正交貪婪演算法與高維迴歸模型選取

中央研究院統計科學研究所　銀慶剛

一、簡　介

高維迴歸模型是近十年來統計學研究中最受矚目的議題之一，其與一般迴歸模型最大的差異在於解釋變數的個數(p)遠超過樣本數(n)，故傳統的統計推論不再適用。然而，隔阻在這些技術困難後面的卻是廣袤無垠的應用世界，包含：生物資訊、訊號及影像處理、金融風險管理、行銷研究、及製造工程等等。這些有趣的應用，驅使著統計學家們竭盡所能地突破高維度所帶來的障礙。

天底下沒有白吃的午餐，要在 p 大 n 小的問題上有所斬獲，統計學家採用的辦法通常是假設迴歸係數滿足某種程度的“稀疏性”。例如，假設非零係數之個數比 n 小很多（此稱爲“強”稀疏性），或者假設大部分係數之值很小，只有少數係數較顯著（此稱爲“弱”稀疏性）。不管是哪一種稀疏性，都容許我們只定焦於模型中相對少量（與 n 比）的顯著係數，進而避開用較少資料估計較多參數的窘境。然而，要在 p 極大的情況下鎖定那些顯著係數，仍有如大海撈針般可望而不可及。事實上，這涉及的是解一道 l_0 最小化問題，爲一非凸(non-convex)最佳化問題。當 p 大時，此類問題是 NP-hard（見[1]及[2]），故在計算上窒礙難行。爲了克服此一困擾，Candès 與 Tao [3]設計了一個 l_1 最小化問題來近似 l_0 最小化問題，此手法通稱爲 l_1 鬆弛(l_1 relaxation)。因 l_1 最小化問題爲一凸(convex)最佳化問題，求解極爲迅速便利，又因近似得宜，故所解出來的非零係數便與眞正的非零（或顯著）係數接近。

Candès 與 Tao 的方法亦稱爲 Dantzig Selector，它與知名的高維選模方法 Lasso (Least Absolute Shrinkage and Selection Operator, Tibshirani [4])漸近等價。它們享有同樣的計算便利，但在理論上也面臨相同的侷限。在強稀疏性下，它們選取變數的一致性（亦即選到眞正非零係數的機率隨著 n 趨近∞而趨近於 1），必須在解釋變數的相關係數矩陣滿足若干頗爲嚴格的條件（例如 irrepresentable 條件或 neighborhood stability 條件，見[5]及[6]）下，才得以確保。另一方面，Dantzig Selector 及 Lasso 在弱稀疏性下則擁有極好的理論性質。特別地，他們在預測上的 minimax 最佳性可在遠較 irrepresentable 或 neighborhood stability 弱的條件下獲得，見[7]。

筆者及合作者過去幾年在上述大海撈針的問題上，則是用一種稱爲正交貪婪演算法(Orthogonal Greedy Algorithm, OGA，見[8])的遞迴方法，逐次挑選重要變數。我們也設計了一個高維訊息準則(High-Dimensional Information Criterion, HDIC，見[9])，作爲此一遞迴算則的停止法則。我們進一步以 HDIC 爲工具，對算則停止前選入的變數作更細緻的篩選(Trim)。在強稀疏性下，我們[9]建立了 OGA+HDIC+Trim 的一致性且無需假設 irrepresentable 或 neighborhood stability 條件。在各式不同的弱稀疏性下，我們[10]建立了 OGA+HDIC 在預測上的最佳性。

二、母體 OGA 及其收斂速度

爲了探討 OGA 的收斂速度我們首先考慮母體(population)迴歸模型

$$y(\mathbf{x}) = \beta_1 x_1 + \cdots + \beta_p x_p, \tag{1}$$

其中 $\mathbf{x} = (x_1, \ldots, x_p)'$，$\mathrm{E}(x_j) = 0$, $j = 1, \ldots, p$ 且對所有 $1 \le i, j \le p$，$\mathbf{E}(x_j^2) = \sigma_j^2$，$\mathrm{E}(x_i x_j)$ 及 $\mathrm{E}(y(\mathbf{x})x_j)$ 均已知。不失一般性地，我們令 $\sigma_j^2 = 1$, $j = 1, \ldots, p$。給定 $0 < \xi \le 1$，母體弱 OGA (WOGA) 進行如下：

起始：$J_{\xi,0} = \varnothing;\ y_{J_{\xi,0}}(\mathbf{x}) = 0;\ u_0 = y(\mathbf{x}).$

遞迴：令 $1 \le j_{\xi,m} \le p$ 爲任意一個滿足

$$\left|E(u_{m-1}x_{j_{\xi,m}})\right| \ge \xi \max_{1\le j\le p}\left|E(u_{m-1}x_j)\right|$$

的整數。定義 $J_{\xi,m} = J_{\xi,m-1} \bigcup \{j_{\xi,m}\}$，

$y_{J_{\xi,m}}(\mathbf{x})$ 爲將 $E(y(\mathbf{x})-\sum_{j\in J_{\xi,m}}\lambda_j x_j)^2$ 最小化的最佳線性預測子(predictor)，且 $u_m = y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x})$ 。

中止：若 m 未達事先設定的遞迴次數上界，令 $m = m+1$ 並回到上一步；若 m 已達上界，輸出 $J_{\xi,m}$（入選變數指標構成的集合）及 $y_{J_{\xi,m}}(\mathbf{x})$（ $y(\mathbf{x})$ 的預測子）。

當 $\xi = 1$ 時，母體 WOGA 則被稱爲母體 OGA。在如下的弱稀疏性條件成立時，

$$\sum_{j=1}^{p}\left|\beta_j\sigma_j\right| < M, \tag{2}$$

其中 M 爲一有限正數，Temlyakov [8]證明了

$$\mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 < C_\xi m^{-1}, \tag{3}$$

其中 C_ξ 是隨 ξ 遞增而遞減的正數。(3)式十分容易解釋，它說明了母體 WOGA 的收斂速度取決於其遞迴次數與貪婪的程度。Gao, Ing 及 Yang [11]考慮較(2)式更一般的弱稀疏條件，

$$\sum_{j=1}^{p}\left|\beta_j\sigma_j\right|^{1/\gamma} < M, \tag{4}$$

其中 M 爲一有限正數且 $\gamma \geq 1$。當 $\gamma = 1$ 時，(4)式與(2)式相同，當 $\gamma > 1$ 時，(4)式可用來描述比(2)式更稀疏的模型，且 γ 愈大模型愈稀疏。Gao, Ing 及 Yang [11]在(4)成立的前提下證明了

$$\mathrm{E}(y(\mathbf{x}) - y_{J_{\xi,m}}(\mathbf{x}))^2 < C_{1,\xi} m^{-2\gamma+1}, \tag{5}$$

其中 $C_{1,\xi}$ 是隨 ξ 遞增而遞減的正數。相較於(3)式，(5)式更進一步地說明了模型愈稀疏母體 WOGA 的收斂速度愈快。事實上，Ing 及 Lai [10]設計了一個比(4)式更一般的稀疏條件：存在 $\gamma \geq 1$ 及 $0 < C_\gamma < \infty$ 使得對任何 $J \subseteq \{1, \cdots, p\}$

$$\sum_{j\in J}|\beta_j\sigma_j| \leq C_\gamma\{\sum_{j\in J}(\beta_j\sigma_j)^2\}^{(\gamma-1)/(2\gamma-1)}, \tag{6}$$

並在此條件下證明(5)式依舊成立，他們甚至證明了(5)式右邊的收斂速度與最佳 m 項逼近(best m-term approximation)的收斂速度相同。因好的收斂速度意味著顯著係數會被很快挑選到，以上結果讓我們相信(W)OGA 確實是執行大海撈針這項艱鉅工作的最佳利器之一。

三、樣本 OGA 及 HDIC

在實際問題中，模型(1)不僅伴隨一隨機干擾項且變數的（共）變異數未知，故母體 OGA 無法執行。然而透過收集觀測值，我們仍能以估計的參數來取代未知參數，進而執行樣本 OGA（以下逕稱 OGA）。更精確地說，我們考慮以下迴歸模型，

$$y_t = \sum_{j=1}^{p}\beta_j x_{tj} + \varepsilon_t := y(\mathbf{x}_t) + \varepsilon_t, t = 1, 2, \ldots, n, \tag{7}$$

其中 $\mathbf{x}_t = (x_{t1}, \ldots x_{tp})'$ 是 $\mathbf{x}$（定義於(1)）的獨立拷貝(independent copy)，n 是觀測值個數，ε_t 是獨立且同態(i.i.d.)的隨機干擾，且對所有 t 滿足 $\mathrm{E}(\varepsilon_t) = 0$ 及 $E(\varepsilon_t^2) = \sigma^2 > 0$. 令 $\mathbf{Y}_n = (y_1, \ldots, y_n)'$ 及 $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})'$, $1 \leq j \leq p$ ，OGA 執行的方式如下：

起始：$\hat{J}_0 = \varnothing; \hat{y}_{\hat{J}_0}(\mathbf{x}) = 0; \widehat{\mathbf{U}}_0 = \mathbf{Y}_n.$

遞迴：令 $\hat{j}_m = \arg\min_{1\leq j\leq p}\left|\mathbf{Y}_n'\mathbf{X}_j / \left\|\mathbf{X}_j\right\|\right|, \hat{J}_m = \hat{J}_{m-1}\cup\{\hat{j}_m\}, \widehat{\mathbf{U}}_m = (I - M_{\hat{J}_m})\mathbf{Y}_n$，且 $\hat{y}_{\hat{J}_m}(\mathbf{x}) = \mathbf{x}'(\hat{J}_m)\hat{\beta}(\hat{J}_m)$ ，其中 $M_{\hat{J}_m}$ 是 $\mathrm{span}\{\mathbf{X}_j, j \in \hat{J}_m\}$ 上的正交投影矩陣， $\mathbf{x}(\hat{J}_m) = (x_j, j \in \hat{J}_m)$ ，而

$$\hat{\beta}(\hat{J}_m) = (\sum_{t=1}^{n}\mathbf{x}_t(\hat{J}_m)\mathbf{x}_t'(\hat{J}_m))^{-1}\sum_{t=1}^{n}\mathbf{x}_t(\hat{J}_m)y_t,$$

其中 $\mathbf{x}_t(\hat{J}_m) = (x_{tj}, j \in \hat{J}_m)$ 。

中止：若 m 未達事先設定的遞迴次數上界，令 $m = m+1$ 並回到上一步；若 m 已達上界，輸出 $\hat{J}_m$ （入選變數的指標構成的集合）及 $\hat{y}_{\hat{J}_m}(\mathbf{x})$ （ $y(\mathbf{x})$ 的預測子）。

雖然， $\hat{y}_{\hat{J}_m}(\mathbf{x})$ 及上一節中的 $y_{J_{\xi,m}}(\mathbf{x})$ 都是 $y(\mathbf{x})$ 的預測子，但唯有前者可實際使用。另一方面相較於後者，前者多了估計及選模的誤差（不確定性），這些誤差將如何反應到預測誤差上呢？Ing 及 Lai 在[9]中回答了此一問題。假設稀疏條件(2)成立且 $p >> n$，在 $p = o(\exp(n))$ 及 ε_t

是 sub-exponential 分配等的限制下，他們證明了若遞迴次數滿足

$$m = O(\sqrt{n/\log p}), \tag{8}$$

則

$$\frac{E[\{y(\mathbf{x}) - \hat{y}_{\hat{J}_m}(\mathbf{x})\}^2 \mid y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]}{m^{-1} + n^{-1} m \log p} = O_p(1). \tag{9}$$

對照(9)及(3)式，我們不難看出 OGA 所涉及的估計及選模誤差可由 $n^{-1}m\log p$ 來表示。其中 m 反應的是估計誤差而 $\log p$ 則是搜尋（選擇）變數時付出的代價，p 愈大則搜尋的代價愈高。所幸的是，我們要求 $p = o(\exp(n))$ 及(8)成立，故 $n^{-1}m\log p$ 仍會隨 $n \to \infty$ 而趨近於 0。Ing 及 Lai [10]在更一般的稀疏條件(4)（或(6)）下證明了，

$$\frac{E[\{y(\mathbf{x}) - \hat{y}_{\hat{J}_m}(\mathbf{x})\}^2 \mid y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]}{m^{-2\gamma+1} + n^{-1} m \log p} = O_p(1). \tag{10}$$

(10)式與(9)式一樣，多出一項（相對於(5)式），$n^{-1}m\log p$，反應估計及選模誤差。

讓(10)式分母中的兩項相等,我們得到預測誤差最佳的收斂速度 $(\log p/n)^{1-(1/2\gamma)}$，此時 $m \asymp (n/\log p)^{1/2\gamma}$ 是 OGA 最佳遞迴次數(停止規則)。然而，由於模型的稀疏狀況未知（γ 未知），吾人無法藉由此一停止規則得到最小（常數項不計）預測誤差。Ing 及 Lai [9]因此藉由高維訊息準則(HDIC)來達到此一目的。對模型 $J \subset \{1, ..., p\}$，HDIC 之值定義如下：

$$\mathrm{HDIC}(J) = n\log\hat{\sigma}_J^2 + \#(J)s\log p, \tag{11}$$

其中 $\hat{\sigma}_J^2$ 爲模型 J 的殘差均方(residual mean square)而 s 爲一正值並容許隨 n 改變。HDIC 與傳統的訊息準則最大不同之處在於多了一個懲罰項 $\log p$，它是用來防止在高維環境下，由於過多變數產生的虛假相關(sporious correlation)所導致的模型誤判。定義

$$\hat{m} = \arg\min_{1 \le m \le K_n} \mathrm{HDIC}(\hat{J}_m),$$

其中 $K_n \asymp \sqrt{n/\log p}$，Ing 及 Lai [10]證明了當 s 爲一固定正數且滿足適當條件時，

$$\frac{E[\{y(\mathbf{x}) - \hat{y}_{\hat{J}_{\hat{m}}}(\mathbf{x})\}^2 \mid y_1, \mathbf{x}_1, \cdots, y_n, \mathbf{x}_n]}{(\log p/n)^{1-(1/2\gamma)}} = O_p(1). \tag{12}$$

(12)式說明了，不管模型的稀疏程度爲何（γ 有多大），HDIC 皆能自動校正遞迴次數，使得對應的預測誤差有最佳的收斂速度。這裡另外值得一提的是 $(\log p/n)^{1-(1/2\gamma)}$ 亦爲一 minimax 最佳速度，見[12]。

四、強稀疏性與 OGA+HDIC+Trim

令 $N_n = \{j : 1 \le j \le p, \beta_j \ne 0\}$。當強稀疏性成立時，意即

$$\#(N_n) < M, \tag{13}$$

其中 M 爲一有限正數，我們很自然會問，是否 OGA 選出的變數會包含 N_n？Ing 及 Lai [9]回答了這個問題，他們證明當 $K_n \asymp \sqrt{n/\log p}$ 時，

$$\lim_{n\to\infty} P(N_n \subseteq \hat{J}_{K_n}) = 1, \tag{14}$$

這個性質被稱爲幾乎確定篩入性(sure screening property)。然而，(14)式無法保證 $\hat{J}_{K_n}$ 不會納入過多的多餘變數，Ing 及 Lai [9]因此建議用 HDIC 來決定 OGA 的遞迴次數，以防止產生過度配適(overfitting)的模型。定義 $\tilde{k} = \min\{k : 1 \le k \le K_n, N \subseteq \hat{J}_k\}$ 若 $N \subseteq \hat{J}_{K_n}$，$\tilde{k} = \infty$ 若 $N \not\subseteq \hat{J}_{K_n}$。Ing 及 Lai [9]證明了當(11)中的 s 隨著 n 緩慢地趨近∞時，

$$\lim_{n\to\infty} P(\hat{J}_{\tilde{k}} = \hat{J}_{\hat{m}}, \tilde{k} \le K_n) = 1. \tag{15}$$

(15)式指出當(13)式成立時，HDIC 將協助 OGA 停在所有非零係數都被納入後的最小遞迴次數，因此相當程度地紓解了過度配適的問題。可惜美中不足的是，(15)無法保證 $\lim_{n\to\infty} P(N_n = \hat{J}_{\hat{m}}) = 1$（這個性質被稱爲選模的一致性），因爲 OGA+HDIC 無法排除在遞迴程序中早期就被選入的多餘變數。爲了排除這樣的變數達到選模的一致性，Ing 及 Lai [9]建議了一個更爲精鍊的集合

表一　四種選模方法在 1000 模擬中正確選中 $\beta_1, ..., \beta_{10}$ 的次數(E)，除 $\beta_1, ..., \beta_{10}$ 外多選入 i 個係數的次數(E+i)，及其預測均方差(MSPE).

Method	E	E+1	E+2	E+3	E+i, $i > 3$	MSPE
OGA+HDIC	0	39	945	16	0	0.035
OGA+HDIC+Trim	1000	0	0	0	0	0.028
Adaptive Lasso	0	0	0	0	0	27.27
Lasso	0	0	0	0	0	2.283

$$\hat{N}_n = \{j : 1 \le j \le \hat{m}, \mathrm{HDIC}(\hat{J}_{\hat{m}}) < \mathrm{HDIC}(\hat{J}_{\hat{m}} - \{j\})\}, \quad (16)$$

並證明了 $\lim_{n\to\infty} P(N_n = \hat{N}_n) = 1$ 。

以下用模擬試驗進一步闡明此一三階段選模法，OGA+HDIC+Trim，的優點。設定高維迴歸模型(7)中的 $n = 400$ ， $p = 4000$ ，且非零係數爲 $(\beta_1, \cdots, \beta_{10})$= (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)。並令 $\sigma = 1$ ， $x_{t1}, \cdots, x_{t10}$ 爲 i.i.d.標準常態，及

$$x_{tj} = d_{tj} + b\sum_{l=1}^{q} x_{tl},\ q+1 \le j \le p, \quad (17)$$

其中 $b = (3/40)^{1/2}$ 而 $(d_{t(q+1)}, \cdots, d_{tp})^\top$ 爲與 x_{tj}, $1 \le j \le 10$ 獨立的 i.i.d.多元常態向量滿足期望向量爲 **0** 和共變異矩陣爲 $(1/4)\mathbf{I}$ 。Ing 及 Lai [9]證明了這些變數間相關性的設定使得 neighborhood stability 條件不成立，因此 Lasso 不具選模一致性。然而，Ing 及 Lai [9]也證明了 OGA 選入的第一個變數是多餘的。如前所述，這個變數無法被 HDIC 刪除。除了 Lasso 及 OGA+HDIC 外，我們也觀察了 OGA+HDIC+Trim 及 Adaptive Lasso 的表現。Adaptive Lasso 是由 Zou [13]率先提出，當 $p >> n$ 時，Adaptive Lasso 通常用原始 Lasso 的估計值來決定第二階段加權 Lasso 的權數，並可在較弱的條件下獲得選模一致性。我們重複了 1000 次選模實驗，並將上述方法正確選中 $\beta_1, \ldots, \beta_{10}$ 的次數(E)， 以及除 $\beta_1, \ldots, \beta_{10}$ 外多選入 i 個係數的次數(E+i)紀錄於表一。爲了更完整的比較，我們也記下了這些選模方法的預測均方差(mean squared prediction error, MSPE)。

從表一中可看出，在 1000 次重複實驗中，OGA+HDIC 的確都無法剛好選入 $\beta_1, \ldots, \beta_{10}$ ，但它可含括所有重要係數（拜幾乎確定篩入性之賜）且至多選進三個多餘係數。另一方面，Lasso 不僅不具一致性，模擬顯示，它在此例中似乎亦不具幾乎確定篩入性，因每次模擬至少有一重要係數未被選入。當 Lasso 表現不佳時，以它作爲起始估計的 adaptive Lasso 表現似乎更爲捉襟見肘，因後者的 MSPE 大過前者約 11 倍。最後，Trim 能準確地刪除 OGA+HDIC 所納入的多餘係數，使得 OGA+HDIC+Trim 在 1000 次模擬中都能剛好選入 $\beta_1, \ldots, \beta_{10}$ 。

五、結　論

本文介紹了以 OGA 及 HDIC 爲基礎的兩(三)階段選模法，並略述其優點。然而，此法在：(1)高維時間序列模型，(2)高維非線性模型，及(3)高維誤設(misspecified)模型，之表現仍未被釐清，值得進一步探討。

參考文獻

[1] E. J. Candès, J. Romberg and T. Tao, *Comm. Pure Appl. Math.*, **59**, 1207 (2006).

[2] E. J. Candès and T. Tao, *IEEE Trans. Inf. Theory*, **51**, 4203 (2005).

[3] E. J. Candès and T. Tao, *Ann. Statist.*, **35**, 2013 (2007).

[4] R. Tibshirani, J. Roy. *Statist. Soc. B*, **58**, 267 (1996).

[5] P. Zhao and B. Yu, *J. Machine Learning Res.*, **7**, 2541 (2006).

[6] N. Meinshausen and P. Bühlmann, *Ann. Statist.*, **34**, 1436 (2006).

[7] P. Bickel, Y. Ritov and A. Tsybakov, *Ann. Statist.*, **37**, 1705 (2009).

[8] V. N. Temlyakov, *Adv. Comput. Math.*, **12**, 213 (2000).

[9] C.-K. Ing and T. L. Lai, *Statist. Sinica*, **21**, 1473 (2011).
[10] C.-K. Ing and T. L. Lai, *Manuscript*, 2015.
[11] F. Gao, C.-K. Ing and Y. Yang, *J. Approx. Theory*, **166**, 42 (2013).
[12] Z. Wang, S. Paterlini, F. Gao and Y. Yang, *J. Machine Learning Res.*, **15**, 1675 (2014).
[13] H. Zou, *J. Amer. Statist. Assoc.*, **101**, 1418 (2006).