非參數模型近展

臺灣大學數學系　鄭明燕

摘　要

利用核估計建立的非參數模型涵蓋迴歸、機率密度、存活函數、條件參數模型及其他領域，近三十年來一直是統計科學理論與方法發展的主流之一，並且已經在實際問題中被廣泛應用。非參數核估計通常建構在某種區域模型之上，所以在樣本數不大時容易受到變異太大影響，我們探討幾種非參數模型的降變異方法，其源於處理非參數迴歸變異數，見 Cheng, et al. [6]，而後一再被發現可以直接使用至各種不同非參數模型，並且保有極佳的有限樣本和大樣本表現。

一、導論

利用核估計建立的非參數迴歸模型有許多理論和應用方面的優點，其中以區域線性迴歸最受重視，它比較重要的優點包含漸近有效性質，自動適應邊界估計，與極佳的解釋性。關於非參數迴歸分析的發展，建議研讀 Wang 和 Jones [1]與 Fan 和 Gijbels [2]兩本書。針對區域線性迴歸，Choi 和 Hall [3]提出降低估計偏差的方法，他們對在估計位置兩旁的區域線性模式做線性組合，Cheng 和 Hall [4]提出降低估計變異的方法，他們的方法使用殘差對各觀察點取不同的寬值。Cheng et al. [6]的降變異法是取在估計位置鄰近等距三點區域線性迴歸估計量的一個特殊線性組合，使得漸近偏差維持不變而漸近變異減少，而取此鄰近等距三點的位置使得漸近變異減少最多。Cheng et al. [6]詳細探討其理論及實際表現。

當解釋變數爲多維時，我們仍然可以構造區域線性迴歸估計量，這時，比起一維解釋變數時，它的漸近估計偏差仍然保持相同的收斂速度，但是它的漸近估計變異數的收斂速度隨著解釋變數的維度變大而變慢，而且在有限樣本時，它的變異增加許多。Cheng 和 Peng [7]提出兩種推廣 Cheng et al. [6]的降變異方法至多維解釋變數時區域線性迴歸的方法，並且推導其漸近偏差與漸近變異。

條件參數模型假設應變數對自變數的條件機率模型爲給定的參數模型，其中參數隨著自變數而變化並且不須參數模型假設。這種模型有許多優點，第一，它可以用來逼近傳統參數模型，第二，它涵蓋許多常見重要的自變數型態，例如：常態、伽瑪、卜瓦松、二項分配，第三，它可以同時對其中所有參數函數做估計，第四，它可以用來做不同形式的迴歸，例如，均值回歸、百分位數迴歸等。條件參數模型的估計可利用線性區域概似估計，見 Aerts 和 Claesken [8]及 Fan et al. [9]。

在最近幾年，逐漸有許多學者使用區域概似模型分析極值和超越值數據，例如，Davison 和 Ramesh [10]及 Hall 和 Tajvidi [11]對樣本極值的分析以及 Beirlant 和 Goegebeur [12]對超越值的分析。在這類環境數據的分析中，因爲觀察對象是特殊現象，樣本數通常不大，如此線性區域模型估計通常變得不穩定，Cheng 和 Peng [5]使用 Cheng et al. [6]的方法於線性區域模型，得到令人滿意的理論與實際應用結果。

二、一維解釋變數的區域線性迴歸

假設$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ 是一組來自二維分佈(X, Y)的樣本，而(X, Y)服從下列的迴歸模型

$$Y = m(X) + \sigma(X)\varepsilon$$

這裡 $m(\cdot)$是迴歸函數，$\sigma^2(\cdot)$是條件變異函數，並且ε是與自變數 X 獨立的隨機誤差。區域線性迴歸估計 $\hat{m}(x)$，用以估計 $m(x)$，是

$$\frac{S_2 T_0 - S_1 T_1}{S_2 T_0 - S_1 S_1},$$

這裡$S_l = h\sum_{i=1}^{n}(x - X_i)^l K_h(x - X_i)$，$l = 0, 1, 2$，$T_l = h\sum_{i=1}^{n}(x - X_i)^l K_h(x - X_i)Y_i$，$l = 0, 1$，$K_h(t) = K(t/h)/h$，K 是核函數，h 是帶寬值。Fan [13]對其分母加 n^{-2} 項以避免 0/0 情形：

$$\hat{m}(x)=\frac{S_2T_0-S_1T_1}{S_2S_0-S_1S_1+n^{-2}},$$

並且，在一些條件下，推導得到，當 $n\to\infty$，$h\to 0$，$nh\to\infty$，

$$E\hat{m}(x)-m(x)=\frac{1}{2}m''(x)\tau_{21}h^2+o\left\{h^2+(nh)^{-\frac{1}{2}}\right\},$$

$$\mathrm{Var}\,\hat{m}(x)=\frac{\sigma^2(x)}{nh\,f(x)}\tau_{02}+o\left\{h^4+(nh)^{-1}\right\},$$

這裡 $\tau_{jk}=\int t^jK(t)^k\,dt$，fx 是 X 的機率密度函數。

Cheng et al. [6]的降變異方法如下。在估計位置 x 附近取三個等距點 $\alpha_j=x-(r+1-j)\omega$，$j=0,1,2$，這裡 $r\in(-1,1)$，$\omega=\delta h$ 而 $\delta>0$ 是一常數。在得到在 $\alpha_0,\alpha_1,\alpha_2$ 的原始區域線性迴歸估計量 $\hat{m}(\alpha_0),\hat{m}(\alpha_1),\hat{m}(\alpha_2)$ 之後，建立通過它們的二次插補函數，最後取此二次插補函數在 x 的值估計 m(x)：

$$\tilde{m}_q(x)=\sum_{j=0,1,2}A_j(r)\hat{m}(\alpha_j),$$

這裡

$$A_0(r)=\frac{r(r-1)}{2},\ A_1(r)=(1-r^2),\ A_2(r)=\frac{r(r+1)}{2}.$$

如此，Cheng et al. [6]在一些條件下，得到，當 $n\to\infty$，$h\to 0$，$nh\to\infty$，

$$E\,\tilde{m}_q(x)-m(x)=\frac{1}{2}m''(x)\tau_{21}h^2+o\left\{h^2+(nh)^{-\frac{1}{2}}\right\},$$

$$\mathrm{Var}\,\tilde{m}_q(x)=\frac{\sigma^2(x)}{nh\,f(x)}\tilde{\tau}_{02}+o\left\{h^4+(nh)^{-1}\right\},$$

這裡 $\tilde{\tau}_{02}=\tau_{02}-r^2(1-r^2)C(\delta)$，$C(\delta)=1.5\,C(0,\delta)-2C(0.5,\delta)+0.5\,C(1,\delta)$，$C(a,\delta)=\int K(t-a\delta)K(t+a\delta)dt$。所以 $\tilde{m}_q(x)$ 與 $\hat{m}(x)$ 有相同的漸近偏差，而 $\tilde{m}_q(x)$ 有較小漸近變異。固定 K，h 和 δ，$\tilde{m}_q(x)$ 的漸近變異在 $r=\pm\sqrt{1/2}$ 時最小，此時我們有

$$\tilde{m}_\pm(x)=\sum_{j=0,1,2}A_j(\pm\sqrt{1/2})\hat{m}\left(x-(\pm\sqrt{1/2}+1-j)\omega\right),$$

而它們的平均值更對稱的估計 m(x)：

$$\tilde{m}_a(x)=\sum_{j=0,1,2}A_j(\pm\sqrt{1/2})\hat{m}\left(x-(\pm\sqrt{1/2}+1-j)\omega\right)$$

在相同條件下，我們有，當 $n\to\infty$，

$$E\,\tilde{m}_\pm(x)-m(x)=\frac{1}{2}m''(x)\tau_{21}h^2+o\left\{h^2+(nh)^{-\frac{1}{2}}\right\},$$

$$\mathrm{Var}\,\tilde{m}_\pm(x)=\frac{\sigma^2(x)}{nh\,f(x)}\{\tau_{02}-C(\delta)/4\}+o\{h^4+(nh)^{-1}\},$$

$$E\,\tilde{m}_a(x)-m(x)=\frac{1}{2}m''(x)\tau_{21}h^2+o\left\{h^2+(nh)^{-\frac{1}{2}}\right\},$$

$$\mathrm{Var}\,\tilde{m}_a(x)=\frac{\sigma^2(x)}{nh\,f(x)}\{\tau_{02}-C(\delta)/4-D(\delta)/2\}+o\{h^4+(nh)^{-1}\},$$

這裡

$$\begin{aligned}D(\delta)=\tau_{02}-\frac{1}{4}C(\delta)&-\frac{1}{16}\left\{4(1+\sqrt{2})C\left(\sqrt{2}-1,\frac{\delta}{2}\right)\right.\\&+(3+2\sqrt{2})C\left(2-\sqrt{2},\frac{\delta}{2}\right)+2C\left(\sqrt{2},\frac{\delta}{2}\right)\\&+4(1-\sqrt{2})C\left(\sqrt{2}+1,\frac{\delta}{2}\right)\\&\left.+(3-2\sqrt{2})C\left(\sqrt{2}+2),\frac{\delta}{2}\right)\right\}.\end{aligned}$$

如果 K 是對稱的，則 $C(\delta)\ge 0$，對任何的 $\delta\ge 0$。如果 K 有唯一的極大值且為 convave，則 $C(\delta)$ 是 $\delta>0$ 的遞增函數。而 $D(\delta)$ 對任何 $\delta\ge 0$ 是非負的。Cheng et al. [6]對 $\hat{m}(x)$，$\tilde{m}_\pm(x)$，$\tilde{m}_a(x)$ 做了詳細的有限樣本比較，得到的結論是 $\delta=1$ 是一個可靠的選擇，取 $\delta=1$ 時，$\tilde{m}_a(x)$ 的最佳平均平方差(Mean Integrated Squared Error; MISE)比 $\hat{m}(x)$ 的最佳平均平方差大約少 10%，不管樣本數是 50，100 或更大，而 Choi 和 Hall [3]的降偏差法只有在大樣本時($n\ge 250$)才會表現優於 $\hat{m}(x)$。

我們的估計 $\tilde{m}_q(x)$，$\tilde{m}_\pm(x)$，$\tilde{m}_a(x)$ 的一個重要優點是，它們都不需要另外帶寬值選擇法，因爲它們的漸近最佳帶帶寬值都與 $\hat{m}(x)$ 的漸近最佳帶寬值相差某個固定已知的常數倍，所以可以直接使用 $\hat{m}(x)$ 的帶寬選擇法(如 Brockmann, et al. [14]; Fan 和 Gijbels [15, 20]; Gasser, et al. [16]; Ruppert [17]; Ruppert, et al. [18])，乘上此常數倍數即可。

三、多維解釋變數的區域線性迴歸估計

假設$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$是獨立且

服從

$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i,\ i = 1, 2, \ldots, n,$

迴歸模型的一組樣本，這裡 $X_i = (X_{i1}, X_{i2}, \ldots X_{id})^T$，$i = 1, 2, \ldots, n$，與隨機誤差$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$獨立。多維區域線性迴歸估計量爲

$$\hat{m}(x) = e^T (X^T) WX)^{-1} X^T WY,$$

這裡 e = (1, 0, …, 0)T 是一個(d + 1)-向量，Y = $(Y_1, Y_2, \ldots, Y_n)$T，

$$W = \operatorname{diag}\left\{\prod_{j=1}^{d} K\left(\frac{x_{1j}-x_j}{b_j h}\right), \ldots, \prod_{j=1}^{d}\left(\frac{x_{nj}-x_j}{b_j h}\right)\right\},$$

$$X = \begin{pmatrix} 1 & \cdots & (X_1 - x)^T \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (X_n - x)^T \end{pmatrix}$$

在一些條件下，Ruppert 和 Wand [19]得到

$$E\{\hat{m}(x) - m(x) | X_1, X_2, \ldots, X_n\} = \frac{1}{2} h^2 \tau_{21} \operatorname{tr}\{BM_2\} + o_p(h^2),$$

$$\operatorname{Var}\{\hat{m}(x) | X_1, X_2, \ldots, X_n\} = \frac{\sigma^2(x)}{nh^d\ f(x)\prod_{i=1}^{d} b_i} \tau_{02}^d \{1 + o_p(1)\},$$

這裡

$$B = \operatorname{diag}\{b_1^2, b_2^2, \ldots, b_d^2\}, M_2\left(\frac{\partial^2}{\partial x_i\, \partial x_j} m(x)\right),$$

Cheng 和 Peng [7]提出兩種降低線性區域估計變異的方法，其中一種方法是將整個估計範圍對每一個解釋變數取等距格子，然後取 x 座落的 d-維格子，然後對此 d-維格子每一維度分別取二次插補係數後取乘積，造線性組合，如此，漸近條件偏差保持不變，漸近變量變小但與 x 的位置有關。另一方法則推廣 Cheng et al. [6]的方法：令

$$r = (r_1, r_2, \ldots, r_d)^T \in \left\{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}^d,$$

$$\alpha_{x,r} = x - h((1+r_1)\delta_1 b_1, \ldots, (1+r_b)\delta_d b_d)^T,$$

$$\Lambda_{x,r} = \{x^*_{k_1,\ldots,k_d} = \alpha_{x,r} + h(k_1\delta_1 b_1, \ldots, k_d\delta_d b_d)^T : (k_1, \ldots, k_d)^T \in \{0, 1, 2\}^d\}$$

而

$$\tilde{m}_r(x;\delta) = \sum_{x^*_{k_1,\ldots,k_d} \in \Lambda_{x,r}} \{\prod_{i=1}^{d} A_{k_1}(r_i)\hat{m}(x^*_{k_1,\ldots,k_d})\}$$

與 $\hat{m}(x)$ 有相同漸近條件偏差且

$$\operatorname{Var}\{\tilde{m}_r(x;\delta) | X_1, X_2, \ldots, X_n\} = \frac{\sigma^2(x)}{nh^d\ f(x)\prod_{i=1}^{d} b_i} \prod_{i=1}^{d}\left\{\tau_{02} - \frac{C(\delta_i)}{4}\right\} \times \{1 + o_p(1)\}$$

這些估計量的平均值

$$\tilde{m}(x;\delta) = 2^{-d} \sum_{r \in \left\{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}^d} \tilde{m}_r(x;\delta)$$

則進一步降低變異：

$$\operatorname{Var}\{\tilde{m}(x;\delta) | X_1, X_2, \ldots, X_n\} = \frac{\sigma^2(x)}{nh^d\ f(x)\prod_{i=1}^{d} b_i} \prod_{i=1}^{d}\left\{\tau_{02} - \frac{C(\delta_i)}{4} - \frac{D(\delta_i)}{2}\right\} \times \{1 + o_p(1)\}.$$

Cheng 和 Peng [7]建議取$\delta_1 = \delta_2 = \ldots = \delta_d = 1$。而 $\tilde{m}(x;\delta)$ 保有所有一維線性區域迴歸降變異估計量 $\tilde{m}_a(x)$ 的各種優點。

四、區域概似模型

假設$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$是一組來自二維分配(X, Y)的樣本，而給定 X = x，Y 的條件機率密度函數是

$$f(y; \theta_1(x), \theta_2(x), \ldots, \theta_k(x)),$$

這裡 f 是已知的，$\theta(x) = (\theta_1(x), \theta_2(x), \ldots, \theta_k(x))^T$是 x 的函數。如此，Y 對 X = x，的相依關係由θ(x)表達，而θ(x)的區域概似函數爲

$$L_n(\theta^*(x)) = \sum_{i=1}^{n} K_h(X_i - x) \log f(Y_i; \theta(x, X_i)),$$

這裡

$$\theta(x, X_i) = \theta(x) + (\theta_1'(x)(X_i - x), \ldots, \theta_k'(x)(X_i - x)^T,$$

$$\theta^*(x) = (\theta_1(x), \theta_1'(x), \ldots, \theta_k(x), \theta_k'(x))^T.$$

如果

$$\hat{\theta}^*(x) = (\hat{\theta}_1(x), \hat{\theta}_1'(x), \ldots, \hat{\theta}_k(x), \hat{\theta}_k'(x))^T = \arg\max_{\theta^*(x)} L_n(\theta^*(x)),$$

則θ(x)的區域限性最大概似估計量(Local Linear Miximum Likelihood Estimator; LLMLE)爲

$$\hat{\theta}(x)=(\hat{\theta}_1(x),\ldots,\hat{\theta}_k(x))^T.$$

爲了簡化符號，我們假設 k = 2。在一些條件之下，Aerts 和 Claeskens [8]得到，當 $n\to\infty$，

$$\sqrt{nh}\left\{\hat{\theta}(x)-\theta(x)-\frac{1}{2}h^2\tau_{21}\theta''(x)\right\}\to_d Z_2,$$

這裡

$$\theta''(x)=(\theta_1''(x),\theta_2''(x))^T,$$
$$Z_2\sim N(0,\Sigma_2),\ \Sigma_2=\tau_{02}f(x)^{-1}I(\theta(x))^{-1},$$
$$I(\theta(x))=E_x\left\{-\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(Y;\theta_1(x),\theta_2(x))\right\}_{i=1,2,j=1,2}$$

Cheng 和 Peng [5]令

$$\tilde{\theta}(x)=\sum_{j=0,1,2}A_j(r)\hat{\theta}(\alpha_j),$$

其中 $A_j(r)$、α_j，j = 0,1,2 如第二章所定義。則 $\tilde{\theta}(x)$ 與 $\hat{\theta}(x)$ 有相同漸近偏差，而其漸近變異爲

$$(nh)^{-1}\Sigma_1=(nh)^{-1}\{\tau_{02}-r^2(1-r^2)C(\delta)\}\ f(x)^{-1}I(\theta(x))^{-1}$$

令

$$\tilde{\theta}_\pm(x)=\sum_{j=0,1,2}A_j(\pm1/\sqrt{2})\hat{\theta}(x-(\pm\sqrt{1/2}+1-j)\delta h),$$

$$\tilde{\theta}_a(x)=\frac{1}{2}\{\tilde{\theta}_+(x)+\tilde{\theta}_-(x)\},$$

則 $\tilde{\theta}_\pm(x)$ 和 $\tilde{\theta}_a(x)$ 皆與 $\hat{\theta}(x)$ 有相同漸近偏差，而 $\tilde{\theta}_\pm(x)$ 和 $\tilde{\theta}_a(x)$ 的漸近變異分別爲

$$(nh)^{-1}\{\tau_{02}-C(\delta)/4\}f(x)^{-1}I(\theta(x))^{-1}$$

及

$$(nh)^{-1}\{\tau_{02}-C(\delta)/4-D(\delta)/2\}f(x)^{-1}I(\theta(x))^{-1}.$$

Cheng 和 Peng [5]在他們的模擬實驗中取 f(x)爲極值機率密度或 Logistic regression 函數，$\delta=1$，得到 $\tilde{\theta}_a(x)$ 的變異皆比 $\hat{\theta}(x)$ 小。在應用於分析荷蘭 De Bilt 氣象觀測站測得 1901 年到 2003 年年最高氣溫數據時，他們發現 $\tilde{\theta}_a(x)$ 的表現比 $\hat{\theta}(x)$ 穩定許多，而此數據的樣本數只有 103。

參考文獻

[1] M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman & Hall, London (1995).

[2] J. Fan and I. Gijbels, *Local polynomial modeling and its applications*, Chapman & Hall, London (1996).

[3] E. Choi and P. Hall, *Biometrika*, **85**, 333 (1998).

[4] M.-Y. Cheng and P. Hall, *Statistica Sinica*, **12**, 429 (2002).

[5] M.-Y. Cheng and L. Peng, *Journal of the American Statistical Association*, **102**, 293 (2007).

[6] M.-Y. Cheng, L. Peng and J.-S. Wu, *The Annals of Statistics*, **35**, 522 (2007).

[7] M.-Y. Cheng and L. Peng, *Journal of Multivariate Analysis*, **97**, 1501 (2006).

[8] M. Aerts and G. Claeskens, *Journal of the American Statistical Association*, **92**, 1536 (1997).

[9] J. Fan, M. Farmen and I. Gijbels, *Journal of the Royal Statistical Society Series B*, **60**, 591 (1998).

[10] A. C. Davison and N. I. Ramesh, *Journal of the Royal Statistical Society Series B*, **62**, 191 (2000).

[11] P. Hall and N. Tajvidi, *Statistical Science*, **15**, 153 (2000).

[12] J. Beirlant and Y. Goegebeur, *Journal of Multivariate Analysis*, **89**, 97 (2004).

[13] J. Fan, *The Annals of Statistics*, **21**, 196 (1993).

[14] M. Brockmann, T. Gasser and E. Herrmann, *Journal of the American Statistical Association*, **88**, 1302 (1993).

[15] J. Fan and I. Gijbels, *Journal of the Royal Statistics Society Series B*, **57**, 371 (1995).

[16] T. Gasser, A. Kneip and W. Kohler, *Journal of the American Statistical Association*, **86**, 643 (1991).

[17] D. Ruppert, *Journal of the American Statistical Association*, **92**, 1049 (1997).

[18] D. Ruppert, S. J. Sheather and M. P. Wand, *Journal of the American Statistical Association*, **90**, 1257 (1995).

[19] D. Ruppert and M. P. Wand, *The Annals of Statistics*, **22**, 1346 (1994).

[20] J. Fan, N. E. Heckman and M. P. Wand, *Journal of the American Statstical Association*, **90**, 141 (1995).