高維度巨量資料的統計模型與分析

台大數學系　鄭明燕

摘　要

近十多年來，由於數據收集與存取的設備快速發展，在不同的研究與應用領域，產生許多不同形態、大量、高維度的資料。如何分析這些資料， 從中獲得新知識，進而從事決策與預測，並且同時能夠確保在學習與應用時的正確性與有效率性質，是十分俱有挑戰性，而且亟需創新力的大問題。統計學習理論與方法，在這些方向持續做了很多突破性的重要貢獻。在這裡，我做一些簡短的回顧，涵蓋針對幾種不同常見資料型態的參數、非參數、以及半參數統計模型利用核估計建立的非參數模型下的統計學習方法及理論的發展。

關鍵詞：統計學習，長期追蹤資料，核估計，區域線性迴歸，非參數迴歸模型，變數選擇。

一、導論

近十多年來，由於數據收集與存取的設備快速發展，在不同的研究與應用領域，產生許多不同形態、大量、高維度的資料，例如，生物資訊著重的基因體資料、大型醫學與流行病學研究中常見的的長期追蹤資料、財務金融中的財務與金融商品時間序列、工業發展中的品質控制資料、網路商業與社群資料、氣象資料等等。如何分析這些龐大而且複雜的資料，從中獲得新知識，進而從事決策與預測，並且能夠確保在學習與應用時的正確性與有效率性質，是十分俱有挑戰性，而且亟需創新力的大問題。統計學習理論與方法，在這些方向持續做了很多突破性的重要貢獻。

首先，考慮傳統的線性迴歸模型，大家都知道，當解釋變數的維度很高，甚至遠超過樣本數時，迴歸模型中參數的個數過大，其最小平方估計量的計算很複雜而且費時，更大的問題是它並不保有一些基本的統計性質，例如收斂性。Tibshirani (1996)提出的 Lasso，用的是 Shrinkage 的統計想法，對最小平方估計施加一個懲罰項，使得許多接近零的估計值被設成零值，也就是去除了許多不重要的解釋變數，而保留少數重要的解釋變數，問題是 Lasso 變數選擇法並沒有所謂的選擇一致性，也就是參數爲零的解釋變數不該保留而參數非零的解釋變數都應該保留。Fan 和 Li (2001)提出的 SCAD，也是利用這種想法，不過它的懲罰項是經過小心設計，使得它俱有變數選擇一致性。Zou (2006)提出的 adaptive Lasso，給予 Lasso 懲罰項中每個解釋變數不同的權重，得以達到變數選擇一致性。Yuan 和 Lin (2006)提出的 group Lasso，進一步將解釋變數分群的結構放入 Lasso 懲罰項中。

變數選擇方法仍有它們不足之處，它們的選擇一致性只有當解釋變數的維度被控制在樣本數的 1/3 次方以內，當解釋變數的維度較這大，甚至遠超過樣本數時，它們是不可行的。爲了突破這個瓶頸，Fan 和 Lv (2008)提出 Sure Independence Screening 的變數篩檢方法，他們的想法是執行個別解釋變數的 Marginal Models，將它們的參數估計值由大到小排序，然後篩選掉後面大部份的不重要的解釋變數，他們也提出和証明了這方法的 Sure Screening 性質，也就是參數非零的解釋變數都應該保留。

時常，傳統的線性迴歸模型等參數模型太過僵硬，無法反應資料中隱含的其它結構，因而產生過大的模型偏差，造成嚴重錯誤的決策與預測。相對的，非參數模型很有彈性，能夠讓資料自己說出資料中隱含的複雜結構，其中，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，而以區域線性迴歸最受重視，它比較重要的優點包含極小極大漸近有效性質、自動適應邊界估計、與自動適應設計性。關於非參數

迴歸分析的發展和區域線性迴歸，建議研讀 Wand 和 Jones (1995)與 Fan 和 Gijbels (1996)兩本書。不過，非參數模型估計的變異數相對比較大，造成它們有相對於參數模型比較慢的收斂速度，而這情形更隨著解釋變數維度的增加而亦形惡化，因此，通常我們至多只用在二維以下的時候，高維度時，是幾乎不可行的。

考量前面提到的兩難處境，好幾種非參數迴歸模型在近十幾年被提出，並且被廣泛應用於實際問題中，其中最成功的是（半）變異係數迴歸模型（Hastie 和 Tibshirani, 1993; Fan 和 Zhang, 1999; Xia, Zhang 和 Tong, 2004）、單指標模型 (Ichimura, 1993)、和相加性模型（Buja, Hastie, Tibshirani, 1989）。Cheng, Honda 和 Zhang (2016) 探討超高維度時變異係數迴歸模型選擇。

長期追蹤資料常見於大型醫學與流行病學研究中，單一個體可能在不同的時間點被觀察到其解釋變數與應變數的值，半變異係數迴歸模型已經是標準的分析模型，在超高維度情形時，Cheng, Honda, Li 和 Peng (2014)建構一套兩步驟方法以達到正確的半變異係數迴歸模型。實際上，很多資料中的解釋變數並不能如以上的變數選擇和變數篩檢方法一分爲二的，理由是它們通常有相關性的，而它們的相關性亦時常是非線性的，Cheng 和 Wu (2013)提出一套當高維解釋變數俱有低維流形結構時的非參數迴歸模型。

以下我簡短介紹 Cheng, Honda 和 Zhang (2016)和 Cheng, Honda, Li 和 Peng (2014)的方法及應用，詳細理論及模擬結果請參考論文本文。

二、超高維下變異係數迴歸模型

假設我們有(p+1)-維度解釋變數 $\mathrm{X} = (1, \mathrm{X}_1, \mathrm{X}_2, \ldots \mathrm{X}_\mathrm{p})^\mathrm{T}$，一個指標變數 T 和單一應變數 Y 的 n 個互相獨立觀察值，記成 $\mathrm{X}_\mathrm{i} = (1, \mathrm{X}_{\mathrm{i}1}, \mathrm{X}_{\mathrm{i}2}, \ldots \mathrm{X}_{\mathrm{ip}})^\mathrm{T}$，$\mathrm{i} = 1,2, \ldots \mathrm{n}$，$\mathrm{T}_\mathrm{i}, \mathrm{i} = 1,2, \ldots, \mathrm{n}$ 和 $\mathrm{Y}_\mathrm{i}, \mathrm{i} = 1,2, \ldots, \mathrm{n}$。假設解釋變數X和應變數$Y$服從變異係數迴歸模型：

$$\mathrm{Y} = \beta(T)^\mathrm{T} X + \varepsilon$$

這裡 $\beta(T)^T = (\beta_0(T), \beta_1(T), \ldots, \beta_\mathrm{p}(T))$ 是 (p+1)-維度指標變數 T 的函數向量，稱爲變異係數函數向量，而ε是隨機誤差項。在超高維度下，大部分的解釋變數都是不重要的，只有小部分的解釋變數的變異係數函數不是零函數。

對任意一個指標集合 M，$\{0\} \subset M \subset \{0,1.2 \ldots., p\}$，令 $\hat{\sigma}^2(M)$ 有指標集合 M 解釋變數的 Spline 變異係數迴歸模型所得到的是隨機誤差項變異數估計值，同時我們定義個模型選擇標準：

$$\mathrm{BIC(M)} = \mathrm{n}\log[\hat{\sigma}^2(M)] + \#M \times L \times \log(n),$$

這裡L是用來估計變異係數函數的 Spline basis 的維度。假使我們在逐步變數選擇過程中已經找到一些重要變數，它們的指標集合爲 S，$\{0\} \subset S \subset \{0,1.2 \ldots., p\}$。我們希望進一步尋找下一個可能的重要變數，方法如下：令 $S(\mathrm{l})$ 代表一個指標集合 S 和 $\{l\}$ 的聯集，這裡 $l \in \{0,1,2 \ldots., p\} \backslash S$。令 $\mathrm{l}^*(S)$ 是所有 $l \in \{0,1,2 \ldots., p\} \backslash S$ 中使得 $\hat{\sigma}^2(S(l))$ 爲最小值者，如此，第 $\mathrm{l}^*(S)$ 個變數就是下一個可能的重要變數。

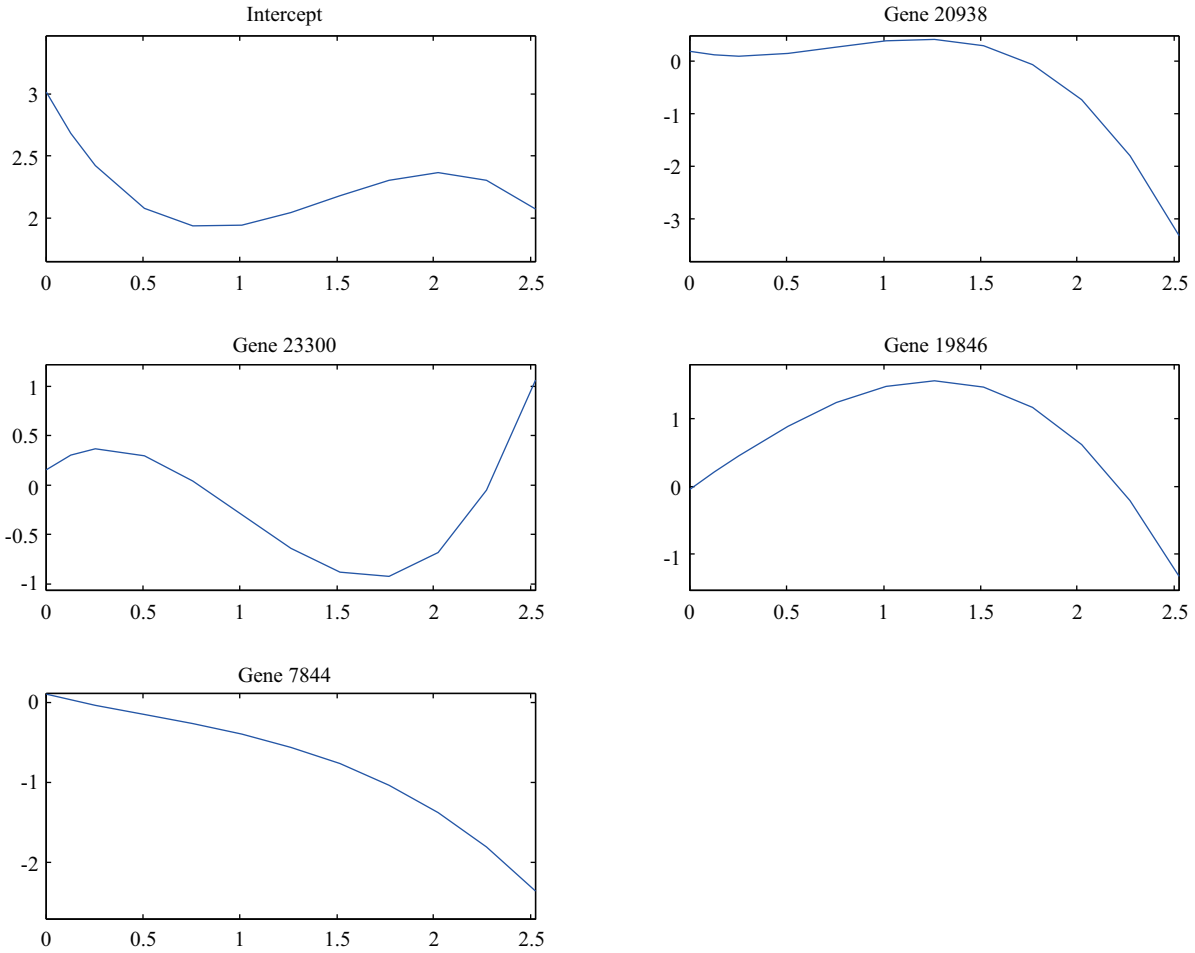
Cheng, Honda 和 Zhang (2016)的逐步變數篩檢法如下：

1. 起始步驟：令 $\mathrm{k} = 0$，選擇指標集合 $\mathrm{S}_1 = \{0\}$，並且計算 $\mathrm{BIC}(\mathrm{S}_1)$ 。
2. 逐步變數選擇：在第 $\mathrm{k}+1$ 步時，計算 $\mathrm{l}^*(\mathrm{S}_\mathrm{k})$。令 $\mathrm{S}_{\mathrm{k}+1} = \mathrm{S}_\mathrm{k} \cup \{\mathrm{l}^*(\mathrm{S}_\mathrm{k})\}$ 並且計算 $\mathrm{BIC}(\mathrm{S}_{\mathrm{k}+1})$。如果 $\mathrm{BIC}(\mathrm{S}_{\mathrm{k}+1}) > \mathrm{BIC}(\mathrm{S}_\mathrm{k})$ 則停止變數選擇，不然，則將 $\mathrm{k}+1$ 改爲 $\mathrm{k}+1$ 然後繼續執行變數選擇。

van't Veer et al. (2002)的乳癌資料紀錄 97 位病人的 24481 個基因的 micro array 表現以及風險因子 age, tumor size, histological grade, angio-invasion, lymphocytic infiltration, estrogen receptor 和 progesterone receptor status。我們以 tumor size 爲應變數，estrogen receptor 爲指標變數，基因的 micro array 表現爲解釋變數，利用以上逐步變數篩檢法，篩檢出四個重要解釋變數，它們對 tumor size 的影響變異係數函數如圖一。我們可以清楚的觀察到圖中的影響變異係數函數都不是常數，亦即，傳統的線性迴歸模型並不合適。

三、超高維長期追蹤資料的半變異係數迴歸模型

如同前段，假設我們有(p+1)-維度解釋變數



圖一　基因表現對乳癌 tumor size 的影響係數，隨 estrogen receptor 指標變化而不同的變化圖

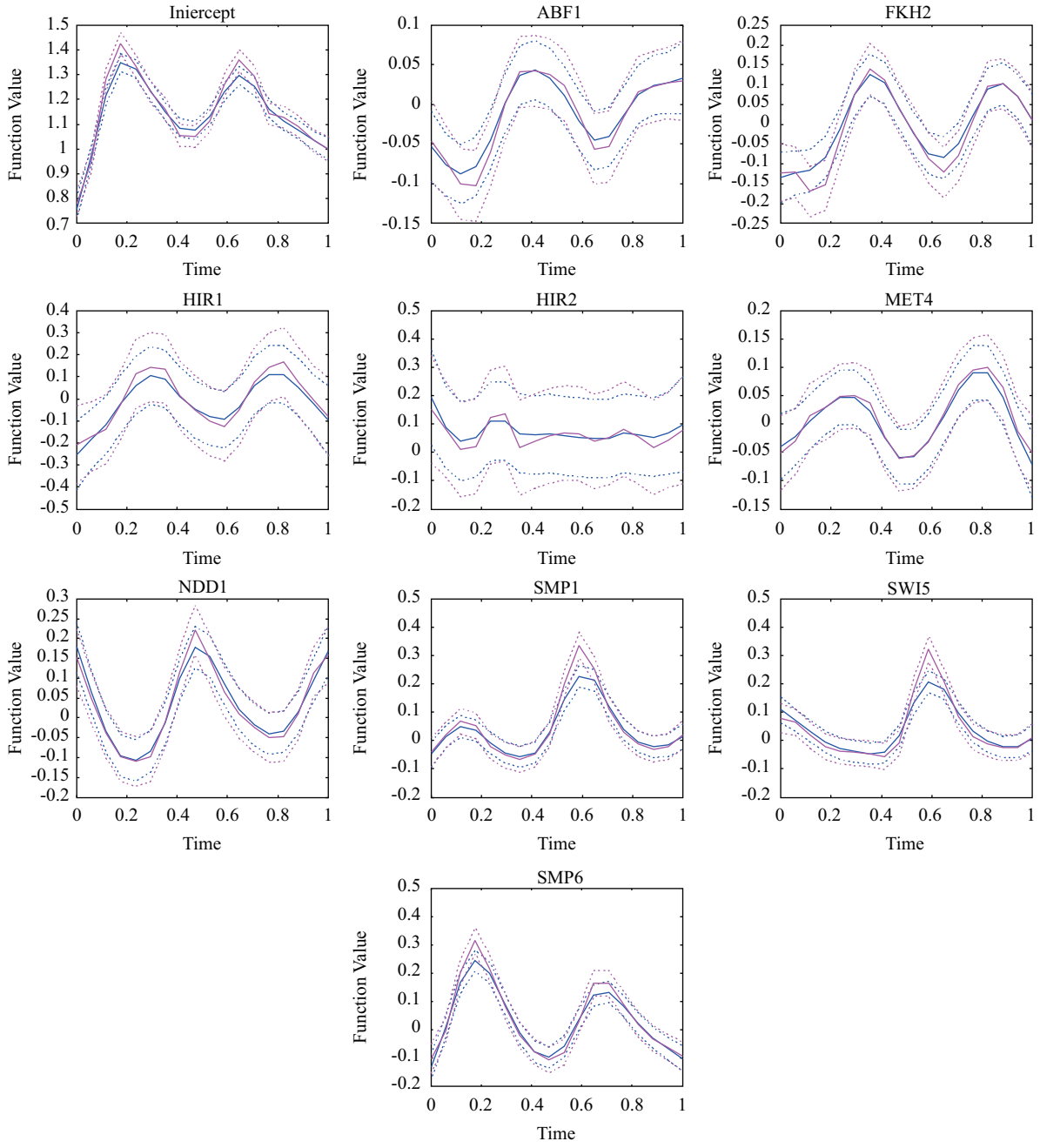
$X = (1, X_1, X_2, \ldots X_p)^T$ 和單一應變數 Y 的 n 個互相獨立觀察個體，在長期追蹤研究裡，會每一個觀察個體在不同時間點做觀察，所以長期追蹤資料包函 $t_i = (t_{i1}, \ldots, t_{im\_i})^T$，$Y_i(t_i)$ 及 $X_i(t_i)$，$i = 1,2, \ldots, n,$ 這裡 t_i 是第 i 個觀察個體的觀察時間點向量。長期追蹤資料的半變異係數迴歸模型是指在每一個觀察時間點，我們有

$$Y(t) = \beta_0(t) + \beta_1(t)^T X(t) + \varepsilon$$

在超高維度下，大部分的解釋變數都是不重要的，只有小部分的解釋變數的變異係數函數不是零函數。而重要的的解釋變數中，又區分爲有常數變異係數者，及有非常數變異係數者。

Cheng, Honda, Li 和 Peng (2014)的兩步驟方法以模型結構決定法則如下：在第一步驟，我們使用 Fan 和 Lv (2008)的 Sure Independence Screening 變數篩檢想法，執行所有個別解釋變數的 Marginal Varying-Coefficient Models，將它們的變異係數函數估計值的 L2 norm 由大到小排序，然後篩選掉後面大部份的不重要的解釋變數。在第二步驟，我們使用 group SCAD 對第一步驟篩選後留下來的解釋變數所配適 spline 變異係數函數估計值 L2 norm 以及偏離常數，分別做 group SCAD 懲罰項，如此，得到的重要解釋變數，自動被區分爲有常數變異係數者，及有非常數變異係數者。

Spellman et al. (1998)的 Yeast Cell Cycle 資料紀錄 297 個基因在 18 個時間點的 micro array 表現，以及 96 個 transcription factors (TF)。我們想知哪些 transcription factors 對 yeast cell cycle 有 gene regularization 作用，使用 Cheng, Honda, Li 和 Peng (2014)的兩步驟方法，我們得到 MCM1 和 RLM12 的常數變異係數估計值分別爲 0.022 和-0.0097，以及非常數變異係數估計如圖二。



圖二　對 yeast cell cycle 影響會隨時間變化的 Transcription factors 的變異係數圖

參考文獻

[1] Buja, A., Hastie, T., and Tibshirani, R. (1989). “Linear Smoothers and Additive Models”, *The Annals of Statistics* 17(2):453-555.

[2] Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. Ann. Statist. 42, 1819-1849.

[3] Cheng, M.-Y., Honda, T., and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. J. Amer. Statist. Assoc., to appear.

[4] Cheng, M.-Y. and Wu, H.-T. (2013). Local linear regression on manifolds and its Geometric interpretation. J. Amer. Statist. Assoc. 108, 1421-1434.

[5] Fan, J. and Gijbels, I. (1996) Local polynomial

modelling and its applications. Chapman & Hall: London.

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348 -1360.

[7] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. J. Royal Statist. Soc. Ser. B 70 849{911.

[8] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coeffi- cient models. Ann. Statist. 27 1491-1518.

[9] Hastie, T. and Tibshirani, R. (1993). Varying-Coefficient models. J. Royal Statist. Soc. Ser. B 55 757-796.

[10] Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models. *Journal of Econometrics*. **58**: 71-120.

[11] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstien, D. and Futcher, B. (1998). Com-prehensive identi_cation of cell cycle-regulated genes of the yeast sacchromyces cerevisiae by microarray hybridization. Molecular Biology of Cell 9 3273-3297.

[12] Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. J. Royal Statist. Soc. Ser. B 58 267-288.

[13] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415:530-536.

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

[15] Xia, Y., Zhang, W. and Tong, H. (2004). E_cient estimation for semivarying-coefficient models. Biometrika 91 661{681.

[16] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Royal Statist. Soc. Ser. B 68 49-67.

[17] Zou, H. (2006). The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101 1418-1429.