



國立中山大學應用數學系

碩士論文

Department of Applied Mathematics

National Sun Yat-sen University

Master Thesis

蒙地卡羅統計方法：積分與優化

Monte Carlo Statistical Methods : **Integration and**

Optimization

研究生：潘恬恬

Tian-Tian Pan

指導教授：郭美惠 博士

Dr. Mei-Hui Guo

中華民國 101 年 6 月

June 2012

國立中山大學研究生學位論文審定書

本校應用數學系碩士班

研究生潘恬恬（學號：M982040011）所提論文

蒙地卡羅統計方法：積分與優化

Monte Carlo Statistical Methods : Integration and Optimization

於中華民國 101 年 6 月 27 日經本委員會審查並舉行口試，符合碩士
學位論文標準。

學位考試委員簽章：

召集人 羅夢娜 羅夢娜

委員 郭美惠 郭美惠

委員 張福春 張福春

委員 黃士峰 黃士峰

委員 _____

委員 _____

指導教授(郭美惠)

郭美惠 (簽名)

誌謝

課業上，首先要感謝的是我的指導老師郭美惠教授，在這三年來對我的指導、關心和體諒，無論是在修課上、生活上或論文書寫上都給了我很多建議與協助，讓我能邊學習邊完成各項工作。再來，要感謝張福春教授、羅夢娜教授和林純穗老師在修課上用心的教導，豐富了我許多統計領域裡的重要知識。最後感謝良靖學長、海唐、家豪學弟及曾經一起學習的同學們，和我一起討論課程內容所遇到的問題，一起完成各項作業或報告，在生活中也不忘適時的提醒我重要的訊息、活動和時間，真的很謝謝你們的關心和協助。

生活上，三年來邊工作邊念碩班的日子，雖然時常覺得非常忙碌、疲憊，有時甚至還懷疑自己是否能力不夠無法完成這些任務，還好有我的家人在身邊幫忙我打理生活瑣事，適時地鼓勵我堅持下去。感謝我的阿公、阿嬤、所有家人和好友，從小到大因為有你們一路相伴及祝福，讓我的學業、工作都能順遂圓滿，這過程雖然辛苦，但深深的覺得自己有滿滿的收穫，也因此為自己的進步感到非常高興。

碩班生涯即將結束，希望這段求學的日子，能成為日後美好的回憶。

恬恬 謹誌 2012.06



摘要

本論文主要的參考書本為蒙地卡羅統計方法（Monte Carlo Statistical Methods, second edition），作者為 Robert 及 Casella (2004)，參考章節為第一章～第五章（不含第四章變異數控制部分）。本論文內容主要目的是將此書前五章的內容進行：1. 中文化 2. 修正錯誤 3. 加入公式推導過程 4. 將例題所用的演算法寫成程式等四大工作項目，以及應用模擬退火演算法處理捨入資料對參數估計的問題。其中利用軟體 Mathematica (7版) 編寫各例題所需要的程式碼，並附上實際操作的結果，可作為提供日後有興趣研讀此書或需要處理相關問題的人士參考的工具書。

關鍵詞：接受拒絕法、包絡接受拒絕法、重要抽樣法、模擬退火法、EM方法

捨入資料

Abstract

This paper is refer to the chapter 1 to chapter 5 (except chapter 4) of the book, **Monte Carlo Statistical Methods (second edition)**, the author is Robert and Casella (2004). The goal is to translate the chapter 1 to chapter 5 contents of this book into Chinese, modify the mistakes, add the details of the examples, translate the algorithm of the examples into Mathematica (7th) codes, and use the Simulated Annealing methods to deal with the estimation of parameters by rounding data, and discuss the results. This paper provides Mathematica (7th) codes of almost every example, and show the actual results, so it can be regarded as a toolbook for those people who are interested in reading this book or may solve some problems related to those examples.

Keywords : Accept-Reject Methods, Envelope Accept-Reject Methods,
Importance Sampling, Simulated Annealing, EM Algorithm,
Rounding Error

目 錄

感謝	i
摘要	ii
Abstract	iii
1 介紹(Introduction)	1
1.1 統計模型(Statistical Models)	1
1.2 概似估計法(Likelihood Methods)	4
1.3 貝氏方法(Bayesian Method)	10
1.4 決定性數值法(Deterministic Numerical Method)	16
1.4.1 最佳化(Optimization)	16
1.4.2 積分(Integration)	18
1.4.3 比較(Comparison)	18
2 隨機變數的生成(Random Variable Generation)	20
2.1 簡介(Introduction)	20
2.1.1 一致分佈的模擬(Uniform Simulation)	20
2.1.2 逆變換(The Inverse Transform)	21
2.1.3 其他例題(Alternatives)	23
2.2 一般變換方法(General Transformation Methods)	23
2.3 接受拒絕法(Accept-Reject Method)	28
2.3.1 模擬的基礎定理(The Fundamental Theorem of Simulation) . .	29
2.3.2 接受拒絕演算法(The Accept-Reject Algorithm)	32
2.4 包絡接受拒絕法(Envelope Accept-Reject Method)	34
2.4.1 夾擠原理(The Squeeze Principle)	34
2.4.2 對數凹密度函數(Log-Concave Densities)	36
3 蒙地卡羅積分(Monte Carlo Integration)	43
3.1 簡介(Introduction)	43
3.2 典型蒙地卡羅積分(Classical Monte Carlo Integration)	47
3.3 重要抽樣(Importance Sampling)	52
3.3.1 原理(Principles)	52
3.3.2 變異數有限之估計量(Finite Variance Estimators)	57
3.3.3 重要抽樣與接受拒絕法比較(Import. Sampling vs. AR method)	64

3.4 拉普拉斯近似(Laplace Approximations)	68
4 蒙地卡羅最佳化(Monte Carlo Optimization)	70
4.1 介紹(introduction)	70
4.2 隨機探索(Stochastic Exploration)	71
4.2.1 基本解(A Basic Solution)	71
4.2.2 梯度法(Gradient Methods)	73
4.2.3 模擬退火演算法(Simulated Annealing)	75
4.2.4 事前回饋(Prior Feedback)	80
4.3 隨機逼近算法(Stochastic Approximation)	82
4.3.1 遺失資料模型及去邊際化(Miss. Data Models and Demarginal.)	82
4.3.2 EM演算法(The EM Algorithm)	84
4.3.3 蒙地卡羅EM(Monte Carlo EM)	91
4.3.4 EM 標準差(EM Standard Errors)	95
5 最佳化問題之模擬研究：捨入資料對最大概似估計量的影響	100
5.1 概似函數之參數估計	100
5.1.1 常態分佈的參數估計	101
5.1.2 柯西分佈的參數估計	102
5.1.3 指數分佈的參數估計	103
5.1.4 Gamma 分佈的參數估計	103
5.2 模擬結果	104
參考文獻	106
附錄	111
勘誤表	111

表 目 錄

114

表 1-1 太空梭起飛時的溫度及0-ring零件的狀態.	114
表 2-1 普魯士軍隊遭受馬匹踢死的數據.	114
表 3-1 常態分佈表.	114
表 3-2 列聯表.	114
表 3-3 卡方檢定的截點.	115
表 3-4 事後期望值之估計值.	115
表 3-5 積分值的近似結果.	115
表 4-1 模擬退火演算法估計最小值結果.	115
表 4-2 模擬退火演算法之接受率.	115
表 4-3 參數估計結果.	116
表 4-4 入學考試成績.	116
表 4-5 期望值之最大概似估計量.	116

圖目錄

117

圖 1-1 柯西分佈的概似函數圖.	117
圖 1-2 牛頓-拉福生法.	117
圖 2-1-1 數對 (Y_n, Y_{n+100}) 的散佈圖.	118
圖 2-1-2 數列 Y_n 的直方圖.	118
圖 2-2 $Jöhnk$ 演算法，接受 (U, V) 的機率.	118
圖 2-3 貝它分佈隨機變數的生成.	118
圖 2-4 來自集合 $\{(x, u) : 0 < u < f(x)\}$ 的一致分佈樣本.	119
圖 2-5 $h(x) = \log f(x)$ 的上下界線.	119
圖 2-6 北方針尾鴨資料.	120
圖 2-7 ARS 演算法生成的 5000 筆樣本.	120
圖 2-8 區域 $[x_2, x_3]$ 所對應的機率、ARS 演算法所得樣本.	121
圖 3-1 蒙地卡羅積分法近似結果.	121
圖 3-2-1 虛無假設下之分佈的直方圖及近似卡方分佈的密度函數.	121
圖 3-2-2 移動經驗百分位數.	122
圖 3-3 經驗累積分佈函數.	122
圖 3-4-1 重要抽樣法估計風險 R_1	122
圖 3-4-2 重要抽樣法估計風險 R_2	123
圖 3-5 重要抽樣法估計 $E_f[h_1(X)]$ 結果.	123
圖 3-6 雙重伽瑪分佈估計 $E_f[h_1(X)]$ 結果.	124
圖 3-7 重要抽樣法估計 $E_f[h_2(X)]$ 結果.	124
圖 3-8 重要抽樣法估計 $E_f[h_3(X)]$ 結果.	124
圖 3-9 估計 $E[h_5]$ 的結果.	125
圖 3-10 估計 $E[h_3(X)] = E[x / (1 + x)]$ 的收斂結果.	125

圖 4-1-1 $h(x)$ 的函數圖.	125
圖 4-1-2 $U[0, 1]$ 樣本計算 $h(x)$ 的結果.	126
圖 4-2 $h(x, y)$ 的函數圖.	126
圖 4-3 梯度法所得 θ 的收斂路徑.	127
圖 4-4 模擬退火演算法所得 ($x(t), h(x(t))$) 的軌跡.	127
圖 4-5 模擬退火演算法所得 (x_t, y_t) 的軌跡.	128
圖 4-6-1 三種概似函數的圖形.	128
圖 4-6-2 EM估計結果.	129
圖 4-7 混合模型的對數概似函數.	129
圖 4-8 遺傳連鎖資料之參數的EM估計結果.	129
圖 5-1-1 常態分佈的模擬結果： $\hat{\mu}$ 的bias.	130
圖 5-1-2 常態分佈的模擬結果： $\hat{\sigma}$ 的bias.	131
圖 5-1-3 常態分佈的模擬結果： $\hat{\mu}$ 的variance.	132
圖 5-1-4 常態分佈的模擬結果： $\hat{\sigma}$ 的variance.	133
圖 5-2-1 柯西分佈的模擬結果： $\hat{\theta}$ 的bias.	134
圖 5-2-2 柯西分佈的模擬結果： $\hat{\gamma}$ 的bias.	135
圖 5-2-3 柯西分佈的模擬結果： $\hat{\theta}$ 的variance.	130
圖 5-2-4 柯西分佈的模擬結果： $\hat{\gamma}$ 的variance.	137
圖 5-3-1 指數分佈的模擬結果： $\hat{\lambda}$ 的bias.	138
圖 5-3-2 指數分佈的模擬結果： $\hat{\lambda}$ 的variance.	139
圖 5-4-1 Gamma分佈的模擬結果： $\hat{\alpha}$ 的bias.	140
圖 5-4-2 Gamma分佈的模擬結果： $\hat{\beta}$ 的bias.	141
圖 5-4-3 Gamma分佈的模擬結果： $\hat{\alpha}$ 的variance.	142
圖 5-4-4 Gamma 分佈的模擬結果： $\hat{\beta}$ 的 variance.	143

1 介紹(Introduction)

直至出現有力且可接受的計算方法之前，試驗者為了得到一個不錯的結果，時常會面臨一些作法上的選擇，例如可能想藉由某現象來描述一個精確的模型(此做法通常是排除計算封閉解的情況)，或是選擇一個能計算的標準模型，但後者存在的問題是可能無法得到真實模型的封閉形式。此難題出現在許多統計學應用的分支上，例如電子工程、航空、生物、網際網路、天文...等。為了能使用真實模型，這些學科的研究員時常為他們所遇到的問題研發原始的模型配適法(特別是物理學家常使用此方法，Markov chain Monte Carlo methods 就是源自於此)，而傳統的分析方法(如一般的數值分析法)就不太適用於此。以下我們來檢視某些特殊的統計模型和有助於發展以模擬為基準之推論的方法。

1.1 統計模型(Statistical Models)

在純統計建構中，計算上的困難在建構機率模型及對模型做統計推論(如：估計、預測、檢定、選擇變數...等)兩大階段都會遇到。在第一種情況中，對於現象的成因做仔細的描述可能會導致機率結構太過複雜以致於不好將模型參數化，此外，對於感興趣的量可能沒有辦法得到封閉形式的估計量。「專家系統」(或稱圖形結構，可見於醫學、物理學、金融...等)是一種與此錯綜性有關的計畫。

另一個因模型錯綜性以致無法明確描述所欲描述之現象的情況，出現在經濟學裡(當然也存在其他領域)那些隱藏變數(或遺失變數)模型的結構問題。給定一個簡單的模型，聚集或移除模型中的某些項有時可能造成這樣複雜的結構出現，此時模擬真的是唯一可以進行推論的方法。在這種情形下一種常用的估計方法是"EM algorithm"(Dempster et al.1997)這將在第三章介紹。接下來的例子我們介紹一個常見的遺失資料的情形，這概念及方法的使用將在本書中重複出現。

例題1.1 削失資料模型(Censored data models)

削失資料模型是指遺失資料的模型其資料的分布並非直接來自抽樣，為了在這種模型中找估計值及做推論，通常需要複雜的計算並排除分析的答案。在典型的簡單統計模型中，我們會從具有分佈為 $f(y|\theta)$ 的母群體中得到獨立的隨機變數 Y_1, Y_2, \dots, Y_n ，這組樣本的分佈為 $\prod_{i=1}^n f(y_i|\theta)$ ，將以此分佈為基礎來對 θ 做推論。

在許多研究裡，特別是醫學統計，我們必須處理削失的隨機變數，也就是我們可能要觀察的是 $\min\{Y_1, \bar{u}\}$ (\bar{u} 表示一個常數)，而不是觀察 Y_1 。舉例來說，若 Y_1 是表示一個病患接受某種特殊治療下的存活時間， \bar{u} 表示此研究進行的時間(設 $\bar{u} = 5$ 年)，若此病人存活時間超過 5 年，則我們會選擇觀察此削失的值 \bar{u} 而不是那存活時間 Y_1 ，此選擇將導

致樣本密度的估計更困難。

以下我們來看一個特別的例子，設 $X \sim N(\theta, \sigma^2)$, $Y \sim N(\mu, \tau^2)$ ，
 $Z = X \wedge Y = \min(X, Y)$ 的分佈是

$$\begin{aligned} F(z) &= P(Z \leq z) = 1 - P(Z \geq z) = 1 - P(X \geq z)P(Y \geq z) \\ &= 1 - P\left(\frac{X - \theta}{\sigma} \geq \frac{z - \theta}{\sigma}\right)P\left(\frac{Y - \mu}{\tau} \geq \frac{z - \mu}{\tau}\right) \\ &= 1 - \left[1 - \Phi\left(\frac{z - \theta}{\sigma}\right)\right] \left[1 - \Phi\left(\frac{z - \mu}{\tau}\right)\right] \end{aligned}$$

故

$$\begin{aligned} f(z) &= \frac{dF(z)}{dz} \\ &= \left[1 - \Phi\left(\frac{z - \theta}{\sigma}\right)\right] \times \tau^{-1} \varphi\left(\frac{z - \mu}{\tau}\right) \\ &\quad + \left[1 - \Phi\left(\frac{z - \mu}{\tau}\right)\right] \sigma^{-1} \varphi\left(\frac{z - \theta}{\sigma}\right) \end{aligned} \tag{1.1}$$

相同地，若 X 是韋伯分佈(Weibull distribution)， $We(\alpha, \beta)$ ，且密度函數為

$$f(x) = \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} \text{ on } \mathbf{R}^+,$$

我們欲觀察的刪失的變數為 $Z = X \wedge \omega = \min(X, \omega)$ ，其中 ω 是常數，則 Z 的密度函數為

$$\begin{aligned} F(z) &= P(Z \leq z) = P(\min(X, \omega) \leq z) \\ &= P(X \leq \omega \leq z) + P(X \leq z \leq \omega) + P(\omega \leq X \leq z) + P(\omega \leq z \leq X) \\ &= \int_0^\omega \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} dx + \int_0^z \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} dx + \int_\omega^z \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} dx \\ &\quad + \int_z^\infty \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} dx \end{aligned}$$

$$f(z) = \frac{dF(z)}{dz} = \alpha \beta z^{\alpha-1} e^{-\beta z^\alpha} I_{z \leq \omega} + \left(\int_\omega^\infty \alpha \beta x^{\alpha-1} e^{-\beta x^\alpha} dx \right) \delta_\omega(z) \tag{1.2}$$

上述 $\delta_a(\cdot)$ 是在 a 的迪拉克密度函數(Dirac mass)。此例子中，迪拉克密度函數的權重， $P(X \geq \omega)$ ，是可以明確地計算出來的。

(1.1)及(1.2)的分佈常見於品管應用上，此處 ω 表示檢驗產品歷經的時間，而感興趣的是檢測到不良品所需的時間。若實驗結束時產線依然正常運作，則想觀察的檢測到不良品所需的時間就被刪除了。類似的例子還有在長期研究某種疾病的例子中，某病患可能因為死亡或失去追蹤而脫離了此研究並導致觀察的時間被刪除。

例題1.2 混合模型(Mixture models)

分佈的混合模型是建立在假設觀測值 X_i 來自 k 個附有機率值 (p_j) 的分佈 (f_j) 其中之一，其密度函數為

$$X \sim p_1 f_1(x) + \dots + p_k f_k(x) \quad (1.3)$$

如果我們觀察一個具有獨立隨機變數 (X_1, \dots, X_n) 的樣本，則樣本的密度函數為

$$\prod_{i=1}^n p_1 f_1(x_i) + \dots + p_k f_k(x_i).$$

當 $f_j(x) = f(x|\theta_j)$ ，在給定 $(\theta_1, \dots, \theta_n, p_1, \dots, p_n)$ 下的概似估計只需要 $\mathcal{O}(kn)$ 次計算量，但後面我們將看到概似估計或貝式推論都需要將上述連乘形式展開(此具有 $\mathcal{O}(k^n)$ 次計算量)，因此在大樣本的情況之下是禁止去計算的。儘管計算這些分佈的動差是可行的(如期望值、變異數、動差估計)，但若想對混合形式描述概似函數，一般而言是不太可能的。最後，我們來看一個特別重要的有關於時間序列的例子，此例子中概似函數將無法明確地被寫出來。

例題1.3 移動平均模型(Moving average model)

一個 $MA(q)$ 模型指變數 X_t 有以下表示法：

$$X_t = \varepsilon_t + \sum_{j=1}^q \beta_j \varepsilon_{t-j} \quad (1.4)$$

其中，對於任一 $i = -q, -(q-1), \dots, \varepsilon_i$ 是獨立隨機變數(iid random variables)且 $\varepsilon_i \sim N(0, \sigma^2)$ ，對於任一 $j = 1, 2, \dots, q$ ， β_j 是未知參數。若此樣本包含觀察值 $((X_0, \dots, X_n), n > q)$ ，則此樣本的分佈為

$$\begin{aligned} & \int_{\mathbb{R}^q} \sigma^{-(n+q)} \prod_{i=1}^q \varphi\left(\frac{\varepsilon_{-i}}{\sigma}\right) \varphi\left(\frac{x_0 - \sum_{i=1}^q \beta_i \varepsilon_{-i}}{\sigma}\right) \varphi\left(\frac{x_1 - \beta_1 \hat{\varepsilon}_0 - \sum_{i=2}^q \beta_i \varepsilon_{1-i}}{\sigma}\right) \dots \\ & \times \varphi\left(\frac{x_n - \sum_{i=1}^q \beta_i \hat{\varepsilon}_{n-i}}{\sigma}\right) d\varepsilon_{-1} \dots d\varepsilon_{-q} \end{aligned} \quad (1.5)$$

其中

$$\begin{aligned} \hat{\varepsilon}_0 &= x_0 - \sum_{i=1}^q \beta_i \varepsilon_{-i}, \\ \hat{\varepsilon}_1 &= x_1 - \sum_{i=2}^q \beta_i \varepsilon_{1-i} - \beta_1 \hat{\varepsilon}_0, \\ &\vdots \end{aligned}$$

$$\hat{\varepsilon}_n = x_n - \sum_{i=1}^q \beta_i \hat{\varepsilon}_{n-i}$$

像 $\hat{\varepsilon}'_i s$ 這種疊代式的定義，不僅阻礙了(1.5)式之積分式子得到明確的結果，也妨礙到對此模型做統計推論。此外，對於 $i = -q, -(q-1), \dots, -1$ ，干擾項 ε_{-i} 可被解釋成遺失的資料(missing data, 可參考第5.3.1節)。

在介紹以模擬為基礎的推論之前，在建構模型時遭遇到的計算上的困難常迫使我們使用所謂”標準”模型及”標準”分佈。一種作法是藉由指數族有許多規律的特性(可參考1.6.1節)，因此可以使用以指數族(exponential families)為基礎的模型(見(1.9)式定義)。另一種作法是放棄用參數來表示非參數化但卻是穩健(robust, 相對於模型誤差而言)的方法。我們注意到化簡分佈(因為計算上的限制而必須要這麼做)不一定要排除那些無法寫成明確表示式或任何統計方法的問題。而我們的重點在於，應用模擬為基礎的方法來幫助更多實際的模型找解答及做推論。

1.2 概似估計法(Likelihood Methods)

我們最常關心的統計方法為最大概似(maximum likelihood)方法、貝式(Bayesian)方法及可由這兩種方法所得的推論。執行這些方法通常都與特別的數學計算有關。前者與最大化(maximization)問題有關，此即把估計量隱含的定義視為是最大化問題的解答；後者與積分問題有關，即把明確的估計量表法視為一個積分式。(可參考以下作者的相關書籍 Berger 1985、Casella 及 Berger 2001、Robert 2001、Lehmann 及 Casella 1998)。最大概似估計法是推導出估計量的一種相當廣用的方法。自具有分佈為 $f(x|\theta_1, \dots, \theta_k)$ 的母群體中取出一組 iid 的樣本 $\mathbf{x} = (x_1, \dots, x_n)$ ，則概似函數為

$$L(\theta|\mathbf{x}) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\theta_1, \dots, \theta_k) \quad (1.6)$$

更一般地，當 $X'_i s$ 不是 iid 時，概似函數就定義成聯合機率密度函數(joint density $f(x_1, \dots, x_n|\theta)$)並將此密度函數視為 θ 的函數。 θ 的值，稱為 $\hat{\theta}$ ，就是所謂的最大概似估計量(maximum likelihood estimator, MLE)，此值將使概似函數 $L(\theta|\mathbf{x})$ 在固定 \mathbf{x} 時達到最大值。值得注意的是，由 MLE 的構成可得知其所定義的範圍與參數的範圍一致。最大概似方法的校正主要在於近似，也就是在正常情況下 MLE 幾乎完全收斂(converge almost surely)到參數真正的值。

例題1.4 Gamma MLE

一個最大概似估計量通常是將概似函數取對數形式後計算最大值而得。設 X_1, \dots, X_n 為

取自 gamma 分佈之 iid 的觀察值，gamma 分佈的概似函數為

$$L(\alpha, \beta | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \alpha, \beta) = \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-\frac{x_i}{\beta}}$$

若此處我們假設 α 已知，並將概似函數取對數，

$$\begin{aligned} \log L(\alpha, \beta | x_1, \dots, x_n) &= \log \prod_{i=1}^n f(x_i | \alpha, \beta) = \log \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-\frac{x_i}{\beta}} \\ &= -n \log \Gamma(\alpha) - n \alpha \log \beta + (\alpha - 1) \sum_{i=1}^n \log x_i - \sum_{i=1}^n \frac{x_i}{\beta}. \end{aligned}$$

解 $\frac{\partial}{\partial \beta} \log L(\alpha, \beta | x_1, \dots, x_n) = 0$ 就可得到 β 的 MLE, $\hat{\beta} = \sum_{i=1}^n \frac{x_i}{n\alpha}$ 。

若 α 也是未知的，則我們需要額外解 $\frac{\partial}{\partial \beta} \log L(\alpha, \beta | x_1, \dots, x_n) = 0$ ，此情形在計算上會產生很難處理的方程組，因此不可能得到一個明確的解。

計算最大概似估計量有時可透過對殘差平方和求最小值而得，這就是最小平方法(method of least squares)的基礎。

例題1.5 Least squares estimators

用最小平方法來做估計最早可追溯到兩位學者Legendre (1805) 及Gauss(1810)。以下我們舉在線性迴歸中的一個特別的例子，我們觀察 $(x_i, y_i), i = 1, \dots, n$ ，其中

$$Y_i = b + ax_i + \varepsilon_i, \quad i = 1, \dots, n \quad (1.7)$$

此處 ε_i 表示誤差，未知參數 (a, b) 的估計是由將下式所給的距離

$$\sum_{i=1}^n (y_i - ax_i - b)^2 \quad (1.8)$$

對 (a, b) 找最小值而得的最小平方估計量。作法如下：

令

$$Q = \sum_{i=1}^n (y_i - ax_i - b)^2,$$

則

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^n 2(y_i - ax_i - b)(-x_i), \quad \frac{\partial Q}{\partial b} = \sum_{i=1}^n 2(y_i - ax_i - b)(-1),$$

令

$$\frac{\partial Q}{\partial a} = 0, \frac{\partial Q}{\partial b} = 0,$$

整理可得

$$\sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - b \sum_{i=1}^n x_i = 0, \quad \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - nb = 0,$$

解出

$$\hat{a} = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}, \quad \hat{b} = \bar{Y} - \hat{a} \bar{X}.$$

若我們加入更多誤差項相關的結構，如 $\varepsilon_i \sim N(0, \sigma^2)$ 、獨立(等價於 $Y_i | x_i \sim N(ax_i + b, \sigma^2)$)，則對 (a, b) 而言的對數概似函數(log-likelihood function)會正比於

$$\log(\sigma^{-n}) - \sum_{i=1}^n (y_i - ax_i - b)^2 / 2\sigma^2,$$

將上式分別對 a, b 偏微分，則可得到與最小平方法相同的 \hat{a}, \hat{b} 。在 (1.8) 式中，如果我們假設 $E(\varepsilon_i) = 0$ 或是假設 $E(Y|x) = ax + b$ ，則以計算的觀點來看，最小化 (1.8) 式，等價於給定 x 下假設 Y 具有常態分佈然後引用最大概似估計法來做估計。

在指數族中，也就是分佈的密度函數型式如下

$$f(x) = h(x) e^{\theta \cdot t(x) - \psi(\theta)}, \quad \theta, x \in \mathbb{R}^k \quad (1.9)$$

將指數族的概似函數取對數， $\log L(\theta) = \log f(x) = \theta \cdot x - \psi(\theta)$ ， $\theta, x \in \mathbb{R}^k$ ，令 $\frac{\partial \log L}{\partial \theta} = 0$ ，可得

$$t(x) = \nabla \psi\{\hat{\theta}(x)\} \quad (1.10)$$

則 θ 的 MLE 即由此式解得。此外，因為動差估計法為 $E_\theta[X] = \nabla \psi(\theta)$ ，式子 (1.10) 也可得到動差估計法的估計量。函數 ψ 是 Log-Laplace 轉換或是 h 的累積生成函數 (cumulative generating function)，亦即 $\psi(t) = \log E[e^{t\theta(X)}]$ 的右邊部分可視為 h 的對數動差生成函數 (Log moment generating function)。

例題1.6 Normal MLE

對常態分佈 $N(\mu, \sigma^2)$ 而言其密度函數可寫成 (1.9) 形式，因為常態分佈的密度函數為

$$\begin{aligned} f(y|\mu, \sigma) &\propto \sigma^{-1} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= \sigma^{-1} \exp\left\{\frac{\mu}{\sigma^2} y - \frac{1}{2\sigma^2} y^2 - \frac{\mu^2}{2\sigma^2}\right\} \end{aligned}$$

令

$$\theta_1 = \frac{\mu}{\sigma^2}, \theta_2 = \frac{-1}{2\sigma^2}$$

則上式可改寫成

$$f(y|\theta_1, \theta_2) \propto \exp\{\theta_1 y + \theta_2 y^2 - [-\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)]\}$$

$$\psi(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$$

其中 θ_1, θ_2 稱為自然參數(natural parameters)。若 $Y_1, \dots, Y_n \text{ iid } \sim N(\mu, \sigma^2)$ ，則聯合機率密度函數

$$f(y_1, \dots, y_n|\mu, \sigma) \propto \sigma^{-n} \exp\left\{\frac{\mu}{\sigma^2} \sum_{i=1}^n y_i - \frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 - \frac{\mu^2}{2\sigma^2}\right\}$$

令 $\theta_1 = \frac{\mu}{\sigma^2}, \theta_2 = \frac{-1}{2\sigma^2}$ ，則 $f(y_1, \dots, y_n|\mu, \sigma)$ 可改寫成

$$f(y_1, \dots, y_n|\theta_1, \theta_2) \propto \exp\left\{\theta_1 \sum_{i=1}^n y_i + \theta_2 \sum_{i=1}^n y_i^2 - [-n\frac{\theta_1^2}{4\theta_2} - n\frac{1}{2}\log(-2\theta_2)]\right\}$$

概似函數 $L(\theta_1, \theta_2|y_1, \dots, y_n) = f(y_1, \dots, y_n|\theta_1, \theta_2)$ ，令 $\frac{\partial \log L}{\partial \theta_1} = 0, \frac{\partial \log L}{\partial \theta_2} = 0$ ，可得

$$\begin{aligned} -n\frac{\theta_1}{2\theta_2} &= \sum_{i=1}^n y_i = n\bar{y} \\ n\frac{\theta_1^2}{4\theta_2^2} - \frac{n}{2\theta_2} &= \sum_{i=1}^n y_i^2 = n(s^2 + \bar{y}^2) \end{aligned} \tag{1.11}$$

又

$$\begin{aligned} \because \mu &= \frac{-\theta_1}{2\theta_2} = \bar{y}, \quad \sigma^2 = \frac{-1}{2\theta_2} = s^2, \quad \text{其中 } s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n \\ \therefore \hat{\mu} &= \bar{y}, \quad \hat{\sigma}^2 = s^2. \end{aligned}$$

以上為能順利解出的例子，不過，許多情況下的 ψ 是無法明確地被計算的。即使 ψ 可以被計算，但(1.10)式的解仍無法明確表示，或者當 θ 有限制時，使(1.9)式有最大值的 θ 並不是(1.10)的解。

例題1.7 Beta MLE

Beta 分佈， $\mathcal{B}e(\alpha, \beta)$ ，也是指數族中一個特別的例子，它的密度函數為

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1}, \quad 0 \leq y \leq 1$$

將此密度函數寫成如(1.9)式形式，並令 $\theta = (\alpha, \beta)$ 、 $x = (\log y, \log(1 - y))$ ，代入式子(1.10)後變成

$$\begin{aligned}\log y &= \Psi(\alpha) - \Psi(\alpha, +\beta), \\ \log(1 - y) &= \Psi(\beta) - \Psi(\alpha, +\beta).\end{aligned}\quad (1.12)$$

其中 $\Psi(z) = d \log \Gamma(z) / dz$ 。如同例題1.6，一般而言我們需要 Y_1, \dots, Y_n 來得到參數的估計量，所以(1.12)可改成

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \log y_i &= \Psi(\alpha) - \Psi(\alpha, +\beta), \\ \frac{1}{n} \sum_{i=1}^n \log(1 - y_i) &= \Psi(\beta) - \Psi(\alpha, +\beta)\end{aligned}$$

可惜的是上式無法解出 α, β 的封閉解。

當參數為一向量形式，如 $\theta = (\lambda, \psi)$ ，其中 ψ 是一個無謂參數(nuisance parameter)，一般的做法是計算出 $\hat{\theta} = (\hat{\lambda}, \hat{\psi})$ ，再用 $\hat{\lambda}$ 來估計 λ 。

例題1.8 Noncentrality parameter

設 $X \sim N_p(\theta, I_p)$ ， $\lambda = \|\theta\|^2$ 是我們感興趣的參數， λ 的 MLE 為 $\hat{\lambda}(x) = \|x\|^2$ ，此估計量的偏移量為常數 p (即 $E(\hat{\lambda}) = \lambda + p$)。觀察值 $Y = \|X\|^2$ 具非中央卡方分佈(noncentral chi square distribution)， $\chi_p^2(\lambda)$ ，其中 λ 的 MLE 並非 Y ，因為它來自以下無法解出 λ 的封閉解的方程式

$$\sqrt{\lambda} I_{(p-2)/2}(\sqrt{\lambda}y) = \sqrt{y} I_{p/2}(\sqrt{\lambda}y), \quad y > p \quad (1.13)$$

此處 I_ν 是修正類型的貝索函數(the modified Bessel function)

$$I_\nu(t) = \frac{(t/2)^\nu}{\sqrt{\pi} \Gamma(\nu + 1/2)} \int_0^\pi e^{t \cos \theta} \sin^{2\nu} \theta d\theta = \left(\frac{t}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(t/2)^{2k}}{k! \Gamma(\nu + k + 1)}$$

因為想解(1.13)式需要先算出 $I_{p/2}$ 、 $I_{(p-2)/2}$ 這種特別的函數，可見即使是屬於指數族中的分佈，我們仍無法避免計算上的問題。

指數族以外，除了如均勻分佈或柏拉圖分佈等其樣本空間與 θ 有關，其餘分佈在使用最大概似估計法時，將因為缺少固定維度的充分統計量(sufficient statistic of fixed dimension)而遇到困難。以下我們來看一些非指數族的例子。

例題1.9 Student's t distribution

將隨機擾動透過具常態分佈的誤差來建模，這點時常被評論為太設限了，而另一個合理的看法是看成 t 分佈， $\mathcal{T}(p, \theta, \sigma)$ ，此法比起其他可能的方法來的穩健(robust)許多。t 分佈($\mathcal{T}(p, \theta, \sigma)$)正比於

$$\sigma^{-1} \left(1 + \frac{(x - \theta)^2}{p\sigma^2} \right)^{-(p+1)/2} \quad (1.14)$$

一般而言， p 是已知的而 θ, σ 是未知的。一組來自上述 t 分佈的獨立樣本 X_1, \dots, X_n ，其概似函數正比於

$$\sigma^{-n} \prod_{i=1}^n \left(1 + \frac{(x_i - \theta)^2}{p\sigma^2} \right)^{-(p+1)/2}$$

對此樣本的某些結構而言，當 σ 已知時，此 $2n$ 次方的多項式可能有 n 個局部極小值，必須透過計算才能確定何時有最大值以求得最大概似估計量。圖 1-1 為已知樣本 $X = \{0, 5, 9\}$ 時， $\text{Cauchy}(\theta, 1)$ 的概似函數圖。

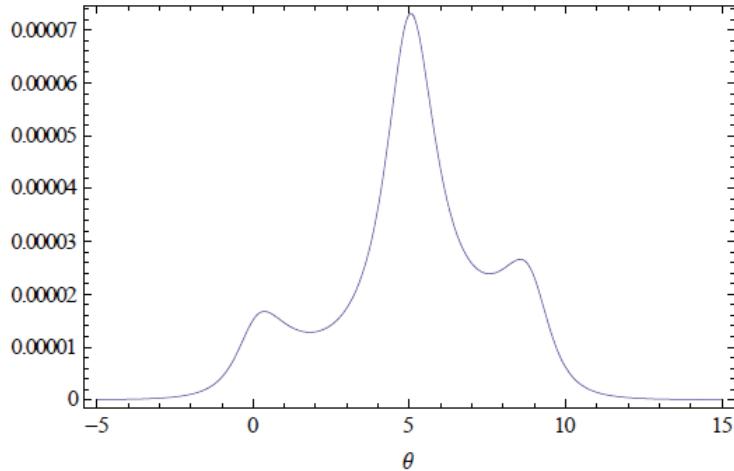


圖 1-1 $\text{Cauchy}(\theta, 1)$ 的概似函數圖

例題1.10 (續例題1.2)

在混合兩個常態分佈的模型裡， $p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\theta, \sigma^2)$ ，一組 iid 樣本 X_1, \dots, X_n 的概似函數正比於

$$\prod_{i=1}^n \left[p\tau^{-1} \varphi \left(\frac{x_i - \mu}{\tau} \right) + (1-p)\sigma^{-1} \varphi \left(\frac{x_i - \theta}{\sigma} \right) \right] \quad (1.15)$$

上式如果展開，會包含 2^n 項將遇到與上一例題相同的有許多局部極值的問題。因為概似函數可能有許多局部極值的問題，以致使用標準的最大化方法來找出絕對極大值的作法常常是失敗的，因此必須設計一些特別的演算法來解決。本例題還另外點出了一個更重

大更困難的問題，因為(1.15)式的展開式中包含

$$p^n \tau^{-n} \prod_{i=1}^n \varphi\left(\frac{x_i - \mu}{\tau}\right) + p^{n-1}(1-p)\tau^{-n+1}\sigma^{-1}\varphi\left(\frac{x_1 - \theta}{\sigma}\right) \prod_{i=2}^n \varphi\left(\frac{x_i - \mu}{\tau}\right) + \dots$$

當 $\theta = x_1$ 且 σ 趨近於 0 時，此式是無界的(unbounded)。

1.3 貝氏方法(Bayesian Method)

前一節我們在最大概似估計法中遇到的問題主要是最佳化問題，本節的貝氏方法則通常是遇到積分問題。在貝氏的例子中由資料 x 提供的訊息，即 $X \sim f(x|\theta)$ ，是結合了先驗分佈(密度函數為 $\pi(\theta)$) 中特定的先驗資訊，並歸納出一個機率分佈 $\pi(\theta|x)$ ，稱此為事後分佈 (posterior distribution)。

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \quad (1.16)$$

其中 $m(x) = \int f(x|\theta)\pi(\theta)d\theta$ 是 X 的邊際機率。

為了估計特別的參數 $h(\theta)$ ，用來做統計推論的決策理論方法 (decision-theoretic approach) 需要一個損失函數 (loss function $L(\delta, \theta)$) 來解釋，損失函數是指用 δ 估計 $h(\theta)$ 時產生的損失，其形式包含二次型損失函數 (quadratic loss function)、絕對值損失函數 (absolute loss function)。用貝氏方法做估計的一個重要的觀點在於最小化貝氏風險 (Bayes risk)：

$$\int \int L(\delta, \theta) f(x|\theta) \pi(\theta) dx d\theta,$$

對上述積分一個直觀的看法是對每個 x 而言，選擇一個能最小化「事後損失」 (posterior loss) 的估計量 δ 。此處「事後損失」是指

$$E[L(\delta, \theta)|x] = \int L(\delta, \theta) \pi(\theta|x) d\theta. \quad (1.17)$$

例如在二次損失 (quadratic loss) 的例子中，

$$L(\delta, \theta) = \|h(\theta) - \delta\|^2$$

其中 $h(\theta)$ 的貝氏估計量就是事後期望值 (posterior mean) $\delta^\pi(x) = E^\pi[h(\theta)|x]$ 。欲計算貝氏估計量 $\delta^\pi(x)$ 通常會遇到兩種困難：

第一、事後分佈 $\pi(\theta|x)$ 的積分式無法透過分析方法完成。

第二、 $\delta^\pi(x)$ 一般而言不見得都能有封閉解。

貝氏方法在計算上的缺點，使得有很長的一段時間裡，貝氏模型中最受喜愛的先驗分佈是那些能明確計算的，稱之為共軛先驗分佈 (conjugate priors)，對應這些共軛先驗分佈的事後分佈與其共軛先驗分佈會屬於相同分佈族(也就是來自相同分佈)。以下舉一個利用共軛先驗分佈做貝氏估計的例題。

例題1.11 二項式貝氏估計量

設一觀察值 X 來自二項式分佈 $\mathcal{B}(n, p)$ ，其共軛先驗分佈為 β 分佈 (Beta distribution)。在二次損失下，欲求得 p 的貝氏估計量我們可以找到使貝氏風險有最小值的 δ ，也就是

$$\min_{\delta} \int_0^1 \sum_{x=1}^n [p - \delta(x)]^2 \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1} (1-p)^{n-x+b-1} dp$$

等價地，由 (1.17) 式的看法我們可以找使事後損失期望值有最小值的 δ ，也就是

$$\min_{\delta} \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(n-x+b)} \int_0^1 [p - \delta(x)]^2 p^{x+a-1} (1-p)^{n-x+b-1} dp,$$

注意上式所寫的 p 的事後分佈 (指給定 x) 就是屬於 $\mathcal{B}e(x+a, n-x+b)$ 分佈。此處貝氏估計量 δ^π 正好就是事後期望值

$$\delta^\pi(x) = \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(n-x+b)} \int_0^1 pp^{x+a-1} (1-p)^{n-x+b-1} dp = \frac{x+a}{a+b+n}.$$

使用二次損失函數來做貝氏估計所得的估計量為事後期望值，這種做法通常簡化了許多計算上的問題；然而，若改成使用絕對損失 (absolute error loss $|p - \delta(x)|$) 或非共軛的先驗分佈，那麼計算上可能就更複雜了。

例題1.12 (續例題1.8)

設 $X \sim N_p(\theta, I_p)$ ， $\lambda = \|\theta\|^2$ 是我們想估計的參數。 θ 的參考先驗分佈 (reference prior) 是 $\pi(\theta) = \|\theta\|^{-(p-1)}$ ，對應的事後分佈是

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \\ &\propto f(x|\theta)\pi(\theta) \\ &\propto \frac{e^{-\|x-\theta\|^2/2}}{\|\theta\|^{p-1}} \end{aligned} \tag{1.18}$$

在二次損失函數的假設下， λ 的貝氏估計量為事後期望值如下

$$\frac{\int_{R^p} \|\theta\|^{3-p} e^{-\|x-\theta\|^2/2} d\theta}{\int_{R^p} \|\theta\|^{1-p} e^{-\|x-\theta\|^2/2} d\theta} \tag{1.19}$$

此積分形式是很難明確計算的。

上述例子中，雖然事後分佈的推導一般是由正比的關係來描述，也就是由貝氏定理(Bayes Theorem)將事後分佈寫成 $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ ，但有時還是有必要知道確切的事後分佈或是 X 的邊際分布的情況。例如在統計模型的貝氏比較之情況，若 $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k$ 是觀察值 X 有可能的模型，這些模型分別具有密度函數 $f_j(\cdot|\theta_j)$ ，若相關於參數 $\theta_1, \theta_2, \dots, \theta_k$ 的先驗分佈為 $\pi_1, \pi_2, \dots, \pi_k$ 及先驗權重為 p_1, p_2, \dots, p_k ，則 X 來自 \mathcal{M}_j 模型的事後機率為

$$\begin{aligned} P(X \in \mathcal{M}_j) &= \frac{P(X = x, X \in \mathcal{M}_j)}{P(X = x)} = \frac{P(X = x, X \in \mathcal{M}_j)}{P(X = x)} \\ &= \frac{P(X \in \mathcal{M}_j)P(X = x|X \in \mathcal{M}_j)}{\sum_{i=1}^k P(X \in \mathcal{M}_i)P(X = x|X \in \mathcal{M}_i)} \\ &= \frac{p_j \int f_j(x|\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_{i=1}^k p_i \int f_i(x|\theta_i)\pi_i(\theta_i)d\theta_i} \end{aligned} \quad (1.20)$$

特別的是，兩種模型 $\mathcal{M}_1, \mathcal{M}_2$ 常透過「貝氏因子」(Bayes factor)來做比較，

$$B^\pi(x) = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2}$$

此比例常數是非常重要的，因此不能像前一個例題所述，省略常數部分直接用正比的方法來做。

例題1.13 邏輯迴歸(Logistic regression)

邏輯模型(logit model)是對二元的反應值(0-1)而言一個好用的迴歸模型，反應變數 Y 在給定解釋變數 $x, x \in \mathcal{R}^n$ 下模型為

$$P(Y = 1) = p = \frac{\exp(\alpha + x\beta)}{1 + \exp(\alpha + x\beta)} \quad (1.21)$$

等價的寫法： $\text{logit}(p) = \log[p/(1-p)] = \alpha + x\beta$

我們來看一個邏輯迴歸的實際例子。1986年一台名為“挑戰人”的太空梭在起飛時爆炸，有七個太空人因此罹難，調查後認為此事件原因可能是來自一個稱為“O-ring”的零件出問題，並且可能與當時周圍環境的溫度過低有關。調查人員合理的懷疑，當溫度下降時，此O-ring零件出錯的機率將增加。

表 1-1 太空梭起飛時的溫度($^{\circ}F$)及當時O-ring零件的狀態(1表示失敗；0表示成功)

flight	14	9	23	10	1	5	13	15	4	3	8
failure	1	1	1	1	0	0	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70
flight	17	2	11	6	7	16	21	19	22	12	20
failure	0	1	1	0	0	0	1	0	0	0	0
Temp.	70	70	70	72	73	75	75	76	76	78	79
											81

合理地，我們可以將上表中數據配適一個邏輯迴歸模型，就像(1.21)式一樣，其中 p 表示O-ring零件失敗的機率， x 代表溫度，對 $\log \alpha$ 給定一個指數型式的先驗分佈，對 β 給定一個flat prior 然後找 α, β 的貝氏估計量。對於此模型，運用蒙地卡羅模擬方法所得的研究結果發現，在太空梭起飛時，當時周圍環境溫度越高則O-ring零件失敗的機率越低；在 $65^{\circ}F$ 時，成功與失敗的機率幾乎各占一半；在 $45^{\circ}F$ 時，失敗的機率幾乎趨近於1。

在貝氏方法中遭遇的問題不僅侷限於積分或標準化常數等計算問題，例如使事後分佈有最大值的信賴域(confidence region、credible region with highest posterior density)的判定，

$$C^\pi(x) = \theta : \pi(\theta|x) \geq k,$$

在預先給定信心水準 γ 下，需要對方程式 $\pi(\theta|x) = k$ 解出 k 值，以滿足以下限制

$$P(\theta \in C^\pi(x)|x) = P(\pi(\theta|x) \geq k|x) = \gamma.$$

我們來看一個與信賴域相關的例子。

例題1.14 貝氏信賴域(Bayes credible regions)

一組來自常態分佈 $\mathcal{N}(\theta, \sigma^2)$ 的獨立樣本 X_1, \dots, X_n ，若 θ 的先驗分佈為 $\mathcal{N}(0, \tau^2)$ ，則 $X = (X_1, \dots, X_n)$ 與 θ 的聯合分佈為

$$\begin{aligned} f(x, \theta) &= f(x|\theta)\pi(\theta) \\ &= (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-(2\sigma^2)^{-1} \sum_{i=1}^n (x_i - \theta)^2\right\} (\sqrt{2\pi}\tau)^{-1} \exp\left\{-(2\tau^2)^{-1}\theta^2\right\} \\ &= (\sqrt{2\pi}\sigma)^{-n} (\sqrt{2\pi}\tau)^{-1} \exp\left\{\frac{n^2\bar{x}^2\tau^4}{2\sigma^2\tau^2(n\tau^2 + \sigma^2)}\right\} \exp\left\{\frac{-(\theta - \frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2})^2}{2\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}}\right\} \end{aligned}$$

且 $X = (X_1, \dots, X_n)$ 的邊際分佈為

$$m(x) = \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-(2\sigma^2)^{-1} \sum_{i=1}^n (x_i - \theta)^2\right\} (\sqrt{2\pi}\tau)^{-1} \exp\left\{-(2\tau^2)^{-1}\theta^2\right\} d\theta \\
&= (\sqrt{2\pi}\sigma)^{-n} (\sqrt{2\pi}\tau)^{-1} \sqrt{2\pi} \frac{\sigma\tau}{\sqrt{n\tau^2 + \sigma^2}} \exp\left\{\frac{n^2\bar{x}^2\tau^4}{2\sigma^2\tau^2(n\tau^2 + \sigma^2)}\right\} \\
&= \int_{-\infty}^{\infty} (\sqrt{2\pi} \frac{\sigma\tau}{\sqrt{n\tau^2 + \sigma^2}})^{-1} \exp\left\{\frac{-(\theta - \frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2})^2}{2\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}}\right\} d\theta \\
&= (\sqrt{2\pi}\sigma)^{-n} (\sqrt{2\pi}\tau)^{-1} \sqrt{2\pi} \frac{\sigma\tau}{\sqrt{n\tau^2 + \sigma^2}} \exp\left\{\frac{n^2\bar{x}^2\tau^4}{2\sigma^2\tau^2(n\tau^2 + \sigma^2)}\right\}
\end{aligned}$$

因此 θ 的事後分佈為

$$\pi(\theta|x) = \frac{1}{\sqrt{2\pi} \frac{\sigma\tau}{\sqrt{n\tau^2 + \sigma^2}}} \exp\left\{\frac{-(\theta - \frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2})^2}{2\frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}}\right\}$$

由此結果可看出 θ 的事後分佈為常態分佈 $\mathcal{N}(\frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2})$ ，且在二次損失函數假設下， θ 的貝氏估計量正好為 $\delta^\pi = \frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2}$ 。若假設 σ^2, τ^2 均已知，則最大事後分佈範圍(highest posterior density region)為

$$\begin{aligned}
C^\pi(x) &= \{\theta; \pi(\theta|x) \geq k\} \\
&= \{\theta : \sqrt{\frac{n\tau^2 + \sigma^2}{2\pi n\tau^2}} \exp[-\frac{n\tau^2 + \sigma^2}{2n\tau^2}(\theta - \delta^\pi)^2] \geq k\}
\end{aligned}$$

等價於

$$\begin{aligned}
\{\theta; (\theta - \delta^\pi(x))^2 \leq k'\} &= \{\theta; -k'' \leq \theta - \delta^\pi(x) \leq k''\} \\
&= \{\theta; \delta^\pi(x) - k'' \leq \theta \leq \delta^\pi(x) + k''\}
\end{aligned}$$

此處 k'' 為一常數。因為 $\delta^\pi(x) = \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x}$ ，所以 θ 的信賴域可利用常態分佈之百分位求得。

像例題1.11(二項式貝氏估計量)的情形，由 p 的事後分佈， $\pi(p|x, a, b)$ ，發現此為 $\text{Be}(x+a, n-x+b)$ 分佈。欲求 p 的 90% 最大事後分佈範圍時，必須先找出兩個極限值 $l(x), u(x)$ 以滿足

$$\int_{l(x)}^{u(x)} \pi(p|x, a, b) dp = 0.9 \quad \text{and} \quad \pi(l(x)|x, a, b) = \pi(u(x)|x, a, b)$$

因為 Beta 分佈不是對稱的，所以此問題是無法用分析方法解決的。

例題1.15 柯西信賴區域(Cauchy confidence regions)

一組來自柯西分佈 $\mathcal{C}(\theta, \sigma)$ 的獨立樣本 X_1, \dots, X_n ， $\pi(\theta, \sigma) = \sigma^{-1}$ 為相關的先驗分佈，

想得到 θ 的事後分佈 $\pi(\theta|x)$ ($X = (X_1, \dots, X_n)$)，必須先對另一個未知參數 σ 做積分。

$$X_i \sim \mathcal{C}(\theta, \sigma), \quad f(x_i|\theta, \sigma) = \frac{1}{\sigma} \frac{1}{\pi(1 + (\frac{x_i - \theta}{\sigma})^2)}$$

X, σ, θ 的聯合分佈為

$$f(x|\sigma, \theta)\pi(\theta, \sigma) = \sigma^{-n}\pi^{-n} \prod_{i=1}^n [1 + (\frac{x_i - \theta}{\sigma})^2]^{-1} \cdot \sigma^{-1}$$

θ 的事後分佈為

$$\pi(\theta|x) \propto \int_0^\infty \sigma^{-n-1} \prod_{i=1}^n [1 + (\frac{x_i - \theta}{\sigma})^2]^{-1} d\sigma$$

因為此積分遇到乘法形式，所以無法明確地被計算出結果。此模型中 θ 的概似信賴區間(likelihood confidence interval)可由縱斷面概似函數(profile likelihood)

$$l^P(\theta|x_1, x_2, \dots, x_n) = \max_\sigma l(\theta, \sigma|x_1, x_2, \dots, x_n)$$

及考慮

$$\{\theta : l^P(\theta|x_1, x_2, \dots, x_n) \geq k\}$$

來獲得。但此處欲求出封閉解也是個困難的問題。

例題1.16 線性校正(Linear calibration)

一般的迴歸模型 $Y = \alpha + \beta x + \varepsilon$ ，我們感興趣的是利用觀測值 x 來估計或預測未知的 Y ；在線性校正模型中，則是想利用觀測值 y 來估計 x 的值。例如：在一個化學實驗中，實驗者可能想將正確卻代價昂貴的測量值 Y 改成較不正確但代價較不昂貴的測量值 x 。此問題可用以下觀測所得的獨立隨機變數來表示

$$Y \sim N_p(\beta, \sigma^2 I_p), Z \sim N_p(x_0 \beta, \sigma^2 I_p), S \sim \sigma^2 \chi_q^2$$

其中 $x_0 \in R$, $\beta \in R^p$, x_0 是此處感興趣的參數。

Kubokawa 及 Robert (1994)針對 (x_0, β, σ) 提出一個指示先驗分佈(reference prior)， $\pi(x_0, \beta, \sigma^2) = (1 + x_0^2)^{-\frac{1}{2}} \sigma^{-p-2}$ ，並得到其聯合事後分佈為

$$\begin{aligned} \pi(x_0, \beta, \sigma^2 | y, z, s) &\propto f(y, z, s | x_0, \beta, \sigma^2) \pi(x_0, \beta, \sigma^2) \\ &= (\sqrt{2\pi}\sigma)^{-p} e^{-(2\sigma^2)^{-1}\|y-\beta\|^2} (\sqrt{2\pi}\sigma)^{-p} e^{-(2\sigma^2)^{-1}\|z-x_0\beta\|^2} \frac{s^{q/2-1}}{\Gamma(q/2)(2\sigma^2)^{q/2}} e^{-\frac{s}{2\sigma^2}} \\ &\times (1 + x_0^2)^{-\frac{1}{2}} \sigma^{-p-2} \end{aligned}$$

$$\propto \sigma^{-(3p+q)-2} \exp\{-(s + \|y - \beta\|^2 + \|z - x_0\beta\|^2)/2\sigma^2\} s^{q/2-1} (1 + x_0^2)^{-\frac{1}{2}} \quad (1.22)$$

將 (1.22) 式整理成

$$\begin{aligned} & \sigma^{-(3p+q)-2} s^{q/2-1} (1 + x_0^2)^{-\frac{1}{2}} \exp\left\{-\frac{s}{2\sigma^2}\right\} \exp\left\{-\frac{\|x_0y - z\|^2}{2(1 + x_0^2)\sigma^2}\right\} \\ & \times \exp\left\{-\frac{\|(1 + x_0^2)\beta - (x_0z + y)\|^2}{2(1 + x_0^2)\sigma^2}\right\} \end{aligned}$$

上式對 β , 積分可得

$$\begin{aligned} \pi(x_0, \sigma^2 | y, z, s) & \propto s^{q/2-1} \sigma^{-(3p+q)-1} \exp\left\{-\frac{s}{2\sigma^2}\right\} \exp\left\{-\frac{\|x_0y - z\|^2}{2(1 + x_0^2)\sigma^2}\right\} \\ & = s^{q/2-1} \sigma^{-(3p+q)-1} \exp\left\{-\frac{[(1 + x_0^2)s + \|x_0y - z\|^2]\sigma^{-2}}{2(1 + x_0^2)}\right\} \end{aligned}$$

上式對 σ^2 積分(可先做變數變換，令 $t = \sigma^{-2}$)，得 x_0 的邊際事後分佈為

$$\begin{aligned} \pi(x_0 | y, z, s) & \propto s^{q/2-1} \left[\frac{1 + x_0^2}{(1 + x_0^2)s + \|x_0y - z\|^2} \right]^{(3p+q)/2} \\ & = s^{q/2-1} \frac{(1 + x_0^2)^{(3p+q)/2}}{\left[(x_0 - \frac{y^t z}{s + \|y\|^2})^2 + \frac{s + \|z\|^2}{s + \|y\|^2} - \frac{(y^t z)^2}{(s + \|y\|^2)^2}\right]^{(3p+q)/2}} \end{aligned}$$

然而，對於 x_0 的貝氏估計量也就是事後期望值之計算，或是信賴域 $\{\pi(x_0 | D) \geq k\}$ 的決定，都是不好處理的問題。

1.4 決定性數值法(Deterministic Numerical Method)

前面我們看過數個範例，包含複雜模型的建構及參數估計，這些例子多數需要特殊技巧來處理，並不是一般分析方法能解決的問題。在開始描述模擬方法前，別忘了對於積分或最佳化問題還存在另一種發展良好的方法，那就是數值法。

1.4.1 最佳化(Optimization)

對於解方程式 $f(x) = 0$ ，一個常用的方法為牛頓演算法(*Newton-Raphson algorithm*)，此法產生一序列 x_n 使其滿足

$$x_{n+1} = x_n - \left(\frac{\partial f}{\partial x} \Big|_{x=x_n} \right)^{-1} f(x_n) \quad (1.24)$$

直到此序列滿足 $|x_{n+1} - x_n| < \epsilon$ ， ϵ 為任意微小常數(在多維度的例子中， $\frac{\partial f}{\partial x}$ 代表著一個矩陣)。關於平滑函數 F 的最佳化問題則是使用這個方法來解方程式 $\nabla F = 0$ ，其中 $\nabla F = 0$ 代表 F 的梯度(gradient of F)。若最佳化問題含有限制 $G(x) = 0$ ，可將 F 取代成 $F(x) - \lambda G(x)$ (Lagrange form)， λ 用來滿足此限制。此關於梯度的方法，

序列 x_n 需滿足

$$x_{n+1} = x_n - (\nabla \nabla^t F)^{-1} \nabla F(x_n), \quad (1.25)$$

$\nabla \nabla^t F = \left[\frac{\partial^2 F}{\partial x_i \partial x_j} \right]$ 表示 F 的二階導數形成的矩陣。

例題1.17 簡單的牛頓-拉福生演算法

(A simple Newton-Raphson Algorithm)

此例題我們來看如何使用牛頓-拉福生演算法來找平方根。若我們感興趣於找 b 的平方根，意即解方程式 $f(x) = x^2 - b = 0$ 引用(1.24)可得下列疊代結果

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})} = x^{(j)} - \frac{x^{(j)2} - b}{2x^{(j)}} = \frac{1}{2} \left(x^{(j)} + \frac{b}{x^{(j)}} \right).$$

若再考慮另一個函數

$$h(x) = [\cos(50x) + \sin(20x)]^2, \quad (1.26)$$

我們來代幾個數值實際操作一遍。關於 $f(x)$ ，取 $b = 2$ ，起始點 x_0 分別用 $0.5, 2, 5$ 代入；關於 $h(x)$ ，起始點 x_0 分別用 $0.25, 0.379, 0.75$ 代入；圖 1-2 上方兩圖為函數 $f(x)$ 及其使用牛頓法疊代後的結果；下方兩圖為函數 $h(x)$ 及其使用牛頓法疊代後的結果。由右上方圖中三條線可看出起始點越接近真實值($\sqrt{2}$)的情況，所需要的疊代次數較少，意即較快逼近到真實值位置。從左下圖可見函數 $h(x)$ 具有多個局部極值，右下圖的三條折線並沒有一起收斂到最大值發生的真實位置($x = 0.379$)，此處可見牛頓演算法有個潛在的問題，疊代所得數列總是往最接近的極值靠近(此極值通常並非真正的最大值或最小值所在位置)，且無法跳脫局部極值所在區域，以致如 $h(x)$ 具多重局部極值的情形，若想使用牛頓法找最大值可能會有問題產生。

文獻上對於牛頓-拉福生法有許多不同的操作手法，此處我們提一下最徒下降法(steepest descent method)，此方法中每次疊代都從單一維度來解決 $F(x_n + td_n)$ 的最佳化問題，其中 $t \in R$ ， d_n 為某既定方向， d_n 的選取通常是使用 ∇F 或使用(1.25)式平滑化的結果如下(Levenberg Marquardt 的版本)

$$[\nabla \nabla^t F(x_n) + \lambda I]^{-1} \nabla F(x_n)$$

其中此平滑化的結果需要同時使得 $\frac{d^2 F}{dt^2}(x_n + td_n) \Big|_{t=0}$ 滿足應有的正負性質。

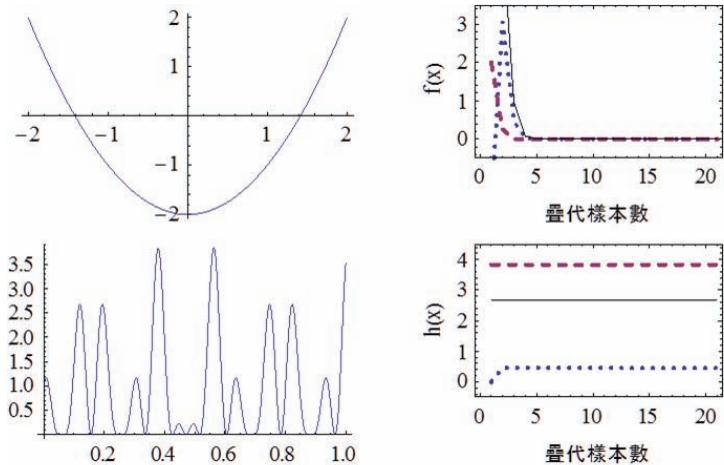


圖 1-2

1.4.2 積分(Integration)

關於積分式 $\mathcal{J} = \int_a^b h(x)dx$ 的數值計算法可使用簡單的黎曼積分法(Riemann integration)或使用更進一步的手法如：梯形法(trapezoidal rule)、辛普森法(Simpson's rule)。

梯形法

$$\hat{\mathcal{J}} = \frac{1}{2} \sum_{i=0}^{n-1} (x_{i+1} - x_i)(h(x_i) + h(x_{i+1}))$$

其中 $x_i, i = 0, 1, \dots, n$ 形成 $[a, b]$ 的一個有序分割(ordered partition)。

辛普森法

$$\bar{\mathcal{J}} = \frac{\delta}{3} \left\{ f(a) + 4 \sum_{i=1}^n h(x_{2i-1}) + 2 \sum_{i=1}^{n-1} h(x_{2i}) + f(b) \right\}$$

此處 $x_i, i = 0, 1, \dots, 2n$ 形成一等距分割，即 $(x_{i+1} - x_i) = \delta$ 。

此外還有其他處理積分問題的方法如：正交多項式(orthogonal polynomial)、弧線(splines)，不過在考慮高維度的情況下這些方法可能無法有太好的表現。

1.4.3 比較(Comparison)

欲將模擬方法及數值分析方法做比較通常是很細微的，因為兩種方法針對不同問題都可提供合適的處理法。此處我們把重點放在各種方法的需求及被引用來處理統計相關問題時所需要的條件。

數值方法

1. 數值積分法只考慮在某特定模型分佈下機率為 0 或機率較低的區域。
2. 數值法常使用高階導數來對逼近產生的誤差提供界限。
3. 處理低維度情況的一般函數通常使用數值法。

模擬方法

1. 模擬方法需要求樣本的隨機性。
2. 模擬方法很少將欲積分或最佳化問題的函數之特殊分析形式列入考慮。
3. 當必須注意統計性質、需要估計數個函數之特性、分佈具多重極值時，模擬方法會較適用。

2 隨機變數的生成(Random Variable Generation)

本章節我們先來考慮模擬所得具一致分佈的隨機變數序列應滿足哪些統計性質，之後再藉由這些具一致分佈的隨機變數，來延伸出其他分佈之隨機變數的生成方法。

2.1 簡介(Introduction)

模擬方法是以生成特定分佈 f 的隨機變數為基礎(f 不一定是明顯已知的某種分佈)，下面將提到偽隨機數生成算子(pseudo-random number generator)的定義，其中隨機變數生成的形式將被正式規範。因為一致分佈 $\mathcal{U}[0, 1]$ 具基本的機率表示，且其他分佈的隨機變數也可能需利用一致分佈的隨機變數來生成，所以我們先來看如何生成 $[0, 1]$ 區間具一致分佈的隨機變數。

2.1.1 一致分佈的模擬(Uniform Simulation)

偽隨機變數的生成有些受限之處，例如由同一演算法生成的兩組隨機變數 (X_1, X_2, \dots, X_n) 、 (Y_1, Y_2, \dots, Y_n) 可能不獨立、不具相同分佈或在機率觀點上是無法做比較的。此外，偽隨機數生成算子的正確性在於看一組樣本 (X_1, X_2, \dots, X_n) 在 $n \rightarrow \infty$ 時的特性，並非看多組樣本 $(X_{11}, X_{12}, \dots, X_{1n})$ 、 $(X_{21}, X_{22}, \dots, X_{2n}) \dots (X_{k1}, X_{k2}, \dots, X_{kn})$ 在 n 固定且 $k \rightarrow \infty$ 時的特性。

定義2.1

一致分佈的偽隨機數生成算子是指一個具起始值 u_0 及遞移算子 D 的演算法，此演算法能生成 $[0, 1]$ 區間內的一組序列 $(u_i) = (D^i(u_0))$ ，而這些數值 (u_1, u_2, \dots, u_n) 要能有一致分佈之獨立隨機變數 (V_1, V_2, \dots, V_n) 應有的行為。

此定義受限於隨機變數之生成是可檢驗的，意即此演算法的正確性包含驗證序列 U_1, U_2, \dots, U_n 是否滿足基本假設 $H_0 : U_1, U_2, \dots, U_n \sim \mathcal{U}[0, 1]$ 。一般的檢驗方法有Kolmogorov-Smirnov 檢定，也可更進一步使用時間序列中的 $ARMA(p, q)$ 模型來檢驗 U_i 及 $(U_{i-1}, \dots, U_{i-k})$ 間的相關性。一個具決定性的系統可以模擬出隨機現象，這個特徵代表著混沌模型(chaotic models)可能可以被用來生成隨機數生成算子。此混沌模型是以動態系統(dynamic system，形式為 $X_{n+1} = D(X_n)$ ，對起始值 X_0 的選取十分敏感)為基礎，並具有複雜的決定性結構。

例題2.2 邏輯函數(The logistic function)

邏輯函數 $D_\alpha(x) = \alpha x(1 - x)$ 在 $\alpha \in [3.57, 4.00]$ 可產生混沌結構(chaotic configuration)，特別是 $\alpha = 4.00$ 時，能使生成的 $[0, 1]$ 區間內的數列 (X_n) ，與來自 arcsine 分

佈之隨機變數具有相同的行爲表現。(arcsine 分佈的密度函數為 $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$) 雖然動態系統 $X_{n+1} = D(X_n)$ 的極限分佈(limit distribution)，也稱平穩分佈(stationary distribution)，有時是被定義的或已知的，但系統的隨機形式並不保證具有相關生成算子的行爲表現。圖 2-1 解釋了以邏輯函數 $D_\alpha(x)$ 為基礎之生成算子的性質，直方圖部分顯示來自成功樣本 $X_{n+1} = D_\alpha(X_n)$ 的轉換變數 $Y_n = 0.5 + \frac{\arcsin(X_n)}{\pi}$ ，可見非常符合一致分佈的形式，散佈圖中數對 (Y_n, Y_{n+100}) 代表的點均勻地填滿了 $[0, 1] \times [0, 1]$ 單位方格，似乎提供了一個良好的隨機逼近，但對於 $(Y_n, Y_{n+1}), (Y_n, Y_{n+10})$ 而言，其分佈形式並沒有符合一致分佈的特徵，且在許多檢定方法下，隨機性的假設都是被拒絕的。這點顯示來自混沌函數定理的典型例子通常無法得到偽隨機數生成算子。

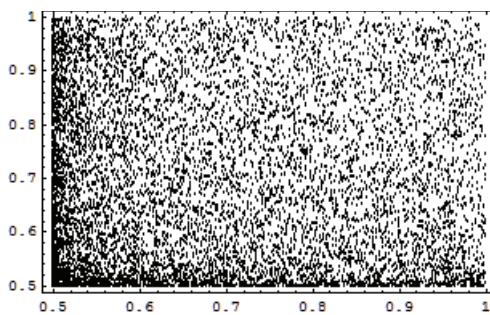


圖 2-1-1 數對 (Y_n, Y_{n+100}) , $n = 1, 2, \dots, 9899$ 所成的散佈圖

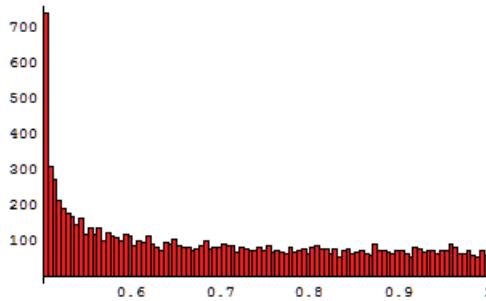


圖 2-1-2 數列 Y_n , $n = 1, 2, \dots, 9899$ 所成的直方圖

2.1.2 逆變換(The Inverse Transform)

在描述隨機變數的狀態空間結構時，常將機率狀態空間 (Ω, \mathcal{F}, P) (Ω 代表全體空間， \mathcal{F} 代表 Ω 上的一個 $\sigma - algebra$ ， P 代表機率測度) 以 $([0, 1], \mathcal{B}, \mathcal{U}[0, 1])$ 來表示 (\mathcal{B} 為 $[0, 1]$ 區間上的 Borel sets)，也就是將 $\omega \in \Omega$ 視為 $[0, 1]$ 區間裡一致分佈的變量。隨機變數 X 表示從 $[0, 1]$ 到 \mathcal{X} 的函數，也就是由廣義的反函數(generalized inverse function) 轉換而得的具一致變量的函數。

定義2.3

實數上一個非遞減的函數 F ，其廣義的反函數 F^- 定義如下

$$F^-(u) = \inf\{x : F(x) \geq u\} \quad (2.1)$$

以下的輔助定理(有時稱為機率積分轉換)，提供我們如何由一致分佈的隨機變數的轉換來表示任意隨機變數的方法。

輔助定理2.4

若 $U \sim \mathcal{U}[0, 1]$ ，則隨機變數 $F^-(U)$ 來自分佈 F 。

證明：

對所有 $u \in [0, 1]$ 及 $x \in F^-[0, 1]$ ，其廣義的反函數滿足

$$F(F^-(u)) \geq u \quad \text{及} \quad F^-(F(x)) \leq x$$

令 $A = \{(u, x) : F^-(u) \leq x\}$ 、 $B = \{(u, x) : F(x) \geq u\}$ ，

若 $(u, x) \in A$ ，則 $F^-(u) \leq x \Rightarrow FF^-(u) \leq F(x)$ ，又已知 $F(F^-(u)) \geq u$ ，

可得 $u \leq F(x) \Rightarrow (u, x) \in B \Rightarrow A \subseteq B$ ；

若 $(u, x) \in B$ ，則 $F(x) \geq u$ ，因為 F 為非遞減函數以致 F^- 亦為非遞減函數，

所以 $F^-(F(x)) \geq F^-(u) \Rightarrow x \geq F^-(u) \Rightarrow (u, x) \in A \Rightarrow B \subseteq A$ 。

由此可得知 $\{(u, x) : F^-(u) \leq x\} = \{(u, x) : F(x) \geq u\}$ ，

且 $P(F^-(U) \leq x) = P(U \leq F(x)) = F(x)$ ，故隨機變數 $F^-(U)$ 來自分佈 F 。

由上述輔助定理可得知，想生成具分佈 F 的隨機變數，可以先生成 $U \sim \mathcal{U}[0, 1]$ ，並將 u 轉換成 $x = F^-(u)$ 就可以得到具分佈 F 的隨機變數。

例題2.5 指數分佈隨機變數的生成(Exponential variable generation)

若 $X \sim \mathcal{E}xp(1)$ ，分佈函數為 $F(x) = 1 - e^{-x}$ ，則從 $u = 1 - e^{-x}$ 可解出 $x = -\log(1 - u)$ 。因此，若 $U \sim \mathcal{U}[0, 1]$ ，則隨機變數 $X = -\log U$ 將具有指數分佈的特性(若 $U \sim \mathcal{U}[0, 1]$ ，則 $1 - U \sim \mathcal{U}[0, 1]$)。

一致分佈隨機變數的生成在其他機率分佈模擬方法中扮演著關鍵的決定性角色，因為這些機率分佈可以表示成一致分佈隨機變數轉換後的形式。然而實際上此方法僅適用於累積分佈函數是明顯可得時(例如：指數分佈，雙重指數分佈(double exponential distribution)，韋伯分佈(Weibull distribution))，並無法涵蓋所有例子。2.3節將介紹“接受-拒絕法”(Accept-Reject method)，此為更一般化且不需利用太強的分佈特性的

方法，因此可處理更多例子及更高維度的問題。

2.1.3 其他例題(Alternatives)

雖然蒙地卡羅法的計算可以被視為是精確的(exact)計算，但通常還是被視為一種逼近法，因此數值逼近法也可視為是另一種層面的蒙地卡羅法，在必要時也可加以使用來處理特別的問題。以下為利用數值法來計算常態機率值的例子。

例題2.6 常態機率值(Normal probabilities)

常態分佈的累積機率函數 $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\{-z^2/2\} dz$ ，雖然 Φ 無法有明顯的表示法，但對於 Φ 及 Φ^- 都存在近似函數。例如 Abramowitz 及 Stegum 在1964年提出的近似函數

$$\Phi(x) \simeq 1 - \varphi(x)[b_1t + b_2t^2 + b_3t^3 + b_4t^4 + b_5t^5] \quad (x > 0)$$

其中 φ 為常態分佈密度函數， $t = (1 + px)^{-1}$ ， $p = 0.2316419$ ， $b_1 = 0.31938$ ， $b_2 = -0.35656$ ， $b_3 = 1.78148$ ， $b_4 = -1.82125$ ， $b_5 = 1.33027$ 。同樣地，也有以下 Φ^- 的近似函數

$$\Phi^-(\alpha) \simeq t - \frac{a_0 + a_1t}{1 + b_1t + b_2t^2}$$

其中 $t^2 = \log(\alpha^{-2})$ ， $a_0 = 2.30753$ ， $a_1 = 0.27061$ ， $b_1 = 0.99229$ ， $b_2 = 0.04481$ 。這兩個近似函數的誤差值為 10^{-8} ，若對 $\mathcal{N}(0, 1)$ 尾端的正確性不嚴格要求且沒有其他更快的模擬方法時，這兩個近似函數應該是可以被使用的。(例題2.8有更快更正確的演算法)。

2.2 一般變換方法(General Transformation Methods)

當一個分佈 f 可以透過一個簡單的方法與一個較容易模擬的函數作連結，這種關係通常可被引用來建構一個演算法，模擬來自 f 的隨機變數。本章節我們要介紹其他技巧來生成非一致分佈的隨機變數，這些方法中有些是較為特別的，因為它們需要依賴原分佈的特性及原分佈與其他分佈的關係，因此可能只適用於某些例子而不好被推廣。以下我們從較容易生成的分佈開始。

例題2.7 指數分佈隨機變數的應用

(Building on exponential random variables)

例題2.5 我們介紹過如何從一致分佈中生成指數分佈隨機變數，此處我們來看一些可由指數分佈生成的隨機變數。若 X_i 為來自 $Exp(1)$ 的獨立隨機變數，也可視為 $X_i \sim Ga(1, 1)$ ，則 $2X_i \sim Ga(1, 2)$ (即 χ_2^2)，且 $\beta X_i \sim Ga(1, \beta)$ 。透過伽瑪分佈的可加性，可得到 $Y = 2 \sum_{j=1}^{\nu} X_j$ 具自由度為 2ν 的卡方分佈($\chi_{2\nu}^2$, $\nu \in N^*$)， $Y = \beta \sum_{j=1}^a X_j$ 具參數為 a, β 的伽瑪分佈($Ga(a, \beta)$, $a \in N^*$)。若 $U = \sum_{j=1}^a X_j \sim Ga(a, 1)$, $V =$

$\sum_{j=a+1}^{a+b} X_j \sim Ga(b, 1)$ ，則 $Y = \frac{U}{U+V}$, $Z = U$ 經由變數變換可得 Y, Z 的聯合分佈為

$$f_{Y,Z}(y, z) = \left(\frac{z^{a-1} e^{-z}}{\Gamma(a)} \right) \left(\frac{[(1/y - 1)z]^{b-1} e^{-[(1/y - 1)z]}}{\Gamma(b)} \right) z/y^2$$

將 $f_{Y,Z}(y, z)$ 整理成

$$f_{Y,Z}(y, z) = \left(\frac{z^{a+b-1} e^{-z/y}}{\Gamma(a+b)y^{a+b}} \right) \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (1-y)^{b-1} y^{a-1} \right)$$

並對 Z 積分，可得 Y 的邊際分佈為 $g_Y(y) = \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (1-y)^{b-1} y^{a-1} \right)$ ，亦即 $Y = \frac{\sum_{j=1}^a X_j}{\sum_{j=1}^{a+b} X_j}$ 為具參數 a, β 的貝塔分佈 ($Be(a, \beta)$, $a, b \in N^*$)。

雖然以上轉換是簡單的，但它們在實際應用上卻受限於可生成之變數所在的範圍及效率性。例如：對伽瑪分佈及貝塔分佈都存在有更有效率的演算法；指數分佈無法生成非整數參數型態的伽瑪分佈的隨機變數，如 χ_1^2 。下面我們繼續介紹”Box-Muller”演算法來生成具標準常態分佈 ($\mathcal{N}(0, 1)$) 的隨機變數。

例題2.8 常態分佈隨機變數的生成(Normal variable generation)

若 X_1, X_2 表示兩個獨立且具標準常態分佈的隨機變數，又 R, Θ 為直角座標 (X_1, X_2) 的極座標(polar coordinates)，則 R, Θ 滿足 $R^2 = X_1^2 + X_2^2$, $\Theta = \tan^{-1} \frac{X_2}{X_1}$ 。令 $R^2 = D$, D, Θ 的聯合分佈為

$$f_{D, \Theta}(d, \theta) = \frac{1}{2\pi} e^{-d/2} \frac{1}{2} = g(\theta)h(d), \quad 0 < d < \infty, 0 < \theta < 2\pi$$

由邊際分佈 $h(d) = \frac{1}{2} e^{-d/2}$ 、 $g(\theta) = \frac{1}{2\pi}$ 得知 $D = R^2$ 與 Θ 獨立，且 $D = R^2 \sim Exp(1/2)$ 、 $\Theta \sim \mathcal{U}_{[0, 2\pi]}$ 。若兩獨立隨機變數 U_1, U_2 均來自 $\mathcal{U}_{[0, 1]}$ ，

令 $X_1 = \sqrt{-2 \log(U_1)} \cos(2\pi U_2)$, $X_2 = \sqrt{-2 \log(U_1)} \sin(2\pi U_2)$ ，則 X_1, X_2 的聯合分佈為

$$f_{X_1, X_2}(x_1, x_2) = 1 \cdot \frac{1}{2\pi} exp\{-(x_1^2 + x_2^2)/2\} = \frac{1}{\sqrt{2\pi}} exp\{-x_1^2/2\} \frac{1}{\sqrt{2\pi}} exp\{-x_2^2/2\}$$

由此可見 X_1, X_2 為具有標準常態分佈的兩獨立隨機變數。

以上討論過程可發現：由極座標表示直角坐標時， $X_1 = R \cos \Theta$, $X_2 = R \sin \Theta$ ，若 $U \sim \mathcal{U}_{[0, 1]}$ ，則 $D = R^2 \sim Exp(1/2)$ 可用 $-2 \log U$ 來代替(例題2.5中介紹過 $-2 \log U \sim Exp(1/2)$)； $\Theta \sim \mathcal{U}_{[0, 2\pi]}$ 可用 $2\pi U$ 來代替，亦即我們可以利用一致分佈來生成標準常態分佈的隨機變數。其相關的演算法如下：

演算法 1 Box-Muller 演算法

1.生成兩獨立隨機變數 $U_1, U_2 \sim \mathcal{U}_{[0,1]}$ ；

2.定義

$$x_1 = \sqrt{-2 \log(u_1)} \cos(2\pi u_2), \quad x_2 = \sqrt{-2 \log(u_1)} \sin(2\pi u_2)$$

3.得到的序列 x_1, x_2 可視為來自標準常態分佈。

相較於一般以中央極限定理為基礎的演算法，此Box-Muller 演算法是精確的，唯一的缺點是必須計算如 \log, \cos, \sin 等函數可能會減慢生成速度。

例題2.9 卜瓦松分佈隨機變數的生成(Poisson generation)

卜瓦松分佈可經由卜瓦松過程與指數分佈作連結，若 $N \sim \mathcal{P}(\lambda)$, $X_i \sim \mathcal{Exp}(\lambda)$, $i \in N^*$ ，則

$$P_\lambda(N = k) = P_\lambda(X_1 + \dots + X_k \leq 1 < X_1 + \dots + X_{k+1})$$

也就是說，卜瓦松分佈可由生成指數分佈隨機變數至其總和超過 1 時所得的隨機變數個數來模擬。此方法是簡單的，但僅於參數 λ 的值較小的時候好用，因為所需要之指數分佈隨機變數的個數與 λ 有關，較大的 λ 值將使此法不適用。此外，當 $Y \sim \mathcal{Ga}(n, (1-p)/p)$, $X|y \sim \mathcal{P}(y)$ ，則 X, Y 的聯合分佈為

$$\begin{aligned} f_{X,Y}(x, y) &= g_Y(y)h(X|Y=y) = \frac{y^{n-1}e^{-py/(1-p)}}{\Gamma(n)[(1-p)/p]^n} \frac{y^x e^{-y}}{x!} \\ &= \frac{y^{n+x-1}}{\Gamma(n)x!} [p/(1-p)]^n e^{-y/(1-p)} \\ &= \frac{y^{n+x-1}e^{-y/(1-p)}}{\Gamma(n+x-1)(1-p)^{n+x}} \frac{\Gamma(n+x)}{\Gamma(n)x!} p^n (1-p)^x \end{aligned}$$

對 Y 積分後可得 X 的邊際分佈， $g_X(x) = \frac{\Gamma(n+x)}{\Gamma(n)x!} p^n (1-p)^x = \binom{n+x-1}{x} p^n (1-p)^x$ ，即 $X \sim \mathcal{Neg}(n, p)$ 。由此我們發現卜瓦松分佈隨機變數的生成算子，可再生成具負二項式分佈(negative binomial distribution)的隨機變數。

例題2.10 一般離散型隨機變數的生成(Discrete random variables)

欲生成隨機變數 $X \sim P_\theta$ ，我們可以先計算出

$$p_0 = P_\theta(X \leq 0), p_1 = P_\theta(X \leq 1), p_2 = P_\theta(X \leq 2), \dots$$

再繼續生成 $U \sim \mathcal{U}_{[0,1]}$ 。若 $p_{k-1} < U < p_k$ ，則取 $X = k$ 。例如：欲生成 $X \sim \mathcal{Bin}(10, 3)$ ，可先取得機率值

$$p_0 = 0.028, p_1 = 0.149, p_2 = 0.382, \dots, p_{10} = 1;$$

或欲生成 $X \sim \mathcal{P}(7)$ ，可先取得機率值

$$p_0 = 0.0009, p_1 = 0.0073, p_2 = 0.0296, \dots$$

以上機率數列將停止於機率值接近 1 時(例如: $p_{20} = 0.999985$)。

例題2.11 貝它分佈的生成(Beta generation)

令 $U_1, U_2, \dots, U_n iid \sim \mathcal{U}_{[0,1]}$ ， $U_{(1)} \leq \dots \leq U_{(n)}$ 表示一組有序樣本， $U_{(i)}$ 的邊際分佈函數為 $f_{U_{(i)}}(u_i) = \frac{n!}{(i-1)!(n-i)!}[F(u_i)]^{i-1}[1-F(u_i)]^{n-i}$ ，可見 $U_{(i)} \sim \mathcal{Be}(i, n-i+1)$ 且 $(U_{(i_1)}, U_{(i_2)} - U_{(i_1)}, \dots, U_{(i_k)} - U_{(i_{k-1})}, 1 - U_{(i_k)}) \sim \mathcal{D}(i_1, i_2 - i_1, \dots, n - i_k + 1)$ (\mathcal{D} 指 Dirichlet 分佈)。雖然以上具一致分佈之樣本的機率性質可以直接被利用來生成貝他分佈或狄利克雷分佈(Dirichlet distribution)的隨機變數，但由於將資料排序可能很耗時間，所以不算是個有效率的演算法；此外，此法僅適用於貝他分佈的參數是整數時。

Jöhnk's Theorem 提供了另一種貝塔分佈隨機變數的生成方法，其內容為

若 $U, V iid \sim \mathcal{U}_{[0,1]}$ ，令 $X = \frac{U^{1/\alpha}}{U^{1/\alpha} + V^{1/\beta}}$ ， $Y = U^{1/\alpha} + V^{1/\beta}$ ，則 X, Y 的聯合分佈為

$$f_{X,Y}(x, y) = \alpha\beta x^{\alpha-1}(1-x)^{\beta-1}y^{\alpha+\beta-1}$$

當 $U^{1/\alpha} + V^{1/\beta} \leq 1$ (即 $y \leq 1$) 時， $\frac{U^{1/\alpha}}{U^{1/\alpha} + V^{1/\beta}} \sim \mathcal{Be}(\alpha, \beta)$ 。但是如圖 2-2 所示，當 $\alpha = \beta$ 的值逐漸變大時，接受數對 (U, V) 的機率會快速下降，因此對 $U^{1/\alpha} + V^{1/\beta}$ 的限制，將使得此演算法在參數 α, β 的值較大時不太好用。

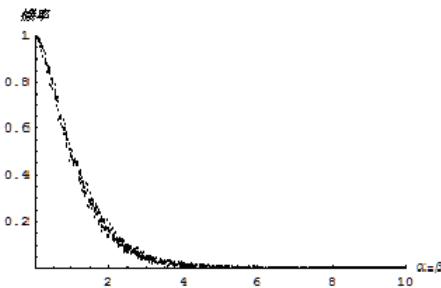


圖 2-2 *Jöhnk* 演算法中，當 $\alpha = \beta$ ，時接受 (U, V) 的機率

例題2.12 伽瑪分佈的生成(Gamma generation)

給定一個貝它分佈隨機變數的生成算子，我們可由以下的方法推導出伽瑪分佈($\mathcal{Ga}(\alpha, 1)$, $\alpha < 1$)隨機變數的生成算子。若 $Y \sim \mathcal{Be}(\alpha, 1 - \alpha)$, $Z \sim \mathcal{Exp}(1)$ ，

令 $X = YZ$, $W = Z$ ，則 X, W 的聯合密度函數為

$$f_{X,W}(x, w) = f_{Y,Z}(x/w, w)|J| = \frac{\Gamma(1)}{\Gamma(\alpha)\Gamma(1-\alpha)}\left(\frac{x}{w}\right)^{\alpha-1}\left(1 - \frac{x}{w}\right)^{-\alpha}e^{-w}\frac{1}{w}$$

X 的邊際密度函數為

$$\begin{aligned} f(x) &= \frac{\Gamma(1)}{\Gamma(\alpha)\Gamma(1-\alpha)} \int_x^\infty \left(\frac{x}{w}\right)^{\alpha-1}\left(1 - \frac{x}{w}\right)^{-\alpha}e^{-w}\frac{1}{w}dw \\ &= \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)}x^\alpha \int_x^\infty (w-x)^{(1-\alpha)-1}e^{-w}dw \\ &= \frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x} \end{aligned} \quad (2.2)$$

可見 $X \sim Ga(\alpha, 1)$ 。

另一種方法是若我們從伽瑪分佈的隨機變數開始，一個更有效率的生成 $Ga(\alpha, 1)$, ($\alpha < 1$) 的方法為若 $Y \sim Ga(\alpha+1, 1)$, $U \sim \mathcal{U}[0, 1]$ 且 Y, U 獨立，令 $X = YU^{1/\alpha}$, $W = Y$ ，則 X, W 的聯合密度函數為

$$f_{X,W}(x, w) = f_{Y,U}(w, (\frac{x}{w})^\alpha)|J| = \frac{\Gamma(1)}{\Gamma(\alpha+1)}w^\alpha e^{-w}(\alpha w^{-\alpha}x^{\alpha-1})$$

X 的邊際密度函數為

$$f(x) \propto \int_x^\infty w^\alpha e^{-w}(\frac{x}{w})^{\alpha-1}w^{-1}dw = x^{\alpha-1}e^{-x} \quad (2.3)$$

可見 $X \sim Ga(\alpha, 1)$ 。上面第(2.2)式可視為混和分佈的一種特殊例子。混和表達式的原則是將密度函數 f 寫成另一分佈的邊際密度函數，形如

$$f(x) = \int_{\mathcal{Y}} g(x, y)dy \quad \text{或} \quad f(x) = \sum_{i \in \mathcal{Y}} p_i f_i(x) \quad (2.4)$$

此表示法取決於 \mathcal{Y} 是連續的或離散的。這種表示法不僅可衍生出有效率的模擬方法，也與往後第九、第十章的方法有關。也就是說，若聯合分佈 $g(x, y)$ 的隨機變數是可以簡單地被模擬出來的，則我們想要的隨機變數 X 可取自從聯合分佈 $g(x, y)$ 中模擬所得的數對 (X, Y) ；另一方面，若第 i 個分佈函數 $f_i(x)$ 可以簡單地被模擬出來，則可先選擇分佈函數 f_i 及其對應的機率 p_i ，再從 f_i 生成我們想要的隨機變數 X 。

例題2.13 t 分佈的生成(Student's t generation)

第(2.4)式的另一個好用的表法為

$$f(x) = \int_{\mathcal{Y}} g(x, y)dy = \int_{\mathcal{Y}} h_1(x|y)h_2(y)dy \quad (2.5)$$

其中 h_1, h_2 分別為 $X|Y = y, Y$ 的條件、邊際密度函數。例如我們可以將具自由度 ν 的 t 分佈之密度函數，寫成與隨機變數 $X|y \sim \mathcal{N}(0, \nu/y)$ 及 $Y \sim \chi_\nu^2$ 相關的表示法。因為 $X|y = \frac{Z}{\sqrt{y/\nu}}$ ，此 $Z \sim \mathcal{N}(0, 1)$ ，所以當已知 $Y \sim \chi_\nu^2$ 時，可得 $\frac{Z}{\sqrt{Y/\nu}} \sim t_\nu$ 。

像第(2.5)式這種表示法在離散分佈情形也是好用的。例如我們在例題 2.9 提過的負二項式分佈 $X \sim Neg(n, p)$ 的生成法，可由隨機變數 $X|y \sim \mathcal{P}(y)$ 及 $Y \sim Ga(n, \beta), \beta = (1-p)/p$ 寫成第(2.5)式的形式

$$\begin{aligned} P(X = x) &= \int_0^\infty \frac{y^x e^{-y}}{x!} \frac{y^{n-1} e^{-y/\beta}}{\Gamma(n)\beta^n} dy \\ &= \frac{\Gamma(n+x)(\beta/(1+\beta))^{n+x}}{x!\Gamma(n)\beta^n} \int_0^\infty \frac{y^{n+x-1} e^{-\frac{1+\beta}{\beta}y}}{\Gamma(n+x)(\beta/(1+\beta))^{n+x}} dy \\ &= \binom{n+x-1}{n-1} p^n (1-p)^x \end{aligned}$$

所得到的 $P(X = x) = \binom{n+x-1}{n-1} p^n (1-p)^x$ 就是 $Neg(n, p)$ 的分佈函數。有一值得注意的地方是離散型分佈的機率混和表達式不一定是離散型表示法，須取決於 \mathcal{Y} 是連續的或離散的隨機變數。

例題2.14 非中心卡方分佈的生成(Noncentral chi squared generation)

由於非中心卡方分佈 $\chi_p^2(\lambda)$ 可視為一些卡方分佈的和(即若 $X_i \sim \mathcal{N}(\theta_i, 1), i = 1, 2, \dots, p$ ，則 $\sum_{i=1}^p X_i^2 \sim \chi_p^2(\lambda)$)，其中 $\sum_{i=1}^p \theta_i^2 = \lambda$)，所以其密度函數也可用混合表達式來表示。依照第(2.5)式的寫法，通常取 $h_1(x|y = k)$ 為卡方分佈 χ_{p+2k}^2 的密度函數(即給定 $Y = k$ 時， $X \sim \chi_{p+2k}^2$)， $h_2(y)$ 為卜瓦松分佈 $\mathcal{P}(\lambda/2)$ 的密度函數(即 $Y \sim \mathcal{P}(\lambda/2)$)，則 $f(x) = \sum_{k=0}^\infty h_1(x|k)h_2(k)$ 即為非中心卡方分佈的密度函數。或者也可以透過 $Z \sim \chi_{p-1}^2, Y \sim \mathcal{N}(\sqrt{\lambda}, 1)$ 組合出 $Z + Y^2 \sim \chi_p^2(\lambda)$ ，此方法會比前述方法有效率。不論是用混合表達式來表示或用修正的貝索函數(modified Bessel function)來描述非中心卡方分佈的密度函數，均無法得到封閉型式的結果。

2.3 接受拒絕法(Accept-Reject Method)

前一節我們介紹過逆變換法，但許多分佈可能很難甚至無法直接由逆變換法來模擬。再者，在某些情況下，我們無法將分佈表示成可用的形式(例如：一種轉換的形式或混合的形式)，如此情況想直接利用機率性質來開創一個模擬方法是不太可能的。因此，我們來看另一種類型的方法，這種方法我們只需要由”乘法常數(multiplicative constant)”來了解感興趣分佈的密度函數 f 具有何種函數形式，並不需要對 f 做深入的分析探討。此方法的關鍵為利用一個較簡單的函數 g 來模擬 f ， g 為工具密度函

數(instrumental density)， f 稱為目標密度函數。上述方法稱為接受拒絕法(Accept-Reject Method)。

2.3.1 模擬的基礎定理(The Fundamental Theorem of Simulation)

以下介紹接受拒絕法的基礎觀念，此觀念在之後第八章建立切片抽樣時也扮演著重要的角色。若 $f_X(x)$ 為感興趣分佈的密度函數，令 $X \sim f_X(x)$, $U | X = x \sim \mathcal{U}(0, f_X(x))$ ，則

$$f_{X,U}(x, u) = f_{U|X}(u | x)f_X(x) = \frac{1}{f_X(x)}f_X(x) = 1, \text{ 其中 } 0 < u < f_X(x)$$

即

$$(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\} \quad (2.6)$$

因此，對任意函數 $f(x)$ 我們可以把 $f(x)$ 表示成

$$f(x) = \int_0^{f(x)} 1 du \quad (2.7)$$

也就是說， f 可視為 (X, U) 的聯合分佈中 X 的邊際密度函數。

第(2.7)式中，具一致性的輔助變數 U 帶來另一種不同的觀點：因為第(2.6)式為 (X, U) 的聯合分佈，且 X 的邊際密度函數正好為目標密度函數 f ，所以我們可以藉由從集合 $\{(x, u) : 0 < u < f(x)\}$ 中生成具一致分佈的隨機變數，取得來自目標密度函數 f 的隨機變數 X 。此方法我們除了需計算函數值 $f(x)$ 外，並不需要使用到函數 f 太多的分析性質。上述重要方法我們統整於以下定理：

定理2.15 模擬的基礎定理(The Fundamental Theorem of Simulation)

欲模擬 $X \sim f(x)$ 等價於模擬 $(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\}$ 。

證明：

(\Rightarrow)

若 $X \sim f_X(x)$ 且 $U | X = x \sim f_{U|X}(u | x)$ ，則 $f_{X,U}(x, u) = f_{U|X}(u | x)f_X(x) = \frac{1}{f_X(x)}f_X(x) = 1$ 其中 $0 < u < f_X(x)$, $x \in D(f)$ ，故 $(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f_X(x)\}$ 。

(\Leftarrow)

若 $(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f_X(x)\}$ ，即 $f_{X,U}(x, u) = 1$, $0 < u < f_X(x)$, $x \in D(f)$ ，則 $\int_0^{f_X(x)} f_{X,U}(x, u) du = \int_0^{f_X(x)} 1 du = f_X(x)$ ，故 $X \sim f_X(x)$ 。

然而，此定理在許多層面都是很基本的，但因為這組具一致分佈的隨機變

數 (X, U) 的模擬通常不那麼直接，所以本定理在此階段多數以一個正規化的表示法來呈現。例如，我們可以模擬 $X \sim f(x)$ 及 $U|X = x \sim \mathcal{U}(0, f(x))$ ，但這種做法將導致整個表示是無用的；或者是採用對稱於前面敘述的方法，從 U 的邊際分佈模擬出 U ，然後再從給定 $U = u$ 的條件分佈中模擬出 X ，不過這種做法最後並無法得到可計算的結果。對此問題的解答是，在一個較大的集合中，一次模擬出整個數對 (X, U) ，再從這些模擬出的數對中取出滿足限制($\{(x, u) : 0 < u < f(x)\}$)的部分。舉例來說，維度是一維的情況下，假設 $\int_a^b f(x) dx = 1$ 且 f 的上界為 m ，令 $Y \sim \mathcal{U}(a, b)$ 及 $U|Y = y \sim \mathcal{U}(0, m)$ ，因為

$$\begin{aligned} P(X \leq x) &= P(Y \leq x | U < f(Y)) = \frac{P(Y \leq x, U < f(Y))}{P(U < f(Y))} \\ &= \frac{\int_a^x \int_0^{f(y)} \frac{1}{m} \cdot \frac{1}{b-a} dudy}{\int_a^b \int_0^{f(y)} \frac{1}{m} \cdot \frac{1}{b-a} dudy} = \int_a^x f(y) dy \end{aligned} \quad (2.8)$$

以上結果顯示，若 $0 < u < f(y)$ ，則數對 (Y, U) 中的 Y 的確具有正確的分佈 f (f 為原定的目標函數)。因此，我們可以藉由模擬 $Y \sim \mathcal{U}(a, b)$ 及 $U|Y = y \sim \mathcal{U}(0, m)$ ，並取出滿足 $0 < u < f(y)$ 的部分數對 (Y, U) ，其中的 Y 即為來自目標函數 f 的隨機變數。此外，以上做法也說明了若 $A \subset B$ ，且我們從 B 集合中生成一致分佈的樣本，之後僅保留落在 A 集合中的部分，則這些被保留的樣本即為 A 集合中具一致分佈的樣本。

例題2.16 貝它分佈的模擬(Beta Simulation)

在例題2.11我們看到要直接模擬貝它分佈的隨機變數是困難的，然而當參數 $\alpha \geq 1, \beta \geq 1$ 時，我們可以使用定理2.15簡單的模擬出貝它分佈。為了生成 $X \sim Be(\alpha, \beta)$ ，我們取 $Y \sim \mathcal{U}[0, 1]$ 及 $U \sim \mathcal{U}[0, m]$ ，其中 m 為貝它分佈的密度函數之最大值($m \approx 2.67$)。圖2-3顯示的是，當 $\alpha = 2.7, \beta = 6.3$ 時，模擬所得的1000組數對 (Y, U) 。圖中落在密度函數曲線下的數對就是被接受的 $X = Y$ 的情況，至於落在密度函數曲線外的數對即為被拒絕的部分。此外，在 $[0, 1] \times [0, m]$ 之內做模擬的接受機率為

$$P(Accept) = P(U < f(Y)) = \int_0^1 \int_0^{f(y)} \frac{1}{m} dudy = \frac{1}{m}$$

以例題2.16為例，若 $m = 2.67$ ，則接受機率約為 $1/2.67 = 37\%$ 。關於第(2.8)式相關的討論，只要在較大集合上做一致分佈的模擬是可行的，則此模擬便可以簡單地推廣成較大集合不是盒狀的情況。然而此推廣可引用於函數 f 的支集或其最大值是無界限的例子。

若較大集合的形式為

$$\mathcal{L} = \{(y, u) : 0 < u < m(y)\}$$

其限制式為 $m(x) \geq f(x)$ ，且希望在 \mathcal{L} 上做一致分佈的模擬是可行的。明顯地，

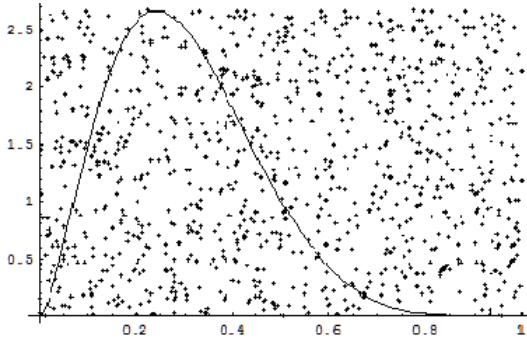


圖 2-3 貝它分佈隨機變數的生成。圓點為使用定理 2.15 模擬所得 1000 筆 (Y, U) ，曲線下的圓點為被接受的 (Y, U) ，約有 368 個。

為了使模擬是有效率的， m 與 f 應該盡可能愈接近愈好，以避免模擬上的浪費。另一個需要注意的重點是，限制式 $m(x) \geq f(x)$ 中的 $m(x)$ ，可能不是一個機率密度函數。因為 $m(x)$ 必須是可積分的(若 $m(x)$ 是不可積分的，則 \mathcal{L} 就沒有有限的範圍，且無法在 \mathcal{L} 上做一致分佈的模擬)，因此我們將 $m(x)$ 改寫成 $m(x) = Mg(x)$ ，其中 $\int_{\mathcal{X}} m(x) dx = \int_{\mathcal{X}} Mg(x) dx = M$ ， $g(x)$ 為一個機率密度函數。因此，我們將模擬基本定理更一般化的應用整理如下：

系理 2.17

令 $X \sim f(x)$ 及 $g(x)$ 為一個密度函數並滿足 $f(x) \leq Mg(x)$ ，其中 $M \geq 1$ 。欲模擬 $X \sim f$ ，可生成 $Y \sim g$ 及 $U|Y=y \sim \mathcal{U}_{(0, Mg(y))}$ ，直到滿足 $0 < u < f(y)$ ，則接受此 y 值。

證明：

對任意可測集合 \mathcal{A} ，由於

$$\begin{aligned} P(X \in \mathcal{A}) &= P(Y \in \mathcal{A}|U < f(Y)) = \frac{P(Y \in \mathcal{A}, U < f(Y))}{P(U < f(Y))} \\ &= \frac{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int_{\mathcal{X}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} = \int_{\mathcal{A}} f(y) dy \end{aligned}$$

可見如果我們先模擬 $Y \sim g$ ，再模擬出 $U|Y=y \sim \mathcal{U}_{(0, Mg(y))}$ ，最後只接受滿足 $u < f(y)$ 的那些 y ，並令 $x = y$ ，則得到的 X 確實是來自目標函數 f 的隨機變數。

圖 2-4 利用 $f(x) \propto \exp(-x^2/2)(\sin^2 6x + 3\cos^2 x \sin^2 4x + 1)$ ，及 $g(x) = \exp(-x^2/2)/\sqrt{2\pi}$ (常態分佈密度函數)，來說明系理 2.17。

由系理 2.17 我們可得兩個結論。

第一、它提供了模擬任一密度函數 f 的一般做法，此密度函數 f 取決於某個乘法因子(multiplicative factor)，也就是說 f 的標準化常數 M (normalizing constant) 是不需

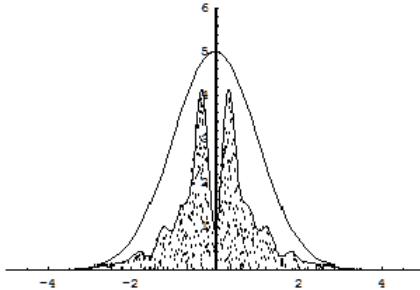


圖 2-4 此圖為來自集合 $\{(x, u) : 0 < u < f(x)\}$ 的一致分佈樣本。其中目標函數為 $f(x) \propto \exp(-x^2/2)(\sin^2 6x + 3\cos^2 x \sin^2 4x + 1)$ ，上界函數 $m(x) = 5\exp(-x^2/2)$ 。

要知道的，因為此方法只需輸入與標準化常數無關的比值 f/M (如圖2.4所舉例子)。此性質在貝氏計算中特別重要。貝氏方法中一個被感興趣的量為事後分佈

$$\pi(\theta|x) \propto \pi(\theta)f(x|\theta) \quad (2.9)$$

由此可見上述與標準化常數無關的性質是適用於貝氏事後分佈的。

第二、接受機率正好為 $1/M$ (幾何分佈中的等待時間)；直到接受第一個隨機變數之前所需的期望試驗次數為 M 。因此，對工具密度函數 g_1, g_2, \dots 的選擇，可藉由比較各別上界 M_1, M_2, \dots 的大小來決定；而最小的上界 M_i 所對應的函數 g_i 則為工具密度函數 g 的最佳選擇。

2.3.2 接受拒絕演算法(The Accept-Reject Algorithm)

系理 2.17 的應用即為所謂的接受拒絕法，我們將它稍做修改，得到下面等價形式的接受拒絕演算法。

演算法 2 接受拒絕演算法

1. 生成 $Y \sim g, U \sim \mathcal{U}[0, 1]$ ；
2. 若 $U \leq f(Y)/Mg(Y)$ ，則接受 $X = Y$ ；
3. 若 $U > f(Y)/Mg(Y)$ ，則回到 1. 重來一次。

證明：

設目標函數為 $f(x), x \in \mathcal{X}$ ，則

$$\begin{aligned} P(X \in \mathcal{A}) &= P(Y \in \mathcal{A} \mid U \leq f(Y)/Mg(Y)) = \frac{P(Y \in \mathcal{A}, U \leq f(Y)/Mg(Y))}{P(U \leq f(Y)/Mg(Y))} \\ &= \frac{\int_{\mathcal{A}} \int_0^{f(y)/Mg(y)} 1 \cdot g(y) dy dy}{\int_{\mathcal{X}} \int_0^{f(y)/Mg(y)} 1 \cdot g(y) dy dy} = \int_{\mathcal{A}} f(y) dy \end{aligned}$$

可見滿足條件 $U \leq f(Y)/Mg(Y)$ 之下的 Y ，的確能使 X 是來自目標函數 f 的隨機變

數。

在 f 、 g 均已標準化(即 f 、 g 均為機率密度函數)的例子中， M 必須是大於 1 的常數。 M 的大小(等同於上述演算法的效率)成為一個描述 g 能多麼近似地模擬 f 的函數，特別是在分佈的尾端之處。由於 f/g 必須有界，所以 g 必須比 f 厚尾。舉例來說，在接受拒絕演算法中我們不能用常態分佈的密度函數 g 來模擬科西分佈的密度函數 f 。此外，我們可以在參數族中尋找工具密度函數 g ，然後決定哪個參數會使上界 M 最小，以得到較佳的接受拒絕演算法。類似前述比較方法，更細微的是兩個參數族間的比較，我們必須考慮由 g 生成一個隨機變數所需要的計算時間。

例題2.18 由雙指數分佈生成常態隨機變數

(Normals from double exponential)

在接受拒絕演算法中，若目標函數 $f(x)$ 為標準常態分佈($\mathcal{N}(0, 1)$)，工具密度函數 $g(x | \alpha) = \frac{\alpha}{2} e^{-\alpha|x|}$, $x \in R$ 為雙指數分佈($\mathcal{L}(\alpha)$)，則 $f(x)/g(x) = \frac{\sqrt{2}}{\sqrt{\pi}\alpha} \exp\{-x^2/2 + \alpha|x|\}$ ，將此兩函數之比值對 x 一次微分並令微分結果為零，可解出當 $x = \alpha$ 時， $\text{Max}(f(x)/g(x)) = \frac{\sqrt{2}}{\sqrt{\pi}\alpha} e^{\alpha^2/2}$ ，而若將此結果對 α 一次微分並令微分結果為零，便可找到使得此上界最小的參數為 $\alpha = 1$ ，此時所得的最小上界為 $\sqrt{\frac{2e}{\pi}}$ 。這表示此演算法的接受機率為 $\sqrt{\frac{\pi}{2e}} \approx 0.76$ ，且平均而言需要 $1/0.76 \approx 1.3$ 個一致分佈的隨機變數才能得到一個具常態分佈的隨機變數。

例題2.19 伽瑪分佈隨機變數的生成

前面例題 2.7 提過，當伽瑪分佈 $Ga(\alpha, \beta)$ 的參數 $\alpha \in \mathcal{N}$ 時，伽瑪分佈可視為 α 個指數分佈 $\epsilon_i \sim \mathcal{E}(\beta)$ 的和，並藉由 $\epsilon_i = -\log(U_i)/\beta$, $U_i \sim \mathcal{U}[0, 1]$ 可以簡單的模擬出伽瑪分佈的隨機變數。但在一般 α 非自然數的情況下，這種表示法便不能成立。接受拒絕演算法可以解決上述一般化的問題。對於伽瑪分佈 $Ga(\alpha, 1)$ 的模擬(不失一般性，我們假設 $\beta = 1$)，可以使用工具密度函數為 $Ga(a, b)$ 的密度函數，取 $a = \lfloor \alpha \rfloor$, $\alpha \geq 1$ ，則 $f(x)/g(x) \propto b^{-a} x^{\alpha-a} \exp\{-(1-b)x\}$ ，由一次微分可得，當 $x = \frac{\alpha-a}{1-b}$ 時， $\text{Max}(f(x)/g(x)) \propto b^{-a} \left(\frac{\alpha-a}{(1-b)e}\right)^{\alpha-a}$ ，即 $M = b^{-a} \left(\frac{\alpha-a}{(1-b)e}\right)^{\alpha-a}$ 且 $b < 1$ ，又當 $b = \frac{a}{\alpha}$ 時，上界 M 有最小值為 $\frac{\alpha^\alpha}{a^a}$ ，可見參數 b 的最佳選擇為 $b = \frac{a}{\alpha}$ ，且此時 $Ga(a, b)$ 與 $Ga(\alpha, 1)$ 具有相同的期望值。

例題2.20 截斷常態分佈(Truncated normal distributions)的生成

在 $X \geq \underline{\mu}$ 的限制下，截斷常態分佈的密度函數

$$f(x) = \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x-\mu)^2/2\sigma^2\} \right) / \int_{\underline{\mu}}^{\infty} f(x) dx$$

$$\propto \exp\{-(x - \mu)^2/2\sigma^2\} I_{\{x \geq \underline{\mu}\}}$$

不失一般性，我們假設 $\mu = 0, \sigma = 1$ ，一個可用的工具密度函數為平移的指數分佈 $\mathcal{E}xp(\alpha, \underline{\mu})$ 的密度函數，

$$\begin{aligned} g_\alpha(z) &= (\alpha e^{-\alpha z}) / \int_{\underline{\mu}}^{\infty} g(z) dz \\ &= \alpha e^{-\alpha(z-\underline{\mu})} I_{\{z \geq \underline{\mu}\}} \end{aligned}$$

$\frac{f}{g_\alpha}(z) \propto \frac{1}{\alpha} e^{-z^2/2} e^{\alpha(z-\underline{\mu})}$ ，由一次微分可得知，當 $z = \alpha$ 時， $\frac{f}{g_\alpha}(z)$ 的上界分別為

$$\begin{cases} \frac{1}{\alpha} \exp\{\alpha^2/2 - \alpha \underline{\mu}\} & \text{當 } \alpha > \underline{\mu} \\ \frac{1}{\alpha} \exp\{-\underline{\mu}^2/2\} & \text{當 } \alpha \leq \underline{\mu} \end{cases}$$

在 $\alpha > \underline{\mu}$ 的情況下，若取 $\alpha^* = (\underline{\mu} + \sqrt{\underline{\mu}^2 + 4})/2$ ，則上述第一種上界有最小值產生；在 $\alpha \leq \underline{\mu}$ 的情況下，若取 $\tilde{\alpha} = \underline{\mu}$ ，則上述第二種上界有最小值產生。此兩種特殊的參數值將使接受拒絕演算法更加有效率。

2.4 包絡接受拒絕法(Envelope Accept-Reject Method)

2.4.1 夾擠原理(The Squeeze Principle)

在許多情況裡，可能因為密度函數 f 本身的複雜性以致於想模擬與 f 相關的分佈是很耗時間的。以例題 1.9 為例， t 分佈($\mathcal{T}(p, \theta, \sigma)$)，密度函數 $f \propto \frac{1}{\sigma} (1 + \frac{(x-\theta)^2}{p\sigma^2})^{-(1+p)/2}$)以貝氏方法來看 θ 的事後分佈 $\pi(\theta | x, \sigma) \propto \prod_{i=1}^n \left[1 + \frac{(x_i-\theta)^2}{p\sigma^2}\right]^{-(p+1)/2}$ ，其中 σ 為已知，此事後分佈 $\pi(\theta | x)$ 的計算包含 n 項乘積，因此在計算上相當費時。欲加速上述問題的模擬速度，可利用包絡演算法(envelope algorithm)，此方法可視為是系理 2.17 的延伸，除了需要目標函數 f 的一個上界函數 g_m 之外，還需要再估計 f 的一個較簡單的下界函數 g_l 。我們可利用 f 的泰勒展開式來協助尋找 f 上下界函數 Mg_m 及 g_l 。

引理 2.21

若存在一密度函數 g_m ，一個函數 g_l ，及常數 M ，並滿足

$$g_l(x) \leq f(x) \leq Mg_m(x)$$

則下列演算法可以產生來自目標函數 f 的隨機變數。

演算法 3 包絡接受拒絕法(Envelope Accept-Reject Method)、夾擠原理(Squeeze Principle by Marsaglia (1977))

1. 生成 $X \sim g_m(x)$, $U \sim \mathcal{U}[0, 1]$ ；

2. 若 $U \leq g_l(X)/Mg_m(X)$, 則接受 X ;
3. 否則, 若 $U \leq f(X)/Mg_m(X)$, 則接受 X 。

由 f 的下界函數 g_l , 我們可推導出此模擬的接受機率爲

$$\begin{aligned} P(\text{Accept}) &= P(U \leq g_l(X)/Mg_m(X)) \\ &= \int \int_0^{g_l(x)/Mg_m(x)} g_m(x) \cdot 1 du dx = \frac{1}{M} \int g_l(x) dx \end{aligned}$$

而模擬所需要的期望次數則爲 $M / \int g_l(x) dx$ 。以上可看出無論是模擬的接受機率或模擬所需要的期望次數都與 f 本身無關, 僅與下界函數 g_l 有關。因此, 當 f 形式複雜時, 此方法可以避免直接計算 f , 如此則可以縮短模擬所需時間。

例題2.22 常態分佈的下界(Lower bound for normal generation)

已知 $e^{-x^2/2}$ 的泰勒展開式爲

$$e^{-x^2/2} = \sum_{n=0}^{\infty} \frac{1}{n!} (-x^2/2)^n \geq 1 - x^2/2,$$

標準常態分佈的密度函數 $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, 則 $f(x) \geq \frac{1}{\sqrt{2\pi}} (1 - x^2/2)$, 可見在模擬標準常態分佈時, 可取下界函數 $g_l(x) = \frac{1}{\sqrt{2\pi}} (1 - x^2/2)$ 。

例題2.23 由邏輯變數取得卜瓦松變數

(Poisson variables from logistic variables)

例題2.9介紹過可利用卜瓦松過程及指數分佈來生成卜瓦松分佈的隨機變數, 但此作法是比較沒有效率的。此處我們介紹由 Atkinson (1979) 所提出的較簡單的作法, 他透過卜瓦松分佈($\mathcal{P}(\lambda)$) 與邏輯分佈之間的相關性來解決此問題。已知邏輯分佈的密度函數及分佈函數分別爲

$$f(x) = \frac{1}{\beta} \frac{\exp\{-(x-\alpha)/\beta\}}{[1 + \exp\{-(x-\alpha)/\beta\}]^2}, \quad F(x) = \frac{1}{1 + \exp\{-(x-\alpha)/\beta\}}$$

且分佈函數 F 的反函數爲 $F^{-1}(x) = \alpha - \beta \log\left(\frac{1-u}{u}\right)$ 。爲了將離散型分佈與連續型分佈做更好的連結, 我們考慮 $N = \lfloor x + 0.5 \rfloor$ 。又因爲邏輯分佈隨機變數的範圍是 $(-\infty, \infty)$, 為了使其範圍能更符合卜瓦松分佈的範圍 $(0, \infty)$, 我們將邏輯分佈隨機變數的範圍限制在 $[-1/2, \infty)$, 則隨機變數 N 的分佈函數如下

$$\begin{aligned} P(N = n) &= P(-1/2 \leq X < n + 1/2) \\ &= [P(X < n + 1/2) - P(X < n - 1/2)] / P(-1/2 < X < \infty) \end{aligned}$$

$$= \left[\frac{1}{1 + \exp\{-(n + 0.5 - \alpha)/\beta\}} - \frac{1}{1 + \exp\{-(n - 0.5 - \alpha)/\beta\}} \right] \\ \times \left(\frac{1 + \exp\{(-0.5 - \alpha)/\beta\}}{\exp\{(-0.5 - \alpha)/\beta\}} \right)$$

卜瓦松隨機變數及隨機變數 N 的分佈函數之比值為

$$\frac{\lambda^n e^{-\lambda}}{n!} / P(N = n) \quad (2.10)$$

此比值並不好找上下界，Atkinson 提出若取 $\alpha = \lambda$, $\beta = \pi/\sqrt{3\lambda}$ ，將使此比值有最佳化結果。在分析上無法對此比值的上下界找到最佳化結果，不過可使用數值方法來得到上界 $c = 0.767 - 3.36/\lambda$ 。此方法的演算法如下：

演算法 4 Atkinson 之卜瓦松分佈模擬 (Atkinson's Poisson Simulation)

0. 定義 $\beta = \pi/\sqrt{3\lambda}$, $\alpha = \lambda\beta$, $k = \log c - \lambda - \log \beta$ ；
1. 生成 $U_1 \sim \mathcal{U}[0, 1]$, 並計算 $X = \{\alpha - \log(\frac{1-u_1}{u_1})\}/\beta$, 直到 $X > -0.5$ 為止；
2. 定義 $N = \lfloor X + 0.5 \rfloor$, 並生成 $U_2 \sim \mathcal{U}[0, 1]$ ；
3. 若 $\alpha - \beta X + \log(u_2/\{(1 + \exp(\alpha - \beta X))\}^2) \leq k + N \log \lambda - \log N!$, 則接受 $N \sim \mathcal{P}(\lambda)$ 。

雖然此演算法所得結果算是精確的，但在參數 α, β 的選擇，及上下界函數和分佈函數比值之計算上，它使用了一些“近似”的手法，並非真正透過分析方法得到上述結果。此外，此演算法還需計算 $N!$ ，這將花費許多計算時間。Devroye (1985) 所提出的演算法，雖然較為複雜但可能更適用。

2.4.2 對數凹密度函數(Log-Concave Densities)

對數凹密度函數(指密度函數取對數後呈現凹函數形式)的特殊例子可以建構有效率的演算法。以下我們先介紹一些對數凹密度函數的例子。

例題2.24 對數凹密度函數(Log-concave densities)

回憶前面第一章曾提過的指數族，

$$f(x) = h(x)e^{\theta x - \psi(\theta)}, \quad \theta, x \in R^k$$

對指數族而言，若密度函數 f 滿足

$$\begin{aligned}\frac{\partial^2}{\partial x^2} \log f(x) &= \frac{\partial^2}{\partial x^2} (\log h(x) + \theta x - \psi(\theta)) = \frac{\partial^2}{\partial x^2} \log h(x) \\ &= \frac{h(x)h''(x) - [h'(x)]^2}{h^2(x)} < 0\end{aligned}$$

則此密度函數即為對數凹密度函數。舉例來說，若 $X \sim \mathcal{N}(\theta, 1)$ ，則 $f(x) = \frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}e^{\theta x - \theta^2/2}$ ，其中 $h(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ ，且 $\partial^2 \log h(x)/\partial x^2 = -1$ ，可見常態分佈之密度函數的確是對數凹密度函數。

接下來我們來看看由 Gilks 及 Wild 在 1992 年所提出的演算法，此演算法是以包絡及相關的接受拒絕演算法為基礎建構而成，又稱為適應拒絕抽樣(adaptive rejection sampling (ARS))。當函數 $h = \log f$ 是凹函數時，此方法為密度函數 f 提供一序列的上下界限。

令集合 $S_n = \{x_0, x_1, \dots, x_{n+1}\}$ ，其中 $x_i, i = 0, 1, 2, \dots, n+1$ 為 f 的支集(support)中滿足 $h(x_i) = \log f(x_i)$ 的已知點。令 $L_{i,i+1}$ 為通過 $(x_i, h(x_i)), (x_{i+1}, h(x_{i+1}))$ 的直線，由於 h 為凹函數， $L_{i,i+1}$ 在 $[x_i, x_{i+1}]$ 範圍內會低於 h 的圖形，在 $[x_i, x_{i+1}]$ 範圍外會高於 h 的圖形(見圖2-5)。對任一 $x_i \in S_n$ ，定義

$$\begin{aligned}x \in [x_0, x_1], \bar{h}_n(x) &= L_{1,2}(x) \text{ 且 } \underline{h}_n(x) = L_{0,1}(x); \\ x \in [x_i, x_{i+1}], i = 1, \dots, n-1, \bar{h}_n(x) &= \min\{L_{i-1,i}(x), L_{i+1,i+2}(x)\} \text{ 且} \\ \underline{h}_n(x) &= L_{i,i+1}(x); \\ x \in [x_n, x_{n+1}], \bar{h}_n(x) &= L_{n-1,n}(x) \text{ 且 } \underline{h}_n(x) = L_{n,n+1}(x)\end{aligned}$$

若 $x \in [x_0, x_{n+1}]^c$ ，則定義

$$\bar{h}_n(x) = \min\{L_{0,1}(x), L_{n,n+1}(x)\} \quad \text{and} \quad \underline{h}_n(x) = -\infty$$

在函數 f 的支集上， h 的上下界限為

$$\underline{h}_n(x) \leq h(x) \leq \bar{h}_n(x) \tag{2.11}$$

因 $\bar{f}_n(x) = \exp \bar{h}_n(x)$ and $\underline{f}_n(x) = \exp \underline{h}_n(x)$ ，由 (2.11) 式可延伸出

$$\underline{f}_n(x) \leq f(x) \leq \bar{f}_n(x) = \varpi_n g_n(x)$$

其中 g_n 是一個機率密度函數， ϖ_n 是標準化常數(normalized constant)。因為每個區間所取的上界均為直線所構成，對於 $i = 0, 1, \dots, n$ ，若我們將區間 (x_i, x_{i+1}) 的上界

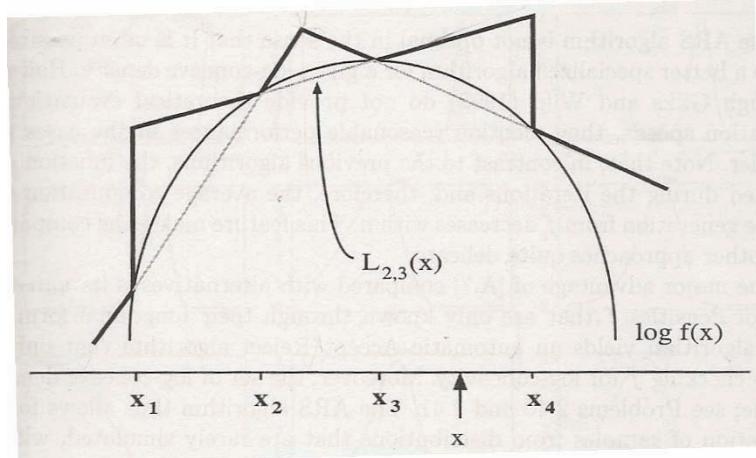


圖 2-5 內容顯示 $h(x) = \log f(x)$ 的上下界線，其中 f 為對數凹密度函數。
(來源：*Gilks et al. 1995*)

記作 $a_i + b_i x$ ，並以 $a_{-1} + b_{-1}x$ 、 $a_{n+1} + b_{n+1}x$ 分別表示區間 $(-\infty, x_0)$ 、 $(x_{n+1}, +\infty)$ 的上界，則上述函數 $\bar{h}_n(x)$ 、 ϖ_n 及 $g_n(x)$ 可詳細寫成

$$\begin{aligned}\bar{h}_n(x) &= \sum_{i=0}^n (a_i x + b_i) I_{[x_i, x_{i+1}]}(x) + (a_{-1} x + b_{-1}) I_{(-\infty, x_0]}(x) \\ &\quad + (a_{n+1} x + b_{n+1}) I_{[x_{n+1}, +\infty)}(x)\end{aligned}$$

$$\varpi_n = \int_{-\infty}^{x_0} e^{a_{-1}x + b_{-1}} dx + \sum_{i=0}^n \int_{x_i}^{x_{i+1}} e^{a_i x + b_i} dx + \int_{x_{n+1}}^{+\infty} e^{a_{n+1}x + b_{n+1}} dx$$

$$\begin{aligned}g_n &= \varpi_n^{-1} \bar{f}_n(x) = \varpi_n^{-1} \left\{ \sum_{i=0}^n e^{a_i x + b_i} I_{[x_i, x_{i+1}]}(x) + e^{a_{-1}x + b_{-1}} I_{(-\infty, x_0]}(x) \right. \\ &\quad \left. + e^{a_{n+1}x + b_{n+1}} I_{[x_{n+1}, +\infty)}(x) \right\}\end{aligned}$$

此 ARS 演算法如下：

演算法 5 ARS 演算法

1. 給定起始值 n 及 S_n ；
2. 生成 $X \sim g_n(x)$ 及 $U \sim \mathcal{U}[0, 1]$ ；
3. 若 $U \leq \underline{f}_n(X)/\varpi_n g_n(X)$ ，則接受 X ；
4. 否則，若 $U \leq f(X)/\varpi_n g_n(X)$ ，則接受 X ，並將 S_n 更新為 $S_{n+1} = S_n \cup \{X\}$ 。

對於 S_n 起始值的給定有個必要的條件為 $\varpi_n < \infty$ 。為達到這個要求，若 f 支集的左側不是有界的，則 $L_{0,1}$ 的斜率必須為正值；若 f 支集的右側不是有界的，則 $L_{n,n+1}$ 的

斜率必須為負值。

例題2.25 捉放模型(Capture and recapture models)

生態學的研究中經常遭遇到的第一個問題是在指定的研究區域中需要先知道不同野生族群的規模大小，為了正確地推論族群的規模，已有許多不同的捉放抽樣方法廣泛地被使用。在一個異質的(heterogeneous)捉放模型中，動物在時間 i 時被捕獲的機率為 p_i ， N 是未知的。令 I 為捕捉次數， n_i 為第 i 次捕捉期間捕捉到的動物個數， r 為所有被捕獲到的相異動物之個數，則此相關的概似函數如下：

$$L(p_1, \dots, p_I | N, n_1, \dots, n_I) = \frac{N!}{(N-r)!} \prod_{i=1}^I p_i^{n_i} (1-p_i)^{N-n_i}$$

若 $N \sim P(\lambda)$ (卜瓦松分佈)， p_i s 來自常態邏輯模型(normal logistic model)，George and Robert (1992) 說 $\log \frac{p_i}{1-p_i} \sim \mathcal{N}(\mu_i, \sigma^2)$ ，則 α_i 的事後分佈滿足

$$\pi(\alpha_i | N, n_1, \dots, n_I) \propto \pi(\alpha_i) f(N, n_1, \dots, n_I | \alpha_i) \propto \exp\left\{\alpha_i n_i - \frac{1}{2\sigma^2} (\alpha_i - \mu_i)^2\right\} / (1 + e^{\alpha_i})^N$$

此事後分佈取對數的結果為

$$\alpha_i n_i - \frac{1}{2\sigma^2} (\alpha_i - \mu_i)^2 - N \log(1 + e^{\alpha_i}) \quad (2.12)$$

(2.12)式對 α_i 二次微分可得 $-1/\sigma^2 - Ne^{\alpha_i}/(1 + e^{\alpha_i})^2 < 0$ ，可見 α_i 的事後分佈為一個凹函數，若欲模擬來自此事後分佈的隨機變數，ARS演算法將可適用於此。

Johnson 及 *Hoeting* 在 2003 年所提出的文章中，描述關於針尾鴨數量的研究，其中 $(n_1, \dots, n_{11}) = (32, 20, 8, 5, 1, 2, 0, 2, 1, 1, 0)$ 表示於 1956 年被做標記的 $N = 1612$ 隻針尾鴨在 1957 ~ 1968 年間被發現的數量。圖 2-6 提供了 1957 ~ 1965 年間 $\alpha_1, \alpha_2, \dots, \alpha_9$ 的事後分佈。ARS 演算法可分別被使用於這些分佈中。以 1960 年為例， $-10, -6, -3$ 可作為集合 S 的起始點，然後隨著疊代次數增加 S 也跟著更新。此演算法由 α_i 的事後分佈可得到一個正確的模擬，如圖 2-7 所示。

例題2.25 也說明了由 $\log \pi(\theta | x) = \log \pi(\theta) + \log f(x | \theta) + c$ (c 為與 θ 無關的常數) 中檢查 $\log \pi(\theta)$ 及 $\log f(x | \theta)$ 是否為凹函數，可以決定貝氏事後分佈($\pi(\theta | x)$)是否為對數凹函數。

例題2.26 卜瓦松迴歸(Poisson regression)

一組解釋變數為 x_i 反應變數 Y_i 為整數的樣本 $(Y_1, x_1), \dots, (Y_n, x_n)$ ，其中 Y_i 及 x_i 經由卜

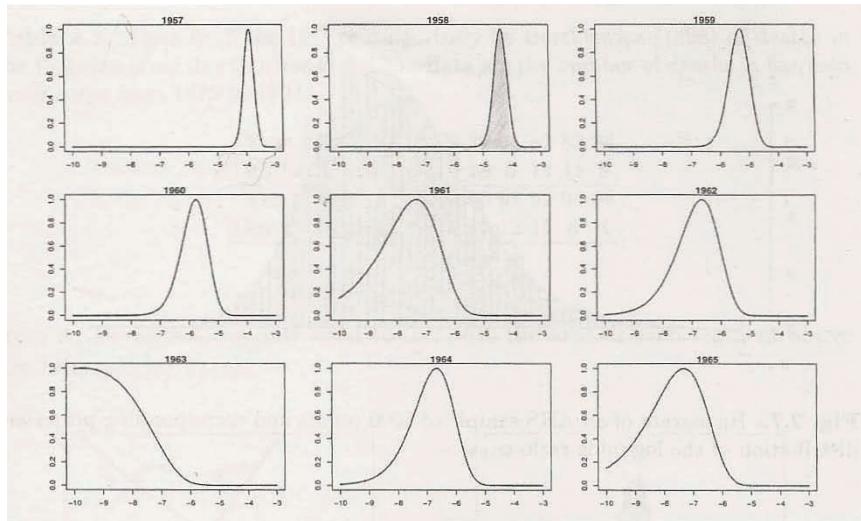


圖 2-6 由 1957 ~ 1965 年間的北方針尾鴨資料之 α_i 的事後分佈

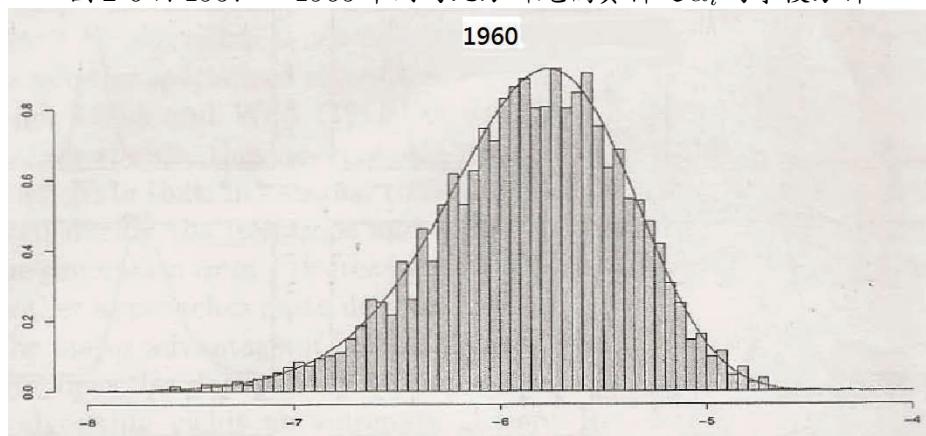


圖 2-7 由 ARS 演算法生成的 5000 筆樣本之直方圖及對應的 α_{1960} 的事後分佈

瓦松分佈來連結的關係如下

$$Y_i \mid x_i \sim \mathcal{P}(\exp\{a + bx_i\})$$

若 (a, b) 的先驗分佈為常態分佈 $\mathcal{N}(0, \sigma^2) \times \mathcal{N}(0, \tau^2)$ ，則 (a, b) 的事後分佈為

$$\begin{aligned}\pi(a, b \mid \mathbf{x}, \mathbf{y}) &\propto f(\mathbf{y}, a, b \mid \mathbf{x}) \\ &= \pi(a, b) f(\mathbf{y} \mid \mathbf{x}, a, b) \\ &= (1/\sqrt{2\pi}\sigma)e^{-a^2/2\sigma^2} (1/\sqrt{2\pi}\tau)e^{-b^2/2\tau^2} \prod_{i=1}^n (e^{(a+bx_i)y_i} e^{-e^{a+bx_i}})/y_i! \\ &\propto \exp\{a \sum_{i=1}^n y_i + b \sum_{i=1}^n y_i x_i - e^a \sum_{i=1}^n e^{x_i b}\} e^{-a^2/2\sigma^2} e^{-b^2/2\tau^2}\end{aligned}$$

通常我們也對模擬兩條件機率 $\pi(a \mid \mathbf{x}, \mathbf{y}, b)$ 及 $\pi(b \mid \mathbf{x}, \mathbf{y}, a)$ 感興趣。因為

$$\begin{aligned}\log \pi(a \mid \mathbf{x}, \mathbf{y}, b) &\propto a \sum_{i=1}^n y_i - e^a \sum_{i=1}^n e^{x_i b} - a^2/2\sigma^2 \\ \log \pi(b \mid \mathbf{x}, \mathbf{y}, a) &\propto b \sum_{i=1}^n y_i x_i - e^a \sum_{i=1}^n e^{x_i b} - b^2/2\tau^2\end{aligned}$$

且

$$\begin{aligned}\frac{\partial^2}{\partial a^2} \log \pi(a \mid \mathbf{x}, \mathbf{y}, b) &= -e^a \sum_{i=1}^n e^{x_i b} - \sigma^{-2} < 0 \\ \frac{\partial^2}{\partial b^2} \log \pi(b \mid \mathbf{x}, \mathbf{y}, a) &= -e^a \sum_{i=1}^n x_i^2 e^{x_i b} - \tau^{-2}\end{aligned}$$

可見 $\pi(a \mid \mathbf{x}, \mathbf{y}, b)$ 及 $\pi(b \mid \mathbf{x}, \mathbf{y}, a)$ 的確為對數凹函數，可利用ARS演算法來做模擬。

接著我們來看表 2.1，這資料是由 *von Bortkiewicz* 在 1898 年所收集的數據，記錄著普魯士軍隊遭受馬匹踢死的數據，其中感興趣的是死亡人數是否有隨時間變化的趨勢。以下我們來看該如何由ARS 方法模擬具分佈函數 $\pi(a \mid \mathbf{x}, \mathbf{y}, b)$ 的觀察值。使用 ARS 演算法前，先注意到兩點。

第一、若 $f(x)$ 是簡單可計算的(如本例題)，則不必建構下界 $f_n(x)$ ，使用 ARS 演算法時可省略此步驟。

第二、我們不需建構 g_n ，只需知道如何由 g_n 來模擬；意即，我們只需計算每個區間 $[x_i, x_{i+1}]$ 上的線段所在的範圍。

表 2-1

Year	75	76	77	78	79
Deaths	3	5	7	9	10
Year	80	81	82	83	84
Deaths	18	6	14	11	9
Year	85	86	87	88	89
Deaths	5	11	15	6	11
Year	90	91	92	93	94
Deaths	17	12	15	8	4

圖 2-8 的左圖中， $[x_2, x_3]$ 是 $f(x)$ 支集的一部分，灰色區域面積會正比於區域 $[x_2, x_3]$ 所對應的機率。圖中灰色區域面積爲

$$\begin{aligned}\omega_2 &= \int_{x_2}^{\frac{a_1-a_3}{b_3-b_1}} e^{a_1+b_1 x} dx + \int_{\frac{a_1-a_3}{b_3-b_1}}^{x_3} e^{a_3+b_3 x} dx \\ &= \frac{e^{a_1}}{b_1} \left[e^{\frac{a_1-a_3}{b_3-b_1} b_1} - e^{b_1 x_2} \right] + \frac{e^{a_3}}{b_3} \left[e^{b_3 x_3} - e^{\frac{a_1-a_3}{b_3-b_1} b_3} \right]\end{aligned}$$

因此，若想從 g_n 抽樣，我們只需選擇一個對應於 ω_i 的區間 $[x_i, x_{i+1}]$ ，並生成 $U \sim \mathcal{U}[0, 1]$ ，然後取 $X = x_i + U(x_{i+1} - x_i)$ 即為來自 g_n 的樣本點。圖 2-8 的右圖顯示 ARS 演算法所得樣本構成的直方圖，其中參數 $b = 0.025$ 、 $\sigma^2 = 5$ ，此與密度函數 g_n 的曲線是很近似的。

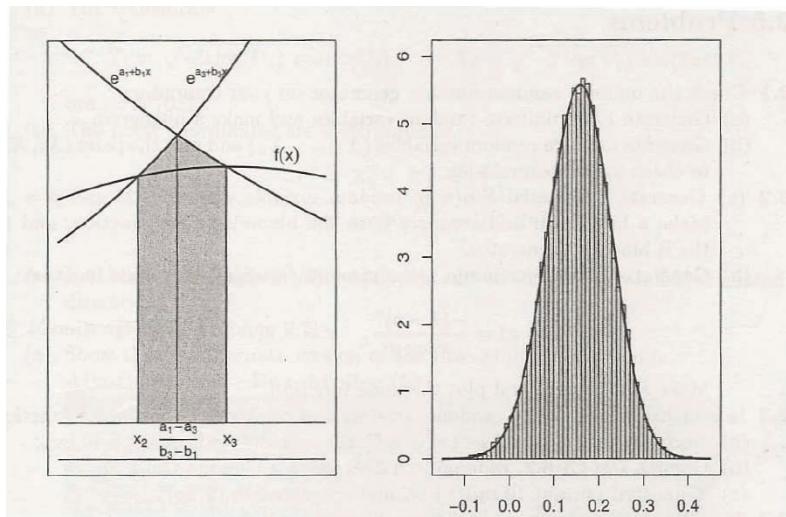


圖 2.8

3 蒙地卡羅積分(Monte Carlo Integration)

第二章我們把重點放在隨機變數的抽樣方法，本章節我們要介紹蒙地卡羅方法的中心概念，也就是利用電腦模擬所得的隨機變數來得到積分的近似值。

3.1 簡介(Introduction)

在統計分析上遇到的兩個主要問題，一個是最佳化問題，另一個是積分問題。雖然最佳化問題一般與概似方法有關，積分問題與貝氏方法有關，但兩問題並無嚴謹的分類。如例題1.1-1.15所提到的，並非任何情況都能推導出明顯的機率模型(explicit probability model)，因此有些時候當然就不太可能以分析方法計算出估計量；再者，某些統計方法，如拔靴法(bootstrap method)，雖與貝氏方法無關，但仍與經驗分佈函數(empirical cdf)的積分有關；相同地，相對於一般標準的概似方法，邊際概似方法需要對擾亂參數(nuisance parameters)積分，這也將遇到積分問題。

貝氏方法中，選定好損失函數(Loss function) $L(\theta, \delta)$ 及先驗分佈 π ，則貝氏估計量為使事後期望損失(the posterior expected loss) $E[L(\theta, \delta)|x]$ 達到最小值的 δ^π ，因為事後分佈 $\pi(\theta|x) \propto f(x|\theta) = \pi(\theta)f(x|\theta)$ ，所以貝氏估計量可視為下列最小化問題的解

$$\min_{\delta} \int_{\Theta} L(\theta, \delta) \pi(\theta) f(x|\theta) d\theta \quad (3.1)$$

若損失函數為二次損失函數(quadratic loss function) $\|\theta - \delta\|^2$ ，則貝氏估計量為事後期望值(posterior expectation) $E(\theta|x)$ ；若損失函數為絕對損失函數 $|\theta - \delta|$ ，則貝氏估計量為事後中位數(posterior median)。然而一般的損失函數多數會使得(1)式無法以分析方法做積分，而必須使用數值法或模擬法來得到貝氏估計量的近似解。在正式介紹使用模擬方法求積分值之前，我們來看看一些特別的貝氏方法所遇到的積分問題。

例題3.1 絕對損失函數

令 $\theta \in R$, $L(\theta, \delta) = |\theta - \delta|$ ，此時貝氏估計量 $\delta^\pi(x)$ 為 $\pi(\theta|x)$ 的事後中位數，也就是下列方程式的解

$$\int_{\theta \leq \delta^\pi(x)} \pi(\theta) f(x|\theta) d\theta = \int_{\theta \geq \delta^\pi(x)} \pi(\theta) f(x|\theta) d\theta \quad (3.2)$$

若 $X \sim \mathcal{N}_p(\theta, I_p)$ (多變量常態分佈)，令 $\lambda = \|\theta\|^2$ ，

使用參考先驗分佈(reference prior) $\pi(\theta) = \|\theta\|^{-(p-1)}$ ，並將 θ 用極座標表示為

$$\theta = (\sqrt{\lambda} \cos \varphi_1, \sqrt{\lambda} \sin \varphi_1 \cos \varphi_2, \sqrt{\lambda} \sin \varphi_1 \sin \varphi_2 \cos \varphi_3, \dots, \sqrt{\lambda} \sin \varphi_1 \sin \varphi_2 \cdots \sin \varphi_{p-1})$$

則變數變換所需的Jacobin行列式為

$$\begin{aligned} \det & \left(\begin{array}{cccc} \frac{1}{2\sqrt{\lambda}} \cos \varphi_1 & -\sqrt{\lambda} \sin \varphi_1 & \cdots & 0 \\ \frac{1}{2\sqrt{\lambda}} \sin \varphi_1 \cos \varphi_2 & \sqrt{\lambda} \cos \varphi_1 \cos \varphi_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{2\sqrt{\lambda}} \sin \varphi_1 \cdots \sin \varphi_{p-1} & \sqrt{\lambda} \cos \varphi_1 \cdots \sin \varphi_{p-1} & \cdots & \sqrt{\lambda} \sin \varphi_1 \cdots \cos \varphi_{p-1} \end{array} \right) \\ & = \frac{1}{2} (\sqrt{\lambda})^{p-1} \prod_{i=1}^{p-2} \sin(\varphi_i)^{p-i-1} \end{aligned}$$

則 λ 的參考先驗分佈為

$$\begin{aligned} \pi(\lambda) & = \int \lambda^{-(p-1)/2} \cdot \frac{1}{2} (\sqrt{\lambda})^{p-1} \prod_{i=1}^{p-2} \sin(\varphi_i)^{p-i-1} d\varphi_1 \cdots d\varphi_{p-1} \\ & = \int \prod_{i=1}^{p-2} \sin(\varphi_i)^{p-i-1} d\varphi_1 \cdots d\varphi_{p-1} \end{aligned}$$

且 λ 的事後分佈如下

$$\begin{aligned} \pi(\lambda|x) & \propto \pi(\lambda) f(x|\lambda) \\ & \propto \int e^{-\|x-\theta\|^2/2} \prod_{i=1}^{p-2} \sin(\varphi_i)^{p-i-1} d\varphi_1 \cdots d\varphi_{p-1} \end{aligned}$$

由此事後分佈複雜的積分形式，可知道若將此事後分佈代入第(3.1)式的積分式中想求出貝氏估計量，一定是很困難的。

例題3.2 分段線性及二次損失函數

(piecewise linear and quadratic loss functions)

給定損失函數為

$$L(\theta, \delta) = \sum_i w_i (\theta - \delta)^2 I_{\theta-\delta \in [a_i, a_{i+1})}, \quad w_i > 0 \quad (3.3)$$

其事後期望值為

$$E[L(\theta, \delta)|x] = \sum_i \int_{a_i}^{a_{i+1}} w_i (\theta - \delta)^2 \pi(\theta|x) d\theta$$

將上式對 δ 微分後令其結果等於 0 可得

$$\sum_i w_i \int_{a_i}^{a_{i+1}} (\theta - \delta^\pi(x)) \pi(\theta|x) d\theta = 0$$

由此可解出貝氏估計量的形式是

$$\delta^\pi(x) = \frac{\sum_i w_i \int_{a_i}^{a_{i+1}} \theta \pi(\theta) f(x|\theta) d\theta}{\sum_i w_i \int_{a_i}^{a_{i+1}} \pi(\theta) f(x|\theta) d\theta}$$

此貝氏估計量的分子分母分別是限制在各區間 $[a_i, a_{i+1})$ 的事後期望值及事後機率，若想得到 $\delta^\pi(x)$ 的顯明形式，仍然與積分有關。很類似地，若考慮分段線性損失函數

$$L(\theta, \delta) = \sum_i w_i \mid \theta - \delta \mid I_{\theta-\delta \in [a_i, a_{i+1})}, \quad w_i > 0$$

或 Huber (1972) 提出的損失函數

$$L(\theta, \delta) = \begin{cases} \rho(\theta - \delta)^2 & \text{若 } |\theta - \delta| < c \\ 2\rho c \{|\theta - \delta| - c/2\} & \text{其它} \end{cases}$$

(ρ, c為特定常數)

也會如上述分段二次損失函數般，解出仍然含有積分式的貝氏估計量 δ^π ，而這些積分若想以分析法處理可能是相當複雜的。

相較於一般的貝氏方法，經驗貝氏方法(empirical Bayes method)也是很常被使用的。給定分佈函數 $f(x|\theta)$ 及共軛先驗分佈 $\pi(\theta|\lambda, \mu)$ ，經驗貝氏方法會先使用最大概似計法，從邊際分佈

$$m(x|\lambda, \mu) = \int f(x|\theta) \pi(\theta|\lambda, \mu) d\theta$$

中求出超參數(hyperparameter) λ, μ 的最大概似估計量 $\hat{\lambda}, \hat{\mu}$ ，再利用 $\pi(\theta|\hat{\lambda}, \hat{\mu})$ 來估計 θ 。下面例子要介紹的是使用經驗貝氏方法求經驗貝氏估計量時可能遇到的問題。

例題3.3 經驗貝氏估計量

令 $X \sim N_p(\theta, I_p)$ ，共軛先驗分佈為 $\theta \sim N_p(\mu, \lambda I_p)$ ，其中通常會先給定超參數 $\mu = 0$ ，而另一個超參數 λ 則以最大概似估計量 $\hat{\lambda}$ 來取代。 X 的邊際分佈為

$$\begin{aligned} m(x|\lambda) &= \int f(x|\theta) \pi(\theta|\lambda) d\theta \\ &= \int \frac{1}{(\sqrt{2\pi})^p} \exp\{-\|x - \theta\|^2/2\} \frac{1}{(\sqrt{2\pi\lambda})^p} \exp\{-\|\theta\|^2/2\lambda\} d\theta \\ &= \int \frac{1}{(\sqrt{2\pi(\lambda+1)})^p} \exp\{-\|x\|^2/2(\lambda+1)\} \frac{1}{(\sqrt{2\pi\lambda/(\lambda+1)})^p} \\ &\quad \times \exp\{-(\lambda+1)\|\theta - \lambda x/(\lambda+1)\|^2/2\lambda\} d\theta \\ &= \frac{1}{(\sqrt{2\pi(\lambda+1)})^p} \exp\{-\|x\|^2/2(\lambda+1)\} \end{aligned}$$

結果顯示 X 的邊際分佈為常態分佈 $N_p(0, (\lambda+1)I_p)$ 。接著由概似函數對 λ 微分並令微

分後式子等於 0，

$$\begin{aligned} dL(\lambda|x)/d\lambda \\ = \frac{1}{(\sqrt{2\pi(\lambda+1)})^p} \exp\{-\|x\|^2/2(\lambda+1)\} \left[\frac{-p}{2(\lambda+1)^{-p/2-1}} + \frac{\|x\|^2}{2(\lambda+1)^{2+p/2}} \right] = 0 \end{aligned}$$

可解得超參數 λ 的最大概似估計量 $\hat{\lambda} = (\|x\|^2/p - 1)^+$ 。

給定 λ 下， θ 的事後分佈為

$$\begin{aligned} \pi(\theta|x, \lambda) &= \frac{\pi(\theta|\lambda)f(x|\theta, \lambda)}{m(x|\lambda)} \\ &= \frac{1}{(\sqrt{2\pi\lambda/(\lambda+1)})^p} \exp\{-(\lambda+1)\|\theta - \lambda x/(\lambda+1)\|^2/2\lambda\} \end{aligned}$$

此 θ 的事後分佈正好為常態分佈 $N_p(\lambda x/(\lambda+1), [\lambda/(\lambda+1)]I_p)$ ，而經驗貝氏方法則以 θ 的擬事後分佈(pseudo posterior distribution) $N_p(\hat{\lambda}x/(\hat{\lambda}+1), [\hat{\lambda}/(\hat{\lambda}+1)]I_p)$ 做為基礎進行推論。舉例來說，若感興趣的是 $\|\theta\|^2$ ，配合二次損失函數可得經驗貝氏估計量為事後期望值如下

$$\begin{aligned} \delta^{eb}(x) &= E(\|\theta\|^2|x) = E(\theta|x)^2 + Var(\theta|x) \\ &= \left(\frac{\hat{\lambda}}{(\hat{\lambda}+1)} \right)^2 \|x\|^2 + \frac{\hat{\lambda}}{(\hat{\lambda}+1)} \times p \\ &= \left[\left(1 - \frac{p}{\|x\|^2} \right)^+ \right]^2 \|x\|^2 + p \left(1 - \frac{p}{\|x\|^2} \right)^+ \\ &= (\|x\|^2 - p)^+ \end{aligned}$$

此經驗貝氏估計量有兩點重要特質：

第一、 $\delta^{eb}(x)$ 是 $\|\theta\|^2$ 的最佳不偏估計量。

第二、若 $\|X\|^2 \sim \chi_p^2(\|\theta\|^2)$ ，則 $\delta^{eb}(x)$ 是 $\|\theta\|^2$ 的最大概似估計量。

關於第一點，因為 $X \sim N_p(\theta, I_p)$ ， $E[\delta^{eb}(X)] = E(\|X\|^2 - p) = E(\|X\|^2) - p = Var(X) + (E(X))^2 - p = p + \|\theta\|^2 - p = \|\theta\|^2$ ，可見 $\delta^{eb}(x)$ 是 $\|\theta\|^2$ 的最佳不偏估計量；而關於第二點，由於非中心化卡方分佈的密度函數為

$$f(x \mid \|\theta\|^2) = \frac{1}{2} (x/\lambda)^{(p-2)/4} I_{(p-2)/2}(\sqrt{\|\theta\|^2 x}) e^{-(\|\theta\|^2+x)/2}$$

其中 $I_{(p-2)/2}(\sqrt{\|\theta\|^2 x})$ 為含有積分式且形式較複雜的貝索函數(Bessel function)，因此關於第二點的驗證，直接使用分析方法是很困難的。

以上數個例題所遇到的計算上的困難，一般都是透過模擬真實分佈或近似分佈來計算感興趣的量，以解決前述分析方法不好處理的問題。

3.2 典型蒙地卡羅積分(Classical Monte Carlo Integration)

為了處理期望值

$$E_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx \quad (3.4)$$

的積分問題，常見的方法是從密度函數 f 中生成一組樣本 (X_1, \dots, X_m) ，並透過強大數法則(Strong Law of Large number)來保證平均值 $\bar{h}_m = \frac{1}{m} \sum_{j=1}^m h(x_j)$ 將幾乎必然地(almost surely)收斂到 $E_f[h(X)]$ 。再者，若 $E_f[h^2(X)]$ 是有限的，則變異數

$$\begin{aligned} Var(\bar{h}_m) &= \frac{\sum_{i=1}^m Var[h(X_i)]}{m^2} = \frac{mVar[h(X)]}{m^2} = \frac{Var[h(X)]}{m} \\ &= \frac{1}{m} \int_{\mathcal{X}} (h(x) - E_f[h(x)])^2 f(x)dx \end{aligned}$$

也可利用樣本變異數 $v_m = \frac{1}{m^2} \sum_{j=1}^m [h(x_j) - \bar{h}_m]^2$ 來估計，藉此我們可以評估 \bar{h}_m 收斂到 $E_f[h(x)]$ 的速度。此外，由中央極限定理可得：若 m 夠大，隨機變數 $\frac{\bar{h}_m - E_f[h(x)]}{\sqrt{v_m}}$ 將近似標準常態分佈，此結果有助於對 $E_f[h(x)]$ 的估計量建構收斂的檢定方法及信賴界。

例題3.4

回憶例題 1.17 中提及的函數 $h(x) = [\cos(50x) + \sin(20x)]^2$ ， $h(x)$ 在 $(0, 1)$ 區間上真實的積分值為

$$\int_0^1 h(x)dx = \int_0^1 (1 + \frac{1}{2}\cos 100x - \frac{1}{2}\sin 40x + \sin 70x - \sin 30x)dx \approx 0.965$$

使用蒙地卡羅積分法：生成隨機變數 $U_1, \dots, U_n \sim \mathcal{U}(0, 1)$ ，以 $\frac{1}{n} \sum_{i=1}^n h(U_i)$ 來近似 $\int h(x)dx$ 。圖 3-1 中，左圖為 $h(x)$ 在 $(0, 1)$ 區間上的函數圖形，中間為 $h(x)$ 的函數值形成的直方圖，右圖則顯示疊代 10000 次所得的移動平均數及標準差，由此看出蒙地卡羅積分法的結果約收斂到 0.963，而此積分的真實值為 0.965。

例題3.5

已知常態分佈的累積分佈函數為

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

此分佈函數無法寫成封閉解的形式，若想建構常態分佈表，就必須透過模擬方法來完成。我們先利用例題 2.8 介紹的 Box-Muller 演算法生成 n 個具常態分佈的樣本

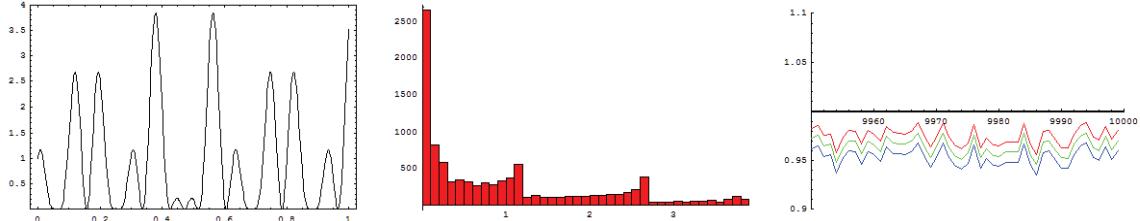


圖 3-1

n/Z_α	0.	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.54	0.775	0.815	0.885	0.96	1.	0.98	1.	1.
10^3	0.494	0.743	0.8185	0.901	0.949	0.991	0.9935	0.997	1.
10^4	0.5024	0.74355	0.79995	0.8986	0.9505	0.99055	0.99475	0.99895	0.99985
10^5	0.49905	0.746445	0.80064	0.899225	0.950325	0.98961	0.99543	0.99902	0.99986
10^6	0.49967	0.748358	0.799234	0.899851	0.950484	0.989853	0.995096	0.999002	0.99989
10^7	0.500055	0.748578	0.799559	0.899708	0.950486	0.989831	0.99507	0.999008	0.999901

表 3-1

(x_1, \dots, x_n) ，由蒙地卡羅方法可得

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^n I_{x_i \leq t}$$

其中每個隨機變數 $I_{x_i \leq t}$ 來自成功機率為 $\Phi(t)$ 的伯努力分佈，所以 $\hat{\Phi}(t)$ 的變異數為 $\Phi(t)(1 - \Phi(t))/n$ 。當 t 很接近 0，此變異數會很接近 $1/4n$ ；當模擬次數約為 $n = 2 \times 10^8$ 時，此估計值將準確到小數點後第四位。表 3-1 為對不同 t 值，利用 $\hat{\Phi}(t)$ 模擬而得的累積機率近似值。

許多檢定會使用到漸進常態假設，例如概似比檢定 (the likelihood ratio test)，給定具有 r 個獨立限制之 $\theta \in \mathcal{R}^k$ 的虛無假設 H_0 ， $\hat{\theta}$ 及 $\hat{\theta}^0$ 分別表示 θ 的未受限制及受限制 (H_0) 的最大概似估計量，則概似比 $l(\hat{\theta}|x)/l(\hat{\theta}^0|x)$ 會滿足

$$2 \log \left[\frac{L(\hat{\theta}|x)}{L(\hat{\theta}^0|x)} \right] = 2 \left\{ \log L(\hat{\theta}|x) - \log L(\hat{\theta}^0|x) \right\} \xrightarrow{\mathcal{L}} \chi_r^2 \quad (3.5)$$

例題3.6 列聯表(contingency table)

表 3-2

	惡性腫瘤 受控制	惡性腫瘤 不受控制	
外科手術	21	2	23
輻射	15	3	18
	36	5	41

上表內容為比較 41 個咽喉癌患者透過外科手術及輻射治療的成果。令每個觀察值 X_i 來自多項式分佈，即 $X_i \sim \mathcal{M}_4(1, \mathbf{p})$, $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$, $\sum_{ij} p_{ij} = 1$, $i = 1, \dots, n$, y_{ij} 表示落在第 ij 格的數，則概似函數可寫成

$$L(\mathbf{p}|\mathbf{y}) \propto \prod_{ij} p_{ij}^{y_{ij}}$$

此虛無假設欲檢定的是獨立性，也就是說治療方式與惡性腫瘤是否受控制無關。

對照表 3-2，將此敘述轉換成參數型式如下：

$$\begin{array}{cc|c} p_{11} & p_{12} & p_1 \\ p_{21} & p_{22} & 1 - p_1 \\ \hline p_2 & 1 - p_2 & 1 \end{array}$$

$H_0 : p_{11} = p_1 p_2$ ，概似比檢定統計量為

$$\lambda(\mathbf{y}) = \frac{\max_{\mathbf{p}: p_{11}=p_1 p_2} L(\mathbf{p}|\mathbf{y})}{\max_{\mathbf{p}} L(\mathbf{p}|\mathbf{y})}$$

分子、分母的最大值發生處分別為 $\hat{p}_1 = \frac{y_{11}+y_{12}}{n}$ 及 $\hat{p}_{ij} = \frac{y_{ij}}{n}$ 。

如同前面敘述，在 H_0 的假設下，此檢定統計量 $-2 \log \lambda$ 具漸進卡方分佈 χ_1^2 ，但是因為只有 41 個觀察值，此漸進分佈並不適用。解決方法有兩種：第一，排列檢定；第二，蒙地卡羅法。以下針對第二點加以討論。

我們利用蒙地卡羅法模擬虛無假設下 $-2 \log \lambda$ 或 λ 的虛無分佈(null distribution)，以得到假設檢定的截點(cutoff point)，將此虛無分佈記作 $f_0(\lambda)$ ，若顯著水準設為 α ，則由積分值

$$\int_0^{\lambda_\alpha} f_0(\lambda) d\lambda = 1 - \alpha \quad (3.6)$$

可反解出 λ_α 。對於此問題，標準的蒙地卡羅法是生成隨機變數 $\lambda^t \sim f_0(\lambda)$, $t =$

$1, \dots, M$ ，將所得樣本排序 $\lambda^{(1)} \leq \lambda^{(2)} \leq \dots \leq \lambda^{(M)}$ 並挑出 $(1 - \alpha)$ 經驗百分位數(empirical percentile) $\lambda^{(\lfloor (1-\alpha)M \rfloor)}$ ，然後我們可以得到

$$\lim_{M \rightarrow \infty} \lambda^{(\lfloor (1-\alpha)M \rfloor)} \rightarrow \lambda_\alpha$$

由於參數 p_1, p_2 沒有特定，僅知道 $p_1, p_2 \sim \mathcal{U}(0, 1)$ ，若想從 $f_0(\lambda)$ 中抽樣，需先抽出

$$\begin{aligned} p_i &\sim \mathcal{U}(0, 1), \quad i = 1, 2 \\ \mathbf{X} &\sim \mathcal{M}_4(p_1 p_2, p_1(1-p_2), (1-p_1)p_2, (1-p_1)(1-p_2)) \end{aligned} \quad (3.7)$$

再去計算概似比檢定統計量 $\lambda(\mathbf{X})$ 的值。結果如表 3-3 及圖 3-2 所示。圖 3-2-1 為模擬蒙地卡羅法之虛無分佈所得的直方圖，其形狀近似 χ_1^2 分佈。圖 3-2-2 由下而上分別為模擬 10000 次所得 0.90, 0.95, 0.99 移動經驗百分位數(running empirical percentile)，發現愈上層的百分位數的變動愈大。

表 3-3

α	截點(蒙地卡羅法)	χ_1^2
0.10	2.84	2.705
0.05	3.93	3.841
0.01	6.72	6.635

例題3.7

常態混合(normal mixture)模型

$$p\mathcal{N}(\mu, 1) + (1-p)\mathcal{N}(\mu + \theta, 1)$$

此處限制 $\theta > 0$ 以保證模型的可識別性，在概似比檢定中，此為卡方(χ_r^2)分佈無法被使用的一個特殊例子。混合模型檢定的虛無假設是無法簡單表示的，因為若假設 $H_0 : p = 0$ ，則當不拒絕 H_0 時，將使混合模型的一部分消失而只剩單一分佈 $\mathcal{N}(\mu + \theta, 1)$ ，這將導致概似函數只能估計 $\mu + \theta$ 而無法分別估計出 μ 和 θ 。若將虛無假設修改成 $H_0 : p = 1$ 或 $\theta = 0$ ，則可以解決前述分別估計 μ 和 θ 的問題，而在這樣的虛無假設及其相關的對立假設下，我們想知道概似比檢定中 (3.5) 式的極限分佈。圖 3-3 最上方顯示的為常態混合模型之對數概似比 $2 \left\{ \log L(\hat{p}, \hat{\mu}, \hat{\theta} | x) - \log L(\hat{\mu}^0 | x) \right\}$ 的經驗分佈函數，其中 $\hat{p}, \hat{\mu}, \hat{\theta}, \hat{\mu}^0$ 為模擬 1000 次每次 100 個標準常態分佈樣本所得的最大概似估計量，又最下方曲線為卡方(χ_2^2)分佈的分佈函數，明顯地可以看出此兩分佈函數的圖形並不一致，此處可見第 (3.5) 式對於對數概似比近似於卡方分佈的結果不太符合。若將 (3.5) 式改為相等權重 ($p = 0.5$) 的卡方(χ_2^2)分佈與狄拉克函數於 $x = 0$ 處的混合型式，將可改善上述 χ_2^2 近似不佳的問題，圖 3-3 中間較接近最上方曲線的函數圖就是此改

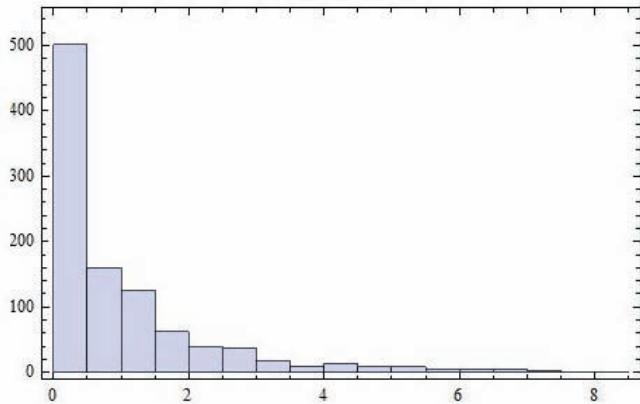


圖 3-2-1 虛無假設下之分佈的直方圖及近似卡方 χ_1^2 分佈的密度函數圖形。

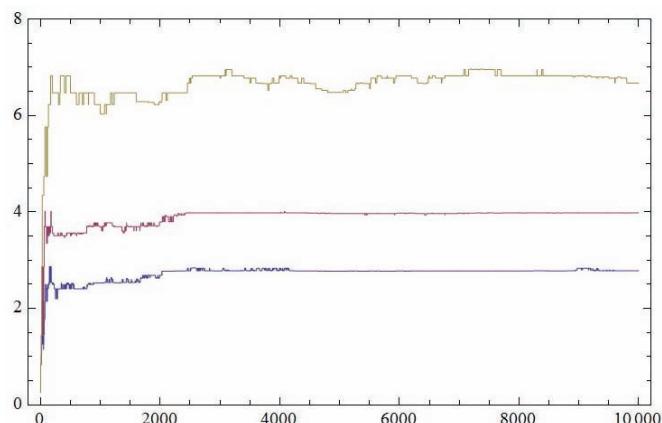


圖 3-2-2 由下而上分別為 10000 次模擬所得之 0.90、0.95、0.99 移動經驗百分位數。

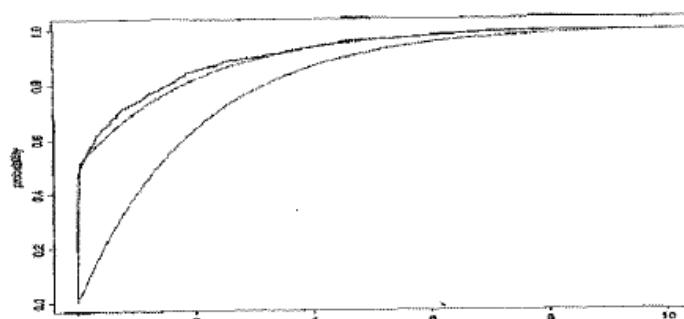


圖 3-3 由上而下分別為常態混合模型、 χ_2^2 與 Dirac mass 機率各半的混合模型、 χ_2^2 等，以上三種分佈的經驗累積分佈函數。

善後的結果。

3.3 重要抽樣(Importance Sampling)

3.3.1 原理(Principles)

令 f 為感興趣的目標函數，若對 f 的積分是不容易處理的，本節欲介紹的解決方法是參考第 3.2 節第 (3.4) 式提及的期望值概念，將對 f 的積分表示式修改為對某特定分佈的期望值，此處我們並非對感興趣的目標函數 f 抽樣，而是從此特定分佈抽樣來處理此積分問題。

例題3.8 柯西分佈尾部機率

令 $X \sim \mathcal{C}(0, 1)$ (柯西分佈)，感興趣的量為 $p = \int_2^\infty \frac{1}{\pi(1+x^2)} dx$ 。若獨立隨機變數 $X_1, \dots, X_m \sim \mathcal{C}(0, 1)$ ，則使用經驗分佈得到的估計值為

$$\hat{p}_1 = \frac{1}{m} \sum_{j=1}^m I_{[X_j > 2]}$$

且

$$\begin{aligned} Var(\hat{p}_1) &= \frac{1}{m^2} \sum_{j=1}^m Var(I_{[X_j > 2]}) \\ &= \frac{1}{m^2} \cdot m \cdot p(1-p) = p(1-p)/m \approx \frac{0.127}{m} \\ &\left(\because p = \int_2^\infty \frac{1}{\pi(1+x^2)} = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} 2 \approx 0.15 \right) \end{aligned}$$

若將 \hat{p}_1 改成

$$\hat{p}_2 = \frac{1}{2m} \sum_{j=1}^m I_{[|X_j| > 2]}$$

則

$$\begin{aligned} Var(\hat{p}_2) &= \frac{1}{4m^2} \sum_{j=1}^m Var(I_{[|X_j| > 2]}) \\ &= \frac{1}{4m^2} \cdot m \cdot 2p(1-2p) = p(1-2p)/2m \approx 0.052/m \end{aligned}$$

此結果小於 $Var(\hat{p}_1)$ 。以上方法有個較不有效率的地方是生成的樣本可能會落在 $[2, \infty)$ 以外的區域，若將 p 寫成

$$p = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx$$

則上述積分部分可視為是 $X \sim \mathcal{U}_{[0,2]}$ ， $h(X) = 2/\pi(1+X^2)$ 的期望值，故 p 的估計量可再改成

$$\hat{p}_3 = \frac{1}{2} - \frac{1}{m} \sum_{j=1}^m h(X_j)$$

此變異數為

$$\begin{aligned} Var(\hat{p}_3) &= (E[(h(X))^2] - E^2[h(X)])/m \\ &= \frac{1}{m} \left[\left(\frac{1}{5\pi^2} + \frac{1}{2\pi^2} \tan^{-1} 2 \right) - \left(\frac{1}{\pi} \tan^{-1} 2 \right)^2 \right] \\ &\approx 0.0285/m \end{aligned}$$

令 $y = 1/x$ ，則 $0 < y < 1/2$ 且 p 可以再改寫成

$$\begin{aligned} p &= \int_2^\infty \frac{1}{\pi(1+x^2)} dx \\ &= \int_0^{1/2} \frac{1}{\pi(1+y^{-2})} y^{-2} dy \\ &= \int_0^{1/2} \frac{1}{4} \frac{2}{\pi(1+y^2)} \cdot 2dy \end{aligned}$$

此時 p 可視為 $Y \sim \mathcal{U}_{[0,1/2]}$ ， $\frac{1}{4}h(Y) = \frac{1}{4}\frac{2}{\pi(1+y^2)}$ 的期望值，故 p 的估計量可再改成

$$\hat{p}_4 = \frac{1}{4m} \sum_{j=1}^m h(Y_j)$$

此變異數為

$$\begin{aligned} Var(\hat{p}_4) &= (E[(h(Y))^2] - E^2[h(Y)])/16m \\ &= \frac{1}{m} \left[\left(\frac{8 + 20\tan^{-1} 1/2 - 80(\tan^{-1} 1/2)^2}{5\pi^2} \right) \right] \\ &\approx 0.95 \times 10^{-4}/m \end{aligned}$$

在準確度相同的情況下，

$$Var(\hat{p}_1)/Var(\hat{p}_4) = \frac{0.127/m_1}{0.95 \times 10^{-4}/m_2} \approx 1 \Rightarrow \frac{m_1}{m_2} \approx 10^3$$

此結果顯示利用 \hat{p}_4 估計 p 所需模擬次數比利用 \hat{p}_1 時所需模擬次數減少約 $\sqrt{1000} \approx 32$ 倍。

有別於直接從 f 抽樣，第(3.4)式有另一個估計方法為重要抽樣法(importance sampling)，定義如下：

定義3.9 紿定一分佈 g ， $supp(g) \supset supp(f)$ ，從 g 中抽樣，得到樣本 X_1, \dots, X_n ，若將第 (3.4) 式寫成另一種形式

$$E_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx = \int_{\mathcal{X}} h(x)\frac{f(x)}{g(x)}g(x)dx = E_g[h(X)\frac{f(X)}{g(X)}] \quad (3.8)$$

則可將此積分視爲

$$E_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j) \quad (3.9)$$

第 (3.9) 式稱爲重要抽樣法的基本恆等式(the importance sampling fundamental identity)，依前述蒙地卡羅法原理，無論 g 選擇何種分佈，只要滿足 $supp(g) \supset supp(f)$ ，則第 (3.8) 式將收斂到第(3.4)式。由於重要抽樣法所選擇的分佈 g 並沒有太多限制，且未必要與原積分式中的函數 h 或原分佈函數 f 有關係，只要 g 本身是好模擬的分佈即可。因此，從 g 中抽樣所得樣本 X_1, \dots, X_n ，可以重複利用於不同函數 h 或不同分佈 f ，此特徵在穩健性(robustness)和貝氏敏感度分析(Bayesian sensitivity analyses)上是非常受到關注的。

例題3.10 指數與對數常態的比較

設目標函數 f 為指數分佈 $Exp(1/\lambda)$, $\lambda > 0$ 的密度函數， $f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}}$ ，而重要抽樣法所選擇的工具分佈函數(instrumental distribution) g 為對數常態分佈 $\mathcal{LN}(0, \sigma^2)$, $\lambda = e^{\sigma^2/2}$ 的密度函數， $g(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\log x)^2}{2\sigma^2}}$ ，若感興趣的參數爲 λ ，估計量 $\hat{\lambda} = X$ ，考慮尺度化平方誤差損失函數 $L(\lambda, \delta) = \frac{(\delta - \lambda)^2}{\lambda^2}$ (此處 $\delta = \hat{\lambda} = X$)，則其風險的理論值爲

$$\begin{aligned} R_1 &= E_f[L(\lambda, X)] \\ &= \int_0^\infty \frac{(x - \lambda)^2}{\lambda^2} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} dx \\ &= \int_{-1}^\infty t^2 e^{-t-1} dt \quad (t = \frac{x}{\lambda} - 1) \\ &= \int_{-1}^0 t^2 e^{-t-1} dt + \Gamma(3)e^{-1} \int_0^\infty t^2 e^{-t}/\Gamma(3)dt \\ &= (1 - 2e^{-1}) + 2e^{-1} = 1 \end{aligned}$$

若以重要抽樣法估計上述風險，需先從函數 g 抽樣，得到一組單一樣本 X_1, \dots, X_T ，則 $R_1 = E_f[L(\lambda, X)] = \int_0^\infty L(\lambda, x) \frac{f(x)}{g(x)} g(x) dx \approx \frac{1}{T} \sum_{t=1}^T \frac{f(x_t)}{g(x_t)} L(\lambda, x_t)$ ，

也就是 R_1 的估計值為

$$\begin{aligned}\hat{R}_1 &= \frac{1}{T} \sum_{t=1}^T \left[\left(\frac{1}{\lambda} e^{-\frac{X_t}{\lambda}} \right) / \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\log X_t)^2}{2\sigma^2}} \frac{1}{X_t} \right) \right] \frac{(X_t - \lambda)^2}{\lambda^2} \\ &= \frac{1}{T\lambda^2} \sum_{t=1}^T X_t e^{-X_t/\lambda} \lambda^{-1} e^{(\log X_t)^2/2\sigma^2} \sqrt{2\pi}\sigma (X_t - \lambda)^2\end{aligned}$$

若目標函數 f 為也是對數常態分佈 $\mathcal{LN}(0, \sigma^2)$, $\lambda = e^{\sigma^2/2}$, 此時 f 恰好等於 g , 則其風險的理論值為

$$\begin{aligned}R_2 &= E_f[L(\lambda, X)] = \int_0^\infty L(\lambda, x) f(x) dx \\ &= \int_0^\infty \frac{(x - \lambda)^2}{\lambda^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\log x)^2}{2\sigma^2}} \frac{1}{x} dx \quad (\text{令 } \log x/\sigma = t) \\ &= \int_{-\infty}^\infty \frac{(e^{\sigma t} - \lambda)^2}{\lambda^2} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\lambda^2} e^{-t^2/2+2\sigma t} - \frac{2}{\lambda} e^{-t^2/2+\sigma t} + e^{-t^2/2} \right) dt \\ &= \frac{1}{\lambda^2} e^{2\sigma^2} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-(t-2\sigma)^2/2} dt \\ &\quad - \frac{2}{\lambda} e^{\sigma^2/2} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-(t-\sigma)^2/2} dt \\ &\quad + \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= \frac{1}{\lambda^2} e^{2\sigma^2} - \frac{2}{\lambda} e^{\sigma^2/2} + 1 \quad (\text{其中 } e^{\sigma^2/2} = \lambda) \\ &= \lambda^2 - 1\end{aligned}$$

而 $R_2 = E_f[L(\lambda, X)] = \int_0^\infty L(\lambda, x) \frac{f(x)}{g(x)} g(x) dx$ (此處 $f(x) = g(x)$) $= \int_0^\infty L(\lambda, x) g(x) dx$, 故 R_2 的估計值為

$$\begin{aligned}\hat{R}_2 &= \frac{1}{T} \sum_{t=1}^T L(\lambda, X_t) \\ &= \frac{1}{T} \sum_{t=1}^T \frac{(X_t - \lambda)^2}{\lambda^2}\end{aligned}$$

結果如圖 3-4 所示。

例題3.11 微小的尾部機率

例題 3.5 我們曾利用蒙地卡羅法估算標準常態分佈的累積機率 $\Phi(t)$, 但發現越接近尾端的估計越不精確, 需要明顯提高模擬次數來改善, 但此作法當越往尾部估計就越不適用。例如 $Z \sim \mathcal{N}(0, 1)$, 感興趣的量值為 $P(Z > 4.5)$ (此為非常小的數值), 若使用蒙地

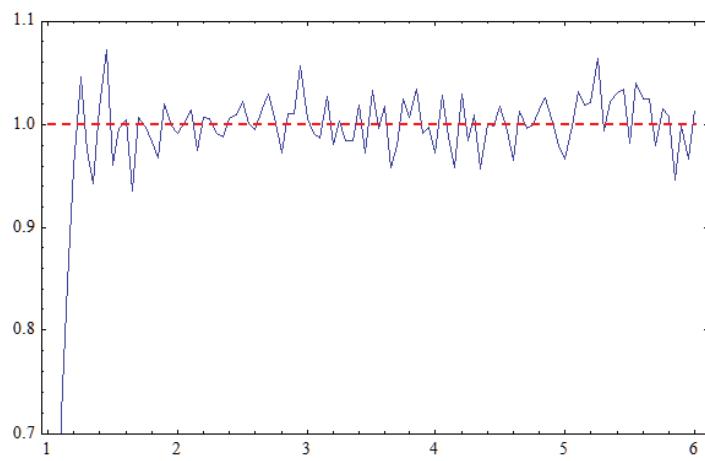


圖 3-4-1 重要抽樣法估計風險 R_1

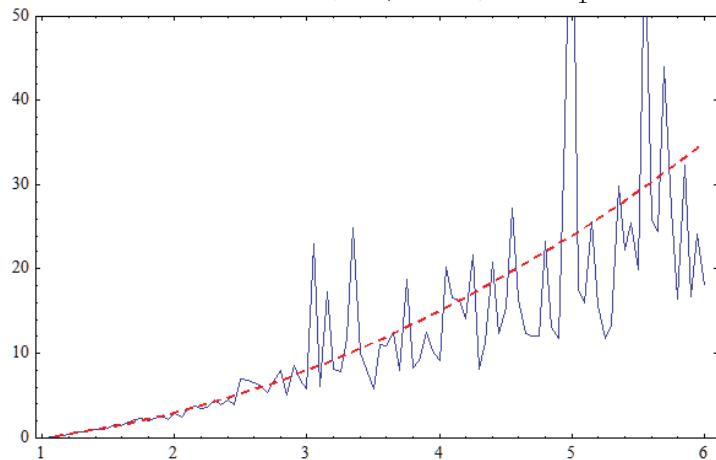


圖 3-4-2 重要抽樣法估計風險 R_2

卡羅法估算，則可以模擬 $Z_i \sim \mathcal{N}(0, 1)$, $i = 1, \dots, M$ ，並計算

$$P(Z > 4.5) \approx \frac{1}{M} \sum_{i=1}^M I(Z^{(i)} > 4.5)$$

但結果是即使樣本數已達 $M = 10000$ ，所得的 $I(Z^{(i)} > 4.5)$ 幾乎都是 0，不容易真正估計出想要的微小機率值。改用重要抽樣法可以大大提高此估計的準確性，作法如下：

令 $Y \sim \mathcal{TE}(4.5, 1)$ (截斷指數分佈)，其密度函數為

$$\begin{aligned} f_Y(y) &= e^{-(y-4.5)}, \quad y \geq 4.5 \\ &= e^{-y} / \int_{4.5}^{\infty} e^{-x} dx, \quad y \geq 4.5 \end{aligned}$$

從 f_Y 中抽樣，並使用重要抽樣法可得

$$\begin{aligned} P(Z > 4.5) &= E[I(Y > 4.5)] \\ &= \int_{4.5}^{\infty} I(Y > 4.5) \frac{\varphi(y)}{f_Y(y)} f_Y(y) dy \quad (\varphi(y) \text{ 是標準常態分佈的密度函數}) \\ &\approx \frac{1}{M} \sum_{i=1}^M \frac{\varphi(Y^{(i)})}{f_Y(Y^{(i)})} I(Y^{(i)} > 4.5) \\ &= 0.000003377 \end{aligned}$$

3.3.2 變異數有限之估計量(Finite Variance Estimators)

重要抽樣法中對函數 g 的選擇幾乎沒有限制，但使用上還是有某些特別的函數 g 是優於其他函數的。對於第(3.4)式的估計，本節想比較不同的函數 g 對變異數的影響。當第(3.9)式的 $\frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$ 確實收斂至第(3.4)式的 $E_f[h(X)]$ ，若其二階動差有限，即

$$\mathbb{E}_g[h^2(X) \frac{f^2(X)}{g^2(X)}] = \mathbb{E}_f[h^2(X) \frac{f(X)}{g(X)}] = \int_X h^2(X) \frac{f^2(X)}{g(X)} dx < \infty \quad (3.10)$$

則其變異數也為有限值。為了使上述二階動差有限，工具函數 g 的選擇必須比目標函數 f 厚尾，以確保 f/g 有界且不會導致 $\mathbb{E}_f[h^2 f/g]$ 發散。特別地，Geweke (1989) 提出兩個充分條件：

- 一、 $f(x)/g(x) < M \quad \forall x \in \mathcal{X}$ 且 $\text{var}_f(h) < \infty$
- 二、支集 \mathcal{X} 為緊緻集合(compact)， $f(x) < F$ 且 $g(x) > \epsilon \quad \forall x \in \mathcal{X}$

將第(3.9)式的估計量之 m 改成權重和 $\sum_{j=1}^m f(x_j)/g(x_j)$ ，可得到另一個滿足變異數有限且更平穩的估計量為

$$\frac{\sum_{j=1}^m h(x_j)f(x_j)/g(x_j)}{\sum_{j=1}^m f(x_j)/g(x_j)}, \quad (3.11)$$

因為當 $m \rightarrow \infty$ ， $(1/m)\sum_{j=1}^m f(x_j)/g(x_j) \rightarrow 1$ ，所以由強大數法則可保證此估計量也收斂至 $E_f[h(X)]$ 。雖然此加權估計量(第(3.11)式)是偏的，但偏量不多，反而是能降低變異數的優點使它更受重視。若給定 h 及固定目標函數 f ，則欲達到變異數有限的眾多工具函數 g 的選擇中，存在 g 最佳化的結果。下面定理介紹的是由 Rubinstein (1981) 提出的結果：

定理3.12

令 $g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(x)|f(x)dx}$ ，則 $g^*(x)$ 將使估計量 $\frac{1}{m} \sum_{j=1}^m \frac{f(X_j)}{g(X_j)} h(X_j)$ 的變異數達到最小值。

證明：

已知

$$Var \left[\frac{h(X)f(X)}{g(X)} \right] = \mathbb{E}_g \left[\frac{h^2(X)f^2(X)}{g^2(X)} \right] - \left(\mathbb{E}_g \left[\frac{h(X)f(X)}{g(X)} \right] \right)^2$$

等式右邊第二項 $\mathbb{E}_g \left[\frac{h(X)f(X)}{g(X)} \right] = \int_{\mathcal{X}} \frac{h(x)f(x)}{g(x)} g(x) dx = \int_{\mathcal{X}} h(x)f(x) dx$ 此結果與 g 無關，欲找到使變異數有最小值的函數 g ，只需把重點放在等式右邊第一項。

由 Jensen 不等式(Jensen's inequality $E(\varphi(X)) \geq \varphi[E(X)]$)，其中 φ 是個凸函數)可得

$$\mathbb{E}_g \left[\frac{h^2(X)f^2(X)}{g^2(X)} \right] \geq \left(\mathbb{E}_g \left[\frac{|h(X)|f(X)}{g(X)} \right] \right)^2 = \left(\int |h(x)|f(x) dx \right)^2$$

此下界可看出變異數的最小值與函數 g 的選擇無關。

又等號成立於

$$\frac{h(x)f(x)}{g(x)} = E \left[\frac{h(X)f(X)}{g(X)} \right] = \int_{\mathcal{X}} h(x)f(x) dx$$

由此可解出使變異數有最小值的 g 為

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(x)|f(x) dx}$$

此最佳化的結果相當正規化，但是當 $h(x) > 0$ 時， g^* 中會含有我們原本想知道的未知積分 $\int_{\mathcal{X}} h(x)f(x) dx$ ，以致於 g^* 並不好用。將第(3.10)式配合定理 3.12 的優點，可得到一個較實用的估計量為

$$\frac{\sum_{j=1}^m h(x_j)f(x_j)/g(x_j)}{\sum_{j=1}^m f(x_j)/g(x_j)}$$

$$\begin{aligned}
&= \frac{\sum_{j=1}^m \left[h(x_j) f(x_j) / \int_{\mathcal{X}} |h(x)| f(x) dx \right]}{\sum_{j=1}^m \left[f(x_j) / \int_{\mathcal{X}} |h(x)| f(x) dx \right]} \\
&= \frac{\sum_{j=1}^m h(x_j) |h(x_j)|^{-1}}{\sum_{j=1}^m |h(x_j)|^{-1}}
\end{aligned} \tag{3.12}$$

其中 $x_i \sim g \propto |h|f$ ，可注意到的是分子部分為 $h(x_j)$ 正的次數與負的次數抵消後的個數，當 $h(x)$ 恒正，則第(3.12)式正好為調和平均數(harmonic mean)。

雖然變異數有限的限制對於第(3.9)、(3.12)式的收斂與否不是必須的條件，但是由第(3.10)式可知道當

$$\int \frac{f^2(x)}{g(x)} dx = +\infty \tag{3.13}$$

重要抽樣法的表現將十分不理想，因此會產生上述情況的函數 g 是不被建議的。以下兩個例題要說明的是，適當地使用重要抽樣法可以帶來比蒙地卡羅法更好的結果，但若變異數有限的條件未滿足，則可能導致一個非常不理想的估計結果。

例題 3.13 t 分佈

令 $X \sim \mathcal{T}(\nu, \theta, \sigma^2)$ ，其密度函數為

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sigma \sqrt{\nu \pi} \Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu \sigma^2}\right)^{-(\nu+1)/2}$$

此處我們取 $\theta = 0$ 、 $\sigma = 1$ ，感興趣的量為 $E_f[h_i(X)]$, $i = 1, 2, 3$, $h_i(X)$ 如下：

$$h_1(x) = \sqrt{\left|\frac{x}{1-x}\right|}, \quad h_2(x) = x^5 I_{[2,1,\infty](x)}, \quad h_3(x) = \frac{x^5}{1+(x-3)^2} I_{x \geq 0}$$

對於工具密度函數 g 我們選擇兩種函數來做比較，一個是柯西分佈 $\mathcal{C}(0, 1)$ ，另一個為常態分佈 $\mathcal{N}(0, \nu/(\nu-2))$ 。後者因為造成下列積分式發散

$$\int_{-\infty}^{\infty} \frac{f^2(x)}{g(x)} dx \propto \int_{-\infty}^{\infty} \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}} dx$$

所以並不是個好的選擇。而前者柯西分佈較 f 厚尾，且保證 $Var(f/g)$ 是有限的，故 g 選擇柯西分佈較佳。

圖 3-5 由左而右顯示分別從目標函數 f 本身、從柯西分佈及從常態分佈中抽樣，此處參數 $\nu = 12$ 、疊代次數(樣本數)為 2000、重複次數 500 次，三張圖分別呈現估計 $E_f[h_1(X)]$ 所得結果。圖中所顯示的平均值都相當平穩，可見平均值並不受重要

抽樣法所選擇的函數 g 而影響。但以函數值分佈的範圍來看，此三張圖均有明顯的大跳動，此現象主要是因為 $x = 1$ 是函數 $h_1(x)$ 的奇異點(singularity)，以致於不僅使 $h_1^2(x)$ 對 f 是無法積分的，也導致無論是由柯西分佈或常態分佈抽樣，所得的估計量的變異數均不是有限的。

為改善上述問題，可以刻意地為 $h_1(x)$ 設計函數 g ，選擇適當的 g 使得 $(1 - x)g(x)$ 在 $x = 1$ 處的表現能更好。例如使用雙重伽瑪分佈(double gamma distribution)摺疊於 $x = 1$ ，也就是說 X 的分佈對稱於 $x = 1$ ，使得

$$|X - 1| \sim \text{Gamma}(\alpha, 1)$$

則當 $\alpha < 1$ 時，比例

$$h_1^2(x) \frac{f^2(x)}{g(x)} \propto |x| f^2(x) |1 - x|^{1-\alpha-1} \exp |1 - x|$$

在 $x = 1$ 附近是可積分的。當然很顯然地，在此比例中指數部分在 $x \rightarrow \infty$ 時將產生問題並導致變異數為無限的情況，但是這對估計量的穩定性的影響較小。圖 3-6 為利用雙重伽瑪分佈 $Ga(0.5, 1)$ 做重要抽樣，樣本數為 2000、重複次數 500 次所得結果，相較圖 3-5，此結果顯示函數值震盪的幅度明顯緩和許多。

關於 $h_2(x) = x^5 I_{\{x \geq 2.1\}}$ ，令 $u = 1/x$ ，則 $0 < u < 1/2.1$ 且

$$E[h_2(X)] = \int_{-\infty}^{\infty} h_2(x) f(x) dx = \int_{2.1}^{\infty} x^5 I_{\{x \geq 2.1\}} f(x) dx = \int_0^{1/2.1} u^{-7} f(1/u) du$$

因此，工具函數可選擇一致分佈 $\mathcal{U}[0, 1/2.1]$ 的密度函數 $g(u) = 2.1$, $0 < u < 1/2.1$ ，並從 $\mathcal{U}(0, 1/2.1)$ 中抽樣，則此重要抽樣法所得的估計量為

$$\delta_2 = \frac{1}{2.1m} \sum_{j=1}^m U_j^{-7} f(1/U_j)$$

圖 3-7 中，點線、長虛線、短虛線、點線分別代表從目標函數 f 及工具函數 $\mathcal{U}(0, 1/2.1)$ 、 $\mathcal{C}(0, 1)$ 、 $N(0, 6/5)$ 中抽樣所得的結果。原期望值的真實值為 6.54，此四種函數所得期望值的估計值分別為 6.75, 6.48, 6.57, 7.06。可見抽樣自 $\mathcal{U}(0, 1/2.1)$ 的部分只需數百次疊代就可收斂到真實值，抽樣自 $\mathcal{C}(0, 1)$ 的部分也蠻平穩的需要疊代更多次數才能達到與 $\mathcal{U}(0, 1/2.1)$ 一樣的準確度。而另外兩個抽樣自目標函數 f 及 $N(0, 6/5)$ 的部分，因為其變異數是無限的，以致於其結果在真實值附近有劇烈變動。

至於 $h_3(x) = \frac{x^5}{1+(x-3)^2} I_{\{x \geq 0\}}$ ，因為

$$\begin{aligned} E[h_3(X)] &= \int_{-\infty}^{\infty} h_3(x)f(x)dx = \int_0^{\infty} \frac{x^5}{1+(x-3)^2}f(x)dx \\ &= \int_0^{\infty} \frac{x^5 e^x}{1+(x-3)^2}f(x)e^{-x}dx \end{aligned}$$

一個合理的工具函數為指數分佈 $\text{Exp}(1)$ 的密度函數 $g(x) = e^{-x}$, $0 < x < \infty$ ，從 $\text{Exp}(1)$ 中抽樣，則此重要抽樣法所得的估計量為

$$\delta_3 = \frac{1}{m} \sum_{j=1}^m h_3(X_j)w(X_j), \text{ 其中 } w(x) = f(x)/g(x) = f(x)e^x$$

圖 3-8 中，實線、短虛線、點線、長虛線分別代表從目標函數 f 及從工具函數 $\mathcal{C}(0, 1)$ 、 $N(0, 6/5)$ 、 $\text{Exp}(1)$ 中抽樣所得的結果。原期望值的真實值為 4.496，此四種函數所得期望值的估計值分別為 4.58, 4.42, 4.99, 4.52。可見 δ_3 是個不錯的估計量，其精確度不僅和使用原函數的情況相同，還比原函數所得結果平穩。取樣自 $\mathcal{C}(0, 1)$ 的結果也是平穩的，但其偏的情況需要較多次疊代才能消失。至於從常態分佈抽樣所得的估計量，因上下波動大嚴重影響其收斂結果。

例題3.14 轉移矩陣

考慮一個具兩種狀態，1, 2，的馬可夫鏈及其對應的轉移矩陣

$$T = \begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

其中

$$P(X_{t+1} = 1 | X_t = 1) = 1 - P(X_{t+1} = 2 | X_t = 1) = p_1$$

$$P(X_{t+1} = 2 | X_t = 2) = 1 - P(X_{t+1} = 1 | X_t = 2) = p_2$$

假設 p_1, p_2 滿足 $p_1 + p_2 < 1$ ，且樣本為 X_1, \dots, X_m ，

則 p_1, p_2 的先驗分佈為

$$\pi(p_1, p_2) = 2I_{\{p_1+p_2<1\}}$$

若 m_{ij} 表示由狀態 i 變成狀態 j 的次數，也就是說 $m_{ij} = \sum_{t=2}^m I_{\{x_t=i\}}I_{\{x_{t+1}=j\}}$ ，

則 p_1, p_2 的事後分佈為

$$\pi(p_1, p_2 | m_{11}, m_{12}, m_{21}, m_{22}) \propto p_1^{m_{11}}(1-p_1)^{m_{12}}p_2^{m_{22}}(1-p_2)^{m_{21}}I_{\{p_1+p_2<1\}}$$

如果我們感興趣的是機率 p_1, p_2 及優勝率 $\frac{p_1}{(1-p_1)}, \frac{p_2}{(1-p_2)}$ 的事後期望值，即 $h_1(p_1, p_2) = p_1, h_2(p_1, p_2) = p_2, h_3(p_1, p_2) = p_1/(1-p_1), h_4(p_1, p_2) = p_2/1-p_2, h_5(p_1, p_2) =$

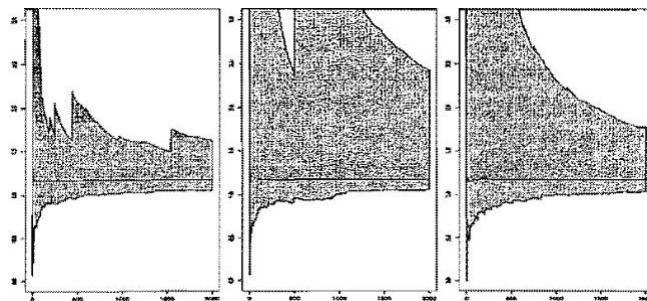


圖 3-5 重要抽樣法估計 $E_f[h_1(X)]$ 結果

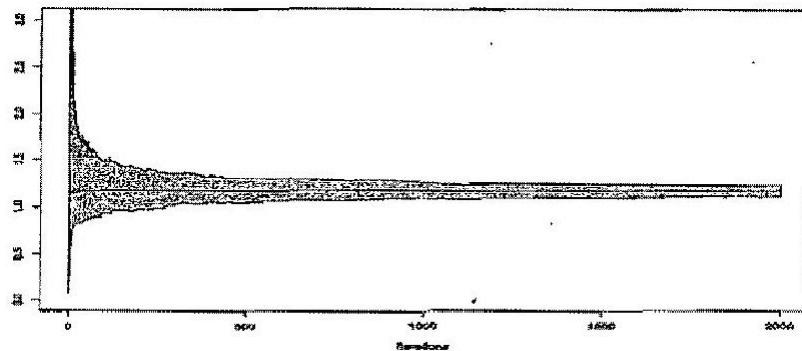


圖 3-6 雙重伽瑪分佈估計 $E_f[h_1(X)]$ 結果

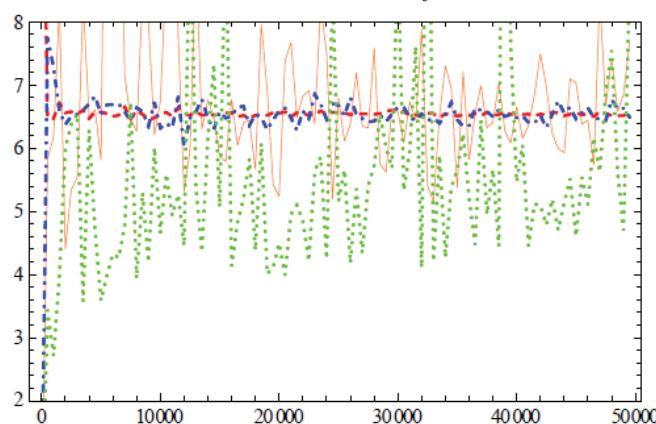


圖 3-7 重要抽樣法估計 $E_f[h_2(X)]$ 結果

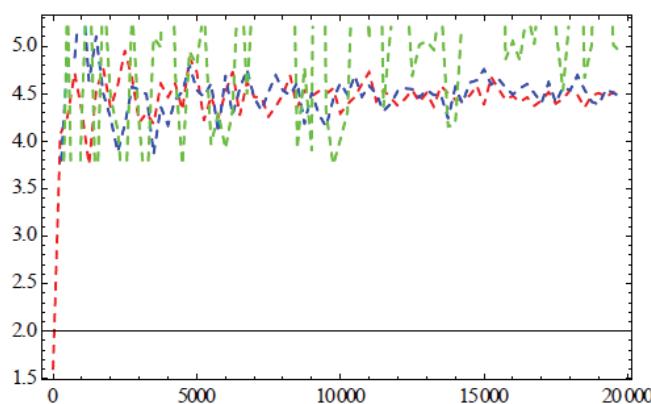


圖 3-8 重要抽樣法估計 $E_f[h_3(X)]$ 結果

$\log \left(\frac{p_1(1-p_2)}{p_2(1-p_1)} \right)$ ，以下我們來看幾種計算事後期望值的方法：

第一、由於此事後分佈 $\pi(p_1, p_2 | \mathcal{D})$ 可視為兩個具限制 $\{(p_1, p_2) : p_1 + p_2 < 1\}$ 之貝他分佈 $Be(m_{11} + 1, m_{12} + 1)$ 、 $Be(m_{22} + 1, m_{21} + 1)$ 的乘積。因此可由這兩個貝他分佈做模擬，直到抽出的兩樣本的和小於 1 為止。

第二、由事後分佈 $\pi(p_1, p_2 | \mathcal{D})$ 的形式可聯想到 Dirichlet 分佈， $D(m_{11} + 1, m_{12} + 1)$ ，其密度函數 $\pi_1(p_1, p_2 | \mathcal{D}) \propto p_1^{m_{11}} p_2^{m_{22}} (1 - p_1 - p_2)^{m_{12} + m_{21}}$ 。然而，因為 $\frac{\pi(p_1, p_2 | \mathcal{D})}{\pi_1(p_1, p_2 | \mathcal{D})}$ 並非有界，且相關的變異數也是無限大，所以這種作法並不算是理想的。

第三、Geweke (1989) 提出對二項式分佈使用常態逼近法，也就是

$$\begin{aligned}\pi_2(p_1, p_2 | \mathcal{D}) &\propto \exp\{-(m_{11} + m_{12})(p_1 - \hat{p}_1)^2 / 2\hat{p}_1(1 - \hat{p}_1)\} \\ &\times \exp\{-(m_{21} + m_{22})(p_2 - \hat{p}_2)^2 / 2\hat{p}_2(1 - \hat{p}_2)\} I_{\{p_1 + p_2 < 1\}}\end{aligned}$$

其中 $\hat{p}_i = m_{ii}/(m_{ii} + m_{i(3-i)})$ 是 p_i 的最大概似估計量。

Geweke (1991) 及 Robert (1995) 提出一個有效率的模擬 π_2 的方法是從 $N(\hat{p}_1, \hat{p}_1(1 - \hat{p}_1)/(m_{12} + m_{11}))$ 生成 p_1 ，並要求 $p_1 \in [0, 1]$ ，再從 $N(\hat{p}_2, \hat{p}_2(1 - \hat{p}_2)/(m_{21} + m_{22}))$ 生成 p_2 ，並限制 $p_2 \in [0, 1 - p_1]$ 。因為有 $\{(p_1, p_2) : p_1 + p_2 < 1\}$ 的限制，所以 $E_\pi[\pi(p_1, p_2 | \mathcal{D})/\pi_2(p_1, p_2 | \mathcal{D})]$ 是有限的。

第四、另一種可能的作法是依然使用 $Be(m_{11} + 1, m_{12} + 1)$ 做為 p_1 的邊際分佈， p_2 的條件分佈原本為 $p_2^{m_{22}}(1 - p_2)^{m_{21}} I_{[p_2 < 1 - p_1]}$ 修改為

$$\pi_3(p_2 | p_1, \mathcal{D}) \propto \frac{2}{(1 - p_1)^2} p_2 I_{\{p_2 < 1 - p_1\}}$$

則重要抽樣法中的比值(權重)

$$\begin{aligned}w(p_1, p_2) &= \pi(p_1, p_2 | \mathcal{D}) / \pi(p_1 | \mathcal{D}) \pi_3(p_2 | p_1, \mathcal{D}) \\ &\propto \frac{p_1^{m_{11}} (1 - p_1)^{m_{12}} p_2^{m_{22}} (1 - p_2)^{m_{21}} I_{\{p_1 + p_2 < 1\}}}{p_1^{m_{11}} (1 - p_1)^{m_{12}} \frac{2}{(1 - p_1)^2} p_2 I_{\{p_2 < 1 - p_1\}}} \\ &= p_2^{m_{22}-1} (1 - p_2)^{m_{21}} (1 - p_1)^2\end{aligned}$$

此結果對 (p_1, p_2) 而言是有界的。

表 3-4

分佈	h_1	h_2	h_3	h_4	h_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

表 3-4 是 h_j 分別對原目標函數 π 及 π_1, π_2, π_3 所得的事後期望值之估計值，相較於 π_1, π_2, π_3 的表現與原目標函數 π 較接近且所需要的模擬次數較少，而 π_1 的表現是最不理想的。

圖 3-9 為利用第(11)式估計 $E[h_5]$ 的結果。圖中 π_1 的跳動狀況為重要抽樣法的估計量在變異數無限時的主要特徵。承第(13)式，若 $E_f[f(X)/g(X)] = \infty$ ，則必須耗費龐大的模擬次數來達到第(9)式的收斂結果，這並不是個有效率的作法。因此，在使用重要抽樣法時，對於工具密度函數的選擇是需要特別注意的。

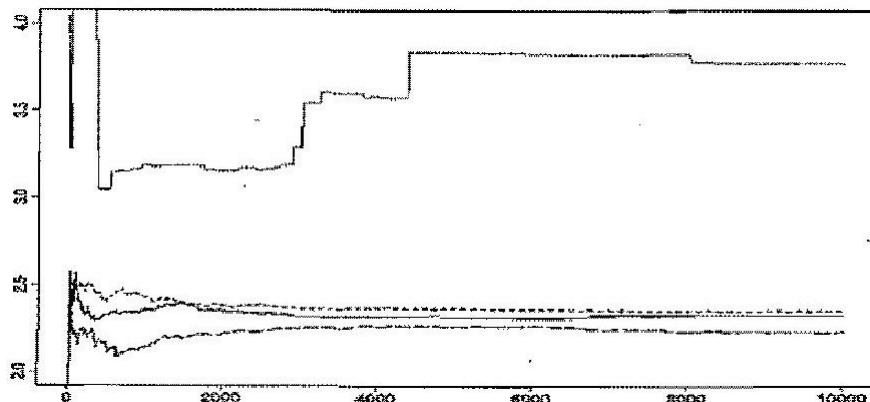


圖 3-9 π : 實線, 收斂值 2.373; π_1 : 點線, 收斂值 3.184;
 π_2 : 長虛線, 收斂值 2.319; π_3 : 短虛線, 收斂值 2.379。

3.3.3 重要抽樣與接受拒絕法比較(Import. Sampling vs. AR method)

第 3.3.2 節中介紹的定理 3.12，因為最佳解 g^* 的內容與我們感興趣的函數 h 有關，這點導致在面對不同函數 h 時，不僅計算時間會明顯地增加，生成的樣本也無法重複被使用。若函數 g 滿足 $f(x) \leq Mg(x)$, $1 < M < \infty$ ，則 g 不僅試用於接受拒絕法，也可以做為重要抽樣法的工具函數。此處沿用函數 g 還有另一個優點，因為 f/g 是有界的，這點可以保證重要抽樣法之估計量具有有限的變異數。

接受拒絕法與重要抽樣法的原始估計量分別為

$$\delta_1 = \frac{1}{n} \sum_{i=1}^n h(X_i) \quad , \quad \delta_2 = \frac{1}{t} \sum_{j=1}^t h(Y_j) \frac{f(Y_j)}{g(Y_j)} \quad (3.14)$$

其中 $\{Y_1, \dots, Y_t\}$ 是從函數 g 中生成的所有樣本，而 $\{X_1, \dots, X_n\} (\subseteq \{Y_1, \dots, Y_t\})$ 為接受拒絕演算法所接受的樣本。若 f/g 只能由某個常數來決定， δ_2 可以修改為

$$\delta_3 = \sum_{j=1}^t h(Y_j) \frac{f(Y_j)}{g(Y_j)} / \sum_{j=1}^t \frac{f(Y_j)}{g(Y_j)}$$

我們將 δ_2 仔細寫成

$$\delta_2 = \frac{n}{t} \left\{ \frac{1}{n} \sum_{i=1}^n h(X_i) \frac{f(X_i)}{g(X_i)} + \frac{t-n}{n} \frac{1}{t-n} \sum_{j=1}^{t-n} h(Z_j) \frac{f(Z_j)}{g(Z_j)} \right\}$$

其中 $\{Y_1, \dots, Y_t\} = \{X_1, \dots, X_n\} \cup \{Z_1, \dots, Z_{t-n}\}$ ， $\{Z_1, \dots, Z_{t-n}\}$ 為接受拒絕演算法所拒絕的樣本。此處需注意 t 是個隨機變數（是接受拒絕演算法的止步規則(stopping rule)）， t 的分佈是負二項式分佈 $\text{Neg}(n, 1/M)$ ，所以 (Y_1, \dots, Y_t) 並不是一組完全由函數 g 決定的 iid 樣本。基於上述理由，對 (Y_1, \dots, Y_t) 的分佈而言， δ_2 有著不正確的表示式。因為 (Y_1, \dots, Y_t) 中被接受的樣本 $\{X_1, \dots, X_n\}$ 可視為與函數 f 具相同的分佈，而被拒絕的樣本 $\{Z_1, \dots, Z_{t-n}\}$ ，其分佈應修正為

$$\begin{aligned} F_Z(z) &= P\left(Z \leq z | U > \frac{f(z)}{Mg(z)}\right) = \frac{\int_{-\infty}^z \int_{\frac{f(y)}{Mg(y)}}^1 g(y) \cdot 1 du dy}{\int_{-\infty}^{\infty} \int_{\frac{f(y)}{Mg(y)}}^1 g(y) \cdot 1 du dy} \\ &= \frac{\int_{-\infty}^z (g(y) - f(y)/M) dy}{1 - 1/M} \end{aligned}$$

Z'_i s 的密度函數為 $dF_Z(z)/dz = \frac{g(z) - f(z)/M}{1 - 1/M} = \frac{Mg(z) - f(z)}{M-1}$ 。因此，對 δ_2 一個合理的修改如下：

$$\begin{aligned} \delta_4 &= \frac{n}{t} \left[\frac{1}{n} \sum_{i=1}^n h(X_i) + \frac{t-n}{n} \frac{1}{t-n} \sum_{j=1}^{t-n} h(Z_j) \frac{f(Z_j)}{\frac{Mg(Z_j) - f(Z_j)}{M-1}} \right] \\ &= \frac{n}{t} \delta_1 + \frac{1}{t} \sum_{j=1}^{t-n} h(Z_j) \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)} \end{aligned} \quad (3.15)$$

上述 δ_4 也是個不偏估計量。以下我們試著表示 $n = 1$ 時 δ_1 及 δ_4 的變異數，先將 δ_4 寫成

$$\delta_4 = \frac{1}{t} h(X_1) + \frac{1}{t} \sum_{j=1}^{t-1} h(Z_j) \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}$$

假設 $\mathbb{E}_f[h(X)] = 0$ ，則 $\text{var}(\delta_1) = \frac{1}{n}\mathbb{E}_f[h^2(X)]$ 且

$$\begin{aligned} \text{var}(\delta_4) &= \mathbb{E}_t \left[\frac{1}{t^2} \mathbb{E}_f[h^2(X)] + \frac{t-1}{t^2} \mathbb{E}_f[h^2(Z)] \left(\frac{(M-1)f(Z)}{Mg(Z) - f(Z)} \right)^2 \right] \\ &= \mathbb{E}_t \left[\frac{t-1}{t^2} \int h^2(x) \left[\frac{(M-1)f(x)}{Mg(x) - f(x)} \right]^2 \frac{Mg(x) - f(x)}{(M-1)} dx + \frac{1}{t^2} \mathbb{E}_f[h^2(X)] \right] \\ &= \mathbb{E}_t \left[\frac{t-1}{t^2} \int h^2(x) \frac{f^2(x)(M-1)}{Mg(x) - f(x)} dx \right] + \mathbb{E}_t \left[\frac{1}{t^2} \mathbb{E}_f[h^2(X)] \right] \\ &= \mathbb{E}_t \left[\frac{t-1}{t^2} \int h^2(x) \frac{f^2(x)(M-1)}{Mg(x) - f(x)} dx \right] + \mathbb{E}_f[h^2(X)] \mathbb{E}_t\left(\frac{1}{t^2}\right) \\ &= \mathbb{E}_t \left[\frac{t-1}{t^2} \int h^2(x) \frac{f^2(x)(M-1)}{Mg(x) - f(x)} dx \right] + n \mathbb{E}_t\left(\frac{1}{t^2}\right) \text{var}(\delta_1) \end{aligned}$$

此變異數形式與函數 f, g, h 相關，以致於不容易比較 $\text{var}(\delta_1)$ 和 $\text{var}(\delta_4)$ 的大小關係。

若我們使用 Z'_i s 的密度函數， $\frac{Mg(z)-f(z)}{M-1}$ ，做為重要抽樣法的工具函數，則重要抽樣法的估計量可表示成

$$\delta_5 = \frac{1}{t-n} \sum_{j=1}^{t-n} h(Z_j) \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}$$

且 $\delta_4 = \frac{n}{t}\delta_1 + \frac{t-n}{t}\delta_5$ ，可見 δ_4 可視為一般蒙地卡羅法的估計量與 δ_5 的加權平均。

此外，若我們僅知道 f 取決於某個積分常數，則 δ_4 可以再改成

$$\delta_6 = \frac{n}{t}\delta_1 + \frac{t-n}{t} \sum_{j=1}^{t-n} \frac{h(Z_j)f(Z_j)}{Mg(Z_j) - f(Z_j)} / \sum_{j=1}^{t-n} \frac{f(Z_j)}{Mg(Z_j) - f(Z_j)} \quad (3.16)$$

例題3.15 伽瑪分佈的模擬

欲模擬 $Ga(\alpha, \beta)$ ，此處使用的工具函數為 $Ga(a, b)$, $a = [\alpha]$, $b = a\beta/\alpha$ 的密度函數，其中 b 的選擇是來自例題 2.19 所得到的能使接受拒絕法的接受機率達到最大值的特殊取法，而 $f(x)/g(x) = w(x) = \frac{\Gamma(a)}{\Gamma(\alpha)} \frac{\beta^\alpha}{b^a} x^{\alpha-a} e^{b-\beta} x$ ，對 x 微分並令微分結果為零，可解得當 $x = \frac{\alpha-a}{\beta-b}$ 時，權重 $w(x)$ 有最大值為

$$\begin{aligned}
M &= \frac{\Gamma(a)}{\Gamma(\alpha)} \frac{\beta^\alpha}{b^a} \left(\frac{\alpha - a}{\beta - b} \right)^{\alpha-a} e^{-(\alpha-a)} \\
&= \frac{\Gamma(a)}{\Gamma(\alpha)} \exp\{\alpha(\log(\alpha) - 1) - a(\log(a) - 1)\}
\end{aligned} \tag{3.17}$$

因為 $a = [\alpha]$ ， $\frac{\Gamma(a)}{\Gamma(\alpha)} \leq 1$ ，所以在模擬時權重 $w(x)$ 的上界的近似值可取成

$$M' = \exp\{\alpha(\log(\alpha) - 1) - a(\log(a) - 1)\}$$

其中 $M'/M = 1 + \epsilon = \frac{\Gamma(\alpha)}{\Gamma([\alpha])}$ 。

設感興趣的函數為

$$h_1(x) = x^3, \quad h_2(x) = x \log x, \quad \text{及} \quad h_3(x) = \frac{x}{1+x}$$

圖 3-10 描述的是當 $\alpha = 3.7$, $\beta = 1$ 的情況下(此時接受拒絕法的接受機率為 $1/M = 0.1$)，分別利用 δ_1 (實線)、 δ_4 (點線)、 δ_6 (虛線) 估計 $E[h_3(X)] = E[x/(1+x)]$ 的收斂結果。 δ_1 收斂到 0.7518， δ_4 收斂到 0.7497， δ_6 收斂到 0.7495 很明顯地， δ_4 、 δ_6 的收斂結果比起經驗平均值 δ_1 不僅平穩許多，收斂速度也快許多，特別是 δ_6 ，約經過 6000 次疊代就能收斂至真實值約為 0.7497。

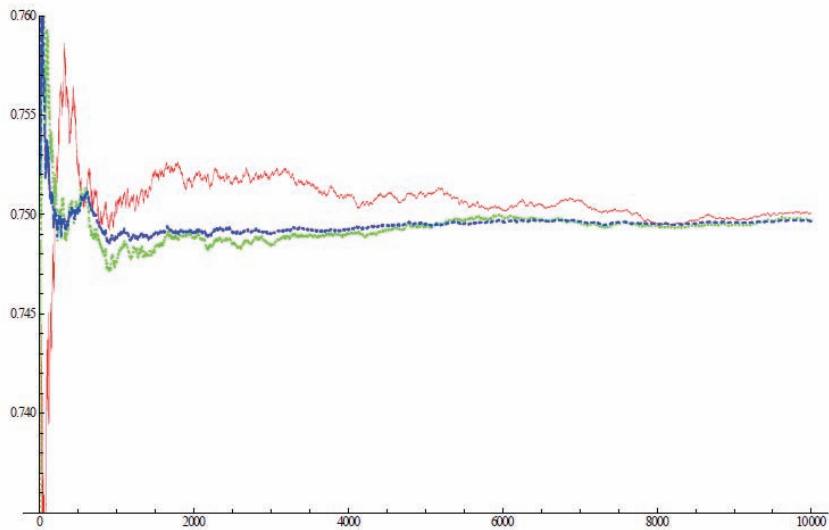


圖 3-10 估計 $E[h_3(X)] = E[x/(1+x)]$ 的收斂結果

3.4 拉普拉斯近似(Laplace Approximations)

假設我們感興趣的積分值為

$$\int_A f(x|\theta) dx \quad (3.18)$$

其中 θ 是固定的參數， f 必須是非負且可積分的函數。令 $f(x|\theta) = \exp\{nh(x|\theta)\}$ ，其中 n 可以是樣本個數或是個無界的參數。將 $h(x|\theta)$ 在 $x = x_0$ 處寫出其泰勒展開式

$$\begin{aligned} h(x|\theta) &\approx h(x_0|\theta) + (x - x_0)h'(x_0|\theta) + \frac{(x - x_0)^2}{2!}h''(x_0|\theta) \\ &\quad + \frac{(x - x_0)^3}{3!}h'''(x_0|\theta) + R_n(x) \end{aligned} \quad (3.19)$$

剩餘項(remainder) $R_n(x)$ 滿足 $\lim_{x \rightarrow x_0} R_n(x)/(x - x_0)^3 = 0$ 。令 $h'(x|\theta) = 0$ ，在給定 θ 下解出使 $h(x|\theta)$ 產生最大值的 x ， $x = \hat{x}_\theta$ ，若選擇 $x_0 = \hat{x}_\theta$ ，則第(3.19)式中的一次項 $h'(x_0|\theta) = 0$ ，且在以 \hat{x}_θ 為中心的附近鄰域(neighborhood)內，第(3.18)式可寫成

$$\begin{aligned} \int_A f(x|\theta) dx &= \int_A \exp\{nh(x|\theta)\} dx \\ &\approx e^{nh(\hat{x}_\theta|\theta)} \int_A \exp \left\{ n \left[\frac{(x - \hat{x}_\theta)^2}{2!}h''(\hat{x}_\theta|\theta) + \frac{(x - \hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) \right] \right\} dx \end{aligned}$$

回想一下 e^y 在 $y = 0$ 處的泰勒展開式為 $e^y \approx 1 + y + y^2/2!$ ，藉此若將上述積分式中的立方項 $\exp \left\{ n \frac{(x - \hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) \right\}$ 在 \hat{x}_θ 處寫出其泰勒展開式

$$\exp \left\{ n \frac{(x - \hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) \right\} \approx 1 + n \frac{(x - \hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) + n^2 \frac{(x - \hat{x}_\theta)^6}{2!(3!)^2} [h'''(\hat{x}_\theta|\theta)]^2$$

則原本第(3.18)式又可以再改寫成

$$\begin{aligned} \int_A f(x|\theta) dx &= \int_A \exp\{nh(x|\theta)\} dx \\ &\approx \exp\{nh(\hat{x}_\theta|\theta)\} \int_A \exp \left\{ n \frac{(x - \hat{x}_\theta)^2}{2!}h''(\hat{x}_\theta|\theta) \right\} \\ &\quad \times \left[1 + n \frac{(x - \hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) + n^2 \frac{(x - \hat{x}_\theta)^6}{2!(3!)^2} [h'''(\hat{x}_\theta|\theta)]^2 \right] dx \quad (3.20) \end{aligned}$$

取近似積分值時，由第(3.20)式右式的積分式由左而右被使用的項數，我們分別將此作法命名為一階近似、二階近似及三階近似。由第(3.20)式的一階近似

$$\int_A f(x|\theta) dx = \int_A \exp\{nh(x|\theta)\} dx \approx \exp\{nh(\hat{x}_\theta|\theta)\} \int_A \exp \left\{ n \frac{(x - \hat{x}_\theta)^2}{2!}h''(\hat{x}_\theta|\theta) \right\} dx$$

可發現此積分式與常態分佈 $\mathcal{N}(\hat{x}_\theta, -\frac{1}{nh''(\hat{x}_\theta|\theta)})$ 的密度函數之積分有關。因此，若 $A = [a, b]$ 、 $\Phi(\cdot)$ 表示標準常態分佈的累積分佈函數，則第 (3.20) 式的一階近似可以寫成

$$\begin{aligned}
 \int_A f(x|\theta) dx &= \int_A \exp\{nh(x|\theta)\} dx \\
 &\approx \exp\{nh(\hat{x}_\theta|\theta)\} \int_a^b \exp\left\{n\frac{(x-\hat{x}_\theta)^2}{2!}h''(\hat{x}_\theta|\theta)\right\} dx \\
 &= \exp\{nh(\hat{x}_\theta|\theta)\} \sqrt{\frac{2\pi}{-nh''(\hat{x}_\theta|\theta)}} \\
 &\quad \times \int_a^b \sqrt{\frac{-nh''(\hat{x}_\theta|\theta)}{2\pi}} \exp\left\{n\frac{(x-\hat{x}_\theta)^2}{2!}h''(\hat{x}_\theta|\theta)\right\} dx \\
 &= \exp\{nh(\hat{x}_\theta|\theta)\} \sqrt{\frac{2\pi}{-nh''(\hat{x}_\theta|\theta)}} \int_{\sqrt{-nh''(\hat{x}_\theta|\theta)}(a-\hat{x}_\theta)}^{\sqrt{-nh''(\hat{x}_\theta|\theta)}(b-\hat{x}_\theta)} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
 &= \exp\{nh(\hat{x}_\theta|\theta)\} \sqrt{\frac{2\pi}{-nh''(\hat{x}_\theta|\theta)}} \\
 &\quad \times \{\Phi[\sqrt{-nh''(\hat{x}_\theta|\theta)}(b-\hat{x}_\theta)] - \Phi[\sqrt{-nh''(\hat{x}_\theta|\theta)}(a-\hat{x}_\theta)]\} \quad (3.21)
 \end{aligned}$$

以下給一個拉普拉斯近似的簡單例子。

例題3.16 伽瑪的近似

考慮 Gamma 分佈 $Ga(\alpha, 1/\beta)$ 相關的積分

$$\int_a^b \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx \quad (3.22)$$

令 $h(x) = -x/\beta + (\alpha - 1) \log x$ ，則 $h(x)$ 在 x_0 處的二階泰勒展開式為

$$\begin{aligned}
 h(x) &\approx h(x_0) + (x - x_0)h'(x_0) + \frac{(x - x_0)^2}{2!}h''(x_0) \\
 &= -x_0/\beta + (\alpha - 1) \log x_0 + \left(\frac{\alpha - 1}{x_0} - 1/\beta\right)(x - x_0) - \frac{\alpha - 1}{2x_0^2}(x - x_0)^2
 \end{aligned}$$

令 $h'(x) = \frac{\alpha-1}{x} - 1/\beta = 0$ ，解出使 $h(x)$ 產生最大值的 $\hat{x}_\theta = (\alpha - 1)\beta$ ，並選擇 $x_0 = \hat{x}_\theta = (\alpha - 1)\beta$ ，則上述 $h(x)$ 的二階泰勒展開式可寫成

$$h(x) \approx -\hat{x}_\theta/\beta + (\alpha - 1) \log \hat{x}_\theta - \frac{\alpha - 1}{2\hat{x}_\theta^2}(x - \hat{x}_\theta)^2$$

若 $n = 1$ ，將上式代入第 (3.21) 式即可得到拉普拉斯近似結果如下

$$\int_a^b \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx$$

$$\approx \frac{1}{\Gamma(\alpha)\beta^\alpha} \hat{x}_\theta^{\alpha-1} e^{-\hat{x}_\theta/\beta} \sqrt{\frac{2\pi\hat{x}_\theta^2}{\alpha-1}} \times \left\{ \Phi \left[\sqrt{\frac{\alpha-1}{\hat{x}_\theta^2}}(b-\hat{x}_\theta) \right] - \Phi \left[\sqrt{\frac{\alpha-1}{\hat{x}_\theta^2}}(a-\hat{x}_\theta) \right] \right\}$$

當 $\alpha = 5$, $\beta = 2$, $\hat{x}_\theta = 8$ 時，可得到最佳的拉普拉斯近似結果。表 3-5 紀錄不同區間之積分值的近似結果，可看出在 $\hat{x}_\theta = 8$ 附近的積分值較準確，而尾端部分近似結果則不盡理想。

表 3-5

區間	近似值	實際值
(7, 9)	0.193351	0.193341
(6, 10)	0.375046	0.37477
(2, 14)	0.848559	0.823349
(15.981, ∞)	0.0224544	0.100005

4 蒙地卡羅最佳化(Monte Carlo Optimization)

本章內容與前面第三章的內容類似，主要目的在解決最佳化的問題。這裡我們要分辨兩種由電腦生成的隨機變數之使用方法：

第一、4.2節中，我們將提到如何製造隨機的技巧來求得函數的極值，並設計隨機探索的方法在函數表面避開局部極大值以求得最大值。

第二、4.3節中，類似第三章所介紹的，我們將利用一個近似的函數來做最佳化，並介紹此方法中常用的 EM(Expectation-Maximization) 演算法。

4.1 介紹(introduction)

類似前面章節看過的積分問題，及此處欲討論的最大化問題

$$\max_{\theta \in \Theta} h(\theta) \quad (4.1)$$

一般的解決方法為數值方法或模擬方法，兩方法的差別在於對函數 h 的處理方法不同。數值方法中，目標函數在分析上的性質，如：凸性convexity、邊界boundedness、平滑性smoothness，通常是很重要的；而模擬方法中，我們主要是以機率的觀點來看函數 h ，此處理方法有別於用分析的方法。雖然數值方法比模擬方法擁有更長久的歷史，但模擬方法因為其對定義域 Θ 及函數 h 的限制較寬鬆，特別是在遇到 h 不容易計算時，模擬方法的優點更顯現其方便性，因此而備受矚目。

例題4.1 信號處理(Signal processing)

一個信號處理相關資料的模型為

$$x_i = \alpha_1 \cos(\omega t_i) + \alpha_2 \sin(\omega t_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, N.$$

其中 $\alpha = (\alpha_1, \alpha_2)$, ω 及 σ 為未知參數，而 t_1, \dots, t_N 為觀察的時間。

又 $\mathbf{x}^t = (x_1, \dots, x_N)$ 的概似函數為

$$L(\alpha, \omega, \sigma | \mathbf{x}) \propto \sigma^{-N} \exp\left(-\frac{(\mathbf{x} - G\alpha^t)^t(\mathbf{x} - G\alpha^t)}{2\sigma^2}\right),$$

此處 $G = \begin{pmatrix} \cos(\omega t_1) & \sin(\omega t_1) \\ \vdots & \vdots \\ \cos(\omega t_N) & \sin(\omega t_N) \end{pmatrix}$ 。若 α, ω, σ 的先驗分佈為 $\pi(\alpha, \omega, \sigma) = \sigma^{-1}$ ，則 ω 的事後邊際分佈為

$$\begin{aligned} \pi(\omega | \mathbf{x}) &= \frac{\int_{\sigma} \int_{\alpha} \pi(\alpha, \omega, \sigma) f(\mathbf{x} | \alpha, \omega, \sigma) d\alpha d\sigma}{\int_{\omega} \int_{\sigma} \int_{\alpha} \pi(\alpha, \omega, \sigma) f(\mathbf{x} | \alpha, \omega, \sigma) d\alpha d\sigma d\omega} \\ &\propto \int_{\sigma} \int_{\alpha} \pi(\alpha, \omega, \sigma) f(\mathbf{x} | \alpha, \omega, \sigma) d\alpha d\sigma \\ &\propto (x^t x - x^t G (G^t G)^{-1} G^t x)^{(2-N)/2} (\det G^t G)^{-1/2} \end{aligned} \quad (4.2)$$

上述積分式的結果((4.2)式)在計算上並不是件容易的事。此外， ω 的事後邊際分佈是個具多重局部極值的函數，因此也不容易找到絕對極值正確的位置。

接下來我們想分辨兩種蒙地卡羅最佳化方法：

第一、探索性方法，蒙地卡羅的觀點傾向於由整個定義域 Θ 來將函數 h 最佳化，而函數本身的性質則不是那麼重要。

第二、以機率上的近似來對函數 h 做最佳化，此觀點提供了一個可接受的近似結果，並可省去分析定義域 Θ 的動作。後面章節我們將看到此方法與遺失資料方法(如：EM演算法)有密切的關係。

4.2 隨機探索(Stochastic Exploration)

4.2.1 基本解(A Basic Solution)

探索方法適用於許多情況。若 Θ 是有界的(有時可透過重新參數化而達到有界的特性)，第一種方法是生成 $U_1, \dots, U_m \sim \mathcal{U}_{\Theta}$ (均勻分佈)，並透過 $h_m^* = \max(h(U_1), \dots, h(U_m))$ 來求得(4.1)式的結果。此方法在 m 趨近於無限大時是會收斂的，但收斂速度過慢，因為它並未考慮到 h 本身的任何特性。若能選擇其他除了均勻分佈以外較接近 h 的分佈，將

可以改善此問題。

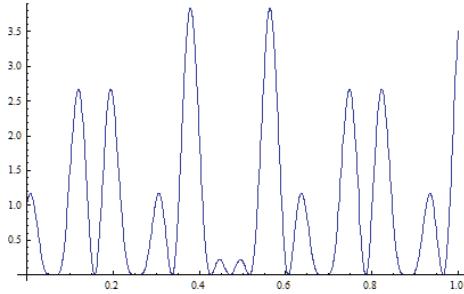


圖 4-1-1 $h(x)$ 的函數圖

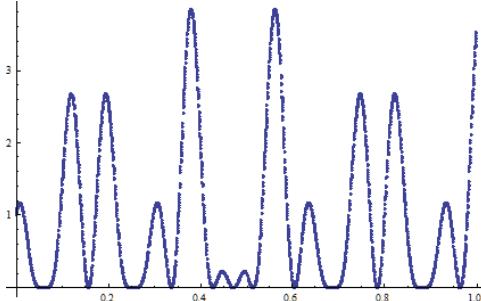


圖 4-1-2 利用 5000 筆 $\mathcal{U}[0, 1]$ 樣本計算 $h(x)$ 的結果

例題4.2 第一型蒙地卡羅最大值(A first Monte Carlo maximization)

回顧例題 3.4， $h(x) = [\cos(50x) + \sin(20x)]^2$, $x \in [0, 1]$ ，此函數的定義域是有界的，我們仿照上述方法模擬 $U_1, \dots, U_m \sim \mathcal{U}[0, 1]$ ，由近似值 $h_m^* = \max(h(U_1), \dots, h(U_m))$ 來找真正的最大值。結果如圖 4-1 所示。由圖可見模擬的函數圖形與真實的函數圖形非常接近，且得到的蒙地卡羅最大值為 3.832 也幾乎等於經煩瑣計算而得到的真實最大值。當然，這題是個小問題，因此第一種處理方法是可行的，但對於其他較複雜或維度較高的函數而言，這可能就不太適用了。

為了解決較複雜的問題，我們要介紹第二種較有成效的方法，此方法將 h 與某些機率分佈做連結，例如若 $h(\theta) > 0$ 且 $\int_{\Theta} h(\theta) d\theta < +\infty$ ，則 (4.1) 式等同於求分佈函數 h 的眾數(modes)。一般而言，若 h 沒滿足前述條件，我們可以將函數 $h(\theta)$ 轉換成 $H(\theta)$ ，而此 $H(\theta)$ 需滿足

(i) H 是非負函數且 $\int H < \infty$ 、(ii)使 $H(\theta)$ 有最大值的解即為 (4.1) 式的解。

例如：可取 $H(\theta) = \exp(h(\theta)/T)$ 或 $H(\theta) = \exp(h(\theta)/T)/(1 + \exp(h(\theta)/T))$ ，其中 T 的選擇可加速收斂速度或避免產生局部極大值(見 4.2.3 節)。若此問題被表示成統計相關形式，則很自然地我們將從 h 或 H 中生成 $\theta_1, \dots, \theta_m$ ，並引用標準的眾數估計方法來解決問題(在某些例子中，我們可以將 $h(\theta)$ 分解成 $h(\theta) = h_1(\theta)h_2(\theta)$ 並利用 $h_1(\theta)$ 來做模擬)。

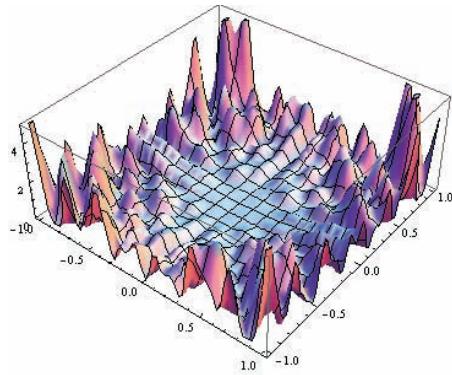


圖 4-2 $h(x, y)$ 的函數圖

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x)x) \\ + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y)$$

例題4.3 最小化複合函數(Minimization of a complex function)

在 R^2 上我們考慮一函數

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x)x) \\ + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y)$$

此函數在 $(x, y) = (0, 0)$ 處會產生最小值為 0。此函數有許多局部極小值，如圖 4-2，以致於一般的探索作法不容易找到絕對最小值發生處。我們可以從另一方面來看此問題，原函數 $h(x, y)$ 正比於 $\exp(-h(x, y))$ ，若將 $\exp(-h(x, y))$ 關於 \cosh, \sinh 的部分省略，得到一個形式再簡單一點的函數 $h_1(x, y) = \exp\{-(x \sin(20y) + y \sin(20x))^2 - (x \cos(10y) - y \sin(10x))^2\}$ ，因為 $h_1(x, y)$ 的分佈是可以被模擬的，所以可結合蒙地卡羅方法和 $h(x, y)$ 之最小值的收斂逼近法來解決此問題。

探索的作法特別是在 Θ 不是凸集合或不是連通集合時會更加困難。在這種情形下，透過模擬一組樣本 $(\theta_1, \dots, \theta_m)$ 來處理 (4.1) 式的問題，比起以數值方法來做會快速許多。當 h 可被表示成 $h(\theta) = \int H(x, \theta) dx$ 時，模擬方法的吸引力將更明顯。若 $H(x, \theta)$ 是一個分佈函數且可以被模擬，則 (4.1) 式的解就是 θ 之邊際機率下的眾數(mode)。以下我們來看探索方法中求最大值的方法。

4.2.2 梯度法(Gradient Methods)

前面 1.4 節所提梯度法對於 (4.1) 式的問題而言，是一種確定性的數值方法。若定義域 $\Theta \subset \mathcal{R}^d$ 且函數 $(-h)$ 是凸函數，則此方法可以產生一個能收斂到 (4.1) 式之真

值(θ^*)的序列(θ_j)，此序列彼此間有一個遞迴關係

$$\theta_{j+1} = \theta_j + \alpha_j \nabla h(\theta_j), \alpha_j > 0 \quad (4.3)$$

其中 ∇h 是函數 h 的梯度。對於不同的 α_j 序列，此演算法都能收斂到唯一的最大值。在更一般化的過程中，(4.3)式可加入隨機擾動(stochastic perturbations)來做修正，其中一種方式是選擇第二個序列 β_j ，並定義

$$\theta_{j+1} = \theta_j + \frac{\alpha_j}{2\beta_j} \Delta h(\theta_j, \beta_j \zeta_j) \zeta_j \quad (4.4)$$

其中， ζ_j 是在單位圓 $||\zeta|| = 1$ 上屬於均勻分佈的隨機變數，

$$\Delta h(x, y) = h(x + y) - h(x - y) \approx 2||y||\nabla h(x)$$

相對於(4.3)式所述方法，此做法不用沿著 θ_j 的斜率前進，因此能避免落入 h 的鞍點(saddle points)或局部極大值發生處。序列(θ_j)如何收斂到真值(θ^*)取決於(α_j)與(β_j)的選擇。若對 α_j 與 β_j 有足夠的條件限制，例如 α_j 需遞減到0， α_j/β_j 需遞減到一個非零常數，如此將保證(θ_j)會收斂。

例題4.4 (續例題4.3)

我們引用(4.4)式選擇不同的 α'_j s與 β'_j s來處理像 $h(x, y)$ 這種多重極值的函數。因為不同的起始值，或不同的 α'_j s與 β'_j s的選擇，此演算法將收斂到不同的局部極小值發生處。圖4-3、表4-1表示固定起始點為(0.65, 0.8)，使用不同的(α_j, β_j)作梯度法所得的(θ_j)的收斂路徑，其停止規則(stopping rule)為 $\|\theta_T - \theta_{T-1}\| < 10^{-5}$ 。第一種情況， $\alpha_j = \beta_j = 1/(10j)$ ，因為 α_j 下降速度太快導致收斂路徑產生大幅度的跳動，因此估計所得的最小值結果較差；第二種情況， $\alpha_j = \beta_j = 1/(100j)$ ，因減緩了 α_j 的下降速度，所以估計出來的最小值結果較第一種情況好；第三種情況， $\alpha_j = 1/(10 \log(1+j))$ ， $\beta_j = 1/j$ ， α_j 的下降速度是緩慢的，得到的結果也是較佳的。此處顯示梯度法的一個特徵， α_j 的下降速度愈緩慢，得到的最小值結果將越準確。

表 4-1

α_j	β_j	θ_T	$h(\theta_T)$	$\min_t h(\theta_t)$	Iteration T
$\frac{1}{10j}$	$\frac{1}{10j}$	(-0.815, 0.326)	1.321	0.1796	125
$\frac{1}{100j}$	$\frac{1}{100j}$	(0.898, 0.749)	0.000139	0.000139	121
$\frac{1}{10 \log(1+j)}$	$\frac{1}{j}$	(0.00002, -0.125)	6.19×10^{-9}	6.19×10^{-9}	253

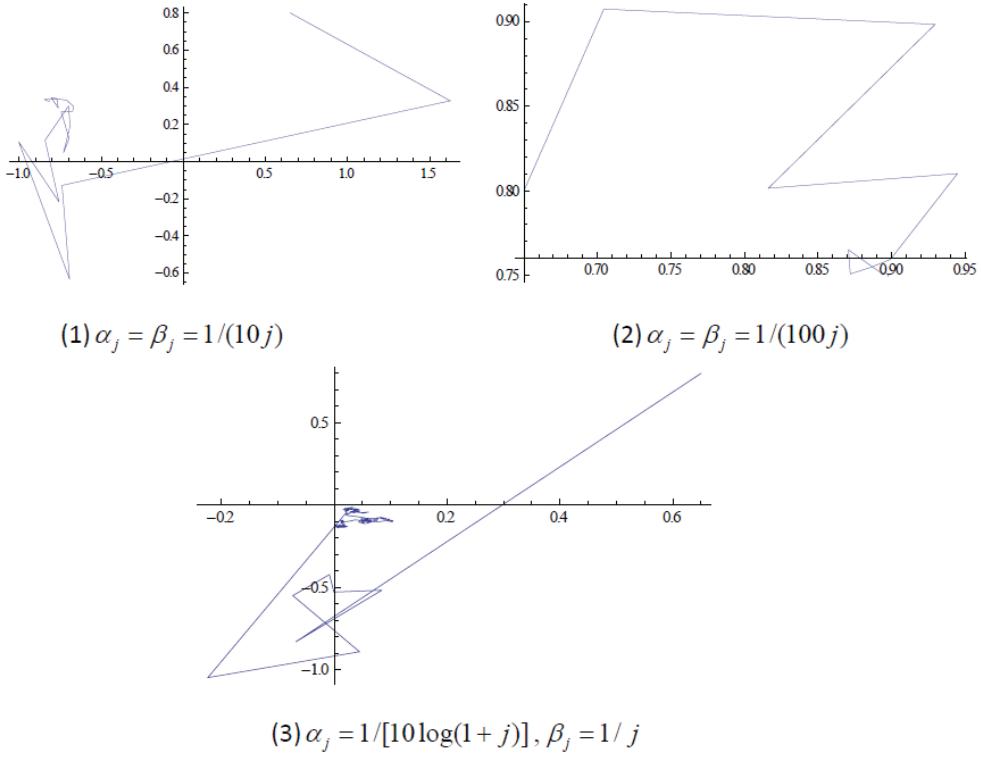


圖 4-3 圖中的路徑分別為給定不同的 α_j 及 β_j ，
以 $(0.65, 0.8)$ 為起始點，使用梯度法所得到的 θ_j 的收斂路徑。

4.2.3 模擬退火演算法(Simulated Annealing)

由 Metropolis 等人在 1953 年所提之模擬退火演算法，是為了解決在元素個數龐大的有限集合中將目標函數最小化的問題；此方法也可應用於模擬方面或在連續集合上求最佳解的問題。此方法的基本概念是藉由改變尺度(此處視為溫度(temperature)，若尺度為一負數則稱為能量(energy))的大小，以加速在函數 h 的圖形表面找出最大值並避免落入局部最大值所在位置。

給定一溫度參數 $T > 0$ ，自分佈 $\pi(\theta) \propto \exp(h(\theta)/T)$ 生成一組樣本 $\theta_1^T, \theta_2^T, \dots$ ，透過 4.2.1 節所述之探索方法可得 h 的最大值之近似值，但此作法在遇到多重局部極值的問題上較不適用。當我們考慮 Metropolis 等人在 1953 年所提之模擬退火演算法時，會明顯發現此方法受到局部極值影響的程度較小，其基本概念是若給定起始點 θ_0 ，可從 θ_0 的鄰域 $V(\theta_0)$ 上的一個均勻分佈，或是更一般地從分佈 $g(|\zeta - \theta_0|)$ 中生成 ζ ，則新的 θ 可以這樣定義

$$\theta_1 = \begin{cases} \zeta & \text{接受機率 } \rho = \exp(\Delta h/T) \wedge 1 \\ \theta_0 & \text{接受機率 } 1 - \rho \end{cases}$$

其中 $\Delta h = h(\zeta) - h(\theta_0)$, $\rho = \exp(\Delta h/T) \wedge 1 = \min\{\exp(\Delta h/T), 1\}$ 。若 $h(\zeta) \geq h(\theta_0)$ 則 ζ 被接受的機率為 1，意即 θ_0 一定改變成 ζ ；若 $h(\zeta) < h(\theta_0)$ ，則 ζ 被接受的機率為 ρ ($\rho \neq 0$)， ρ 的大小亦與 T 的選擇有關。若 θ_0 為局部最小值發生處，由此可見此演算法可避免 θ 被局限於 θ_0 附近(此方法事實上即為第七章所介紹的Metropolis 演算法。Metropolis 演算法是在模擬正比於 $e\{h(\theta)/T\}$ 的密度函數，而 $\theta_0, \theta_1, \dots$ 的極限分佈正好為此密度函數)。

一般使用模擬退火演算法時，在每次疊代的過程中都會修正 T 的大小，演算法內容如下：

演算法 6 模擬退火演算法

- 1.自密度函數 $g(|\zeta - \theta_i|)$ 的分佈中生成 ζ ；
- 2.在 $\rho_i = \exp(\Delta h_i/T_i) \wedge 1$ 的機率大小下接受 $\theta_{i+1} = \zeta$ ，其餘情況則取 $\theta_{i+1} = \theta_i$ ；
- 3.將 T_i 更新成 T_{i+1} 。

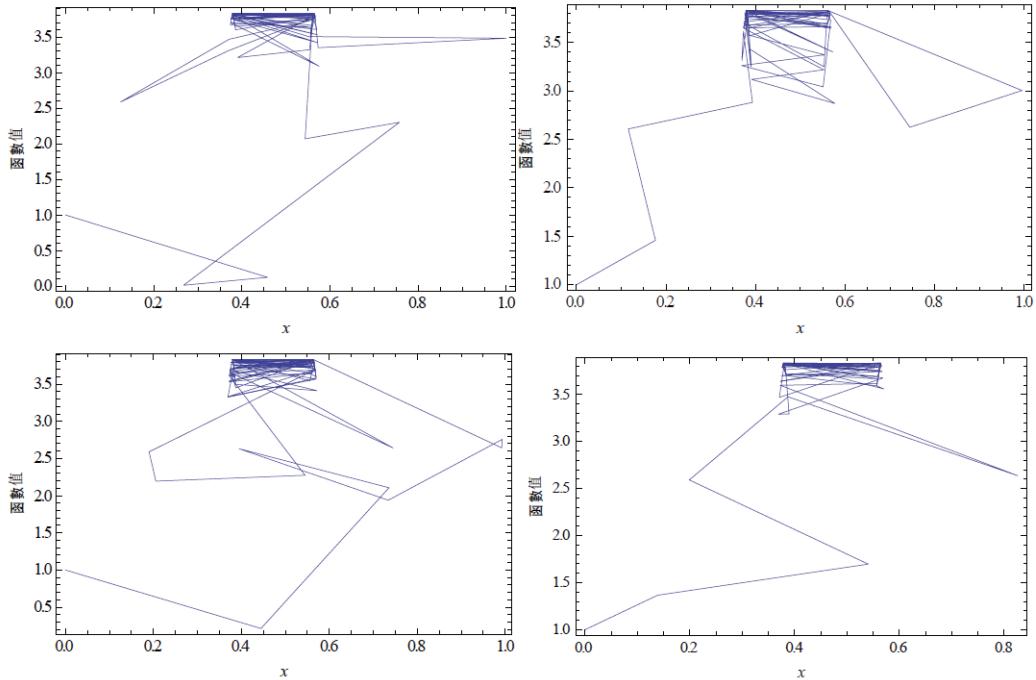


圖 4-4

四張圖均為使用模擬退火演算法所得的 2500 組序對 $(x^{(t)}, h(x^{(t)}))$ 形成的軌跡。

例題4.5

回顧例題 4.2， $h(x) = [\cos(50x) + \sin(20x)]^2$ ，利用以下特別的模擬退火演算法來找出最大值。假設至第 t 次疊代後，此演算法停在 $(x^{(t)}, h(x^{(t)}))$ 處，後續如下：

1. 令 $a_t = \max(x^{(t)} - r, 0)$, $b_t = \min(x^{(t)} + r, 1)$, 並自一致分佈 $U(a_t, b_t)$ 中生成一數值 u 。
2. 在 $\rho^{(t)} = \min\{\exp\left(\frac{h(u) - h(x^{(t)})}{T_t}\right), 1\}$ 的機率大小下接受 $x^{(t+1)} = u$, 其餘情況則取 $x^{(t+1)} = x^{(t)}$;
3. 將 T_i 更新成 T_{i+1} 。

在 $r = 0.5$ 、 $T_t = 1/\log(t)$ 時，模擬退火演算法所得結果如圖4-4所示。此四張圖呈現出 2500 組 $(x^{(t)}, h(x^{(t)}))$ 形成的不同軌跡，可見此方法如何快速地朝最大值發生處靠近(最大值的實際值約為 3.832，最大值發生處的 x 位於 $0.4 \sim 0.6$ 之間)並停留在兩個最大值發生處左右震盪(因為 h 對稱於 $1/2$)，其中 r 的大小是控制 $x^{(t)}$ 的範圍， T_t 的大小是控制 $h(x^{(t)})$ 收斂的速度。

定義4.6

設 \mathcal{E} 為一個有限狀態空間，在此有限狀態空間中，欲找函數 h 的最大值，

1. 若存在一序列 e_1, e_2, \dots, e_n 可以聯繫某兩個狀態 e_i, e_j ，並滿足 $h(e_k) \geq \underline{h}$, $k = 1, 2, \dots, n$ ，則稱狀態 e_j 可由狀態 e_i 在高度為 \underline{h} 時抵達。
2. 狀態 e_i 的高度為 d_i 的最大值，此 d_i 將使得存在某狀態 e_j ，滿足 $h(e_j) > h(e_i)$ ，且可由狀態 e_i 在高度為 $h(e_i) + d_i$ 時抵達。

因此， $h(e_i) + d_i$ 是連接 e_i, e_j 的最大高度。此外，當 e_i 是絕對最大值發生處則取 $d_i = -\infty$ 。

設 \mathcal{O} 表示集合 E 之下包含局部極大值發生點的集合， $\underline{\mathcal{O}} \subset \mathcal{O}$ 表示包含絕對極大值發生點的集合，Hàjek 於1988年提出下列定理

定理4.7

考慮一系統，此系統滿足任意兩狀態間都可能找到一個有限的狀態序列加以連接此兩狀態，若對於任意高度 $\underline{h} > 0$ 及任意狀態序對 (e_i, e_j) ， e_i 可由 e_j 在高度 \underline{h} 時抵達， e_j 也可由 e_i 在高度 \underline{h} 時抵達，且 (T_i) 遲減到零(模擬退火演算法中提到的”溫度”，用來加速 (θ_i) 收斂的參數)，則序列 (θ_i) 將滿足

$$\lim_{i \rightarrow \infty} P(\theta_i \in \underline{\mathcal{O}}) = 1$$

若且爲若

$$\sum_{i=1}^{\infty} \exp\{-D/T_i\} = +\infty$$

其中 $D = \min\{d_i : e_i \in \mathcal{O} - \underline{Q}\}$ 。

對於溫度 (T_i) 遲減的速率，定理 4.7 提供了模擬退火演算法能收斂到絕對極大值的充分必要條件。舉例來說，若 $T_i = \Gamma / \log i$ ，則模擬退火演算法能收斂到絕對極大值的充分必要條件爲 $\Gamma \geq D$ 。溫度遞減的速率除了前述的對數速率之外，實際上幾何速率 $T_i = \alpha_i T_0$ ($0 < \alpha < 1$) 也時常被採用。

除了以上模擬方法，近似方法在有限的狀態空間中對於解決最佳化問題也是必須的，因爲某些模型所包含的有限空間可能是含有非常龐大數量的樣本點，例如： 256×256 畫數的電視影像將對應於一個基數(cardinality)爲 $2^{256 \times 256} \approx 10^{20000}$ 的狀態空間；DNA 序列的分析可能包含 6×10^5 個基底(A、C、G、T)，並對應到大小爲 4^{600000} 的狀態空間。由此可知，雖然是在有限的狀態空間中討論最佳化問題，近似方法的確還是必要的。

例題4.8 易行模型(Ising model)

一般在電磁學或影像處理問題中常見的易行模型，主要在將一個二維表格 s 建模(可設 s 的大小爲 $D \times D$ ，其中每項均爲 1 或 -1)建模，表格數據整體的分佈相關於下列函數

$$h(s) = -J \sum_{(i,j) \in \mathcal{N}} s_i s_j - H \sum_i s_i \quad (4.5)$$

其中 i 表示表格 s 中各項的下標， \mathcal{N} 表示等價的鄰域關係(例如： i 與 j 可能是垂直或水平的鄰近位置關係)，而 J 和 H 假設爲已知參數。(4.5) 式在加入條件 $\tilde{s}_i = (s_i + 1)/2$ 後將等價於一個邏輯模型

$$P(\tilde{s}_i = 1 | s_j, j \neq i) = \frac{e^g}{1 + e^g} \quad (4.6)$$

其中 $g = g(s_j) = 2(H + J \sum_j s_j)$ 表示 i 的鄰近數字的和。對於已知的參數 J 和 H ，欲推論的問題可能是想獲得此系統最有可能的結構，也就是 $h(s)$ 的最小值。

此過程中，Metroplis 方法的使用在於對原始表格 s 的每個位置，利用 (4.6) 式的條件分佈逐一修正，並在每一步配合遞減的溫度參數 T 重複此方法，其初始值爲 s^0 ，機率爲 $\exp(-\Delta h/T)$ ，最後修正完的結果爲表格 s^1 。

例題4.9 (承例題4.3)

關於例題4.3所提及的函數

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x)x) \\ + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y)$$

我們可以使用模擬退火演算法來找尋 h 的最小值，此等價於找 $\exp(-h(x, y)/T_i)$ 的最大值。對於函數 g ，我們選擇 $[-0.1, 0.1]$ 上的一致分佈；對於溫度序列 (T_i) ，可選擇不同的遞減速率，如 $T_i = \Gamma / \log(i+1)$ 。表 4-2 及圖 4-5 的結果可見不同的溫度下降速率，將導致不同的收斂結果。

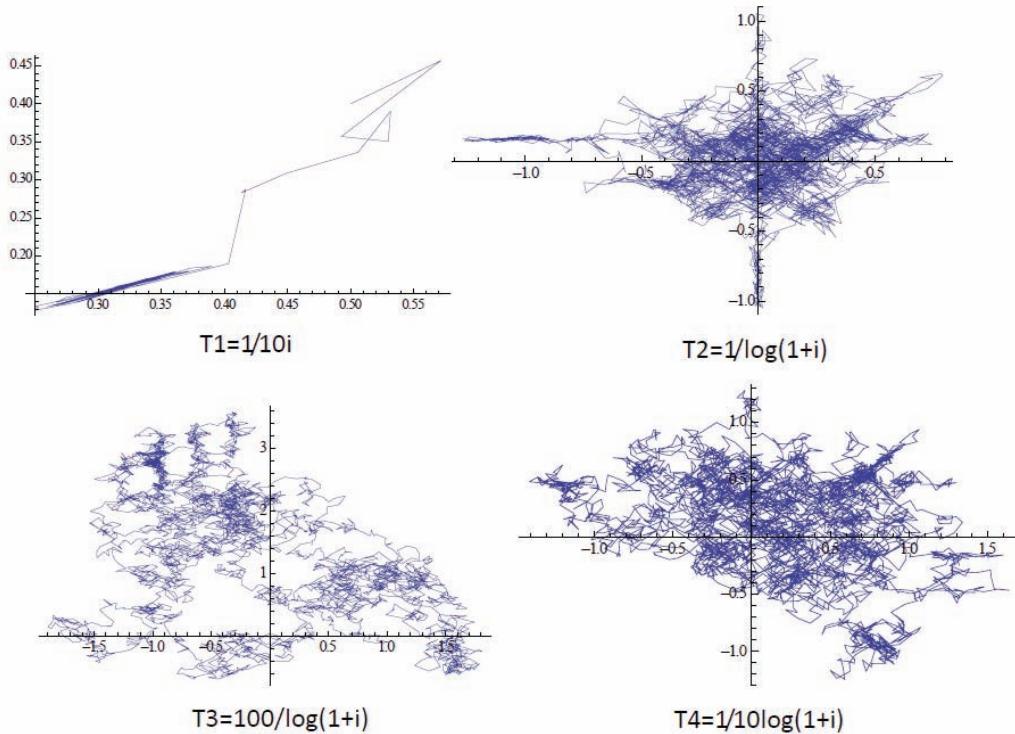


圖 4-5 對四種溫度 T_i 分別使用模擬退火演算法，得到 5000 組序對 (xt, yt) 所形成的軌跡。起點均為 $(0.5, 0.4)$ ，目標為尋找 $h(x, y)$ 的最小值發生處。

表 4-2

情況	T_i	θ_T	$h(\theta_T)$	$\min_t h(\theta_t)$	接受率
1	$\frac{1}{10i}$	$(0.271, 0.136)$	5.10×10^{-5}	2.68745×10^{-7}	0.0124
2	$\frac{1}{\log(1+i)}$	$(-0.176, 0.088)$	0.0204164	3.95×10^{-8}	0.6086
3	$\frac{100}{\log(1+i)}$	$(0.128, 0.489)$	0.270502	1.86×10^{-7}	0.8112
4	$\frac{1}{10\log(1+i)}$	$(-0.192, 0.056)$	0.0311471	3.33×10^{-7}	0.7532

4.2.4 事前回饋(Prior Feedback)

另一個尋找函數 $h(\theta)$ 之最大值的方法為 *Gibbs* 方法，也稱遞迴積分法或 Prior Feedback (Robert (1993))。*Gibbs* 方法是由 Hwang (1980) 提出，此方法的基礎概念是在 $h(\theta)$ 產生最大值的集合上， $\exp\{h(\theta)/T\}$ 也將收斂到一致分佈的結果(對 T 而言)。

定理4.10

設 h 為實數值函數，定義域為封閉且有界的集合 $\Theta (\subset \mathcal{R}^p)$ ，若存在唯一的 θ^* 滿足

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta})$$

且 h 在 θ^* 位置連續，則

$$\lim_{\lambda \rightarrow \infty} \frac{\int_{\Theta} \theta e^{\lambda h(\theta)} d\theta}{\int_{\Theta} e^{\lambda h(\theta)} d\theta} = \theta^*$$

更多關於此定理的內容請參考 Pincus (1968)。下面接著介紹一個與定理 4.10 相關的系理，此系理將遞迴積分法稍作修改，結果變為以貝氏方法來求對數概似函數 $\ell(\theta|x)$ 的最大值。

系理4.11

設 π 為 Θ 上的正密度函數 (positive density)，若存在唯一的最大概似估計量 θ^* ，則 θ^* 滿足

$$\lim_{\lambda \rightarrow \infty} \frac{\int \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta} = \theta^*$$

系理 4.11 和定理 4.10 一樣在分子分母部分均使用了 Laplace 近似法，系理 4.11 最後結果主要是說明了最大概似估計量可寫成一序列貝氏估計量的極限，而這些貝氏估計量會與任意分佈 π 及對應於 λ 次方的概似函數 ($\exp\{\lambda \ell(\theta|x)\}$) 之實際觀察值有關。當 $\lambda \in \mathcal{N}$ ，則

$$\delta_{\lambda}^{\pi}(x) = \frac{\int \theta e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}{\int e^{\lambda \ell(\theta|x)} \pi(\theta) d\theta}$$

即為與事後分佈 π 及重複 λ 次的樣本 x 有關的貝氏估計量，而對於系理 4.11 的結果，有一個直觀的看法是，當樣本個數趨近於無窮時，事後分佈的影響將消失，且隨著 λ 的增加，關於 $e^{\lambda \ell(\theta|x)} \pi(\theta)$ 的分佈將愈來愈集中於 $\ell(\theta|x)$ 的最大值發生處。

以一個實用的觀點來看，遞迴積分法可藉由計算貝氏估計量 $\delta_{\lambda_i}^{\pi}(x)$, $i = 1, 2, \dots$ ，直

到它變的平穩為止，即可得到結果。當使用遞迴演算法來計算 $\delta_\lambda^\pi(x)$ ，前一次 $\delta_\lambda^\pi(x)$ 的解可當作下一次計算較大一點的 λ 所對應的 $\delta_\lambda^\pi(x)$ 的起始值，此特徵與模擬退火演算法很類似。而此作法有別於模擬退火演算法的地方為：

1. 對於特定溫度 $(\frac{1}{\lambda})$ ，此演算法會收斂到一個定值 δ_λ^π 。
2. $(\frac{1}{\lambda})$ 的連續遞減在統計上是無意義的。
3. λ 增加到 $+\infty$ 的速度對於 δ_λ^π 收斂到 θ^* 不會造成影響。
4. 因為 λ 的意思是樣本的重複次數，所以此方法的統計意涵是較強烈的。
5. 此方法在分析上唯一的限制是 $\ell(\theta|x)$ 必須存在最大值發生處 θ^* 。

例題4.12 伽瑪分佈之形狀參數估計

考慮伽瑪分佈 $G(\alpha, \beta)$ ， α 是未知的形狀參數， β 為已知參數(不失一般性可設 $\beta = 1$)，對於 α 的常數先驗分佈而言(常數先驗分佈(constant prior distribution)，也稱非正常先驗分佈(improper prior distribution))，其事後分佈為

$$\pi_\lambda(\alpha|x) \propto x^{\lambda(\alpha-1)} e^{-\lambda x} \Gamma(\alpha)^{-\lambda}$$

給定 λ ，則欲計算期望值 $\mathbb{E}[\alpha|x, \lambda]$ 可使用Metropolis-Hastings 演算法(第七章) 透過工具密度函數 $Exp\left(\frac{1}{\alpha^{(n-1)}}\right)$ 來模擬，其中 $\alpha^{(n-1)}$ 表示此相關的馬可夫鏈的前一個數值。表4-3呈現的是 $x = 1.5$ ，對應於不同 λ 所得的 $\delta_\lambda^\pi(x) = \mathbb{E}[\alpha|x, \lambda]$ 。透過Mathematica 的驗證顯示 $\frac{x^\alpha}{\Gamma(\alpha)}$ 在 $x = 1.5$ 時的最大值發生處， α 會接近 2.0。

表4-3

λ	5	10	100	1000	5000	10^4
δ_λ^π	2.02	2.04	1.89	1.98	1.94	2.00

例題4.13 保序迴歸函數(Isotonic regression function)

考慮一組具常態分佈的觀察值 $X_{i,j} \sim \mathcal{N}(\theta_{i,j}, \frac{1}{n_{i,j}})$ ，

	$\theta_{i-1,j}$	
$\theta_{i,j-1}$	$\theta_{i,j}$	$\theta_{i,j+1}$
	$\theta_{i+1,j}$	

其中，期望值滿足

$$\theta_{i+1,j} \vee \theta_{i,j+1} \leq \theta_{i,j} \leq \theta_{i-1,j} \wedge \theta_{i,j-1}$$

在這樣的期望值限制下，雖然Dykstra and Robertson (1982) 曾提出一個有效率的演算法來最大化概似函數，不過前述遞迴積分法對此問題，可以不需研究額外的定理也不需高超的程式技術就可以找到概似估計量 $\theta = (\theta_{i,j})$ 的最大值。表 4-4 顯示的是由Robertson 等人(1988)收集的美國愛荷華大學大一學生的兩次入學考試成績，雖然這些數據是有界的，但若此處欲最小化的函數是最小平方法準則(the least squares criterion, $\sum_{i=1}^n (y_i - ax_i - b)^2$)，則可用常態模型來描述。表 4-5 為Robert 、Hwang (1996)及Robertson 等人 (1988)使用遞迴積分法所得期望值之最大概似估計量的結果。

HSR decile	ACT score								
	1-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
91 ≤ HSR ≤ 99	1.57 (4)	2.11 (5)	2.73 (18)	2.96 (39)	2.97 (126)	3.13 (219)	3.41 (232)	3.45 (47)	3.51 (4)
81 ≤ HSR ≤ 90	1.80 (6)	1.94 (15)	2.52 (30)	2.68 (65)	2.69 (117)	2.82 (143)	2.75 (70)	2.74 (8)	(0)
71 ≤ HSR ≤ 80	1.88 (10)	2.32 (13)	2.32 (51)	2.53 (83)	2.58 (115)	2.55 (107)	2.72 (24)	2.76 (4)	(0)
61 ≤ HSR ≤ 70	2.11 (6)	2.23 (32)	2.29 (59)	2.29 (84)	2.50 (75)	2.42 (44)	2.41 (19)	(0)	(0)
51 ≤ HSR ≤ 60	1.60 (11)	2.06 (16)	2.12 (49)	2.11 (63)	2.31 (57)	2.10 (40)	1.58 (4)	2.13 (1)	(0)
41 ≤ HSR ≤ 50	1.75 (6)	1.98 (12)	2.05 (31)	2.16 (42)	2.35 (34)	2.48 (21)	1.36 (4)	(0)	(0)
31 ≤ HSR ≤ 40	1.92 (7)	1.84 (6)	2.15 (5)	1.95 (27)	2.02 (13)	2.10 (13)	1.49 (2)	(0)	(0)
21 ≤ HSR ≤ 30	1.62 (1)	2.26 (2)	1.91 (5)	1.86 (14)	1.88 (11)	3.78 (1)	1.40 (2)	(0)	(0)
HSR ≤ 20	1.38 (1)	1.57 (2)	2.49 (5)	2.01 (7)	2.07 (7)	(0)	0.75 (1)	(0)	(0)

表 4-4

HSR decile	ACT score								
	1-12	13-15	16-18	19-21	22-24	25-27	28-29	31-32	34-36
91-99	1.87	2.18	2.73	2.96	2.97	3.13	3.41	3.45	3.51
81-89	1.87	2.17	2.52	2.68	2.69	2.79	2.79	2.80	
71-99	1.86	2.17	2.32	2.53	2.56	2.57	2.72	2.76	
61-69	1.86	2.17	2.29	2.29	2.46	2.46	2.47		
51-59	1.74	2.06	2.12	2.13	2.24	2.24	2.24	2.27	
41-49	1.74	1.98	2.05	2.13	2.24	2.24	2.24		
31-39	1.74	1.94	1.99	1.99	2.02	2.06	2.06		
21-29	1.62	1.93	1.97	1.97	1.98	2.05	2.06		
00-20	1.38	1.57	1.97	1.97	1.97		1.97		

表 4-5

4.3 隨機逼近算法(Stochastic Approximation)

4.3.1 遺失資料模型及去邊際化(Miss. Data Models and Demarginal.)

前幾章節中，我們遇過某些因為遺失資料，而使已觀察得的模型變的非常複雜的結構，如刪失資料模型，混合模型(我們沒有觀察生成觀察值之項目的指標)，邏輯迴歸(觀察值 Y_i 可被解釋為一個指標其中具邏輯分佈的連續隨機變數少於 $X_i^t \beta$)。

刪失資料模型的概似函數可表示為

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta) dz. \quad (4.7)$$

或更一般地將欲最佳化的函數 $h(x)$ 表示成

$$h(x) = E[H(x, Z)]. \quad (4.8)$$

例題4.14 刪失資料的概似函數

假設我們觀察到 Y_1, \dots, Y_n 是 iid 來自分佈 $f(y - \theta)$ ，將此觀察值排序後得到沒有被刪失的資料為 $\mathbf{y} = (y_1, \dots, y_m)$ ，被刪失的資料為 (y_{m+1}, \dots, y_n) (均視為 a)。此概似函數為

$$L(\theta|\mathbf{y}) = [1 - F(a - \theta)]^{n-m} \prod_{i=1}^m f(y_i - \theta) \quad (4.9)$$

其中 F 是相關於 f 的 cdf。若我們有觀察到後 $n-m$ 個值，記作 $\mathbf{z} = (z_{m+1}, \dots, z_n)$ 且 $z_i > a$ ($i = m + 1, \dots, n$)，則可建構完備資料(complete data)的概似函數

$$L^c(\theta|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^m f(y_i - \theta) \prod_{i=m+1}^n f(z_i - \theta)$$

又兩概似函數之關係為

$$L(\theta|\mathbf{y}) = E[L^c(\theta|\mathbf{y}, \mathbf{z})] = \int L^c(\theta|\mathbf{y}, \mathbf{z}) d\mathbf{z}$$

若 $f(y - \theta) = N(\theta, 1)$ ，圖 4-6 顯示三種概似函數的圖形。

以上為遺失資料模型(missing data model)， $L^c(\theta|\mathbf{y}, \mathbf{z}) = f(\mathbf{y}, \mathbf{z}|\theta)$ 為對應完備資料(\mathbf{y}, \mathbf{z})之觀察值的完備模型(complete-model)或完備資料(complete-data)的概似函數。一般而言，我們將(4.7)式稱為「去邊際化(demarginalization)」，此作法就是將一個我們感興趣的函數表示成一個更容易處理的量的積分形式。

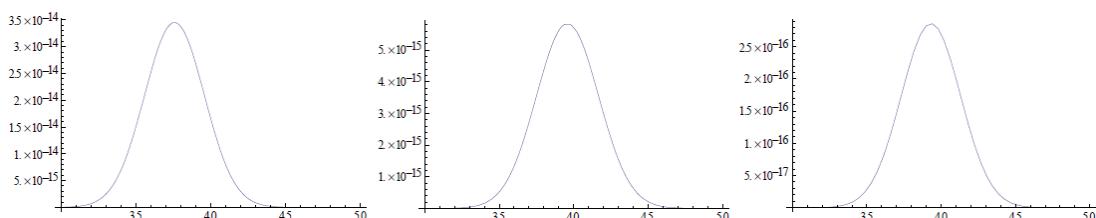


圖 4-6-1 三種概似函數的圖形

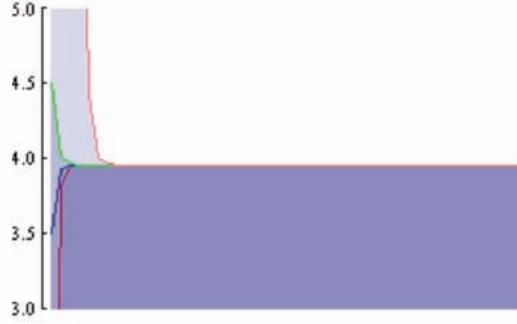


圖 4-6-2 EM 估計結果

4.3.2 EM演算法(The EM Algorithm)

EM(Expectation-Maximization) 演算法最早是由Dempster et al.(1971) 提出，是為了解決最大化概似函數的問題。此法利用了(4.7)式的表示方法，並找出一序列較簡單的最大化問題的解，而這些解的極限就是原問題的解。假設觀察值 X_1, \dots, X_n 是 iid 來自分佈 $g(x|\theta)$ ，欲找 $\hat{\theta}$ 使 $L(\theta|x) = \prod_{i=1}^n g(x_i|\theta)$ 達到最大值。將樣本擴充，增加 \mathbf{Z} ， $\mathbf{X}, \mathbf{Z} \sim f(x, z|\theta)$ ，以下為 EM 演算法中的基本關係式

$$k(\mathbf{z}|\theta, \mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{z}|\theta)}{g(\mathbf{x}|\theta)} \quad (4.10)$$

$k(\mathbf{z}|\theta, \mathbf{x})$ 是遺失資料 \mathbf{Z} 在給定 \mathbf{x} 下的條件分佈。

承 (4.10)式，

$$g(\mathbf{x}|\theta) = \frac{f(\mathbf{x}, \mathbf{z}|\theta)}{k(\mathbf{z}|\theta, \mathbf{x})} \implies \log g(\mathbf{x}|\theta) = \log f(\mathbf{x}, \mathbf{z}|\theta) - \log k(\mathbf{z}|\theta, \mathbf{x})$$

對任意 θ_0 ，我們將上式對 \mathbf{z} 取期望值(先乘上 $k(\mathbf{z}|\theta_0, \mathbf{x})$ 再對 \mathbf{z} 做積分)，

$$\begin{aligned} E_{\theta_0}[\log g(\mathbf{x}|\theta)] &= E_{\theta_0}[\log f(\mathbf{x}, \mathbf{z}|\theta) - \log k(\mathbf{z}|\theta, \mathbf{x})] \\ &\implies \log g(\mathbf{x}|\theta) = E_{\theta_0}[\log f(\mathbf{x}, \mathbf{z}|\theta)] - E_{\theta_0}[\log k(\mathbf{z}|\theta, \mathbf{x})] \\ &\implies \log L(\theta|\mathbf{x}) = E_{\theta_0}[\log L^c(\theta|\mathbf{x}, \mathbf{z})] - E_{\theta_0}[\log k(\mathbf{z}|\theta, \mathbf{x})] \end{aligned} \quad (4.11)$$

由 (4.11)式發現，EM 演算法可視為是一個去邊際化模型。特別的是，欲使 $\log L(\theta|\mathbf{x})$ 達到最大值，我們只需處理 $E_{\theta_0}[\log L^c(\theta|\mathbf{x}, \mathbf{z})]$ ，而後半部 $E_{\theta_0}[\log k(\mathbf{z}|\theta, \mathbf{x})]$ 可忽略不計。

一般EM的標記法為

$$Q(\theta|\theta_0, \mathbf{x}) = \mathbb{E}_{\theta_0}[\log L^c(\theta|\mathbf{x}, \mathbf{z})]$$

然後再將 $Q(\theta|\theta_0, \mathbf{x})$ 最大化，若 $\hat{\theta}_1$ 是使 $Q(\theta|\theta_0, \mathbf{x})$ 達到最大的 θ 值，則可將 $Q(\theta|\theta_0, \mathbf{x})$ 中的 θ_0 用 $\hat{\theta}_1$ 取代；依此類推我們將得到一序列的估計量 $\hat{\theta}_j, j = 1, 2, \dots, \hat{\theta}_j$ 為最大

化 $Q(\theta|\hat{\theta}_{j-1}, \mathbf{x})$ 之 θ 值，也就是

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \mathbf{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \mathbf{x}) \quad (4.13)$$

重複上述疊代過程，直到得到固定的 Q 值為止。 (4.13) 式包含了求期望值及求最大值的步驟，正好說明了EM演算法的命名由來。例如：在第 j 次疊代時，我們先計算 (4.12) 式的期望值，並用 $\hat{\theta}_{j-1}$ 取代 (4.12) 中的 θ_0 ，此步驟稱為E步驟(E-step)，再最大化期望值，此步驟稱為M步驟(M-step)。

演算法 7 EM演算法(The EM Algorithm)

1.(E步驟)對應於 $k(\mathbf{z}|\hat{\theta}_m, \mathbf{x})$ ，計算下面所列之期望值

$$Q(\theta|\hat{\theta}_{(m)}, \mathbf{x}) = \mathbb{E}_{\hat{\theta}_{(m)}} [\log L^c(\theta|\mathbf{x}, \mathbf{z})]$$

2.(M步驟)找最大化 $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ 的 θ ，記作 $\theta_{m+1} = \arg \max_{\theta} Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ 。

EM演算法的理論基礎為在每一步最大化 $Q(\theta|\hat{\theta}_{(m)}, \mathbf{x})$ 的過程中， $\log L(\theta|\mathbf{x})$ 的函數值也隨之變大。我們來看一下由 Dempster 等學者在 1977 年提出的定理。

定理4.15

(4.13) 式所表示之序列($\hat{\theta}_{(j)}$)滿足

$$L(\hat{\theta}_{(j+1)}|\mathbf{x}) \geq L(\hat{\theta}_{(j)}|\mathbf{x})$$

其中等號成立於

$$Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, \mathbf{x}) = Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, \mathbf{x})。$$

證明：

首先我們先定義所謂的 f 與 g 之間對應於 f 的entropy distance：

$$E_f [\log(f(X)/g(X))] = \int \log(f(x)/g(x))f(x)dx$$

由Jensen's不等式可推得

$$\begin{aligned} E_f [\log(f(X)/g(X))] &= \int \log(f(x)/g(x))f(x)dx \\ &= \int -\log(g(x)/f(x))f(x)dx \\ &\geq -\log \int (g(x)/f(x))f(x)dx = 0 \end{aligned}$$

$$\implies E_f[\log f(X)] \geq E_f[\log g(X)]$$

在一連串的疊代過程中，由定義得知 $\hat{\theta}_{(j+1)}$ 滿足

$$Q(\hat{\theta}_{(j+1)} | \hat{\theta}_{(j)}, \mathbf{x}) \geq Q(\hat{\theta}_{(j)} | \hat{\theta}_{(j)}, \mathbf{x})$$

若我們可以證明

$$\mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j+1)}, \mathbf{x})] \leq \mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j)}, \mathbf{x})] \quad (4.14)$$

則由 (4.11) 式可得知概似函數 $L(\theta | \mathbf{x})$ 將隨著每次疊代而遞增。

藉由上面用到的 Jensen 不等式，我們得到

$$\begin{aligned} \mathbb{E}_{\hat{\theta}_{(j)}} \left[\log \left(\frac{k(\mathbf{Z} | \hat{\theta}_{(j+1)}, \mathbf{x})}{k(\mathbf{Z} | \hat{\theta}_{(j)}, \mathbf{x})} \right) \right] &\leq \log \mathbb{E}_{\hat{\theta}_{(j)}} \left[\frac{k(\mathbf{Z} | \hat{\theta}_{(j+1)}, \mathbf{x})}{k(\mathbf{Z} | \hat{\theta}_{(j)}, \mathbf{x})} \right] = 0 \\ \implies \mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j+1)}, \mathbf{x})] &\leq \mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j)}, \mathbf{x})] \end{aligned} \quad (4.15)$$

因此

$$\begin{aligned} L(\hat{\theta}_{(j+1)} | \mathbf{x}) &= Q(\hat{\theta}_{(j+1)} | \hat{\theta}_{(j)}, \mathbf{x}) - \mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j+1)}, \mathbf{x})] \\ &\geq Q(\hat{\theta}_{(j)} | \hat{\theta}_{(j)}, \mathbf{x}) - \mathbb{E}_{\hat{\theta}_{(j)}} [\log k(\mathbf{Z} | \hat{\theta}_{(j)}, \mathbf{x})] \\ &= L(\hat{\theta}_{(j)} | \mathbf{x}) \end{aligned}$$

我們完成了此證明。雖然定理 4.15 保證了概似函數 $L(\theta | \mathbf{x})$ 將隨著每次疊代而遞增，但我們仍不能確定序列 $(\hat{\theta}_{(j)})$ 是否會收斂到最大概似估計量。

下述定理由 Boyles 及 Wu 在 1983 年提出，此定理可算是一個最便於應用的條件來保證序列 $(\hat{\theta}_{(j)})$ 會收斂到平穩點 (stationary point)，使一階導數為零的點)。

定理 4.16

若完備資料之概似函數的期望值 $Q(\theta | \hat{\theta}_0, \mathbf{x})$ 在 θ 及 $\hat{\theta}_0$ 均連續，則 EM 序列 $(\hat{\theta}_{(j)})$ 的每個極限點 (limit point) 即為 $L(\theta | \mathbf{x})$ 的一個平穩點，且 $L(\hat{\theta}_{(j)} | \mathbf{x})$ 將遞增地收斂到 $L(\hat{\theta} | \mathbf{x})$ ，其中 $\hat{\theta}$ 是某個平穩點。

例題 4.17 刪失資料之EM演算法 (EM for censored data)

設 $Y_i \sim N(\theta, 1)$ 且在 $Y_i > a$ 處發生刪失情形。此完備資料的概似函數為

$$L^c(\theta | \mathbf{y}, \mathbf{z}) \propto \prod_{i=1}^m \exp\{-(y_i - \theta)^2/2\} \prod_{i=m+1}^n \exp(-(z_i - \theta)^2/2)$$

而刪失資料 $\mathbf{z} = (z_{m+1}, \dots, z_n)$ 為截尾常態分佈

$$\mathbf{Z} \sim k(\mathbf{z}|\theta, y) = \left(\frac{1}{1 - \Phi(a - \theta)} \right)^{(n-m)} \times \frac{1}{(2\pi)^{(n-m)/2}} \exp \left\{ - \sum_{i=m+1}^n (z_i - \theta)^2 / 2 \right\} \quad (4.16)$$

將完備資料的對數概似函數對刪失資料 \mathbf{z} 取期望值(E步驟)

$$\begin{aligned} \log L^c(\theta|\mathbf{y}, \mathbf{z}) &\propto \sum_{i=1}^m -(y_i - \theta)^2 / 2 - \sum_{i=m+1}^n (z_i - \theta)^2 / 2 \\ \implies Q(\theta|\theta') &\propto -\frac{1}{2} \sum_{i=1}^m (y_i - \theta)^2 - \frac{1}{2} \sum_{i=m+1}^n \mathbb{E}_{\theta'}[(Z_i - \theta)^2]. \end{aligned}$$

接著將 $Q(\theta|\theta')$ 對 θ 微分，以求出使 $Q(\theta|\theta')$ 有最大值的估計量 $\hat{\theta}$ (M步驟)

$$\frac{\partial Q(\theta|\theta')}{\partial \theta} = \sum_{i=1}^m (y_i - \theta) - \frac{1}{2} \sum_{i=m+1}^n \frac{\partial \mathbb{E}_{\theta'}[(Z_i - \theta)^2]}{\partial \theta} \quad (a)$$

其中

$$\begin{aligned} \frac{\partial \mathbb{E}_{\theta'}[(Z_i - \theta)^2]}{\partial \theta} &= \frac{\partial}{\partial \theta} \int_a^\infty (z_i - \theta)^2 k(\mathbf{z}|\theta', y) dz \\ &= \int_a^\infty -2(z_i - \theta) k(\mathbf{z}|\theta', y) dz \\ &= -2 \left[\int_a^\infty z_i k(\mathbf{z}|\theta', y) dz - \theta \int_a^\infty k(\mathbf{z}|\theta', y) dz \right] \\ &= -2 [\mathbb{E}_{\theta'}(Z_i) - \theta] \end{aligned} \quad (b)$$

且

$$\begin{aligned} \mathbb{E}_{\theta'}(Z_i) &= \int_a^\infty z_i k(\mathbf{z}|\theta', y) dz_i \\ &= \int_a^\infty \frac{z_i}{(1 - \Phi(a - \theta')) \sqrt{2\pi}} e^{-(z_i - \theta')^2 / 2} dz_i \\ &= \frac{1}{1 - \Phi(a - \theta')} \left[\frac{1}{\sqrt{2\pi}} \int_a^\infty (z_i - \theta') e^{-(z_i - \theta')^2 / 2} dz_i \right. \\ &\quad \left. + \theta' \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-(z_i - \theta')^2 / 2} dz_i \right] \\ &= \frac{1}{1 - \Phi(a - \theta')} \left[\frac{1}{\sqrt{2\pi}} e^{-(a - \theta')^2 / 2} + (1 - \Phi(a - \theta')) \theta' \right] \\ &= \frac{1}{1 - \Phi(a - \theta')} [\phi(a - \theta') + (1 - \Phi(a - \theta')) \theta'] \\ &= \frac{\phi(a - \theta')}{1 - \Phi(a - \theta')} + \theta' \end{aligned} \quad (c)$$

將(b)、(c)的結果代回(a)，並令 $\frac{\partial Q(\theta|\theta')}{\partial \theta} = 0$

$$\begin{aligned}\frac{\partial Q(\theta|\theta')}{\partial \theta} &= \sum_{i=1}^m (y_i - \theta) - \frac{1}{2} \sum_{i=m+1}^n (-2 [\mathbb{E}_{\theta'}(Z_i) - \theta]) = 0 \\ \implies m\bar{y} - m\theta + (n-m)\mathbb{E}_{\theta'}(Z_1) - (n-m)\theta &= 0 \\ \implies \hat{\theta} &= \frac{m\bar{y} + (n-m)\mathbb{E}_{\theta'}(Z_1)}{n} \\ \implies \hat{\theta} &= \frac{m}{n}\bar{y} + \frac{n-m}{n} \left[\theta' + \frac{\phi(a-\theta')}{1-\Phi(a-\theta')} \right]\end{aligned}$$

其中 ϕ 、 Φ 分別表示常態分佈的 pdf、cdf。上式結果可寫成EM序列

$$\hat{\theta}^{(j+1)} = \frac{m}{n}\bar{y} + \frac{n-m}{n} \left[\hat{\theta}^{(j)} + \frac{\phi(a-\hat{\theta}^{(j)})}{1-\Phi(a-\hat{\theta}^{(j)})} \right] \quad (4.17)$$

例題4.18 行動電話方案(Cellular phone plans)

行動電話業者常提出一些可供客戶選擇的方案，可能一個價格就包含有四至五種選擇(如：簡訊、來電顯示等服務)，或者也可能是對單一功能分別出售。某家電信業者在某區域提出了一個四方案的組合，在另一區提出了一個五方案的組合(內容包含前述四合一方案，只是再多加一項服務)。兩地區的客戶將被詢問喜歡哪一種方案，並將客戶選擇的結果記錄下來，看哪一種方案較被大眾所喜愛，以協助電信業者進行定價動作。以下我們來將完備資料建模。若在區域 i 有 n_i 個客戶每位客戶都從五種方案中選擇其一，對於第 j 個客戶的選擇內容，我們用 $Z_i^{(j)} = (Z_{i1}, \dots, Z_{i5})$ 表示，而 $Z_i^{(j)}$ 屬於多項式分佈 $\mathcal{M}(1, (p_1, \dots, p_5))$ 。假設客戶之間是獨立的則在區域 i 客戶群的選擇內容可記為

$$T_i = (T_{i1}, \dots, T_{i5}) = \sum_{n_i}^{j=1} Z_i^{(j)} \sim \mathcal{M}(n_i, (p_1, \dots, p_5))$$

若前 m 個區域在第五種方案的資料(Z_{i5})發生遺失狀況，將遺失資料記作 x_i ，則完備資料的概似函數為

$$L(\mathbf{p}|\mathbf{T}, \mathbf{x}) = \prod_{i=1}^m \binom{n_i + x_i}{T_{i1}, \dots, T_{i4}, x_i} p_1^{T_{i1}} \dots p_4^{T_{i4}} p_5^{x_i} \times \prod_{i=m+1}^n \binom{n_i}{T_{i1}, \dots, T_{i5}} \prod_{j=1}^5 p_j^{T_{ij}}$$

其中 $\mathbf{p} = (p_1, \dots, p_5)$, $\mathbf{T} = (T_1, \dots, T_5)$, $\mathbf{x} = (x_1, x_2, \dots, x_m)$ ，且 $\binom{n}{n_1, n_2, \dots, n_k}$ 是多項式分佈的係數 $\frac{n!}{n_1!n_2!\dots n_k!}$ 。由於已觀察到的資料 \mathbf{T} 的概似函數為

$$\begin{aligned}L(\mathbf{p}|\mathbf{T}) &= \sum_{\mathbf{x}} L(\mathbf{p}|\mathbf{T}, \mathbf{x}) \\ &= \sum_{\mathbf{x}} \prod_{i=1}^m \binom{n_i + x_i}{T_{i1}, \dots, T_{i4}, x_i} p_1^{T_{i1}} \dots p_4^{T_{i4}} p_5^{x_i} \times \prod_{i=m+1}^n \binom{n_i}{T_{i1}, \dots, T_{i5}} \prod_{j=1}^5 p_j^{T_{ij}}\end{aligned}$$

$$\begin{aligned}
&= \prod_{i=m+1}^n \binom{n_i}{T_{i1}, \dots, T_{i5}} \prod_{j=1}^5 p_j^{T_{ij}} \sum_{\mathbf{x}} \prod_{i=1}^m \binom{n_i + x_i}{T_{i1}, \dots, T_{i4}, x_i} p_1^{T_{i1}} \dots p_4^{T_{i4}} p_5^{x_i} \\
&= \prod_{i=m+1}^n \binom{n_i}{T_{i1}, \dots, T_{i5}} \prod_{j=1}^5 p_j^{T_{ij}} \prod_{i=1}^m \frac{1}{T_{i1}! \dots T_{i4}!} p_1^{T_{i1}} \dots p_4^{T_{i4}} \sum_{\mathbf{x}} \prod_{i=1}^m \frac{(n_i + x_i)!}{x_i!} p_5^{x_i}
\end{aligned}$$

可推得遺失資料的概似函數為

$$\begin{aligned}
k(\mathbf{x}|\mathbf{T}, \mathbf{p}) &= \frac{L(\mathbf{p}|\mathbf{T}, \mathbf{x})}{L(\mathbf{p}|\mathbf{T})} \\
&= \frac{\prod_{i=1}^m \frac{(n_i + x_i)!}{x_i!} p_5^{x_i}}{\sum_{\mathbf{x}} \prod_{i=1}^m \frac{(n_i + x_i)!}{x_i!} p_5^{x_i}} \\
&= \frac{\prod_{i=1}^m \frac{(n_i + x_i)!}{n_i! x_i!} p_5^{x_i} (1 - p_5)^{n_i+1}}{\sum_{\mathbf{x}} \prod_{i=1}^m \frac{(n_i + x_i)!}{n_i! x_i!} p_5^{x_i} (1 - p_5)^{n_i+1}}
\end{aligned}$$

因為 $\frac{(n_i + x_i)!}{n_i! x_i!} p_5^{x_i} (1 - p_5)^{n_i+1}$ 為負二項式分佈 $NB(n_i + 1, p_5)$ 的機率密度函數，所以分母的值為 1，也就是

$$\sum_{\mathbf{x}} \prod_{i=1}^m \frac{(n_i + x_i)!}{n_i! x_i!} p_5^{x_i} (1 - p_5)^{n_i+1} = 1$$

故

$$k(\mathbf{x}|\mathbf{T}, \mathbf{p}) = \frac{L(\mathbf{p}|\mathbf{T}, \mathbf{x})}{L(\mathbf{p}|\mathbf{T})} = \prod_{i=1}^m \frac{(n_i + x_i)!}{n_i! x_i!} p_5^{x_i} (1 - p_5)^{n_i+1}$$

設 $W_j = \sum_{i=1}^n T_{ij}$, $j = 1, \dots, 4$, $W_5 = \sum_{i=1}^m T_{i5}$, 則完備資料的對數概似函數為

$$\begin{aligned}
\log L(\mathbf{p}|\mathbf{T}, \mathbf{x}) &= \sum_{i=1}^m \log \frac{1}{T_{i1}! \dots T_{i4}!} + \sum_{i=m+1}^n \log \binom{n_i}{T_{i1}, \dots, T_{i5}} \\
&\quad + \sum_{i=1}^n (T_{i1} \log p_1 + \dots + T_{i4} \log p_4) \\
&\quad + \left(\sum_{i=m+1}^n T_{i5} + \sum_{i=1}^m x_i \right) \log p_5 + \sum_{i=1}^m \log \frac{(n_i + x_i)!}{x_i!} \\
&= \sum_{i=1}^m \log \frac{1}{T_{i1}! \dots T_{i4}!} + \sum_{i=m+1}^n \log \binom{n_i}{T_{i1}, \dots, T_{i5}} + \sum_{j=1}^4 W_j \log p_j \\
&\quad + (W_5 + \sum_{i=1}^m x_i) \log(1 - p_1 - p_2 - p_3 - p_4) + \sum_{i=1}^m \log \frac{(n_i + x_i)!}{x_i!}
\end{aligned}$$

先對 \mathbf{x} 取期望值，

$$Q(\mathbf{p}|\mathbf{p}') = \sum_{i=1}^m \log \frac{1}{T_{i1}! \dots T_{i4}!} + \sum_{i=m+1}^n \log \binom{n_i}{T_{i1}, \dots, T_{i5}} + \sum_{j=1}^4 W_j \log p_j$$

$$\begin{aligned}
& + (W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}')) \log(1 - p_1 - p_2 - p_3 - p_4) \\
& + \sum_{i=1}^m E(\log \frac{(n_i + X_i)!}{X_i!} |\mathbf{p}')
\end{aligned}$$

再分別對 p_1, p_2, p_3, p_4 偏微分，

$$\frac{\partial Q(\mathbf{p}|\mathbf{p}')}{\partial p_j} = \frac{W_j}{p_j} - \frac{W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}')}{1 - p_1 - p_2 - p_3 - p_4}$$

令 $\frac{\partial Q(\mathbf{p}|\mathbf{p}')}{\partial p_j} = 0, j = 1, 2, 3, 4$ 可得

$$\begin{aligned}
(W_1 + W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}'))p_1 + W_1p_2 + W_1p_3 + W_1p_4 &= W_1, \\
W_2p_1 + (W_2 + W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}'))p_2 + W_2p_3 + W_2p_4 &= W_2, \\
W_3p_1 + W_3p_2 + (W_3 + W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}'))p_3 + W_3p_4 &= W_3, \\
W_4p_1 + W_4p_2 + W_4p_3 + (W_4 + W_5 + \sum_{i=1}^m E(X_i|\mathbf{p}'))p_4 &= W_4
\end{aligned}$$

整理後得到

$$\hat{p}_j^{(t+1)} = \frac{W_j}{\sum_{i=1}^m E(X_i|\mathbf{p}^{(t)}) + \sum_{j=1}^5 W_j}, j = 1, 2, 3, 4$$

其中 $E(X_i|\mathbf{p}^{(t)}) = (n_i + 1) \frac{\hat{p}_5^{(t)}}{1 - \hat{p}_5^{(t)}}$ 。

例題4.19 平均混合常態分佈的 EM 演算法

(EM for mean mixtures of normal distributions)

我們考慮兩個常態分佈的混合模型 $p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2)$ ，除了 μ_1, μ_2 之外，其他參數均為已知。圖 4-7 下半部分表示此模型的對數概似函數之表面圖形，由此分佈(分別給定參數的起始值 $p = 0.7, \sigma = 1, (\mu_1, \mu_2) = (0, 3.1)$)模擬出 500 個觀察值。我們可輕易地看出圖中有兩個局部極大值，其中一個的位置接近真實的參數值，而另一個的位置則位於 $(2, -0.5)$ 。我們隨機採用不同的起始值作五次EM演算法，其中三個EM演算法所得的序列會朝較大的局部極大值處靠近，而另外兩個則朝較小的局部極大值處靠近。後者主要是因為所給定的起始值正好落在較小的局部極大值的定義域裡。圖 4-7 上半部分表示此混合模型的對數概似函數值，這些函數值分別朝向上述兩個不同的局部極大值靠近，此問題可由增加疊代次數來改善。

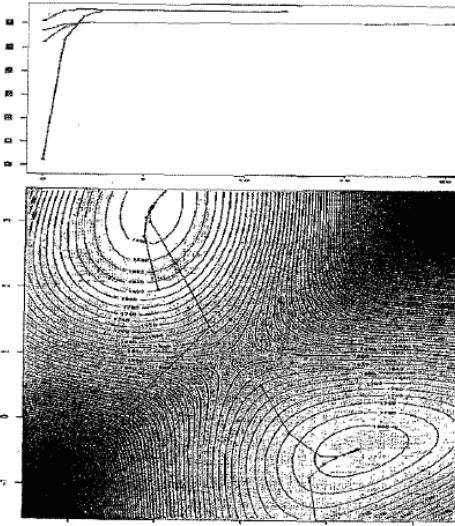


圖 4-7 上半部分表示此混合模型的對數概似函數值，
下半部分表示此模型的對數概似函數之表面圖形。

4.3.3 蒙地卡羅EM(Monte Carlo EM)

前一節所介紹的EM演算法中，因為每個”E步驟”都需要計算對數概似函數的期望值 $Q(\theta|\theta_0, x)$ ，所以有時在計算此期望值時將遇到一些困難。Wei 及 Tanner (1990a,b)建議用蒙地卡羅方法(MCEM)來克服此問題；也就是說，藉由從條件分佈 $k(z|x, \theta)$ 中生成 Z_1, \dots, Z_m 然後再將下面所列之近似的完備資料對數概似函數最大化

$$\hat{Q}(\theta|\theta_0, x) = \frac{1}{m} \sum_{i=1}^m \log L^c(\theta|x, z) \quad (4.18)$$

當 m 趨近於無限大，上述函數將收斂到 $Q(\theta|\theta_0, x)$ ，且此蒙地卡羅EM演算法的極限型式就是一般的EM演算法。雖然最大化(4.18)式一般還是蠻複雜的(因為(4.18)式是總和形式)，但指數族的設定將能使此問題得到封閉解。

例題4.20 刪失資料的MCEM

例題 4.17 的EM解為 $\hat{\theta} = \frac{m\bar{y} + (n-m)E_{\hat{\theta}'}(Z_1)}{n}$ 。相對於此EM數列

$$\hat{\theta}^{(j+1)} = \frac{m\bar{y} + (n-m)E_{\hat{\theta}^{(j)}}(Z_1)}{n}$$

MCEM的解為將 $E_{\hat{\theta}^{(j)}}(Z_1)$ 改成 $\frac{1}{M} \sum_{i=1}^M Z_i$, $Z_i \sim k(z|\hat{\theta}^{(j)}, y)$ 。圖 5.6 右側部分可見 MCEM 數列，其收斂速度不像 EM 數列那樣快速，且其波動大小決定於 M 的選取；較大的 M 值對應到的波動會較小。

例題4.21 遺傳連鎖(Genetic linkage)

遺傳的問題是EM演算法中一個典型的例子。設觀察值 (x_1, x_2, x_3, x_4) 來自多項式分佈(multinomial distribution)

$$\mathcal{M}(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$$

我們將 x_1 分成兩部分，並創造一個增廣的模型

$$(z_1, z_2, x_2, x_3, x_4) \sim \mathcal{M}(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4})$$

其中 $x_1 = z_1 + z_2$ 。此完備資料的概似函數可簡單的表示成 $\theta^{z_2+x_4}(1-\theta)^{x_2+x_3}$ ，已觀察到的資料之概似函數為 $(2+\theta)^{x_1}\theta^{x_4}(1-\theta)^{x_2+x_3}$ 。將完備資料之對數概似函數取期望值

$$\begin{aligned} Q(\theta|\theta_0, x) &\propto E_{\theta_0}[(Z_2 + x_4) \log \theta + (x_2 + x_3) \log(1 - \theta)] \\ &= \left(\frac{\theta_0}{2 + \theta_0}x_1 + x_4\right) \log \theta + (x_2 + x_3) \log(1 - \theta), \\ \text{其中 } Z_2 &\sim \text{Bin}(x_1, \frac{\theta_0}{2 + \theta_0}). \end{aligned}$$

令

$$\frac{\partial}{\partial \theta} Q(\theta|\theta_0, x) \propto \frac{\left(\frac{\theta_0}{2 + \theta_0}x_1 + x_4\right)}{\theta} - \frac{(x_2 + x_3)}{1 - \theta} = 0$$

整理後可得

$$\hat{\theta}_1 = \frac{\frac{\theta_0}{2 + \theta_0}x_1 + x_4}{\frac{\theta_0}{2 + \theta_0}x_1 + x_2 + x_3 + x_4}$$

如果改用蒙地卡羅EM演算法，則期望值 $E_{\theta_0}(Z_2) = \frac{\theta_0}{2 + \theta_0}x_1$ 將被取代成平均 $\bar{z}_m = \frac{1}{m} \sum_{i=1}^m z_i$ ，其中 z'_i s是從二項式分佈 $\text{Bin}(x_1, \frac{\theta_0}{2 + \theta_0})$ 模擬而得。由MCEM所得的結果為 $\hat{\theta}_1 = \frac{\bar{z}_m + x_4}{\bar{z}_m + x_2 + x_3 + x_4}$ 。

例題4.22 重複捕取模型回顧(Capture-recapture models revisited)

對一個重複捕取模型的概述，也就是假設某動物 $i, i = 1, 2, \dots, n$ 在時間 $j, j = 1, 2, \dots, t$ 時，在 m 個地點中的某一處可被捕捉，其中此 m 個地點呈現多項式分佈 $H \sim \mathcal{M}_m(\theta_1, \dots, \theta_m)$ ，當然此動物不一定被捉到(可能是因為沒被發現或是已經死亡)。當我們經由時間來追蹤各個動物時可藉由兩個隨機變數來將此過程建模。表示地點的隨機變數 H 是由集合 $\{1, 2, \dots, m\}$ 中取值，其發生機率對應於集合 $\{\theta_1, \dots, \theta_m\}$ 。給定 $H = k$ ，隨機變數 X 具白努力分佈($X \sim \mathcal{B}(p_k)$)，此 p_k 是指在地點 k 有捉到此動物的機率。

舉例來說，當 $t = 6$ 時，所觀察的捕捉地點 \mathbf{h} 可記作 $\mathbf{h} = (4, 1, -, 8, 3, -)$ ，捕捉結

果 \mathbf{x} 可記作 $\mathbf{x} = (1, 1, 0, 1, 1, 0)$ 。對某動物 i 而言，我們定義隨機變數

$$X_{ijk} = \begin{cases} 1 & \text{若在時間為 } j \text{、地點為 } k \text{ 時，有捕捉到動物 } i \\ 0 & \text{其他情況} \end{cases}$$

已觀察到的資料 $Y_{ijk} = \mathbb{I}(H_{ij} = k)\mathbb{I}(X_{ijk} = 1)$ ；刪失的資料 $Z_{ijk} = \mathbb{I}(H_{ij} = k)\mathbb{I}(X_{ijk} = 0)$ 。令 $X_{1jk} \sim Ber(p_k)$, $X_{2jk} \sim Ber(p_k)$, ..., $X_{njk} \sim Ber(p_k)$ ，在地點 k 處，完備資料的聯合密度函數為

$$p_k^{\sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times (1 - p_k)^{nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times \theta_k^{\sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + z_{ijk})}$$

綜合 m 個不同地點所得完備資料的概似函數為

$$\begin{aligned} L(\theta_1, \dots, \theta_m, p_1, \dots, p_m | \mathbf{y}, \mathbf{x}) &= \sum_{\mathbf{z}} L(\theta_1, \dots, \theta_m, p_1, \dots, p_m | \mathbf{y}, \mathbf{x}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} \prod_{k=1}^m [p_k^{\sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times (1 - p_k)^{nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times \theta_k^{\sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + z_{ijk})}] \end{aligned}$$

此式子看來是個複雜的期望值問題，不過第一步我們可以用EM方法整理，先得到完備資料的概似函數 $L(\theta_1, \dots, \theta_k, p_1, \dots, p_k | \mathbf{y}, \mathbf{x}, \mathbf{z})$ ，然後再用MCEM方法來計算期望值。作法如下：

完備資料之對數概似函數為

$$\begin{aligned} \log L(\theta_1, \dots, \theta_m, p_1, \dots, p_m | x, y, z) &= \sum_{k=1}^m \log [p_k^{\sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times (1 - p_k)^{nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}} \times \theta_k^{\sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + z_{ijk})}] \\ &= \sum_{k=1}^m \left[\sum_{i=1}^n \sum_{j=1}^t x_{ijk} \log p_k + (nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}) \log(1 - p_k) \right. \\ &\quad \left. + \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + z_{ijk}) \log \theta_k \right] \end{aligned}$$

上述完備資料之對數概似函數對 \mathbf{z} 取期望值為

$$\begin{aligned} Q(\theta_1, \dots, \theta_m, p_1, \dots, p_m | x, y) &= E[\log L(\theta_1, \dots, \theta_m, p_1, \dots, p_m | x, y, z)] \\ &= \sum_{k=1}^m \left[\sum_{i=1}^n \sum_{j=1}^t x_{ijk} \log p_k + (nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}) \log(1 - p_k) \right. \\ &\quad \left. + \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) \log \theta_k \right] \end{aligned}$$

因為 $\sum_{k=1}^m \theta_k = 1$ ，令

$$\begin{aligned} Q(\theta_1, \dots, \theta_m, p_1, \dots, p_m, \lambda | x, y) \\ = \sum_{k=1}^m \left[\sum_{i=1}^n \sum_{j=1}^t x_{ijk} \log p_k + (nt - \sum_{i=1}^n \sum_{j=1}^t x_{ijk}) \log(1 - p_k) \right. \\ \left. + \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) \log \theta_k \right] + \lambda (\sum_{k=1}^m \theta_k - 1) \end{aligned}$$

其中 λ 為拉格朗日乘數(Lagrange multiplier)。

將 $Q(\theta_1, \dots, \theta_m, p_1, \dots, p_m, \lambda | x, y)$ 分別對 θ_k, λ 微分，並令為分結果為 0，

$$\frac{\partial Q}{\partial \theta_k} = \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) \cdot \frac{1}{\theta_k} + \lambda = 0 \implies \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) + \lambda \theta_k = 0 \quad (a)$$

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} &= \sum_{k=1}^m \theta_k - 1 = 0 \implies \sum_{k=1}^m \theta_k = 1 \\ \therefore \sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) + \lambda \sum_{k=1}^m \theta_k &= 0 \\ \therefore \lambda &= - \sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) \end{aligned}$$

因為上式是對 Z 取期望值，而 y_{ijk} 可視為常數，即 $E(y_{ijk}) = y_{ijk}$ ，故

$$\begin{aligned} \lambda &= - \sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + E(z_{ijk})) = - \sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^t (E(y_{ijk}) + E(z_{ijk})) \\ &= -E[\sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + z_{ijk})] = -E[\sum_{i=1}^n \sum_{j=1}^t \sum_{k=1}^m I(H_{ij} = k)] \\ &= -E[\sum_{i=1}^n \sum_{j=1}^t 1] = -nt \end{aligned}$$

將 $\lambda = -nt$ 代回(a)，可解得

$$\hat{\theta}_k = \frac{1}{nt} \sum_{i=1}^n \sum_{j=1}^t [y_{ijk} + E(Z_{ijk})]$$

在 MCEM 方法中，我們將 $E(Z_{ijk})$ 取代成 $\hat{z}_{ijk} = \frac{1}{L} \sum_l^L \hat{z}_{ijkl}$ 可得

$$\hat{\theta}_k = \frac{1}{nt} \sum_{i=1}^n \sum_{j=1}^t [y_{ijk} + \hat{z}_{ijk}]$$

對於以上 $\theta_1, \dots, \theta_k$ 的估計，我們整理出下面的演算法。

演算法 8 重複捕取MCEM演算法 (Capture-recapture MCEM Algorithm)

1. (M步驟) 取 $\hat{\theta}_k = \frac{1}{nt} \sum_{i=1}^n \sum_{j=1}^t (y_{ijk} + \hat{z}_{ijk})$ ；
2. (蒙地卡羅E步驟) 若 $x_{ijk} = 0$ ，對 $l = 1, 2, \dots, L$ ，生成 $\hat{z}_{ijkl} \sim \mathcal{M}_k(\hat{\theta}_1, \dots, \hat{\theta}_m)$ ，並計算 $\hat{z}_{ijk} = \frac{1}{L} \sum_1^L \hat{z}_{ijkl}$ 。

最後，我們須注意到MCEM方法不再像EM方法一樣具有單調性(monotonicity)，甚至可能存在有是否平滑(smoothness)的問題。此外，若遇到更複雜的概似函數，可使用馬可夫蒙地卡羅方法來生成那些刪失的資料，並因此得到更多額外的相關結構。

4.3.4 EM 標準差(EM Standard Errors)

對於EM演算法的標準差而言，存在有許多演算法及公式來估算其標準差；其中由Oakes (1999)所提出的方法，配合蒙地卡羅的觀點，是既簡單又好用的方法。由最大概似估計量的一致性可得 $\hat{\theta}$ (θ的MLE)的標準差滿足

$$Var\hat{\theta} \approx \frac{1}{nI(\theta)}$$

其中 $I(\theta)$ 為 Fisher 信息 (Fisher information)。因為

$$\begin{aligned} I(\theta) &= -E \left[\frac{\partial^2}{\partial \theta^2} \log L(\theta|x) \right] \\ &\approx -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log L(\theta|x_i) = -\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{x}), \quad \mathbf{x} = (x_1, x_2, \dots, x_n) \end{aligned}$$

故 $\hat{\theta}$ 的標準差可視為

$$Var\hat{\theta} \approx \left[-\frac{\partial^2}{\partial \theta^2} \log L(\theta|x) \right]^{-1}$$

由(4.10)式我們知道 $k(z|\theta, x)$ 是遺失資料 Z 在給定已觀察得的資料 x 下的條件分佈，又 $\int_{-\infty}^{\infty} k(z|\theta, x) dz = 1$ ，假設對 θ 微分與對 z 積分是可以對調順序的，則

$$\begin{aligned} E \left[\frac{\partial \log k(z|\theta, x)}{\partial \theta} \right] &= \int_{-\infty}^{\infty} \left[\frac{\partial \log k(z|\theta, x)}{\partial \theta} \right] k(z|\theta, x) dz \\ &= \int_{-\infty}^{\infty} \left[\frac{\partial k(z|\theta, x)/\partial \theta}{k(z|\theta, x)} \right] k(z|\theta, x) dz = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} k(z|\theta, x) dz \\ &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} k(z|\theta, x) dz = 0 \end{aligned}$$

將上式繼續對 θ 做第二次偏微分

$$\begin{aligned}
& \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} \log k(z|\theta, x) k(z|\theta, x) dz + \int_{-\infty}^{\infty} \frac{\partial \log k(z|\theta, x)}{\partial \theta} \frac{\partial}{\partial \theta} k(z|\theta, x) dz = 0 \\
\implies & \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} \log k(z|\theta, x) k(z|\theta, x) dz \\
& + \int_{-\infty}^{\infty} \frac{\partial \log k(z|\theta, x)}{\partial \theta} \left[\frac{\partial k(z|\theta, x)}{\partial \theta} / k(z|\theta, x) \right] k(z|\theta, x) dz = 0 \\
\implies & E \left[\frac{\partial^2}{\partial \theta^2} \log k(z|\theta, x) \right] + E \left[\left(\frac{\partial}{\partial \theta} \log k(z|\theta, x) \right)^2 \right] = 0 \\
\implies & E \left[\left(\frac{\partial}{\partial \theta} \log k(z|\theta, x) \right)^2 \right] = -E \left[\frac{\partial^2}{\partial \theta^2} \log k(z|\theta, x) \right]
\end{aligned}$$

承接先前(4.11)、(4.12)式， $\log L(\theta'|x) = Q(\theta|\theta', x) - E[\log k(z|\theta', x)]$ ，

將 $\log L(\theta'|x) = Q(\theta|\theta', x) - E[\log k(z|\theta', x)]$ 先對 θ' 做二次微分

$$\begin{aligned}
\frac{\partial \log L(\theta'|x)}{\partial \theta'} &= \frac{\partial Q(\theta|\theta', x)}{\partial \theta'} - \frac{\partial}{\partial \theta'} E[\log k(z|\theta', x)] \\
\implies \frac{\partial \log L(\theta'|x)}{\partial \theta'} &= \frac{\partial Q(\theta|\theta', x)}{\partial \theta'} - E \left[\frac{\partial \log k(z|\theta', x)}{\partial \theta'} \right]
\end{aligned} \tag{a}$$

其中

$$\begin{aligned}
& \frac{\partial}{\partial \theta'} E[\log k(z|\theta', x)] \\
&= \frac{\partial}{\partial \theta'} \int_{-\infty}^{\infty} \log k(z|\theta', x) k(z|\theta, x) dz \\
&= \int_{-\infty}^{\infty} \frac{\partial \log k(z|\theta', x)}{\partial \theta'} k(z|\theta, x) dz \\
&= E \left[\frac{\partial \log k(z|\theta', x)}{\partial \theta'} \right]
\end{aligned}$$

將(a)繼續對 θ' 做第二次偏微分

$$\frac{\partial^2 \log L(\theta'|x)}{\partial \theta'^2} = \frac{\partial^2 Q(\theta|\theta', x)}{\partial \theta'^2} - E \left[\frac{\partial^2 \log k(z|\theta', x)}{\partial \theta'^2} \right] \tag{b}$$

此外，將 $\log L(\theta'|x) = Q(\theta|\theta', x) - E[\log k(z|\theta', x)]$ 對 θ 做第一次偏微分

$$\begin{aligned}
\frac{\partial \log L(\theta'|x)}{\partial \theta} &= \frac{\partial Q(\theta|\theta', x)}{\partial \theta} - \frac{\partial}{\partial \theta} E[\log k(z|\theta', x)] \\
\implies \frac{\partial \log L(\theta'|x)}{\partial \theta} &= \frac{\partial Q(\theta|\theta', x)}{\partial \theta} - E \left[\log k(z|\theta', x) \frac{\partial \log k(z|\theta, x)}{\partial \theta} \right]
\end{aligned} \tag{c}$$

其中

$$\frac{\partial \log L(\theta'|x)}{\partial \theta} = 0$$

且

$$\begin{aligned}
& \frac{\partial}{\partial \theta} E[\log k(z|\theta', x)] \\
&= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \log k(z|\theta', x) k(z|\theta, x) dz \\
&= \int_{-\infty}^{\infty} \log k(z|\theta', x) \frac{\partial k(z|\theta, x)}{\partial \theta} dz \\
&= \int_{-\infty}^{\infty} \log k(z|\theta', x) \frac{\partial k(z|\theta, x)/\partial \theta}{k(z|\theta, x)} k(z|\theta, x) dz \\
&= \int_{-\infty}^{\infty} \log k(z|\theta', x) \frac{\partial \log k(z|\theta, x)}{\partial \theta} k(z|\theta, x) dz \\
&= E \left[\log k(z|\theta', x) \frac{\partial \log k(z|\theta, x)}{\partial \theta} \right]
\end{aligned}$$

(c) 繼續對 θ' 做第二次偏微分

$$\begin{aligned}
& \frac{\partial}{\partial \theta'} \left(\frac{\partial \log L(\theta'|x)}{\partial \theta} \right) = \frac{\partial^2 Q(\theta|\theta', x)}{\partial \theta' \partial \theta} - \frac{\partial}{\partial \theta'} E \left[\log k(z|\theta', x) \frac{\partial \log k(z|\theta, x)}{\partial \theta} \right] \\
\implies & 0 = \frac{\partial^2 Q(\theta|\theta', x)}{\partial \theta' \partial \theta} - E \left[\frac{\partial \log k(z|\theta', x)}{\partial \theta'} \frac{\partial \log k(z|\theta, x)}{\partial \theta} \right] \tag{d}
\end{aligned}$$

將(b)、(d)式中 θ' 代換成 θ ，則

$$\begin{aligned}
& \frac{\partial^2 \log L(\theta'|x)}{\partial \theta'^2} \Big|_{\theta'=\theta} = \frac{\partial^2 \log L(\theta|x)}{\partial \theta^2}, \\
& E \left[\frac{\partial^2 \log k(z|\theta', x)}{\partial \theta'^2} \right] \Big|_{\theta'=\theta} = E \left[\frac{\partial^2 \log k(z|\theta, x)}{\partial \theta^2} \right] = -E \left[\left(\frac{\partial \log k(z|\theta, x)}{\partial \theta} \right)^2 \right], \\
& E \left[\frac{\partial \log k(z|\theta', x)}{\partial \theta'} \frac{\partial \log k(z|\theta, x)}{\partial \theta} \right] \Big|_{\theta'=\theta} = E \left[\left(\frac{\partial \log k(z|\theta, x)}{\partial \theta} \right)^2 \right],
\end{aligned}$$

且

$$\begin{aligned}
(b) + (d) \implies & \frac{\partial^2 \log L(\theta|x)}{\partial \theta^2} = \frac{\partial^2 Q(\theta|\theta', x)}{\partial \theta'^2} + \frac{\partial^2 Q(\theta|\theta', x)}{\partial \theta' \partial \theta} \\
\implies & \frac{\partial^2}{\partial \theta^2} \log L(\theta|x) = \left\{ \frac{\partial^2}{\partial \theta'^2} E[\log L(\theta'|x, z)] + \frac{\partial^2}{\partial \theta' \partial \theta} E[\log L(\theta'|x, z)] \right\} \Big|_{\theta'=\theta} \tag{4.19}
\end{aligned}$$

以上為 Oakes (1999) 證明出的結果，這將可以幫助我們計算EM演算法所得的MLE之標準差。

對蒙地卡羅 EM 演算法而言，(4.19)式必須把期望值提到外面，這樣才能用蒙地卡羅法改成求和的形式。所以我們分析一下(4.19)式內容

$$\frac{\partial^2}{\partial \theta'^2} E[\log L(\theta'|x, z)]|_{\theta'=\theta} = E \left(\frac{\partial^2}{\partial \theta'^2} \log L(\theta'|x, z) \right) |_{\theta'=\theta} = E \left(\frac{\partial^2}{\partial \theta^2} \log L(\theta|x, z) \right) \quad (\text{e})$$

又

$$\begin{aligned} \frac{\partial}{\partial \theta} E[\log L(\theta'|x, z)] &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \log L(\theta'|x, z) k(z|\theta, x) dz \\ &= \int_{-\infty}^{\infty} \log L(\theta'|x, z) \frac{\partial k(z|\theta, x)/\partial \theta}{k(z|\theta, x)} k(z|\theta, x) dz \\ &= \int_{-\infty}^{\infty} \log L(\theta'|x, z) \frac{\partial \log k(z|\theta, x)}{\partial \theta} k(z|\theta, x) dz \\ &= \int_{-\infty}^{\infty} \log L(\theta'|x, z) \left(\frac{\partial \log L(\theta|x, z)}{\partial \theta} - \frac{\partial \log L(\theta|x)}{\partial \theta} \right) k(z|\theta, x) dz \end{aligned}$$

$$\begin{aligned} &\frac{\partial^2}{\partial \theta' \partial \theta} E[\log L(\theta'|x, z)]|_{\theta'=\theta} \\ &= \left\{ \int_{-\infty}^{\infty} \frac{\partial \log L(\theta'|x, z)}{\partial \theta'} \left(\frac{\partial \log L(\theta|x, z)}{\partial \theta} - \frac{\partial \log L(\theta|x)}{\partial \theta} \right) k(z|\theta, x) dz \right\} |_{\theta'=\theta} \\ &= \int_{-\infty}^{\infty} \left(\frac{\partial \log L(\theta|x, z)}{\partial \theta} \right)^2 k(z|\theta, x) dz - \frac{\partial \log L(\theta|x)}{\partial \theta} \int_{-\infty}^{\infty} \frac{\partial \log L(\theta|x, z)}{\partial \theta} k(z|\theta, x) dz \\ &= \int_{-\infty}^{\infty} \left(\frac{\partial \log L(\theta|x, z)}{\partial \theta} \right)^2 k(z|\theta, x) dz - \left(\int_{-\infty}^{\infty} \frac{\partial \log L(\theta|x, z)}{\partial \theta} k(z|\theta, x) dz \right)^2 \\ &= E \left[\left(\frac{\partial}{\partial \theta} \log L(\theta|x, z) \right)^2 \right] - \left[E \left(\frac{\partial}{\partial \theta} \log L(\theta|x, z) \right) \right]^2 \quad (\text{f}) \end{aligned}$$

由(e)、(f)，可將(4.19)重新改寫成

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \log L(\theta|x) &= E \left(\frac{\partial^2}{\partial \theta^2} \log L(\theta|x, z) \right) \\ &\quad + E \left[\left(\frac{\partial}{\partial \theta} \log L(\theta|x, z) \right)^2 \right] - \left[E \left(\frac{\partial}{\partial \theta} \log L(\theta|x, z) \right) \right]^2 \\ &= E \left(\frac{\partial^2}{\partial \theta^2} \log L(\theta|x, z) \right) + var \left(\frac{\partial}{\partial \theta} \log L(\theta|x, z) \right) \quad (4.20) \end{aligned}$$

將上式加入蒙地卡羅估計

$$\frac{\partial^2}{\partial \theta^2} \log L(\theta|x) = \frac{1}{M} \sum_{j=1}^M \frac{\partial^2}{\partial \theta^2} \log L(\theta|x, z^{(j)})$$

$$+ \frac{1}{M} \sum_{j=1}^M \left(\frac{\partial}{\partial \theta} \log L(\theta|x, z^{(j)}) - \frac{1}{M} \sum_{j'=1}^M \frac{\partial}{\partial \theta} \log L(\theta|x, z^{(j')}) \right)^2,$$

其中 $z^{(j)}, j = 1, \dots, M$ 由刪失資料的分佈模擬而得。

例題4.23 遺傳連鎖問題的標準差 (Genetic linkage standard errors)

承接例題 4.21，我們已知完備資料的概似函數為 $\theta^{z_2+x_4}(1-\theta)^{x_2+x_3}$ ，將完備資料的概似函數取對數，並對 θ 分別做一階及二階偏微分

$$\begin{aligned}\log L(\theta|\mathbf{x}, \mathbf{z}) &= (z_2 + x_4) \log \theta + (x_2 + x_3) \log(1 - \theta), \\ \frac{\partial}{\partial \theta} \log L(\theta|\mathbf{x}, \mathbf{z}) &= \frac{z_2 + x_4}{\theta} - \frac{x_2 + x_3}{1 - \theta}, \\ \frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{x}, \mathbf{z}) &= -\frac{z_2 + x_4}{\theta^2} + \frac{x_2 + x_3}{(1 - \theta)^2}\end{aligned}$$

則引用 (4.20) 式的寫法得到

$$\frac{\partial^2}{\partial \theta^2} \log L(\theta|x) = \frac{-(1-\theta)^2 E(Z_2) - x_4(1-\theta)^2 + (x_2+x_3)\theta^2}{\theta^2(1-\theta)^2} + \frac{var(Z_2)}{\theta^2}$$

其中 $E(Z_2) = \frac{x_1\theta}{(2+\theta)}$, $var(Z_2) = \frac{2\theta x_1}{(2+\theta)^2}$ 。實際上，我們會利用最大概似估計量 $\hat{\theta}$ 的收斂值來估計此期望值，結果如圖 4-8。同時，我們計算 $\hat{\theta}$ 的變異數

$$Var\hat{\theta} \approx \left[\frac{(1-\theta)^2 E(Z_2) + x_4(1-\theta)^2 - (x_2+x_3)\theta^2}{\theta^2(1-\theta)^2} - \frac{var(Z_2)}{\theta^2} \right]^{-1}.$$

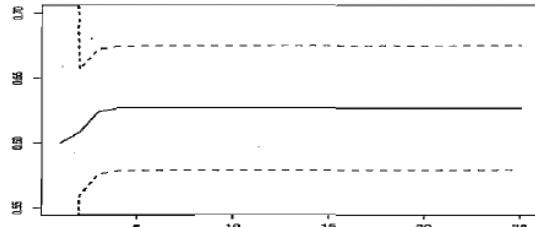


圖 4-8 遺傳連鎖資料之參數的 EM 估計結果。

5 最佳化問題之模擬研究：捨入資料對最大概似估計量的影響

由於平時數據的調查者，常會為了方便記錄而將數據做四捨五入，當後人欲分析這些數據時，通常都不是最原始的數據，也就是所謂的捨入資料，此時在統計分析上若想估算某些感興趣的參數之最大概似估計量，可能必須考慮捨入資料對估計結果的影響。我們利用第四章介紹的模擬退火演算法，對常態分佈、科西分佈、 t 分佈、指數分佈及伽瑪分佈的參數之最大概似估計量進行捨入資料對估計結果的討論。

5.1 概似函數之參數估計

原始隨機變數 $\mathbf{X} = (X_1, \dots, X_n)$ ，相對應的捨入後隨機變數為 $\mathbf{Y} = (Y_1, \dots, Y_n)$ ，其中

$$Y_t = y_t \quad \text{iff} \quad X_t = x_t \quad \text{且} \quad x_t \in \left[y_t - \frac{h}{2}, y_t + \frac{h}{2} \right], \quad t = 1, \dots, n$$

上述 h 表示捨入的區間寬度。

令

$$Y_t = X_t + U_t$$

其中隨機變數 U_t 表示捨入誤差，且 U_t s.i.i.d. $\sim \mathcal{U}\left[\frac{-h}{2}, \frac{h}{2}\right]$ 。

若可取得真實資料 $\mathbf{x} = (x_1, \dots, x_n)$ ，其對應的聯合密度函數為 $f(\mathbf{x}; \beta)$ ，則 β 的最大概似估計量即為使 $f(\mathbf{x}; \beta)$ 產生最大值的 β 。多數情況通常只能得到捨入資料 $\mathbf{y} = (y_1, \dots, y_n)$ ，此時由忽略捨入效應的聯合密度函數 $f(\mathbf{y}; \beta)$ ，解方程式 $\frac{\partial \ln f(\mathbf{y}; \beta)}{\partial \beta} = 0$ ，得到的解稱為 β 的偽最大概似估計量 $\hat{\beta}_0$ (pseudo mle)。

若考慮捨入效應，由捨入資料 $\mathbf{y} = (y_1, \dots, y_n)$ 得到 β 的真實概似函數為

$$L(\beta | \mathbf{y}) = h^{-n} \int_{y_n-h/2}^{y_n+h/2} \cdots \int_{y_1-h/2}^{y_1+h/2} f(\mathbf{u}; \beta) d u_1 \cdots d u_n$$

將 $L(\beta | \mathbf{y})$ 取對數，則使對數概似函數 $L_I(\beta | \mathbf{y})$ 最大化的 β 即為 β 的最大概似估計量 $\hat{\beta}_I$ ，但此處 n 重積分的形式導致計算上十分困難。

Lindley (1950) 在 f 是單一變數分佈函數的情況，將概似函數 $L(\beta | \mathbf{y})$ 推導成在 $h = 0$ 處的 Maclaurin 展式。Tallis (1967) 將 Lindley 的 Maclaurin 展式推廣至下列多變數的

情況

$$L(\beta|\mathbf{y}) = f(\mathbf{y}; \beta) + \frac{h^2}{24} \sum_{t=1}^n \frac{\partial^2 f(\mathbf{y}|\beta)}{\partial y_t^2} + O(h^3)$$

將上述 Tallis 展式取對數，可得到真實對數概似函數的近似

$$L_T(\beta|\mathbf{y}) = \ln [L(\beta|\mathbf{y})] \approx \ln f(\mathbf{y}; \beta) + \ln \left[1 + \frac{h^2}{24} \sum_{t=1}^n \frac{\partial^2 f(\mathbf{y}|\beta)/\partial y_t^2}{f(\mathbf{y}; \beta)} \right]$$

若以 pseudo mle $\hat{\beta}_0$ 為起始點，利用牛頓-拉福生法疊代一次，可得 $\hat{\beta}_0$ 的修正

$$\hat{\beta}_A = \hat{\beta}_0 - A^{-1}b$$

其中若參數 $\beta = (\beta_1, \dots, \beta_p)$ ，則上述 $A = [a_{ij}]$, $b = [b_j]$ ，且

$$\begin{aligned} a_{ij} &= \left. \frac{\partial^2 L_T(\beta|\mathbf{y})}{\partial \beta_i \partial \beta_j} \right|_{\beta=\hat{\beta}_0}, \quad i, j = 1, \dots, p; \\ b_j &= \left. \frac{\partial L_T(\beta|\mathbf{y})}{\partial \beta_j} \right|_{\beta=\hat{\beta}_0}, \quad j = 1, \dots, p. \end{aligned}$$

我們將前述三種對數概似函數 $L_0(\beta|\mathbf{y}) = \log f(\mathbf{y}; \beta)$ 、 $L_I(\beta|\mathbf{y})$ 、 $L_T(\beta|\mathbf{y})$ 分別利用模擬退火演算法來找常態分佈、柯西分佈、指數分佈及伽瑪分佈的參數之最大概似估計量並比較所得結果的差異性；此外，我們也計算 $\hat{\beta}_0$ 、 $\hat{\beta}_A$ 的值來和模擬退火演算法所得結果做比較。

5.1.1 常態分佈的參數估計

公式推導

常態分佈的聯合機率密度函數：

$$f(\mathbf{x}|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^2)}$$

捨入資料(\mathbf{Y})的偽對數概似函數：

$$L_0(\mu, \sigma|\mathbf{y}) = -n \log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}$$

捨入資料(\mathbf{Y})的真實對數概似函數：

$$\begin{aligned} L_I(\mu, \sigma|\mathbf{y}) &= \log \left[\int_{-0.5}^{0.5} \cdots \int_{-0.5}^{0.5} \frac{1}{\sqrt{2\pi}\sigma} e^{-\sum_{i=1}^n (y_i + h_i - \mu)^2 / (2\sigma^2)} dh_1 \cdots dh_n \right] \\ &= \sum_{i=1}^n \log \left[\Phi \left(\frac{0.5 - (\mu - y_i)}{\sigma} \right) - \Phi \left(\frac{-0.5 - (\mu - y_i)}{\sigma} \right) \right] \end{aligned}$$

$L_I(\mu, \sigma | \mathbf{y})$ 的近似(Tallis 展式)：

$$\begin{aligned} L_T(\mu, \sigma | \mathbf{y}) &\approx \log(f(\mathbf{y} | \mu, \sigma)) + \log[1 + \frac{1}{24} \sum_{i=1}^n \frac{\partial^2 f(\mathbf{y} | \mu, \sigma) / \partial y_i^2}{f(\mathbf{y} | \mu, \sigma)}] \\ &= -n \log(\sqrt{2\pi}\sigma) + \frac{1 - 12\sigma^2}{24\sigma^4} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{24\sigma^2} \end{aligned}$$

pseudo mle : $\hat{\beta}_0 = (\hat{\mu}_0, \hat{\sigma}_0) = (\bar{\mathbf{y}}, \sqrt{\frac{n-1}{n}}S)$ ，其中 $S = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{\mathbf{y}})^2}{n-1}}$ 為樣本標準差；
牛頓拉福生法所得 mle : $\hat{\beta}_A = (\hat{\mu}_A, \hat{\sigma}_A)$

$$\begin{aligned} \hat{\mu}_A &= \bar{\mathbf{y}}; \hat{\sigma}_A^2 &= \hat{\sigma}_0^2 - \frac{24\hat{\sigma}_0^4}{96\hat{\sigma}_0^2 - 288\hat{\sigma}_0^4 - n} \end{aligned}$$

5.1.2 柯西分佈的參數估計

公式推導

柯西分佈的聯合機率密度函數：

$$f(\mathbf{x} | \theta, \gamma) = (\pi\gamma)^{-n} \prod_{i=1}^n \left[1 + \left(\frac{x_i - \theta}{\gamma} \right)^2 \right]^{-1}$$

捨入資料(\mathbf{Y})的偽對數概似函數：

$$L_0(\theta, \gamma | \mathbf{y}) = -n \log(\pi\gamma) - \sum_{i=1}^n \log \left[1 + \left(\frac{y_i - \theta}{\gamma} \right)^2 \right]$$

捨入資料(\mathbf{Y})的真實對數概似函數：

$$\begin{aligned} L_I(\theta, \gamma | \mathbf{y}) &= \log \left[\int_{-0.5}^{0.5} \cdots \int_{-0.5}^{0.5} (\pi\gamma)^{-n} \prod_{i=1}^n \left[1 + \left(\frac{y_i + h_i - \theta}{\gamma} \right)^2 \right]^{-1} dh_1 \cdots dh_n \right] \\ &= \log \left[\pi^{-n} \prod_{i=1}^n \left[\tan^{-1} \left(\frac{y_i + 0.5 - \theta}{\gamma} \right) - \tan^{-1} \left(\frac{y_i - 0.5 - \theta}{\gamma} \right) \right] \right] \\ &= -n \log \pi + \sum_{i=1}^n \log \left[\tan^{-1} \left(\frac{y_i + 0.5 - \theta}{\gamma} \right) - \tan^{-1} \left(\frac{y_i - 0.5 - \theta}{\gamma} \right) \right] \end{aligned}$$

$L_I(\theta, \gamma | \mathbf{y})$ 的近似(Tallis 展式)：

$$\begin{aligned} L_T(\theta, \gamma | \mathbf{y}) &\approx -n \log \pi\gamma - \sum_{i=1}^n \log \left[1 + \left(\frac{y_i - \theta}{\gamma} \right)^2 \right] \\ &\quad + \log \left[1 + \frac{1}{24} \sum_{i=1}^n \frac{2}{\gamma^2} \frac{3[(y_i - \theta)/\gamma]^2 - 1}{[1 + ((y_i - \theta)/\gamma)^2]^2} \right] \end{aligned}$$

5.1.3 指數分佈的參數估計

公式推導

指數分佈的聯合機率密度函數：

$$f(\mathbf{x}|\lambda) = \lambda^{-n} e^{-\sum_{i=1}^n x_i/\lambda}$$

捨入資料(\mathbf{Y})的偽對數概似函數：

$$L_0(\lambda|\mathbf{y}) = -n \log \lambda - \frac{\sum_{i=1}^n y_i}{\lambda}$$

捨入資料(\mathbf{Y})的真實對數概似函數：

$$\begin{aligned} L_I(\lambda|\mathbf{y}) &= \log \left[\int_{-0.5}^{0.5} \cdots \int_{-0.5}^{0.5} \lambda^{-n} e^{-\sum_{i=1}^n (y_i + h_i)/\lambda} dh_1 \cdots dh_n \right] \\ &= \sum_{i=1}^n \log [e^{-(y_i - 0.5)/\lambda} - e^{-(y_i + 0.5)/\lambda}] \end{aligned}$$

$L_I(\lambda|\mathbf{y})$ 的近似(Tallis 展式)：

$$\begin{aligned} L_T(\mu, \sigma|\mathbf{y}) &\approx -n \log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \\ &\quad + \log \left[1 + \frac{1}{24} \left(\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^4} - \frac{n}{\sigma^2} \right) \right] \end{aligned}$$

pseudo mle : $\hat{\beta}_0 = \hat{\lambda}_0 = \bar{\mathbf{y}}$;

牛頓拉福生法所得 mle :

$$\hat{\beta}_A = \hat{\lambda}_A = \bar{\mathbf{y}} - \left[\frac{-n}{\bar{\mathbf{y}}^2} + \frac{2n^2 + 144n\bar{\mathbf{y}}^2}{(24\bar{\mathbf{y}}^3 + n\bar{\mathbf{y}})^2} \right]^{-1} \left(\frac{-2n}{24\bar{\mathbf{y}}^3 + n\bar{\mathbf{y}}} \right).$$

5.1.4 Gamma 分佈的參數估計

公式推導

Gamma 分佈的聯合機率密度函數：

$$f(\mathbf{x}|\alpha, \beta) = \frac{\prod_{i=1}^n x_i^{\alpha-1} e^{-\sum_{i=1}^n x_i/\beta}}{(\Gamma(\alpha)\beta^\alpha)^n}$$

捨入資料(\mathbf{Y})的對數概似函數：

$$L_0(\alpha, \beta|\mathbf{y}) = -n \log(\Gamma(\alpha)\beta^\alpha) + (\alpha - 1) \sum_{i=1}^n \log y_i - \frac{\sum_{i=1}^n y_i}{\beta}$$

捨入資料(\mathbf{Y})的真實對數概似函數：

$$L_I(\alpha, \beta|\mathbf{y}) = \log \left[(\Gamma(\alpha)\beta^\alpha)^{-n} \int_{-0.5}^{0.5} \cdots \int_{-0.5}^{0.5} (y_i + h_i)^{\alpha-1} e^{-(y_i + h_i)/\beta} dh_1 \cdots dh_n \right]$$

$$\begin{aligned}
&= \log \left[(\Gamma(\alpha)\beta^\alpha)^{-n} \prod_{i=1}^n \beta^\alpha \int_{(-0.5+y_i)/\beta}^{(0.5+y_i)/\beta} t_i^{\alpha-1} e^{-t_i} dt_i \right] \\
&= \left[\Gamma(\alpha)^{-n} \prod_{i=1}^n (\Gamma(\alpha, (-0.5+y_i)/\beta) - \Gamma(\alpha, (0.5+y_i)/\beta)) \right] \\
&= -n \log \Gamma(\alpha) + \sum_{i=1}^n \log [\Gamma(\alpha, (-0.5+y_i)/\beta) - \Gamma(\alpha, (0.5+y_i)/\beta)]
\end{aligned}$$

其中 $\Gamma(\alpha, (-0.5+y_i)/\beta) = \int_{(-0.5+y_i)/\beta}^{\infty} t_i^{\alpha-1} e^{-t_i} dt_i$ 表示 Gamma function $L_I(\alpha, \beta|\mathbf{y})$ 的近似(Tallis 展式)：

$$\begin{aligned}
L_T(\alpha, \beta|\mathbf{y}) &\approx -n \log(\Gamma(\alpha)\beta^\alpha) + (\alpha-1) \sum_{i=1}^n \log y_i - \frac{\sum_{i=1}^n y_i}{\beta} \\
&\quad + \log \left[1 + \frac{1}{24} \sum_{i=1}^n ((\alpha-1)(\alpha-2)y_i^{-2} - \frac{2}{\beta}(\alpha-1)y_i^{-1} + \frac{1}{\beta^2}) \right]
\end{aligned}$$

5.2 模擬結果

以下各種分佈，我們一致取樣本數為 150 個、模擬退火法疊代次數為 750 次、重複模擬 50 次，依據模擬所得圖示(附於附錄中，請參考圖 5-1-1 ~ 5-4-4)，模擬結果整理如下。

(一) 常態分佈 $\mathcal{N}(\mu, \sigma^2)$

	Bias	Variance
$\hat{\mu}$	σ 大: $\mu_T > \mu_0 = \mu_A > \mu_I$	σ 大: $\mu_0 = \mu_A \geq \mu_I \approx \mu_T$
	σ 小: $\mu_T \approx \mu_0 = \mu_A > \mu_I$	σ 小: $\mu_0 = \mu_A \geq \mu_I \approx \mu_T$
$\hat{\sigma}$	σ 大: $\sigma_0 > \sigma_A \approx \sigma_I > \sigma_T$	σ 大: $\sigma_0 = \sigma_A \geq \sigma_I \approx \sigma_T$
	σ 小: $\sigma_I > \sigma_T \approx \sigma_A > \sigma_0$	σ 小: $\sigma_0 = \sigma_A > \sigma_T > \sigma_I$

上表中， $>$ 表示前者表現優於後者(並非指數值上前者大於後者)，又各下標表示意思為： T : by Tallis approximation ; I : by Integration ; A : by 牛頓法 ; 0 : Pseudo mle。

(二) 柯西分佈 $\text{Cauchy}(\theta, \gamma)$ ，位置參數 θ ，尺度參數 γ

	Bias	Variance
$\hat{\theta}$	γ 大: $\theta_I > \theta_T \geq \theta_0$	γ 大: $\theta_I \approx \theta_T \approx \theta_0$
	γ 小: $\theta_0 > \theta_I > \theta_T$	γ 小: $\theta_0 > \theta_I > \theta_T$
$\hat{\gamma}$	γ 大: $\gamma_T > \gamma_I > \gamma_0$	γ 大: $\gamma_I \geq \gamma_T > \gamma_0$
	γ 小: $\gamma_I > \gamma_T > \gamma_0$	γ 小: $\gamma_I > \gamma_T > \gamma_0$

上表中， \succ 表示前者表現優於後者(並非指數值上前者大於後者)，又各下標表示： T ： by Tallis approximation； I ： by Integration； 0 ： Pseudo mle。

(三) 指數分佈 $Exp(\lambda)$

$\hat{\lambda}$	Bias	Variance
	λ 大: $\lambda_0 > \lambda_A \approx \lambda_T > \lambda_I$	λ 大: $\lambda_0 = \lambda_A \approx \lambda_I \approx \lambda_T$
	λ 小: $\lambda_0 \approx \lambda_A > \lambda_T > \lambda_I$	λ 小: $\lambda_0 = \lambda_A > \lambda_T > \lambda_I$

上表中， \succ 表示前者表現優於後者(並非指數值上前者大於後者)，又各下標表示意思為： T ： by Tallis approximation； I ： by Integration； A ： by 牛頓法； 0 ： Pseudo mle。

(四) Gamma 分佈 $Gam(\alpha, \beta)$

	Bias	Variance
$\hat{\alpha}$	β 大: $\alpha_0 \geq \alpha_I > \alpha_T$	β 大: $\alpha_I \approx \alpha_T \geq \alpha_0$
	β 小: 均不佳	β 小: 均不佳
$\hat{\beta}$	β 大: $\beta_0 > \beta_I \approx \beta_T$	β 大: $\beta_I \geq \beta_T \geq \beta_0$
	β 小: $\beta_0 \approx \beta_I \geq \beta_T$	β 小: $\beta_0 \geq \beta_T > \beta_I$

上表中， \succ 表示前者表現優於後者(並非指數值上前者大於後者)，又各下標表示： T ： by Tallis approximation； I ： by Integration； 0 ： Pseudo mle。

參考文獻

- [1] Aarts, E. and Kors, T. (1989). *Simulated Annealing and Boltzman Machines: A Stochastic Approach to Combinatorial Optimisation and Neural Computing*. John Wiley, New York.
- [2] Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion). *Statist. Science*, 7:131-177.
- [3] Ahrens. J. and Dieter, U. (1974). Computer methods for sampling from gamma, beta, Poisson and binomial distributions. *Computing*, 12:223-246.
- [4] Aldous, D. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing. *Pr. Eng. Inform. Sciences*, 1:33-46.
- [5] Atkinson, A. (1979). The computer generation of Poisson random variables. *Appl. Statist.*, 28:29-35.
- [6] Ball, F., Cai, Y., Kadane, J., and O'Hagan, A. (1999). Bayesian inference for ion channel gating mechanisms directly from single channel recordings, using Markov chain Monte Carlo. *Proc. Royal Society London A*, 455:2879-2932.
- [7] Barnett, G., Kohn, R., and Sheather, S. (1996). Bayesian estimation of an autoregressive model using Markov chian Monte Carlo. *J. Econometrics*, 74:237-254.
- [8] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition.
- [9] Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference*, 25:303- 328.
- [10] Berger, J. (1994). An overview of robust Bayesian analysis(with discussion). *TEST*, 3:5-124.
- [11] Berger, J. and Bernardo, J. (1992). On the development of the reference prior method. In Berger, J., Bernardo, J., Dawid, A., and Smith, A., editors, *Bayesian Statistic 4*, pages 35-49. Oxford University Press, London.
- [12] Berger, J. and Pericchi, L. (1998). Accurate and stable Bayesian model selection: the median intrinsic Bayes factor. *Sankhya B*, 60:1-18.

- [13] Berger, J., Philippe, A., and Robert, C. (1998). Estimation of quadratic functions: reference priors for non-centrality parameters. *Statistica Sinica*, 8(2):359-375.
- [14] Bernardo, J. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Royal Statist. Soc. Series B*, 41:113- 147.
- [15] Bernardo, J. and *Giròn*, F. (1986). A Bayesian approach to cluster analysis. In *Second Catalan International Symposium on Statistics, Barcelona*, Spain.
- [16] Bernardo, J. and Giron, F. (1988). A Bayesian analysis of simple mixture problems. In Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., editors, *Bayesian Statistics 3*, pages 67-78. Oxford University Press, Oxford.
- [17] Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.
- [18] Berthelsen, K. and Moller, J. (2003). Likelihood and non-parametric Bayesian MCMC infcrence for spatial point processes based on perfect simulation and path sampling. *Scandinavian J. Statist.*, 30:549-564.
- [19] Billio, M., Monfort, A., and Robert, C. (1998). The simulated likelihood ratio method. Technical Report 9821, CREST, INSEE, Paris.
- [20] Box. G. and Muller, M. (1958). A note on the generation of random normal variates. *Ann. Mathemat. Statist.*, 29:610-611.
- [21] Boyles, R. (1983). On the convergence of the EM algorithm. *J. Royal Statist. Soc. Series B*, 45:47-50.
- [22] Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *J. American Statist. Assoc.*, 88:9-25.
- [23] Carlin, B. and Louis, T. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York, second edition.
- [24] Casella, G., and Berger, R. (2001). *Statistical Inference*. Wadsworth, Belmont, CA.
- [25] Castledine, B. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67:197-210.

- [26] Chen, M. and Shao, Q. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, 25:1563-1594.
- [27] Cheng, R. (1977). The generation of gamma variables with non-integral shape parameter. *Applied Statistics (Ser. C)*, 26:71-75.
- [28] Cheng, R. and Fest, G. (1979). Some simple gamma variate generators. *Appl. Statist.*, 28:290-295.
- [29] Cipra, B. (1987). An introduction to the Ising model. *American Mathematical Monthly*, 94:937-959.
- [30] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Soc. Series B*, 39:1-38.
- [31] D'Epifanio. (1996). Notes on a recursive procedure for point estimate. *TEST*, 5:203-225.
- [32] Devroye, L. (1981). The computer generation of Poisson random variables. *Computing*, 26:197-207.
- [33] Devroye, L. (1985). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- [34] Eberly, L. E. and Casella, G. (2003). Estimating Bayesian credible intervals. *J. Statist. Plann. Inference*, 112:115-132.
- [35] Evans, M. and Swartz, T. (1995). Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems. *Statist. Science*, 10:254-272.
- [36] Geweke, J. (1991). Efficient simulation from the multivariate normal and student t-distributions subject to linear constraints. *Computer Sciences and Statistics: Proc. 23d Symp. Interface*.
- [37] Geyer, C. (1996). Estimation and Optimization of functions. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in Practice*, pages 241-258. Chapman and Hall, New York.

- [38] Haario, H. and Sacksman, E. (1991). Simulated annealing in general state space. *Adv. Appl. Probab.*, 23:866-893.
- [39] H  jek, B. (1988). Cooling schedules for optimal annealing. *Math. Operation. Research*, 13:311-329.
- [40] Hesterberg, T. (1998). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37:185-194.
- [41] Hwang. C. (1980). Laplace's method revisited: Weak convergence of probability measures. *Ann. Probab.*, 8:1177-1182.
- [42] Johnson, D. and Hoeting, J. (2003). Autoregressive models for capture-recapture data: A Bayesian approach. *Biometrics*, 59(2):341-350.
- [43] Laird, N., and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J. American Statist. Assoc.*, 82:97-105.
- [44] Lavielle, M. and Moulines, E. (1997). On a stochastic approximation version of the EM algorithm. *Statist. Compute.*, 7:229-236.
- [45] Liu. J. (1996a). Metropolized independent sampling with comparisions to rejection sampling and imporlance sampling. *Statistics and Computing*, 6:113-119.
- [46] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087-1092.
- [47] Murray, G. (1977). Comments on "Maximum likelihood from incomplete data via the EM algorithm". *J. Royal Statist. Soc. Series B*, 39:27-28.
- [48] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Mathemat. Statist.*, 22:400-407.
- [49] Robert, Christian P. and Casella, George. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- [50] Robert, C. (1991). Generalized inverse normal distributions. *Statist. Prob. Lett.*, 11:37-41.

- [51] Robert, C. (1993). Prior feedback: A Bayesian approach to maximum likelihood estimation. *Comput Statist.*, 8:279-294.
- [52] Rubinstein, R. (1981). *Simulation and the Monte Carlo Method*. John Wiley, New York.
- [53] Scherrer, J. (1997). Monte Carlo estimation of transition probabilities in capture-recapture data. Technical report, Biometrics Unit, Cornell Univ., Ithaca, New York. Masters Thesis.
- [54] Seber, G. (1983). Capture-recapture methods. In Kotz, S. and Jonhson, N., editors, *Encyclopedia of Statistical Science*. John Wiley, New York.
- [55] Wu, C. (1983). On the convergence properties of the EM algorithm. *Ann Statist.*, 11:95-103.
- [56] 李根良 (2010), 「四捨五入型資料參數估計之研究」。國立中山大學應用數學系碩士論文。

附錄-勘誤表

1. page 7. 原式 $y_i = a + bx_i + \epsilon_i$ (1.7)

修正後 $y_i = b + ax_i + \epsilon_i$ (1.7)

2. page 8. 指數族 $f(x) = h(x)e^{\theta x - \varphi(\theta)}$ (1.9)

修正後 $f(x) = h(x)e^{\theta t(x) - \varphi(\theta)}$ (1.9)

3. page 8. 指數族 $x = \nabla \varphi\{\hat{\theta}(x)\}$ (1.10)

修正後 $t(x) = \nabla \varphi\{\hat{\theta}(x)\}$ (1.10)

4. page 8. 例題 1.6 $\varphi(\theta) = \frac{-\theta_1^2}{4\theta_2} + \log(-\theta_2/2)$

修正後 $\varphi(\theta) = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)$

5. page 9. 例題 1.8 $\sqrt{\lambda}I_{(p-1)/2}(\sqrt{\lambda y}) = \sqrt{y}I_{p/2}(\sqrt{\lambda y}), \quad y > p$ (1.13)

修正後 $\sqrt{\lambda}I_{(p-2)/2}(\sqrt{\lambda y}) = \sqrt{y}I_{p/2}(\sqrt{\lambda y}), \quad y > p$ (1.13)

6. page 14. 例題 1.12 $\frac{\int_{RP} ||\theta||^{2-p} e^{-||x-\theta||^2/2} d\theta}{\int_{RP} ||\theta||^{1-p} e^{-||x-\theta||^2/2} d\theta}$ (1.19)

修正後 $\frac{\int_{RP} ||\theta||^{3-p} e^{-||x-\theta||^2/2} d\theta}{\int_{RP} ||\theta||^{1-p} e^{-||x-\theta||^2/2} d\theta}$ (1.19)

7. page 17. 例題 1.14 θ 的事後分佈為 $\mathcal{N}(\frac{n\tau^2\bar{x}}{n\tau^2+\sigma^2}, \frac{n\tau^2\sigma^2}{n\tau^2+\sigma^2})$

修正後 $\mathcal{N}(\frac{n\tau^2\bar{x}}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2})$

8. page 20. 例題 1.17 圖 1.3 標示及縱軸函數值有誤。

9. page 49.

$$P(X \in \mathcal{A}) = P(Y \in \mathcal{A} | U < f(Y)) = \frac{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} = \int_{\mathcal{A}} f(y) dy$$

修正後

$$P(X \in \mathcal{A}) = P(Y \in \mathcal{A} | U < f(Y)) = \frac{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int_{\mathcal{X}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} = \int_{\mathcal{A}} f(y) dy$$

10. page 53. 例題 2.20 $\frac{f}{g_\alpha}(z) = e^{-z^2/2} e^{-\alpha(z-\underline{\mu})}$

修正後 $\frac{f}{g_\alpha}(z) = e^{-z^2/2} e^{\alpha(z-\underline{\mu})}$

11. page 54. 例題 2.22 $(1 - x^2/2) \leq f(x)$

修正後 $\frac{1}{\sqrt{2\pi}}(1 - x^2/2) \leq f(x)$

12. page 55. 例題 2.23 $P(N = n) = \left[\frac{1}{1+\exp\{-(n+0.5-\alpha)/\beta\}} - \frac{1}{1+\exp\{-(n-0.5-\alpha)/\beta\}} \right]$

修正後

$$P(N = n) = \left[\frac{1}{1+\exp\{-(n+0.5-\alpha)/\beta\}} - \frac{1}{1+\exp\{-(n-0.5-\alpha)/\beta\}} \right] \cdot \left(\frac{1+\exp\{(-0.5-\alpha)/\beta\}}{\exp\{(-0.5-\alpha)/\beta\}} \right)$$

13. page 61. 例題 2.26 圖 2.8 標示有誤。

14. page 80. 例題 3.1 $\theta_1 = \lambda \cos \varphi_1, \theta_2 = \lambda \sin \varphi_1 \cos \varphi_2, \dots$

修正後 $\theta_1 = \sqrt{\lambda} \cos \varphi_1, \theta_2 = \sqrt{\lambda} \sin \varphi_1 \cos \varphi_2, \dots$

15. page 86. 概似比近似卡方分佈 $\log \left[\frac{L(\hat{\theta}|x)}{L(\hat{\theta}^0|x)} \right] \quad (3.5)$

修正後 $2 \log \left[\frac{L(\hat{\theta}|x)}{L(\hat{\theta}^0|x)} \right] \quad (3.5)$

16. page 88. 表 3.3 卡方分佈的百分位數有誤。

17. page 92. 例題 3.10 $\hat{R}_1 = \frac{1}{T\lambda^2} \sum_{t=1}^T X_t e^{-X_t/\lambda} \lambda^{-1} e^{(\log X_t)^2/2\sigma^2} \sqrt{2\pi\sigma}(X_t - \lambda)^2$

修正後 $\hat{R}_1 = \frac{1}{T\lambda^2} \sum_{t=1}^T X_t e^{-X_t/\lambda} \lambda^{-1} e^{(\log X_t)^2/2\sigma^2} \sqrt{2\pi\sigma}(X_t - \lambda)^2$

18. page 97. 例題 3.13 $h_1(x) \frac{f^2(x)}{g(x)} \propto \sqrt{x} f^2(x) |1-x|^{1-\alpha-1} \exp |1-x|$

修正後 $h_1^2(x) \frac{f^2(x)}{g(x)} \propto |x| f^2(x) |1-x|^{1-\alpha-1} \exp |1-x|$

19. page 98. 例題 3.13 $g(x) = e^{-x} I_x \geq 0 \quad \text{修正後} \quad g(x) = e^{-x} I_{\{x \geq 0\}}$

20. page 104. $\delta_1 = f(y_t) \quad \text{修正後} \quad \delta_1 = h(y_t)$

21. page 106. 例題 3.15 $M' = \exp\{a(\log(a) - 1) - \alpha(\log(\alpha) - 1)\}$

修正後 $M' = \exp\{\alpha(\log(\alpha) - 1) - a(\log(a) - 1)\}$

22. page 172. 例題 5.13 $X_{i,j} \sim \mathcal{N}(\theta_{i,j}, 1)$ 及 $\theta_{i-1,j} \vee \theta_{i,j-1} \leq \theta_{i,j} \leq \theta_{i+1,j} \wedge \theta_{i,j+1}$
 修正後 $X_{i,j} \sim \mathcal{N}(\theta_{i,j}, \frac{1}{n_{i,j}})$ 及 $\theta_{i+1,j} \vee \theta_{i,j+1} \leq \theta_{i,j} \leq \theta_{i-1,j} \wedge \theta_{i,j-1}$

23. page 175. 例題 5.14 $L(\theta|\mathbf{y}) = E[L^c(\theta|\mathbf{y}, \mathbf{z})] = \int L^c(\theta|\mathbf{y}, \mathbf{z}) f(\mathbf{z}|\mathbf{y}, \theta) d\mathbf{z}$
 修正後 $L(\theta|\mathbf{y}) = E[L^c(\theta|\mathbf{y}, \mathbf{z})] = \int L^c(\theta|\mathbf{y}, \mathbf{z}) d\mathbf{z}$

24. page 178.

例題 5.17

$$L^c(\theta|\mathbf{y}, \mathbf{z}) \propto \prod_{i=1}^m \exp\{-(y_i - \theta)^2/2\} \prod_{i=m+1}^n \exp\{(z_i - \theta)^2/2\}$$

$$\text{修正後 } L^c(\theta|\mathbf{y}, \mathbf{z}) \propto \prod_{i=1}^m \exp\{-(y_i - \theta)^2/2\} \prod_{i=m+1}^n \exp\{-(z_i - \theta)^2/2\}$$

25. page 178. 例題 5.17 刪失資料 $\mathbf{z} = (z_{n-m+1}, \dots, z_n)$

$$\text{修正後 } \mathbf{z} = (z_{m+1}, \dots, z_n)$$

26. page 178.

例題 5.17

$$k(\mathbf{z}|\theta, by) = \frac{1}{(2\pi)^{(n-m)/2}} \exp\left\{\sum_{i=m+1}^n (z_i - \theta)^2/2\right\} \quad (5.16)$$

修正後

$$k(\mathbf{z}|\theta, y) = \left(\frac{1}{1 - \Phi(a - \theta)}\right)^{(n-m)} \times \frac{1}{(2\pi)^{(n-m)/2}} \exp\left\{-\sum_{i=m+1}^n (z_i - \theta)^2/2\right\}$$

27. page 179. 例題 5.17 $\hat{\theta}^{(j+1)} = \frac{m}{n}\bar{y} + \frac{1}{n} \left[\hat{\theta}^{(j)} + \frac{\phi(a - \hat{\theta}^{(j)})}{1 - \Phi(a - \hat{\theta}^{(j)})} \right]$ (5.17)

$$\text{修正後 } \hat{\theta}^{(j+1)} = \frac{m}{n}\bar{y} + \frac{n-m}{n} \left[\hat{\theta}^{(j)} + \frac{\phi(a - \hat{\theta}^{(j)})}{1 - \Phi(a - \hat{\theta}^{(j)})} \right] \quad (5.17)$$

附錄-表

表 1-1 太空梭起飛時的溫度($^{\circ}F$)及當時O-ring零件的狀態(1表示失敗；0表示成功)

flight	14	9	23	10	1	5	13	15	4	3	8
failure	1	1	1	1	0	0	0	0	0	0	0
Temp.	53	57	58	63	66	67	67	67	68	69	70
flight	17	2	11	6	7	16	21	19	22	12	20
failure	0	1	1	0	0	0	1	0	0	0	0
Temp.	70	70	70	72	73	75	75	76	76	78	79
											81

表 2-1

Year	75	76	77	78	79
Deaths	3	5	7	9	10

Year	80	81	82	83	84
Deaths	18	6	14	11	9

Year	85	86	87	88	89
Deaths	5	11	15	6	11

Year	90	91	92	93	94
Deaths	17	12	15	8	4

表 3-1

n/Z_{α}	0.	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.54	0.775	0.815	0.885	0.96	1.	0.98	1.	1.
10^3	0.494	0.743	0.8185	0.901	0.949	0.991	0.9935	0.997	1.
10^4	0.5024	0.74355	0.79995	0.8986	0.9505	0.99055	0.99475	0.99895	0.99985
10^5	0.49905	0.746445	0.80064	0.899225	0.950325	0.98961	0.99543	0.99902	0.99986
10^6	0.49967	0.748358	0.799234	0.899851	0.950484	0.989853	0.995096	0.999002	0.99989
10^7	0.500055	0.748578	0.799559	0.899708	0.950486	0.989831	0.99507	0.999008	0.999901

表 3-2

	惡性腫瘤		
	受控制	不受控制	
外科手術	21	2	23
輻射	15	3	18
	36	5	41

表 3-3

α	截點(蒙地卡羅法)	χ_1^2
0.10	2.84	2.705
0.05	3.93	3.841
0.01	6.72	6.635

表 3-4

分佈	h_1	h_2	h_3	h_4	h_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

表 3-5

區間	近似值	實際值
(7, 9)	0.193351	0.193341
(6, 10)	0.375046	0.37477
(2, 14)	0.848559	0.823349
(15.981, ∞)	0.0224544	0.100005

表 4-1

α_j	β_j	θ_T	$h(\theta_T)$	$\min_t h(\theta_t)$	IterationT
$\frac{1}{10j}$	$\frac{1}{10j}$	(-0.815, 0.326)	1.321	0.1796	125
$\frac{1}{100j}$	$\frac{1}{100j}$	(0.898, 0.749)	0.000139	0.000139	121
$\frac{1}{10 \log(1+j)}$	$\frac{1}{j}$	(0.00002, -0.125)	6.19×10^{-9}	6.19×10^{-9}	253

表 4-2

情況	T_i	θ_T	$h(\theta_T)$	$\min_t h(\theta_t)$	接受率
1	$\frac{1}{10i}$	(0.271, 0.136)	5.10×10^{-5}	2.68745×10^{-7}	0.0124
2	$\frac{1}{\log(1+i)}$	(-0.176, 0.088)	0.0204164	3.95×10^{-8}	0.6086
3	$\frac{100}{\log(1+i)}$	(0.128, 0.489)	0.270502	1.86×10^{-7}	0.8112
4	$\frac{1}{10 \log(1+i)}$	(-0.192, 0.056)	0.0311471	3.33×10^{-7}	0.7532

表 4-3

λ	5	10	100	1000	5000	10^4
δ_λ^π	2.02	2.04	1.89	1.98	1.94	2.00

表 4-4

HSR decile	ACT score								
	1-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
91 ≤ HSR < 99	1.57 (4)	2.11 (5)	2.73 (18)	2.96 (39)	2.97 (126)	3.13 (219)	3.41 (232)	3.45 (47)	3.51 (4)
81 ≤ HSR < 90	1.80 (6)	1.94 (15)	2.52 (30)	2.68 (65)	2.69 (117)	2.82 (143)	2.75 (70)	2.74 (8)	(0)
71 ≤ HSR < 80	1.88 (10)	2.32 (13)	2.32 (51)	2.53 (83)	2.58 (115)	2.55 (107)	2.72 (24)	2.76 (4)	(0)
61 ≤ HSR < 70	2.11 (6)	2.23 (32)	2.29 (59)	2.29 (84)	2.50 (75)	2.42 (44)	2.41 (19)	(0)	(0)
51 ≤ HSR < 60	1.60 (11)	2.06 (16)	2.12 (49)	2.11 (63)	2.31 (57)	2.10 (40)	1.58 (4)	2.13 (1)	(0)
41 ≤ HSR < 50	1.75 (6)	1.98 (12)	2.05 (31)	2.16 (42)	2.35 (34)	2.48 (21)	1.36 (4)	(0)	(0)
31 ≤ HSR < 40	1.92 (7)	1.84 (6)	2.15 (5)	1.95 (27)	2.02 (13)	2.10 (13)	1.49 (2)	(0)	(0)
21 ≤ HSR < 30	1.62 (1)	2.26 (2)	1.91 (5)	1.86 (14)	1.88 (11)	3.78 (1)	1.40 (2)	(0)	(0)
HSR ≤ 20	1.38 (1)	1.57 (2)	2.49 (5)	2.01 (7)	2.07 (7)	(0)	0.75 (1)	(0)	(0)

表 4-5

HSR decile	ACT score								
	1-12	13-15	16-18	19-21	22-24	25-27	28-29	31-32	34-36
91-99	1.87	2.18	2.73	2.96	2.97	3.13	3.41	3.45	3.51
81-89	1.87	2.17	2.52	2.68	2.69	2.79	2.79	2.80	
71-99	1.86	2.17	2.32	2.53	2.56	2.57	2.72	2.76	
61-69	1.86	2.17	2.29	2.29	2.46	2.46	2.47		
51-59	1.74	2.06	2.12	2.13	2.24	2.24	2.24	2.27	
41-49	1.74	1.98	2.05	2.13	2.24	2.24	2.24		
31-39	1.74	1.94	1.99	1.99	2.02	2.06	2.06		
21-29	1.62	1.93	1.97	1.97	1.98	2.05	2.06		
00-20	1.38	1.57	1.97	1.97	1.97		1.97		

附錄-圖

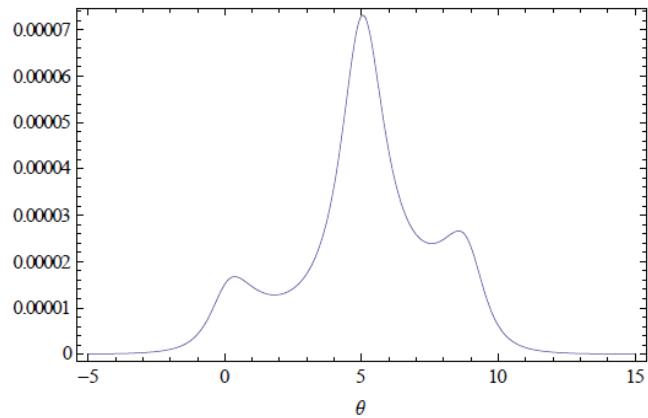


圖 1-1 $\mathcal{C}auchy(\theta, 1)$ 的概似函數圖。

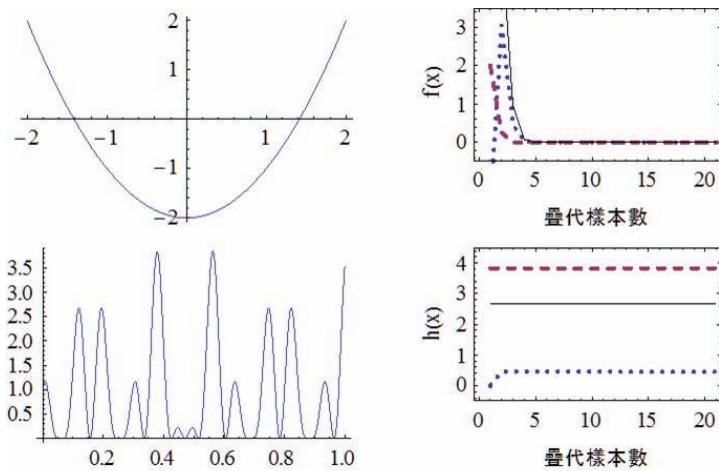


圖 1-2 牛頓-拉福生法

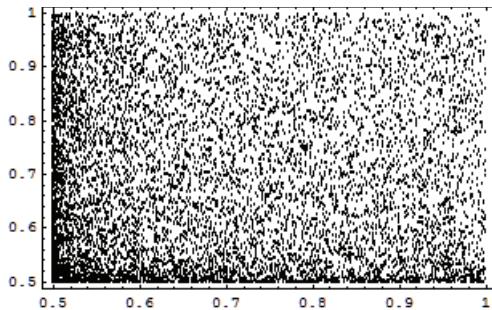


圖 2-1-1 數對 (Y_n, Y_{n+100}) , $n = 1, 2, \dots, 9899$ 所成的散佈圖。

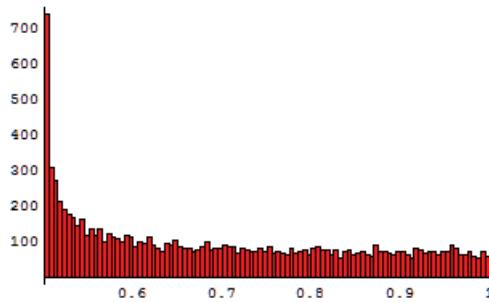


圖 2-1-2 數列 Y_n , $n = 1, 2, \dots, 9899$ 所成的直方圖。

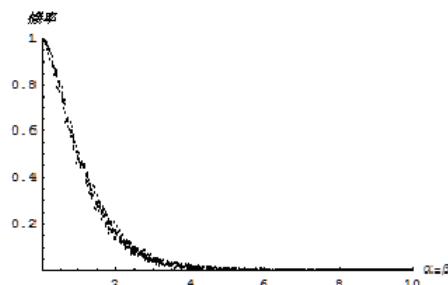


圖 2-2 $Jöhnk$ 演算法中，當 $\alpha = \beta$ ，時接受 (U, V) 的機率。

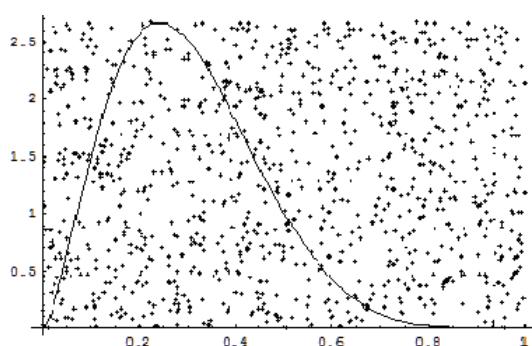


圖 2-3 貝它分佈隨機變數的生成。圓點為使用定理 2.15 模擬所得 1000 筆 (Y, U) ，曲線下的圓點為被接受的 (Y, U) ，約有 368 個。

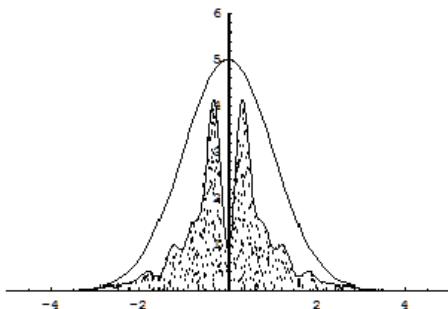


圖 2-4 此圖為來自集合 $\{(x, u) : 0 < u < f(x)\}$ 的一致分佈樣本。其中目標函數為 $f(x) \propto \exp(-x^2/2)(\sin^2 6x + 3\cos^2 x \sin^2 4x + 1)$ ，上界函數 $m(x) = 5\exp(-x^2/2)$ 。

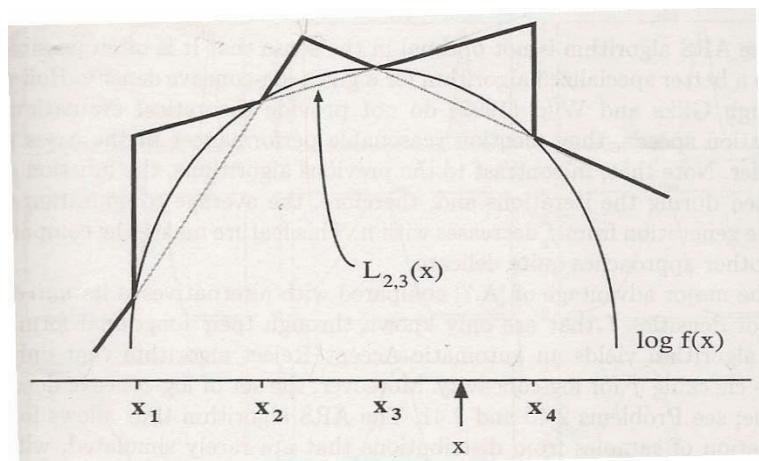


圖 2-5 內容顯示 $h(x) = \log f(x)$ 的上下界線，其中 f 為對數凹密度函數。
(來源：*Gilks et al. 1995*)

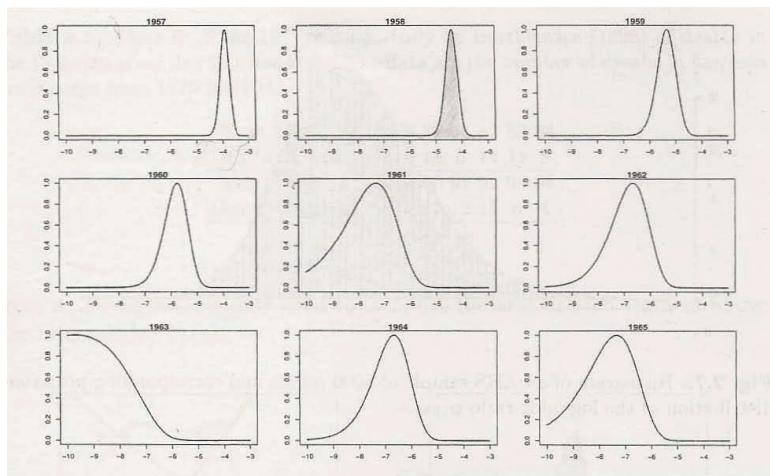


圖 2-6 由 1957 ~ 1965 年間的北方針尾鴨資料之 α_i 的事後分佈。

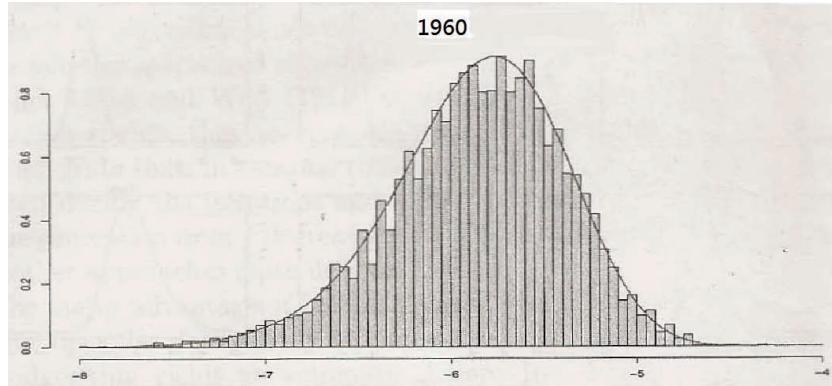


圖 2-7 由 ARS 演算法生成的 5000 筆樣本之直方圖及對應的 α_{1960} 的事後分佈。

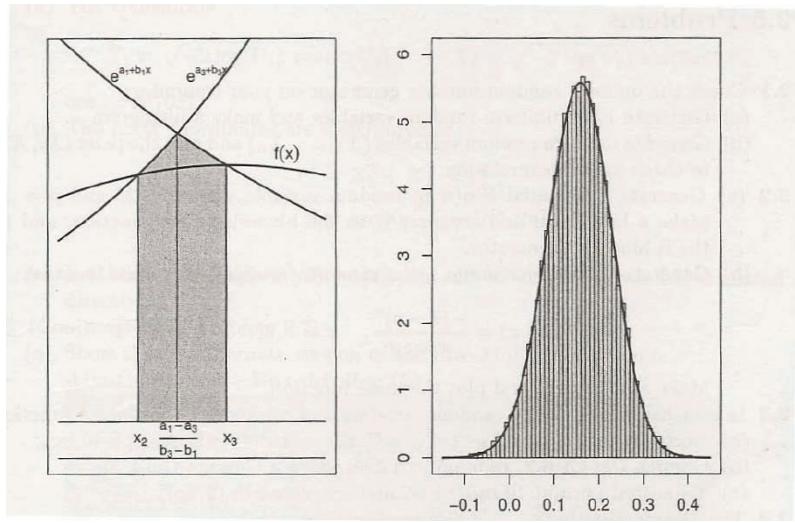


圖 2-8 左圖為區域 $[x_2, x_3]$ 所對應的機率、右圖為ARS演算法所得樣本。

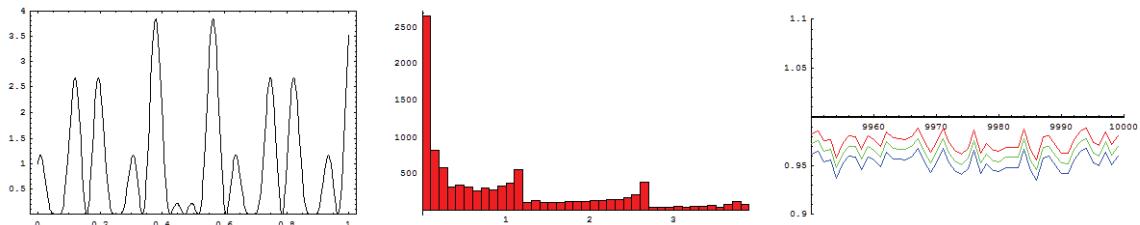


圖 3-1 蒙地卡羅積分法近似結果。

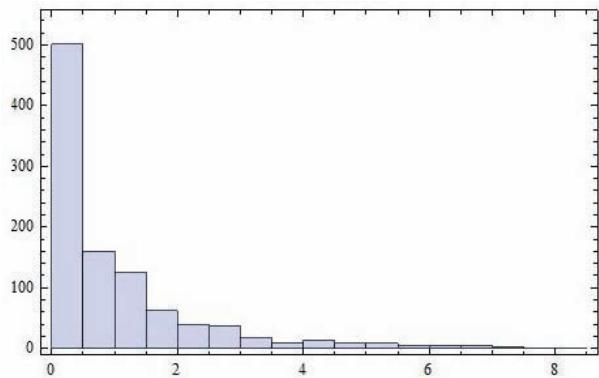


圖 3-2-1 虛無假設下之分佈的直方圖及近似卡方 χ^2_1 分佈的密度函數圖形。

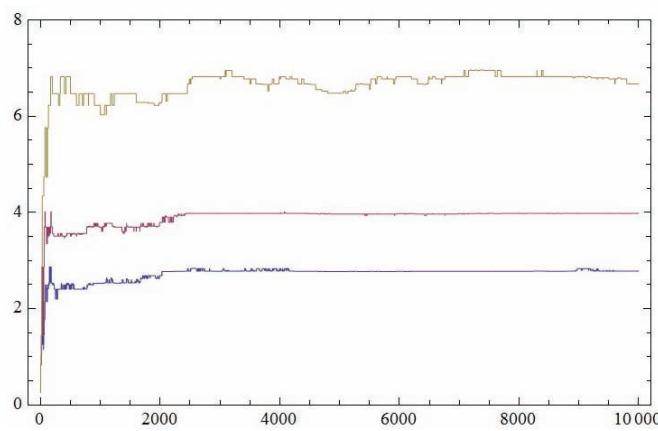


圖 3-2-2 由下而上分別為 0.90、0.95、0.99 移動經驗百分位數。
模擬次數 10000 次。

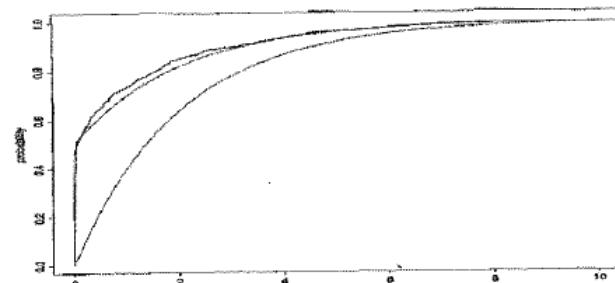


圖 3-3 由上而下分別為常態混合模型、 χ^2 與 Dirac mass 機率各半的混合模型、 χ^2 等，以上三種分佈的經驗累積分佈函數。

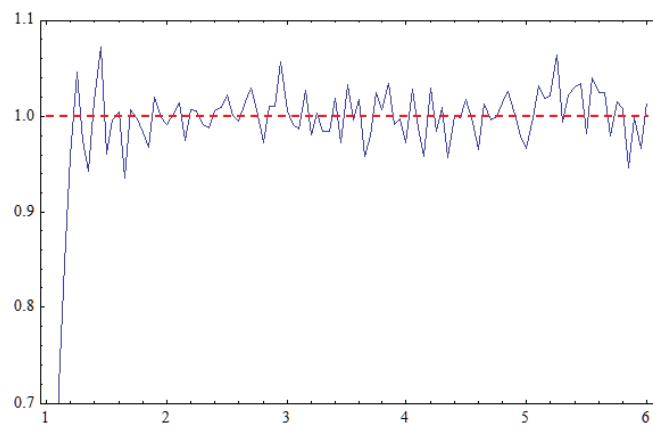


圖 3-4-1 重要抽樣法估計風險 R_1

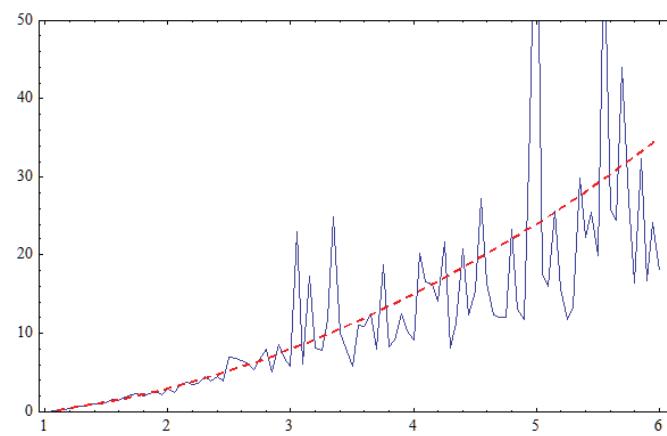


圖 3-4-2 重要抽樣法估計風險 R_2

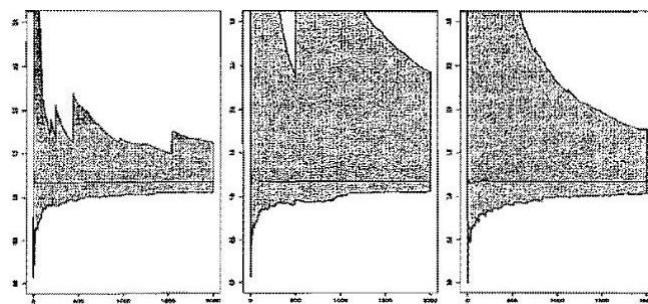


圖 3-5 重要抽樣法估計 $E_f[h_1(X)]$ 結果。

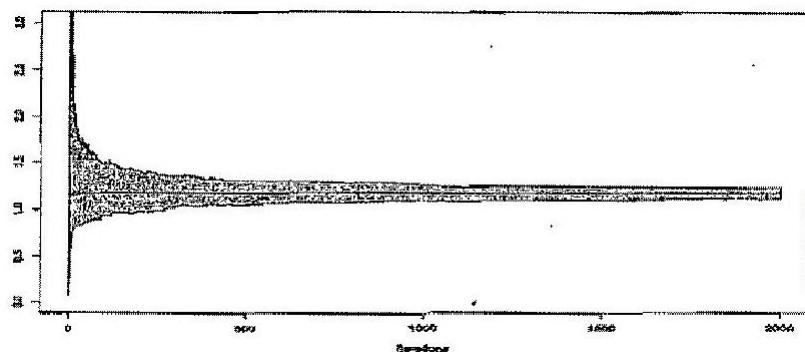


圖 3-6 雙重伽瑪分佈估計 $E_f[h_1(X)]$ 結果。

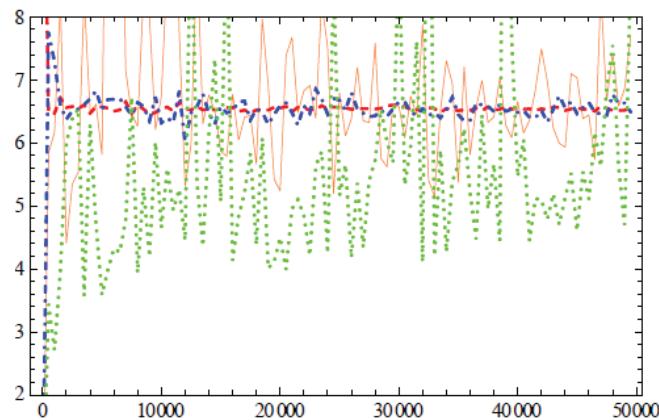


圖 3-7 重要抽樣法估計 $E_f[h_2(X)]$ 結果。

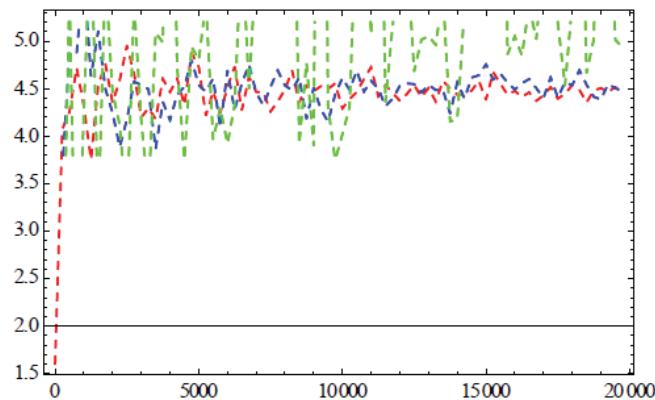


圖 3-8 重要抽樣法估計 $E_f[h_3(X)]$ 結果。

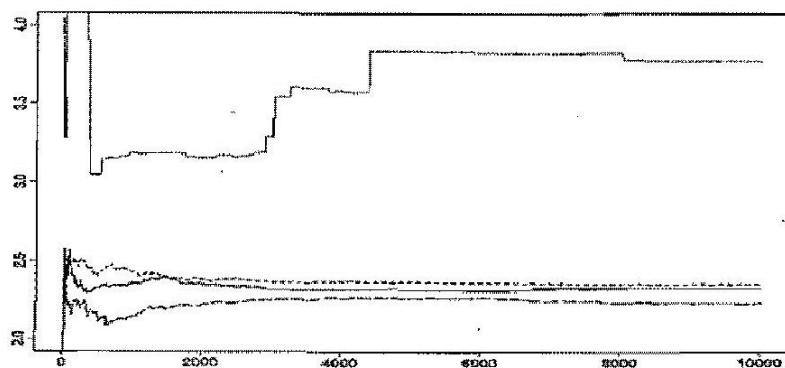


圖 3-9 π : 實線, 收斂值 2.373; π_1 : 點線, 收斂值 3.184;
 π_2 : 長虛線, 收斂值 2.319; π_3 : 短虛線, 收斂值 2.379。

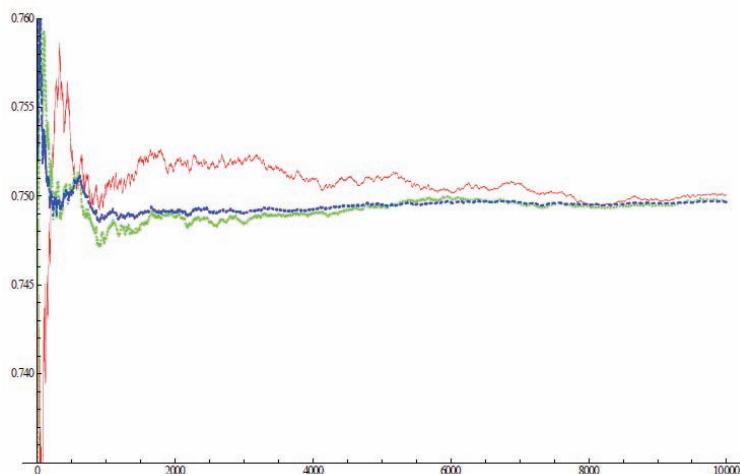


圖 3-10 估計 $E[h_3(X)] = E[x/(1+x)]$ 的收斂結果。

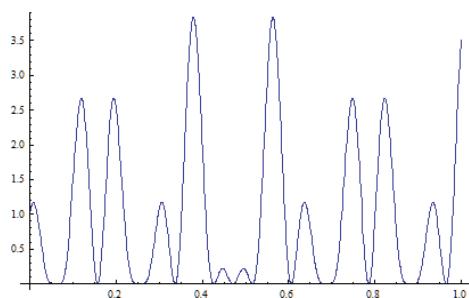


圖 4-1-1 $h(x)$ 的函數圖。

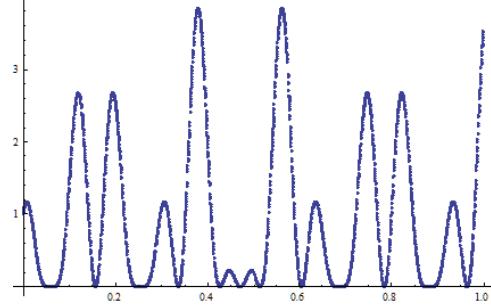


圖 4-1-2 利用 5000 筆 $\mathcal{U}[0, 1]$ 樣本計算 $h(x)$ 的結果。

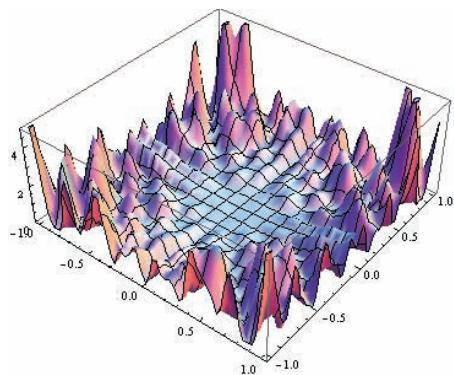


圖 4-2 $h(x, y)$ 的函數圖

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x)x) \\ + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y)$$

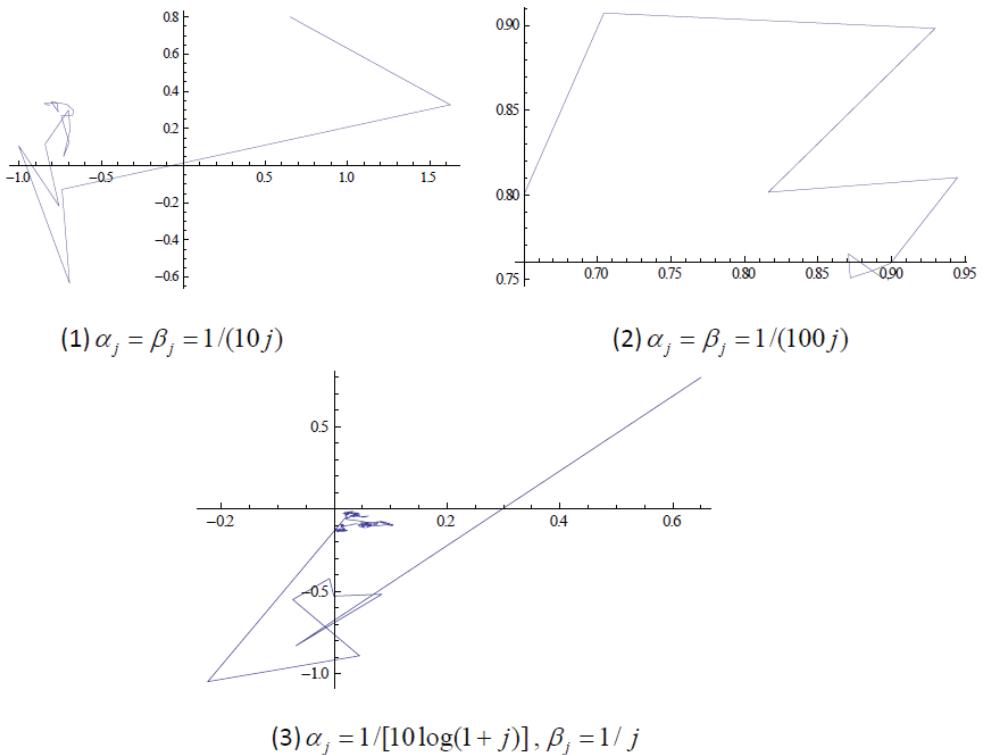


圖 4-3 圖中的路徑分別為給定不同的 α_j 及 β_j ，
以 $(0.65, 0.8)$ 為起始點，使用梯度法所得到的 θ_j 的收斂路徑。

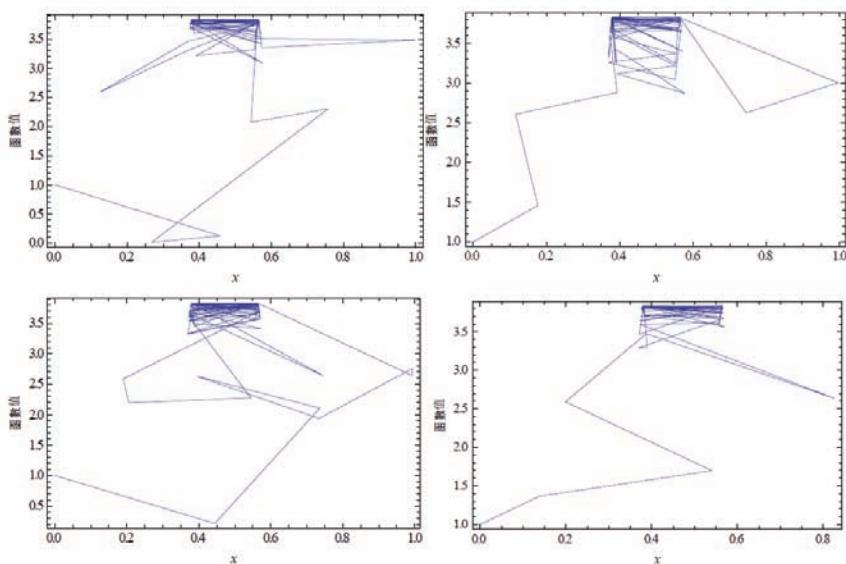


圖 4-4 四張圖均為使用模擬退火演算法所得的 2500 組序對 $(x^{(t)}, h(x^{(t)}))$ 形成的軌跡。

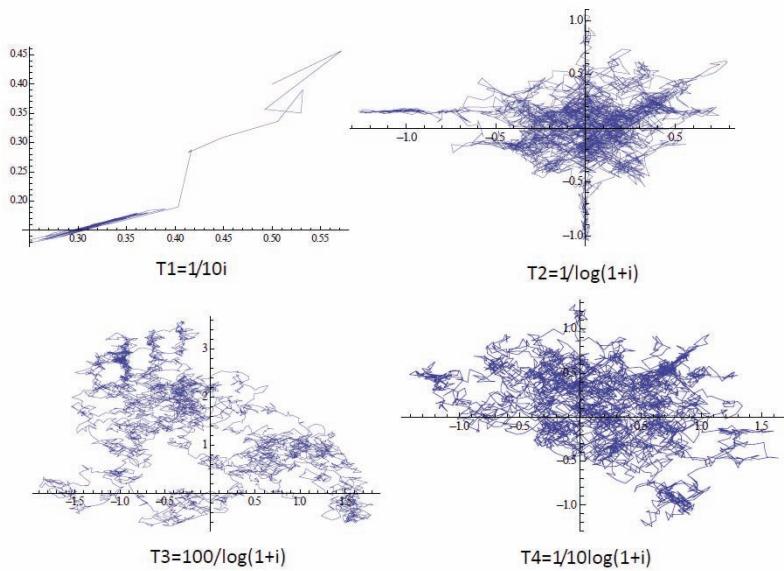


圖 4-5 對四種溫度 T_i 分別使用模擬退火演算法，得到 5000 組序對 (xt, yt) 所形成的軌跡。
起點均為 $(0.5, 0.4)$ ，目標為尋找 $h(x, y)$ 的最小值發生處。

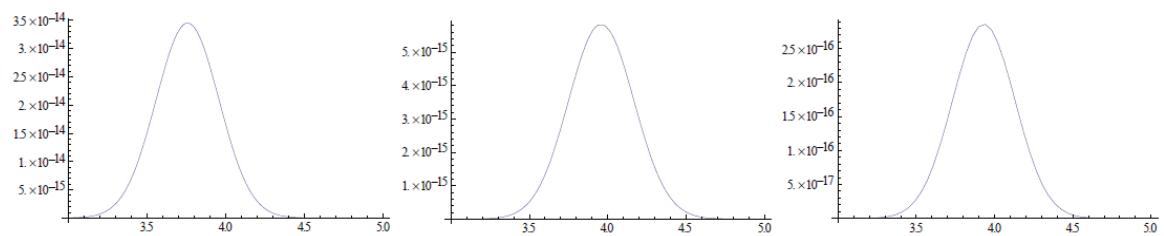


圖 4-6-1 三種概似函數的圖形。

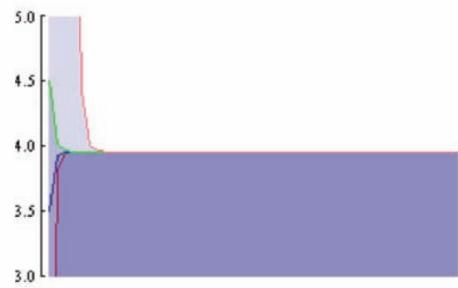


圖 4-6-2 EM 估計結果。

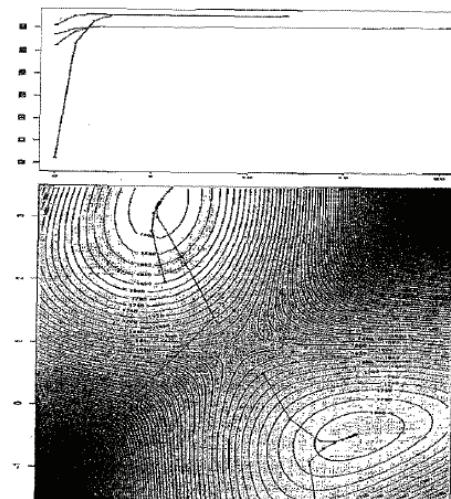


圖 4-7 上半部分表示此混合模型的對數概似函數值，
下半部分表示此模型的對數概似函數之表面圖形。

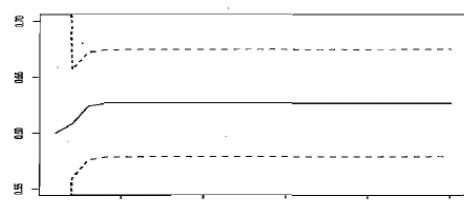


圖 4-8 遺傳連鎖資料之參數的 EM 估計結果。

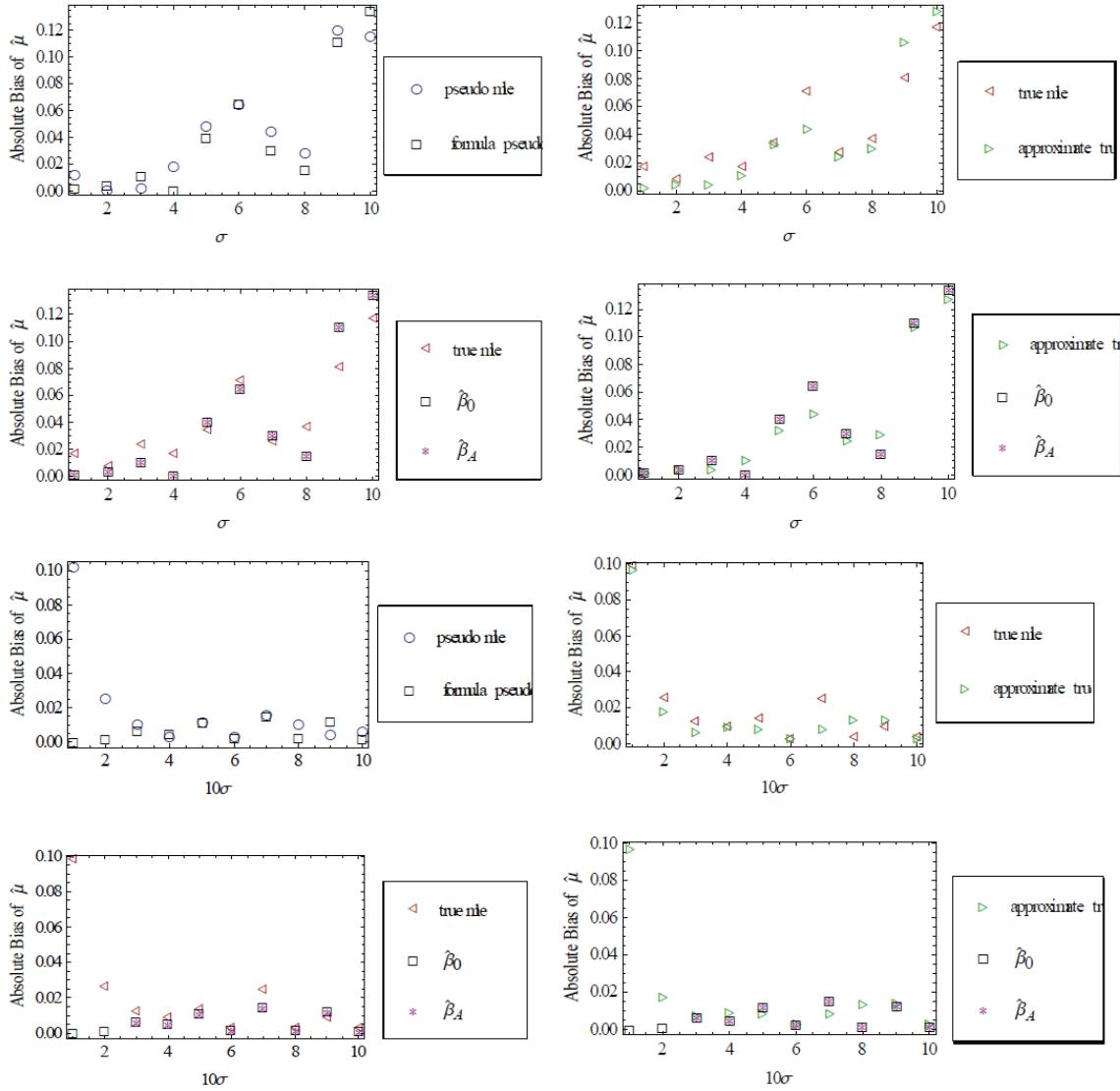


圖 5-1-1 常態分佈的模擬結果： $\hat{\mu}$ 的 bias

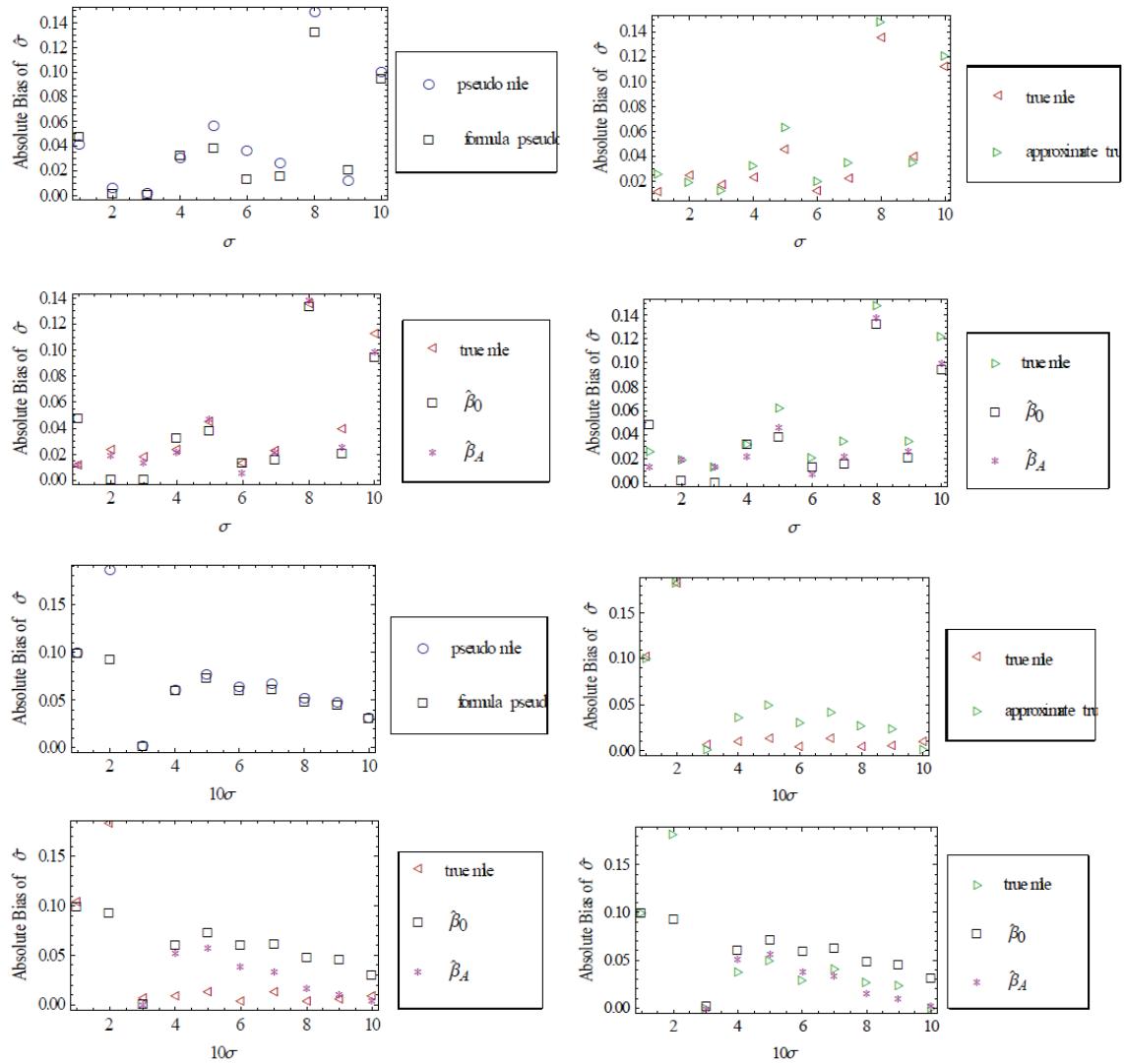


圖 5-1-2 常態分佈的模擬結果： $\hat{\sigma}$ 的 bias

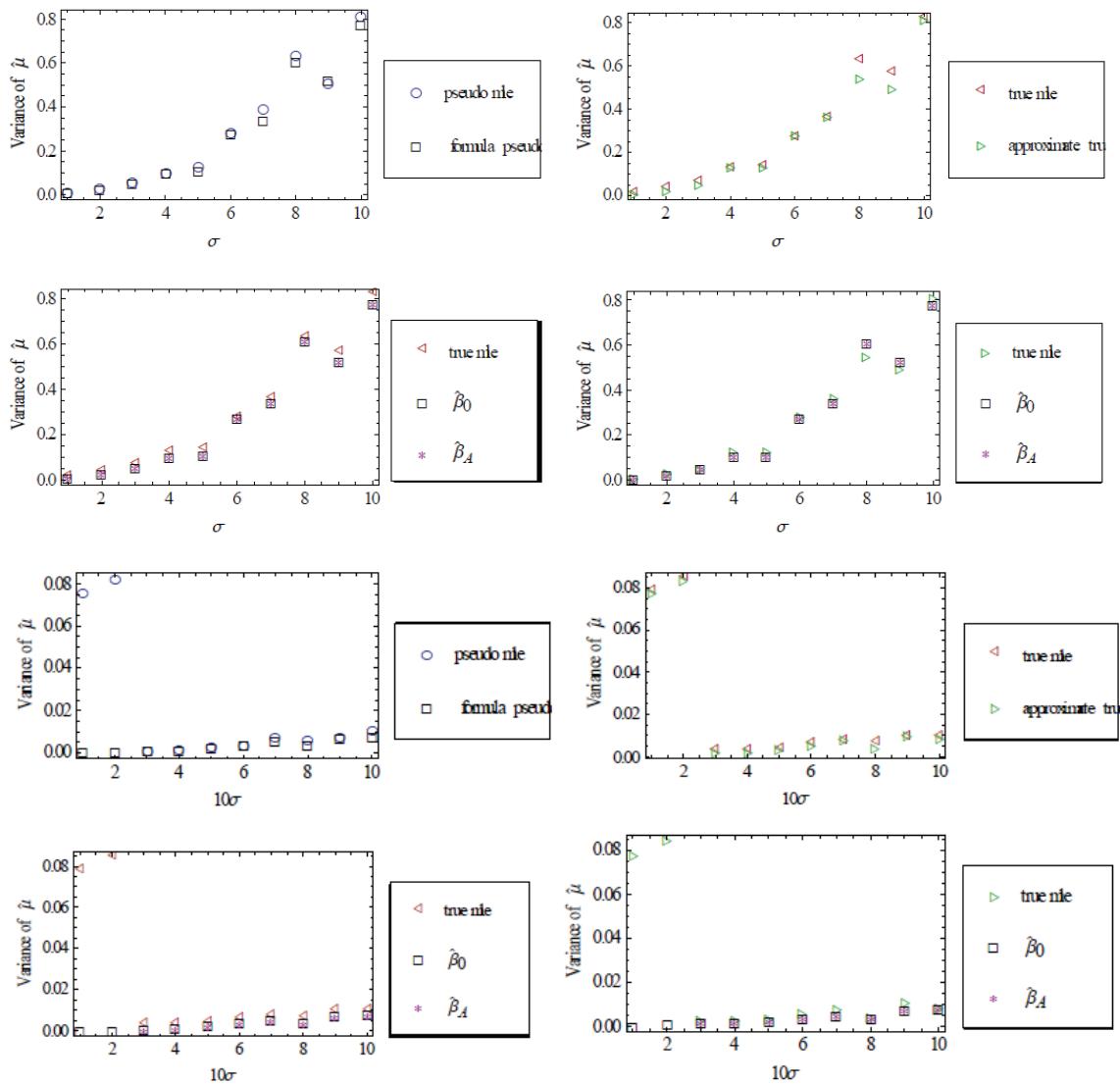


圖 5-1-3 常態分佈的模擬結果： $\hat{\mu}$ 的 variance

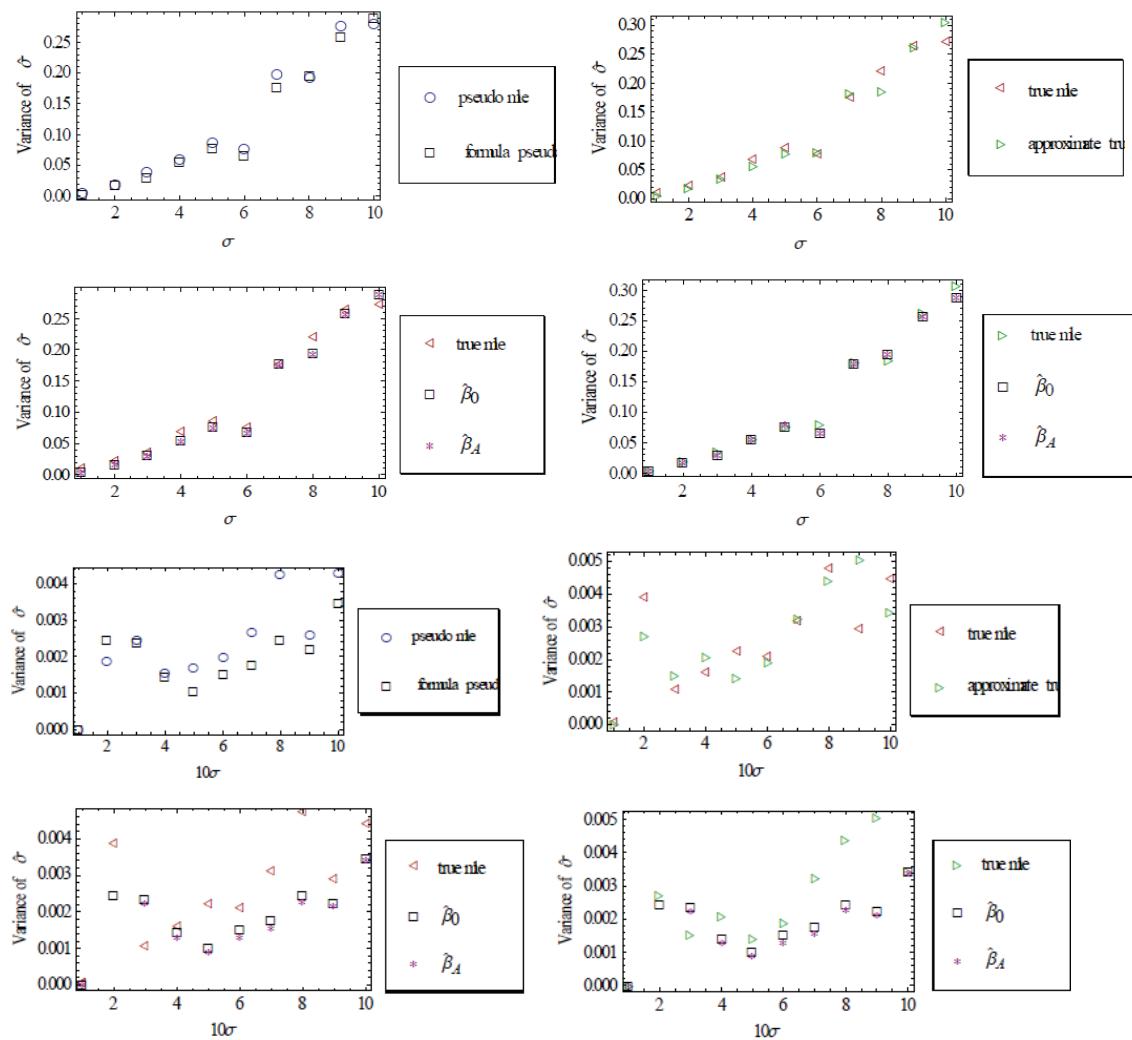


圖 5-1-4 常態分佈的模擬結果： $\hat{\sigma}$ 的 variance

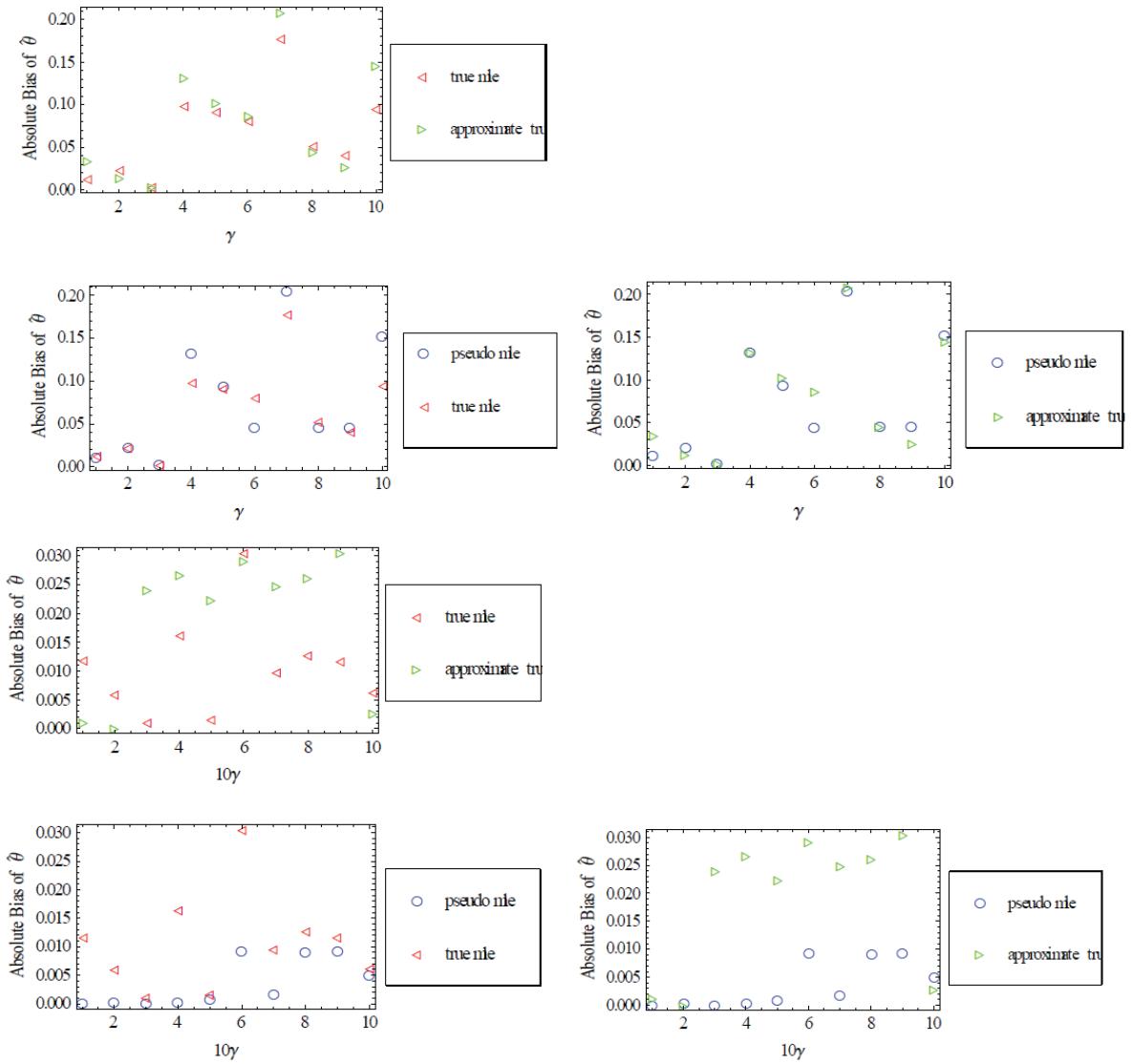


圖 5-2-1 柯西分佈的模擬結果： $\hat{\theta}$ 的 bias

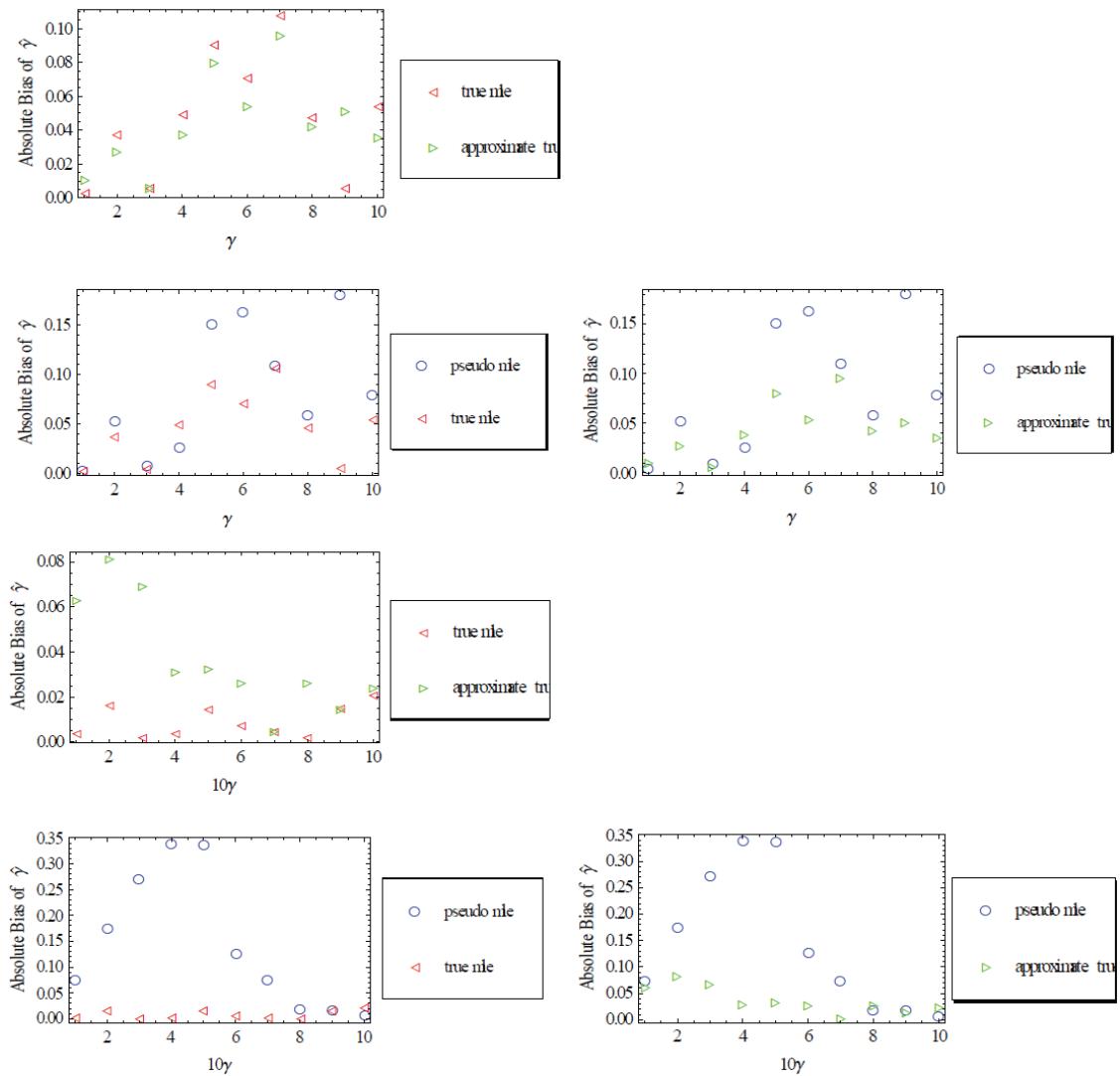


圖 5-2-2 柯西分佈的模擬結果： $\hat{\gamma}$ 的 bias

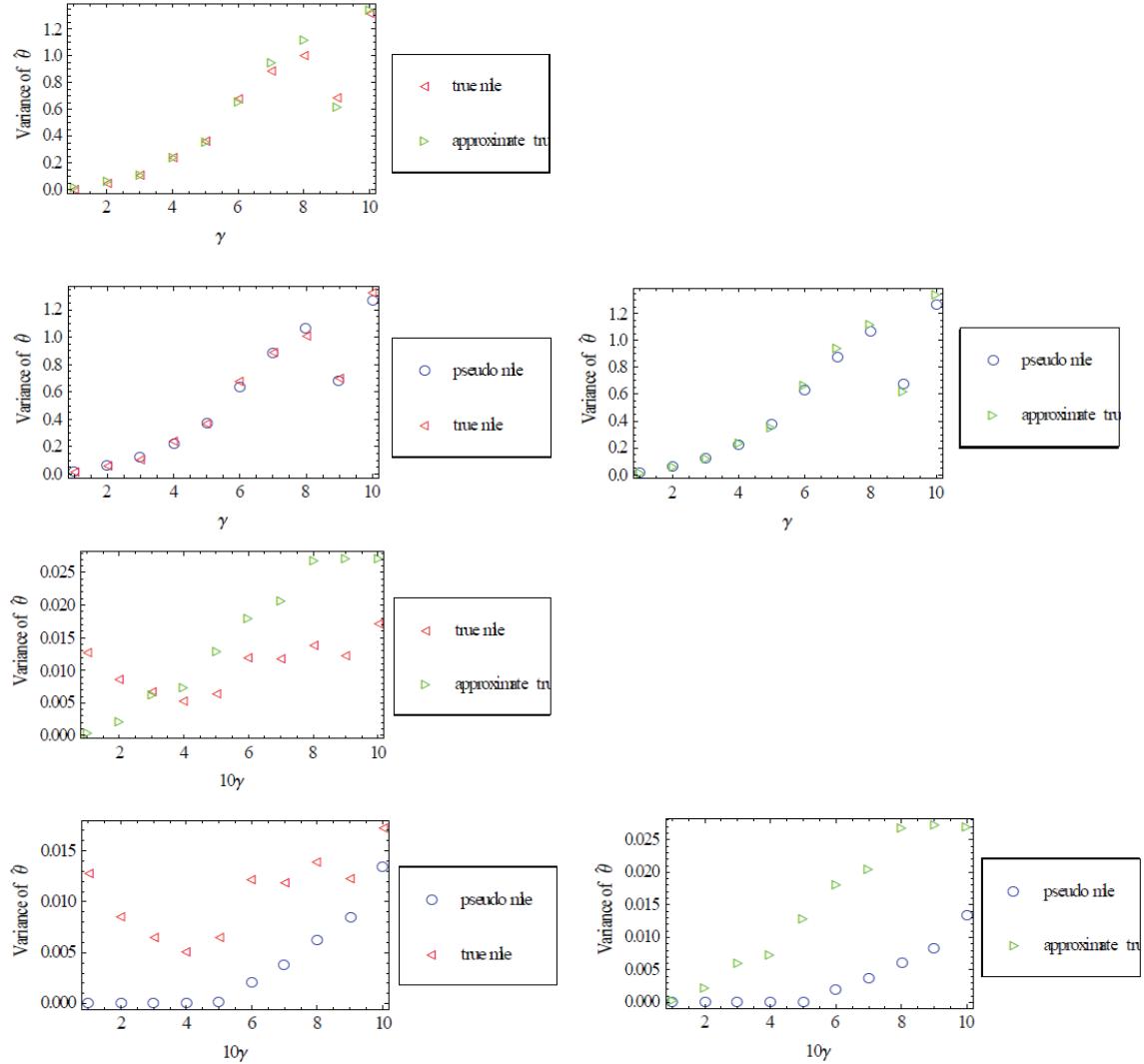


圖 5-2-3 柯西分佈的模擬結果： $\hat{\theta}$ 的 variance

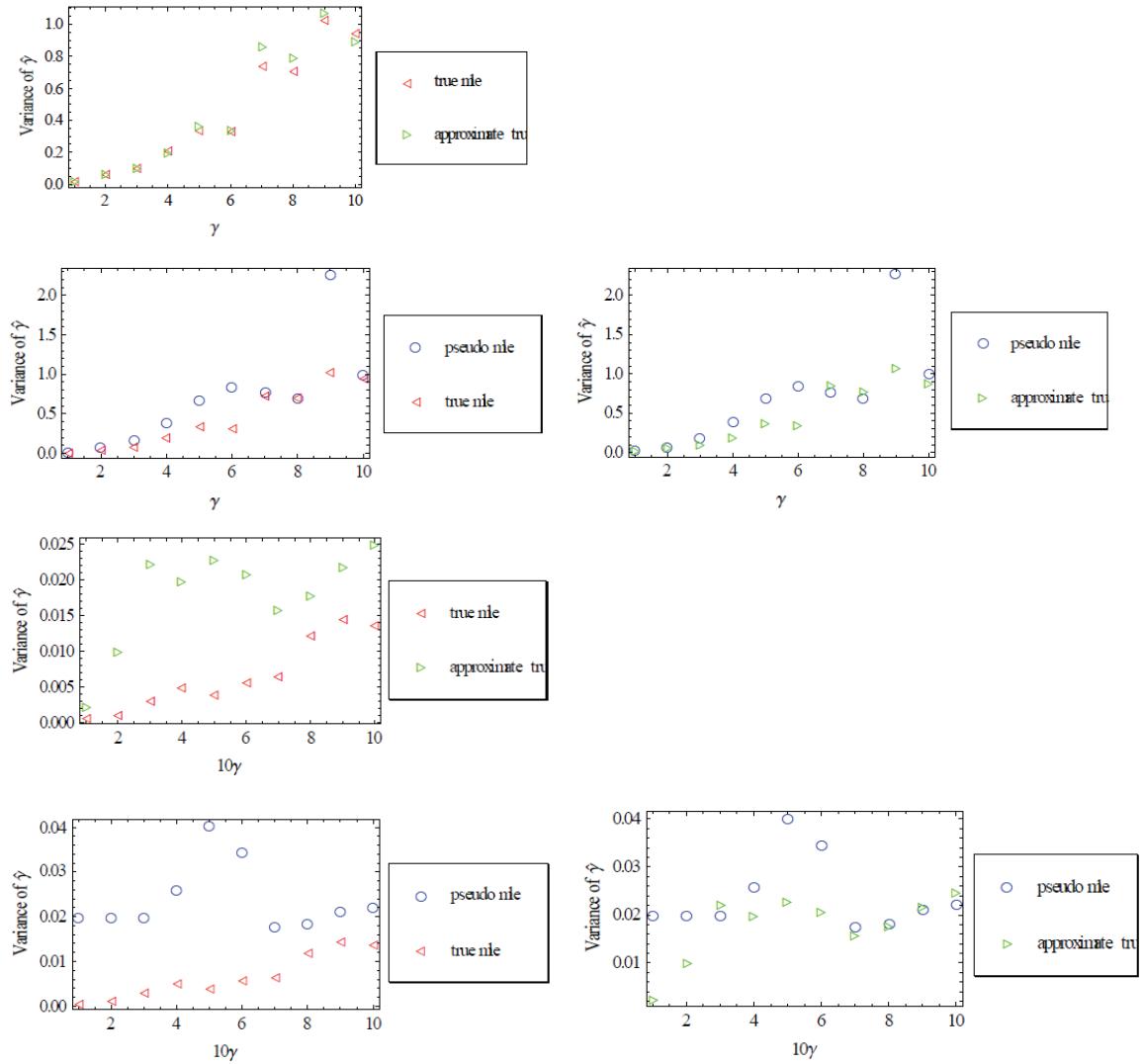


圖 5-2-4 柯西分佈的模擬結果： $\hat{\gamma}$ 的 variance

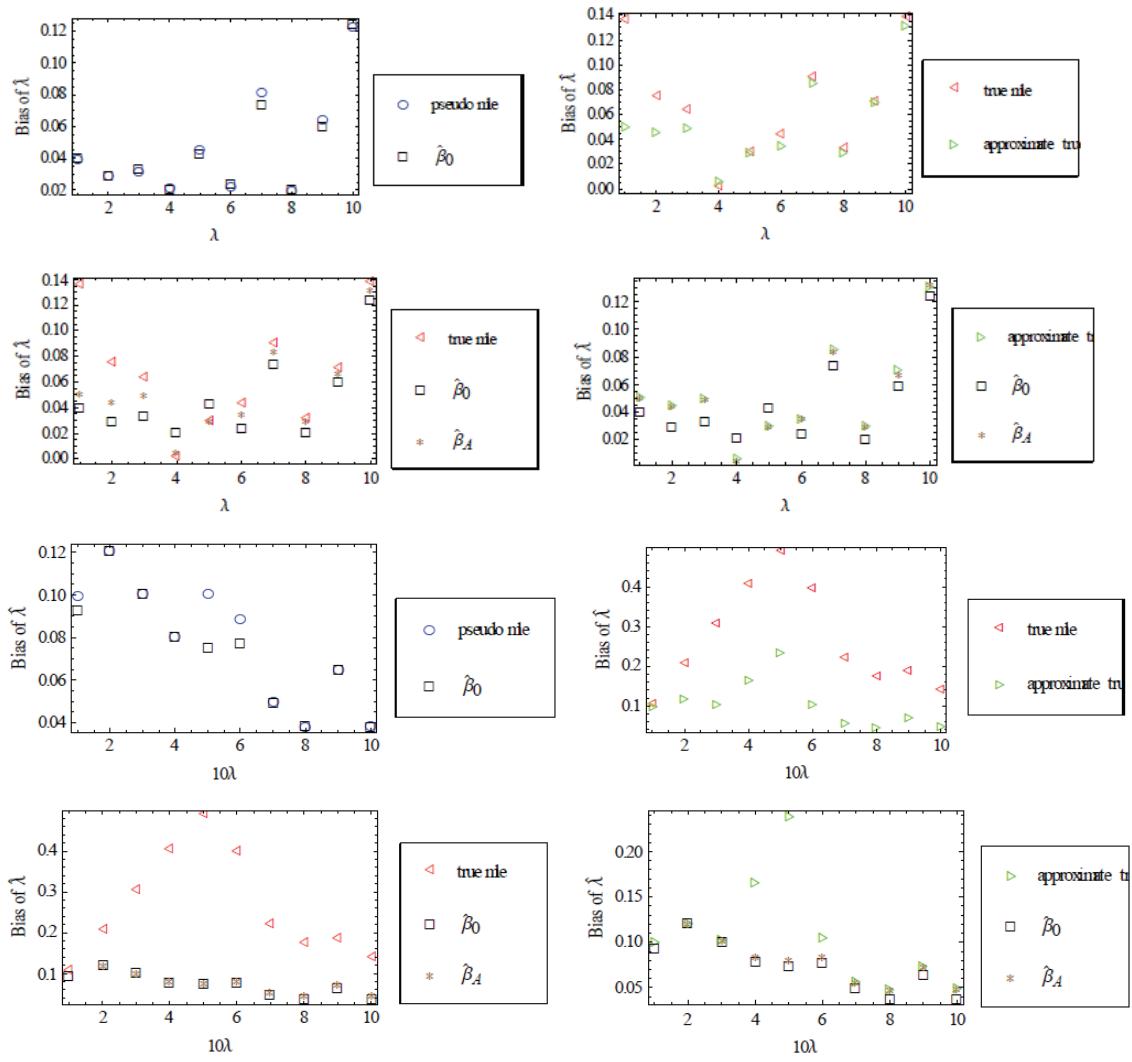


圖 5-3-1 指數分佈的模擬結果： $\hat{\lambda}$ 的 bias

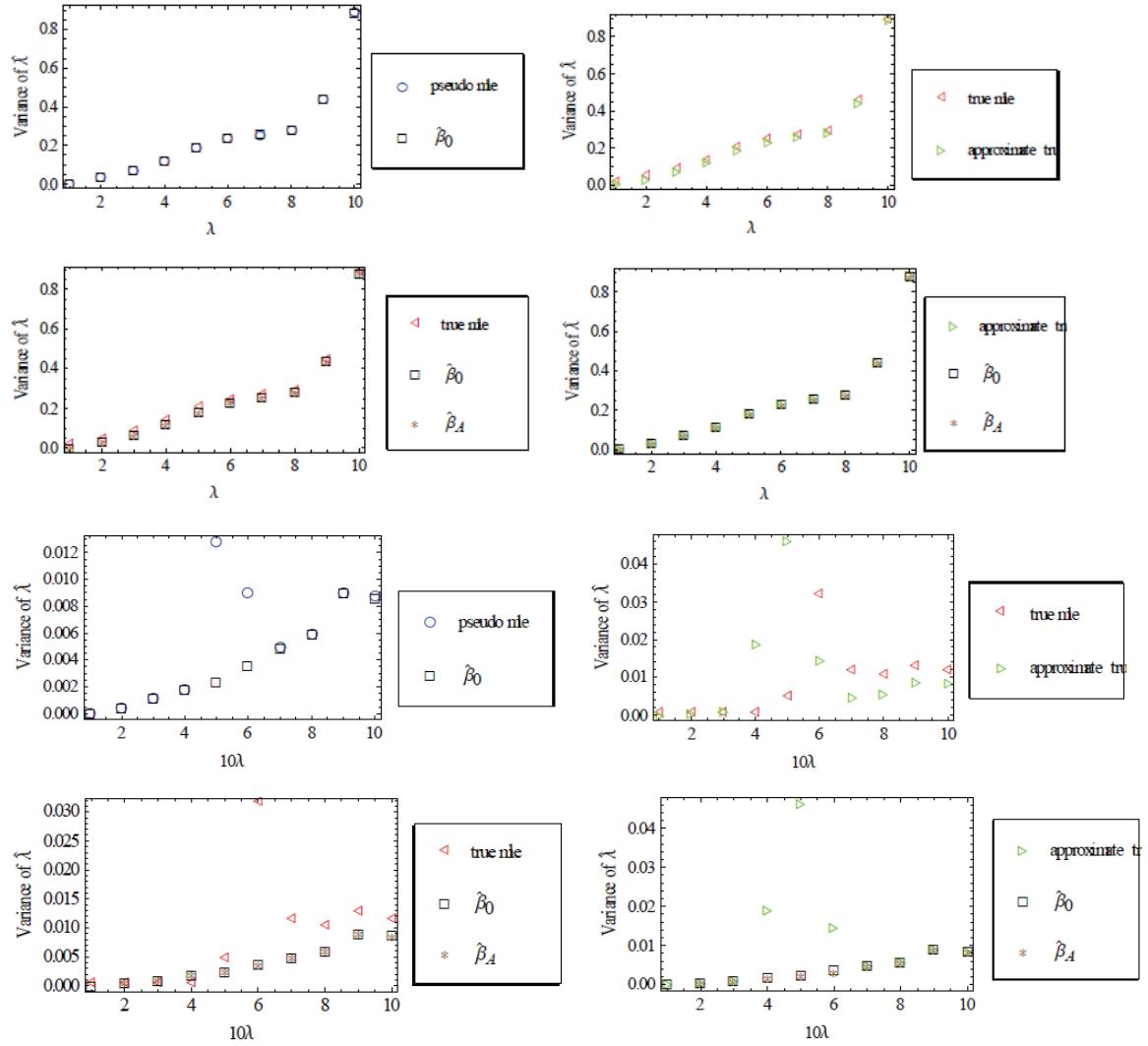


圖 5-3-2 指數分佈的模擬結果： $\hat{\lambda}$ 的 variance

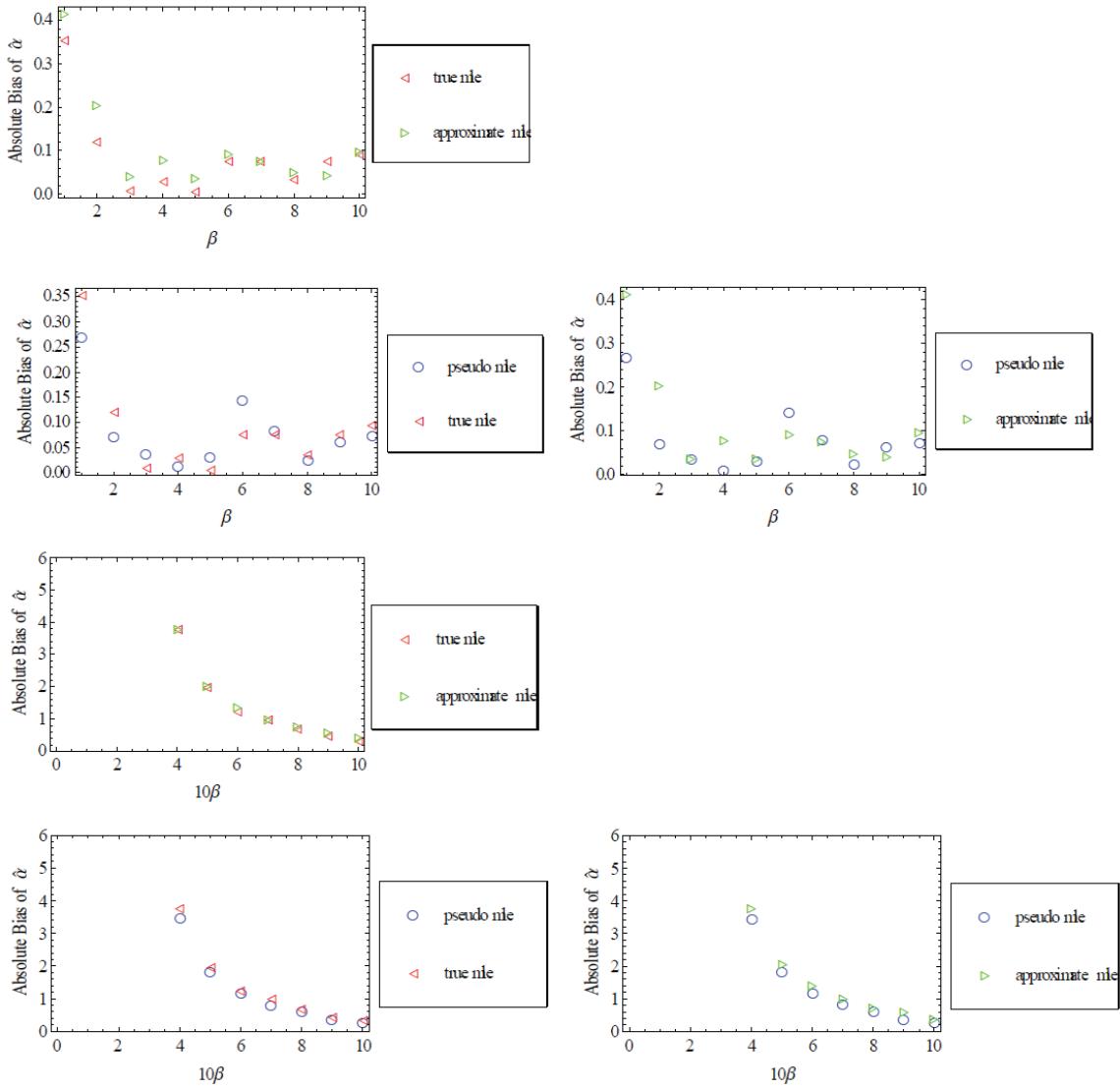


圖 5-4-1 Gamma 分佈的模擬結果： $\hat{\alpha}$ 的 bias

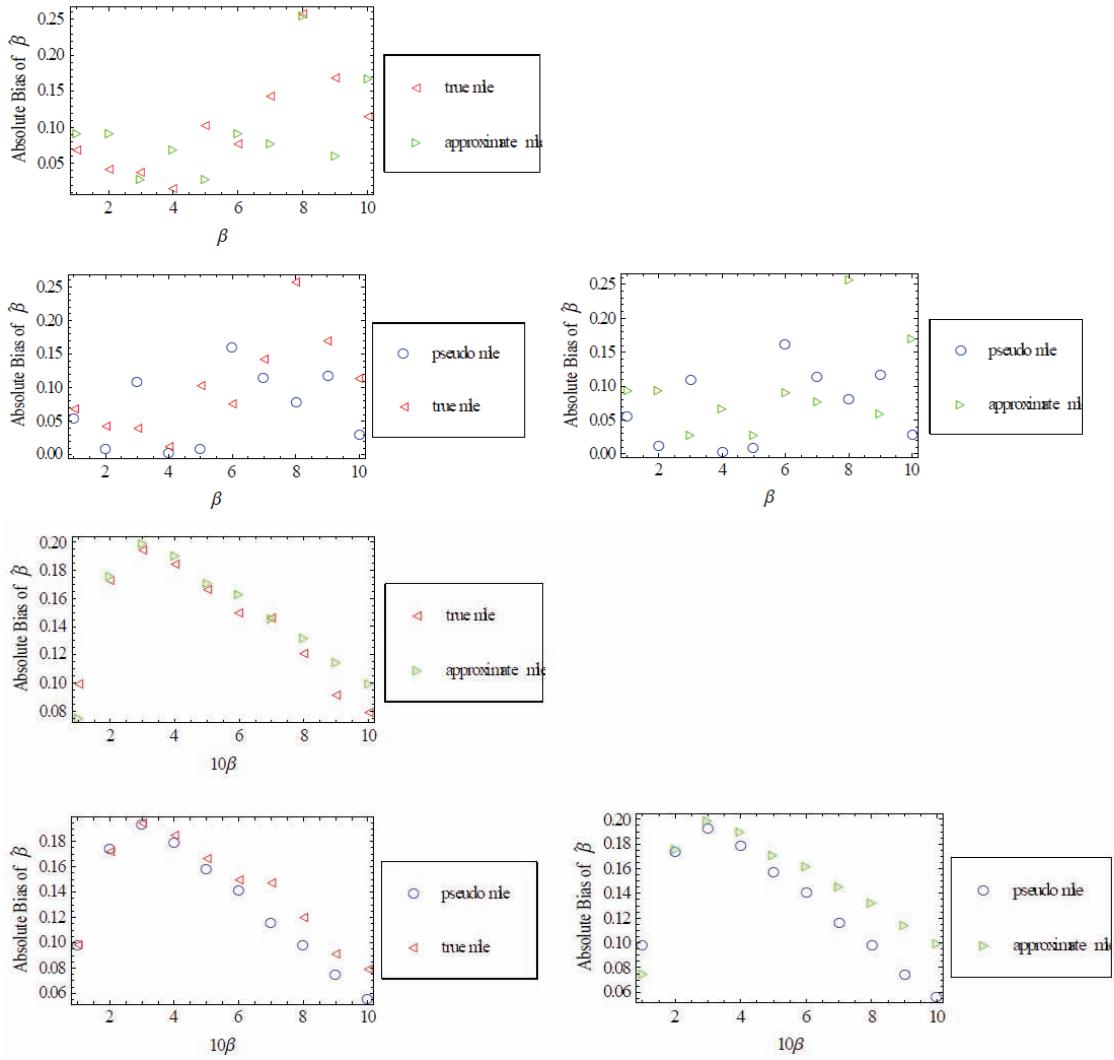


圖 5-4-2 Gamma 分佈的模擬結果： $\hat{\beta}$ 的 bias

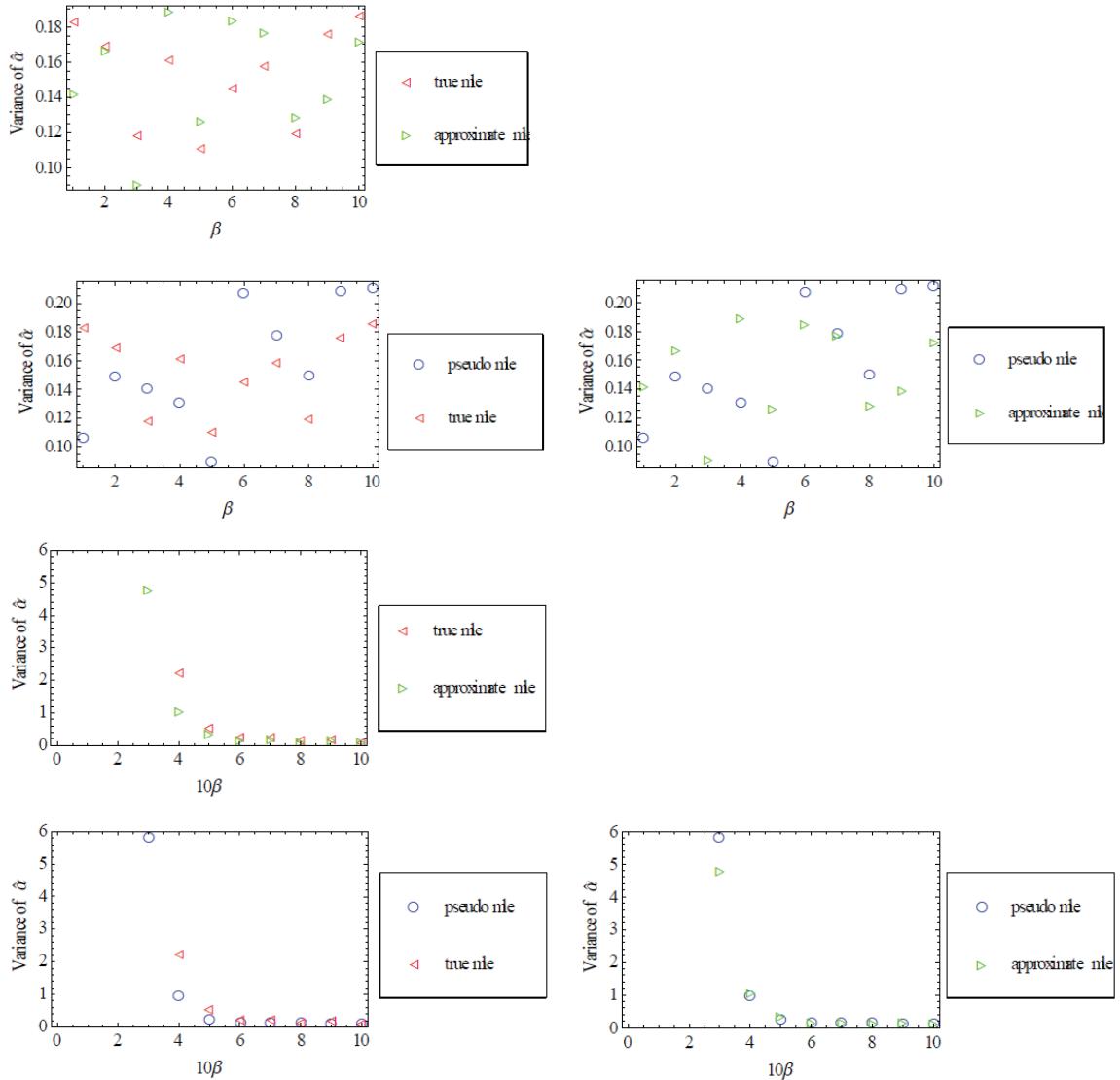


圖 5-4-3 Gamma 分佈的模擬結果： $\hat{\alpha}$ 的 variance

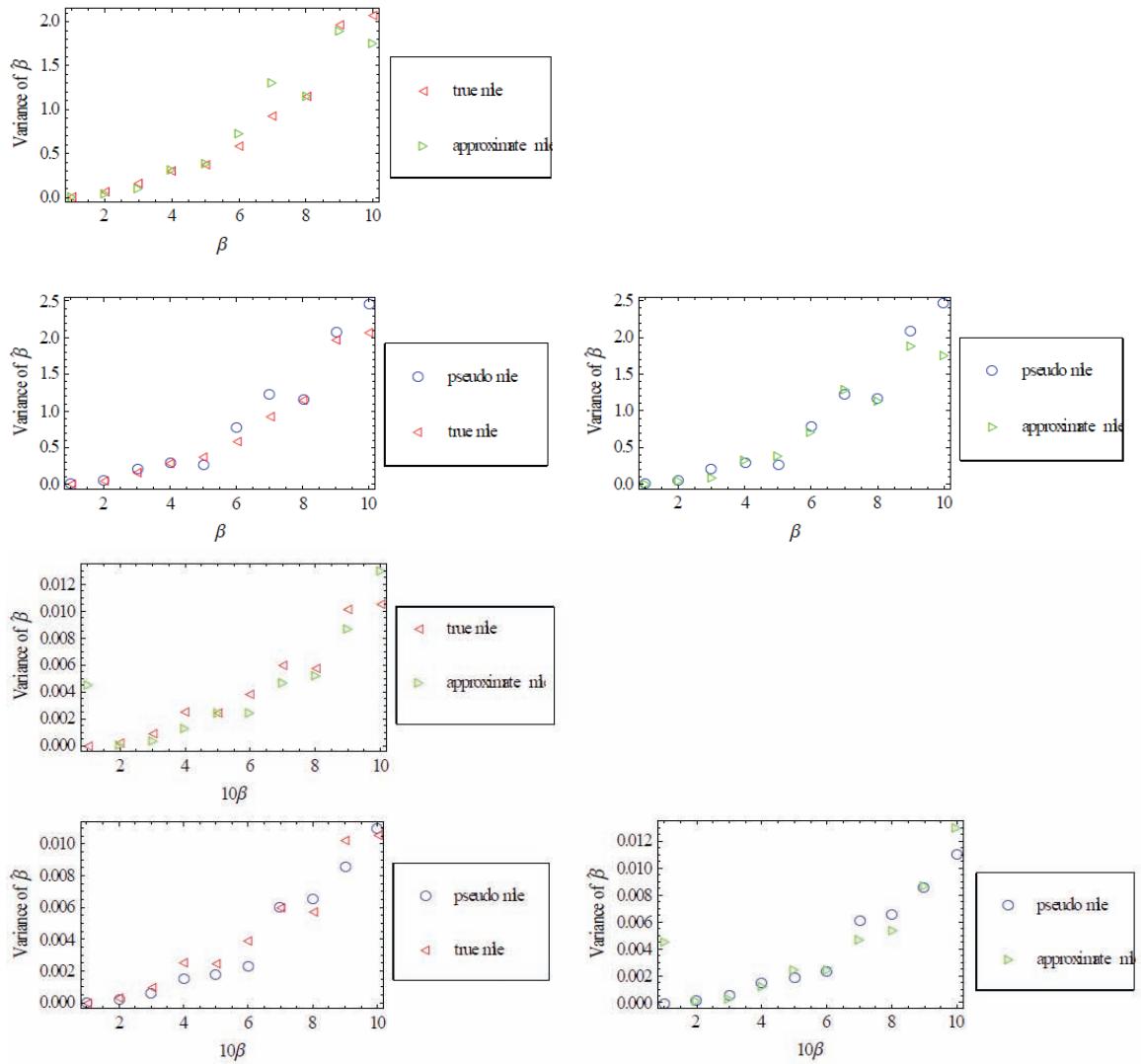


圖 5-4-4 Gamma 分佈的模擬結果： $\hat{\beta}$ 的 variance