國立政治大學統計學研究所

碩士學位論文

交叉驗證用於迴歸樣條的模型選擇之探討

指導教授 ： 黃子銘 博士

研究生 ： 謝式斌 撰

中華民國 一百零七 年 一 月

謝誌

首先真的非常謝謝這一年來黃子銘老師的指導，在很多不理解的問題上，老師都非常有耐心的教導我，讓我從最初摸索的階段，能夠很快的進入狀況，從中也明白自律的重要性。還要謝謝口試委員，陳素雲教授和翁久幸教授對於本篇論文的建議。 另外也要感謝研究所的同學，無論是在課業或是各種外在壓力下能夠一起挺過來，我真的很感激。

最後要感謝我家人的支持，獨自一人上來北部念書也讓家裡多了些負擔，現在我也完成學業了，希望能夠與你們分享這份喜悅。

摘要

在無母數的迴歸當中，因爲原始的函數類型未知，所以常用已知特定類型的函數來近似未知的函數，而spline函數也可以用來近似未知的函數，但是要估計spline函數就需要設定節點(knots)，越多的節點越能準確近似原始函數的內容，可是如果節點太多有較多的參數要估計, 就會變得比較不準確，所以選擇適合節點個數就變得很重要。

在本研究中，用交叉驗證的方式來尋找適合的節點個數， 考慮了幾種不同切割資料方式來決定訓練資料和測試資料, 並比較不同切割資料的方式下選擇節點的結果與函數估計的效果。

Abstract

In this thesis, I consider the problem of estimating an unknown regression function using spline approximation.

Splines are piecewise polynomials jointed at knots. When using splines to approximate unknown functions, it is crucial to determine the number of knots and the knot locations. In this thesis, I determine the knot locations using least squares for given a given number of knots, and use cross-validation to find appropriate number of knots. I consider three methods to split the data into training data and testing data, and compare the estimation results.

目 錄

圖 目 錄

表 目 錄

1 緒論

無母數迴歸當中，我們觀察 $(x_i.y_i)$, $i = 1, \ldots, n$. 被解釋變數y_i和解釋變數x_i的關係爲:

$$y_i = f(x_i) + \varepsilon_i, i = 1, ..., n \tag{1.1}$$

其中

$$\varepsilon_i \sim N(0, \sigma^2)$$

在無母數的迴歸當中，經常用已知特定類型的函數來近似未知的原始函數f，而spline函數也可以用來近似未知的函數，但是要估計spline函數就需要設定節點(knots)，在本研究中, 固定節點數後, 使用最小平方法估計節點的位置。 由這些節點算出B-spline基底後，就可以根據資料用最小平方法估計基底係數得到函數估計，最後再觀察函數估計效果如何。 所以上述估計方法要先決定節點個數。

當節點用的個數越多，就可能會產生過度配適(overfitting)的問題而產生較大的估計誤差，如果太少就會有函數近似效果不好的問題, 所以節點數的選擇對於spline函數的近似來說是很重要的。 節點數的選取也可以當作是模型選取的問題(詳見第三章),而交叉驗證是一種常用模型選取的方法, 所以這裡嘗試使用交叉驗證的方式選取節點的個數。

使用交叉驗證(cross-validation)選取模型時，可以透過不同的抽樣方式將資料切割成訓練資料與測試資料，本研究中考慮三種不同切割資料的方法, 進行交叉驗證, 會針對方法與樣本大小影響節點數的選取的效果做比較，同時也比較不同的節點數下函數的估計效果。

2 文獻探討

2.1 Cross Validation

Cross-validation由Stone[9]和Geisser[1]發表的模擬方法，在有限的資料下進行重複抽樣，分成訓練資料跟測試資料，用測試資料來測試訓練資料建立的模型可靠度有多少，在Shao[8]中使用了多種CV來選擇線性模型，其中有leave-one-out跟Monte Carlo method的leave-n-out抽樣方法，Monte Carlo cross-validation是Picard and Cook[4]發表的方法，在樣本較小的情況下適合使用，選取一部分的樣本進行隨機的抽樣，重複次數到期望的個數當作訓練資料。

越多的資料建立的模型，通常會較完整，但是會花費大量的時間，相對的如果少量資料建立模型，預測能力就會降低，在Racine[5]的例子中，抽取訓練資料的數量是$n^{0.5}$個，實驗後發現在n大於250的時候就有顯著的穩定性。

2.2 Spline節點的選擇

樣條迴歸(spline regressoin)中節點(knots)的多少會影響迴歸函數的估計. Meyer[3]有提到節點的個數和位置會影響迴歸估計的效果和估計函數的平滑程度，因爲在節點較多的狀況，可以將資料切割成更小的片段，就可以更完整的表現出眞實曲線的細節，如果眞實曲線的轉折較多, 估計出來的曲線也會較不平滑。但如果節點太少, 即使眞實曲線轉折較多, 估計出的曲線也會變平滑。 此外也提到, 節點越多則迴歸函數的參數會越多，近似誤差會變小，但是如果太多就會出現過度配適(overfitting)的狀況, 估計誤差就會變大，所以在此篇文章中就想探討同時兼顧近似與估計誤差的方法。

Spline函數爲分段多項式(piesewise polynomial)，分段點稱爲節點。Spline函數可以透過基底函數組成, 其中一種常用的基底函數是B-spline基底。B-spline在Schoenberg[7]中

有提到是如何構成的基底。估計spline函數時需要設定節點，選擇的方法很多種，像是Halpern[2]跟Ruppert[6]是平均的將節點放在區間中，決定節點後生成spline函數，再觀察配適程度。

3 研究方法

3.1 迴歸函數估計

無母數迴歸當中，我們觀察 $(x_i.y_i)$, $i = 1, \ldots, n$. 被解釋變數y_i和解釋變數x_i的關係如(1.1). 若使用資料估計f得到$\hat{f}$, 可以用ISE(Intergrate Squares Error)來檢測估計效果如何。ISE的定義如下:

$$ISE = \int_a^b (\hat{f}(x) - f(x))^2 dx.$$

(1.1)中，f 可以由一些基底函數的線性組合來近似, 而基底係數的部分可以用最小平方法(Least Squares Method)來估計. 本研究中我們選擇使用spline函數近似原始f函數。以下簡單介紹spline函數。

Spline函數是一種分段多項式，假設分段點爲t_1,...,t_k,

$$t_1 < t_2 < \cdots < t_k,$$

則t_1,...,t_k稱做節點. 節點向量$t = (t_1,\ldots,t_k)$, 階數m的spline 由以下$(m + k)$個基底函數

$$x^0, x^1, ..., x^{m-1}, (x - t_1)_+^{m-1}, ..., (x - t_k)_+^{m-1}$$

組成. 其中

$$(x - t_i)_+^{m-1} = \begin{cases} (x - t_i)^{m-1} & \text{當 } x > t_i; \\ 0 & \text{當 } x \leq t_i, \end{cases}$$

$i \in \{1, \ldots, k\}$, 若g爲節點向量$t = (t_1,\ldots,t_k)$, 階數m的spline函數,則g可以表示成

$$g(x) = \sum_{j=1}^{m+k} \beta_j N_j(x|t)$$

其中

$$N_j(x|t) = \begin{cases} x^{j-1} & \text{當 } 1 \le j \le m; \\ (x - t_{j-m})_+^{m-1} & \text{當 } m+1 \le j \le m+k. \end{cases}$$

在(1.1)中, 當 f 近似於 $g = \sum_{j=1}^{m+k} \beta_j N_j(\cdot|t)$, 我們表示成

$$y_i \approx \sum_{j=1}^{m+k} \beta_j N_j(x_i|t) + \varepsilon_i.$$

所以在節點向量$t = (t_1,\ldots,t_k)$已知下, 基底函數組成係數向量

$$\beta = (\beta_1, \ldots, \beta_{m+k})$$

可用最小平方法估計。 此處選擇$m = 4$. 在節點向量$t = (t_1,\ldots,t_k)$已知下, β的最小平方估計量爲:

$$\hat{\beta}(t) = (X(t)^T X(t))^{-1} X(t)^T Y,$$

其中$X(t)$ 是一個 $n \times (m+k)$ 的矩陣, 而 Y 是一個 $n \times 1$ 的行向量,

$$X(t) = \begin{pmatrix} 1 & x_1 & \ldots & x_1^{m-1} & (x_1 - t_1)_+^{m-1} & \ldots & (x_1 - t_k)_+^{m-1} \\ & & & \vdots & & & \\ 1 & x_n & \ldots & x_n^{m-1} & (x_n - t_1)_+^{m-1} & \ldots & (x_n - t_k)_+^{m-1} \end{pmatrix}$$

而$Y = (y_1 \ldots y_n)^T$.

在節點向量$t = (t_1,\ldots,t_k)$未知但節點個數固定下, 節點位置可以透過最小平方法估計如下: 令

$$rss(t_1, \ldots, t_k) = (Y - X(t)\hat{\beta}(t))^T (Y - X(t)\hat{\beta}(t)),$$

解節點位置$t_1, \ldots, t_k$使$rss(t_1, \ldots, t_k)$最小, 則$\hat{t}_1,\ldots,\hat{t}_k$爲得到的節點位置解。 令 $\hat{t} = (\hat{t}_1,\ldots,\hat{t}_k)$ 則β的估計量爲

$$\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{m+k})^T = \left(X\left(\hat{t}\right)^T X\left(\hat{t}\right) \right)^{-1} X\left(\hat{t}\right)^T Y$$

而f的估計爲$\hat{f}(\cdot) = \sum_{j=1}^{m+k} \hat{\beta}_j N_j(\cdot|t)$.

在節點的個數確定時，可以用以上方法估計出節點的位置並得到f的估計，節點個數的選擇, 我們是使用交叉驗證幫忙挑選。

3.2 資料切割及交叉驗證

在交叉驗證時，需要將資料分成訓練資料跟測試資料兩種，假設M_T是訓練資料，M_V是測試資料，這裡考慮下列1–3三種抽法:

1. 將n個資料隨機抽出c個建立模型(Monte Carlo method):

$$\underbrace{1, 2, \cdots, n-1, n}_{\text{抽取c個}}$$

2. 將n個資料切割成c份，每一份會有n/c個，把每份的第ℓ個取出組成一組訓練資料$M_T^{(\ell)}$，其中$1 \le \ell \le n/c$:

$$\underbrace{1, 2, \cdots, n/c}_{n/c} \underbrace{\cdots}_{n/c}, \cdots, \underbrace{(c-1)n/c+1, \cdots, n-1, n}_{n/c}$$

用矩陣可以表示成

$$\begin{pmatrix} 1 & 2 & \cdots & n/c \\ \vdots & & & \vdots \\ (c-1)n/c+1 & \cdots & n-1 & n \end{pmatrix}$$

3. 將資料切割成c份，每次從每一份隨機抽取一個，集合起來當作訓練資料 重複m次, 得到m次訓練資料 $M_T^{(1)}$,..., $M_T^{(m)}$:

$$\underbrace{1, 2, \cdots, n/c}_{\text{抽1}}, \underbrace{\cdots}_{\text{抽1}}, \cdots, \underbrace{(c-1)n/c+1, \cdots, n-1, n}_{\text{抽1}}$$

上述的第2,3種切割方法, 因爲$x_1, \ldots, x_n$是遞增排序, 所以藉由這種切割方法可以均勻的獲得各部分的資料。

一般來說, 正常的交叉驗證選擇節點個數的方法，是在每次抽取資料後，用這筆抽取的資料在計算節點的位置，然後再用不同的節點個數的位置建立模型在做配適選擇較好的節點數，可是如果這樣會耗費大量的時間，所以在這邊我們是先在最初就用全部的資料找出節點的位置，然後抽取資料後用事先計算好的節點位置, 根據訓練資料來估計係數並建立預測模型，可以節省計算節點位置的時間。

4 模擬資料分析

在本章中，我們資料的生成方式是根據(1.1)，再由第三章中介紹的三種切割方法，來選擇訓練資料跟測試資料，其中$x_1, \cdots, x_n$是由區間$[0,1]$中等距排列而成，同時$[0,1]$也是spline節點的範圍，生成資料的次數爲100次。三種切割方法的細節如下，第一種方法是生成函數後直接抽取訓練資料做了五次，第二種方法是生成函數後，等距切割將同一個位置的取出，但是最多只會得到五組訓練資料，將五組都進行驗證得到最佳knots數，第三種方法是等距切割$[0,1]$使得每一段有5個x_i，每一段隨機抽出一個x_i，組合起來產生一組訓練資料，抽樣重複五次,得到五組訓練資料，將五組都進行驗證得到最佳knots數, 選擇完knots個數再估計spline函數，最後用ISE來觀察估計的效果。

在下兩節會考慮非spline跟spline兩種不同的f，並呈現對應的節點選擇效果以及函數估計的效果。

4.1 f非spline函數

本節中先考慮f爲非spline函數:

$$f(x) = x\sin(20x),$$

根據1.1生成資料100次, σ設定爲0.2.

使用上述f是因爲我們想要觀察在較穩定的函數下的估計結果, $\sin(x)$在$[-1,1]$之間, 而且波型穩定沒有尖銳的轉折, 所以選用sin函數。

(表4.1)–(表4.3)呈現三種不同切割方法下的估計效果 ,每一個表呈現三個不同的n之下,平均100次ISE的結果:

n	3	4	5	6	7	Time	平均ISE
300	0	0	47	34	19	3m11s	0.002144
500	0	1	48	32	19	6m18s	0.001466
1000	0	2	53	25	20	28m1s	0.000942

表 4.1: 非spline隨機抽取訓練資料下的knots分布

n	3	4	5	6	7	Time	平均ISE
300	0	0	56	30	14	3m23s	0.002158
500	0	0	55	31	14	6m18s	0.001395
1000	0	0	23	55	22	28m1s	0.000705

表 4.2: 非spline等距分割訓練資料下的knots分布

n	3	4	5	6	7	Time	平均ISE
300	0	0	44	24	14	3m8s	0.002161
500	0	0	31	45	24	5m8s	0.001333
1000	0	0	32	35	31	22m56s	0.000618

表 4.3: 非spline等距抽取訓練資料下的knots分布

從上面三種方法得到的結果發現, 幾乎都是選到5, 6, 7個的knots, 但是在隨機抽取時, 不管n爲多少, 較容易選到5個knots ,而另外兩種方法在n爲1000時比較容易選到6、7個knots, (圖4.1)–(圖4.3)爲同一組資料在這三組不同的knots估計下函數估計的結果。 從(圖4.1)可以發現knots用5個時，右邊的重疊率還滿高的, 但是左邊

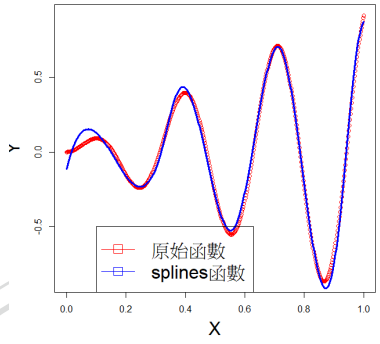

圖 4.1: 非spline下5個knots的函數與原始函數

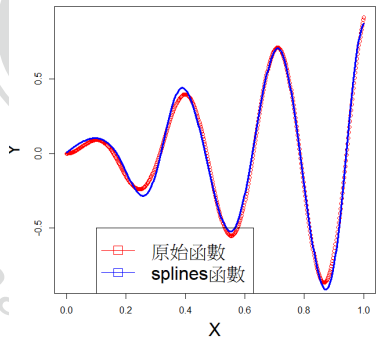

圖 4.2: 非spline下6個knots的函數與原始函數

還是差異很大。 把(圖4.1)和(圖4.2)做比較時會發現，在$[0, 0.2]$的區間有差異，但是(圖4.2)和(圖4.3)，也就是當knots選到6或7的時候, 就沒有太明顯的差異, 所以從估計的角度來看, 雖然knots選擇越多，解釋變數也會增加，解釋能力越好, 但是必要性可能就沒有那麼高。

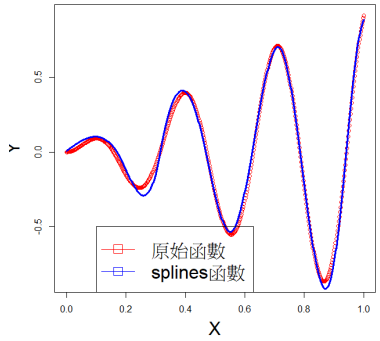

圖 4.3: 非spline下7個knots的函數與原始函數

4.2 f爲spline函數

本節中考慮f爲一個spline函數, 使用B-spline基底, 20個knots, 係數隨機生成。 f圖形如下: 根據1.1生成資料時, 爲了跟上節結果做比較，生成資料維持跟之前一樣的 信噪

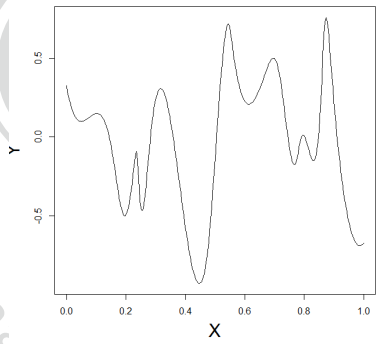

圖 4.4: 20個knots生成的函數

比(signal-to-noise ratio)，生成資料的次數仍然是100次。

(表4.4)–(表4.6)呈現三種不同切割方法下的估計效果 ,每一個表呈現三個不同的n之下,平均100次ISE的結果。 從(表4.4)–(表4.6)發現，雖然生成的knots數是20個，但是選擇後卻多數集中在6個，雖然有隨著n越大後更容易抽取較多的個數，可是還是集中

在6個，平均ISE的部分不像第一個函數一樣，隨著n越大而有明顯的下降。

n	3	4	5	6	7	Time	平均ISE
300	0	0	5	89	6	3m39s	0.024491
500	0	0	5	80	15	5m37s	0.023691
1000	0	0	0	76	24	26m19s	0.022777

表 4.4: spline隨機抽取訓練資料下的knots分布

n	3	4	5	6	7	Time	平均ISE
300	0	0	3	58	39	3m18s	0.024119
500	0	0	0	67	33	5m33s	0.023154
1000	0	0	0	47	53	11m1s	0.022696

表 4.5: spline等距分割訓練資料下的knots分布

n	3	4	5	6	7	Time	平均ISE
300	0	0	2	48	50	3m40s	0.024052
500	0	0	1	62	37	5m8s	0.023256
1000	0	0	0	56	44	22m56s	0.022705

表 4.6: spline等距抽取訓練資料下的knots分布

(圖4.5)–(圖4.7)爲同一組的資料下，這三個不同的knots數估計後的結果。從三張圖可以發現，在[0.6 ∼ 1]之間，(圖4.5)是較平滑的往下，而且尾端沒有勾起，但是(圖4.6)跟(圖4.7)有多了一個更符合原始函數的曲折，在尾端也有勾起，單純觀察6個knots跟7個konts的圖形也並沒有明顯的不同。

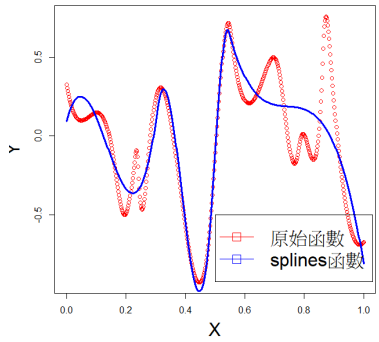

圖 4.5: spline下5個knots的函數與原始函數

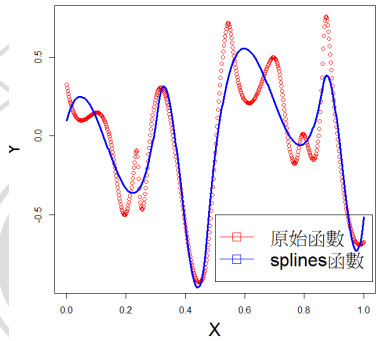

圖 4.6: spline下6個knots的函數與原始函數

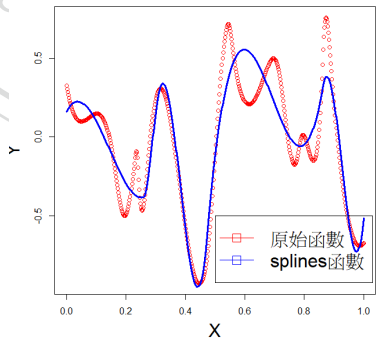

圖 4.7: spline下7個knots的函數與原始函數

5 結果討論與建議

從第四章模擬結果發現, 用交叉驗證選擇knots時我們發現不一定會選到個數最多的knots,也就是說不一定會選到近似效果最好的模型。下面的(表5.1)顯示節點數越多時, 近似效果是越好的, 此處生成的資料是沒有加誤差的。

函數	3	4	5	6	7
非spline	0.02198	0.00852	0.00036	2.015×10^{-5}	2.011×10^{-5}
spline	0.07455	0.05094	0.03065	0.02238	0.02231

表 5.1: 零誤差時的ISE

用交叉驗證不一定選到近似效果最好的模型, 原因可能是使用knots個數太多的模型估計效果未必是最好的, 下面的(表5.2)顯示節點數越多時, 估計效果不一定是越好的, 此處生成的資料是有加誤差的。這邊我們用$n = 300$的樣本，執行10次計算平均ISE.

函數	3	4	5	6	7
非spline	0.02295	0.00966	0.00196	0.00178	0.00189
spline	0.08329	0.05238	0.03433	0.02589	0.02464

表 5.2: 有誤差時10次的平均ISE

單純觀察(表5.2)，我們預期使用交叉驗證選knots時, 非spline函數會選到6個knots, spline函數會選到7個knots. 事實上模擬結果在非spline函數的情況下, 在樣本數多, 使用等距切割及等距抽樣的方法下較容易選到6個knots, 但是如果只是隨機抽取, 仍然最常選到5個knots. 在spline函數的情況下, 可能選到的knots數是5–7個, 但比較不會選

到5個knots, 可能是因爲5個knots估計誤差明顯比較大, 而6與7的估計誤差差距比較小, 交叉驗證可能不太容易分辨出差異。

在(表5.2)中, 不同knots數下估計誤差沒有那麼明顯改變的情況, 圖形也不會有太明顯的不同, 例如像(圖4.2)與(圖4.3)以及(圖4.6)與(圖4.7)。在這種情況下, 交叉驗證也不容易分出差距。

在建議方面，在本文中測試的三種切割方法中, 第三種等距抽樣可以達到最佳的ISE, 可以避免掉出現離群值的狀況，同時又可以重複很多次增加準確度，因此比較推薦這種切割方法。

另外, 本文中只測試了兩種不同的原始函數, 建議未來研究的可以用多種不同的原始函數來測試交叉驗證的效果。也可以考慮測試其他的抽樣方法。

參 考 文 獻

[1] S Geisser. A predictive sample reuse method with application. *Journal of the American Statistical Association*, 70:320–8, 1975.

[2] E.F. Halpern. Bayesian spline regression when the number of knots is unknown. *Journal of the Royal Statistical Society B*, 35:347–60, 1973.

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

[4] R Picard, R and D Cook, R. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:575–583, 1984.

[5] Jeff Racine. Feasible cross-validatory model selection for general stationary processes. *Journal of Applied Econometrics*, 12(2):169–179, 1997.

[6] David Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.

[7] Issac Jacob Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions ,part b: On the problem of osculatory interpolation, a second class of analytic appoximation formulae. *Quart. Appl. Math*, 4:112–141, 1983.

[8] J Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–95, 1993.

[9] M Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147, 1974.