第一章、緒論

1.1 商業智慧之核心－Data Mining

隨著資訊科技的進步及全球化市場的高度競爭衝擊下，商業智慧已成為今天企業提升其產業競爭力及獲利的解決方案。如何結合資訊技術與專業知識將資源轉變成具決策價值的商業智慧，成為每個企業最關心的議題。而其中客戶是企業最重要的資源。因此，如何獲取最佳的客戶、了解他們的需要、提供他們最個人化的產品與服務、及防止既有客戶的流失，便成為企業努力的目標，也因而產生了 Data Mining 之技術及應用。

何謂 Data Mining？簡單來說，Data Mining 就是在龐大的資料庫中尋找出有價值的隱藏事件，藉由統計及人工智慧的科學技術，將資料做深入分析，找出其中的知識，並根據企業的問題建立不同的模型，以提供企業進行決策時的依據。舉例來說：銀行或信用卡公司可藉由 Data Mining 的技術將其龐大的顧客資料做篩選、分析、推演及預測，找出哪些是最有貢獻的顧客，哪些是高流失率族群，或是預測一個新的產品或促銷活動可能帶來的回應率，能夠在適當的時間提供適合的產品及服務。也就是說，透過 Data Mining 企業得以了解他的顧客，掌握他們的喜好，滿足他們的需要。

一般 Data Mining 較常被應用的領域包括金融業、保險業、零售業、直效行銷業、通訊業、製造業、以及醫療服務業等。下面是一些 Data Mining 在這些產業運用的類型：

(1)在直效行銷業方面(Direct Marketing)

Data Mining 被廣泛的應用在郵寄活動上，藉由 Data Mining 可以預測出哪些人最有可能回覆郵寄活動和購買產品，企業可針對較少、較有可能消費的潛在客戶進行行銷活動。亦即如何利用最少的行銷成本，將行銷訊息傳遞到正確的顧客身上。

(2)在財務金融方面，進行財務預測(Financial Prediction)

利用時間序列分析(Time Series Analysis)或是類神經網路(Neural Network)，可以建立起季節性或是非季節性的財務數字預測，同時也能夠預估進行促銷活動對於銷售數字及獲利的影響，讓企業界儘量不做出錯誤的決策。

(3)在銀行業、保險業、信用卡公司等行業方面

近年來對於詐欺行為的偵測(Fraud Detection)非常關心，因為每年這些行業因詐欺行為而造成的損失都非常可觀。Data Mining 可以從一些信用不良的客戶資料中找出相似特徵並預測可能的詐欺行為，達到減少損失的目的。

(4)在零售業方面

利用 Data Mining 找出哪些顧客最有可能購買新產品以及哪些產品通常會一起被購買。

1.2 Data Mining 的技術

Data Mining 提供的技術很多，以下列舉幾項常用的功能：

1.分類 (Classification)

分類就是分析資料的所有特質，再將其指派至一個現有的群集中。例如，將信用狀況區分為高風險、中度風險及低風險，或是將顧客區分為高貢獻度族群、高忠誠度族群等。藉由分類可以對不同族群給予不同的產品及服務。它使用的 Data Mining 技術有決策樹(Decision Tree)，記憶基礎理解(Memory-Based Reasoning)等。

2.預測 (Prediction)

預測是根據對象屬性之過去觀察值來推估該屬性未來之值。例如由過去行銷活動所產生的反應來預測未來新活動的回應率，或是由顧客的職業、年齡、收入等人口屬性特質及其消費行為來預測可能的流失率等。使用的 Data Mining 技術包括時間序列分析(Time Series Analysis)、類神經網路(Neural Network)、決策樹(Decision Tree)、迴歸分析(Regression Analysis)、無母數迴歸分析(Nonparametric Regression Analysis)等。

3.關聯分組 (Affinity Grouping)又稱購物籃分析 (Market Basket Analysis)

關聯分組的功能是去發掘哪些事物總是同時發生。舉例來說，買 A 商品的通常同時購買 C 商品。美國一個應用 Data Mining 做購物籃分析的有名實例是零售連鎖商 Walmart 發現的「星期四、尿布和啤酒」。也就是由購物籃分析發現在禮拜四晚上，消費者通常會同時購買尿布和啤酒。這樣的發現提供了 Walmart 更多可與此結合的行銷點子。而實際上，購物籃分析就是達成交叉銷售的方法。

4.群集化 (Clustering)

群集化就是將一群異質的群體區隔為同質性較高的群體或是子群。它與分類不同的是，群集化沒有依靠事先明確定義的類別來進行分類，資料是根據自身的相近性而群集的。因此，群集化可說是分類的前置作業，它也是進行市場區隔的第一步。使用的技術為 K 平均法(K-means method)。

近年來，Data Mining 已成為企業熱門的話題，愈來愈多的企業想導入 Data Mining 的技術，美國的研究報告更是將 Data Mining 視為二十一世紀十大明星產業。例如，Time 時代雜誌就預估：Data Mining 將是 21 世紀最熱門之五大新興行業之一；而麻省理工學院 2000 年元月號?科技評論?(Technology Review)亦預測：未來會改變世界的十大新興科技趨勢中—Data Mining 名列前矛，可見它的重要性。如何讓數字說話，讓資料發光，成了二十一世紀最重要的一項任務。

面對二十一世紀的新知識經濟時代，商業智慧是提高企業競爭力的最終解決方案。而 Data Mining 正是商業智慧的核心，如何將 Data Mining 的技術結合企業領域的知識，真正達到提高利潤、提昇競爭力，將是未來企業努力的目標。

第二章、Data Mining 與統計方法

在本論文中, 主要以”無母數迴歸分析”來對我們有興趣的例子加以分析並對”區別分析”、 ”群集分析之 K-組平均數分群法 (K-means)”作一介紹。

1.區別分析 (Discriminant Analysis)：

有 2 個主要目標, 目標一：從數個已知的蒐集物(母體)而得到的物件(觀察值), 藉由圖形上(3 維或較小維度)或代數上來描述其不同的特徵, 並儘可能使得蒐集物的數值被區隔出來。目標二：把物件(觀察值)分類到 2 個或多個已歸類的類群中, 要強調的是導出一個規則能夠用來將一個新的物件最佳化地分配到已歸類的類群中。舉例說明如下, (1)商業上：由貸款者過去的存貸款、房地產等等資料, 以區別貸款者能否如期還款(2)行銷上：由顧客的收入、性別、教育水準等等資料, 以區別顧客是否會消費該商品。

觀察值(X_1, X_2, Λ, X_n), 分別來自兩個母體p_1, p_2, X_i 為 $p \times 1$的向量(共有 p 個因素), 已知其歸屬之母體。假設有一新的 $p \times 1$的向量 X_0, 如何根據上面的資訊來幫助我們判別它歸屬哪個母體？

利用 p 個因素用來定義某函數 $L : R^p \to R$, 稱為區別函數(discrimination function), 並決定判別區域 R_1 和 R_2。今有一新的觀察值, 若 $L(X_0)$的值落在 $R_1 \Rightarrow$ 將之歸類成p_1；若 $L(X_0)$的值落在 $R_2 \Rightarrow$ 將之歸類成p_2, 並且可藉由 $L(X_0)$我們可算出 X_0 屬於p_1 或p_2 之機率為多少。

兩個母體p_1, p_2, 其平均數各為u_1, u_2 而且變異數各為Σ_1, Σ_2, 其中 $u_1 = E(X|p_1), u_2 = E(X|p_2)$, 假設 $\Sigma_1 = \Sigma_2 = \Sigma$

考慮線性組合$Y = l'X$，$u_{1y} = l'u_1$，$u_{2y} = l'u_2$，$s_y^2 = l\Sigma l'$，令$d = u_1 - u_2$，

考慮一個比例為$\frac{y\text{之間}mean\text{的距離平方}}{y\text{之}\,\text{var}\,iance} = \frac{(u_{1y} - u_{2y})^2}{s_y^2} = \frac{(ld)^2}{(l\Sigma l)}$，此值愈大表示$Y$愈能區分$p_1, p_2$。

問題：如何找l使得此 ratio 為最大？由定理可知$l = c\Sigma^{-1}(u_1 - u_2)$，對於任何一個$c \neq 0$。而$Y = c(u_1 - u_2)'\Sigma^{-1}X$稱為”Fisher linear discrimination function”。

新的觀察值X_0，其 discrimination function 值為$y_0 = (u_1 - u_2)'\Sigma^{-1}X_0$。

令m為p_1, p_2轉換後平均值之中點：

$$m = \frac{1}{2}(u_{1y} + u_{2y}) = \frac{1}{2}(l'u_1 + l'u_2) = \frac{1}{2}(u_1 - u_2)\Sigma^{-1}(u_1 + u_2)$$

若$y_0 > m \Rightarrow X_0 \in p_1$，$y_0 < m \Rightarrow X_0 \in p_2$

因為u_1, u_2, Σ為未知的，而用已知的$\overline{X}_1, \overline{X}_2, S_p$代替，並令$\hat{l} = (\overline{X}_1 - \overline{X}_2)S_p^{-1}$

，由定理可知$\hat{l}$能使得此比例$\frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\hat{l}'d)^2}{(\hat{l}'S_p\hat{l})}$，其中$d = \overline{X}_1 - \overline{X}_2$。

2.群集分析 (Cluster Analysis)：

基本目標為發現在項目(或變數)之中自然的分群(Grouping)，其中依其觀察值的相似性(similarities)或距離(distances)較近的方法使得群內個體差異小，而群間個體差異大。主要有兩大型式：分層法(Hierarchical)與非分層法(Nonhierarchical)，分層法又分成凝聚分層法(Agglomerative)與分離分層法(Divisve)；其中凝聚分層法開

始時每一個體為一群，然後最近的兩個體合成一群，依次結合使群組愈變愈少，最後所有個體結成一群，其中依照距離計算方式的不同而有最近法(又稱單一聯結法 Single Linkage)、最遠法(又稱完全聯結法 Complete Linkage)、平均法(Average Linkage)、中心法(Centroid)與華德法(Ward)。非分層法以 K-組平均數分群法(K－means)為代表。舉例說明如下，想要知道台灣上市電子類股其財務績效，進而作為投資人投資的參考依據，蒐集 20 家主要公司得到七個變數(1.本益比 2.投資報酬率 3.資產負債比 4.銷售成長率 5.每股獲利成長率 6.純利百分比 7.每股紅利)，將這 20 家公司分成?鼓勵積極投資?、?一般投資?、?不值得投資?3 個群組。

K-組平均數分群法 (K-means)：

麥昆(J.B. MacQueen)於 1967 年提出，常使用於多維向量中$(X_1, X_2, X_3, \Lambda, X_n)$，為了方便說明，我們以二維向量$(X_1, X_2)$來描述如下：

步驟一：選擇 K 個資料點作為種子，並作為群集的質心

步驟二：將每一資料點分配到距離質心最接近的群集中

步驟三：計算每一個群集的質心，只要將群集中每一個點的位置加以平均即可。

步驟四：找出新群集，每一點再次被分配到距離質心最接近的群集中。

步驟五：重複進行直到群集邊界不再變動為止。

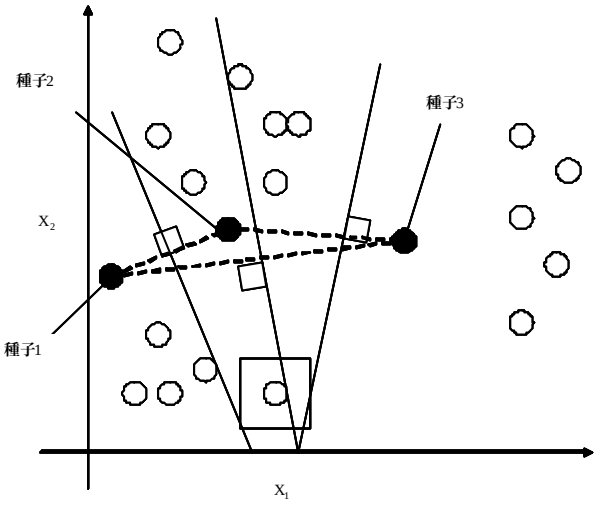


圖10.3 初始種子決定了初始的群集邊界

起初的三個種子由虛線連起來，而實線代表這三個種子所購成群集的邊界。注意到一個被方形圍起來的那一個點。

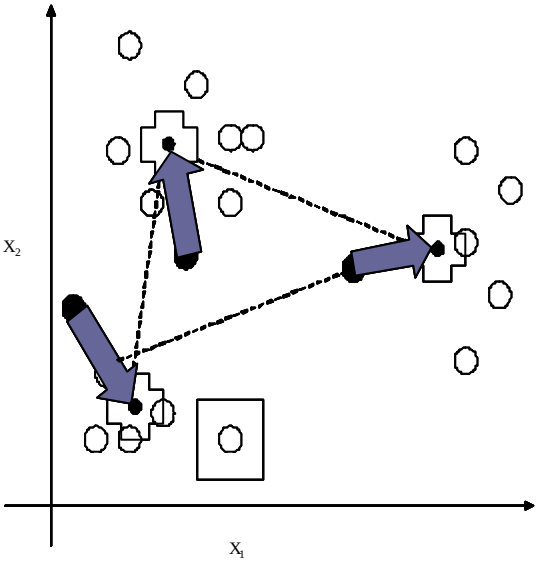


圖10.4 計算新群集的質心

新的質心由一個十字形圖樣來標示，箭頭代表原本的種子從原來的位置移動到新位置的情況。

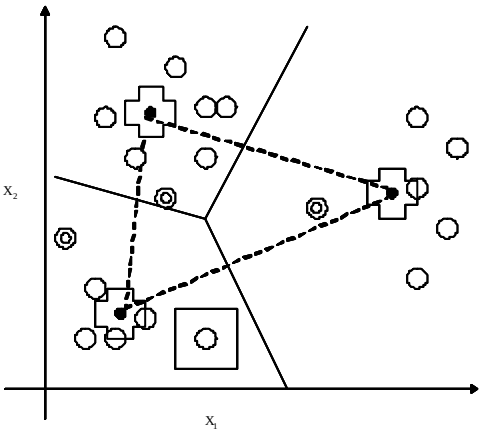


圖10.5 每一次重複的過程中，所有群集分配都必須重新計算一次

顯示出新群集的邊界，這個界線是由與兩個質心距離相等的點所構成。注意到一個被方形圍起來的那一個點，它原本屬於第二個群集，現在被重新分配到第一個群集。

＊注意到區別分析與群集分析不同：區別分析是已知每個觀察個體所屬的群體，並利用區別函數建立區別規則，亦即在分析之前就已知共分為幾群；而群集分析之前並不知道觀察個體所屬的群體，亦即在分析之前並不知道分為幾群。

蕭凱芳(2001)利用區別分析與 K-組平均數分群法對於 Fisher's iris(鳶尾花)的資料進行分析，並驗証加權法的利弊，針對一些原本被分錯群的觀察值進行修正，使其可回歸於原本所屬之群體，但仍是有些許的觀察值無法修正，此加權法雖不完備，但仍有可議討論之處。

3.無母數迴歸分析 (Nonparametric Regression Analysis)：

我們首先考慮以下模型，

$$y_k = f(t_k) + e_k, \qquad k = 1,\mathrm{K}\,,n$$

y_k 為在可控制時間點 t_k 的觀察值，$f(t)$ 為一平滑曲線並且 e_k 其平均數為零且不相關的隨機變數。有很多可用來估計 $f(t)$ 曲線的方法，例如以核為基礎的方法(kernel-based methods)以及曲線樣條法(smoothing splines methods)。而在近年來，無母數迴歸使用樣條的方法已快速成為統計上的一個分支。

假設估計值 $\hat{f}$ 能夠使得(1)式之數值為最小

$$\sum_{k=1}^{n}\{y_k - f(t_k)\}^2 + l\int f''(t)^2\,dt, \qquad (1)$$

在所有二次可微函數 f 的一類中，l 被當作是一個平滑參數。l 扮演一個關鍵角色於控制由殘差平方和 $\left(\sum_{k=1}^{n}\{y_k - f(t_k)\}^2\right)$ 所表現的適合度(goodness of fit)以及藉由積分二次導數平方項 $\left(\int f''(t)^2\,dt\right)$ 所表現的平滑性(smoothness)之間的交換(trade-off)。

第三章、無母數迴歸的診斷及其強穩性

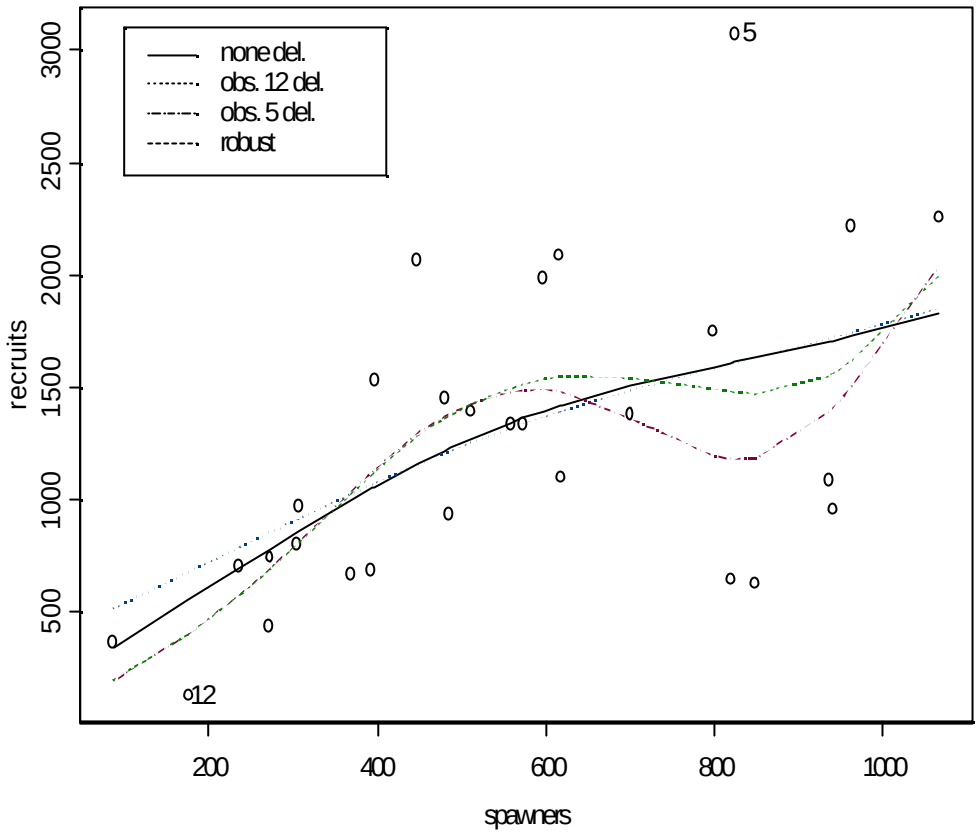
3.1 介紹

一種關鍵的影響包含平滑參數的估計。從貝氏的觀點來看，平滑參數扮演一個在相關混合效應模型中隨機效應項中的先驗參數角色(Speed，1991)。於是，平滑參數的選擇是相同於先驗參數的選擇。所以，平滑參數的影響可以當作在貝氏分析中先驗選擇的影響。因此，對於一個在平滑參數估計中有極大影響力的觀察值，其在“先驗的選擇上”亦是有影響力的。也就是說，對於在平滑的程度上，這種有影響力的觀察值(influential observation)可能在先驗的選擇上有重大的影響。一般在迴歸領域裡使用刪除一點法(delete-one method)法在平滑參數估計上做診斷，然而還沒在無母數迴歸的環境下討論(Thomas，1991)。我們解釋了刪除一點法(delete-one method)可能不能反映一些有影響力的觀察值在無母數迴歸中其“真正”的影響。我們提出其它的刪除診斷方法當作另一個選擇。其它在文獻中還未被討論到的重要事情就是估計影響的精確性。在無母數迴歸中一個主要的目標就是給予精確的配適。信賴區或是預測區能夠傳達關於曲線估計精確性的訊息。刪除一個有影響力的觀察值可能會導致不同的信賴區。此外，當檢定 $f(t)$ 是一個簡單參數迴歸估計的虛無假設時，由於刪除一個有影響力的觀察值可能會得到不同的檢定結果。這些有影響力的觀察值而因此在推論上是有影響力的。我們提出一個以抽樣為基礎的程序來辨別這些在推論上是有影響力的觀察值。

就如同 Carroll 與 Ruppert(p.175,1988)指出，當只有診斷或是只有強穩的方法不能如同將兩者方法適當組合來得較為有用。關於有影響力的觀察值，我們學習到的愈多，我們愈有可能發展出一個合理且強穩的方法。本論文的目標為發展出數個明顯的影響方法來辨別上述討論關於有影響力的觀察值並且發展出一個合理且強穩的無母數迴歸方法藉由適當地把這些有影響力的觀察值向下加權。這些有影響力的方法只能被用來察覺出與此模型不一致的觀察值。更進一步，相關強穩的方法能提供一個替代性的配適。在 3.2 節將給一個引發動機的例子來解釋這些有趣的問題。在 3.3 節將發展出一些有影響力的診斷方法，而在 3.4 節將有相關強穩的方法。

3.2 引發動機的例子

資料為從 Carroll 與 Ruppert(1988)取得。記錄 28 年(1940-1967)來每一年成熟的母鮭魚與其子魚數量。在圖一，實線為原始的配適。平滑參數估計值為 0.0488。注意資料在刪除第 5 個觀察值之後，配適曲線的行為以及平滑參數同時地改變，而完全地不同於原始的配適。另外，資料在刪除第 5 個觀察值後，重新配適所得新的配適曲線的行為是完全地不同於當去除第 12 個觀察值時所重新配適的曲線。在沒有第 12 個觀察值下，配適曲線行為是類似於原始配適曲線的。然而，藉著檢查不同配適的 GCV 估計值，相當驚訝的是刪除第 5 個觀察值與刪除第 12 個觀察值其 GCV 估計值的改變是非常接近的。在沒有第 12 個觀察值之下，平滑參數估計值為 0.0958；而在沒有第 5 個觀察值之下，平滑參數估計值為 0.0015。刪除第 5 個觀察值而得之 GCV 估計值與原始 GCV 估計值之差異的絕對值為 0.0473，然而刪除第 12 個觀察值而得之 GCV 估計值與原始 GCV 估計值之差異的絕對值為 0.0470。這意味著藉由差異而測量 GCV 估計值的改變並非總是與配適曲線行為的改變有關。事實上，原始 GCV 估計值與不含第 5 個觀察值之 GCV 估計值而得之比例為 32.49；然而原始 GCV 估計值與不含第 12 個觀察值之 GCV 估計值而得之比例為 0.50。因此，我們應合理地使用比例所提供的資訊來發展我們的方法。也要注意的是刪除第 5 個觀察值而得之平滑參數估計值遠小於刪除第 12 個觀察值而得之平滑參數估計值。這似乎意味著配適曲線在小的平滑參數估計比起在大的平滑參數估計有較明顯的改變。



圖一

表示為母鮭魚與其子魚數量的圖形。並可得到在不同的配適之下 GCV **的估計值：實線(非強穩配適，** 0.0488)；**點線(刪除第** 12 **個觀察值，** 0.095)；**破折線(緊密間隔的，刪除第** 5 **個觀察值，** 0.0015)；**破折線(強穩配適，** 0.00146)

3.3 診斷

在這一節，我們提出診斷來辨別那些在平滑參數選擇以及統計推論上有重大影響力的觀察值。令 $f(t)=\sum_{j=1}^{p} a_j B_j(t)$, **其中** p **為適當選擇基底函數的個數，通常至少要足夠大以保證近似的精確性，例如，** $p=n+2$。 $B_j(t)$ **為基底函數，一般常使用** B-**樣條(**B-splines)。**因此，(1)式可簡化為** $(y-Ba)^t(y-Ba)+\lambda a^t P a,$

其中 $y=(y_1,\mathrm{K},y_n)^t, B=(b_1,\mathrm{K},b_n)^t=\{B_j(t_i)\}_{ij}$ 為 $n\times p$ 矩陣,

$a=(a_1,\mathrm{K},a_p)^t$, 且 $P=\left\{\int B_l''(t)B_j''(t)dt\right\}_{lj}$ 為 $p\times p$ 矩陣。

a 係數的估計值為 $\hat{a}(l)=(B^tB+lP)^{-1}B^ty$.於是,

最小估計值為 $\hat{f}(l)=\{\hat{f}(t_1,l),\mathrm{K},\hat{f}(t_n,l)\}^t=B(B^tB+lP)^{-1}B^ty=H(l)y,$ (2)

其中 $H(l)=\{h_{ik}(l)\}_{ik}=B(B^tB+lP)^{-1}B^t$ 為帽子矩陣, $i=1,\mathrm{K},n,\quad k=1,\mathrm{K},n.$

這個方法計算上較為簡單(Hastie 與 Tibshirani,1990,pp.27-29;Wahba,1990,chapter7), 與有限維度的貝氏型相關(Silverman,1985)以及其類似一個簡單的脊迴歸公式。

在配適過程中, 選擇一個好的平滑參數估計是非常重要的。數個一般常用來平滑參數估計為使得目標函數 $\log\{\hat{s}^2(l)\}+f\{H(l)\}$ 最小,

其中 $\hat{s}^2(l)=y^t\{I-H(l)\}^t\{I-H(l)\}y/[n-tr\{H(l)\}]$ 為變異數估計值而且 $f(\bullet)$ 為對於樣條配適之平滑性的處罰函數, 並且 $tr(A)$ 為 A 矩陣的跡(trace)。在本論文中, 令 $\hat{l}$ 為 GCV 的估計值(Golub,Heath,and Wahba,1979)；也就是使得目標函數當 $j\{H(l)\}=-\log\{1-tr\{H(l)\}/n\}$ 的情形下為最小。

一個有局部帶寬變化之樣條平滑法相當近似一個藉由以核為基礎來估計的平滑法。樣條配適也可表示為 $\hat{f}(t,l)=n^{-1}\sum_{k=1}^{n}G(t,t_k,l)y_k$ (Silverman,1985), 其中 $G(t,s,l)$ 為某種加權函數。注意當 $|t-s|$ 增加時, $G(t,s,l)$ 指數地遞減為 0。因此, 預測值 $\hat{f}(t_k,l)$ 主要地依靠在 t_k 附近的觀察點。設 $\hat{f}_{(i)}(t,l)$ 為一個刪除第 i 個觀察

值的配適函數。因為 $\hat{f}(t_k,l)-\hat{f}_{(i)}(t_k,l)=h_{ik}(l)e_i(l)/\{1-h_{ii}(l)\}$ 而且 $h_{ik}(l)\approx n^{-1}G(t_k,t_i,l)$，$\hat{f}(t_k,l)-\hat{f}_{(i)}(t_k,l)\approx 0$ 當 $|t_k-t_i|$ 之值很大時，並且 $e_i(l)=y_i-\hat{f}(t_i,l)$。設 t_k 遠離 t_i 點，以及 $\hat{l}_{(i)}$ 為刪除觀察點 i 之平滑參數估計值。因為 $\hat{f}\left(t_k,\hat{l}_{(i)}\right)-\hat{f}_{(i)}\left(t_k,\hat{l}_{(i)}\right)$ 之差很小，如此一來 $\hat{f}\left(t_k,\hat{l}\right)$ 與 $\hat{f}_{(i)}\left(t_k,\hat{l}_{(i)}\right)$ 其數值之差異，$\hat{f}\left(t_k,\hat{l}\right)-\hat{f}_{(i)}\left(t_k,\hat{l}_{(i)}\right)=\left\{\hat{f}\left(t_k,\hat{l}\right)-\hat{f}\left(t_k,\hat{l}_{(i)}\right)\right\}+\left\{\hat{f}\left(t_k,\hat{l}_{(i)}\right)-\hat{f}_{(i)}\left(t_k,\hat{l}_{(i)}\right)\right\}$，主要依賴 $\hat{f}\left(t_k,\hat{l}\right)-\hat{f}\left(t_k,\hat{l}_{(i)}\right)$。

經由泰勒一階近似，可得到

$$\hat{f}\left(t_k,\hat{l}\right)-\hat{f}\left(t_k,\hat{l}_{(i)}\right)\approx\frac{1}{2}\left[\left\{\frac{\partial\hat{f}(t_k,l)}{\partial l}\right\}_{l=\hat{l}}+\left\{\frac{\partial\hat{f}(t_k,l)}{\partial l}\right\}_{l=\hat{l}_{(i)}}\right]\left(\hat{l}-\hat{l}_{(i)}\right).$$

因此，由於去除觀察值 i 使得平滑參數估計與配適值同時地改變而反映在藉由 $\hat{l}-\hat{l}_{(i)}$ 項造成的平滑參數估計改變與關於平滑參數數值在配適值改變比率上。由於去除觀察值 i 或是因為平滑參數估計在配適曲線有很高的敏感性使得在平滑參數估計上有明顯的改變而導致 $\hat{f}\left(t_k,\hat{l}\right)-\hat{f}_{(i)}\left(t_k,\hat{l}_{(i)}\right)$ 數值之差異可能是大的。這就解釋了為什麼當去除某些觀察值而同時地改變平滑參數估計值，可能不僅在影響配適曲線的局部(local)行為而且在配適曲線的整體(global)行為。這也解釋了在圖一內這兩個配適曲線上不同的行為(實線與緊密間隔破折線)。因為 $\partial G(s,t,l)/\partial l=O\left(l^{-5/4}\right)$，此加權函數 $G(s,t,l)$ 對於 l 的一階導數之數值為當 l 接近 0 時其值接近無窮大而當 l 接近無窮大時其值接近 0。因此，當 l 接近 0 時，$\left\{\partial\hat{f}(t_k,l)/\partial l\right\}_{l=\hat{l}}$ 接近無窮大。也就是說，配適曲線對於小的平滑參數估計值特別地敏感。這導致了在配適曲線的行為上，一個小的平滑參數估計值上小小的改變比較起一個大的平滑參數估計值上大大的改變可能會有較大的影響。這解釋了在圖一上不同配適曲線的行為(緊密間隔破折線與點線)。因此，在配適曲線的行為

上，一般使用刪除診斷量，如同 $\hat{l} - \hat{l}_{(i)}$，可能不總是在辨別最有影響力的觀察值上為有效的。在 Hastie 與 Tibshirani (1990)的庫克距離(Cook ' s distance)描述中有固定的平滑參數為

$$C_i(l) = \frac{\left\| \hat{f}(l) - \hat{f}_{(i)}(l) \right\|^2}{tr\{H^t(l)H(l)\}\hat{s}^2(l)}$$

其中 $\hat{f}_{(i)}(l) = \{\hat{f}_{(i)}(t_1, l), \mathrm{K}, \hat{f}_{(i)}(t_n, l)\}^t$ 而且 $\|\bullet\|$ 表示歐氏長度(Euclidean norm)。

一個類似的影響方法為考慮到平滑參數之改變，為 $\frac{\left\| \hat{f}(\hat{l}) - \hat{f}_{(i)}(\hat{l}_{(i)}) \right\|^2}{tr\{H^t(l)H(\hat{l})\}\hat{s}^2(\hat{l})}$

其中 $\hat{f}_{(i)}(\hat{l}_{(i)}) = \{\hat{f}_{(i)}(t_1, \hat{l}_{(i)}), \mathrm{K}, \hat{f}_{(i)}(t_n, \hat{l}_{(i)})\}^t$。

因為 $\left\| \hat{f}(\hat{l}) - \hat{f}_{(i)}(\hat{l}_{(i)}) \right\|^2 \approx \frac{1}{4}\sum_{k=1}^{n}\left[\left\{ \frac{\partial \hat{f}(t_k, l)}{\partial l} \right\}_{l=\hat{l}} + \left\{ \frac{\partial \hat{f}(t_k, l)}{\partial l} \right\}_{l=\hat{l}_{(i)}} \right]^2 (\hat{l} - \hat{l}_{(i)})^2$，　而且 $\partial \hat{f}(t_k, l)/\partial l = O(l^{-5/4})$，我們將提出下列的方法以替代刪除一點診斷量

$$S_i = \left\{ \left(\frac{1}{\hat{l}^{5/4}} + \frac{1}{\hat{l}_{(i)}^{5/4}} \right)(\hat{l} - \hat{l}_{(i)}) \right\}^2 \qquad (3)$$

這些診斷量能被看作是一個加權的刪除一點診斷量。在一個非常小的平滑參數估計改變上施以較多的權重就如同藉由比例項 $1/\hat{l}^{5/4} + 1/\hat{l}_{(i)}^{5/4}$ 反映一樣。將 S_i 常態化為 $S_i \Big/ \left(\sum_{k=1}^{n} S_k^2 \right)^{1/2}$，而當觀察點之數值超過 $\pm 2/\sqrt{n}$ 時，此觀察點即診斷為有影響力的點。

3.4 強穩的 GCV 估計

在這一節，我們首先提出辨別有影響力的觀察值，然後應用一個將這些有影響力觀察值向下加權的強穩方法。為了要得到一個更強穩的平滑參數估計值，我們提出一個於目標函數上將有影響力的觀察值整體向下加權的方法。因此，提出的強穩平滑參數估計值為目標函數的最小數量式，

$$\log\left\{\hat{s}_v^2(l)\right\}-\log\left[1-tr\left\{\underset{\sim}{H}_v(l)\right\}\Big/ n\bar{v}\right], \tag{4}$$

其中

$$\hat{s}_v^2(l)=\frac{y^t\left\{\underset{\sim}{I}-\underset{\sim}{H}_v(l)\right\}\underset{\sim}{V}\left\{\underset{\sim}{I}-\underset{\sim}{H}_v(l)\right\}y}{\left[n\bar{v}-tr\left\{\underset{\sim}{H}_v(l)\right\}\right]^2}$$

而其中

$$\underset{\sim}{H}_n(l)=\underset{\sim}{B}\left[\underset{\sim}{B}^t\left(diag\left\{n\left(\frac{S_i^{1/2}-m}{s}\right)\right\}\right)\underset{\sim}{B}+l\,\underset{\sim}{P}\right]^{-1}\underset{\sim}{B}^t\left(diag\left[n\left(\frac{S_i^{1/2}-m}{s}\right)\right]\right),$$

$$\bar{v}=\sum_{i=1}^{n}v\left(\frac{S_i^{1/2}-m}{s}\right)$$

而其中 $v(t)$ 為加權函數並且大的 S_i 數值之觀察值藉由 $v(t)$ 來向下加權，m 與 s 分別代表 $S_1^{1/2},\Lambda\ ,S_n^{1/2}$ 樣本中之樣本中位數與樣本中位數絕對差異。數個 $v(t)$ 的選擇被建議，包含 Box kernel,

$$n_b(t)=\begin{cases}1 & \|t\|\le 1\\ 0 & \|t\|>1\end{cases}$$

Triangular kernel,

$$n_t(t)=\begin{cases}1-\|t\| & \|t\|\le 1\\ 0 & \|t\|>1\end{cases}$$

Parzen type of kernel,

$$n_p(t)=\begin{cases}1-\dfrac{3t^2}{16} & \|t\|\le 2\\ \dfrac{-t^2+5\|t\|-6}{4} & 2<\|t\|\le 3\\ 0 & \|t\|>3\end{cases}$$

Gaussian kernel,

$$n_g(t)=\frac{1}{\sqrt{2p}}e^{-\frac{t^2}{2}}$$

Biweight kernel,

$$n_{bi}(t)=\begin{cases}\dfrac{15}{16}\{1-t^2\}^2 & \|t\|\le 1\\ 0 & \|t\|>1\end{cases}$$

以及 Epanechnikov kernel,

$$n_e(t)=\begin{cases}\dfrac{3\sqrt{5}}{4}\left(1-0.2t^2\right) & \|t\|\le\sqrt{5}\\ 0 & \|t\|>\sqrt{5}\end{cases}$$

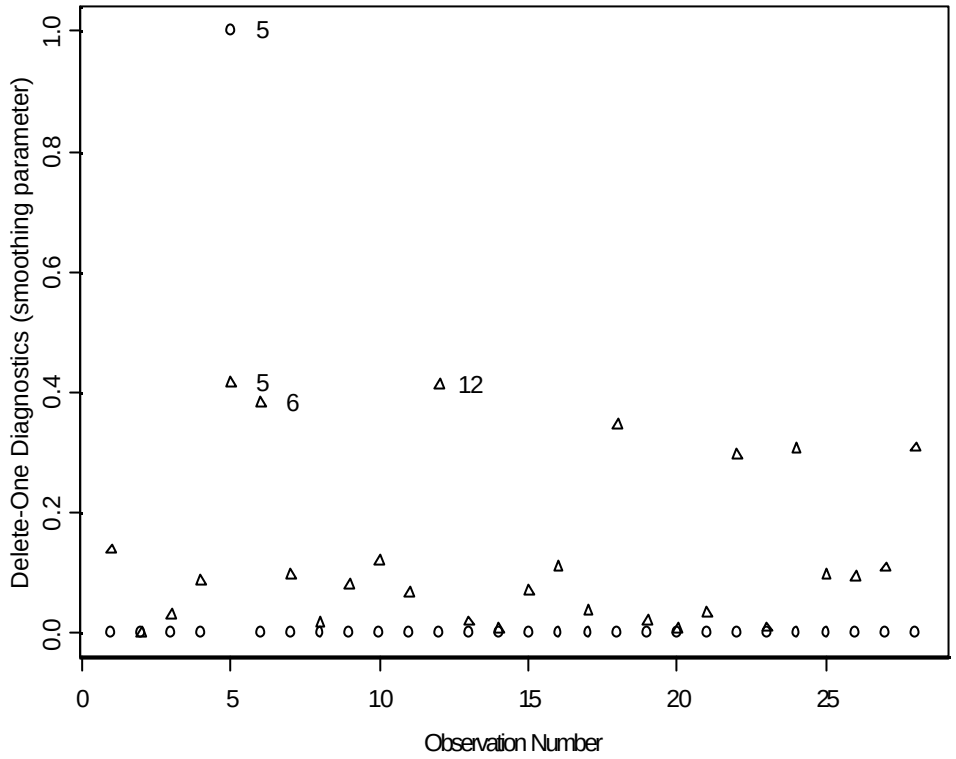
原則上，加權函數$v(t)$扮演一個如同在核估計(Kernel-type estimate)中核的角色一般。在估計的過程中，有較大S_i數值之觀察值被向下加權。如此一來在(2)式中的平滑樣條配適的平滑參數估計值可以取為強穩的 GCV 估計值。

第四章、實例及模擬

在這一節，我們呈現出一個模擬的例子以及實際資料的例子。在這些例子中，之前在 3.3 節討論到關於有影響力的觀察值將會被辨別出來。此外，在 3.4 節給定的強穩配適在某些觀察值的影響之下比起在第二個以非強穩描述來模擬的例子更加的穩定。在實際的例子中，經由強穩的配適來解釋資料有些地方不同於非強穩的方式。

4.1 資料例子

在 3.3 節所給定的(3)式 S_i 現在應用到鮭魚資料的例子上。當刪除一點診斷量 $\hat{l} - \hat{l}_{(i)}$ 應用到此資料的例子上時，數個觀察值被辨別出來。然而當檢查 S_i 值時，只有第 5 個觀察值脫穎而出。在圖二給定 $\hat{l} - \hat{l}_{(i)}$ 與 S_i 的圖形。當檢查觀察值之中刪除其中一點而得資料配適曲線的行為時，只有當第 5 個觀察值被刪除而得的資料配適顯著的不同於原始的配適。這證實了 S_i 比 $\hat{l} - \hat{l}_{(i)}$ 更為適合去診斷有影響力的觀察值。注意到第 5 個觀察值比起它附近的觀察值有明顯較大的子魚數。在 3.4 節所發展出強穩的方法也將應用於這些資料之上。以加權函數 $v(t)$ 為基礎的強穩 GCV 估計值為 0.00146。用強穩 GCV 估計值而得配適曲線的行為是與原始 GCV 估計值不同的。在母魚 600 之前的配適曲線上強穩的估計值比較起非強穩的估計值顯示有較大的斜率。在母魚 600 之後，強穩的估計值顯示出在母魚與子魚之間有非線性的關係，然而原始的估計值卻顯示出有線性的關係。用強穩 GCV 估計值而得配適曲線的行為也是不同於當資料刪除第 5 個觀察值之後而得配適曲線的行為。強穩的估計值在母魚 600 之後比較起當刪除第 5 個觀察值之後用非強穩的方法顯示有較小的曲率。在圖一也給定了以加權函數 $v_g(t)$ 為基礎之強穩配適。注意到以加權函數 $v_g(t)$、$v_p(t)$、$v_e(t)$、$v_b(t)$ 為基礎之強穩 GCV 估計值而得配適曲線的行為是非常相似的。



圖二

表示對於母鮭魚與其子魚的資料而言，常用的刪除一點診斷量 $\hat{l} - \hat{l}_{(i)}$ **以及所提出的刪除診斷量** S_i **對觀察值的圖形。** $\hat{l} - \hat{l}_{(i)}$ **與** S_i **以不同的符號表示：用** S_i **來診斷，以。表示；用** $\hat{l} - \hat{l}_{(i)}$ **來診斷，以△表示。**

4.2 模擬

在科學的研究上使用到很多的分配。而有一些分配在理論上是有良好根據的。其中之一為 Lorentzian 分配函數，也稱為柯西分配(Cauchy distribution)。Lorentzian 分配以$C(m,\Gamma)$表示，其中 m 為平均數且為位置參數，而Γ為尺度參數。

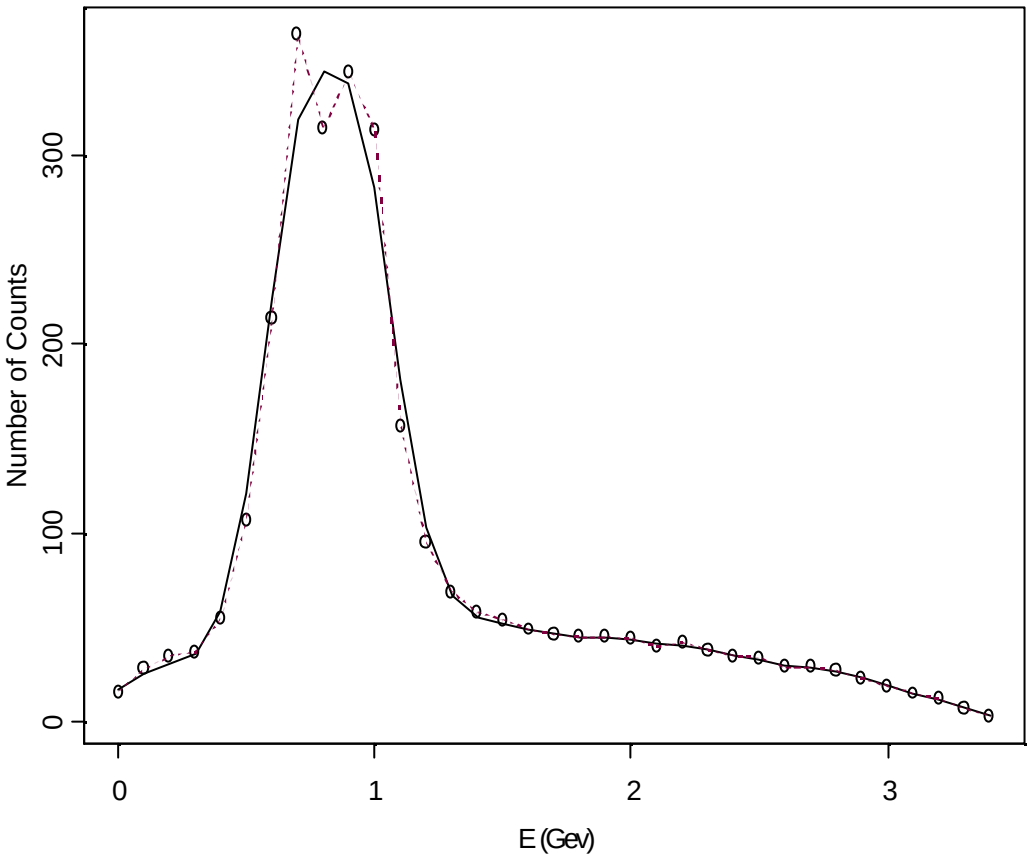
這個分配對於用來描述物理上的共振現象是非常適當的，例如在 Mossbauer 效應核子混合物或是粒子反應中輻射吸收等具有能量的變動(參考 Bevington 與 Robinson，1992)。Lorentzian 分配為單位模型而且其眾數(也稱為頂峰)為 m。我們

產生一些資料能夠被用來描述在一個基本的粒子交互反應下一個共鳴的狀態。我們也試圖在一個以模擬資料點為基礎的隨機訊號或是干擾去辨別從一個平滑背景的相反訊號。從一個隨機亂數產生器模擬出 2000 筆獨立且屬於相同分配 $C(1,0.1)$ 的變數資料。然後，在一個長條寬為 0.1 之下，2000 筆模擬的事件次數(直方圖)能夠得到。二元的資料，$Z_{1i}, i=1,\Lambda,35$，使用範圍為從 0 到 3.5。同樣地，其它由隨機亂數產生器模擬出 2000 筆獨立且屬於相同分配的 $C(0.75,0.1)$，$C(1.25,0.1)$ 以及 $C(1.5,0.1)$ 變數資料。在一個相同的長條寬之下，其它 35 筆二元的資料點能夠被使用，Z_{2i}, Z_{3i} 以及 $Z_{4i}, i=1,\Lambda,35$，個別地相對應 $C(0.75,0.1)$，$C(1.25,0.1)$ 以及 $C(1.5,0.1)$ 分配。資料點 $y_i=-1+45x_i-13x_i^2+0.5z_{1i}$，在一個平滑地變動背景之下，其中 y_i 對應於計算的數字，$x_i=(i-0.5)/10$ 於能量的水準範圍為從 0(GeV)到 3.5(GeV)，z_{1i} 可以描述具有一個頂峰而且二階多項式函數說明對於一個平滑背景合理的近似。這個模型為 $y_i=f(x_i)+e_i, i=1,\Lambda,35$，而且 $f(\bullet)$ 可以藉由平滑樣條法來配適。在圖三，此線為非強穩的平滑樣條配適而且強穩的平滑樣條配適使用第(4)式所給定的強穩 GCV。兩個配適插入資料。相反的訊號可藉由這兩個方法被辨別出來。

然而，考慮資料點 $y_i=-1+45x_i-13x_i^2+0.5z_{1i}+0.5z_{2i}$，其中 z_{1i} 與 z_{2i} 可以描述一個具有兩個緊密地間隔頂峰大約在 0.75 與 1。在圖三，實線為非強穩的平滑樣條配適，然而點線為平滑的樣條配適並使用在第(4)式所給定的強穩 GCV。平滑的參數估計值在兩個配適分別為 1.57×10^{-5} 以及 1.1×10^{-15}。只有一個在閉區間 $[0.8,0.9]$ 的頂峰已藉由非強穩的方法來預測。注意到以非強穩的方法所得到的第 8 個觀察值以及第 9 個觀察值(相對應於在閉區間 $[0.7,0.9]$ 計算的數字)的預測數值顯著地小於或大於資料的數值。第 8 個以及第 9 個觀察值比較起其它的觀察值有較大的 S_i 值。這意味著這些觀察值可能導致太過平滑的傾向。這個強穩的 GCV 方法將這些有影響力的觀察值向下加權。因此，藉由強穩的方法，兩個頂峰被正確地預測。然而，考慮資料 $y_i=-1+45x_i-13x_i^2+0.5z_{2i}+0.5z_{4i}$，其中 z_{2i} 以及 z_{4i} 可以描述一個大約在 0.75 以及 1.5 個別地兩個頂峰的狀態。相同於第一個模擬，非強穩以及強穩的平滑樣條配適都插入資料以及正確地預測頂峰。

然而，考慮資料 $y_i = -1 + 45x_i - 13x_i^2 + 0.5z_{1i} + 0.5z_{2i} + 0.5z_{3i}$，其中 z_{1i}、z_{2i} 以及 z_{3i} 可以描述一個具有參個緊密地間隔頂峰的狀態。只有一個在閉區間 $[0.9,1.1]$ 的頂峰可藉由非強穩的平滑樣條來預測，然而第 3 個頂峰藉由強穩的平滑樣條來正確地預測。以非強穩的方法所得到的第 8 個以及第 9 個觀察值(相對應於在閉區間 $[0.7,0.9]$ 計算的數字)以及第 12 個以及第 13 個觀察值(相對應於在閉區間 $[1.1,1.3]$ 計算的數字)的預測值顯著地小於或大於資料的數值。第 9 個、第 12 個以及第 13 個觀察值比較起其它的觀察值有較大的 S_i 值。對於資料，

$y_i = -1 + 45x_i - 13x_i^2 + 0.5z_{1i} + 0.5z_{2i} + 0.5z_{4i}$，非強穩以及強穩的平滑樣條配適都插入資料以及正確地預測頂峰，其中 z_{1i}、z_{2i} 以及 z_{4i} 可以描述一個具有兩個緊密地間隔頂峰以及一個個別的頂峰的狀態。上面的結果似乎意味著非強穩的平滑樣條方法可以在具有緊密地間隔頂峰存在之下不正確地配適資料。在頂峰附近有影響力的觀察值可能不能以非強穩的平滑樣條來正確地配適，然而一個更精確的結果可以藉由使用強穩平滑樣條而向下加權這些觀察值來得到。注意到具有強穩 GCV 估計值在上面以不同的加權函數為基礎的模擬之下，其配適曲線的行為幾乎是相同的。因為在核能以及粒子物理的領域上，分離緊密地間隔頂峰是個重要的問題，有效的診斷以及強穩的方法可能是有用的。



圖三

表示在二階多項式背景之下對應於兩個 Lorentzian 頂峰的模擬資料得到強穩以及非強穩的平滑樣條配適。實線對應於非強穩配適；而點線對應於強穩的配適。

第五章、討論

統計學家 Efron 曾提出”統計是最成功的資訊科學”，要如何能藉由統計而從資料中萃取出一些資訊，進而幫助我們更加瞭解資料並對問題作一些決策，這是值得我們好好思考的。比如，在本論文中主要針對”無母數迴歸分析”而得到有影響力觀察值之資訊。

如同在 3.2 節給定一個例子來解釋，辨別有影響力的觀察值在配適曲線的行為上可能有很大的影響以及對於資料不同的解釋而得到進一步結果。此外，強穩的方法藉由將有影響力的觀察值向下加權能夠提供一個替代性的配適，而通常似乎比起非強穩的方法要來得更穩定。

在迴歸診斷之中一個主要的問題就是單一有影響力觀察值的偽裝(或甚至是一群有影響力的觀察值)。因此，考慮在無母數迴歸設定類似是一個影響的診斷可能是必要的，在本論文有影響力觀察值其貢獻的顯著性已藉由粗糙的警告限制來評估。一般使用影響測量的分配理論還未被發展。在 3.3 與 3.4 節的方法能夠用來對於其它平滑參數估計的發展上(參考 Hurich , Simonoff ,and Tasi ,1998)。最後，不僅在核子與粒子物理而且在其它的領域上，緊密間隔頂峰的分開是一個重要的問題。影響診斷法以及強穩的方法可能有廣泛的應用。

參考文獻

中文部份：

1. 彭文正(1997)譯，資料採礦－顧客關係管理暨電子行銷之應用
2. 吳旭智、賴淑貞(2000)譯，資料採礦－顧客關係管理的技巧與科學
3. 蕭凱方(2001)，資料擷取與區別分析，東海大學統計學研究所碩士論文
4. 涂吟樺(2001)，商業智慧之核心－Data Mining(資料採礦)，台新銀行月刊 90 年 8 月號
5. 林傑斌、劉明德(2002)，資料採掘與 OLAP 理論與實務

英文部份：

1. Bevington, P.R. & Robinson, D. K. (1992),Data reduction and error analysis for the physical science,WCB McGraw-Hill.
2. Carroll, R. J. & Ruppert, D. (1988),Transformations and weighting in regression, New York: Chapman and Hall.
3. Golub, G., Heath, M., & Wahba, G. (1979),'Generalized cross validation as a method for choosing a good ridge parameter',Technometrics 21,215-224.
4. Green, P. J. & Silverman, B. W. (1994),Nonparametric regression and generalized linear models: a rough penalty approach, London: Chapman and Hall.
5. Hastie, T. J. & Tibshirani, R. J. (1990), Generalized additive models, New York: Chapman and Hall.
6. Hurvich, C. M., Simonoff, J. S. & Tasi, C. L. (1998),'Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion',Journal of Royal Statistical Society, Series B 60, 271-293.
7. Nychka, D. (1988), 'Bayesian confidence intervals for smoothing splines', Journal of the American Statistical Association 83, 1134-1143.

8. O'Sullivan, F., Yandell, B., &Raynor, W. (1986),'Automatic smoothing of regression function in generalized linear models', Journal of the American Statistical Association 81, 96-103.
9. Silverman, B. W. (1985),'Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion)', Journal of Royal Statistical Society, Series B 47, 1-52.
10. Simonoff, J. S. (1996), Smoothing methods in statistics, New York: Springer-Verlag.
11. Speed, T. P. (1991),'Comment on " That BLUP is a good thing: the estimation of random effects " ', Statistical Science 6, 42-44.
12. Thomas, W. (1991), 'Influence diagnostics for the cross-validated smoothing parameter in spline smoothing', Journal of the American Statistical Association 86, 693-698.
13. Wahba, G. (1983), 'Bayesian confidence intervals for the cross-validated smoothing spline', Journal of Royal Statistical Society, Series B 45, 133-150.
14. Wahba, G. (1990), Spline models for observational data, Philadelphia: Society for Industrial and Applied Mathematics.

附錄: 本論文所使用的程式

```
#
#          Computing  S_i  and deleone-one diagnostics
#
influence.spline<-function(x,y){
        len<-length(y)
        si<-rep(0,len)
        deleone<-rep(0,len)
        oriobj<-smooth.spline(x,y,all.knots=T)
      orispar<-oriobj$spar
     for(i in 1:len){
       delobj<-smooth.spline(x[-i],y[-i],all.knots=T)
       spari<-delobj$spar
       deleone[i]<-orispar-spari
     si[i]<-(((1/spari^(1.25))+(1/orispar^(1.25)))*(orispar-spari))^(2)
                                }
       derivlist<-list(si=si,deleone=deleone)
         invisible(derivlist)
}
```

```
#
#          Kernel functions used in robust smoothing
#
weight<-function(w){
    u <-median(w)
     sdv<-mad(w)
     sw<-(w-u)/sdv
     absw<-abs(sw)
     nweight<-exp((-0.5)*(absw^2))
     bweight<-ifelse(absw<=1,1,0)
     tweight<-ifelse(absw<=1,(1-absw),0)
     pweight<-ifelse(absw<=2,1-(3/16)*absw^(2),0)
     pweight<-ifelse((absw>2)&(absw<=3),-absw^(2)/4+1.25*absw-1.5,pweight)
     eweight<-ifelse(absw<=sqrt(5),(1-0.2*absw^2),0)
     biweight<-ifelse(absw<=1,(15/16)*(1-absw^(2))^(2),0)
     weightlist<-list(nw=nweight,bw=bweight,tw=tweight,pw=pweight,ew=eweight,biw=bi
     weight)
     invisible(weightlist)
     }
```

```
#
#          Robust GCV method
#
rgcv<-function(x,y,w){
     absw<-abs(w)
     lambda<-seq(1e-6,1e-3,by=1e-6)
     lenlamb<-length(lambda)
     rgcv<-rep(0,lenlamb)
     ratio<-sum(absw)/length(x)
     for(i in 1:lenlamb){
          spltem<-smooth.spline(x,y,w=absw,spar=lambda[i],all.knots=T)
          rgcv[i]<-weighted.mean((y-predict(spltem,x)$y)^2,absw)/
                         ((1-mean(spltem$lev)/ratio)^2)
                                        }
          index<-order(rgcv)
         robustgcv<-index[1]*(1e-6)
          if(robustgcv==1e-6){
          print("GCV again")
          rgcvtem<-rgcv1(x,y,w)
          robustgcv<-rgcvtem$robustgcv
          rgcv<-rgcvtem$rgcv
               }
```

```
    gcvlist<-list(robustgcv=robustgcv,rgcv=rgcv)
     invisible(gcvlist)
}
```