UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI PIANO DEGLI ESPERIMENTI II

PROBLEMI DI ALLOCAZIONE OTTIMA

Alberto Cavalin

ANNO ACCADEMICO 2006–2007

Indice

Introduzione

Se per la stima di due parametri β_1, β_2 si hanno a disposizione diverse potenziali osservazioni (dette "sorgenti") del tipo $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \eta_i$, ognuna delle quali è indefinitamente ripetibile, il problema che sorge per lo sperimentatore è quello di decidere quali di queste utilizzare ed in quali proporzioni.

Sotto opportune condizioni di ottimalità, la soluzione è la seguente: per la stima di una singola quantità della forma $\theta = \alpha_1\beta_1 + \alpha_2\beta_2$, per l'allocazione dell'ottimo servono solo due sorgenti; per la stima di entrambi i parametri ne servono invece due o tre.

Capitolo 1

Formalizzazione del problema

Si consideri uno sperimentatore che vuole determinare due quantità sconosciute β_1, β_2, e si assuma per questo scopo che si abbiano a disposizione r differenti potenziali osservazioni, il cui esito è del tipo:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \eta_i \ , \qquad (i = 1, \ldots, r) \tag{1.1}$$

dove x_{i1}, x_{i2} sono coefficienti noti, e η_i l'errore a media nulla e varianza σ^2. Si assuma inoltre che le osservazioni siano tra loro incorrelate, e che lo sperimentatore possa effettuare ciascuna di esse in una qualsiasi quantità (0, 1, ...).

Fissato un numero $n \leq r$ di osservazioni da effettuare, sorge poi il quesito su quali scegliere tra le possibili r ed in quale quantità. Per distinguere le potenziali osservazioni da quelle attualmente effettuate, verranno indicate come *sorgenti* le prime e come *osservazioni* le seconde. Le sorgenti che vengono utilizzate nella soluzione di un qualunque problema di allocazione ottima vengono dette *rilevanti*, le rimanenti *irrilevanti*.

Una sorgente è essenzialmente descritta dai coefficitenti del vettore $x_i = (x_{i1},\ x_{i2})$, perciò la si chiamerà sorgente x_i, o più semplicemente sorgente i-esima.

Sia $n_i = np_i$ il numero di osservazioni allocate per l'i-esima sorgente, dove i p_i sono dei multipli di $1/n$ che soddisfano le condizioni:

$$p_i \geq 0, \qquad \sum p_i = 1 \tag{1.2}$$

La media delle osservazioni dell'i-esima sorgente assume la forma:

$$\overline{y}_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \frac{\overline{\eta}_i}{\sqrt{p_i}}\,, \qquad i \in \{1, \ldots, r\} \tag{1.3}$$

dove il termine d'errore $\overline{\eta}_i/\sqrt{p_i}$ ha varianza σ^2/n_i, e quindi $\overline{\eta}_i$ ha varianza σ^2/n. Se un certo p_i è nullo, la corrispondente equazione sarà lasciata fuori dal sistema.

Per n grande, i p_i possono variare in modo praticamente continuo nell'insieme (1.2), definendo così un problema di grande campionamento indipendete da σ ed n; si assumerà inoltre per semplicità $\sigma^2/n = 1$.

Con queste assunzioni il problema diventa ora il seguente: *"dato un criterio di ottimalità ed un insieme predefinito di osservazioni del tipo (1.3), dove* $\mathbf{E}(\overline{\eta}_i) = 0$ *e* $\mathbf{Var}(\overline{\eta}_i) = 1$*, quali sono i valori ottimali da assegnare ai pesi* p_i*?"*.

Capitolo 2

Stima di una singola quantità

Si vuole stimare una combinazione lineare dei parametri del tipo:

$$\theta = a_1\beta_1 + a_2\beta_2 \tag{2.1}$$

Se si considerano tutte le combinazioni lineari $t = \sum c_i \overline{y}_i$ che portano a stime non distorte di θ, allora esistono infiniti set di c_i che soddisfano la condizione:

$$\mathbf{E}(t) = \sum c_i(x_{i1}\beta_1 + x_{i2}\beta_2) = a_1\beta_1 + a_2\beta_2 \qquad \Leftrightarrow \qquad \sum c_i \underline{x}_i = \underline{a} \tag{2.2}$$

inoltre, ***fissato*** un set di p_i ed utilizzando le stime ai minimi quadrati pesati $\hat{\beta}_1$ e $\hat{\beta}_2$ che minimizzano la quantità:

$$\sum p_i(\overline{y}_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \tag{2.3}$$

è possibile individuare una stima t a varianza minima. Questa stima è perciò funzione dei pesi p_i, e si vuole ora cercare un set di quest'ultimi tale da trovare la *più piccola varianza minima*: $\min_p \min_c \mathbf{Var}(t)$.

Invertendo l'ordine delle minimizzazioni, si trova facilmente che:

$$\min_p \mathbf{Var}\left(\sum c_i \overline{y}_i\right) = \min_p \sum c_i^2/p_i = \left(\sum |c_i|\right)^2 = k_c^2 \qquad (2.4)$$

dove per i pesi p_i vengono utilizzati i valori $p_{ci} = |c_i|/k_c$.

Ora resta da minimizzare la (2.4) rispetto ai c_i. Riscrivendo equivalentemente la condizione (2.2) nel modo seguente:

$$\underline{a} = \sum c_i \underline{x}_i = k_c \sum p_{ci} \cdot \text{sgn}(c_i) \cdot \underline{x}_i = k_c \underline{a}_c \qquad (2.5)$$

ed osservando che (vedi fig. 2.1):

- i fattori k_c sono tutti positivi $\Rightarrow$ $\underline{a}_c$ ha la stessa direzione di $\underline{a}$
- $p_{ci} \geq 0$, $\forall i$ e $\sum p_{ci} = 1$ $\Rightarrow$ $\underline{a}_c$ è contenuto nel poligono convesso Π
- k_c è praticamente il rapporto tra le lunghezze di $\underline{a}$ ed $\underline{a}_c$

allora la (2.4) è minina quando $\underline{a}_c$ termina in A^*, cioè l'intersezione tra Π e $\underline{a}$ (o la sua estensione).

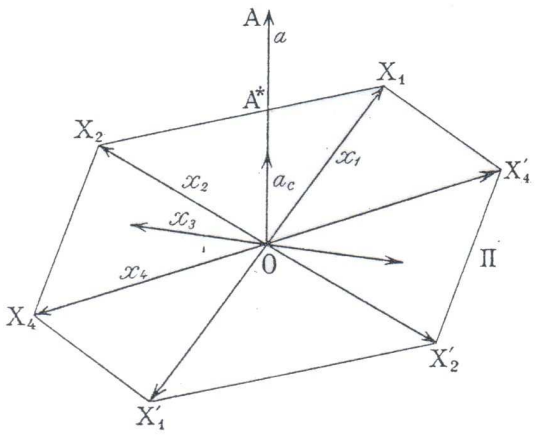

Figura 2.1: Rappresentazione geometrica dei vettori.

Se, ad esempio, $A^* \in \overline{X_1 X_2}$ allora i pesi ottimi p_{ci} per la (2.5) saranno proporzionali ai segmenti $\overline{A^* X_1}$, $\overline{A^* X_2}$ per $i = 1, 2$, e varranno zero per $i = 3, \ldots, r$. La più piccola varianza minima sarà data da $(\overline{OA}/\overline{OA^*})^2$. Si noti inoltre che nella costruzione del poligono Π, vengono scartati i vettori che vi cadono completamente all'interno.

Con questi risultati si deduce che per la stima di una singola quantità (2.1), servono solamente due sorgenti usate nelle proporzioni sopra indicate. Generalizzando il problema a tre parametri, Π diviene un poliedro convesso a facce triangolari, e si identificano 3 sorgenti rilevanti. Per più di tre parametri sono richiesti metodi algebrici.

Capitolo 3

Stima di entrambi i parametri

Dato un set di osservazioni, cioè con p_i fissati, il metodo dei minimi quadrati genera delle stime la cui accuratezza non è controllabile. In questo contesto i p_i sono variabili, ed il metodo per gestire la precisione delle stime, consiste nel minimizzare la somma delle varianze degli stimatori:

$$q = \mathbf{E}\{(\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2\} = \mathbf{Var}(\hat{\beta}_1) + \mathbf{Var}(\hat{\beta}_2) \tag{3.1}$$

rispetto ai p_i. La matrice di covarianza $\mathbf{\Lambda}$ dei due stimatori coincide con l'inversa della matrice d'informazione $\mathbf{M}$:

$$\mathbf{M} = \begin{bmatrix} \sum p_i x_{i1}^2 & \sum p_i x_{i1} x_{i2} \\ \sum p_i x_{i1} x_{i2} & \sum p_i x_{i2}^2 \end{bmatrix} = \sum p_i \underline{x}_i' \underline{x}_i = \mathbf{\Lambda}^{-1} \tag{3.2}$$

e si vuole perciò minimizzarne la traccia $q = \mathbf{Tr}(\mathbf{\Lambda}) = \lambda_{11} + \lambda_{22}$ rispetto ai pesi.

La funzione obiettivo q delle variabili p_i ($i = 1, \ldots, r$) è caratterizzata da ammettere un punto di minimo usualmente sulla frontiera del dominio (1.2), e questo implica che alcuni p_i saranno nulli. Per ogni soluzione di questo problema, esiste una costante $-k^2$ tale che $\partial q / \partial p_i = -k^2$ per tutte le sorgenti rilevanti, e

$\partial q/\partial p_i \geq -k^2$ per quelle irrilevanti. Da queste due affermazioni, si deduce che:

$$\sum_{i=1}^{r} p_i \frac{\partial q}{\partial p_i} = -k^2 \sum_{i=1}^{r} p_i = -k^2$$

e, per l'identità di Eulero, k^2 è il valore minimo assunto da q.

Differenziando l'equazione $\mathbf{M}\mathbf{\Lambda} = \mathbf{I}$ si ottiene il seguente risultato:

$$\frac{\partial \mathbf{\Lambda}}{\partial p_i} = -\mathbf{\Lambda}\underline{x}_i'\underline{x}_i\mathbf{\Lambda} = -(\mathbf{\Lambda}\underline{x}_i')(\mathbf{\Lambda}\underline{x}_i')'$$

il quale torna utile per il calcolo della generica derivata parziale di q:

$$\frac{\partial q}{\partial p_i} = \frac{\partial \mathbf{Tr}(\mathbf{\Lambda})}{\partial p_i} = -\mathbf{Tr}\left[(\mathbf{\Lambda}\underline{x}_i')(\mathbf{\Lambda}\underline{x}_i')'\right] = -||\mathbf{\Lambda}\underline{x}_i'||^2 \tag{3.3}$$

dove $||\cdot||$ corrisponde alla norma euclidea. Si noti che $||\mathbf{\Lambda}\underline{x}_i'||^2$ è una forma quadratica definita positiva rispetto alle componeti di $\underline{x}$, e ponendola pari ad una costante, essa delinea una ellisse centrata nell'origine.

Combinando tutti i risultati fin qui incontrati, è ora possibile enunciare il seguente teorema.

TEOREMA 1 *Ad ogni set di p_i che minimizza la funzione (3.1) corrisponde un'ellisse centrata nell'origine, tale che i punti delle sorgenti rilevanti cadono sulla sua frotiera, mentre i rimanenti non cadono al di fuori di essa.*

Sono sufficienti tre punti per identificare una conica centrata nell'origine, ed è quindi possibile affermare che servono altrettante sorgenti per minimizzare q.

È possibile generalizzare il problema con un numero s arbitrario di parametri, utilizzando un iperelissoide in $\mathbb{R}_s$, ed ottenendo così al più $s(s+1)/2$ sorgenti rilevanti[1].

[1] Incontrando purtroppo grossi problemi computazionali per $s \geq 3$

Capitolo 4

Come scegliere i pesi

È dimostrato che casi con due o tre sorgenti rilevanti accadono realmente. Si supponga per ora di sapere come selezionare queste sorgenti, e si voglia trovare la distribuzione dei pesi ed il minimo valore di q.

4.1 Caso con 2 sorgenti rilevanti

Avendo due sorgenti rilevanti, ad esempio $i = 1, 2$, le stime di β_1 e β_1 si trovano risolvendo il sistema si equazioni $\overline{y}_i = x_{i1}\beta_1 + x_{i2}\beta_2$ con $i = 1, 2$. Risolvendo tale sistema, calcolando le varianze, e trasformando in coordinate polari ρ, θ, si trova:

$$q = \mathbf{Var}(\hat{\beta}_1) + \mathbf{Var}(\hat{\beta}_2) = \frac{\rho_2^2/p_1 + \rho_1^2/p_2}{\rho_1^2\rho_2^2 \sin^2(\theta_2 - \theta_1)} \tag{4.1}$$

la quale è minima per i seguenti pesi:

$$p_1 = \rho_2/(\rho_1 + \rho_2),\ p_2 = \rho_1/(\rho_1 + \rho_2) \quad \Rightarrow \quad q_{\text{min}} = \frac{\rho_1^{-1} + \rho_2^{-1}}{\sin(\theta_2 - \theta_1)} \tag{4.2}$$

4.2 Caso con 3 sorgenti rilevanti

Avendo tre sorgenti rilevanti, ad esempio $i = 1, 2, 3$, è più conveniente lavorare con una diversa funzione obiettivo:

$$q = L/M = \left[\mathbf{Tr}(\mathbf{M}) \cdot (p_1 + p_2 + p_3)\right] / \mathbf{Det}(\mathbf{M})$$

l'interesse viene poi spostato sulla ricerca delle proporzioni da assegnare ai tre pesi, minimizzando L con il vincolo $M =$ constante. Utilizzando il teorema dei moltiplicatori di Lagrange e differenziando $L - \lambda M$ si ottiene il seguente sistema di equazioni:

$$\sum_{j=1}^{3} (l_{ij} - \lambda m_{ij}) p_j = 0 \ , \qquad (i = 1, 2, 3) \tag{4.3}$$

$$l_{ij} = r_i^2 + r_j^2 \quad , \quad m_{ij} = r_i^2 r_j^2 \sin^2(\theta_j - \theta_i)$$

Se λ è un autovalore del sistema (4.3) e P il corrispondete autovettore, allora L/M in P vale λ. Se ne deduce che q_{min} coincide col più piccolo autovalore di (4.3), per il quale tutte le componenti del corrispondete autovettore sono del medesimo segno. I pesi ottimali sono dati da queste componenti normalizzate in modo che sommino ad uno.

Capitolo 5

Come selezionare le sorgenti

Se si scartano le sorgenti non rilevanti utilizzando il teorema visto nel capitolo 3, è molto probabile che rimangano a disposizione più di tre sorgenti. In linea di principio è sempre possibile esaminare una tripletta di sorgenti alla volta, calcolare il valore q_{min} corrispondente come visto nel capitolo 4, ed infine scegliere quella che genera il più piccolo.

La maggior parte delle triplette si ridurrà a delle paia, due rilevanti ed una no, utilizzando il seguente criterio:
Una sorgente x_3, combinata con x_1, x_2, è irrilevante se e solo se x_3 cade all'interno dell'ellisse passante per x_1, x_2 avente equazione parametrica

$$x = \frac{x_1 \sin(t - \theta_1) + x_2 \sin(t - \theta_2)}{\sin(\theta_2 - \theta_1)} \qquad (0 \leq t \leq 2\pi) \tag{5.1}$$

Capitolo 6

Osservazioni a costi differenti

Si consideri il caso in cui le potenziali sorgenti hanno un costo unitario di osservazione $c_1, \ldots, c_r$, e debbano essere osservate con numerosità $n_1, \ldots, n_r$; se il budget a nostra disposizione è fissato ad una soglia C, si deve quindi aggiungere il vincolo $\sum n_i c_i = C$ anzichè $\sum n_i = n$. Dividendo le equazioni di regressione di $\overline{y}_i$ per $\sqrt{v_i}$ si ottiene un nuovo set di equazioni:

$$\overline{y}_i^* = x_{i1}^* \beta_1 + x_{i2}^* \beta_2 + \frac{\overline{\eta}_i^*}{\sqrt{p_i^*}} , \qquad i \in \{1, \ldots, r\} \tag{6.1}$$

con

$$\overline{y}_i^* = \overline{y}_i / \sqrt{v_i} \qquad x_{ij}^* = x_{ij} / \sqrt{v_i} \qquad p_i^* = v_i n_i / C \tag{6.2}$$

dove $\overline{\eta}_i^*$ è una variabile casuale di media nulla e varianza σ^2/C, ed i pesi p_i^* sono soggetti al vincolo (1.2). In questo modo ci si è ricondotti alla situazione di partenza e si possono applicare i passi descritti nei capitoli precedenti, tenendo però presente che i risultanti p_i^* forniscono l'allocazione ottima dei costi, non delle osservazioni.

Bibliografia

[1] G. Elfving, Optimum allocation in linear regression theory
http://links.jstor.org/