國立政治大學商學院統計研究所

碩士論文

傘型迴歸函數估計

Estimation of umbrella shaped regression function

指導教授：黃子銘　博士

研究生：林似蓉 撰

中華民 101 年 7 月

謝 誌

不經一番寒徹骨，焉得論文撲鼻香。歷經這段令人心力交瘁的日子後，終於完成此篇論文。此時除了感謝黃子銘老師平時耐心的教導外，還要謝謝黃貞瑛老師及鄭宗記老師兩位口委對於本拙作提供適當及切中的建議。

在此，特別感謝政大統研所的同學這兩年來的陪伴。跟你們一起打球、玩樂、上課以及一起打拼論文的時光總是過的特別快，希望以後還能夠一起嘴砲及出遊。另外，還要謝謝統計系的學弟妹，平時總是會聽我的抱怨並給我一些建議和溫暖。真的很想跟你們說：有你們真好!還有，感謝我的大學同學們，六年來，無論是在手機上或是 Facebook 上聊天，你們總是會關心我並替我加油打氣。特別是劉香吟，一起分享彼此的痛苦及笑容，無論是在寫論文或者是找工作，我們也會互相鼓勵對方。還有陳紓綺總是在半夜時陪伴我，讓我在夜間工作時感到不孤單。

最後，感謝我的家人在精神上以及經濟上的幫助，我才能夠如此順利的完成此篇論文。

林似蓉 謹誌於
國立政治大學統計研究所
中華民國一〇一年七月

摘 要

傘型迴歸函數是類似傘的形狀的迴歸函數，只要符合先上升後下降的趨勢皆爲傘型迴歸函數。無母數迴歸函數中最常見的方法之一是樣條(Splines)迴歸函數。樣條爲充分平滑分段多項式函數，而節點(knots)爲平滑多項式函數連接的地方。在本論文中，將節點以等距離擺放並以AIC(Akaike information criterion)值得到合理的節點數。用三種方法的樣條迴歸函數去估計傘型函數。第一種爲RSPL(restrictted spline regression)，也就是有形狀限制時的樣條迴歸函數。第二種是CSPL(concave spline regression)，是參考Meyer寫的樣條迴歸函數，此樣條迴歸函數爲凹函數(concave function)。最後一種則稱SPL(spline regression)，爲沒有形狀限制也不是凹函數的樣條函數。以IMSE爲評估標準，IMSE越小，則代表此方法估計的越好。由模擬結果，在估計先上升後下降的函數時，用RSPL的方法去估計會得到最小的IMSE；而在估計凹函數時，則是CSPL會得到最小的IMSE。利用RSPL和SPL兩個方法估計由中央氣象局蒐集最近13年(1998-2010)的月均溫資料並探討最近幾年的月均溫資料趨勢是否有改變。未來假如需要估計傘型函數時，則可利用本篇所述的方法去估計。

Abstract

In this thesis, we consider the problem of estimating a regression function assuming the regression function is unimodal. The proposed method is to model the regression function as linear combination of B-spline basis functions with equally spaced knots, and the number of knots is determined using AIC (Akaike information criterion). Specific constraints are placed on the coefficients of basis functions to ensure that estimated regression function is unimodal. The coefficients are estimated using least square method.

The proposed method is refered as RSPL and is compared with two other methods: SPL and CSPL, where SPL is similar to RSPL except that the coefficients of basis functions are estimated without any constraints, and CSPL gives concave regression function estimates. Simulation results show that RSPL outperforms SPL and CSPL when the true regression function is unimodal but not concave, and CSPL outperforms RSPL and SPL when the true regression function is concave. Also, RSPL is applied to temperature data to estimate temperature trend within one year.

目 錄

圖 目 錄

表 目 錄

1 緒論

在迴歸分析中，解釋變數X和反應變數Y之間的關係可以用一個函數f來解釋，

$$Y \approx f(X).$$

在某一些應用上，會假設此迴歸函數f有形狀限制，例如：單調性(monotonic)、凸性(convexity)。在本論文中，考慮的是估計f爲單峰(unimodal)迴歸函數。單峰迴歸函數，也是大家所熟知的傘型迴歸函數，定義如下：現在令f的範圍介在一個區間$I = [a, b]$。假如現在存在一個數字m在此區間(I)內，在$[a, m]$之間，f爲非遞減函數(nondecreasing function)；而在$[m, b]$的區間內，f則是非遞增函數(nonincreasing function)，則f稱爲傘型迴歸函數。

傘型迴歸函數可以應用在許多方面。在生物方面，隨著時間不同，藥物濃度一開始會先上升，在達到最大值之後又下降。而在經濟方面，Lipsay & Steiner(1972) 書中也有許多例子。比較重要的幾個，例如：邊際生產力和勞工數的關係以及劣質商品的需求和家庭收支的關係。此外，有隨著時間變動的經濟指標單峰迴歸函數也可以用來估計經濟衰退和經濟趨向之間變化的時間點。圖1.1爲邊際生產力和勞工數的關係。

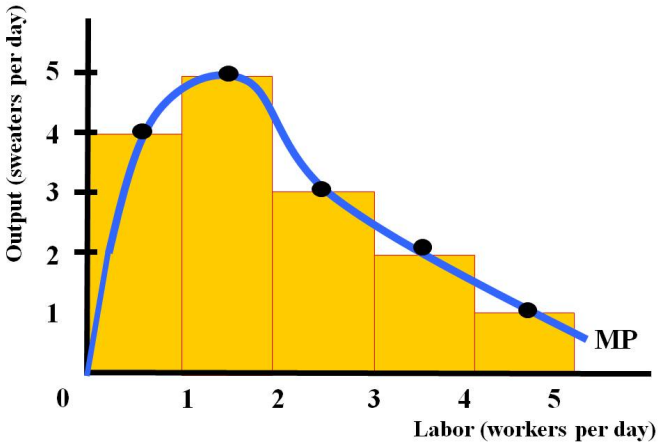

圖 1.1: 邊際生產力和勞工數的關係

在無母數迴歸[2]中常見的估計方法有：樣條迴歸(regression spline)、平滑樣條(spline smoothing)和核估計(kernel estimation)。因爲樣條函數具有好的近似能力且具有平滑性，故本篇是以B-樣條(B-splines)迴歸函數去近似迴歸函數f。樣條是於1946年首次由Schoenberg提出。在數學上，樣條爲平滑分段多項式函數(piecewise polynomial function)，而節點(knots)爲平滑多項式函數連接的地方。B-樣條函數爲B-樣條基底函數的線性組合。假如對B-樣條基底函數採用合適的係數限制，有時可使組成的的B-樣條函數滿足特定形狀限制。在本論文中，是採用特殊型式的係數去保證組合成的B-樣條函數爲單峰。

在第二章及第三章會介紹不同的基底函數。利用這些基底函數的線性組合去近似迴歸函數，以得到特定形狀的迴歸函數。

2 文獻回顧

在這一章，會介紹文獻上樣條迴歸中節點個數的選取方法，以及有形狀限制時的迴歸函數估計。

2.1 節點個數選取

Meyer (2008)[5]中提到，樣條迴歸對於節點(knots)的個數以及擺放位置是敏感的。因爲節點個數會影響無母數迴歸估計的平滑程度。假如個數越多，分段多項式函數的個數也會增加，使得誤差變小。但是個數太多，則會有過度配適(overfitting)的問題。所以在此篇文章中，如何選擇節點及節點的個數是重要的議題。Keele (2008)[7] 提到，在相同條件的樣條迴歸函數下，最小的AIC(Akaike)[1] 值提供一個最合適且節點數最少的衡量方法。此外，Miyata和Shen (2005) 提到，BIC(Schwarz)[10]也是一種選擇節點個數的方法。Osborne ,Presnell ,Turlach三位學者則提出用LASSO[7]的方法來選擇節點及節點個數。下面會詳述這三種方法。

首先介紹AIC(Akaike information criterion)：AIC是一種衡量統計模型配適優良性的一種標準。$AIC = 2k - 2\ln(L)$，k代表模型的參數個數，而L則代表模型的概似函數(likelihood function)。假設條件是模型的誤差服從獨立常態分佈。令n爲觀察值個數，RSS(residual sum of squares)爲剩餘平方和，則$AIC = 2k + n\ln(\frac{RSS}{n})$。增加自由參數個數是爲了提高配適優良性，但是要避免過度配適。因爲AIC是一種尋找可以最好解釋但包含最少自由參數模型的方法，所以AIC最小的模型是我們優先考量的模型。

Shibata(1976)證明AIC準則對模型參數個數會產生高估的現象，故Schwarz發展一套利用貝氏方法得到最小AIC過程的標準，稱爲BIC(Bayesian information criterion)。BIC的定義爲：$BIC = -2\ln(L) + k\ln(n)$，k代表模型的參數個數，n爲觀察值個數，L則代表模型的概似函數。與AIC最大的差異在懲罰項不同，目的就是避免過度配適。BIC同樣也是衡量統計模

型配適優良性的一種標準，所以BIC最小的模型是我們優先考慮的模型。

最後介紹LASSO(Tibshirani,1996) [12]方法：LASSO(least absolute shrinkage and selection operator)是一種對線性迴歸的特徵值縮減和變數選擇的方法。假設現在有一組資料$\langle x_i, y_i\rangle$，$i = 1, \ldots, n$，其中，x_i是單變量且將x_i遞增排序。而將設計矩陣(X)建構成

$$(x - x_2)_+^p, (x - x_3)_+^p, \ldots, (x - x_{n-1})_+^p, 1_n, (x - x_1)^1, \ldots, (x - x_1)^p.$$

此時設計矩陣爲一個 $n \times (n + p - 1)$矩陣，n代表基底數，p則代表樣條迴歸函數的次數。其限制式爲

$$\underset{\beta \in \Re^m}{\text{minimize}} \quad f(\beta) = \frac{1}{2}\,(y - X\beta)^T (y - X\beta)$$

$$\text{subject to} \quad g(\beta) = t - \|\beta\|_1 \geq 0$$

接著，用 M. R. Osborne , B. Presnell , B. A. Turlach三位學者所寫的演算法得到LASSO的估計值(β)。在此，$\beta \neq 0$的個數就代表選取後的節點個數，會比原本設定的節點個數少。這些選出來的節點，可以避免過度配適的問題。所以，這也是一種選擇節點個數的方法。

2.2 有形狀限制時的迴歸函數估計

此節要介紹有形狀限制時的迴歸函數估計，形狀限制考慮單調迴歸函數及凸迴歸函數兩種。即考慮對一組二元資料(x_i, y_i)配適迴歸模型

$$y_i \approx f(x_i).$$

而f爲單調函數或凸函數。以下提到的估計方法中，均假設f可表示成一些基底函數的線性組合，而以最小平方法在適當係數限制下估計基底對應的係數。當f爲單調函數時，Ramsay(1988)[8]提出單調樣條函數(monotone spline function)作爲基底函數以組成f。令$a = min(x_1, \ldots, x_n)$，$b = max(x_1, \ldots, x_n)$，則對應格點$u_1, \ldots, u_l \in (a, b)$的$k$次單調樣條基底函數$M_i^{(k)}$定義如下：令$t_1 = \ldots = t_k = a$，$(t_{k+1}, \ldots, t_{l+k}) = (u_1, \ldots, u_l)$，$t_{l+k+1} = \ldots = t_{l+2k} = b$.則一次單調樣條迴歸函數是階梯函數

$$M_i^{(1)}(x) = \begin{cases} \frac{1}{t_{i+1} - t_i}, & \text{for } t_i \leq x \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, l+1$。 而k次單調樣條迴歸函數則是

$$M_i^{(k)}(x) = \begin{cases} \frac{k[(x-t_i)M_i^{(k-1)}(x)+(t_{i+k}-x)M_{i+1}^{(k-1)}(x)]}{(k-1)(t_{i+1}-t_i)}, & \text{for } t_i \le x \le t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

將$M_i^{(k)}$, $i = 1, \ldots, l+1$線性組合成f時，係數要≥ 0，組合出的函數才有單調性。 再將$M_i^{(k)}$積分得到$I_i^{(k)}$:

$$I_i^{(k)}(x) = \int_{t_1}^{x} M_i^{(k)}(u)du \text{ for } i = 1, \ldots, l+k = m, \text{ for } x \in [x_1, x_n].$$

假設現在存在一組f爲$f = \Sigma a_i I_i$，且f介於$[0,1]$之間，並令t_i爲節點序列。此時要求$a_i \geq 0$且$\Sigma a_i = 1$，可解釋成$f(1) = 1$。

當f爲凸函數(convex function)時，Meyer[5]提出使用。凸樣條函數基底來組成f。基底形式爲

$$C_i^{(k)}(x) = \int_{t_1}^{x} I_i^{(k)}(u)du \text{ for } i = 1, \ldots, l+k = m, \text{ for } x \in [x_1, x_n].$$

使用$C_i^{(k)}$基底線性組合成f時，基底係數要≥ 0，再加上一組常數函數以及特定的函數$g(x) = x$去做線性組合，組合出來的f才會是凸函數。然而，在本論文中，欲估計的迴歸函數爲凹函數，所以將基底函數取負號，也就是將凸函數改成凹函數。由於這個轉換過的基底函數爲凹函數，則可利用此基底函數進行估計。將此方法命名爲CSPL(concave spline regression)。

3 研究方法

本章最主要是說明此論文所使用的研究方法。如同前一章所述，樣條函數爲本論文最主要使用的方法。在此，詳細的描述樣條函數，主要參考Spline Functions: Basic Theory[9]這本書。並詳述如何應用樣條函數去估計傘型函數。

3.1 選取節點個數

第二章有提到，如何選擇節點及節點的個數在本篇是重要的問題。因爲AIC及BIC較常見，故先考慮AIC及BIC的方法。以同樣的函數f得到AIC及BIC的節點個數，將得到的節點個數代回估計的函數，比較AIC及BIC的估計哪個較佳。將模擬結果繪製成圖3.1，此圖爲BIC的IMSE與AIC的IMSE相減(BIC-AIC)得到的直方圖。由於圖形不好判斷哪一個較好，故將結果加總，發現爲負值，代表AIC的IMSE較小。也就是AIC提供的個數較合適，故本論文以AIC值去得到合理的節點數。

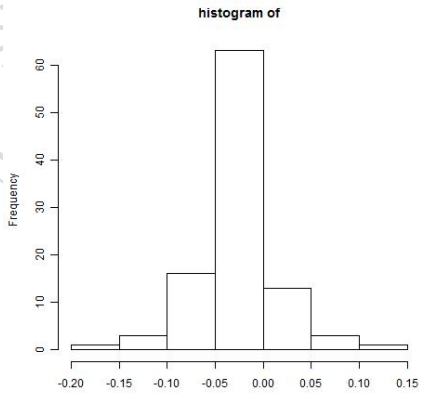

圖 3.1: (BIC-AIC)的IMSE差異直方圖

3.2 B-樣條函數

一開始，先定義m階B-樣條(B-splines)函數。首先說明一些符號：令$\{u_i\}_1^m$爲一組定義在I集合的函數，且設$t_1,\ldots,t_m$爲在I集合上的點，並令$t_1 < t_2 < \ldots < t_m$。接著，定義與$\{u_i\}_1^m$及$\{t_i\}_1^m$有關的矩陣：

$$M\begin{pmatrix} t_1,\ldots,t_m \\ u_1,\ldots,u_m \end{pmatrix} = \begin{bmatrix} u_1(t_1) & u_2(t_1) & \ldots & u_m(t_1) \\ u_1(t_2) & u_2(t_2) & \ldots & u_m(t_2) \\ \ldots & & & \\ u_1(t_m) & u_2(t_m) & \ldots & u_m(t_m) \end{bmatrix}$$

定義D矩陣如下：

$$D\begin{pmatrix} t_1,\ldots,t_m \\ u_1,\ldots,u_m \end{pmatrix} = detM\begin{pmatrix} t_1,\ldots,t_m \\ u_1,\ldots,u_m \end{pmatrix}$$

接著，均差(divided difference)的定義爲：在給定$t_1,\ldots,t_{r+1}$這些點並假設 t爲遞增排序及函數f，其r階均差爲

$$[t_1,\ldots,t_{r+1}]f = \frac{D\begin{pmatrix} t_1,\ldots,t_{r+1} \\ 1,x,\ldots,x^{r-1},f \end{pmatrix}}{D\begin{pmatrix} t_1,\ldots,t_{r+1} \\ 1,x,\ldots,x^{r} \end{pmatrix}}$$

定義B-樣條函數：令$\ldots \le y_{-1} \le y_0 \le y_1 \le y_2 \le \ldots$爲一實數序列，在給定整數$i$及$m > 0$及對所有的$x$，令$Q_i^{(m)}$爲：

$$Q_i^{(m)}(x) = \begin{cases} (-1)^m[y_i,\ldots,y_{i+m}](x-y)_+^{(m-1)}, & \text{if } y_i \le x < y_{i+m} \\ 0, & \text{otherwise} \end{cases}$$

其中，$Q_i^{(m)}$爲第m階及節點爲$y_i,\ldots,y_{i+m}$的樣條迴歸函數。而B-樣條函數爲：

$$N_i^{(m)}(x) = (y_{i+m} - y_i)Q_i^{(m)}(x)$$

其中，$N_i^{(m)}$是節點爲$y_i,\ldots,y_{i+m}$的B-樣條迴歸函數。假如現在爲一次B-樣條迴歸函數($m = 1$)且$y_i < y_{i+1}$，則

$$N_i^{(1)}(x) = \begin{cases} 1, & \text{for } y_i \le x < y_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

而根據[9]書中定理4.9，B-樣條函數的單位分割形式爲

$$\sum_{i=j+1-m}^{j} N_i^{(m)}(x) = 1, \tag{3.1}$$

對所有的$y_j \le x < y_{j+1}$。故對所有$m \ge 1$和所有的$x \in \Re$,

$$0 \le N_i^{(m)}(x) \le 1.$$

現在考慮$[a,b]$爲一個有限閉區間，而$x_1 < \ldots < x_k$爲$[a,b]$個子集合，且這些子集合是$[a,b]$中的k個點。令

$$y_1 = \ldots = y_m = a,$$

$$b = y_{m+k+1} = \ldots = y_{2m+k},$$

$$(y_{m+1}, \ldots, y_{m+k}) = (x_1, \ldots, x_k).$$

則$N_i^{(m)}$，$i = 1, \ldots, m+k$形成$[a,b]$上的一組B-樣條基底函數，而內部節點爲$(x_1, \ldots, x_k)$。在本論文中，將$[a,b]$設爲$[0,1]$，且內部節點爲等距離擺放，也就是$y_{i+1} - y_i$爲一個定值，$i = m, \ldots, m+k+1$。且$k+1 \ge m$，則 $N_i^{(m)}$，$i = m, \ldots, k+1$。代表在這裡均勻分布的節點，會使得B-樣條基底函數曲線形狀相同$(i = m, \ldots, k+1)$，且僅在x軸上平移一個節點的增量值。

假如次數爲0，則這些基底函數皆爲階梯函數(step function)。換句話說，基底函數$N_i^{(1)}(x) = 1$，對第i次節點長度$[y_i, y_{i+1})$。舉例來說，現在有四個節點，分別是$y_0 = 0$、$y_1 = 1$、$y_2 = 2$和$y_3 = 3$。節點長度則爲$[0,1)$、$[1,2)$和$[2,3)$。B-樣條迴歸函數是基底函數(basis function)的線性組合，故不同的基底函數會產生不同的B-樣條迴歸函數。而基底函數的形狀則是由節點的位置去決定且具有局部控制(local control)的能力。基底函數爲

$$N_i^{(1)}(x) = \begin{cases} 1, & \text{for } y_i \le x < y_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

其中，$i = 0, 1, 2$。換句話說，基底函數皆> 0。以下面的圖來看會比較清楚:

3.3 RSPL

在本論文中，將自己寫的方法命名爲RSPL(restricted spline regression)，而在B-樣條迴歸函數爲沒有限制的情況下，稱爲SPL(spline regression)。最主要的事情是去進行傘型迴

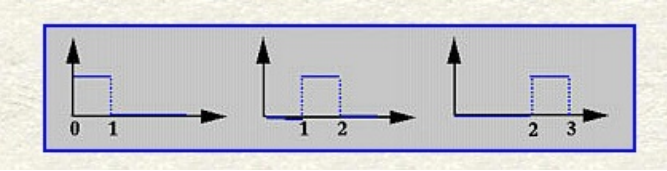

圖 3.2: 基底函數

歸函數的估計，所以接下來要研究如何利用B-樣條迴歸函數做估計。由[9]書中定理5.9:
令$s=\sum_{i=1}^{m+k} c_i N_i^{(m)}$和假設$1 \le d \le m$，對所有$y_m \le x < y_{m+k}$，

$$D_+^{d-1} s(x) = \sum_{i=d}^{m+k} c_i^{(d)} N_i^{(m-d+1)}(x)$$

，其中，$c_i^{(1)} = c_i, i = 1, 2, \ldots, m+k$。

$$c_i^{(j)} = \begin{cases} (m-j+1)\frac{c_i^{(j-1)} - c_{i-1}^{(j-1)}}{y_{i+m-j+1} - y_i}, & \text{if } (y_{i+m-j+1} - y_i) > 0 \\ 0, & \text{otherwise} \end{cases}$$

，對$i = j, \ldots, m+k$和$j = 2, 3, \ldots, d$。將一次導數代入，也就是將$d=2$代入，可以得到

$$D_+ s(x) = (m-1) \sum_{i=2}^{m+k} c_i^{(2)} N_i^{(m-1)}(x). \tag{3.3}$$

其中，D_+代表微分後取右極限。在此，D_+的意思爲當係數遞增時，迴歸函數也爲遞增。

考慮

$$s(x) = \sum_{i=1}^{m+k} c_i N_i^{(m)}(x)$$

且係數滿足$c_1 \le \ldots \le c_{i^*}$以及$c_{i^*} \ge \ldots \ge c_{m+k}$的情況。由(3.3)及$N_i^{(m)}$在$[y_i, y_{i+m}]$外爲0的特性，

$$D_+ s(x) = \begin{cases} (m-1)\sum_{i=2}^{i*} c_i^{(2)} N_i^{(m-1)}(x), & \text{if } x \in [a, y_{i^*+1}] \\ (m-1)\sum_{i=i^*+1}^{m+k} c_i^{(2)} N_i^{(m-1)}(x), & \text{if } x \in [y_{i^*+m-1}, b] \end{cases} \tag{3.4}$$

由(3.2)和(3.4)以及

$$c_i - c_{i-1} = \begin{cases} \ge 0, & i = 2, \ldots, i^* \\ \le 0, & i = i^*+1, \ldots, m+k \end{cases} \tag{3.5}$$

可知$s(x)$在$[a, y_{i^*+1}]$上爲遞增；而在$[y_{i^*+m-1}, b]$上爲遞減。然而，在(y_{i^*+1}, y_{i^*+m-1})上，雖然不能證明$s(x)$的遞增遞減情況，但測試了許多滿足(3.5)的c_i後發現，當$m = 3$及$k = 2$時，$s(x)$均爲傘型。

4 模擬結果與實證分析

這一章最主要是要說明模擬與實證分析。會分成三節，前面兩節與模擬結果有關，最後一節則是說明實證資料分析及結果。

4.1 模擬結果1

先定義三種方法：第一種爲第三章所提的方法，也就是RSPL；第二種則是在第二章提起，是將Meyer的方法改寫，稱爲CSPL法；第三種是沒有限制條件下的方法，以SPL稱之。評估標準爲IMSE:

$$\int (\hat{f} - f)^2.$$

其中，f爲母體迴歸函數，而$\hat{f}$則是估計母體迴歸模型。將積分範圍定在$[0,1]$。IMSE越小，代表$\hat{f}$與f越接近，也就是估計的越佳。現在，先生成一條符合先上升後下降的函數當爲基底函數，將此函數加上誤差項去得到100條符合先上升後下降的函數，分別用這三種方法去估計，而同一條函數再重複估計100次。此時，在同一條函數下，各方法都會有100個IMSE，將這100個IMSE取平均，再去比較三種方法的平均IMSE孰大孰小，較小的平均IMSE代表用此方法得到的估計值較佳。以SPL爲基準，將RSPL與SPL的差距(在圖4.1上稱爲h1)和CSPL與SPL的差距(以h2稱之)繪製成兩張直方圖。從圖4.1可以看出，兩者相減都是負的，代表RSPL的IMSE較SPL的IMSE小，也就是RSPL估計的比SPL好。而圖4.2則顯示兩者之間的差距皆不小於0，與上圖結果相反，CSPL的IMSE較SPL的IMSE大，則是說明以SPL得到的估計值較CSPL的估計值佳。從上述模擬結果可看出，在估計先上升後下降的函數時，以RSPL的方法去估計會得到最小的IMSE。

4.2 模擬結果2

上述結果雖滿意，但因爲CSPL法的限制條件較強，故將先上升後下降的函數考慮改成凹函數(concave function)。與上一節的做法相同，先產生一條具有凸性的基底函數，再加上誤差項去產生100條函數，將這些函數做一次微分和二次微分。設一次微分爲0($f' = 0$)且二次微分函數值爲負($f'' < 0$)，再積分以確保具有凸性。以三種方法估計，再去針對每一條函數重複估計100次，將同一條函數的IMSE得到平均並做比較。同樣，以SPL爲基準，將RSPL與SPL的差距(在圖上改稱h3)和CSPL與SPL的差距(令爲h4)以直方圖的方式呈現。 從h3的直方圖(圖4.3)可以發現，此時RSPL與SPL的差距皆爲負值，代表RSPL估計的比SPL好。再來看h4的直方圖(圖4.4)，發現CSPL與SPL的差距也都是負的，所以接著去畫RSPL與CSPL差距(以h5稱之)的直方圖。從圖4.5可以看出，RSPL與CSPL的差距爲正的，也就是RSPL的IMSE較CSPL的IMSE大，故此時CSPL估計的較RSPL好。由以上模擬結果，需要估計具有凸性的函數時，優先考慮CSPL。

4.3 實證分析

本論文中考慮使用傘型迴歸函數以分析氣溫資料。根據中央氣象局的介紹：臺灣之氣候，一般被稱爲副熱帶或海洋候區。而且由於台灣的地理位置特殊，是在亞熱帶地區，一年四季溫度適宜。冬季受大陸冷氣團及東北季風之影響，南部乾燥北部濕冷，當寒潮爆發時全省均有可能遭受低溫災害之機會。冬季通常以1月下旬至2月中感覺最冷。南部較接近熱帶氣候，日照充足，冬天及夏天的溫度變化比北部來得小，也就是說北部地區的最高氣溫與最低氣溫的相差比較大，南部地區一年四季氣溫的變化比較小。現在設x軸爲月份，y軸爲月平均溫度，把這些點連在一起，就會是一條傘型函數。蒐集從1998年到2011年的資料，估計每一年的傘型迴歸函數，並分析是否有氣候暖化及氣候異常的現象。除了年資料外，還可以檢查不同地區是否有差異。現在有六個地區，分別是台北、淡水、新竹、台中、高雄及花蓮。按照緯度高低，預期淡水應是所有地區裡最低溫，高雄則是最高溫。

在本節，分析由中央氣象局蒐集最近13年(1998-2010)的月均溫資料，且以RSPL估計一年中的月均溫資料趨勢。爲了有足夠資料進行估計並了解各地月均溫趨勢最近幾年是否有改變，故針對每個地區將前五年(1998-2002)的資料合併，並與後五年的資料(2007-2011)去做比較。在月均溫資料並不是凹函數，故不考慮CSPL方法，僅用RSPL與SPL兩種方法。實線

爲RSPL，而虛線則是代表SPL。由於在寫程式時，將x軸的定義域定在$[0,1]$，在此，也將月份經過轉換，0爲一月，而1爲十二月。分別去比較台北、高雄、淡水、新竹、台中及花蓮這六個地區前五年與後五年的差異。估計結果顯示不同地區的月均溫有所差異。淡水氣溫明顯偏低，而高雄則是裡面最熱的地區，符合預期。從圖4.6到圖4.17，可以看出，後五年兩個方法得到的估計線較爲相近，而前五年的則不太相同。由於希望估計線看起來較平滑，故在此認爲RSPL法較SPL法的估計線平滑。

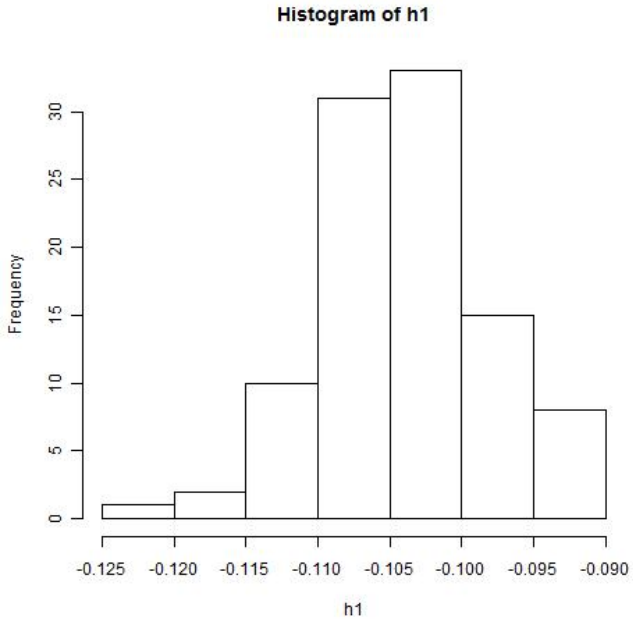

圖 4.1: (RSPL-SPL)的IMSE差異直方圖

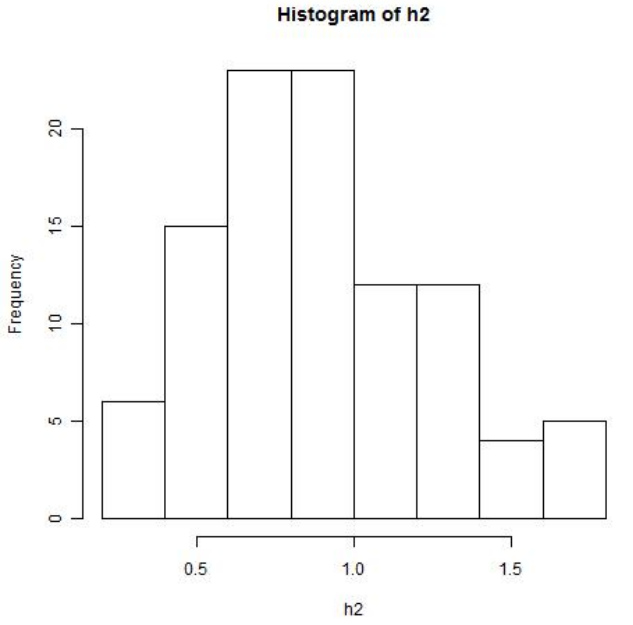

圖 4.2: (CSPL-SPL)的IMSE差異直方圖

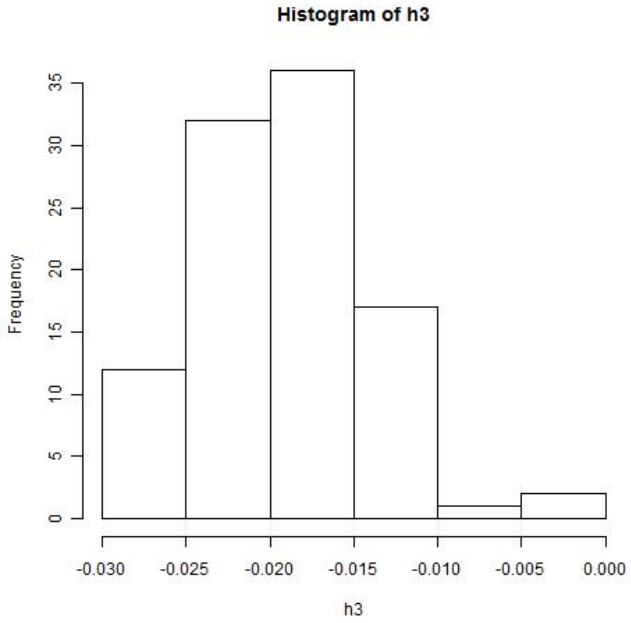

圖 4.3: (RSPL-SPL)在concave function下的IMSE差異直方圖

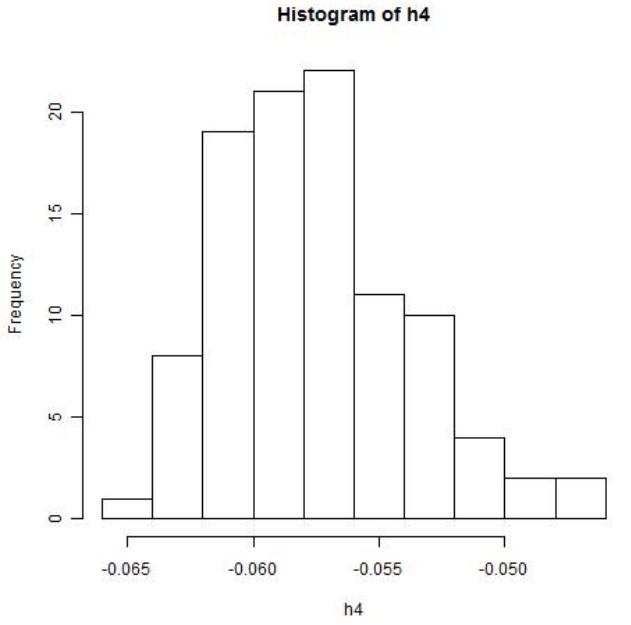

圖 4.4: (CSPL-SPL)在concave function下的IMSE差異直方圖

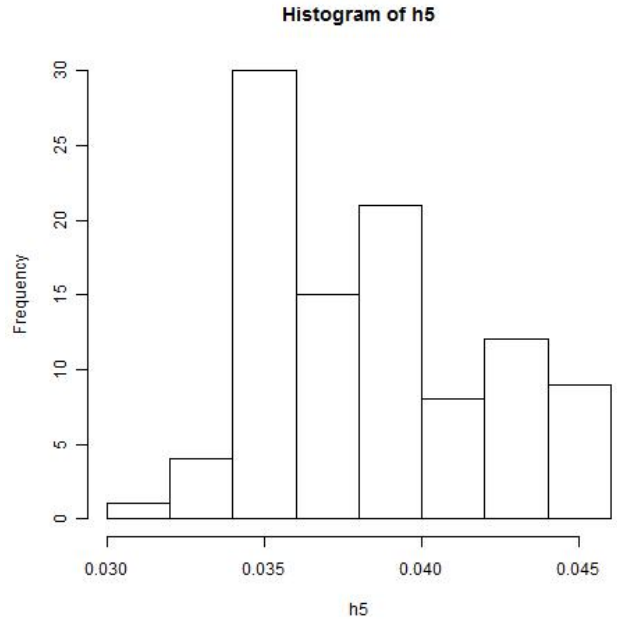

圖 4.5: (RSPL-CSPL)在concave function下IMSE差異直方圖

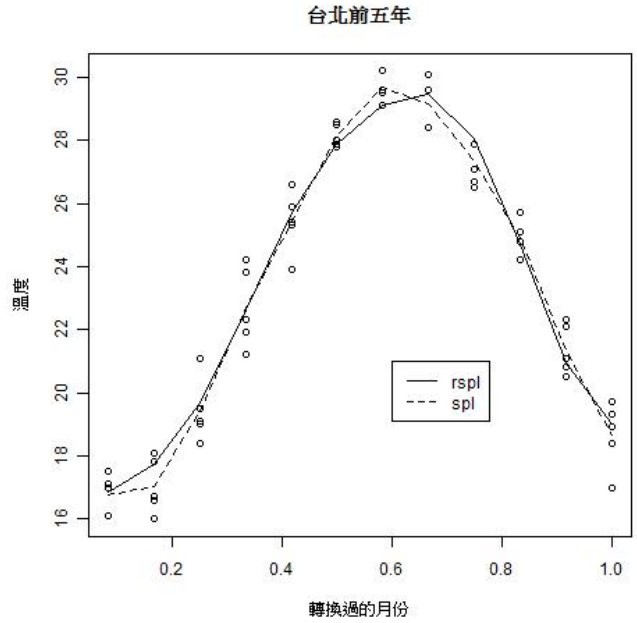

圖 4.6: 1998-2002年台北月均溫估計圖

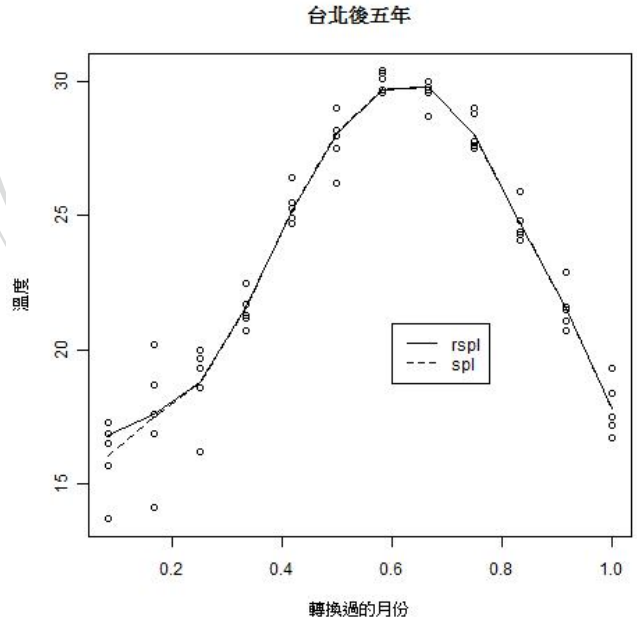

圖 4.7: 2007-2011年台北月均溫估計圖

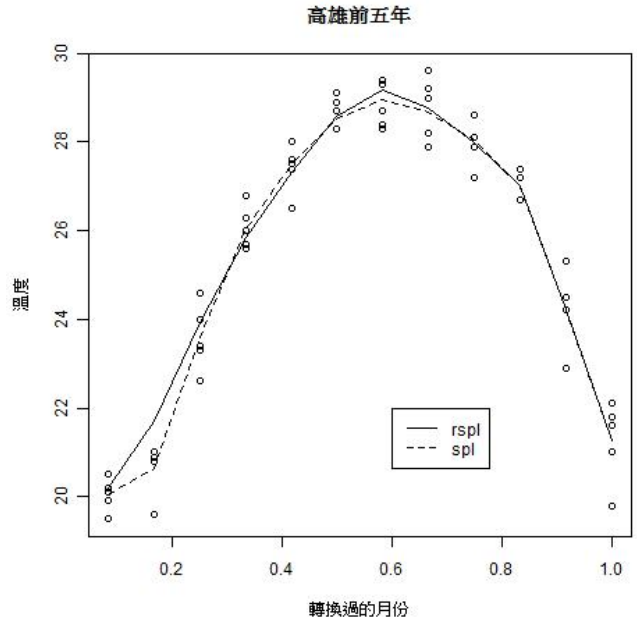

圖 4.8: 1998-2002年高雄月均溫估計圖

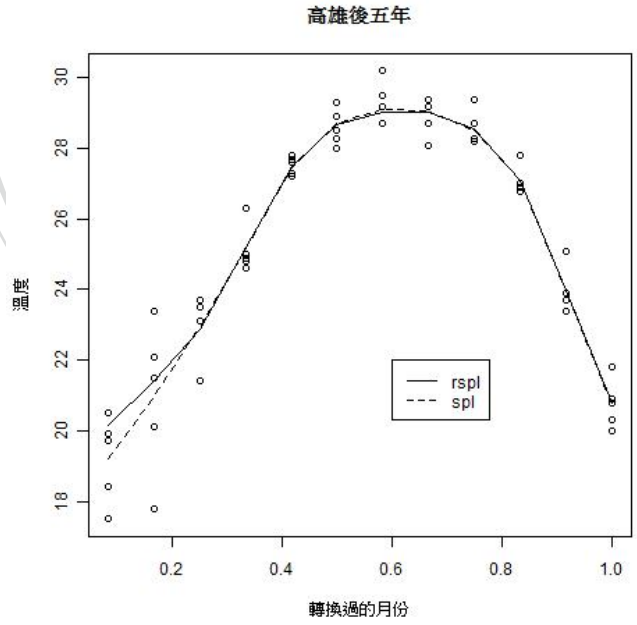

圖 4.9: 2007-2011年高雄月均溫估計圖

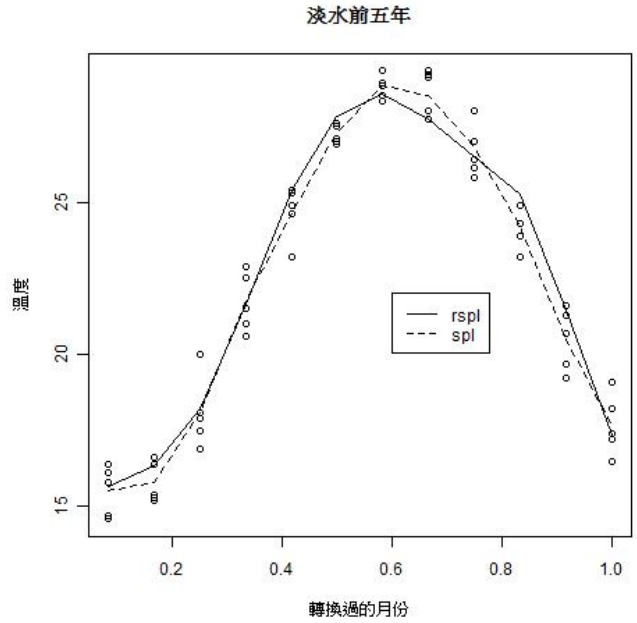

圖 4.10: 1998-2002年淡水月均溫估計圖

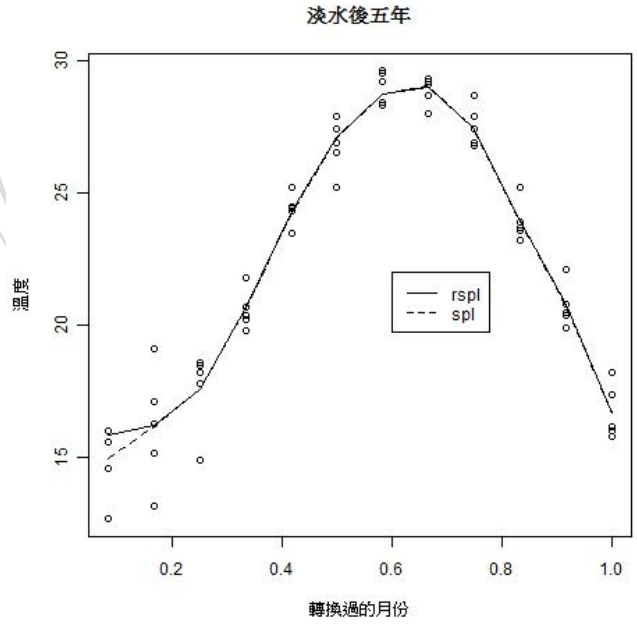

圖 4.11: 2007-2011年淡水月均溫估計圖

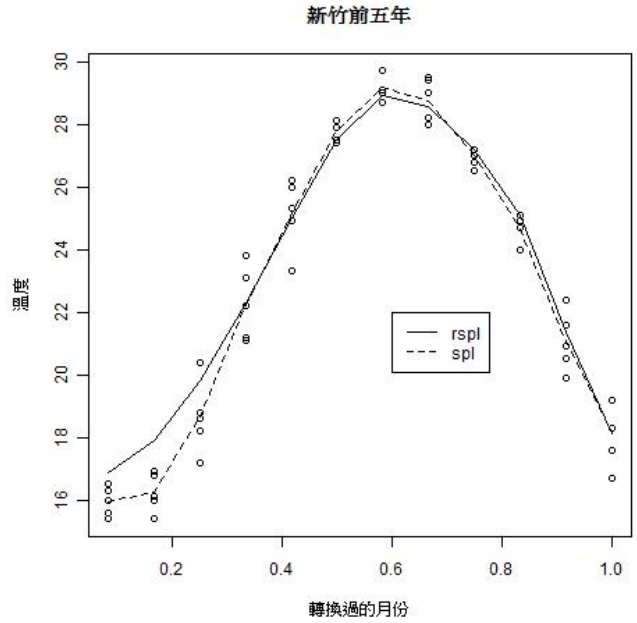

圖 4.12: 1998-2002年新竹月均溫估計圖

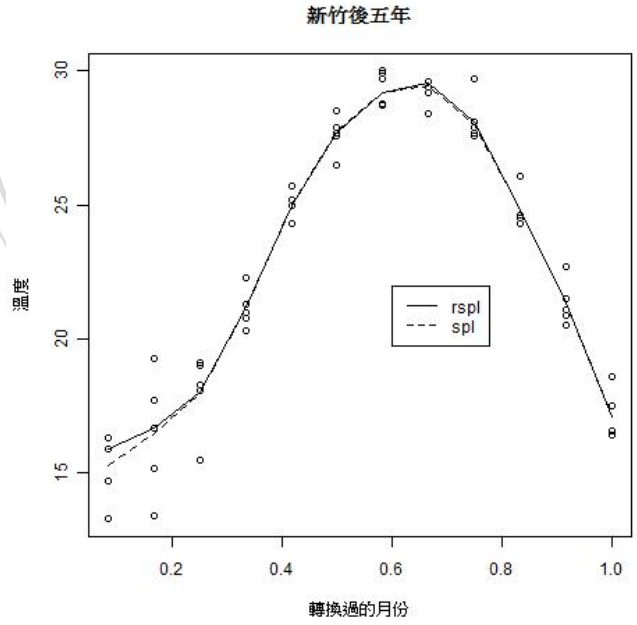

圖 4.13: 2007-2011年新竹月均溫估計圖

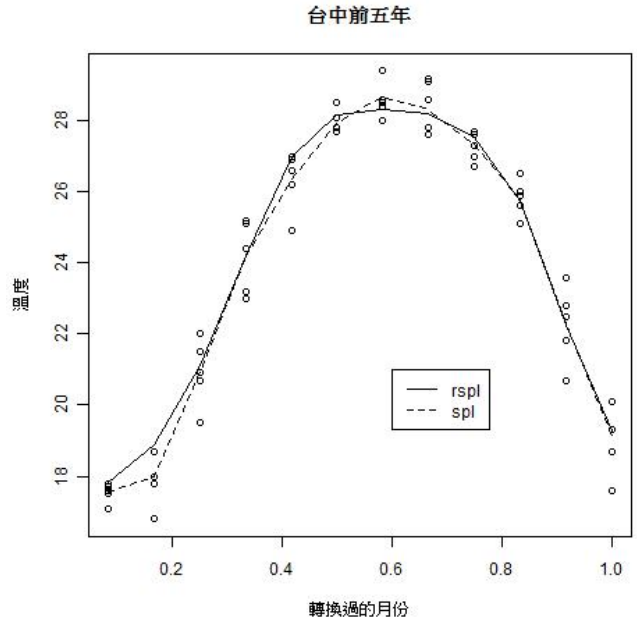

圖 4.14: 1998-2002年台中月均溫估計圖

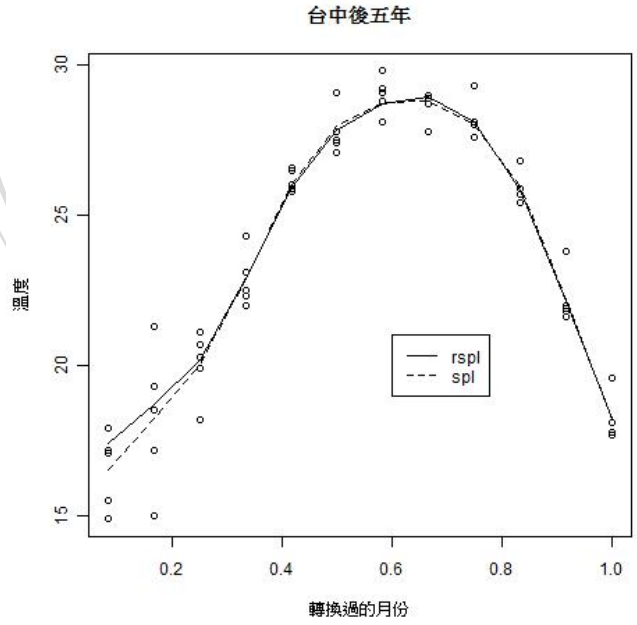

圖 4.15: 2007-2011年台中月均溫估計圖

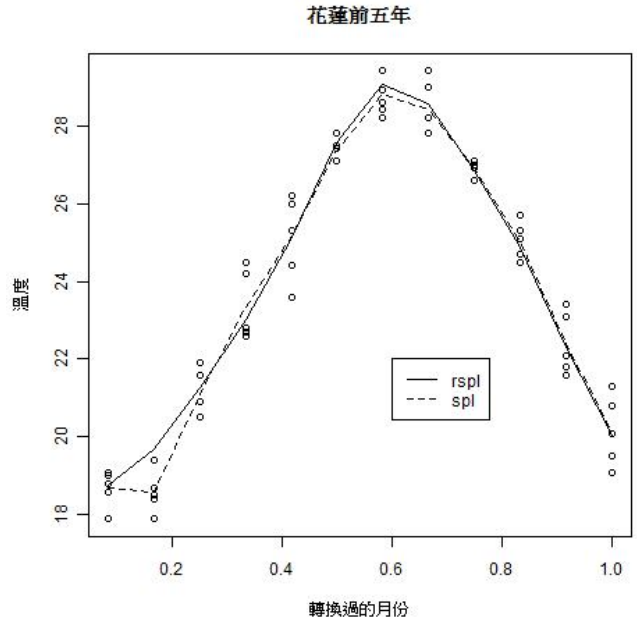

圖 4.16: 1998-2002年花蓮月均溫估計圖

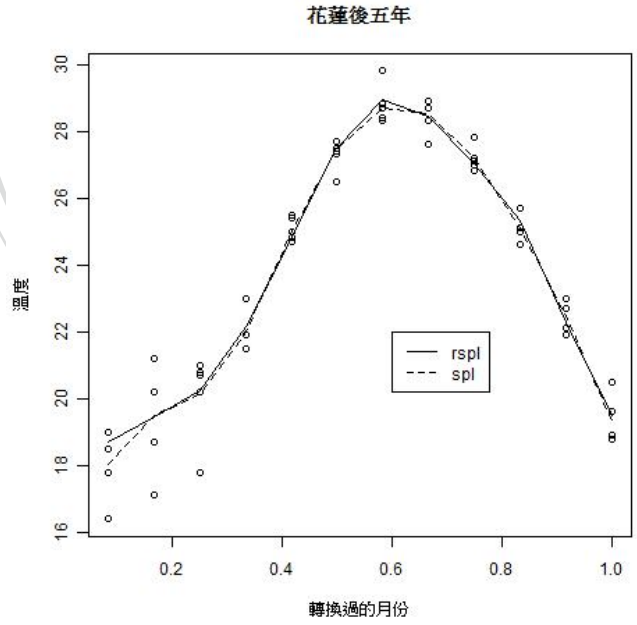

圖 4.17: 2007-2011年花蓮月均溫估計圖

5 結論與建議

本論文中提出以RSPL方法估計傘型迴歸函數，並和CSPL、SPL二種方法估計結果進行比較。根據模擬結果，如果迴歸函數爲凹函數，CSPL法會得到較佳的結果。如果傘型迴歸函數不爲凹函數，只是符合先上升後下降的形狀，則RSPL法效果較佳。至於在節點選擇上，本論文中使用等距節點並以AIC決定節點數。未來可考慮使用不同的放置節點方式再進行RSPL法估計。

本論文欲探討最近13年(1998-2011)的月均溫資料趨勢。因爲月均溫資料不屬於凹函數，故不考慮CSPL法，僅利用RSPL與SPL兩方法去進行估計。發現在前五年，這兩種方法估計的不太一樣，而在後五年，這兩種方法得到的估計線較爲相似。推測，後五年的月均溫表現較爲異常，冬天的時候變的很冷，而在夏天時溫度則不斷飆高，較符合傘型迴歸函數的定義，故估計的較爲相似。然而無法從圖形推測氣候暖化的現象。

除了月均溫外，從中央氣象局的氣候資料發現日照時數也是呈現傘型函數。假如未來還有類似的例子，則可以用本篇所述的方法去估計傘型迴歸函數。

參 考 文 獻

[1] H. Akaike. A new look at the statistical model identification. *Institute of Statistical Mathematics, Minato-ku, Tokyo, Japan*, 19 , Issue: 6:716– 723, 1974.

[2] Wolfgang Härdle. *Applied nonparametric regression.* Cambridge University Press, 1990.

[3] Luke Keele. *Semiparametric Regression for the Social Sciences.* Wiley, Chichester, UK, 2008. ISBN 978-0470319918.

[4] E. Mammen and C. Thomas-agnan. Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, 26:239–252, 1998.

[5] Mary C. Meyer. Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, 2(3):1013–1033, 2008.

[6] Satoshi Miyata and Xiaotong Shen. Free-knot splines and adaptive knot selection. *J. Japan Statist. Soc.*, Vol. 35 No. 2:303—324, 2005.

[7] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. Knot selection for regression splines via the LASSO. In *Computing Science and Statistics. Dimension Reduction, Computational Complexity and Information. Proceedings of the 30th Symposium on the Interface*, pages 44–49, 1998.

[8] J. O. Ramsay. Monotone regression splines in action (C/R: p442-461). *Statistical Science*, 3:425–441, 1988.

[9] Larry L. Schumaker. *Spline Functions: Basic Theory.* Cambridge University Press, 2007.

[10] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[11] E. V. Shikin and Alexander I. Plis. *Handbook on Splines for the User*. CRC Press, 1995.

[12] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.