^t$. Therefore by taking natural logs, the expression is linearized into a rate equation, $\ln(Y) = \ln c + t$. If t is constant across biological replications, then variation in expression is due to linear differences in the rate constant c, the gene regulatory factor. Differences due to treatments are then interpreted as linear differences in gene regulatory factors (rate constants).

The vast majority of array data will require such a transformation, however, curiously these data better met the assumption when non-transformed. To check this assumption for any data, compute the within gene variance for each gene (the residual error variance in the ANOVA), then plot that against the average expression level for that gene. Any slope significantly different from

[6]part no. 510429, Affymetrix, Santa Clara, CA

zero (a zero slope is parallel to the x axis) indicates that the data require a transformation before the analysis proceeds.

For a given gene, because each treatment combination was randomized onto each of 4 biological replicates, the experiment as detailed above is a 2×2 factorial arrangement of treatments in a completely randomized design (CRD). The ANOVA for this design with treatment effects partitioned is given in Table 1.

Table 1: ANOVA table with partitions.

Source of Variation	Degrees of Freedom	Mean Square
Treatments	$t-1$	MS(T)
Genotypes (C_1)	1	MS(C_1)
Inhibitor (C_2)	1	MS(C_2)
Interaction of Genotype x Inhibitor (C_3)	1	MS(C_3)
Within Error	$t(r-1)$	MS(E)

The mean squares for the partitions can be found using the following formula along with the contrast coefficients given in Table 2

Table 2: Coefficients for partitions of treatment effects.

Treatment	Treatment Combination		Contrast Coefficients (C_{mj})		
	Genotype	Inhibitor	C_{1j}	C_{2j}	C_{3j}
1- pkl	*pickle*	None	1	1	1
2- Upkl	*pickle*	Uniconazole	1	-1	-1
3- wt	wild type	None	-1	1	-1
4- Uwt	wild type	Uniconazole	-1	-1	1

$$MS(C_m) = r(\sum_j C_{mj}\bar{Y}_{ij})^2 I(\sum_j) C_{mj}^2). \tag{3.1}$$

The F test, which is distributed as F with 1 and $t(r-1)$ degrees of freedom, is then computed as the ratio of $F = MS(C_m)/MS(E)$. This test is equivalently computed as t

$$T_m = \frac{\sum_j C_{mj}\bar{Y}_{ij\cdot}}{\sqrt{\frac{1}{r}MS(E)\sum_j C_{mj}^2}} \tag{3.2}$$

which has $t(r-1)$ degrees of freedom. From these formula it is easy to verify that the calculated value $F = t^2$ and from tables one can verify corresponding critical values, i.e. $F_{1,t(r-1)} = (t_{t(r-1)})^2$.

However, when calculated as a t-test the sign of the contrast is preserved, thus allowing a one tailed test. This approach will extend to any contrast for any number of treatments, provided the sum of the coefficients for that contrast is zero. To be orthogonal with other contrasts the sum of the cross products must also sum to zero.

For this analysis, our hypothesis was that one or more genes existed for which the expression level was elevated in pickle mutants, regardless of uniconazole treatment. This hypothesis was based on an expression pattern similar to that of LEC1 and LEC2. Thus the primary contrast of interest was the main effect of genotype (C_1). Because we were only looking for a similar pattern (up regulation), the power to detect up regulated genes increased. Use of prior information to increase power is more cost effective than increasing the number of biological replicates. In other experiments additional contrasts may be of equal or greater importance, this may be particularly true of the interaction of genotypes with uniconazole treatment (C_3), which test the hypothesis that application of uniconazole has a different effect on one genotype than the other.

The critical value of t depends on a number of factors, including one- vs. two-sided alternatives, degrees of freedom (df) for estimation of error variance, and acceptable type I error rates. Choosing an acceptable Type I error rate is discussed in the next section.

3.2 Generalized experiment wise error rate (GFWER(k))

Experimenters have long recognized that if a comparison wise type I error rate (CWER) is used across a great number of tests, a large proportion of declared significant differences would be false. For example analysis of array data involves thousands of comparisons, consequently, if a per comparison error rate of 0.05 were used for our analysis, more than 413 of the 8256 tests would be expected to be declared significant by chance alone. The most widely used approaches to control Type I errors in multiple tests is based on controlling the family wise Type I error rate (FWER) (Fernando *et al.*, 2004). The FWER is the probability of rejecting one or more true null hypotheses, i.e. the probability of accepting one or more false positives. A common method for controlling the FWER is the Bonferroni or Sidak (1967) adjustments.

However, the FWER with those adjustments is too conservative if the cost of false negatives is high relative to the cost of false positives, i.e. they sacrifice power to avoid accepting false positives. Methods have been developed to address this issue by allowing for some false positives among those declared significant, such as the false discovery rate (FDR, Benjamini and Hochberg, 1995; Reiner. *et al.*, 2003; see Nguyen, 2005 for general discussion on this issue). Alternatives to the FDR have since been proposed that take into account the expected number of false null hypothesis and other modifications (see Fernando *et al.*, 2004 for review). However, all methods used to estimate an FDR make assumptions about the distribution of truly expressed genes. As a result, the FDR will either be too liberal or conservative.

Here we present an alternative that does not attempt to establish an FDR. Rather the method is an extension of the FWER methodology to allow for a higher family wise error rate. The development is as follows: Assume a strictly null distribution from which N independent test statistics are computed, from which N independent decisions are made at the same critical threshold level. The probability that any one decision is incorrect is p. An incorrect decision is defined as rejecting a true null hypothesis. With multiple tests, the probability of exactly m incorrect and $N - m$ correct decisions is

$$P(m = \text{ incorrect } | N = \text{ decisions }) = \binom{N}{m} p^m (1-p)^{N-m} \tag{3.3}$$

The usual FWER $= \xi(1)$ is the probability of rejecting 1 or more true null hypotheses found as:

$$\xi(1) = \sum_{m=1}^{N} p^m (1-p)^{N-m} \tag{3.4}$$

or equivalently, 1 minus the probability of no incorrect decisions,

$$\xi(1) = 1 - (1-p)^N \tag{3.5}$$

which is Sidak's (1967) equation. The value of p per comparison (CWER) is found such that the $\xi(1)$ is achieved, i.e.

$$p = 1 - e^{[\ln\{1-\xi(1)\}/N]} \tag{3.6}$$

Stated in the reverse, there is a 1-FWER probability of *no* incorrect decisions among the N decisions made, i.e.

$$\omega(k) = \sum_{m=1}^{k-1} \binom{N}{m} p^m (1-p)^{N-m} \tag{3.7}$$

A generalization of this procedure is to divide the total probability of making Type I errors into parts associated with how many errors are likely to be made at a given probability. Among the N decisions made, define $\xi(k)$ as the probability of rejecting k or more true null hypotheses and $\omega(k)$ as the probability of rejecting fewer than k true null hypotheses, $\xi(k) + \omega(k) = 1$, where

$$\xi(k) = \sum_{m=k}^{N} p^m (1-p)^{N-m} \tag{3.8}$$

$$\omega(k) = \sum_{m=0}^{k-1} p^m (1-p)^{N-m} \tag{3.9}$$

If for a given k, the value for $\xi(k)$ is set to a small value, then among those tests declared significant, one accepts that there will be a high probability of $k-1$ false positives plus a low probability of k or more false positives. Therefore, a new type of error rate is defined as GFWER(k), which is strictly the probability of making k or more incorrect decisions at a given level of p, and ignores the probability of less than k Type I errors. The latter type of errors are considered acceptable in order to gain power and decrease the Type II error rate. For a more general development of the generalized family wise error rate see van der Laan (2004). The GFWER(k) cannot be solved for directly, but solutions can be found by iteration. SAS source code used to compute adjusted p values for any $\xi(k)$ and N is given at our web site. However, Equations (3.8) and (3.9) can also be approximated by the normal as follows: If X is binomial with n trials and probability of success p, then

$$P[X > r] \approx \Phi\left(\frac{r - np}{\sqrt{np(1-p)}}\right),$$

where Φ is the cumulative distribution of standard normal distribution.

Tables 3 and 4 give p values for, respectively, a one- and two-tailed alternative, and $\xi(k) = .05$. Associated critical values of t are given in Tables 5 and 6 for experiments with 6 and 60 df for estimating error variance. Note that for all k values, the critical value of t for a one-sided test

is between 12 and 13% smaller, with corresponding increases in power. Table 3 shows that by allowing for 2 or more false positives in a 1 sided t-test increases the adjusted p value by 6.5 times, and thereby also increasing the power of the test. Results presented in Tables 3 show that for the range of N examined (i.e. $N > 1000$) the ratios of p values for GFWER(k) to that of GFWER(1) are independent of N. Thus, for such chips, once the Sidak p values are found, GFWER(k) can be found by multiplication using the constants given in the table.

Table 3: Adjusted p values for $k = 1$ to 5, chips of size 1,000 and 50,000 and a one-tailed GFWER(k)=5%.

k	Number of Tests			
	1,000		50,000	
	p-value$\times 10^6$	Ratio*	p-value $\times 10^6$	Ratio*
1	51.29		1.025	
2	335.02	6.5	6.70	6.5
3	783.41	15.3	15.66	15.3
4	1320.01	25.7	26.38	25.7
5	1913.31	37.3	38.23	37.3

* Ratio of p-values to that of GFWER(1)

Table 4: Adjusted p-values for $k = 1$ to 5, chips of size 1,000 and 50,000 and a two-tailed GFWER(k)=5%.

k	Number of Tests	
	1,000	50,000
	p-value$\times 10^6$	p-value $\times 10^6$
1	25.63	.521
2	167.5	3.35
3	391.7	7.83
4	660.0	13.19
5	956.65	19.15

An important issue is what value of k should one use. The value of k should be set as small as possible without sacrificing too much power. For an experiment of a given size, the rate at which power increases is dependent on the critical value of t. Examination of Tables 5 and 6 shows that the greatest decrease in the critical value of t, with either large or small experiment, comes from increasing k from 1 to 2. For large experiments increasing k beyond 2, or for small increasing k beyond 3, brings about much smaller incremental decreases in t. From these results, some general guidelines can be deduced for choice of k. Regardless of the number of spots on a chip, a k value of 2 or 3 should be adequate for large and small experiments respectively.

Table 5: The six genes for which the qRT-PCR assay detected no expression in untreated wild type seed. Transcripts were detected in untreated *pickle* seeds. Transcripts were also detected for both wild type and *pickle* seeds (Uwt and Upkl) germinated in the presence of uniconazole-*p*, thus permitting calculation of Upickle fold change relative to Uwt. The mean values from the arrays are included for illustration. #Pr is the number of times Affymetrix Microarray Suite software (v. 5.0) labeled a gene 'present' for the 16 gene chips used for this investigation.

AGI Code		Mean values (4 chips)				qRT-PCR fold change				Putative ID/function
	#Pr	wt	pickle	Uwt	Upickle	wt	pickle	Uwt	Upickle	
At3g16410	16	4535	14225	3630	18074	−	+	1	204.3	Jacalin type lectin
At4g27140	15	1000	2365	417	3214	−	+	1	8.95	2S1 seed storage protein
At1g67330	4	170	701	97	1255	−	+	1	3.75	uncharacterized
At5g13930	16	18459	37623	18471	46304	−	+	1	1.75	TT4/chalcone synthase
At3g23220	16	1611	2606	1364	2560	−	+	1	1.43	ERF1/transcription factor
At1g09750	16	2005	3200	2178	5193	−	+	1	0.58	nucleoid-like protein

Table 6: Presence of uniconazole-*p* increases derepression of *PICKLE*-dependent genes in *pickle* seedlings.

	qPCR Ratios			Putative Function
AGI Code	pkl/Wt	Upk/Wt	Upk/pkl	
At5g01600	2.90	8.82	3.04	maturation
At3g16420	4.75	11.41	2.40	defense
At1g73190	1.79	2.96	1.65	maturation
At2g28790	1.77	2.90	1.64	
At1g20620	2.33	3.51	1.50	maturation
At5g54740	2.97	4.10	1.43	
At4g19810	2.43	3.33	1.37	
At3g52500	5.55	7.55	1.36	
At3g16430	1.55	2.08	1.34	defense
At1g05510	1.88	2.53	1.34	
At3g16460	4.64	5.22	1.13	defense
At4g08685	2.61	2.78	1.06	
At2g35810	4.24	4.24	1.00	
At2g19590	2.50	2.24	0.90	
At4g37410	2.38	1.95	0.82	
At5g12030	5.36	2.68	0.50	desiccation

4. Biological Verification: QRT-PCR analysis

Those genes found significant with ANOVA were re-analyzed using qRT-PCR to compare results. The qRT-PCR method, while more precise than the chip analysis, is still subject to error. The method is based on PCR amplification of mRNA in the sample until a pre-determined threshold is obtained. Because the amplification is a doubling with each cycle, the accuracy of the method is questionable if there exists less than a 2 fold difference in mRNA between the two treatments. qRT-PCR is also subject to biological variability between samples and should therefore also be replicated and treated to statistical analysis. However, replicated qRT-PCR analysis for

each gene would be extremely expensive and time consuming. Therefore within the limitations of this experiment, and recognizing those limitations, we defined confirmation of *PICKLE*-dependent expression as a two-fold or greater increase in expression level of a given gene in *pickle* versus wild-type seed when grown in either the absence or presence of uniconazole-P. qRT-PCR was used to compare transcript levels in *pickle* versus wild-type seed grown in the absence of uniconazole-P as well as transcript levels in *pickle* versus wild-type seed grown in the presence of uniconazole-P.

Quantitative RT-PCR was performed on an ABI sequence detection system using RNA from one of the biological replicates previously generated (Rider *et al.*, 2003). Oligonucleotide primer sequences and primer concentrations used are listed in supplementary Table 2S available at the web site.

5. Results and Discussion

5.1 Statistical issues

For this experiment, we used $\xi(2) = .05$. Allowing for one false positive raised the adjusted p value from 6.21×10^{-6} to 4.1×10^{-5} and correspondingly increased the power of the test The ANOVA method selected 43 genes, less than one of which was expected to be a false positive based on the experimentwise selection criteria that we employed ($8256 \times 4.1 \times 10^{-5} = .33$). Our qRT-PCR analysis supported 36 of the 43 genes (Figure 1). A surprising result of this study was that qRT-PCR did not detect transcripts in wild-type seeds for 6 of the 43 genes identified as having expression differences based on analysis of the array data (Table 5). Although this observation is consistent with the hypothesis that PICKLE represses expression of these genes in wild-type seeds to facilitate the developmental transition from embryo to seedling, the array expression values did not suggest absence of transcripts in wild-type seeds.

figure 1 about here

There are at least two possible explanations for the elevated number of observed false positives. Affymetrix constructed this GeneChip when the sequence of the Arabidopsis genome was only partially completed. Inflated expression values for some oligos may have arisen from cross hybridization to unintended targets. In fact, two of the false positives were false because qRT-PCR detected no expression in germinating seeds under any condition. Alternatively, as previously discussed, the discrepancy may be due to different criteria used to determine success for each method. The qRT-PCR data should only be viewed as supporting evidence, not confirmatory.

5.2 Biological Inferences

PICKLE is necessary to repress expression of embryonic traits in Arabidopsis seedlings. Previous analysis of genes that exhibit *PICKLE*-dependent repression identified genes associated with various stages of seed development. ANOVA identified genes associated with seed development, including 2S albumin genes, HSP17.6, and several lectin-like genes (Guerche *et al.*, 1990; Lenman *et al.*, 1993; Ruuska *et al.*, 2002; Sun *et al.*, 2001). In all, 10 of the genes (28%) identified and confirmed by qRT-PCR analysis were associated with embryo development or exhibit sequence similarity to genes involved in embryo development (Table 1S, available at the web site). Additional studies will be necessary to determine if the other 26 genes that showed PICKLE-dependent expression in the germinating seed are also involved in some aspect of embryo development. Previous expression analysis did not suggest a specific role for uniconazole-P in increasing penetrance

of the pickle root phenotype in *pickle* seedlings (Rider *et al.*, 2003). This analysis revealed that the extent of derepression of many of the genes that exhibit *PICKLE*-dependent repression is enhanced by the presence of uniconazole-P (Figure 1, black bars versus white bars). The magnitude of this enhancement, however, was often due in large part to the fact that the presence of uniconazole-P resulted in decreased expression of the gene relative to wild-type seed imbibed in the absence of uniconazole-P (data not shown).

In order to examine the effect of combining the *pickle* mutation with exposure to uniconazole-P, we compared the fold change values of genes in *pickle* versus wt seedlings (pkl/wt) and the fold change values of genes in *pickle* treated with uniconazole-P versus wt seedlings (Upkl/wt) as determined by qRT-PCR (Table 6). In order to make this analysis comparable to previous analysis of the dataset, either the ratio pkl/wt or the ratio Upkl/wt or both had to be ≥ 2 for a gene to be included in this analysis. Genes for which a transcript was not detected in wild-type seedlings were excluded from this analysis. Sixteen genes identified with ANOVA met these expression criteria. We found that the presence of uniconazole-P did increase expression of many of these genes in *pickle* seedlings; the transcript level of 10 genes increased 33% or more when *pickle* seeds were imbibed in the presence of uniconazole-P. In contrast, a previous analysis of the same array data identified no genes for which the corresponding transcript was increased by treatment of *pickle* seeds with uniconazole-P (Rider *et al.*, 2003).

Uniconazole-P increases the probability that primary roots of the *pickle* mutant will express embryonic differentiation traits (Ogas *et al.*, 1997). Genes associated with seed development exhibit elevated expression in *pickle* seedlings, suggesting that the expression of these genes contributes to the ability of *pickle* seedling to express embryonic traits after germination (Rider *et al.*, 2003). The discovery that the presence of uniconazole-P enhances the expression of 10 genes in *pickle* seedlings, 5 of which (50%) are involved in seed development or exhibit sequence similarity to genes involved in seed development, suggests for the first time that the increased penetrance of embryonic traits in *pickle* seedlings treated with uniconazole-P may be mediated in part through changes in gene expression. Specifically, our results are consistent with the hypothesis that GA acts in concert with PICKLE during germination to repress expression of genes that promote embryonic traits. Further characterization of the genes identified here may facilitate subsequent genetic and biochemical analysis of the GA signal transduction pathway that mediates this response.

5.3 Utility

We have shown that a simple ANOVA method can identify a manageable number of candidate genes for differential expression from a gene expression array, most of which were real. Although we only applied the approach to an experiment that incorporated a simple class-by-treatment design, it is applicable to any full factorial design and is computationally straightforward. Previous analysis of the array data employed a modified fold change (MFC) approach (Rider *et al.*, 2003) and failed to detect many of the genes identified by ANOVA. In addition, our current analysis demonstrates that treatment of *pickle* seedlings with uniconazole-P enhances the derepression of *PICKLE*-dependent genes during germination. These results reinforce the power of ANOVA versus a method that emphasizes fold-change.

The practical utility of the GFWER(k) method is derived from allowing the user to influence the number of genes identified by selecting the appropriate value for k, the number of false positives allowed above the threshold significance level. A critical question is what value of k will result in the greatest increase in power with the lowest number of Type I errors. A simple power analysis showed that regardless of the number of spots on a chip, a k value of 2 or 3 should be adequate for large and small experiments respectively. Although the GFWER(k) and FDR are closely related and greatly increase the power of the experiment by relaxing the Type 1 error rate, the application of the GFWER(k) does not attempt to project an FDR, rather, we only set the maximum number of false

positives under the null hypothesis. Calculations for an exact FDR would require knowledge of 1) the number of truly expressed genes, 2) the signal to noise ratio, and 3) their distribution. Without knowing these factors, the FDR as calculated by any of the current methods is an approximation. As a result, the GFWER(k) may be more or less conservative than FDR methods, depending on the particular experiment. However, the GFWER(k) is constant and independent of the experiment, which in itself is appealing. This gives rise to another interesting difference between the methods. The expected number of false positives can be determined *a priori* with the GFWER(k) because the rate is independent of the data, whereas with the FDR (and newer methods as reviewed by Fernando *et al.*, 2004) calculations are dependent on the data and one has to wait until the list is generated to determine what the expected number of false positives will be. This difference could be critical in the planning stage of an experiment.

Acknowledgements

We thank Guilherme Rosa for helpful comments and discussion during preparation of the manuscript, Jim Henderson for helping to generate the RNA used for this study and Howard Edenberg for chip hybridization and calculation of expression values. **Funding**: Partial Funding for this research was provided by: JO National Institutes of Health (R01GM059770-01A1 and 5R01GM59770-02); JRS The Indiana 21-st Century Research and Development Fund and Purdue Research Foundation; SDR BASF ; WM USDA/NRI 9803430 Animal Genetic Mechanisms. This is journal paper number 17605 of the Purdue University Agricultural Experiment Station. **Supplemental tables** used in the analysis are available at:
http://www.biochem.purdue.edu/research/ogas_lab/arrays/JPmethod/.

References

Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15**, 63-78.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate — A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289-300.

Black M. A., and Doerge R. W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**, 1609-1616.

Craig B. A., Black M. A., and Doerge R. W. (2003). Gene expression data: The technology and statistical analysis. *J. Agricultural Biological and Environmental Statistics* **8**, 1-28.

Faccioli, P., Lagonigro, M., Cecco, L., de, Stanca, A., Alberici, R., Terzi, V., and de Cecco, L. (2002). Analysis of differential expression of barley ESTs during cold acclimatization using microarray technology. *Plant Biol.* **4**, 630-639.

Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller M. (2004). Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611-619.

Fujita, N., Jaye, D. L., Kajita, M., Geigerman, C., Moreno, C. S., and Wade, P. A. (2003). MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell* **113**, 207-219.

Gill, J. L. (1978). *Design and Analysis of Experiments in the Animal and Medial Sciences. Volume 1.* The Iowa State University Press.

Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol* **124**, 1570-1581.

Goda, H., Shimada, Y., Asami, T., Fujioka, S., and Yoshida, S. (2002). Microarray analysis of brassinosteroid-regulated genes in arabidopsis. *Plant Physiol* **130**, 1319-1334.

Gu, C. C., Rao, D. C., Stormo, G., Hicks, C., T., and Province, M. A. (2002). Role of gene expression microarray analysis in finding complex disease. *Genes. Genet Epidemiol* **23**, 37-56.

Guerche, P., Tire, C., De Sa, F. G., De Clercq, A., Van Montagu, M., and Krebbers, E. (1990). Differential expression of the arabidopsis 2S albumin genes and the effect of increasing gene family Size. *Plant Cell* **2**, 469-478.

Izumi, K., Kamiya, Y., Sakurai, A., Oshio, H., and Takahashi, N. (1985). Studies of sites of action of a new plant growth retardant (E)-1-(4-chlorophenyl)-4,4-dimethyl-2-(1,2,4-triazol-1-yl)-1-penten-3-ol (S-3307) and comparative effects of its stereoisomers in a cell-free system from Cucurbita maxima. *Plant Cell Physiol* **26**, 821-827.

Kerr M. K., andChurchill G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123-128.

Kerr M. K., Martin M., and Churchill G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computational Biology* **7**, 819-837.

Kim, J. W., and Wang, X. W. (2003). Gene expression profiling of preneoplastic liver disease and liver cancer: A new era for improved early detection and treatment of these deadly diseases? *Carcinogenesis* **24**, 363-369.

Lee, M. T., Whitmore, G. A., Yukhananov, R. Y. (2003). Analysis of unbalanced microarray data. *Journal of Data Science* **1**, 103-121.

Lenman, M., Falk, A., Rodin, J., Hoglund, A. S., Ek, B., and Rask, L. (1993). Differential expression of myrosinase gene families. *Plant Physiol* **103**, 703-711.

Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z., and Peng, J. (2003). 15,000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Research* **13**, 455-466.

Lotan, T., Ohto, M., Yee, K. M., West, M. A., Lo, R., Kwong, R. W., Yamagishi, K., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (1998). Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* **93**, 1195-1205.

Lundquist, H., Oredsson, S., and Akesson, B. (2002). Effect of quercetin on gene expression in human cells as measured by microarrays — a pilot study. Paper presented at: Health promoting compounds in vegetables and fruit. Proceedings of a workshop in Karrebaeksminde, Denmark, 6 8 November, 2002. DIAS Report, Horticulture. 2002, No.29, 77 80; 10 ref. (Danmarks JordbrugsForskning; Tjele; Denmark).

Martinez, J. M., Afshari, C. A., Bushel, P. R., Masuda, A., Takahashi, T., and Walker, N. J. (2002). Differential toxicogenomic responses to 2,3,7,8-tetrachlorodibenzo-p-dioxin in malignant and nonmalignant human airway epithelial cells. *Toxicol Sci.* **69**, 409-423.

Nguyen, D. V., Wang, N. and Carroll, R. J. (2004). Evaluation of missing value estimation for microarray data. *Journal of Data Science* **2**, 347-370.

Nguyen, D. V. (2005). A unified computational framework to compare direct and sequential false discovery rate algorithms for exploratory DNA microarray studies. *J. Data Science* (In Press).

Nitin, J., Thatte, j. Braciale, T., Ley, K., O'Connell, M. and Lee. J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 1945-1951.

Ogas, J., Cheng, J. C., Sung, Z. R., and Somerville, C. (1997). Cellular differentiation regulated by gibberellin in the arabidopsis thaliana pickle mutant. *Science* **277**, 91-94.

Ogas, J., Kaufmann, S., Henderson, J., and Somerville, C. (1999). PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in arabidopsis. *Proc. Natl. Acad. Sci. USA* **96**, 13839-13844.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* **2**, 418-429.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-375.

Rider, S. D., Henderson, J. T., Jerome, R. E., Edenberg, H. J., Romero-Severson, J., and Ogas, J. (2003). Coordinate repression of regulators of embryonic identity by PICKLE during germination in arabidopsis. *Plant J.* **35**, 33-43.

Ruuska, S. A., Girke, T., Benning, C., and Ohlrogge, J. B. (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* **14**, 1191-1206.

Sidak, Z. (1967). Rectangular confidence regions for means of multivariate normal distributions. *J. Amer. Statis. Asso.* **62**, 626-********

Sokal, R. R., and Rohlf, F. J. (1995). *Biometry, Third edition.* W. H. Freeman and Company.

Stone, S. L., Kwong, L. W., Yee, K. M., Pelletier, J., Lepiniec, L., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2001). LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. USA* **98**, 11806-11811.

Storey, J. D. and Tibshiran, R. (2003a). Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci, USA* **100**, 9440-9445.

Storey, J. D. and R. Tibshirani, R. (2003b). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (Edited by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S.L . Zeger). Springer.

Sun, W., Bernard, C., van de Cotte, B., Van Montagu, M., and Verbruggen, N. (2001). At-HSP17.6A, encoding a small heat-shock protein in Arabidopsis, can enhance osmotolerance upon overexpression. *Plant J.* **27**, 407-415.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*[7]

Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Computational Biology* **8**, 625-637.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**, 579-588.

Zhang, H., Yu, C. Y., and Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *Proc. Natl. Acad. Sci. USA* **100**, 4168-4172.

[7]See http://www.bepress.com/sagmb/vol3/iss1/art14/

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爲: 將 likelihood function $L(\phi|\mathbf{x})$ 找出來。因爲一旦找了出來, 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, 多因爲不完整數據有時是因爲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 。

讀我的註脚需參考原始論文。見

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

以下爲我的註脚。

---

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, 並有實際的爲法來解, 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。因爲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. 這一句和摘要是同一意思, 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, 因爲明擺著不肯用心, 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, 因爲他們的身份和品味擺在那兒。新手以不做爲宜。但 "specify, generality, wide range of examples" 是用得好的。

p.1, introduction, line 6. 在此先做 exponential family , 是因爲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。但眞的怎樣來? 用嘴巴是說不清楚的, 必須用數學式子, 因爲後面要做證明, 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 $X_1, \ldots, X_{100}$ 算是 complete data, 它的 likelihood function 是 $f(x_1, \ldots, x_{100}|\phi)$ 。若我們丟掉了 x_{100} , 則不完整的數據就是 $X_1, \ldots, X_{99}$, 而我們可以用積分的方法找到這二者者的關係:

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

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) 未必爲 unique 。當然, 我們可在 $X_1, \ldots, X_{101}$ 中, 丟掉 X_{100}, X_{101} , 然後再得到 $X_1, \ldots, X_{99}$ 。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^* (眞正的收斂極限), 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, 是眞有實驗的。

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. 此一段敍述 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" 。這等於是說別人都沒有進一步的理論 —— 一直都在說自己好, 卻不帶火氣。

p.3, line 10. 注意到 "always increse 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 bonomial, 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, 這些話都是爲「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。因爲若沒有它, 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" , 因爲它可看成爲一組方程式。你當然可以用 "equation" , 將之視作一個用 vector 來表的 eqution 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. "familar form" 此詞用得精準, 因爲 (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 $\log, \sin, \tan, \ldots$ 而不是 $log, sin, tan, \ldots$ 。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 $\mathbf{x}$ 是整數, 但 $E(\mathbf{x})$ 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。

[6] L. 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。愼之, 愼之。

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, 因爲基本上, 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 $\phi^{(p)} \to \phi^*$, 最後還得去說 $\phi^{(p)} \to MLE$ 。

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 (因爲太 detail), 但研究上你該都理清楚。這四個定理都在說甚麼? 眞正所需的條件爲何? 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。因爲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。

這些都是號稱爲了嚴謹而玩的文字遊戲。注意到眞正的要點是 GEM 基本上是 by definition, it works 。這些定理都是化妝後的結果: 眞正用 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眞的證明過, 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。爲何用到 second derivative 時, 便會 speed up convergence? (3.19) or (3.26) 說的是甚麼? 爲何與收斂速度有關? 此類言語, 不容易說。因爲說得不對時, 內行人便看得出來, 便成爲畫虎類犬了。

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, lien 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)。因爲 $Q(\phi'|\phi)$ 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, 就不足爲奇了。

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。它的眞實例容其實只有 (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?

September 4, 2006

第十部分: 人在江湖

武林弟子藝成之日, 拜別師父下山, 就進入江湖。以後是否成名立萬, 全靠自己的努力。

當你拿到博士的那一天, 如果你還預備搞科研, 如果你不是去那種不准對外發表的機密單位, 那麼你就算是進入了「以發表論文來求生存」的江湖。江湖中有黑道白道。當然, 你是名門弟子, 一開始算是白道。

白道有白道的規矩。下面所列, 不全是白道的規矩。有些是你師父在平常就該教的。如果他忘了, 你不妨拿來參考。

不弄虛作假

寫論文的目的 是爲研究的成果做紀錄, 做紀錄的目的是讓後面的人據此可以更向前進。我們所做的這類學問是累積型的[1], 牛頓說他的成績, 都是因爲他能站在巨人的肩上的原故。要看高看遠看深, 當然要站得高。

因爲是累積型的工作, 所以要正確。若你的結果毫無用處, 別人根本不會用, 就算錯了也不會對科學有影響。但若是你的結果表面甚好, 但其中有不能瀰補的大漏洞[2], 別人容易由此可以得到更好的結果, 這樣就會有大問題了。

學術論文的基本要求是誠實和嚴謹。準此, 論文裡的小錯可以原諒, 但不可以故意弄錯, 尤其不能在關鍵處故意弄錯[3]。

引用明確

論文中所引用的文獻, 主要以專書和其它學利期刊的論文爲主[4]。至於一般的技術報告, 開會的紀錄 (conference proceedings)[5], 專書的一章, 網路上所貼的文章, 某處的碩士論文等 ..., 這類比較不入流的事物, 能避則避[6] —— 這些都是塗在臉化妝品, 你總不希望評審人覺得你在用劣質化妝品罷。

引用論文時, 可以不必提你所引用的在文中何處 (有時你還是要提在 section 3.4 或者 equation (4.2) 之類)。但引用專書時, 則最好要提章節或頁碼, 以方便讀者去查。這一方面是禮貌, 一方面則表示「 你眞的讀過」。

因此, 你引用的文獻, 應該是你讀過的 (當然不必是全部)。若引用的是冷門資料, 有

[1]像老子的《道德經》之類, 是非累積型的。

[2]邏輯這回事是有趣的, 若是你的錯誤結果被廣泛利用, 從這條路走下去, 不久就會發現你論文裡有問題。這時, 無心的錯或是有心的誤導便一目了然。

[3]另一個理由是: 在關鍵處, 這類錯太容易被內行抓住。例如我也在做同一問題, 有同一瓶頸, 爲甚麼你做得出來但我做不出來? 競爭者總會用心查一查。

[4]較弱期刊的論文效引用多了, 會減少你論文的份量。

[5]這有例外: 如 Proceedings of the Berkeley Symposium , 雖是開會紀錄, 卻是極有聲望的。其它圈內的幾篇有名的技術報告 (如 Efron 在前述 bootstrap 的文章裡所引用的關於 infinitesimal jackknife 的 Bell Labs TM —— 因爲反而讓你顯得內行。) 則不在此限。

[6]因此你就知道爲甚麼要少在三流期刊上投稿, 因爲會較少人引用, 而 citation 的數目, 暗示你論文重要的程度。

些期刊就會要求你寄一仍複印去 (給評審人參考)。

誰先做出

"Priority" 是學術工作者最在乎的事情之一。因爲研究要求的就是求新求變, 因此「這件事是誰先做出來的」就不能含糊。例如引用到 jackknife 時, 你就需要提到 Quenoulli-Tukey 。但爲甚麼 Tukey (1958)[7] 明明在 Quenoulli (1956) 之後, 卻要提起? 你若是內行, 就知道這兩篇文章有基本的不同, 而 Tukey 的看法要深刻得多。

準此, 你也要保護你的 priority 。明明是你做的, 被人搶先一步發表, 那你的論文 (至少在好期刊上) 就泡湯了。

關於 Priority 的保護, 口說無憑。我曾聽到的故事說: 有一篇投出的文章被拒, 但過了一陣, 這個作者被要求評審一篇類似的文章, 眞是「類似」, 因爲這篇新文章和他的「舊作」, 相差只在他的 α 一律被改爲 β 。類似的可怕故事並不是沒有, 投稿到有聲望的期刊一般不會遇到這樣事, 當你小有名氣之後別人也不見得眞敢這樣做。但是, 把你的文章壓一壓, 同時再做一篇比你的內容更多的文章吃掉你大部分可以做的 extension , 卻是可能的。

保護的方法是: 要求你的工作單位建立一套有時間、有文號的技術報告系列。這樣, 你的工作出來的成績, 因爲你工作單位的背書, 就比較受到保護。例如你大可寄信給期刊的主編、對方的學術領導、或者相關國際學會的理事長等, 說你在 1999 年 7 月已做出該結果來, 有本校的技術報告爲證。這樣的信當然會有影響[8]。如果那人抄得十分明確, 他的日子不會好過。因爲學術圈子其實不大, 這樣的作爲會貽笑國際也。

不可硬拗

若是你的文章有錯被別人挑出來, 當然第一步是確認是否眞是你的錯[9]。若沒錯, 當然是頂回去, 用詞不必強烈, 維持學術上的樸克面孔就好。別忘了還是要感謝別人讀了你的文章。若有不能修改的錯, 你也只好承認[10](並還得在用詞上感謝別人)。

長袖可舞

人際關係永遠是重要的, 在學術圈子裡亦然。學術裡會看實力, 學問上的和學術資源上的。例如假如你手上有一大筆預算支援一些項目, 或者你是某不錯期刊的主編, 總會多受一點尊重[11]。沒有辦法 —— 大家都要吃飯。杜甫詩: 同學少年皆不賤, 五陵裘馬自輕肥。現在大家都差不多, 說不定你的學問還好些。但十年後爲甚麼有所不同? 用心的話, 錢愈花愈多, 學術資源也是愈用愈多。你若是老啃著那一畝三分地, 除非裡面能冒出石油, 長遠下去, 場面怎會變大?

「在家靠父母, 出外靠朋友」。單人隻劍的獨行俠, 走江湖可不容易。煮酒放歌, 交友論劍, 也是白道人物可以做的。

---

[7] T 的文章, 其實只是一個半頁的摘要。

[8] 這樣的信, 若是由你的單位領導來寫, 就有「捍衛本單位人員的學珊權益」的意思。

[9] 注意到, 若是別人有錯在先, 你因爲引用錯的結果而犯錯, 這也算你的錯 —— 誰叫你不察, 不夠用功嘛。

[10] 一般好期刊都允許作者去信更正已發現的錯, 這類事當然是愈少愈好。

[11] 但這類事隨著你的權力而來, 也隨著你的權力而去。你最好是做莊周的時候用心做莊周, 做蝴蝶的時候用心做蝴蝶。

熱心而肯辦事的學者機會會多一點。你若是有這樣的服務性格, 不妨多入世一點。有些雜務不妨交給自己的學生做[12], 但這裡有分寸, 不要過分。需知你的學生並不傻, 這類事英文叫 “give and take” , 中文是願打願挨。

退一步沒有海闊天空

做研究指的是在第一線做研究。意思是說你總是在做最新的工作, 要發明或發現新的東西, 而這是要和全世界的一小撮最好的人在比。好期刊不會因爲你的行政職位高低來影響是否接受你的論文。即使你曾經在第一線, 但學問如逆水行舟, 不進則退。你若是有三年的停滯, 你就是有三年的停滯。要追回世界水準, 也許你還有功力, 不需三年。但是你就是要重下功夫才行。

別人也不見得多利害

世上是有一些高手怪傑之類, 吾等小民硬是比不過。但不要怕, IQ 的分布大概還是差不多 normal 。做研究工作就不能沒有氣性。所謂的「氣不可衰」, 前已言及。

多半的人都是在軟木頭上鑽洞。你若是一心搶快, 就不要咪咪摸摸的。將學術資源搞起來 (如要求本單位訂閱重要期刊), 固定去查閱可能的對手有甚麼新的技術報告, 儘量去找一些別人沒有的數據 ....。總之, 別閑著。

別一個方法是做較難一點的。對於較難的問題[13], 因爲投入/產出比不高, 一般精明的搶快型學者是不太肯做的。你若是在偏遠地區, 學術資源少, 不妨試這條路。當然, 你需要注意保護你的 priority 。

自己也沒有多利害

意思是, 不要選自己做不出來的題目[14]。研究工作雖有時候要搶快, 但基本上還是要靠穩扎穩打。選題要 (1) 題目要有一點趣味, (2) 至少看起來要有影響, (3) 自己做得出來 (4) 如果你還有餘力, 留一點空間給別人做。

September 5, 2006

[12]例如在主辦學術會議時, 總需要人辦事。

[13]注意不建議你去做無謂的難題, 要選難而有影響的。收尾型的題目往往難而功勞不大, 畢竟, 把一個房子打掃掉最後的 tough 垃圾再關上門, 是無趣的事。

[14]你可以問: 不去做怎知做不出? Well, 用心試一兩個月無妨, 但是「知止」也是重要的品行。