1 LARS

Il LARS (Least Angle Regression) è una versione semplificata della procedura Stagewise la quale usa una semplice formula matematica per accelerare il calcolo computazionale. In soli m passi si giunge all'insieme delle soluzioni, dove m è il numero delle covarianze.
Come nel più classico algoritmo Forward Selection si parte con tutti i coefficienti della stima β uguali a zero e si trova il predittore x_{j_1} maggiormente correlato con la risposta. Si cerca di muoversi lungo la direzione di x_{j_1} il più possibile finché non subentra un altro predittore x_{j_2} che ha maggior correlazione con il residuo. Da qui in poi il LARS procede lungo una nuova direzione equiangolare rispetto a x_{j_1} e x_{j_2}. Si percorre tale direzione finché un nuovo predittore x_{j3} diventa il più correlato, a questo punto la direzione che viene presa è equiangolare rispetto a x_{j_1}, x_{j_2} e x_{j_3}. Ovviamente si procede con questa iterazione di k passi di volta in volta avanzando lungo una direzione equiangolare ai predittori precedentemente individuati e aggiornando la direzione nel momento in cui si trova un predittore maggiormente correlato con il residuo.
Si darà ora una veloce descrizione geometrica di questo processo. Con il LARS si cerca di dare una stima di $\widehat{\mu} = X\widehat{\beta}$ in una successione di passi, aggiungendo ad ogni passaggio una covarianza del modello, in maniera che dopo k passi il vettore $\widehat{\beta}_j$ abbia k componenti non nulle.

Figura 1.1:

In figura 1 viene illustrato il caso con due covarianze $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$ cioè con m=2. In questo caso la correlazione corrente diventa:

$$\mathbf{c}(\hat{\mu}) = X'(\mathbf{y} - \hat{\mu}) = X'(\bar{\mathbf{y}}_\mathbf{2} - \hat{\mu}) \tag{1.1}$$

dimostrando una dipendenza solo da $\bar{\mathbf{y}}_2$ nello spazio lineare $\mathcal{L}(X)$ generato da $\mathbf{x}_1$ e $\mathbf{x}_2$. Come precedentemente spiegato si parte con $\hat{\mu}_0 = \mathbf{0}$. In figura 1 si ha che $\bar{\mathbf{y}}_2 - \hat{\mu}_0$ produce un angolo più piccolo con $\mathbf{x}_1$ che con $\mathbf{x}_2$ ovvero si ha $c_1(\hat{\mu}_0) > c_2(\hat{\mu}_0)$; per cui il LARS aumenta $\widehat{\mu}_0$ nella direzione di $\mathbf{x}_1$ ottenendo:

$$\hat{\mu}_\mathbf{1} = \hat{\mu}_\mathbf{0} + \widehat{\gamma}_1 \mathbf{x}_1. \tag{1.2}$$

A questo punto Stagewise sceglierebbe $\widehat{\gamma}_1$ uguale ad un qualche velore ϵ (molto piccolo), e poi ripetere il processo molte volte. Il classico Forward Selection prenderebbe $\widehat{\gamma}_1$ sufficientemente grande per rendere $\hat{\mu}_1$ equivalente a $\bar{\mathbf{y}}_1$, la proiezione di $\mathbf{y}$ su $\mathcal{L}(\mathbf{x}_1)$. Invece LARS usa un valore intermedio di $\widehat{\gamma}_1$, il valore che rende $\bar{\mathbf{y}}_2 - \hat{\mu}$ ugualmente correlato con $\mathbf{x}_1$ e $\mathbf{x}_2$, in maniera che $\bar{\mathbf{y}}_2 - \hat{\mu}_1$ sia la retta secante l'angolo compreso fra $\mathbf{x}_1$ e $\mathbf{x}_2$, ottenendo $c_1(\hat{\mu}_1) = c_2(\hat{\mu}_1)$.
Sia $\mathbf{u}_2$ il versore della bisettrice, il nuovo passo del LARS sarà:

$$\hat{\mu}_{\mathbf{2}} = \hat{\mu}_{\mathbf{1}} + \widehat{\gamma}_2 \mathbf{u}_2, \tag{1.3}$$

con $\widehat{\gamma}_2$ scelta per rendere $\widehat{\mu}_2 = \bar{\mathbf{y}}_2$ nel caso di m=2. Con più di 2 covarianze $\widehat{\gamma}_2$ sarebbe più piccola in conformità con un altro cambio di direzione come illustrato in figura 1.

Figura 1.2: Sc

Da questo processo si evince che il calcolo teorico di $\widehat{\gamma}_1$, $\widehat{\gamma}_2$, ..., $\widehat{\gamma}_m$ sia molto più veloce per il LARS rispetto al calcolo con l'algoritmo Stagewise che avanza solo per piccoli passi.
Si presume ora di assumere i vettori covarianze $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_m$ linearmente indipendenti. Per $\mathcal{A}$, un sottoinsieme degli indici $\{1, 2, \ldots, m\}$, si definisce la matrice:

$$\mathbf{X}_{\mathcal{A}} = (\ldots s_j \mathbf{x}_j \ldots)_{j \in \mathcal{A}} \tag{1.4}$$

dove la variabile segno s_j può assumere i valori ± 1. Si prende

$$\mathcal{G}_{\mathcal{A}} = X'_{\mathcal{A}} X_{\mathcal{A}} \quad \text{con} \quad \mathcal{A}_{\mathcal{A}} = (1'_{\mathcal{A}} \mathcal{G}_{\mathcal{A}} 1_{\mathcal{A}})^{-\frac{1}{2}} \tag{1.5}$$

dove $1_{\mathcal{A}}$ è un vettore di 1 con lunghezza pari $|\mathcal{A}|$, la dimensione di $\mathcal{A}$.
Sia

$$\mathbf{u}_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}} \quad \text{con} \quad w_{\mathcal{A}} = \mathcal{A}_{\mathcal{A}} {\mathcal{G}_{\mathcal{A}}}^{-1} 1_{\mathcal{A}} \tag{1.6}$$

il vettore equiangolare ovvero il vettore unitario che rende gli angoli (minori di 90°) con le colonne della matrice $\mathcal{A}$ tutti uguali,

$$X'_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} = \mathcal{A}_{\mathcal{A}} 1_{\mathcal{A}} \quad \text{e} \quad ||\mathbf{u}_{\mathcal{A}}||^2 = 1. \tag{1.7}$$

Ora descriveremmo i passi dell'algoritmo LARS. Si inizia sempre con $\widehat{\mu}_0 = \mathbf{0}$ e si costruisce $\widehat{\mu}$ cercando di percorrere passi più grandi rispetto a quelli compiti dal LARS. Supponendo che $\widehat{\mu}_{\mathcal{A}}$ sia la stima corrente con

$$\hat{\mathbf{c}} = X'(\mathbf{y} - \hat{\mu}_{\mathcal{A}}) \tag{1.8}$$

il vettore delle correlazioni correnti come definito dalla generica $\widehat{\mathbf{c}} = \mathbf{c}(\widehat{\mu}) = X'(\mathbf{y} - \widehat{\mu})$. La selezione attiva degli indici $\mathcal{A}$ corrisponde con le covarianze che hanno in assoluto il valore maggiore della correlazione corrente,

$$\widehat{C} = \max_j |\hat{c}_j| \quad \text{e} \quad \mathcal{A} = \{j : |\widehat{c}_j| = \widehat{C}\}. \tag{1.9}$$

Prendendo

$$s_j = sign\{\widehat{c}_j\} \quad \text{con} \quad j \in \mathcal{A} \tag{1.10}$$

si calcola $X_{\mathcal{A}}$, $A_{\mathcal{A}}$ ed $\mathbf{u}_{\mathcal{A}}$, e si calcola il prodotto interno

$$\mathbf{a} \equiv X'\mathbf{u}_{\mathcal{A}}. \tag{1.11}$$

Al prossimo passo l'algoritmo LARS aggiorna $\widehat{\mu}_{\mathcal{A}}$, definita come

$$\widehat{\mu}_{\mathcal{A}_+} = \widehat{\mu}_{\mathcal{A}} + \widehat{\gamma}\mathbf{u}_{\mathcal{A}} \tag{1.12}$$

dove

$$\widehat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+}\{\frac{\widehat{C} - \widehat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\widehat{C} + \widehat{c}_j}{A_{\mathcal{A}} + a_j}\} \tag{1.13}$$

con $= \min^+$ a significare che il minimo viene preso solo tra i componenti positivi delle scelte j in 1.13. Usando le formule 1.12 e 1.13 si definisce

$$\mu(\gamma) = \widehat{\mu}_{\mathcal{A}} + \gamma\mathbf{u}_{\mathcal{A}} \tag{1.14}$$

con $\gamma > 0$, in maniera che la correlazione corrente sia

$$c_j(\gamma) = \mathbf{x}_j'(\mathbf{y} - \mu(\gamma)) = \widehat{c}_j - \gamma a_j. \tag{1.15}$$

Per $j \in \mathcal{A}$ usando 1.7 1.9

$$|c_j(\gamma)| = \widehat{C} - \gamma A_{\mathcal{A}} \tag{1.16}$$

mostrando che tutte le correlazioni correnti declino nella stessa maniera. Per $j \in \mathcal{A}^c$ uguaglio l'equazioni 1.15 con la 1.16, ottenendo che $c_j(\gamma)$ è uguale al massimo valore di $\gamma = (\widehat{C}_j - \widehat{c}_j)/(A_{\mathcal{A}} - a_j)$; alla stessa maniera per $-c_j(\gamma)$ la correlzione corrente per la covarianza inversa $-\mathbf{x}_j$ raggiunge il massimo per $(\widehat{C}_j + \widehat{c}_j)/(A_{\mathcal{A}} + a_j)$.
Quindi la $\widehat{\gamma}$ in 1.13 è il più piccolo valore positivo di γ tale che il nuovo indice $\widehat{j}$ venga aggiunto all'insieme attivo degli indici; $\widehat{j}$ è indice che minimizza la 1.13,