UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

DOTTORATO DI RICERCA IN STATISTICA ED APPLICAZIONI

FUNZIONALI STATISTICI NELLA CLASSE DELLE EQUAZIONI DI STIMA GENERALIZZATE

Tommaso Lando

Matricola: 707881

Relatore: Chiar.mo Prof. Lucio Bertoli-Barsotti

XXII Ciclo

Indice

Lista degli acronimi

GEE	*Generalized estimating equation* (Equazioni di stima generalizate)
MALD	*Minimum "amplified" likelihood divergence* (Minima divergenza di verosimiglianza "amplificata")
MCDD	*Minimum "depenalized" Chi^2 divergence* (Minima divergenza di Chi^2 "depenalizzata")
MDE	*Minimum disparity estimator* (Stimatore di minima disparità)
MGSP	*Maximum generalized spacing product* (Massimo prodotto di spacing generalizzato)
MHDE	*Minimum Hellinger divergencee estimator* (Stimatore della minima divergenza di Hellinger)
MHPD	*Minimum "penalized" Hellinger divergence* (Minima distanza penalizzata di Hellinger)
MLE	*Maximum likelihood estimator* (Stimatore di massima verosimiglianza)
MSP	*Maximum spacing product* (Massimo prodotto di spacing)
MSE	*Mean squared error* (Errore quadratico medio)
PWD	*Power divergence* (Divergenza potenziata)
RAF	*Residual adjustment function* (Funzione di adeguamento dei residui)

Introduzione

In statistica, gran parte delle procedure di stima si basano su un principio piuttosto semplice: si cerca di individuare la distribuzione che, all'interno di una certa classe, sia il più "simile" possibile alla distribuzione empirica (costruita con le osservazioni campionarie).

Concetti vaghi come "somiglianza" o "dissomiglianza" non sono sicuramente semplici da definire e analizzare, da un punto di vista matematico. Lo strumento ideale per affrontare tali temi è la *maggiorazione vettoriale*, pre-ordinamento grazie al quale si possono confrontare vettori con certe caratteristiche in termini di dissomiglianza.

Uno degli obiettivi principali della tesi è quello di approfondire e generalizzare il più possibile il concetto di maggiorazione, estendendolo anche a funzioni e a distribuzioni di probabilità, per poi capire sotto quali condizioni un funzionale, utilizzato per misurare la dissomiglianza, rispetti effettivamente tale pre-ordinamento. I funzionali ritenuti adatti a misurare la dissomiglianza tra distribuzioni prendono il nome di *distanze* o *divergenze*. Si desidera che questi funzionali siano *Schur-convessi*, ossia "coerenti", tali che al crescere della dissomiglianza cresca anche la divergenza.

Si cercherà quindi di definire una classe di funzionali statistici coerenti con la maggiorazione tra distribuzioni. Tali misure di divergenza potranno essere utilizzate nella stima: si cercherà di individuare la distribuzione con la minore divergenza dalla distribuzione empirica: da questo metodo si otterranno equazioni di stima nella classe *GEE* (*generalized estimating equation*) e diversi stimatori, con proprietà desiderabili quali consistenza, efficienza e robustezza.

Lo scopo di questo studio è, oltre a quello di inquadrare da un punto di vista teorico comune la maggior parte dei metodi di stima conosciuti, quello di avere una visone generale piuttosto ampia, che permetta di risolvere in maniera adeguata il problema della stima in situazioni particolari.

Ad esempio si vedrà che, nel caso di campioni non sufficientemente numerosi, lo studio della dissomiglianza tramite la maggiorazione risulta meno chiaro e più complicato: questo suggerirà di modificare le più note misure di divergenza per cercare di migliorare i risultati delle stime. Tali divergenze "modificate" verranno

proposte nel capitolo *6*, prendendo spunto anche dalla *divergenza di Hellinger penalizzata* (Basu, Harris, 1997).

Nel capitolo *7* invece si affronterà, come caso di studio, il modello di *Rasch*. In primo luogo si confronteranno le stime corrispondenti a due delle principali divergenze: stima di massima verosimiglianza e di minimo Chi^2 (senz'altro insolita per questo modello). Si cercherà di dare un'interpretazione alla stima di Chi^2 e di spiegarne la logica, anche da un punto di vista matematico. In secondo luogo si cercherà di risolvere alcuni problemi di stima del modello Rasch legati a due casi particolari: "punteggi estremi" e "mal-condizionamento" delle matrici dei dati. In tali casi, infatti, i più noti metodi di stima falliscono, poiché le stime divergono. Nel programma *Winsteps* ® (Linacre, 2009) è disponibile un'opzione per risolvere i problemi di stima nel caso di punteggi estremi. Prendendo spunto da questa soluzione, ne verrà proposta una variante per ottenere risultati accettabili anche nel caso di matrici mal-condizionate. Inoltre verrà proposta una soluzione alternativa, adatta in entrambi i casi, basata proprio sul concetto di "distanza" tra distribuzioni.

Si osservi che la lettura del primo capitolo necessita conoscenze matematiche piuttosto approfondite: per questo motivo, nell'appendice *A* sono riportati e riassunti i concetti ed i teoremi utilizzati.

Le appendici *B* e *C*, invece, contengono simulazioni, esemplificazioni e calcoli: i risultati riportati sono stati ottenuti con *Mathematica 5.0* (Wolfram, 2003).

Capitolo 1

Maggiorazione

1.1 Il pre-ordinamento di maggiorazione

Si supponga che su un insieme $\Im$ esista una relazione definita dal simbolo $\leq$ tale che :

$$x \leq x\,, \qquad \forall\, x \in \Im;$$

$$x \leq z \text{ e } y \leq z \quad \text{implica} \quad x \leq z\,, \quad x, y, z \in \Im.$$

Allora la relazione $\leq$ è detta *pre-ordinamento* su $\Im$.
Una relazione di pre-ordinamento che soddisfa anche:

$$x \leq y \text{ e } y \leq x \quad \text{implica} \quad x = y\,, \quad x, y \in \Im,$$

è detta *ordinamento parziale* su $\Im$.
Se, oltre alle precedenti condizioni, la relazione $\leq$ è tale da poter stabilire:

$$x \leq y \text{ o } y \leq x \quad \forall x, y \in \Im,$$

allora $\leq$ è detta *ordinamento totale* su $\Im$.
Confrontando due vettori x e y di uguale dimensione, si può osservare se, intuitivamente, le componenti di un vettore sono "più diverse" o "più uguali" tra loro. Si dice che x è *maggiorato* da y se le componenti di y sono caratterizzate da una "maggior disuguaglianza", il che si traduce, come sarà mostrato in seguito, in una serie di condizioni matematiche (disequazioni) che devono essere soddisfatte. Si verificherà che la relazione di maggiorazione è un pre-ordinamento. Si osservi quindi che la maggiorazione tra due vettori non è sempre verificabile.
La definizione più generale possibile è quella di maggiorazione rispetto ad una particolare misura μ. Grazie a questa definizione, il concetto di maggiorazione può essere esteso dal caso discreto (vettoriale) al caso continuo, in cui si ha a che fare con funzioni, considerate come vettori con un'infinità continua di componenti.

La teoria della maggiorazione porta allo studio di funzioni, o funzionali (a seconda dei casi) che preservino tale pre-ordinamento. In seguito, si parlerà infatti di funzioni *Schur-convesse*. Tali funzioni però sono state studiate approfonditamente solo nel caso discreto (Marshall, Olkin, 1979). Nel caso continuo invece (si parla di funzionali in questo caso) la generalizzazione si basa principalmente sul teorema dimostrato da Karamata nel 1932 e indipendentemente da Hardy, Littlewood e Pòlya nel 1929.
Il paragrafo *1.2* contiene la dimostrazione del teorema di Karamata più una sua importante estensione e un corollario (*1.2.2*). Questa premessa sarà fondamentale per il seguito: nel paragrafo *1.3* infatti, basandosi su questi teoremi sarà possibile dare una definizione generale di maggiorazione.

1.2 Teorema di Karamata

1.2.1 Dominio limitato

Sia μ una misura reale su un intervallo $[a,b]$. Si desidera trovare le condizioni necessarie e sufficienti che deve soddisfare μ affinché valga la seguente disuguaglianza:

$$\int_a^b \phi(x)d\mu \geq 0\,, \tag{1}$$

per ogni funzione ϕ convessa in (a,b).
(Karamata 1932, Hardy Littlewood e Pòlya 1929)

Teorema 1.2.A
Se ϕ è convessa in un intervallo (a,b), allora può essere approssimata uniformemente da una combinazione lineare positiva di funzioni (convesse) di tipo:

- lineare;
- $(x\text{-}c)^{+}=\max(0,x\text{-}c)$.

(Hardy, Littlewood e Pòlya, 1929)

Si osservi che le funzioni $(x\text{-}c)^+$ possono essere sostituite da funzioni tipo $(c\text{-}x)^+$, ottenute con le trasformazione lineare seguente:

$$(c\text{-}x)^+=c\text{-}x+(x\text{-}c)^+;$$

Come dimostrato da Karamata nel 1932, si otterrà che la (1) è valida $\forall\,\phi$ convessa se e solo se è valida per queste speciali funzioni convesse.

Si supponga innanzi tutto che (1) valga $\forall\,\phi$ convessa.

Ponendo $\phi(x)=\pm 1$ e $\phi(x)=\pm x$ si ottengono le seguenti condizioni necessarie:

$$\int_a^b d\mu = 0\,; \tag{2}$$

$$\int_a^b x\,d\mu = 0\,. \tag{3}.$$

Si osservi che la (2) equivale a $\mu(a,b)=0$, mentre, dalla (2) e la (3), integrando per parti, si ottiene:

$$0 = \int_a^b x\,d\mu = \left[x\mu(a,x)\right]_a^b - \int_a^b \mu(a,x)\,dx = -\int_a^b \mu(a,x)\,dx\,.$$

Allora le condizioni (2) e (3) possono essere riformulate così:

$$\mu(a,b)=0; \tag{2}'$$

$$\int_a^b \mu(a,x)\,dx = 0\,. \tag{3}'$$

Ponendo ora $\phi(x)=(x\text{-}c)^+$, $c\in(a,b)$, e tenendo valida la (4), si ottiene un'ulteriore condizione necessaria:

$$\int_a^b (x-c)^+\,d\mu = \int_c^b (x-c)\,d\mu = \int_c^b x\,d\mu - c\mu(c,b) = \left[x\mu(a,x)\right]_c^b - \int_c^b \mu(a,x)\,dx - c\mu(c,b) =$$

$$= b\mu(a,b) - c\left[\mu(a,c)+\mu(c,b)\right] - \int_c^b \mu(a,x)\,dx \geq 0\,, \text{ cioè } \int_c^b \mu(a,x)\,dx \leq 0\,.$$

Avvalendosi ora della (3)′ la precedente condizione equivale a:

$$\int_a^x \mu(a,t)dt \geq 0\,, \quad \forall x \in (a,b). \tag{4}$$

Riassumendo, (2)′, (3)′e (4) sono condizioni necessarie perché valga (1).

Si supponga ora che μ soddisfi (2)′, (3)′e (4): in virtù del teorema *1.2.1.1* è possibile costruire una successione $\{\phi_n\}$ di combinazioni lineari positive di funzioni dei tipi indicati dal teorema, che permetta la seguente rappresentazione di $\phi(x)$:

$$\phi(x) = \lim_{n\to\infty} \phi_n(x) = \lim_{n\to\infty} \sum_{i=1}^{n} h_i l_i(x) + \sum_{i=1}^{n} k_i (x - c_i)^+ \qquad h_i,\, k_i \geq 0,$$

dove le l_i sono funzioni lineari.
Si osservi quindi che:

- $\phi(x)=0 \Rightarrow l_i(x)=0,\ (x-c_i)^+=0,\ \forall$i;
- $\phi(x)>0 \Rightarrow l_i(x)>0,\ (x-c_i)^+>0,\ \forall$i;
- $\phi(x)<0 \Rightarrow l_i(x)<0,\ (x-c_i)^+<0,\ \forall$i.

Allora $\phi_n(x)$ è una successione monotòna crescente (in n) nei punti dove $\phi(x)\geq 0$ (insieme $P(\phi)$) e monotòna decrescente (in n) nei punti dove $\phi(x)\leq 0$ (insieme $N(\phi)$). Questo rende possibile l'applicazione del teorema di convergenza monotòna, che giustifica il passaggio del limite sotto il segno di integrale.

$$\int_a^b \phi(x)d\mu = \int_a^b \phi^+(x)d\mu - \int_a^b \phi^-(x)d\mu = \int_{(a,b)\cap P(\phi)} \phi(x)d\mu - \int_{(a,b)\cap N(\phi)} -\phi(x)d\mu =$$

$$= \int_{(a,b)\cap P(\phi)} \lim_{n\to\infty} \phi_n(x)d\mu - \int_{(a,b)\cap N(\phi)} \lim_{n\to\infty}(-\phi_n(x))d\mu = \lim_{n\to\infty} \int_a^b \phi_n(x)d\mu =$$

$$= \sum_{i=1}^{\infty} h_i \int_a^b l_i(x)d\mu + \sum_{j=1}^{\infty} k_j \int_a^b (x-c_i)^+ d\mu \geq 0\,.$$

Infatti, per ipotesi, si sa che l'integrale (rispetto a μ) su (a,b) di una qualsiasi funzione lineare è zero (condizioni (2)′ e (3)′), mentre l'integrale su (a,b) di una qualsiasi funzione di tipo $(x\text{-}c)^+$ assume valori non negativi (condizione(4)).
In conclusione la (1) vale, per ogni funzione convessa, se e solo se μ soddisfa (2)′, (3)′e (4), come scritto nel seguente teorema.

Teorema 1.2.B (di Karamata)

Sia μ una misura reale su un intervallo $[a,b]$. Allora:

$$\int_a^b \phi(x)d\mu \geq 0\,,\ \text{per ogni funzione convessa in } (a,b),$$

se e solo se:

$$\int_a^b d\mu = \int_a^b x d\mu = 0 \ \text{e}\ \int_a^x \mu(a,t)dt \geq 0\,, \quad \forall x \in (a,b).$$

Lo stesso risultato può essere dimostrato restringendo l'attenzione sulle funzioni convesse derivabili due volte (Karlin e Novikoff, 1962). Infatti, indicando con K l'insieme dalle funzioni convesse in $[a,b]$, si osservi che l'insieme $C^2 \cap K$ è debolmente denso in K.

Si è visto precedentemente che, se (1) vale ($\forall \phi$ convessa), allora, ponendo $\phi(x)=\pm 1$ e $\phi(x)=\pm x$, valgono (2) e(3) o, equivalentemente, (2)′ e (3)′. Allora in seguito, per ottenere la (4), si considerino le funzioni ϕ tali che $\phi(a)=0$, $\phi'(a)=0$ e $\phi \in K \cap C^2$ (ad esempio $(x-a)^2$, $(x-a)^3$, …).

Allora, ponendo $\int_a^x \mu(a,t)dt = M(x)$, si ottiene che:

$$\int_a^b \phi(x)d\mu = \left[\phi(x)\mu(a,x)\right]_a^b - \int_a^b \phi'(x)\mu(a,x)dx =$$

$$= 0 - \left[\phi'(x)M(x)\right]_a^b + \int_a^b \phi''(x)M(x)dx = \int_a^b \phi''(x)M(x)dx\,.$$

Ma dato che, per ipotesi, vale la (1) e $\phi \in K \cap C^2$, si avrà senz'altro $\phi''(x) \geq 0$ in (a,b) e, di conseguenza, $M(x) \geq 0$ in (a,b), condizione che chiaramente coincide con la (4).

Per dimostrare che vale anche il viceversa basta osservare che se $\phi \in K \cap C^{2,}$ ($\phi''(x) \geq 0$ in (a,b)) e μ soddisfa (2), (3) e (4) ($M(x) \geq 0$), allora la vale la (1). Inoltre, se la (1) vale per $\phi \in K \cap C^2$, allora vale anche più in generale, per ogni funzione convessa in (a,b).

1.2.2 Dominio illimitato

Il risultato del teorema *1.2.1.2*, riadattando e reinterpretando alcune espressioni, può rimanere valido anche in un intervallo $[a,b]$ illimitato. Innanzi tutto si esprima μ come differenza tra due distribuzioni F e G:

$$\mu(-\infty,x) = F(x)-G(x).$$

Si supponga che valga

$$\int_{-\infty}^{\infty} \phi(x)dF(x) \geq \int_{-\infty}^{\infty} \phi(x)dG(x), \text{ per ogni } \phi \text{ convessa in } \Re. \qquad (1)'$$

Si osservi che, perchè la precedente condizione sia valida, si dovrà verificare almeno che:

$$\int_{-\infty}^{\infty} \phi(x)dG(x) < \infty, \text{ per ogni } \phi \text{ convessa in } \Re.$$

Si ipotizzi inoltre, in vista delle prossime condizioni, che $\int_{-\infty}^{\infty} xdF(x) < \infty$, per ogni ϕ convessa in $\Re$.

Seguendo la dimostrazione precedente passo per passo, si ottengono condizioni necessarie analoghe a (2), (3) e (4).

In particolare, la condizione (2) diventa:

$$F(-\infty)=0,\ F(\infty)=V;\ G(-\infty)=0,\ G(\infty)=V. \qquad (2)''$$

Si noti che la (1)′, integrando per sostituzione, equivale a:

$$\int_0^V \phi(\bar{F}(t))dt \geq \int_0^V \phi(\bar{G}(x))dt \text{ , per ogni } \phi \text{ convessa in } \Re. \qquad (1)''$$

Per quanto riguarda la condizione (3), invece, si osservi innanzitutto che integrando per parti $\int_{-\infty}^{\infty} x dF(x)$ si ottiene una forma indeterminata. Non è possibile quindi riformulare la (3) con un'espressione analoga alla (3)′.

Integrando allora per sostituzione si ottiene che $\int_{-\infty}^{\infty} x dF(x) = \int_{0}^{V} \overleftarrow{F}(t)dt$, di conseguenza la (3) diventa:

$$\int_{0}^{V} \overleftarrow{F}(t)dt = \int_{0}^{V} \overleftarrow{G}(t)dt \,. \qquad (3)''$$

Si consideri infine la condizione (4), che in questo caso si può scrivere:

$$\int_{-\infty}^{\infty} (x-c)^{+} dF(x) \geq \int_{-\infty}^{\infty} (x-c)^{+} dG(x) \,.$$

Integrando per sostituzione:

$\int_{-\infty}^{\infty} (x-c)^{+} dF(x) = \int_{c}^{\infty} (x-c) dF(x) = \int_{F(c)}^{V} \overleftarrow{F}(t)dt - c[V - F(c)]$, quindi la (4) equivale a:

$$\int_{F(c)}^{V} \overleftarrow{F}(t)dt + cF(c) \geq \int_{G(c)}^{V} \overleftarrow{G}(t)dt + cG(c) \,.$$

Si supponga prima che $F(c)>G(c)$, e si riscriva così la precedente disuguaglianza:

$$\int_{F(c)}^{V} \overleftarrow{F}(t)dt + cF(c) \geq \int_{G(c)}^{F(c)} \overleftarrow{G}(t)dt + \int_{F(c)}^{V} \overleftarrow{G}(t)dt + cG(c)$$

Allora, poiché $\int_{G(c)}^{F(c)} \overleftarrow{G}(t)dt + cG(c) \geq cF(c)$, si ottiene:

$$\int_{F(c)}^{V} \overleftarrow{F}(t)dt \geq \int_{F(c)}^{V} \overleftarrow{G}(t)dt \,.$$

Se invece $F(c)<G(c)$, si consideri la disuguaglianza in questi termini:

$$\int_{F(c)}^{V} \overleftarrow{F}(t)dt + cF(c) \geq \int_{F(c)}^{V} \overleftarrow{G}(t)dt - \int_{F(c)}^{G(c)} \overleftarrow{G}(t)dt + cG(c).$$

Allora, poiché $cG(c) - cF(c) \geq \int_{G(c)}^{F(c)} \overleftarrow{G}(t)dt$, si ottiene ancora:

$$\int_{F(c)}^{V} \overleftarrow{F}(t)dt \geq \int_{F(c)}^{V} \overleftarrow{G}(t)dt.$$

Da questa condizione e dalla condizione (2) si ottiene:

$$\int_{0}^{z} \overleftarrow{F}(t)dt \leq \int_{0}^{z} \overleftarrow{G}(t)dt, \text{ dove } z \in [0,V]. \qquad (4)'$$

Allora si è visto che (2)′′,(3)′′ e (4)′ sono condizioni necessarie affinchè valga la (1)′.

Si supponga ora che F e G soddisfino (2)′′,(3)′′ e (4)′ e si ricordi la successione $\{\phi_n\}$ definita precedentemente per approssimare ϕ, in virtù del teorema *1.2.1.1*. Allora:

$$\int_{-\infty}^{\infty} \phi(x)dF(x) = \int_{0}^{V} \phi(\overleftarrow{F}(t))dt = \int_{0}^{V} \phi(\overleftarrow{F}(t))^{+}dt - \int_{0}^{V} \phi(\overleftarrow{F}(t))^{-}dt =$$

$$= \int_{[0,V]\cap P(\phi(\overleftarrow{F}))} \phi(\overleftarrow{F}(t))dt - \int_{[0,V]\cap N(\phi(\overleftarrow{F}))} -\phi(\overleftarrow{F}(t))dt =$$

$$= \int_{[0,V]\cap P(\phi(\overleftarrow{F}))} \lim_{n\to\infty} \phi_n(\overleftarrow{F}(t))dt - \int_{[0,V]\cap N(\phi(\overleftarrow{F}))} -\lim_{n\to\infty} \phi_n(\overleftarrow{F}(t))dt =$$

$$= \lim_{n\to\infty} \int_{0}^{V} \phi_n(\overleftarrow{F}(t))dt = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{0}^{V} \left(l_i(\overleftarrow{F}(t)) + (\overleftarrow{F}(t) + c_i)^{+} \right) dt.$$

Equivalentemente, per quanto riguarda G, si ha:

$$\int_{-\infty}^{\infty} \phi(x)dG(x) = \int_{0}^{V} \phi(\overleftarrow{G}(t))dt = \lim_{n\to\infty} \sum_{i=1}^{n} \int_{0}^{V} \left(l_i(\overleftarrow{G}(t)) + (\overleftarrow{G}(t) + c_i)^{+} \right) dt.$$

Ma, come mostrato precedentemente, (2)″ e (3)″ equivalgono a:

$$\int_0^V l_i(\overleftarrow{F}(t))dt = \int_0^V l_i(\overleftarrow{G}(t))dt\,,$$

mentre la (4)′ equivale a:

$$\int_0^V (\overleftarrow{F}(t) + c_i)^+ dt \geq \int_0^V (\overleftarrow{G}(t) + c_i)^+ dt\,.$$

Allora si può concludere che:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \int_0^V \left(l_i(\overleftarrow{F}(t)) + (\overleftarrow{F}(t) + c_i)^+ \right) dt \geq \lim_{n\to\infty} \sum_{i=1}^{n} \int_0^V \left(l_i(\overleftarrow{G}(t)) + (\overleftarrow{G}(t) + c_i)^+ \right) dt\,,$$

ossia

$$\int_0^V \phi(\overleftarrow{F}(t))dt \geq \int_0^V \phi(\overleftarrow{G}(t))dt\,.$$

In conclusione, il Teorema di Karamata può anche essere espresso nel modo seguente.

Teorema 1.2.C

Siano h_1 e h_2 due funzioni monotòne non decrescenti e continue da sinistra in [0, V]. Allora:

$$\int_0^V \phi(h_1(t))\,dt \geq \int_0^V \phi(h_2(t))\,dt$$

per ogni ϕ convessa se e solo se:

1. $\int_0^z h_1(t)dt \leq \int_0^z h_2(t)dt\,,\ z \in [0, V]$;
2. $\int_0^V h_1(t)dt = \int_0^V h_2(t)dt\,.$

Corollario 1.2.D

Date due funzioni monotòne non decrescenti e continue da sinistra in [0, V] h_1 e h_2, tali che:

$$\int_0^z h_1(t)dt \le \int_0^z h_2(t)dt\,,\ z\in[0,\,V];\ \int_0^V h_1(t)dt = \int_0^V h_2(t)dt\,.$$

Allora, per ogni funzione ϕ strettamente convessa, vale che:

$$\int_0^V \phi(h_1(t))\,dt > \int_0^V \phi(h_2(t))\,dt\,,$$

dove $h_1 \neq h_2$ quasi ovunque rispetto alla misura di Lebesgue.

Dimostrazione

Si supponga, per assurdo, che valga:

$$\int_0^V \phi(h_1(t))\,dt = \int_0^V \phi(h_2(t))\,dt\,,\ \text{per}\ h_1 \neq h_2\,. \qquad (*)$$

Si ponga $h_3 = \alpha h_1 + \overline{\alpha} h_2$, $\alpha\in(0,1)$ e $\overline{\alpha} = 1-\alpha$. Si verifica facilmente che:

$$\int_0^V h_3(t)dt = \int_0^V h_1(t)dt = \int_0^V h_2(t)dt\,;$$

$$\int_0^z h_3(t)dt = \alpha\int_0^z h_1(t)dt + \overline{\alpha}\int_0^z h_2(t)dt \ge \alpha\int_0^z h_1(t)dt + \overline{\alpha}\int_0^z h_1(t)dt = \int_0^z h_1(t)dt\,;$$

$$\int_0^z h_3(t)dt = \alpha\int_0^z h_1(t)dt + \overline{\alpha}\int_0^z h_2(t)dt \le \alpha\int_0^z h_2(t)dt + \overline{\alpha}\int_0^z h_2(t)dt = \int_0^z h_2(t)dt\,,$$

dove $z\in[0,\,V]$.

Allora, per il teorema di Karamata, si ha che:

$$\int_0^V \phi(h_2(t))\,dt \le \int_0^V \phi(h_3(t))\,dt \le \int_0^V \phi(h_1(t))\,dt\,,$$

che, tenendo conto di (*), diventa:

$$\int_0^V \phi(h_2(t))\,dt = \int_0^V \phi(h_3(t))\,dt = \int_0^V \phi(h_1(t))\,dt\ .$$

Ma dato che ϕ è strettamente convessa, si trova anche:

$$\int_0^V \phi(h_3(t))\,dt = \int_0^V \phi(\alpha h_1(t) + \overline{\alpha} h_2(t))\,dt < \alpha \int_0^V \phi(h_1(t))\,dt + \overline{\alpha} \int_0^V \phi(h_2(t))\,dt\ ,$$

tenendo conto di (*), la precedente disuguaglianza diventa:

$$\int_0^V \phi(h_3(t))\,dt < \int_0^V \phi(h_1(t))\,dt\ ,$$

che è assurdo.

1.3 Maggiorazione e Schur-convessità

1.3.1 Riordino di una funzione rispetto ad una misura

Sia f una funzione μ-misurabile in un intervallo I (anche illimitato), dove μ è una misura positiva e finita su I. Il *riordino decrescente* di f rispetto a μ è:

$$f_{\downarrow}^{(\mu)}(t) = \sup\{z:\ \lambda_z > t\},\ \ t \in [0, \mu(I)],$$

dove $\lambda_z = \mu(A_z)$, $A_z = \{t \in I:\ f(t) > z\}$.

(Bertoli-Barsotti, 2001)

$f_{\downarrow}^{(\mu)}$ è una funzione monotòna non crescente e continua da destra.

Si osservi che, a differenza di f e $f_{\downarrow}^{(\mu)}$, è misurabile rispetto a m (misura di Lebesgue) sull'insieme $[0,\mu(I)]$.

Ciò nonostante, f e $f_{\downarrow}^{(\mu)}$ sono funzioni *equimisurabili*, vale a dire:

$$\mu\{t \in I : z_1 < f(t) < z_2\} = m\left\{t \in [0, \mu(I)]: z_1 < f_{\downarrow}^{(\mu)}(t) < z_2\right\}.$$

In particolare $\int_I f d\mu = \int_0^{\mu(I)} f_{\downarrow}^{(\mu)}(t)\,dt$.

Analogamente, è possibile definire il *riordino crescente* di f rispetto a μ, che è:

$$f_{\uparrow}^{(\mu)}(t) = \inf\{z:\ \nu_z \geq t\},\ \ t \in [0, \mu(I)],$$

dove $\nu_z = \mu(B_z)$, $B_z = \{t \in I: f(t) \leq z\}$.

$f_{\uparrow}^{(\mu)}$ è una funzione monotòna non decrescente e continua da sinistra, equimisurabile rispetto a f (e a $f_{\downarrow}^{(\mu)}$).

1.3.2 (μ)-Maggiorazione

Siano *f, g* funzioni μ-integrabili e μ-misurabili su I ($f, g \in L^1(I,\mu)$). Si dice che *g* (μ)-*maggiora f* e si scrive $f \prec_{(\mu)} g$ quando:

1. $\int_0^z f_{\downarrow}^{(\mu)}(t)dt \leq \int_0^z g_{\downarrow}^{(\mu)}(t)dt$, $\forall z \in [0, \mu(I)]$.
2. $\int_I f d\mu = \int_I g d\mu$.

(Bertoli-Barsotti 2001)

Oppure, equivalentemente, $f \prec_{(\mu)} g$ quando:

1. $\int_0^z f_{\uparrow}^{(\mu)}(t)dt \geq \int_0^z g_{\uparrow}^{(\mu)}(t)dt$, $\forall z \in [0, \mu(I)]$.
2. $\int_I f d\mu = \int_I g d\mu$.

Nel caso particolare in cui μ è la misura di Lebesgue, si dice più semplicemente che *g maggiora f*, e si scrive $f \prec g$. Si osservi che, non avendo a che fare con una misura finita, su insiemi illimitati, *f* e *g* devono essere definite su intervalli chiusi. Dal caso discreto (dove μ è la misura del conteggio) si ottiene la definizione di *maggiorazione vettoriale*. Si noti che la misura del conteggio è finita su insiemi di cardinalità finita, *n*. Allora, poiché $f, g \in \Re^n$, si può scrivere:

$$f = (f_1, f_2, ..., f_n);$$

$$g = (g_1, g_2, ..., g_n).$$

Se $f_{[1]} \geq f_{[2]} \geq ..... \geq f_{[n]}$, si dice che *g maggiora f* quando:

1. $\sum_{i=1}^{k} f_{[i]} \le \sum_{i=1}^{k} g_{[i]}$, k =1,2,….n-1.

2. $\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} g_i$.

Oppure, equivalentemente, $f \prec g$ quando, se $f_{(1)} \le f_{(2)} \le ..... \le f(n)$:

1. $\sum_{i=1}^{k} f_{(i)} \le \sum_{i=1}^{k} g_{(i)}$, k =1,2,….n-1

2. $\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} g_i$.

(Marshall, Olkin 1979)

L'idea alla base della maggiorazione è la seguente: se g maggiora f significa che tra i valori assunti da g vi è una "maggiore disuguaglianza" rispetto ai valori assunti da f.

Più in generale, la maggiorazione rispetto ad una generica misura μ, può tener conto, eventualmente, di differenti "pesi", da attribuire ai valori delle funzioni f, g per i riordini, decrescenti o crescenti. Ad esempio, se μ è una misura di probabilità discreta su (0, 1, ….,n), i pesi variano a seconda che la distribuzione associata a μ sia una Uniforme discreta (pesi tutti uguali a 1/n) piuttosto che una Binomiale. Si osservi che non sempre si può dire se $f \prec_{(\mu)} g$ oppure $g \prec_{(\mu)} f$.

Un funzionale ψ: $L^1(I,\mu) \to \Re$ si dice (μ)-*Schur-convesso* se preserva l'ordinamento di (μ)-maggiorazione, vale a dire se:

$$f \prec_{(\mu)} g \Rightarrow \quad \psi(f) \le \psi(f),$$

dove $f, g \in L^1(I,\mu)$.

Analogamente un funzionale ψ: $L^1(I,\mu) \to \Re$ si dice (μ)-*Schur-concavo* se:

$$f \prec_{(\mu)} g \Rightarrow \quad \psi(f) \ge \psi(f),$$

dove $f, g \in L^1(I,\mu)$.

Nel caso particolare in cui μ è la misura di Lebesgue si dice semplicemente che ψ è un funzionale *Schur-convesso* (o *Schur-concavo*). Se invece μ è la misura del

conteggio, allora, come si è visto prima, $f, g \in \Re^n$. In tal caso si dice che ψ è una *funzione Schur-convessa* (o *Schur-concava*).

Teorema di (μ)*-maggiorazione 1.3.A*

Siano $f, g \in L^1(I,\mu)$. Allora $f \prec_{(\mu)} g$ se e solo se:

$$\int_I \phi(f)d\mu \le \int_I \phi(g)d\mu \text{, per ogni funzione convessa } \phi.$$

(Bertoli-Barsotti, 2001)

Dimostrazione

Data l'equimisurabilità tra f e $f_{\uparrow}^{(\mu)}(t)$, g e $g_{\uparrow}^{(\mu)}(t)$, la disuguaglianza è equivalente a:

$$\int_0^{\mu(I)} \phi\left(f_{\uparrow}^{(\mu)}(t)\right)dt \le \int_0^{\mu(I)} \phi\left(g_{\uparrow}^{(\mu)}(t)\right)dt,$$

$f_{\uparrow}^{(\mu)}(t)$ e $g_{\uparrow}^{(\mu)}(t)$ sono funzioni non decrescenti e continue da sinistra. Quindi, applicando il teorema di Karamata *1.2.A*, la disequazione vale se e solo se:

1. $\int_0^z f_{\uparrow}^{(\mu)}(t)dt \ge \int_0^z g_{\uparrow}^{(\mu)}(t)dt$, $\forall z \in [0,\mu(I)]$
2. $\int_0^{\mu(I)} f_{\uparrow}^{(\mu)}(t)dt = \int_0^{\mu(I)} g_{\uparrow}^{(\mu)}(t)dt$,

cioè se e solo se $f \prec_{(\mu)} g$.

Quindi ogni funzionale di tipo $\psi(f) = \int_I \phi(f)d\mu$, con ϕ convessa, è Schur-convesso.

Inoltre, se $f \prec_{(\mu)} g$, si osservi che, per ogni funzione ϕ strettamente convessa, vale:

$$\int_I \phi(f)d\mu < \int_I \phi(g)d\mu \text{, per } f \ne g,$$

(ovviamente non si deve avere anche $g \prec_{(\mu)} f$, caso da non tenere mai in considerazione anche nel seguito).

Questi risultati chiaramente si riducono al teorema di Karamata nel caso in cui μ è la misura di Lebesgue. Nel caso in cui invece μ è la misura del conteggio, ci si riconduce ai seguenti teoremi, dimostrati da Schur (1923), Hardy, Littlewood e Polya (1929; 1934; 1952).

Teorema 1.3.B

Dati i vettori $f = (f_1, f_2, ..., f_n)$; $g = (g_1, g_2, ..., g_n)$, l'intervallo I e una qualsiasi funzione convessa ϕ: $I \to \Re$, allora:

$$\psi(f) = \sum_{i=1}^{n} \phi(f_i)$$

è una funzione Schur-convessa su I^n. Ossia:

$$\sum_{i=1}^{n} \phi(f_i) \leq \sum_{i=1}^{n} \phi(g_i)$$

se e solo se $f \prec g$, per ogni funzione convessa ϕ.

(Marshall Olkin, pag. 64)

Corollario 1.3.C

Se $f \prec g$, allora, per ogni funzione *strettamente* convessa ϕ, vale:

$$\sum_{i=1}^{n} \phi(f_i) < \sum_{i=1}^{n} \phi(g_i)$$

(Marshall Olkin, pag. 64)

Chiaramente tutti i risultati ottenuti valgono per le funzioni o per i funzionali Schur-concavi, con la ovvia condizione che la funzione ϕ sia concava, o strettamente concava.

1.3.3 (μ)-Maggiorazione in senso debole

La (μ)-maggiorazione debole è una definizione che permette di confrontare la dissomiglianza tra funzioni in condizioni meno restrittive. Tuttavia, come dice la parola stretta, la definizione è più "debole": meno chiara, meno naturale della precedente.

Siano *f, g* funzioni μ-integrabili e μ-misurabili su *I* ($f, g \in L^1(I,\mu)$). Si dice che *g* (μ)-*maggiora da sotto f in senso debole* e si scrive $f \prec_{(\mu)w} g$ quando:

$$\int_0^z f_{\downarrow}^{(\mu)}(t)dt \le \int_0^z g_{\downarrow}^{(\mu)}(t)dt \ , \ \forall z \in [0,\mu(I)].$$

Si dice che *g* (μ)-*maggiora da sopra f in senso debole* e si scrive $f \prec_{(\mu)}{}^{w} g$ quando:

$$\int_0^z f_{\uparrow}^{(\mu)}(t)dt \ge \int_0^z g_{\uparrow}^{(\mu)}(t)dt \ , \ \forall z \in [0,\mu(I)].$$

Il seguente teorema permette di determinare funzionali $\psi(f)$ che preservino l'ordinamento di (μ)-maggiorazione in senso debole, cioè tali che:

$$f \prec_{(\mu)w} g \ \ (\text{o } f \prec_{(\mu)}{}^{w} g) \Rightarrow \ \psi(f) \le \psi(f).$$

Teorema di (μ)*-maggiorazione in senso debole 1.3.D*

1. Siano $f, g \in L^1(I,\mu)$. Allora $f \prec_{(\mu)w} g$ se e solo se:

   $\int_I \phi(f)d\mu \le \int_I \phi(g)d\mu$, per ogni funzione crescente e convessa ϕ.

2. Siano $f, g \in L^1(I,\mu)$. Allora $f \prec_{(\mu)}{}^{w} g$ se e solo se:

   $\int_I \phi(f)d\mu \le \int_I \phi(g)d\mu$, per ogni funzione decrescente e convessa ϕ.

Se ϕ è strettamente convessa, nelle precedenti disequazioni vale la disuguaglianza stretta.

Dimostrazione

Per il caso discreto (dove μ è la misura del conteggio) il risultato è stato dimostrato da Tomic (1949) e Weil (1949) (Marshall, Olkin, pag. 64 e 109). Per quanto riguarda il caso continuo (dove μ è la misura di Lebesgue) il risultato si ottiene ripercorrendo la dimostrazione del teorema di Karamata, modificando ovviamente le condizioni. Il caso generale (μ misura qualsiasi) si ottiene quindi in modo analogo a quanto visto per il teorema di (μ)-maggiorazione (*1.3.A*).

Le definizioni ed i teoremi contenuti in questo primo capitolo saranno utili per studiare, dal capitolo in avanti, certi funzionali statistici. Nel capitolo *2* invece verranno studiati funzionali statistici, stimatori e relative proprietà.

Capitolo 2

Stimatori e funzionali statistici

2.1 Funzionali statistici

2.1.1 Funzione di distribuzione empirica

Si supponga di effettuare un campionamento casuale semplice, di dimensione *n*, da una popolazione con distribuzione di probabilità *P*, ottenendo le variabili casuali i.i.d. $(X_1, X_2, ..., X_n)$. Si indica con *F* la *funzione di distribuzione* (o di *ripartizione*) corrispondente a *P*:

$$F(x) = P((-\infty, x)).$$

L'obiettivo dell'inferenza statistica è ottenere informazioni sull'incognita distribuzione *F*, servendosi dei dati campionari. In particolare, lo stimatore naturale di *F* nello spazio $\mathcal{F}$ delle funzioni di distribuzione è dato dalla *funzione di distribuzione* (o di *ripartizione*) *empirica*:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x),$$

dove $I(X_i \le x)$ è la *funzione indicatrice*, tale che:

$$I(X_i \le x) = \begin{cases} 1, & X_i \le x \\ 0, & X_i > x \end{cases}.$$

La misura di probabilità associata a F_n si indica con P_n.

Si osservi che:

$E\{I(X_i \le x)\} = F(x)$, $E\{I^2(X_i \le x)\} = F(x)$, $V\{I(X_i \le x)\} = F(x)(1 - F(x))$.

Essendo quindi $I(X_i \le x)$ una variabile casuale Bernuolliana,

$\sum_{i=1}^{n} I(X_i \le x) = nF_n(x)$ è una Binomiale, $nF_n(x) \sim Bi(F(x); n)$.

Allora $F_n(x)$ è uno stimatore non-distorto di $F(x)$, con varianza data da $F(x)(1-F(x))/n$. Inoltre, $F_n(x)$ gode delle seguenti proprietà.

1. *Consistenza forte*

Per la legge forte dei grandi numeri:

$F_n(x) \overset{q.c.}{\to} F(x)$, per $n \to \infty$.

(Borovkov, Mathematical Statistics, pag.4)

2. *Glivenko-Cantelli:*

$\sup_x \left| F_n(x) - F(x) \right| \overset{q.c.}{\to} 0$, per $n \to \infty$,

ossia $F_n(x)$ converge uniformemente a $F(x)$.

3. *Normalità asintotica*

Per il teorema del limite centrale:

$\sqrt{n}\left(F_n(x) - F(x)\right) \overset{d.}{\to} N\left(0; F(x)[1-F(x)]\right)$.

(Borovkov, 1998, Mathematical Statistics)

2.1.2 Definizione e proprietà di un funzionale statistico

Sia $\mathcal{F}$ lo spazio delle funzioni di distribuzione. Spesso, in statistica, si è interessati alla stima di una certa caratteristica dell'incognita distribuzione $F \in \mathcal{F}$, piuttosto che alla stima della stessa F.

Nella maggior parte dei casi tali caratteristiche di F possono essere espresse da funzionali del tipo:

$$T(F): \mathcal{F} \to \Re^k.$$

Ponendo ad esempio

$$T(F) = \int x^r dF(x),$$

si ottiene il *momento di ordine r* di F.

Oppure, ponendo

$$T(F) = \bar{F}(p) = \inf\{x \in \Re: F(x) \geq p\}, \quad \text{dove } p \in (0,1),$$

si ottiene il *percentile di ordine p* di F.

Se si è invece interessati al valore di un qualsiasi *parametro* θ da cui dipende la distribuzione F basta porre:

$$T(F) = \theta .$$

In tal caso si parla di funzionale *Fisher - consistente*.

Il metodo più naturale e intuitivo per la stima di una caratteristica $T(F)$ è il *metodo della sostituzione*, che consiste semplicemente nel sostituire all'incognita distribuzione F il suo stimatore naturale, ossia la funzione di distribuzione empirica F_n. $T(F_n)$ è quindi lo *stimatore per sostituzione* di $T(F)$, ed è anche detto *funzionale statistico* (Borovkok, 1998, Mathematical Statistics, pag. 52; Shao, 2003, pag. 291).

Sia F_0 è la vera funzione di distribuzione. Spesso $T(F)$ può appartenere alle seguenti classi di funzionali.

1. Funzionali del tipo:

$$T(F) = h\left(\int g(x)dF(x)\right),$$

dove g è una funzione $\mathcal{B}$-misurabile e h una funzione continua nel punto $\int g(x)dF_0(x)$.

2. Funzionali $T(F)$ uniformemente continui in F_0. Ossia tali che:

$$\sup_x \left|F^{(n)}(x) - F_0(x)\right| \to 0 \quad \Rightarrow \quad T(F^{(n)}) \to T(F_0),$$

dove $\left\{F^{(n)}\right\}$ è una successione di distribuzioni con supporti contenuti in quello di F_0.

Si dimostra che, se $T(F)$ è un funzionale di tipo 1) o 2), allora, per $n \to \infty$, $T(F_n) \overset{q.c.}{\to} T(F_0)$.

(Borovkov, 1998, Mathemaical Statistics, pag. 9)

E' possibile inoltre stabilire le condizioni sotto le quali un qualsiasi funzionale $T(F)$ abbia un comportamento asintoticamente normale. Questo deriva da una sorta di generalizzazione del *metodo*-Δ, tecnica che utilizza lo sviluppo di Taylor per determinare la normalità asintotica di una funzione (differenziabile) di una statistica con distribuzione asintoticamente normale (Hogg-McKean-Craig, 2005,

pag. 214-215). In questa circostanza, si ha a che fare con funzionali $T(F)$ e non con semplici funzioni dei dati campionari (statistiche). Quindi, prima di procedere con questa generalizzazione (del metodo-Δ), è necessario definire il differenziale di un funzionale.

2.1.3 Differenziabilità di un funzionale

Sia T un funzionale su $\mathcal{F}_0$, una *famiglia* di funzioni di distribuzione, ossia un sottoinsieme di $\mathcal{F}$. Si consideri l'insieme di funzioni $\mathcal{D} = \{ c(G_1 - G_2) : c \in \Re, G_i \in \mathcal{F}_0, \text{ i=1,2}\}$ e si ricordi che un funzionale $L(G)$ è *lineare* sull'insieme $\mathcal{D}$ se $L(c_1\Delta_1 + c_2\Delta_2) = c_1 L(\Delta_1) + c_2 L(\Delta_2)$, $\forall \Delta_j \in \mathcal{D}$, $\forall\, c_j \in \Re$.

Inoltre si osservi che un funzionale ρ: $\mathcal{F}_0 \times \mathcal{F}_0 \to [0,\infty)$ è detto *distanza* o *metrica* su $\mathcal{F}_0$ quando, per $F_1, F_2, F_3 \in \mathcal{F}_0$, sono soddisfatte le seguenti condizioni:

1. $\rho(F_1, F_2) = 0$ se e solo se $F_1 = F_2$;
2. $\rho(F_1, F_2) = \rho(F_2, F_1)$;
3. $\rho(F_1, F_2) \le \rho(F_1, F_3) + \rho(F_2, F_3)$.

(Shao, 2003, pag. 278)

Ad esempio si considerino le distanze:

$$\rho_\infty(F_1, F_2) = \sup_x \left| F_1(x) - F_2(x) \right| \text{ e}$$

$$\rho_{L_p}(F_1, F_2) = \left[\int \left| F_1(x) - F_2(x) \right|^p dx \right]^{1/p}.$$

La funzione $\|F\|$: $\mathcal{F}_0 \to [0,\infty)$ è detta *norma* su $\mathcal{F}_0$ se soddisfa:

1. $\|F\| \ge 0$ $\forall F \in \mathcal{F}_0$, $\|F\| = 0$ se e solo se $F = 0$.
2. $\|cF\| = |c| \|F\|$, $\forall c \in \Re$.
3. $\|F_1 + F_2\| \le \|F_1\| + \|F_2\|$, $\forall\, F_1, F_2 \in \mathcal{F}_0$.

Ci sono diverse definizioni di differenziabilità per un funzionale $T(F)$.

1. Un funzionale T su $\mathcal{F}_0$ è *Gateaux-differenziabile* in $G \in \mathcal{F}_0$ se esiste un funzionale lineare L_G su $\mathcal{D}$ tale che, se $\Delta \in \mathcal{D}$ e $G + t\Delta \in \mathcal{F}_0$, allora:

$$\lim_{t \to 0} \left[\frac{T(G + t\Delta) - T(G)}{t} - L_G(\Delta) \right] = 0 .$$

2. Sia ρ una distanza su $\mathcal{F}_0$. Si supponga che:

   $|c|\rho(G_1 - G_2) = \|c(G_1 - G_2)\| = \|\Delta\|$, con $\Delta \in \mathcal{D}$, definisca una norma su $\mathcal{D}$.

   Un funzionale T su $\mathcal{F}_0$ è *ρ-Hadamard-differenziabile* in $G \in \mathcal{F}_0$ se esiste un funzionale lineare L_G su $\mathcal{D}$ tale che:

$$\lim_{i \to \infty} \left[\frac{T(G + t_i \Delta_i) - T(G)}{t_i} - L_G(\Delta_i) \right] = 0 ,$$

   per ogni successione di numeri $t_i \to 0$ e di funzioni $\{\Delta, \Delta_i, i = 1,2,...\} \subset \mathcal{D}$ che soddisfano $\|\Delta - \Delta_i\| \to 0$ e $G + t_i \Delta_i \in \mathcal{F}_0$.

3. Sia ρ distanza su $\mathcal{F}_0$. Un funzionale T su $\mathcal{F}_0$ è *ρ-Frechèt-differenziabile* in $G \in \mathcal{F}_0$ se esiste un funzionale lineare L_G su $\mathcal{D}$ tale:

$$\lim_{i \to \infty} \left[\frac{T(G_i) - T(G) - L_G(G_i - G)}{\rho(G_i, G)} \right] = 0 ,$$

   per ogni successione $\{G_i\}$ che soddisfi $G_i \in \mathcal{F}_0$ e $\rho(G_i, G) \to 0$.

In ogni caso, il funzionale lineare L_G è detto *differenziale* di T in G (Shao, 2003, pag. 292).

Se si definisce la funzione $h(t) = T(G + t\Delta)$, allora dire che T è Gateaux-differenziabile in G equivale a dire che h è differenziabile in 0 e $L_G(\Delta)$ equivale a $h'(0)$.

2.1.4 Proprietà asintotiche di funzionali statistici differenziabili. Funzione di influenza e robustezza

Sia $\delta_y(x)$ la funzione di distribuzione degenere nel punto y e si definisca $I_G(y) = L_G(\delta_y - G)$, detta *funzione di influenza* di T in G (Huber, 2004, pag. 13 e 38). Tale funzione risulta di fondamentale importanza nello studio del comportamento asintotico di un funzionale statistico. In particolare è determinante nel calcolo della varianza asintotica dello stimatore.

La funzione di influenza è il differenziale di T in $\delta_y - G$, indica quindi il comportamento dello stimatore in conseguenza alla variazione di un punto del campione. Infatti, sia G la distribuzione che si crede "vera". Si supponga che i dati non seguano esattamente G, bensì che "tendano leggermente" verso un'altra distribuzione, H. In questo caso i dati provengono da una versione "contaminata" di G, data da $(1-\varepsilon)G(x) + \varepsilon H(x)$. Se G è contaminata da δ_y, funzione indicatrice di y, si ponga:

$$G_\varepsilon(x) = (1-\varepsilon)G(x) + \varepsilon\delta_y(x).$$

Allora, la funzione di influenza dello stimatore T in G è data da:

$$I_G(y) = \lim_{\varepsilon \to 0} \frac{T(G_\varepsilon) - T(G)}{\varepsilon},$$

indica la variazione nella stima corrispondente ad una contaminazione infinitesima nel punto campionario y.

Se T è Gateaux-differenziabile in F, allora, ponendo $t = n^{-1/2}$ e $\Delta = \sqrt{n}(F_n - F)$, si ha il seguente sviluppo:

$$\sqrt{n}\left[T(F_n) - T(F)\right] = L_F\left(\sqrt{n}(F_n - F)\right) + R_n.$$

Dato che L_F è un funzionale lineare:

$$L_F\left(\sqrt{n}(F_n - F)\right) = \sqrt{n}\left[\sum_{i=1}^{n} \frac{L_F\left(\delta(X_i) - F\right)}{n}\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_F(X_i).$$

Se si verifica:

$$E\left[I_F(X_1)\right] = 0 \quad \text{e} \quad E\left[I_F(X_1)\right]^2 = \sigma_F^2 < \infty,$$

che è solitamente vero quando I_F è limitata o quando F ha qualche momento finito (Shao, 2003, pag 292), allora, per il teorema del limite centrale:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_F(X_i) \xrightarrow{d} N(0,\sigma_F^2)\,.$$

Quindi, per il teorema di Slutsky, se R_n è $o_p(1)$:

$$\sqrt{n}\left[T(F_n)-T(F)\right] \xrightarrow{d} N(0,\sigma_F^2)\,.$$

La Gateaux-differenziabilità di T non basta a garantire la $R_n = o_p(1)$. Tuttavia questa condizione è solitamente vera, anche se difficile da dimostrare. Per farlo, bisogna servirsi di tipi di differenziabilità più forti, come quella di Hadamard e di Frechèt (ancora più forte). In particolare sono noti i seguenti risultati:

1. Se T è ρ_∞-Hadamard-differenziabile in F, allora R_n è $o_p(1)$.
2. Se T è ρ-Frechèt-differenziabile in F, dove ρ è una distanza tale che:

$$\sqrt{n}\rho(F_n,F) = O_p(1)\,,$$

allora R_n è $o_p(1)$.

(Shao, 2003, pag. 293)

Sfortunatamente, il concetto di Frechèt-differenziabilità risulta troppo forte: in gran parte dei casi il differenziale secondo Frechèt non esiste, o comunque risulta troppo difficile stabilirne l'esistenza.

La funzione di influenza risulta uno strumento molto importante nel determinare la robustezza di uno stimatore. La robustezza è una proprietà di grande utilità in diversi contesti di inferenza parametrica "applicata". Si ipotizza che esista un modello parametrico che rappresenti con buona approssimazione l'incognita distribuzione. Tuttavia non si può sapere con certezza quanto il modello sia vicino alla realtà. Per questo motivo si cerca uno stimatore che non sia troppo sensibile ad eventuali leggere deviazioni dalle assunzioni del modello, fornendo quindi risultati non troppo diversi da quelli forniti nella situazione "ideale". Inoltre, nel caso di deviazioni più consistenti dalle ipotesi, lo stimatore dovrebbe comunque fornire stime se non altro accettabili.

Si supponga che G la sia distribuzione che si crede "vera". La differenza $\Delta T(\varepsilon) = T(G_\varepsilon) - T(G)$ indica quanto il funzionale cambia in conseguenza ad una

certa contaminazione, quantificata da ε. Si supponga che l'obiettivo sia stimare il parametro $\theta = T(G)$. Allora $\Delta T(\varepsilon)$, detta *bias response curve*, rappresenta la distorsione asintotica della stima di θ, dovuta alla contaminazione.

Spesso si ritiene valida la seguente approssimazione:

$$\Delta T(\varepsilon) = T(G_\varepsilon) - T(G) \approx \varepsilon \cdot I_G(y),$$

secondo la quale la funzione di influenza risulta un indicatore della robustezza della stimatore. Tuttavia, per alcuni tipi di stimatori, la precedente approssimazione può risultare fuorviante (Lindsay, 1994), come si vedrà nel capitolo *5.3*.

Un caso particolare in cui la funzione di influenza ha senz'altro un ruolo cruciale nel determinare sia l'efficienza che la robustezza, è il caso in cui essa sia limitata.

Un funzionale ρ_∞-Hadamard-differenziabile T con funzione di influenza limitata e continua è detto *robusto* secondo Hampel (Huber, 2004, pag. 38). Infatti il comportamento asintotico di $T(F_n)$ è determinato da quello di $L_F(F_n - F)$ e quindi da I_F. Una piccola variazione nel campione produce una piccola variazione nello stimatore $T(F_n)$ se I_F è limitata e continua.

2.2 Stimatori L e M

2.2.1 Stimatori *L*

Le due classi più note e importanti di funzionali statistici sono quelle degli stimatori *L* e *M*.

Sia $J(t)$ una funzione su [0,1]. Si definisce un *L-funzionale* con:

$$T(G) = \int xJ(G(x))dG(x), \quad G \in \mathcal{F}.$$

Allora $T(F_n)$ è detto *stimatore*-L di $T(F)$.

Ad esempio, ponendo $J = 1$, si ottiene $T(F_n) = \overline{X}$, media campionaria.

Si consideri che, effettuando il cambio di variabilli $G(x) = t$, si ottiene:

$$T(G) = \int_0^1 J(t)\bar{G}(t)dt \, .$$

Sostituendo poi i valori campionari si ottiene che uno stimatore-*L* non è altro che una combinazione delle statistiche d'ordine $(x_{(1)}, x_{(2)}, ..., x_{(n)})$, quindi:

$$T(F_n) = \int_0^1 J_n(t)\bar{G}_n(t)dt = \frac{1}{n}\sum_{i=1}^{n} J_{n,i}x_{(i)} \, ,$$

dove $J_n(t) = J_{n,t}$ per $t \in ((i-1)/n, i/n]$.

Ad esempio la mediana campionaria e un qualsiasi percentile campionario costituisce uno stimatore-*L*.

E' possibile stabilire le condizioni su *J* che garantiscono la differenziabilità del *L*-funzionale $T(F)$ e di conseguenza la normalità asintotica dello stimatore-*L* $T(F_n)$ (Shao, 2003, pag. 297).

In particolare si consideri uno stimatore-*L* con $J(t) = 0$ per $t \in [0, a] \cup [b, 1]$, $b > a$, *J* limitata altrove. In questo caso si parla di stimatore–*L* "sfrondato" (trimmed) in quanto esclude una percentuale a scelta dei valori estremi del campione.

Si dimostra che uno stimatore-*L* "sfrondato" è asintoticamente normale, con funzione di influenza limitata e continua, quindi robusto secondo Hampel (Shao, 2003, pag. 298). I percentili campionari costituiscono proprio un caso limite di stimatore-L "sfrondato", e possiedono ulteriori proprietà (Shao, 2003, pag.305).

2.2.2 Stimatori *M*

Sia $H(x, \theta)$ una funzione $\Re \times \Theta \to \Re$, dove $\theta \in \Theta$, sottoinsieme aperto di $\Re$. Si chiama $\hat{M}$ *-funzionale* (Borovkov, 2004) il funzionale $T(G)$ tale che:

$$\int H(x, T(G))dG(x) = \min_{\theta \in \Theta} \int H(x, \theta)dG(x) \, , \quad G \in \mathcal{F}.$$

$T(F_n)$ è detto *stimatore-* $\hat{M}$ di $T(F)$.

Un importante caso particolare di stimatore-$\hat{M}$ è lo stimatore di *massima verosimiglianza* (che verrà approfondito in seguito), ottenuto ponendo

$H(x,\theta) = \ln f_\theta(x)$, dove f_θ è la densità corrispondente alla ditribuzione F_θ. Gli stimatori-$\hat{M}$ costituiscono evidentemente una generalizzazione di tale metodo, in cui la funzione $H(x,\theta)$ non dipende necessariamente dalla distribuzione. Ad esempio ponendo $H(x,\theta) = (x-\theta)^2$ si ottiene $T(F_n) = \overline{X}$, media campionaria. Oppure ponendo $H(x,\theta) = |x-\theta|$ si ottiene la mediana campionaria.

Se $H(x,\theta)$ è derivabile rispetto a θ, con $h(x,\theta) = dH(x,\theta)/d\theta$, e se vale la derivazione sotto il segno di integrale (vale senz'altro se *h* è continua), allora un funzionale-$\hat{M}$ $T(G)$ è anche un *funzionale-M*, cioè tale che:

$$\int h(x,T(G))dG(x) = \frac{d}{d\theta}\int H(x,T(G))dG(x) = 0\,.$$

Quindi si definisce stimatore-*M* la soluzione della seguente *equazione di stima*:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i,\theta) = 0\,.$$

Si osservi che non tutti gli stimatori-$\hat{M}$, soluzioni di quest'ultima equazione, sono stimatori-*M*. Ad esempio un eventuale minimo locale costituisce uno stimatore-*M* ma non uno stimatore-$\hat{M}$. Tuttavia gli stimatori-*M* godono di ottime proprietà e vengono utilizzati frequentemente.

Ecco degli esempi di funzioni *H* su cui si basano alcuni importanti stimatori di tipo $\hat{M}$ o *M*.

1. Ponendo:

$$H(x,\theta) = \ln f_\theta(x)\,,$$

allora:

$$h(x,\theta) = d\ln f_\theta(x)/d\theta\,,$$

si ottiene lo stimatore di *massima verosimiglianza*.

2. Ponendo:

$$H(x,\theta) = (x-\theta)^2\,,$$

allora:

$$h(x,\theta) = \theta - x\ ,$$

si ottiene lo stimatore *media campionaria*.

3. Ponendo:

$$H(x,\theta) = \begin{cases} 1/2\,(x-\theta)^2 & |x-\theta| \le C \\ C^2 & |x-\theta| > C \end{cases};$$

allora:

$$h(x,\theta) = \begin{cases} \theta - x & |x-\theta| \le C \\ 0 & |x-\theta| > C \end{cases},$$

si ottiene un tipo di media capionaria "sfrondata" (trimmed), proposta da Huber (1964).

4. Ponendo:

$$H(x,\theta) = |x-\theta|^p / p\ , \quad p \in [1,2),$$

allora:

$$h(x,\theta) = \begin{cases} |x-\theta|^{p-1} & x < \theta \\ 0 & x = \theta \\ -|x-\theta|^{p-1} & x > \theta \end{cases};$$

si ottiene lo stimatore della minima distanza L_p. Per p=1 si ottiene la *mediana campionaria*.

5. Ponendo:

$$H(x,\theta) = \begin{cases} 1/2\,(x-\theta)^2 & |x-\theta| \le C \\ C|x-\theta| - C^2/2 & |x-\theta| > C \end{cases};$$

allora:

$$h(x,\theta) = \begin{cases} C & \theta - x > C \\ \theta - x & |x-\theta| \le C \\ -C & \theta - x < C \end{cases},$$

si ottiene un tipo di stimatore detto media campionaria "*windsorizzata*" (windsorisez).

Si osservi che in alcuni casi la funzione *h* è limitata. E' soprattutto da questa caratteristica da cui derivano le proprietà di robustezza degli stimatori-*M*, come verrà mostrato nel seguente paragrafo.

2.2.3 Proprietà degli stimatori *M*

Sotto certe condizioni, è possibile determinare la normalità asintotica di uno stimatore-*M*. Queste condizioni hanno a che fare con la validità dell'approssimazione:

$$\sqrt{n}\left[T(F_n) - T(F)\right] \cong L_F\left(\sqrt{n}(F_n - F)\right),$$

vista nel paragrafo *2.1.4*.

Si dimostra (Huber, 2004) che la funzione di influenza di un funzionale-*M* è porporzionale alla sua funzione *h*, in particolare:

$$I_F(x) = -\frac{h(x,T(F))}{\int \frac{dh(x,T(F))}{d\theta} dF(x)}.$$

Questo significa che è possibile determinare le proprietà dello stimatore-*M* corrispondente, conoscendo *h*.

Teorema 2.2.A

Sia *T* un funzionale-*M* e sia *h* limitata e continua su $\Re \times \Theta$. Inoltre si supponga che $\int h(x,\theta)dF(x) = \lambda_F(\theta)$ sia derivabile in $\theta = T(F)$ con derivata $\lambda'(T(F)) \neq 0$ e continua. Allora il funzionale T è Hadamard-differenziabile in *F* con funzione di influenza limitata e continua, data da:

$$+I_F(x) = -h(x,T(F))/\lambda'_F(T(F)).$$

(Shao, 2003, pag.300)

Questo risultato garantisce la normalità asintotica dello stimatore-*M* $T(F_n)$, che è anche robusto secondo Hampel.

Sotto certe condizioni, è anche possibile stabilire la consistenza degli stimatori-$\hat{M}$ e *M*.

Per semplicità, si assume che l'insieme Θ sia compatto.

Inoltre è necessario che la funzione $H(x,\theta)$ sia tale $\int H(x,\theta)dF$ raggiunga il massimo in corrispondenza di un unico punto $\bar{\theta}$, cioè:

$$\int \left(H\left(x,\bar{\theta}\right)-H(x,\theta)\right)dF(x)<0 \text{, per ogni } \theta \neq \bar{\theta}.$$

Tuttavia le precedenti condizioni non sono sufficienti, per garantire la consistenza di uno stimatore-$\hat{M}$ sono necessarie due condizioni particolari.
Si definiscano le funzioni:

$$H^{\Delta}(x,\theta)=\sup_{|u|<\Delta} H(x,\theta+u);$$

$$H^{0}(x,\theta)=\limsup_{u\to\theta} H(x,u).$$

Teorema 2.2.B

Uno stimatore-$\hat{M}$ è consistente in senso forte se, per ogni δ>0, esiste un $\Delta=\Delta(\delta)$ tale che, $\forall\theta$, per $\varepsilon>0$, si ha:

$$\int \left(H\left(x,\bar{\theta}\right)-H^{\Delta}(x,\theta)\right)dF(x)<-\varepsilon.$$

(Borovkov, 1998, Mathematical Statistics, pag. 60)

Teorema 2.2.C

Uno stimatore-$\hat{M}$ è consistente in senso forte se valgono le condizioni:

1. $\int \left(H\left(x,\bar{\theta}\right)-H^{0}(x,\theta)\right)dF(x)<0, \quad \forall\theta\neq\bar{\theta}.$
2. $\int \left(H\left(x,\bar{\theta}\right)-H^{\Delta}(x,\theta)\right)dF(x)<\infty, \quad \forall\theta$ e per qualche $\Delta>0$.

Si dimostra infatti che le condizioni 1) e 2) sono equivalenti alla condizione del teorema *2.2.B* (Borovkov, 1998, Mathematical Statistics, pag. 61).

Le condizioni sono più semplici nel caso particolare in cui si considera uno stimatore-*M*, soluzione di un'equazione del tipo:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i,\theta)=0.$$

Teorema 2.2.D

Uno stimatore-*M* è consistente in senso forte se valgono le condizioni:

1. $\sup|h(x,\theta)|\leq\gamma(x)$, dove $\int\gamma(x)dF(x)<\infty$.

2. $h(x,\theta)$ è continua in θ quasi ovunque.
3. L'equazione $\int h(x,\theta)dF(x)=0$ ha un'unica soluzione θ_0, punto interno di Θ.
4. Θ è compatto.

(Borovkov, 1998, Mathematical Statistics, pag. 62).

2.3 Stimatori GEE

Molte procedure inferenziali si riducono nella risoluzione di particolari equazioni di stima, dette *generalized estimating equations* (GEE). Questo metodo è molto generale, dato che include molti degli stimatori studiati precedentemente come casi particolari.

Si supponga che i vettori casuali campionari $(X_1, X_2, ..., X_n)$ siano indipendenti, ma non necessariamente identicamente distribuiti. Si è interessati alla stima dell'incognito vettore di parametri θ, da cui dipende la popolazione esaminata. Si indica con d_i, $i = 1,...,n$ la dimensione di X_i ($\sup_i d_i < \infty$) e con *k* la dimensione di θ, dove $\theta \in \Theta$ (spazio parametrico).

Sia ψ_i una funzione misurabile $\psi_i : \Re^{d_i} \times \Theta \to \Re^k$, per $i = 1,...,n$ e si definisca:

$$s_n(\gamma) = \sum_{i=1}^{n} \psi_i(X_i, \gamma), \qquad \gamma \in \Theta .$$

Si chiama *stimatore GEE* il vettore $\hat{\theta} \in \Theta$ che soddisfa:

$$s_n(\hat{\theta}) = \sum_{i=1}^{n} \psi_i(X_i, \hat{\theta}) = 0 .$$

L'equazione $s_n(\hat{\theta}) = 0$ è detta appunto equazione di stima generalizzata.

Si osservi che, ad esempio, uno stimatore *M* è un caso speciale di stimatore GEE.

Solitamente, l'equazione di stima è scelta in modo da soddisfare:

$$E[s_n(\theta)] = \sum_{i=1}^{n} E[\psi_i(X_i, \theta)] = 0 .$$

In tal caso, la scelta di $\hat{\theta} \in \Theta$ è giustificata dal fatto che $s_n(\hat{\theta}) = 0$ risulta essere l'analogo, da un punto di vista campionario, di $E[s_n(\theta)] = 0$.

Se l'esatto valore atteso non esiste nella precedente espressione, allora *E* può essere sostituito dal valore atteso asintotico *AE*, definito nella maniera seguente.

Sia $\{X_n\}$ una successione di vettori casuali e sia $\{E(X_n)\}$ la successione dei valori attesi corrispondenti. Se $\lim_{n\to\infty} E(X_n) = \mu$, allora $\mu = AE(X_n)$ è detto *valore atteso asintotico* di $\{X_n\}$.

Si indichi con $\hat{\theta}_n$ lo stimatore GEE al variare della numerosità campionaria. Sotto opportune condizioni sulle funzioni ψ_i, è possibile dimostrare la consistenza e la normalità asintotica degli stimatori GEE.

Teorema 2.3.A

Si supponga che $X_1, X_2, ..., X_n$ siano variabili casuali i.i.d. da una distribuzione *F* e che $\psi_i = \psi$ sia una funzione limitata e continua $\psi : \Re^d \times \Theta \to \Re^k$. Si ponga

$$\Psi(t) = \int \psi(x,t) dF(t).$$

Se $\Psi(\theta) = 0$ e se $d\Psi(\theta)/dt$ esiste ed è di rango pieno in $t = \theta$, allora $\hat{\theta}_n \to_p \theta$. (Shao, 2003, pag. 317)

Teorema 2.3.B

Siano $X_1, X_2, ..., X_n$ variabili casuali i.i.d. da una distribuzione *F*, con $\theta \in \Re$. Si supponga che $\Psi(t) = 0$ se e solo se $t = \theta$, che $d\Psi(\theta)/dt = \Psi'(\theta)$ esista e che $\Psi'(\theta) \neq 0$.

Sia $\psi(x,t)$ non crescente rispetto a *t* a sia l'integrale $\int [\psi(x,t)]^2 dF(t)$ finito, per *t* in un intorno di θ, e continuo in $t = \theta$. Allora, ogni successione di stimatori GEE $\{\hat{\theta}_n\}$ soddisfa:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_F^2),$$

dove

$$\sigma_F^2 = \int [\psi(x,t)]^2 dF(t) \Big/ [\Psi'(\theta)]^2.$$

(Shao, 2003, pag.321)

Un caso particolare in cui si possono applicare i precedenti teoremi è il metodo della *massima verosimiglianza*, che verrà illustrato in seguito. In tal caso la funzione ψ, detta *score fuction*, è data da $d \ln f(x,t)/d t$, dove f è la densità di F.

Le proprietà di consistenza e normalità asintotica per i GEE possono essere ottenute anche in casi più generali, come il caso di variabili non i.i.d. (Shao, 2003, pag. 317-325).

Capitolo 3

Stima della minima distanza

3.1 Metodo della minima distanza

3.1.1 Caso generale

Sia $\mathcal{F}$ lo spazio delle funzioni di distribuzione, e sia $\mathcal{F}_0$ una famiglia di funzioni di distribuzione, sottoinsieme di $\mathcal{F}$. Sia ρ una distanza (o metrica) su $\mathcal{F}_0$.

Il metodo della minima distanza è, intuitivamente, un metodo di stima dell'incognita distribuzione $F \in \mathcal{F}_0$, che consiste nell'individuare la distribuzione $F^* \in \mathcal{F}_0$ tale che la distanza tra F^* e la distribuzione empirica $F_n(x)$ sia minima.

Si osservi che non necessariamente $F_n(x) \in \mathcal{F}_0$, caso nel quale chiaramente $F_n(x) = F^*$.

Tuttavia, è consueto l'impiego di funzionali, che, pur non essendo per forza distanze vere e proprie, godano di proprietà indicate per questa procedura di stima. Infatti, è sufficiente considerare funzionali $d(F,G)$, con G e $F \in \mathcal{F}_0$, tali che:

1. $\min_F d(F,G) = d(G,G)$;
2. $d(F,G) > d(G,G)$ per $F \neq G$.

In seguito verranno chiamati (impropriamente) *distanze* anche funzionali di questo tipo (Borovkov, 1998, Mathematical Statistics, pag. 65). Si osservi che non necesariamente $d(G,G) = 0$.

Si supponga che la variabile casuale X abbia un'incognita funzione di distribuzione $F \in \mathcal{F}_0$.

Data la distribuzione G, non necessariamente appartenente a $\mathcal{F}_0$, si indica con $_d(G)_{\mathcal{F}0}$ la distribuzione in $\mathcal{F}_0$ più vicina a G, rispetto alla "distanza" d. Quindi:

$$d({}_d(G)_{\mathcal{F}_0},\, G) = \min_{\Pi} d(\Pi, G)\,, \quad \Pi \in \mathcal{F}_0,$$

si osservi che ${}_d(G)_{\mathcal{F}_0} = G$ se $G \in \mathcal{F}_0$.

Se si pone $G = F_n(x)$ si ottiene lo *stimatore della minima distanza* $F^* = (F_n)_{\mathcal{F}}$ per la distribuzione F, rispetto alla "distanza" d. Ossia:

$$d(F^*, F_n) = \min_{\Pi} d(\Pi, F_n)\,, \quad \Pi \in \mathcal{F}_0,$$

si osservi che $F^* = F_n$ se $F_n \in \mathcal{F}_0$.

(Borovkov, 1998, Mathematical Statistics, pag.66)

In generale, si ricordi che d è una distanza su $\mathcal{F}_0$, quindi non gode necessariamente delle proprietà 1) e 2) nell'intero spazio $\mathcal{F}$ delle funzioni di distribuzione. Spesso infatti $F_n \notin \mathcal{F}_0$, fatto che, talvolta, rende addirittura incalcolabile $d(F, F_n)$.

Ricordando le proprietà dei funzionali (μ)-Schur-convessi, sarà possibile, nel capitolo *4*, individuare una classe di distanze, dette anche *divergenze*, che siano coerenti con un'eventuale relazione di dissomiglianza tra distribuzioni, epressa dalla (μ)-maggiorazione.

3.1.2 Caso parametrico

Si consideri la famiglia parametrica di funzioni di distribuzione $\mathcal{F}_\theta$, con $\theta \in \Theta$, tale che, date F_{θ_1} e $F_{\theta_2} \in \mathcal{F}_\theta$:

$$F_{\theta_1} \neq F_{\theta_2} \text{ se } \theta_1 \neq \theta_2.$$

Essendoci in questo caso una corrispondenza biunivoca tra il parametro θ e la distribuzione F_θ, è possibile definire il funzionale $H: \mathcal{F}_\theta \to \Theta$, tale che:

$$H(F_\theta) = \theta\,.$$

Data la distanza d su $\mathcal{F}_\theta$, si indica ancora con ${}_d(G)_{\mathcal{F}_\theta}$ la distribuzione in $\mathcal{F}_\theta$ più vicina a G, rispetto a d. Quindi:

$$d({}_d(G)_{\mathcal{F}_\theta}, G) = \min_{\Pi} d(\Pi, G), \quad \Pi \in \mathcal{F}_\theta,$$

si osservi che ${}_d(G)_{\mathcal{F}_\theta} = G$ se $G \in \mathcal{F}_\theta$.

Ponendo $G = F_n(x)$ si ottiene $\hat{\theta}$, lo *stimatore* di θ *della minima distanza* rispetto a d, tale quindi che:

$$d(F_{\hat{\theta}}, F_n) = \min_{\theta} d(F_\theta, F_n), \quad F_\theta \in \mathcal{F}_\theta.$$

$\hat{\theta}$ è dato da:

$$H[{}_d(F_n)_{\mathcal{F}_\theta}] = \hat{\theta}.$$

Chiaramente, se $F_n \in \mathcal{F}_\theta$, allora $F_{\hat{\theta}} = F_n$, e $H(F_n) = \hat{\theta}$.

In questo caso lo stimatore della minima distanza costituisce uno stimatore-$\hat{M}$ e, eventualmente, anche uno stimatore-M. Di conseguenza, sotto le opportune condizioni, viste precedentemente, è possibile verificarne la proprietà di consistenza e normalità asintotica.

3.2 Massima verosimiglianza e minima distanza

3.2.1 Verosimiglianza empirica, caso non parametrico

Dato il campione $(x_1, x_2, ..., x_n)$ da una distribuzione incognita $F_0 \in \mathcal{F}$, si definisce *verosimiglianza empirica* della funzione di distribuzione $F \in \mathcal{F}$, il funzionale $L(F)$: $\mathcal{F} \to [0,1]$, tale che:

$$L(F) = \prod_{i=1}^{n} \Delta F(x_i),$$

dove $\Delta F(x_i) = F(x_i) - F(x_i^-) = P_F(X = x_i)$.

$L(F)$ rappresenta la probabilità di ottenere esattamente i valori campionari osservati $x_1, x_2, ..., x_n$ da una qualsiasi distribuzione F. Di conseguenza $L(F)=0$ se F è continua, mentre per avere un valore positivo di $L(F)$, F deve assegnare probabilità positiva ad ogni singola osservazione x_i.

Si supponga che i valori osservati $x_1, x_2, ..., x_n$ siano tutti diversi tra loro. Definendo il funzionale $l(F) = \ln L(F)$, si osserva che:

$$d_1(F, F_n) = -\frac{1}{n} l(F) = \sum_{i=1}^{n} - \ln \Delta F(x_i) \frac{1}{n} = \int - \ln \Delta F dF_n .$$

Si può interpretare il funzionale $d_1(F, F_n) = \int - \ln \Delta F dF_n$ come distanza in $\mathcal{F}$ tra F e F_n. Infatti valgono le condizioni espresse dal seguante teorema.

Teorema 3.2.A

$$\min_F d_1(F, F_n) = d_1(F_n, F_n), \ F \in \mathcal{F};$$

$$d_1(F, F_n) > d_1(F_n, F_n) \ \text{ per } F \neq F_n, \text{ ossia } l(F) < l(F_n).$$

Dove $x_i \neq x_j, \forall i \neq j$.

(Shao, 2003, pag. 282)

Quindi la funzione di distribuzione empirica F_n rappresenta la stima di minima distanza di F rispetto a d_1 o, equivalentemente, la stima di massima verosimiglianza di F.

Questo risultato può essere dimostrato in modo più semplice sfruttando il concetto di Schur-convessità.

Dimostrazione

Per comodità si ponga $\Delta F(x_i) = P(X = x_i) = p_i$. Si è visto che massimizzare la verosimiglianza empirica equivale a minimizzare $-l(F) = \sum_{i=1}^{n} - \ln p_i$. Si osservi che i valori delle probabilità p_i dipendono dall'incognita distribuzione F e che $\sum_{i=1}^{n} p_i = c \leq 1$. $c = 1$ solo nel raro caso in cui si ha un'osservazione per ogni punto

del supporto di *F*. Si chiamerà $\mathcal{F}(c)$ la classe delle distribuzioni tali che $\sum_{i=1}^{n} p_i = c$.

Si ponga $p_i' = \Delta F'(x_i) = P_{F'}(X = x_i)$, dove $F' \neq F$ e $F' \in \mathcal{F}(c)$, quindi $\sum_{i=1}^{n} p_i' = \sum_{i=1}^{n} p_i = c$.

Per la proprietà delle funzioni Schur-convesse si ha che

$$\sum_{i=1}^{n} -\ln p_i \leq \sum_{i=1}^{n} -\ln p_i' ,$$

se e solo se

$$(p_1, p_2, ..., p_n) \prec (p_1', p_2', ..., p_n') .$$

Allora:

$$\min_{F \in F(c)} \sum_{i=1}^{n} -\ln p_i = \sum_{i=1}^{n} -\ln \frac{c}{n} = n\left(-\ln \frac{c}{n}\right).$$

Il minimo nella classe $\mathcal{F}(c)$ si ottiene quindi in corrispondenza della distribuzione tale che $p_i = c/n$, $i = 1,2,....,n$.

Si consideri ora il risultato al variare di *c*. Dato che il valore massimo di *c* è 1:

$$\min_{F \in F} \sum_{i=1}^{n} -\ln p_i = n\left(-\ln \frac{1}{n}\right),$$

cioè il minimo nella classe $\mathcal{F}$ di tutte le possibili distribuzioni si ottiene in corrispondenza della distribuzione tale che $p_i = 1/n$, $i = 1,2,....,n$, che è ovviamente la funzione di ripartizione empirica F_n.

Il risultato può essere esteso anche al caso, senz'altro più comune, in cui i valori osservati $x_1, x_2, ..., x_n$ non siano necessariamente diversi tra loro.

Teorema 3.2.B

$$\min_F d_1(F, F_n) = d_1(F_n, F_n)\,,\ F \in \mathcal{F};$$

$$d_1(F, F_n) > d_1(F_n, F_n) \quad \text{per } F \neq F_n\text{, ossia } l(F) < l(F_n)\,.$$

Dimostrazione

Si supponga che tra i valori $x_1, x_2, ..., x_n$ ci siano $z_1, z_2, ..., z_k$ valori diversi ($k < n$). Si indichi con μ_{F_n} la misura (di probabilità) definita da F_n sull'insieme di punti $z_1, z_2, ..., z_k$, con $\mu_{F_n}(z_j) = f_n(z_j) = n_j / n$, per $j = 1,..,k$.

Si osservi che massimizzare

$$\sum_{i=1}^{n} -\ln \Delta F(x_i) \frac{1}{n} = \sum_{j=1}^{k} -\ln \Delta F(z_j) \frac{n_j}{n}\,,$$

rispetto all'incognita distribuzione *F,* è equivalente a massimizzare:

$$\sum_{j=1}^{k} -\ln\left(\frac{\Delta F(z_j)}{n_j / n}\right)\frac{n_j}{n}\,,$$

dato che $\sum_{j=1}^{k} \ln\left(n_j / n\right)\frac{n_j}{n}$ non dipende da *F*.

Sia $\Delta F(z_i) = P(X = z_i) = p_i$. In questo caso $\mathcal{F}(c)$ è la classe delle distribuzioni tali che $\sum_{i=1}^{k} p_i = c$.

Si ponga $p_i' = \Delta F'(x_i) = P_{F'}(X = x_i)$, dove $F' \neq F$ e $F' \in \mathcal{F}(c)$, quindi $\sum_{j=1}^{k} p_j' = \sum_{i=1}^{k} p_j = c$.

Per la proprietà delle funzioni Schur-convesse rispetto ad una misura, si ha che

$$\sum_{j=1}^{k} -\ln\left(\frac{p_i}{n_j / n}\right)\frac{n_j}{n} = \int -\ln\left(\frac{\Delta F}{f_n}\right) dF_n \leq \int -\ln\left(\frac{\Delta F'}{f_n}\right) dF_n = \sum_{j=1}^{k} -\ln\left(\frac{p_i'}{n_j / n}\right)\frac{n_j}{n}\,,$$

se e solo se

$$(p_1, p_2, ..., p_n) \prec_{(\mu_{F_n})} (p'_1, p'_2, ..., p'_n).$$

Allora:

$$\min_{F \in F(c)} \sum_{j=1}^{k} -\ln\left(\frac{p_j}{n_j/n}\right)\frac{n_j}{n} = \sum_{j=1}^{k} -\ln\left(\frac{c(n_j/n)}{n_j/n}\right)\frac{n_j}{n} = -\ln c .$$

Il minimo nella classe $\mathcal{F}(c)$ si ottiene in corrispondenza della distribuzione tale che $p_j = c(n_j/n)$, $j = 1,2,....,k$. In tal caso si ha infatti che $\frac{p_j}{n_j/n} = c$, $j = 1,2,....,k$.

Il minimo nella classe $\mathcal{F}$ di tutte le possibili distribuzioni si ottiene quindi in corrispondenza della distribuzione tale che $p_j = n_j/n$, $i = 1,2,....,k$, che è ovviamente la funzione di ripartizione empirica F_n.

Come si vedrà in seguito, $\rho_1(F_n, F) = \sum_{j=1}^{k} -\ln\left(\frac{p_j}{n_j/n}\right)\frac{n_j}{n}$ è la *distanza* di *Kullback-Leibler* tra la distribuzione empirica F_n e la distribuzione *F*. Allora F_n può essere vista come stima della minima distanza in $\mathcal{F}$ (ovviamente $F_n \in \mathcal{F}$).

3.2.2 Verosimiglianza nel caso parametrico. Efficienza

Si supponga che $(X_1, X_2, ..., X_n)$ sia un campione da una distribuzione *F* appartenente ad una famiglia parametrica $\mathcal{F}_\theta$, tale che, date F_{θ_1} e $F_{\theta_2} \in \mathcal{F}_\theta$:

$$F_{\theta_1} \neq F_{\theta_2} \text{ se } \theta_1 \neq \theta_2.$$

Inoltre si supponga che esista una misura σ-finita μ, rispetto alla quale ogni distribuzione $F_\theta \in \mathcal{F}_\theta$ sia assolutamente continua, con densità data da $f_\theta(x) = (dF_\theta / d\mu)(x)$ e che il supporto di f_θ non dipenda da θ.

Sia F_θ una distribuzione discreta. La funzione $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$ è detta funzione di *verosimiglianza* di θ e rappresenta la probabilità di ottenere esattamente i valori campionari osservati $X_1, X_2, ..., X_n$ da una distribuzione $F=F_\theta$.

Nel caso continuo la probabilità di ottenere esattamente i valori campionari osservati $X_1, X_2, ..., X_n$ da una distribuzione $F=F_\theta$ è nulla. La verosimiglianza quindi può essere definita (al massimo) come prodotto delle probabilità di intervalli contenti le osservazioni, ad esempio:

$$\prod_{i=1}^{n} P_\theta(x_i, x_i + h_i) = \prod_{i=1}^{n} \int_{x_i}^{x_i+h_i} f_\theta(x)dx ,$$

dove h_i sono quantità positive, piccole a piacere (Cheng, Iles 1987).

Solitamente, grazie a due significative approssimazioni, la funzione di verosimiglianza si riduce ad un prodotto di funzioni di densità. Si tenga però conto che queste approssimazioni non sono sempre accettabili (Cheng, Iles, 1987; Cheng, Amin, 1983).

La prima approssimazione è data da:

$$\int_{x_i}^{x_i+h_i} f_\theta(x)dx \cong f_\theta(x_i)h_i , \quad \text{per } i=1,2,\ldots..,n.$$

La seconda si ottiene facendo tendere a zero la lunghezza (h_i) degli intervalli. In tal modo la prima approssimazione diventa sempre più precisa e la funzione di verosimiglianza risulta, a meno di una cosante moltiplicativa, data da:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) ,$$

espressione nota comunemente come verosimiglianza di θ.

Definendo poi la funzione $l(\theta) = \ln L(\theta)$, si osserva che

$$-\frac{1}{n}l(\theta)=\sum_{i=1}^{n}-\ln f_\theta(x_i)\frac{1}{n}=\int-\ln f_\theta dF_n\ .$$

Si può interpretare il funzionale $d_2(F,G)=\int_S-\ln\ fdG$, dove $f=dF/d\mu$, come distanza tra *F* e *G* nella classe $\mathcal{F}_\mu$ delle distribuzioni assolutamente continue rispetto a μ, con supporto comune *S*.

Infatti valgono le condizioni espresse dal seguente teorema.

Teorema 3.2.C

$$\min_F d_2(F,G)=d(G,G)\ ;\ d_2(F,G)>d_2(G,G)\ \text{ per } F\neq G.$$

(Borovkov, 1998, Mathematical statistics, pag. 68)

Questo risultato può essere dimostrato in pochi passaggi, utilizzando la proprietà dei funzionali schur-convessi.

Dimostrazione

Bisogna dimostrare che:

$$\int_S-\ln\frac{f}{g}dG\geq 0=\int_S-\ln\frac{g}{g}dG\ .$$

Sia μ_G la misura indotta da *G* . Per la proprietà dei funzionali Schur-convessi rispetto ad una misura, la precedente disequazione è soddisfatta se $\frac{g}{g}\prec_{(\mu_G)}\frac{f}{g}$, che è sicuramente vero, dato che ovviamente $g/g=1$ su *S*. Si noti poi che –ln è strettamente convessa, quindi vale la disuguaglianza stretta per $f\neq g$.

Si osservi che, se il supporto di f_θ non dipende da θ, allora $\mathcal{F}_\mu\supset\mathcal{F}_\theta$ e d_2 è una distanza anche nella classe $\mathcal{F}_\theta$.

Ponendo $G=F_n(x)$ si indica con $\hat{\theta}$ lo stimatore di θ della minima distanza rispetto a $d_2(F_\theta,G)$, $\hat{\theta}$ è dato da:

$$H\,[_{d2}(F_n)_{\mathcal{F}\theta}]=\hat{\theta},$$

e tale che:

$$d_2(F_{\hat{\theta}}, F_n) = \min_{\theta} d_2(F_\theta, F_n), \quad F_\theta \in \mathcal{F}_\theta.$$

Chiaramente, se $F_n \in \mathcal{F}_\theta$, allora $F_{\hat{\theta}} = F_n$.

$\hat{\theta}$ è anche detto stimatore di *massima verosimiglianza* di θ ("*maximum likelihood estimator", MLE*), essendo equivalentemente identificato da:

$$\max_{\theta} \sum_{i=1}^{n} \ln f_\theta(x_i) = \sum_{i=1}^{n} \ln f_{\hat{\theta}}(x_i),$$

quindi $\hat{\theta}$ è il parametro che rende massima la funzione di verosimiglianza di $(X_1, X_2, ..., X_n)$.

Si osservi che $\hat{\theta}$ è associato alla seguente equazione di stima (di cui è soluzione), detta *equazione di verosimiglianza*:

$$\sum_{i=1}^{n} \nabla \ln f_\theta(x_i) = 0,$$

dove ∇ indica la derivata rispetto a θ. Questo colloca la stima di massima verosimiglianza nelle seguenti classi di stimatori: $\hat{M}$, *M* e *GEE*.

La funzione $\nabla \ln f_\theta(x)$, detta *score function*, risulta essere di grande importanza nell'inferenza statistica. Innanzitutto valgono le seguenti proprietà.

$$E[\nabla \ln f_\theta(X)] = 0;$$

$$V(\nabla \ln f_\theta(X)) = E[(\nabla \ln f_\theta(X))^2] = E[-\nabla'' \ln f_\theta(X)],$$

dove $E[-\nabla'' \ln f_\theta(X)] = I(\theta)$ è detta *informazione di Fisher*.

Sia $(X_1, ..., X_n)$ una variabile di campionamento da una distribuzione F_θ: allora l'informazione di Fisher contenuta in un campione *n*-dimensionale è data da:

$$V(\nabla \ln L(\theta)) = nI(\theta).$$

Si dimostra che, sotto certe condizioni di regolarità di f_θ (Hogg, McKean, Craig, 2005, pag. 322) uno stimatore non-distorto T di θ soddisfa la seguente disuguaglianza:

$$V(T(F_n)) \geq \frac{1}{nI(\theta)},$$

dove $1/I(\theta)$ è detto *limite inferiore di Rao-Cramèr* della varianza di *T*.

Nel caso valga, in particolare, $V(T(F_n)) = \frac{1}{nI(\theta)}$, si dice che *T* è *efficiente*. Il rapporto tra la varianza di uno stimatore e il limite inferiore di Rao-Cramèr è rappresenta l'*efficienza* dello stimatore.

Data una successione T_n di stimatori di , al variare della numerosità campionaria *n*, si dice che T_n è *asintoticamente efficiente* se:

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right).$$

Si indichi con $\hat{\theta}_n$ la successione di stimatori di massima verosimiglianza, al variare della numerosità campionaria *n*. Si dimostra, sotto ad ulteriori condizioni di regolarità di f_θ (Hogg, McKean, Craig, 2005, pag.325), che lo stimatore di massima verosimiglianza è asintoticamente normale e asintoticamente efficiente.

Esistono definizioni di efficienza più generali. Secondo Rao, uno stimatore consistente T_n è *efficiente al primo ordine* se vale:

$$\sqrt{n}\left|T_n - \theta - \beta(\theta) Z_n(\theta)\right| \xrightarrow{p} 0,$$

dove $\beta(\theta)$ è una funzione che non coinvolge le osservazioni e $Z_n(\theta) = \nabla \ln L(\theta)$ (Kanji, 1983). E' dimostrato che l'efficienza di primo ordine implica l'efficienza asintotica, ma non è vero il viceversa.

Sempre Rao fornisce una definizione di efficienza più astatta di quella del primo ordine. Si misura l' *efficienza del secondo ordine* di uno stimatore con:

$$E_2(T_n) = \min_{\lambda, a} AV\left(Z_n(\theta) - a(\theta)[T_n - \theta] - \lambda(\theta)[T_n - \theta]^2\right),$$

dove α, λ sono funzioni di θ, mentre *AV* indica la varianza asintotica (Lindsay, 1994). In questo caso, un valore più grande di E_2 significa efficienza del secondo ordine minore.

Si dimostra che lo stimatore di massima verosimiglianza è efficiente al secondo ordine (quindi rende minimo il valore di E_2). Il valore minimo dipende dalla distribuzione e dalla parametrizzazione, ma è zero se l'oggetto di stima è la media di una distribuzione appartenente alla famiglia esponenziale.

Nel seguente capitolo verranno presentate le misure di *divergenza*, delle particolari distanze coerenti con un'eventuale relazione di dissomiglianza tra distribuzioni, epressa dalla (μ)-maggiorazione. Applicando quindi il metodo della minima distanza, basato su tali divergenze, si otterranno diversi tipi di stimatori.

Capitolo 4

Misure di divergenza e maggiorazione

4.1 Divergenza relativa

Siano F e G due distribuzioni appartenenti alla classe $\mathcal{F}_\nu$, assolutamente continue rispetto ad una misura ν e con densità, appartenenti alla corrispondente classe di funzioni $\mathcal{D}_\nu$, date da:

$$f = dF / d\nu ; \quad g = dG / d\nu .$$

Si supponga che F e G abbiano lo stesso supporto $S = \{x : f(x) > 0\} = \{x : g(x) > 0\}$.

Si indichi la classe di tali distribuzioni con $\mathcal{F}_\nu(S)$ e la rispettiva classe di densità con $\mathcal{D}_\nu(S)$.

Si osservi che:

$$\int_S \frac{g(x)}{f(x)} dF(x) = \int_S g(x) d\nu(x) = 1 = \int_S \frac{f(x)}{f(x)} dF(x), \ \forall\, g \in \mathcal{D}_\nu(S).$$

Inoltre, considerando che $f/f = I_S$, funzione indicatrice di S, ed indicando con μ_F la misura definita da F, vale:

$$\int_0^z I_S^{(\mu_F)}(t)dt = \int_0^z 1 dt \geq \int_0^z (g/f)^{(\mu_F)}(t)dt , \ \forall z \in [0,1], \ (\mu_F(S) = 1).$$

Quindi si ha che la funzione f/f è maggiorata da g/f, rispetto alla misura μ_F definita da F, cioè:

$$I_S = \frac{f}{f} \prec_{(\mu_F)} \frac{g}{f} .$$

Allora per $g \neq f$, applicando il teorema di μ-maggiorazione, per ogni funzione strettamente convessa ϕ, si ha:

$$\int_S \phi\left(\frac{f}{f}\right)dF < \int_S \phi\left(\frac{g}{f}\right)dF \text{ , per } f \neq g,$$

Quindi, ogni funzionale del tipo:

$$\Phi(F,G) = \int_S \phi\left(\frac{g}{f}\right)dF ,$$

dove ϕ è una funzione strettamente convessa, può essere interpretato come distanza tra le distribuzioni F e G, dato che vale:

$$\Phi(F,G) > \Phi(F,F) \quad \text{per } G \neq F.$$

Si considerino $G_1, G_2 \in \mathcal{F}_\nu(S)$, con corrispondenti densità rispetto a ν date da g_1, g_2. La relazione di μ-maggiorazione

$$\frac{g_1}{f} \prec_{(\mu_F)} \frac{g_2}{f},$$

indica la maggior "dissomiglianza relativa" di G_2, piuttosto che G_1, rispetto a F (Bertoli-Barsotti, 2002). Questa situazione è detta di *divergenza relativa* o *divergenza direzionale*. Come si è visto, funzionali del tipo $\Phi(F,G)$ preservano l'ordine di μ-maggiorazione, vale a dire:

$$\frac{g_1}{f} \prec_{(\mu_F)} \frac{g_2}{f} \Rightarrow \Phi(F,G_1) < \Phi(F,G_2).$$

Per questo motivo i funzionali del tipo $\Phi(F,G)$ vengono chiamati *misure di divergenza relativa* o più semplicemente *divergenze* su $\mathcal{F}_\nu(S)$.

La relazione di μ-maggiorazione rispetto a μ_F costituisce un pre-ordinamento su $\mathcal{F}_\nu(S)$, quindi non è sempre possibile confrontare due diverse distribuzioni G_1 e $G_2 \in \mathcal{F}_\nu(S)$. Laddove è possibile, però, si è visto che $\Phi(F,G)$ è coerente con tale confronto. Questo rende le misure di divergenza particolarmente interessanti e utili, rispetto agli altri tipi di distanza.

4.2 Stima della minima divergenza e maggiorazione

Le misure di divergenza possono essere utilizzate nell'inferenza parametrica su una distribuzione F_θ, con supporto *S*, appartenente alla famiglia $\mathcal{F}_\theta \subseteq \mathcal{F}_\nu(S)$.

Si ottengono due tipologie differenti di misure di divergenza tra F_θ e F_n a seconda che si ponga oppure $F = F_n$ e $G = F_\theta$ oppure $F = F_\theta$ e $G = F_n$ all'interno di $\Phi(F, G)$.

Analogamente a quanto avviene per gli altri tipi di distanza, una misura di divergenza su $\mathcal{F}_\nu(S)$ non gode necessariamente delle sue proprietà (coerenza con la μ-maggiorazione) sull'intero spazio $\mathcal{F}$ delle funzioni di distribuzione. Può accadere infatti che $F_n \notin \mathcal{F}_\nu(S)$: questo avviene se F_θ non è discreta, se il supporto di F_n è incluso nel supporto di F_θ o per entrambi motivi.

Come verrà mostrato in seguito, se $F_n \in \mathcal{F}_\nu(S)$, un funzionale con la forma $\Phi(F_n, F_\theta)$, rimane coerente con il pre-ordinamento di μ-maggiorazione. Inoltre, se $F_n \notin \mathcal{F}_\nu(S)$ ma la distribuzione F_θ è discreta, allora $\Phi(F_n, F_\theta)$ risulta essere coerente con il pre-ordinamento di μ-maggiorazione debole, sotto la condizione che ϕ sia decrescente, oltre che strettamente convessa (paragrafo *3.3.3*). Infine, nel paragrafo *3.3.4* si dimostrerà che anche un funzionale con forma $\Phi(F_\theta, F_n)$ gode, allo stesso modo, di queste proprietà (ovviamente le divergenze non sono distanze simmetriche).

Si supponga che F_θ sia assolutamente continua rispetto alla misura del conteggio ν e che il supporto della distribuzione empirica F_n coincida con il supporto *S* di F_θ, identico al variare di θ. Ovviamente questo è possibile solo nei casi in cui $\nu(S) < \infty$, e presuppone che la numerosità campionaria sia sufficientemente elevata da avere almeno un'osservazione per ogni punto di *S*. Si ha quindi che $F_n, F_\theta \in \mathcal{F}_\nu(S)$.

In questo caso è possibile stabilire un ordinamento di maggiorazione su $\mathcal{F}_\nu(S)$,.

Definendo quindi la misura di divergenza relativa tra F_n e F_θ con

$$\Phi(F_n, F_\theta) = \sum_{x \in S} \phi\left(\frac{f_\theta(x)}{f_n(x)}\right) f_n(x),$$

dove f_n è il rapporto tra la osservazioni che assumono lo stesso valore e la numerosità campionaria *n*, vale:

$$\frac{f_{\theta_1}}{f_n} \prec_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \Rightarrow \Phi(F_n, F_{\theta_1}) < \Phi(F_n, F_{\theta_2}),$$

dove $F_{\theta_1}, F_{\theta_2} \in \mathcal{F}_\theta$..

Si osservi che in questo caso $\mathcal{F}_\theta \subset \mathcal{F}_\nu$, $F_n \in \mathcal{F}_\nu$ e, solitamente, $F_n \notin \mathcal{F}_\theta$.

Questa proprietà suggerisce di stimare θ con il metodo della minima distanza rispetto a $\Phi(F_n, F_\theta)$, ottenendo lo stimatore $\hat{\theta}=H\,[\Phi(F_n)_{\mathcal{F}\theta}]$ tale che:

$$\Phi(F_n, F_{\hat{\theta}}) = \min_\theta \Phi(F_n, F_\theta), \quad F_\theta \in \mathcal{F}_\theta.$$

Chiaramente se si verifica che $F_n \in \mathcal{F}_\theta$, allora F_n costituisce la stima di F_θ ($F_{\hat{\theta}} = F_n$).

Quindi $F_{\hat{\theta}}$ è la distribuzione di $\mathcal{F}_\theta$ che diverge il meno possibile da F_n, rispetto alla misura di divergenza rappresentata dal funzionale Φ.

4.2 Divergenza e maggiorazione in senso debole

Nel paragrafo precedente sono state prese in considerazione solamente distribuzioni F_θ con supporto *S* di cardinalità finita, $\nu(S) < \infty$. Si supponga ora che $\nu(S) = \infty$ oppure che $\nu(S) < \infty$ ma la che numerosità campionaria *n* non sia sufficientemente elevata da avere almeno un'osservazione per ogni punto di *S*. Sotto queste condizioni, si ha che il supporto S_n della funzione di distribuzione empirica F_n è contenuto nel supporto *S* di F_θ, cioè $S_n \subset S$. In tal caso evidentemente non è possibile stabilire un pre-ordinamento di μ-maggiorazione

tra le distribuzioni di $\mathcal{F}_\theta$. Le distribuzioni F_θ e F_n non sono confrontabili, avendo due supporti diversi.

A maggior ragione si consideri che, nel caso continuo, calcolare la misura di divergenza tre F_θ e F_n è impossibile. In tal caso infatti le due distribuzioni differiscono sia per quanto riguarda il supporto che per quanto riguarda la misura (F_θ è assolutamente continua rispetto alla misura di Lebesgue, F_n rispetto alla misura del conteggio). Una semplice soluzione a questo inconveniente consiste nel raggruppare le osservazioni e "discretizzare" F_θ, riducendo il problema ad un confronto tra distribuzioni discrete con uguale supporto. Questo verrà mostrato in seguito (capitolo *5.2*) con riferimento alle tre misure di divergenza più importanti. Le eltre possibili estenzioni al caso continuo saranno illustrate nel capitolo *5.4*.

Tornando al caso discreto, siano $F_{\theta_1}, F_{\theta_2} \in \mathcal{F}_\theta$. Si è visto che, se $S_n \subset S$, allora $F_n \notin \mathcal{F}_V(S)$: l'impossibilità di stabilire una maggiorazione tra f_{θ_1}/f_n e f_{θ_2}/f_n è dovuta al fatto che, generalmente:

$$\int_{S_n} \frac{f_{\theta_1}}{f_n} dF_n = \sum_{x \in S_n} f_{\theta_1}(x) \neq \sum_{x \in S_n} f_{\theta_2}(x) = \int_{S_n} \frac{f_{\theta_2}}{f_n} dF_n ,$$

mentre ovviamente $\sum_{x \in S_n} f_n = 1$.

Definendo:

$$\sum_{x \in S_n} f_\theta(x) = c(\theta) \leq 1 ,$$

è possibile confrontare due distribuzioni solamente nella classe $\mathcal{F}(c_\theta)$ delle distribuzioni F_θ tali che $\sum_{x \in S_n} f_\theta = c_\theta$.

Tuttavia è possibile stabilire un pre-ordniamento di μ-maggiorazione debole tra f_{θ_1}/f_n e f_{θ_2}/f_n. Infatti si sa che:

$$\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} ,$$

se e solo se:

$$\int_0^x \frac{f_{\theta_1}}{f_n} \overset{(\mu_{F_n})}{\uparrow} dt \geq \int_0^x \frac{f_{\theta_2}}{f_n} \overset{(\mu_{F_n})}{\uparrow} dt, \quad \forall x \in [0,1].$$

Si può avere quindi che $\sum_{x \in S_n} f_{\theta_1}(x) \geq \sum_{x \in S_n} f_{\theta_2}(x)$.

Allora , se ϕ è strettamente convessa e decrescente:

$$\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \quad \Rightarrow \quad \Phi(F_n, F_{\theta_1}) < \Phi(F_n, F_{\theta_2}),$$

dove $\Phi(F_n, F_\theta)$ è definito ovviamente su $S_n \subset S$, cioè:

$$\Phi(F_n, F_\theta) = \sum_{x \in S_n} \phi\left(\frac{f_\theta(x)}{f_n(x)}\right) f_n(x).$$

Questo suggerisce l'utilizzo di funzionali con ϕ decrescente oltre che strettamente convessa. Infatti se ϕ non è decrescente non è detto che la precedete relazione sia verificata. Ad esempio, ponendo $\phi(x) = -\ln(x)$, si ottiene un funzionale (distanza di Kullback-Leibler) coerente rispetto alla μ-maggiorazione debole.

Si osservi che $\sum_{x \in S_n} f_\theta(x) = P_\theta(X \in S_n)$. La μ-maggiorazione debole (da sopra) comporta una condizione per la quale si perferisce la distribuzione che a parità di "somiglianza relativa" con F_n presenta una maggior valore di $P_\theta(X \in S_n)$.

4.3 Misure di divergenza con formula "invertita"

Ponendo $F = F_\theta$ e $G = F_n$ all'interno di $\Phi(F, G)$, si ottengono alcune note divergenze, caratterizzate dalla seguente forma:

$$\Phi(F_\theta, F_n) = \sum_{x \in S} \phi\left(\frac{f_n(x)}{f_\theta(x)}\right) f_\theta(x),$$

dato che $S_n \subset S$ si considerino solo funzioni ϕ tali che $\phi\left(f_n(x)/f_\theta(x)\right) < \infty$.

In questo caso i "pesi" sono dati da F_θ, non più da F_n, il che comporta una complicazione.

Considerando infatti due distribuzioni F_{θ_1} e $F_{\theta_2} \in \mathcal{F}_\theta$, non si può stabilire un ordinamento di μ-maggiorazione tra f_n / g_{θ_1} e f_n / g_{θ_2} rispetto alla misura (variabile) definita da F_θ, dato che quest'ultima dipende ovviamente da θ.

Tuttavia, come verrà mostrato in seguito, un funzionale con forma $\Phi(F_\theta, F_n)$, rispetta comunque il pre-ordinamento di maggiorazione debole (da sopra) tra le distribuzioni. Questa è una conseguenza del prossimo teorema.

Teorema 4.3.A

Siano G e H due distribuzioni discrete con supporto comune S e sia F discreta con supporto $S' \subset S$. Siano g, h, f, le rispettive densità rispetto alla misura del conteggio ν. Allora:

$$\int_0^x \frac{g}{f}^{(\mu_F)\uparrow} dt \geq \int_0^x \frac{h}{f}^{(\mu_F)\uparrow} dt, \text{ per } x \in [0,1],$$

se e solo se

$$\int_0^x \frac{f}{g}^{(\mu_G)\uparrow} dt \geq \int_0^x \frac{f}{h}^{(\mu_H)\uparrow} dt \text{ , per } x \in [0,1].$$

Dimostrazione

Innanzitutto si osservi che $\frac{f}{g}$ e $\frac{f}{g}^{(\mu_G)\uparrow}$ sono equimisurabili, cioè:

$$\mu_G\left\{t \in S : z_1 < \frac{f}{g}(t) < z_2\right\} = m\left\{t \in [0,1] : z_1 < \frac{f}{g}^{(\mu_G)\uparrow}(t) < z_2\right\},$$

dove m è la misura di Lebesgue. Lo stesso vale chiaramente per $\frac{f}{h}$ e $\frac{f}{h}^{(\mu_H)\uparrow}$.

Allora, considerando che $f = 0$ su $S - S'$, è evidente che:

$$\int_0^1 \frac{f}{g}^{(\mu_G)}\uparrow dt = \int_S \frac{f}{g}dG = \int_{S'} \frac{f}{g}dG = 1 = \int_{S'} \frac{f}{h}dH = \int_S \frac{f}{h}dH = \int_0^1 \frac{f}{h}^{(\mu_H)}\uparrow dt \,.$$

Si può osservare in modo piuttosto analogo che, per ipotesi, $1 \geq \mu_G(S') \geq \mu_H(S')$.

Si indichi con k la dimensione di S', cioè $\nu(S') = k$.

E' possibile rappresentare l'insieme $S \cap S' = S'$ con il vettore $(w_1, w_2, ..., w_k)$, le cui componenti sono i punti tali che:

$$\frac{g(w_j)}{f(w_j)} \geq \frac{g(w_{j-1})}{f(w_{j-1})}, \text{ per } i = 2,...,k \,;$$

oppure come il vettore $(z_1, z_2, ..., z_k)$, tale che:

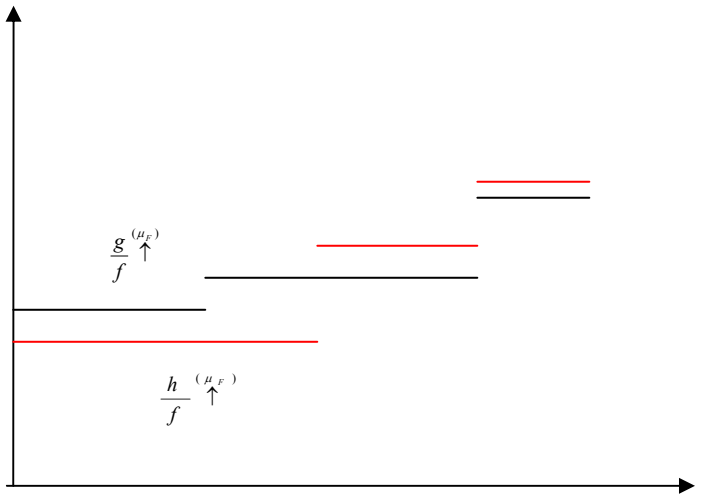
$$\frac{h(z_j)}{f(z_j)} \geq \frac{h(z_{j-1})}{f(z_{j-1})}, \text{ per } j = 2,...,k \,.$$

Allora i riordini crescenti di g/f e di h/f rispetto a μ_F sono funzioni monotone non decrescenti e "a gradini" su [0,1], date rispettivamente da:

$$\frac{g}{f}^{(\mu_F)}\uparrow(t) = \frac{g(w_i)}{f(w_i)} \quad \text{per} \quad t \in [\sum_{m=1}^{i-1} f(w_m), \sum_{m=1}^{i} f(w_m)], \quad i = 1,...,k.;$$

$$\frac{h}{f}^{(\mu_F)}\uparrow(t) = \frac{h(z_j)}{f(z_j)} \quad \text{per} \quad t \in [\sum_{m=1}^{j-1} f(z_m), \sum_{m=1}^{j} f(z_m)], \quad j = 1,...,k.$$

dove $f(w_0) = f(z_0) = 0$ e ovviamente $\sum_{m=1}^{k} f(w_m) = \sum_{m=1}^{k} f(z_m) = 1$.



L'integrale tra 0 e x di una funzione monotona non decrescente e "a gradini" risulta essere una funzione continua, monotona, strettamente crescente e lineare a tratti su [0,1]. Quindi si ottiene rispettivamente:

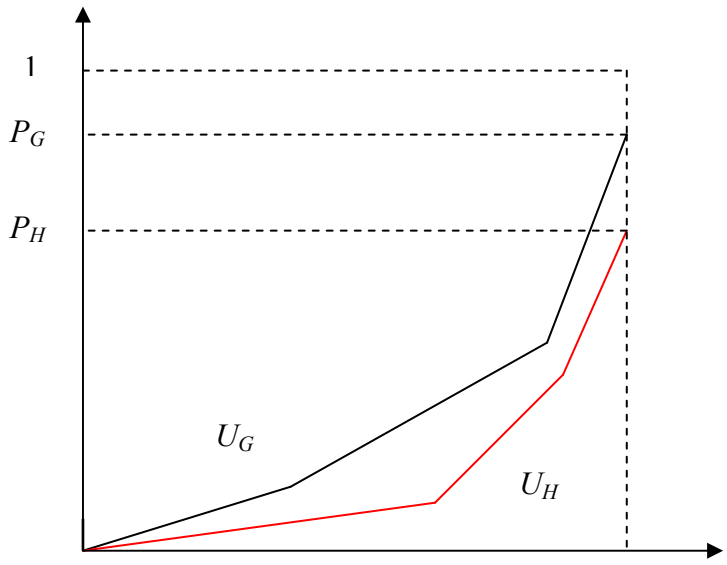
$$U_G(x) = \int_0^x \frac{g}{f}^{(\mu_F)\uparrow} dt = g(w_{i-1}) + \frac{g(w_i)}{f(w_i)}x\,,$$

$$\text{per} \quad t \in [\sum_{m=1}^{i-1} f(w_m), \sum_{m=1}^{i} f(w_m)]\,,\ i = 1,...,k.;$$

$$U_H(x) = \int_0^x \frac{h}{f}^{(\mu_F)\uparrow} dt = h(z_{j-1}) + \frac{g(z_j)}{f(z_j)}x\,,$$

$$\text{per} \quad t \in [\sum_{m=1}^{j-1} f(z_m), \sum_{m=1}^{j} f(z_m)]\,,\ j = 1,...,k.,$$

dove $g(w_0) = h(z_0) = 0$ e, si ricordi, $1 \geq \mu_G(S') = G(w_k) \geq \mu_H(S') = H(z_k)$.
Per comodità, nei grafici si scrive $\mu_G(S') = P_G$ e $\mu_H(S') = P_H$.



$U_G(x)$ è sicuramente invertibile su [0,1], la sua inversa $(U_G)^{-1}(y)$ è una funzione continua, monotona, strettamente crescente e lineare a tratti su [0, $\mu_G(S')$] (infatti si osservi che $\mu_G(S') = U_G(1)$).

$$(U_G)^{-1}(y) = f(w_{i-1}) + \frac{f(w_i)}{g(w_i)}y \quad \text{per } y \in [\sum_{m=1}^{i-1} g(w_m), \sum_{m=1}^{i} g(w_m)], \quad j = 1,...,k.$$

Si consideri ora il riordino decrescente di f/g rispetto alla misura μ_G. Si osservi innanzi tutto che:

$$\frac{g(w_i)}{f(w_i)} \geq \frac{g(w_{i-1})}{f(w_{i-1})} \Leftrightarrow \frac{f(w_{i-1})}{g(w_{i-1})} \geq \frac{f(w_i)}{g(w_i)}, \text{ per } i = 2,...,k .$$

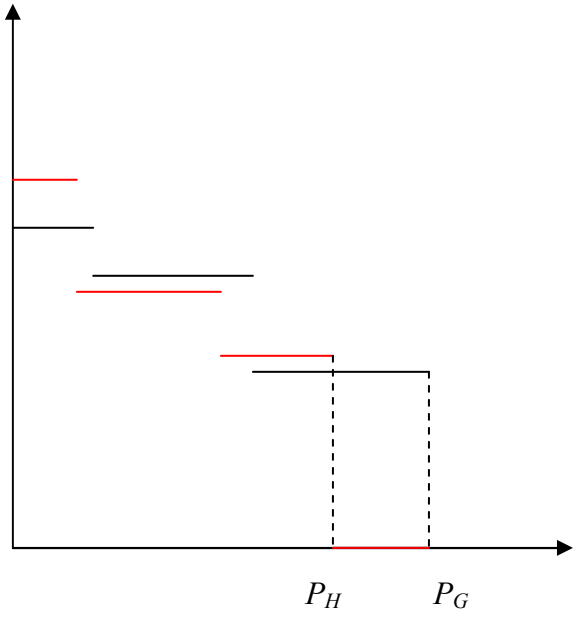
Quindi si ottiene:

$$\frac{f}{g}^{(\mu_G)}\downarrow(t) = \begin{cases} \dfrac{f(w_i)}{g(w_i)} & t \in [\sum_{m=1}^{i} g(w_m), \sum_{m=1}^{i-1} g(w_m)] \\ 0 & t \in [\mu_G(S'),1] \end{cases}, \quad i = 1,...,k ,$$

funzione "a gradini" monotòna non crescente.

Analogamente:

$$\frac{f}{h}^{(\mu_H)}\downarrow(t) = \begin{cases} \dfrac{f(w_i)}{h(w_i)} & t \in [\sum_{m=1}^{i} h(z_m), \sum_{m=1}^{i-1} h(z_m)] \\ 0 & t \in [\mu_H(S'),1] \end{cases}, \quad i = 1,...,k ,$$



L'integrale tra 0 e y di $\frac{f}{g}^{(\mu_G)}\downarrow$ risulta:

$$V_G(y) = \int_0^y \frac{f}{g}^{(\mu_G)}\downarrow dt = \begin{cases} f(w_{i-1}) + \dfrac{f(w_i)}{g(w_i)} y & y \in [\sum_{m=1}^{i-1} g(w_m), \sum_{m=1}^{i} g(w_m)] \\ 1 & y \in [\mu_G(S'),1] \end{cases},$$

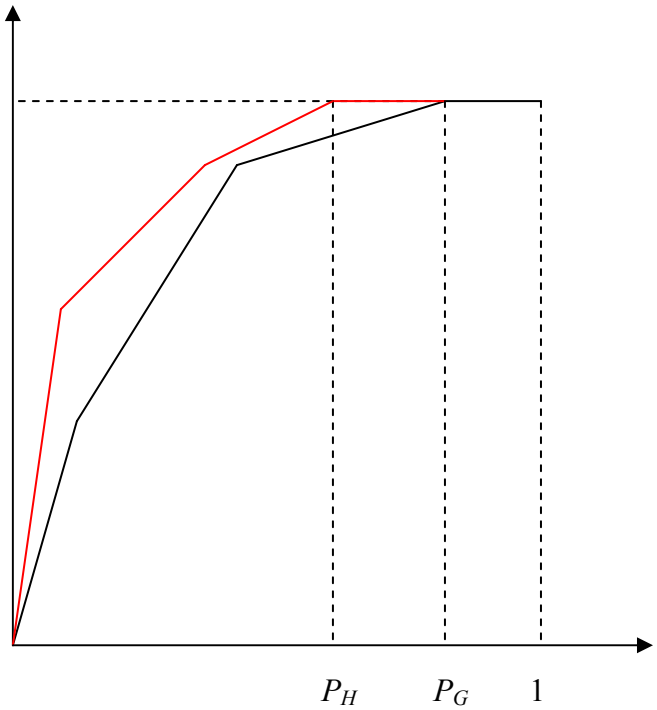
$$i = 1,...,k.$$

In altri termini:

$$V_G(y) = \begin{cases} (U_G)^{-1}(y) & y \in [0, \mu_G(S')] \\ 1 & y \in [\mu_G(S'), 1] \end{cases}.$$

Analogamente, per H si ha:

$$V_H(y) = \begin{cases} (U_H)^{-1}(y) & y \in [0, \mu_H(S')] \\ 1 & y \in [\mu_H(S'), 1] \end{cases}.$$



$U_G(x) \geq U_H(x)$, per $x \in [0,1]$, se e solo se $(U_G)^{-1}(y) \leq (U_H)^{-1}(y)$, per $y \in [0, \mu_H(S')]$. Questo è equivalente a $V_G(y) \leq V_H(y)$ per $y \in [0, 1]$, dato che $V_H(y) = 1$ per $y \in [\mu_H(S'), \mu_G(S')]$ e che $V_G(y) = V_H(y) = 1$ per $y \in [\mu_G(S'), 1]$.

Ma poiché $V_G(1) = V_H(1) = 1$:

$$\int_0^y \frac{f}{g}^{(\mu_G)\downarrow} dt \leq \int_0^y \frac{f}{h}^{(\mu_H)\downarrow} dt \quad \Leftrightarrow \quad \int_0^y \frac{f}{g}^{(\mu_G)\uparrow} dt \geq \int_0^y \frac{f}{h}^{(\mu_H)\uparrow} dt, \quad \text{per } y \in [0, 1],$$

il che dimostra il teorema.

Si torni ora sui funzionali con forma $\Phi(F_\theta, F_n)$. Si sa che:

$$\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n},$$

se e solo se:

$$\int_0^x \frac{f_{\theta_1}}{f_n}^{(\mu_{F_n})\uparrow} dt \geq \int_0^x \frac{f_{\theta_2}}{f_n}^{(\mu_{F_n})\uparrow} dt, \quad \forall x \in [0,1].$$

Per il teorema precedente, l'ultima condizione è equivalente a:

$$\int_0^x \frac{f_n}{f_{\theta_1}}^{(\mu_{F_{\theta_1}})\uparrow} dt \geq \int_0^x \frac{f_n}{f_{\theta_2}}^{(\mu_{F_{\theta_2}})\uparrow} dt, \quad \forall x \in [0,1], \quad \int_0^1 \frac{f_n}{f_{\theta_1}}^{(\mu_{F_{\theta_1}})\uparrow} dt = \int_0^1 \frac{f_n}{f_{\theta_2}}^{(\mu_{F_{\theta_2}})\uparrow} dt.$$

$\frac{f_n}{f_\theta}$ e $\frac{f_n}{f_\theta}^{(\mu_{F_\theta})\uparrow}$ sono equimisurabili, cioè:

$$\mu_{F_\theta}\left\{t \in S : z_1 < \frac{f_n}{f_\theta}(t) < z_2\right\} = m\left\{t \in [0,1] : z_1 < \frac{f_n}{f_\theta}^{(\mu_{F_\theta})\uparrow}(t) < z_2\right\},$$

dove m è la misura di Lebesgue. Inoltre $\frac{f_n}{f_{\theta_i}}^{(\mu_{F_{\theta_i}})\uparrow}$, i=1,2, sono funzioni non decrescenti e continue da sinistra in [0,1] .

Sia ϕ una funzione strettamente convessa. Allora è possibile applicare il teorema di Karamata, per cui si ottiene:

$$\int_0^1 \phi\left(\frac{f_n}{f_{\theta_1}}^{(\mu_{F_{\theta_1}})\uparrow}\right) dt \geq \int_0^1 \phi\left(\frac{f_n}{f_{\theta_2}}^{(\mu_{F_{\theta_2}})\uparrow}\right) dt.$$

Data l'equimisurabilità tra $\frac{f_n}{f_\theta}$ e $\frac{f_n}{f_\theta}^{(\mu_{F_\theta})\uparrow}$, è possibile riscrivere l'ultima disequazione nel seguente modo:

$$\int_S \phi\left(\frac{f_n}{f_{\theta_1}}\right) dF_{\theta_1} \geq \int_S \phi\left(\frac{f_n}{f_{\theta_2}}\right) dF_{\theta_2}.$$

Quindi, riassumendo, se ϕ è una funzione strettamente convessa,

$$\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \quad \Rightarrow \quad \int_S \phi\left(\frac{f_n}{f_{\theta_1}}\right) dF_{\theta_1} \geq \int_S \phi\left(\frac{f_n}{f_{\theta_2}}\right) dF_{\theta_2} .$$

Alcune tra le più note divergenze assumono questa forma, ad esempio Chi^2 e Hellinger, come si vedrà nel capitolo seguente.

Capitolo 5

Stimatori basati su misure di divergenza

5.1 Misure di divergenza potenziate

Siano F e G due distribuzioni discrete con uguale supporto $S = \{x_1,..,x_k\}$. Un'importante classe di misure di divergenza è la famiglia delle *misure di divergenza potenziate* ("*power divergence*", *PWD*; Cressie, Read, 1984) definita da:

$$PWD_\lambda(F,G) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k} f(x_i)\left[\left(\frac{f(x_i)}{g(x_i)}\right)^{\lambda} - 1\right], \qquad \lambda \in \Re .$$

Si verifica anche (Cressie, Read,1984; Lindsay, 1994)che:

$$PWD_\lambda(F,G) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k} g(x_i)\left[\left(1+\frac{g(x_i)-f(x_i)}{f(x_i)}\right)^{\lambda+1} - 1\right], \qquad \lambda \in \Re .$$

Per comodità, è preferibile utilizzare la prima formula per $\lambda < 0$ e la seconda per $\lambda \geq 0$.

Si osservi che per $\lambda = 0$ o per $\lambda = -1$ la precedente espressione non è definita. Tuttavia, ricordando il limite notevole $\ln a = \lim_{x\to 0}(a^x - 1)/x$, si possono calcolare i limiti di $PWD_\lambda(F,G)$ per $\lambda \to 0$ e per $\lambda \to -1$, ottenendo:

$$PWD_0(F,G) = \lim_{\lambda\to 0} PWD_\lambda(F,G) = -\sum_{i=1}^{k} f(x_i)\ln\left(\frac{g(x_i)}{f(x_i)}\right),$$

$$PWD_{-1}(F,G) = \lim_{\lambda\to -1} PWD_\lambda(F,G) = -\sum_{i=1}^{k} g(x_i)\ln\left(\frac{f(x_i)}{g(x_i)}\right).$$

Ponendo $F = F_n$ e $G = F_\theta \in \mathcal{F}_\theta$, si definisce la stima di minima distanza rispetto alla divergenza potenziata, data dal valore del parametro θ che minimizza:

$$PWD_\lambda(F_n,F_\theta) = \frac{1}{\lambda(\lambda+1)}\sum_{i=1}^{k} f_n(x_i)\left[\left(\frac{f_n(x_i)}{f_\theta(x_i)}\right)^{\lambda} - 1\right],$$

o equivalentemente:

$$PWD_\lambda(F_n, F_\theta) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} f_\theta(x_i) \left[\left(1 + \frac{f_n(x_i) - f_\theta(x_i)}{f_\theta(x_i)} \right)^{\lambda+1} - 1 \right].$$

La funzione $\phi(x) = \frac{x^\lambda - 1}{\lambda(\lambda+1)}$ è strettamente convessa (laddove è definita). Quindi applicando i risultati ottenuti nel precedente capitolo, si ha che, se il supporto S_n di F_n coincide con il supporto S di F_θ, allora:

$$\frac{f_{\theta_1}}{f_n} \prec_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \Rightarrow PWD_\lambda(F_n, F_{\theta_1}) < PWD_\lambda(F_n, F_{\theta_2}),$$

dove $F_{\theta_1}, F_{\theta_2} \in \mathcal{F}_\theta$.

Quindi le misure di divergenza potenziate sono coerenti con il pre-ordinamento di maggiorazione (forte) se S_n coincide con S.

Si supponga ora che S_n non coincida necessariamente con S, ma che $S_n \subset S$ (ipotesi senz'altro più realistica). Per $\lambda < 0$ la funzione $\phi(x) = \frac{x^\lambda - 1}{\lambda(\lambda+1)}$ è strettamente convessa e decrescente (laddove esiste). Inoltre la prima formula di $PWD_\lambda(F_n, F_\theta)$, se $\lambda < 0$, è un funzionale con forma $\Phi(F_\theta, F_n)$. Allora:

$$\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \Rightarrow PWD_\lambda(F_n, F_{\theta_1}) < PWD_\lambda(F_n, F_{\theta_2}), \quad \text{per } \lambda < 0,$$

dove $F_{\theta_1}, F_{\theta_2} \in \mathcal{F}_\theta$.

Si consideri ora la seconda formula e si osservi che può essere riformulata così:

$$PWD_\lambda(F_n, F_\theta) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} f_\theta(x_i) \left[\left(\frac{f_n(x_i)}{f_\theta(x_i)} \right)^{\lambda+1} - 1 \right].$$

Considerando che $\phi(x) = \frac{x^{\lambda+1} - 1}{\lambda(\lambda+1)}$ è una funzione strettamente convessa e che, espressa in questo modo, $PWD_\lambda(F_n, F_\theta)$, è un funzionale con forma $\Phi(F_n, F_\theta)$, allora:

$$\frac{f_{\theta_1}}{f_n} \prec^{w_{(\mu_{F_n})}} \frac{f_{\theta_2}}{f_n} \Rightarrow PWD_\lambda(F_n, F_{\theta_1}) < PWD_\lambda(F_n, F_{\theta_2}), \quad \text{per } \lambda \in \Re .$$

Si conclude quindi che le divergenze potenziate sono funzionali in generale coerenti rispetto al pre-ordinamento di maggiorazione debole (da sopra). Gran parte delle misure di divergenza più importanti appartengono a questa classe.

5.2 Principali misure di divergenza e stimatori corrispondenti

5.2.1 Kullback-Leibler

Ponendo $\phi(x) = -\ln(x)$ all'interno di $\Phi(F,G)$ si ottiene la distanza di *Kullback-Leibler* tra le distribuzioni *F* e *G*, data da:

$$\rho_1(F,G) = \int_S -\ln\frac{g}{f}dF = \int_S -\ln\left(\frac{g}{f}\right) f d\nu .$$

Dove F , $G \in \mathcal{F}_0$, classe delle distribuzioni assolutamente continue rispetto ad una misura ν.

Questa divergenza appartiene alla classe delle divergenza potenziate, ponendo $\lambda = -1$ oppure con $\lambda = 0$ nella formula $PWD_\lambda(F,G)$. Si osservi che a seconda di questi due valori di λ si ottengono stimatori basati su divergenze di tipo $\Phi(F_\theta, F_n)$ oppure di tipo $\Phi(F_n, F_\theta)$. In ogni caso, essendo ϕ decrescente e strettamente convessa, si ottiene un funzionale con le proprietà studiate nei paragrafi precedenti.

Nel caso $\lambda = 0$ (il più utilizzato in ambito di stima), si ottiene uno stimatore strettamente legato allo stimatore di massima verosimiglianza, come verrà mostrato in seguito, talvolta infatti tale divergenza è detta *likelihood disparity* (divergenza di verosimiglianza) (Lindsay, 1994). Nel caso $\lambda = -1$ invece si ottiene un funzionale statistico detto *modified likelihood statistic* (Kanji, 1983; Cressie, Read, 1984), senz'alltro meno utilizzato, che non verrà approfondito nel seguito

Si osservi che $\rho_1(F,G)=0$ per $F = G$ e che quindi, ovviamente, si ha $\rho_1(F,G)>0$ per $F \neq G$.

Sia $\mathcal{F}_\theta$, con $\theta \in \Theta$, una famiglia parametrica di funzioni di distribuzione. Si è interessati alla stima di θ secondo il metodo della minima distanza, rispetto a ρ_1.
Si consideri che:

$$\rho_1(F,F_\theta)=\int_S \ln\frac{dF}{d\nu}dF-\int_S \ln f_\theta dF\ ,$$

dove il secondo addendo, definito precedentemente (massima verosimiglianza, paragrafo *3.2.2*) con $d_2(F_\theta,F)$, è l'unico a dipendere da θ.

Si supponga in un primo momento che la distribuzione F_θ sia discreta e, quindi, che ν sia la misura del conteggio. Allora, sostituendo ad F la funzione di distribuzione empirica F_n, si ottiene:

$$\rho_1(F_n,F_\theta)=-\int_S \ln\frac{dF_\theta}{d\nu}dF_n+\int_S \ln\frac{dF_n}{d\nu}dF_n=d_2(F_\theta,F_n)-d_2(F_n,F_n)\,.$$

Dato che solo il primo addendo dipende da θ, si noti che la stima della minima distanza rispetto a ρ_1 equivale a determinare il valore θ che minimizza $d_2(F_\theta,F_n)$, che coincide con lo stimatore di massima verosimiglianza.

Dato che –ln è una funzione strettamente convessa e decrescente, nel caso discreto $\rho_1(F_n,F_\theta)$ è coerente con la maggiorazione se il supporto di F_n coincide con *S*, altrimenti è coerente con la maggiorazione debole.

Se F_θ è continua, è evidentemente impossibile calcolare $\rho_1(F_n,F_\theta)$, dato che F_n e F_θ non sono assolutamente continue rispetto ad una comune misura ν. Infatti si ha che F_n è assolutamente continua rispetto alla misura del conteggio, mentre F_θ è assolutamente continua rispetto alla misura di Lebesgue, λ.

Tuttavia è possibile, approssimando F_θ con una distribuzione discreta, generalizzare il risultato ottenuto nel caso discreto anche al caso continuo.

Innanzitutto si indichi con P_θ la misura di probabilità corrispondente a F_θ. Sia $\Delta_1,\Delta_2,....,\Delta_N$ una partizione del supporto di F_θ e si indichi con n_i il numero di osservazioni campionarie appartenenti all'insieme Δ_i ($P_n(\Delta_i)=n_i/n$). Si

mettono dunque a confronto le due distribuzioni di probabilità discrete rappresentate da $P_n(\Delta_i)$ e $P_\theta(\Delta_i)$. Allora, come stimatore della minima distanza rispetto a ρ_1, si potrebbe scegliere il valore θ che minimizza:

$$\sum_{i=1}^{N} -\ln\left(\frac{P_\theta(\Delta_i)}{n_i/n}\right)\frac{n_i}{n}.$$

Anche nel caso continuo, grazie ad una opportuna approssimazione di F_θ con una distribuzione discreta ${F_\theta}^N$, è possibile ricondurre la stima di minima distanza rispetto a Kullback-Leibler alla stima di massima verosimiglianza (Borovkov, 1998, Mathematical Statistics, pag. 184).

Si supponga che per gli insiemi $\Delta_1, \Delta_2, ..., \Delta_N$, valga l'approssimazione:

$$f_\theta(x) \sim f_\theta(y_i), \text{ per } x \in \Delta_i,$$

dove $y_i \in \Delta_i$, i =1,2,….,N.

Si supponga inoltre che:

$$\sum_{i=1}^{N} f_\theta(y_i)\lambda(\Delta_i) = \int f_\theta(x)d\lambda(x) = 1$$

A questo punto è possibile definire una distribuzione discreta ${F_\theta}^N$ e la corrispondente misura di probabilità ${P_\theta}^N$, avente come supporto l'insieme dei punti $y_1, y_2, ..., y_N$ e tale che:

$${P_\theta}^N(y_i) = f_\theta(y_i)\lambda(\Delta_i).$$

L'approssimazione assunta precedentemente implica che valga anche la seguente:

$$P_\theta(\Delta_i) = \int_{\Delta_i} f_\theta(x)d\lambda(x) \sim f_\theta(y_i)\lambda(\Delta_i) = {P_\theta}^N(y_i),$$

riconducibile tra l'altro alla prima delle due approssimazioni che giustificano la definizione di funzione di verosimiglianza come prodotto di densità (nel caso continuo), come si è visto nel capitolo *3.2.2*.

Raggruppando quindi le osservazioni provenienti dagli insiemi Δ_i, è possibile costruire la funzione di distribuzione empirica corrispondente a ${F_\theta}^N$, indicata con ${F_n}^N$, e la misura di probabilità associata a ${F_n}^N$, indicata con ${P_n}^N$. La misura di probabilità di un singolo insieme Δ_i sarà data dal rapporto tra il numero di osservazioni appartenenti a Δ_i, che verrà chiamato n_i, e la numerosità campionaria n, ossia: ${P_n}^N(y_i) = P_n(\Delta_i) = n_i / n$.

Si noti ora che minimizzare la distanza di Kullback-Leibler tra le distribuzioni ${F_\theta}^N$ e ${F_n}^N$, $\rho_1({F_n}^N, {F_\theta}^N)$, equivale a minimizzare:

$$d_2({F_n}^N, {F_\theta}^N) = \int -\ln\left({P_\theta}^N(y_i)\right) d{F_n}^N = \int -\ln\left(f_\theta(y_i)\lambda(\Delta_i)\right) d{F_n}^N =$$

$$= \int -\ln f_\theta(y_i)\, d{F_n}^N + \int -\ln \lambda(\Delta_i)\, d{F_n}^N =$$

$$= \sum_{i=1}^{N} -\ln f_\theta(y_i) P_n(\Delta_i) + \sum_{i=1}^{N} -\ln \lambda(\Delta_i) P_n(\Delta_i).$$

Dato che il secondo addendo non dipende da θ, è sufficiente minimizzare il primo addendo.

Si ricordi che $f_\theta(y_i) \sim f_\theta(x)$, per $x \in \Delta_i$. Inoltre, al crescere di N, si noti che ${F_n}^N \to F_n$. Allora vale questa ulteriore approssimazione:

$$\sum_{i=1}^{N} -\ln f_\theta(y_i) P_n(\Delta_i) \sim \int -\ln f_\theta(x)\, dF_n = d_2(F_n, F_\theta),$$

riconducibile alla seconda delle due approssimazioni che giustificano la definizione di funzione di verosimiglianza come prodotto di densità (nel caso continuo).

Questo risultato può anche essere spiegato nella maniera seguente. Si indichi con $\hat{\theta}_N$ lo stimatore che minimizza la distanza $\rho_1({F_n}^N, {F_\theta}^N)$. Allora, al crescere di N, $\hat{\theta}_N$ si avvicina sempre di più a $\hat{\theta}$, stimatore della minima distanza rispetto a $d_2(F_n, F_\theta)$, equivalentemente detto stimatore di massima verosimiglianza.

Si osservi in conclusione che, nel caso continuo, gli argomenti che giustificano teoricamente l'utilizzo del metodo della massima verosimiglianza permettono anche di interpretare la stima di massima verosimiglianza come stima della minima distanza rispetto a Kullback-Leibler. Sostanzialmente, infatti, queste argomentazioni si trattano di semplificazioni basate sulle stesse due approssimazioni. E' importante tenerne conto, sia perché ogni tipo di approssimazione comporta ovviamente una perdita di precisione, sia perché le approssimazioni in questione possono non essere valide in alcune circostanze (Cheng, Iles 1987; Cheng, Amin 1983).

5.2.2 Chi-quadrato

Ponendo $\phi(x) = (x-1)^2$ all'interno di $\Phi(F,G)$ si ottiene la divergenza di *Chi^2* (χ^2) tra le distribuzioni *F* e *G*, data da:

$$\rho_2(F,G) = \int_S \left(\frac{g}{f} - 1\right)^2 dF = \int_S \frac{(g-f)^2}{f} d\nu .$$

Dove F , $G \in \mathcal{F}_0$, classe delle distribuzioni assolutamente continue rispetto ad una misura ν.

Anche χ^2 appartiene alla classe delle divergenze potenziate, ponendo nella formula $PWD_\lambda(F,G)$ $\lambda = 1$ oppure con $\lambda = -2$. Si parla rispettivamente di χ^2 *di Pearson* (senz'altro più conosciuto) oppure di χ^2 *modificato di Neyman* (Neyman, 1949; Kanji, 1983).

ρ_2 gode delle proprietà viste per ρ_1, infatti si noti che $\rho_2(F,G) = 0$ per $F = G$ e $\rho_2(F,G) > 0$ per $F \neq G$.

Sia $\mathcal{F}_\theta$, con $\theta \in \Theta$, una famiglia parametrica di funzioni di distribuzione. Si è interessati alla stima di θ secondo il metodo della minima distanza, rispetto a ρ_2, nella forma di Pearson (la più importante ed utilizzata). Il funzionale da minimizzare, in questo caso, è posto nella forma $\Phi(F_\theta, F_n)$ ed è quindi coerente con l'ordinamento di maggiorazione debole (da sopra).

Chiaramente, $\rho_2(F_\theta, F_n)$ è determinabile solo se F_θ è discreta. Infatti, come si è già visto per ρ_1, per poter calcolare $\rho_2(F_\theta, F_n)$, F_n e F_θ devono essere assolutamente continue rispetto ad una comune misura ν, cosa impossibile se F_θ è continua.

Si ipotizzi quindi che F_θ sia discreta, con supporto *S* dato dai punti x_i, *i*=1,2,…. e si indichi con n_i il numero di osservazioni campionarie corrispondenti al punto x_i. In tal modo si ha che $P_\theta(\{x_i\}) = f_\theta(x_i)$ e $P_n(\{x_i\}) = n_i/n = f_n(x_i)$. Si osservi che, nei punti di *S* in cui non è stato osservato nessun valore, si ha $f_n = 0$.

Si chiama stimatore del *minimo* χ^2 (di Pearson) il valore di θ che minimizza:

$$\rho_2(F_\theta, F_n) = \sum_{x \in S} \frac{(f_n(x) - f_\theta(x))^2}{f_\theta(x)} .$$

La proprietà di coerenza rispetto alla maggiorazione debole (da sopra) si può intuire meglio riscrivendo così la formula:

$$\rho_2(F_\theta, F_n) = \sum_{x \in S_n} \frac{(f_n(x) - f_\theta(x))^2}{f_\theta(x)} + (1 - P_\theta(x \in S_n)) .$$

Quindi tanto più cresce $P_\theta(x \in S_n)$ tanto più diminuisce la distanza. Questo è evidente se si osserva il secondo addendo della precedente formula, ma è vero anche per quanto riguarda il primo addendo.

Si consideri un punto $x' \in S_n$ e due distribuzioni discrete F_{θ_1} e F_{θ_2}. Si supponga che

$$(f_n(x') - f_{\theta_1}(x'))^2 = (f_n(x') - f_{\theta_2}(x'))^2 ,$$

ma che $f_{\theta_1}(x') > f_{\theta_2}(x')$.

Attribuendo un maggior peso ad x', F_{θ_1} fornisce un contributo inferiore alla distanza. Infatti si ha che :

$$(f_n(x') - f_{\theta_1}(x'))^2 / f_{\theta_1}(x') < (f_n(x') - f_{\theta_2}(x'))^2 / f_{\theta_2}(x') .$$

Nel caso in cui F_θ è continua si può procedere in un modo analogo (per certi versi) a quanto visto per quanto riguarda la stima di minima distanza rispetto a Kullback-Leibler. Si approssima quindi F_θ con una distribuzione discreta per poi

calcolare la distanza tra quest'ultima e la distribuzione empirica (Borovkov, Mathematical Statistics, pag. 67)

Sia $\Delta_1, \Delta_2, ..., \Delta_N$ una partizione del supporto di F_θ e si indichi con n_i il numero di osservazioni campionarie appartenenti all'insieme Δ_i ($P_n(\Delta_i) = n_i / n$). Allora si chiama stimatore del *minimo* χ^2 il valore di θ che minimizza:

$$\sum_{i=1}^{N} \frac{\left((n_i/n) - P_\theta(\Delta_i)\right)^2}{P_\theta(\Delta_i)}.$$

Tornando al caso discreto, esiste come è stato detto precedentemente, una versione modificata dello stimatore del minimo χ^2, data dal valore di θ che minimizza:

$$\rho_2(F_n, F_\theta) = \sum_{x \in S} \frac{\left(f_n(x) - f_\theta(x)\right)^2}{f_n(x)},$$

Nella definizione di questo stimatore si suppone che il valore di f_n non sia mai zero, ossia che S coincida con S_n (Neyman, 1949).

In questo caso, essendo $\phi(x) = (x-1)^2$ una funzione strettamente convessa, si ottiene una misura di divergenza tra F_n e F_θ coerente con il pre-ordinamento di μ-maggiorazione, ossia:

$$\frac{f_{\theta_1}}{f_n} \prec_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n} \Rightarrow \rho_2(F_n, F_{\theta_1}) < \rho_2(F_n, F_{\theta_2}).$$

L'ipotesi $f_n \neq 0$ però è piuttosto restrittiva, tuttavia definire la distanza $\rho_2(F_n, F_\theta)$ sul supporto $S_n \subset S$, così come avviene per la distanza di Kullback.Leibler, non è indicato. In tal caso infatti:

$$\rho_2(F_n, F_\theta) = \sum_{x \in S_n} \frac{\left(f_n(x) - f_\theta(x)\right)^2}{f_n(x)}$$

non risulta coerente rispetto alla μ-maggiorazione debole (da sopra), come accade invece per $\rho_1(F_n, F_\theta)$. Questo è dovuto al fatto che $\phi(x) = (x-1)^2$ è strettamente

convessa ma non è decrescente. A questo proposito si consideri il seguente esempio.

Esempio 5.2.A

Un campione casuale semplice di dimensione $n = 4$ assume i valori (ordinati) $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$, con $x_{(2)} = x_{(3)}$. Si ponga $x_1 = x_{(1)}$, $x_2 = x_{(2)} = x_{(3)}$ e $x_3 = x_{(4)}$, quindi $S_n = \{x_1, x_2, x_3\}$. I valori di $f_n(x_i)$, $i = 1,2,3$, si possono allora rappresentare in forma vettoriale con $f_n(x) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

Si considerino due distribuzioni F_{θ_1} e F_{θ_2}. Si rappresentino coi vettori $f_{\theta_1}(x)$ e $f_{\theta_2}(x)$ i valori rispettivamente di $f_{\theta_1}(x_i)$ e $f_{\theta_2}(x_i)$, $i = 1,2,3$ e si supponga che:

$$f_{\theta_1}(x) = (0.1,\ 0.3,\ 0.3) \quad \text{e} \quad f_{\theta_2}(x) = (0.1,\ 0.3,\ 0.2).$$

Si possono ottenere senza difficoltà i riordini crescenti di $\frac{f_{\theta_1}}{f_n}(x)$ e $\frac{f_{\theta_2}}{f_n}(x)$ rispetto a μ_{F_n}, dati da:

$$\frac{f_{\theta_1}}{f_n}(x)^{(\mu_{F_n})\uparrow} = \begin{cases} 0.4 & se \quad 0 \le x < 0.25 \\ 0.6 & se \quad 0.25 \le x < 0.75 \\ 1.2 & se \quad 0.75 \le x \le 1 \end{cases};$$

$$\frac{f_{\theta_2}}{f_n}(x)^{(\mu_{F_n})\uparrow} = \begin{cases} 0.4 & se \quad 0 \le x < 0.25 \\ 0.6 & se \quad 0.25 \le x < 0.75 \\ 0.8 & se \quad 0.75 \le x \le 1 \end{cases}$$

Si verifica agevolmente che:

$$\int_0^x \frac{f_{\theta_1}}{f_n}^{(\mu_{F_n})\uparrow} dt \ge \int_0^x \frac{f_{\theta_2}}{f_n}^{(\mu_{F_n})\uparrow} dt, \quad \forall x \in [0,1],$$

quindi $\frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n}$.

Come dimostrato nel paragrafo *3.3.3)*, la distanza di Kullback-Leibler rispetta tale pre-ordinamento, infatti:

$$\rho_1(F_n, F_{\theta_1}) = 0.409031 < 0.46019 = \rho_1(F_n, F_{\theta_2}).$$

La distanza di la distanza di χ^2 è anch'essa coerente con l'ordinamento, come dimostrato nel paragrafo *3.3.4*, infatti:

$$\rho_2(F_{\theta_1}, F_n) = 0.\overline{6} < 0.77083\overline{3} = \rho_2(F_{\theta_2}, F_n).$$

Per quanto riguarda la divergenza di Chi^2 a formula "invertita", definita precedentemente (e calcolata su S_n) si ha invece:

$$\rho_2(F_n, F_{\theta_1}) = \rho_2(F_n, F_{\theta_2}) = 0.18.$$

A maggior ragione, si consideri ora la distribuzione F_{θ_3} tale che:

$$f_{\theta_3}(x) = (0.1,\ 0.3,\ 0.4).$$

In questo caso:

$$\frac{f_{\theta_3}}{f_n}(x) \overset{(\mu_{F_n})}{\uparrow} = \begin{cases} 0.4 & se \quad 0 \le x < 0.25 \\ 0.6 & se \quad 0.25 \le x < 0.75 \\ 1.4 & se \quad 0.75 \le x \le 1 \end{cases},$$

quindi:

$$\frac{f_{\theta_3}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_1}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_2}}{f_n}.$$

Tuttavia:

$$\rho_2(F_n, F_{\theta_1}) = 0.21 > \rho_2(F_n, F_{\theta_1}) = \rho_2(F_n, F_{\theta_2}) = 0.18.$$

5.2.3 Hellinger

Ponendo $\phi(x) = (\sqrt{x} - 1)^2$ all'interno di $\Phi(F, G)$ si ottiene la *divergenza di Hellinger* tra le distribuzioni *F* e *G*, data da:

$$\rho_3(F, G) = \int_S \left(\frac{\sqrt{g}}{\sqrt{f}} - 1 \right)^2 dF = \int_S \left(\sqrt{g} - \sqrt{f} \right)^2 d\nu.$$

Dove F , $G \in \mathcal{F}_0$, classe delle distribuzioni assolutamente continue rispetto ad una misura ν.

Anche la distanza di Hellinger appartiene alla classe delle divergenze potenziate, ponendo $\lambda = -1/2$ nella formula $PWD_\lambda(F, G)$.

Si nota subito che, a differenza di ρ_1 e ρ_2, la distanza di Hellinger è un funzionale simmetrico di F, G. ρ_3 infatti è una distanza vera e propria (o metrica) (Borovkov, 1998, Mathematical Statistics, pag. 178).

In maniera del tutto analoga a quanto visto per ρ_2, per F_θ discreta, con supporto dato dai punti x_i, i=1,2,…. , si definisce lo stimatore della minima distanza rispetto a ρ_3 come il valore di θ che minimizza:

$$\rho_3(F_\theta, F_n) = \sum_{x \in S} \left(\sqrt{f_n(x)} - \sqrt{f_\theta(x)} \right)^2 .$$

Tale stimatore è detto *MHDE* (*minimum Hellinger divergence estimator*).

Il funzionale da minimizzare, come il χ^2 di Pearson, è posto nella forma $\Phi(F_\theta, F_n)$ ed è quindi coerente con l'ordinamento di maggiorazione debole (da sopra). Anche la distanza di Hellinger, come il χ^2, può essere scritta come:

$$\rho_3(F_\theta, F_n) = \sum_{x \in S_n} \left(\sqrt{f_n(x)} - \sqrt{f_\theta(x)} \right)^2 + \sum_{x_i \notin S_n} f_\theta(x) .$$

Anche in questo caso, tanto più diminuisce $P_\theta(x \notin S_n)$ (o tanto più cresce $P_\theta(x \in S_n)$) tanto più diminuisce la distanza

Tornando alla formula principale della distanza di Hellinger, si osservi che:

$$\rho_3(F_\theta, F_n) = \sum_{x \in S} \left(\sqrt{f_n(x\ )} - \sqrt{f_\theta(x)} \right)^2 = 2\left(1 - \sum_{x_i \in S} \sqrt{f_n(x)} \sqrt{f_\theta(x)} \right)$$

(Basu, Basu, Chaudhuri, 1997), perciò minimizzare la distanza di Hellinger equivale a massimizzare:

$$s_{n,\theta} = \sum_{x \in S} \sqrt{f_n(x)} \sqrt{f_\theta(x)} .$$

Indicando con ∇ la derivata rispetto a θ, è possibile scrivere l'equazione di stima corrispondente in forma "standardizzata", ossia:

$$\sum_{x \in S} \nabla \ln f_\theta(x) \frac{\sqrt{f_n(x)} \sqrt{f_\theta(x)}}{s_{n,\theta}} .$$

Si parla di forma standardizzata perché, dividendo $\sqrt{f_n(x)} \sqrt{f_\theta(x)}$ per $s_{n,\theta}$ si ottiene una distribuzione di probabilità, che da somma 1 sul supporto S.

Confrontando la precedente equazione di stima con l'equazione di verosimiglianza:

$$\sum_{x \in S} \left(\nabla \ln f_\theta(x) \right) f_n(x) = 0 ,$$

si può notare come la stima della minima distanza di Hellinger consista in una media della "score function" $\nabla \ln f_\theta(x)$ rispetto a $s_{n,\theta}^{-1} \sqrt{f_n} \sqrt{f_\theta}$. Ciò è in contrasto con la massima verosimiglianza, in cui la media è rispetto a f_n.

Riflettendo sulla differenza tra queste due equazioni di stima si possono intuire facilmente delle importanti proprietà della stimatore basato sulla distanza di Hellinger. Da un lato si osservi che, se il modello è valido, le due equazioni tendono a coincidere, al crescere della numerosità campionaria, infatti per $n \to \infty$, $s_{n,\theta}^{-1} \sqrt{f_n} \sqrt{f_\theta} \to f_\theta$. Dall'altro lato, l'effetto di una grossa deviazione dei dati delle assunzioni del modello è decisamente sotto-pesato nell'equazione relativa al MHDE piuttosto che nell'equazione di verosimiglianza. Infatti, se il dato x' è ritenuto poco probabile dal modello, il suo contributo all'equazione di stima MHDE sarà:

$$s_{n,\theta}^{-1} \sqrt{f_n(x')} \sqrt{f_\theta}(x') < f_n(x') .$$

Si intuisce allora che lo stimatore MHDE gode di proprietà efficienza asintotica e di robustezza, come è stato dimostrato (Beran, 1977; Simpson, 1987).

Nel caso continuo, invece, definita la consueta partizione $\Delta_1, \Delta_2, ..., \Delta_N$, si può definire lo stimatore della minima distanza rispetto a ρ_3 come il valore di θ che minimizza:

$$\sum_{i=1}^{N} \left(\sqrt{(n_i / n)} - \sqrt{P_\theta(\Delta_i)} \right)^2 ,$$

dove n_i è il numero di osservazioni campionarie appartenenti all'insieme Δ_i ($P_n(\Delta_i) = n_i / n$).

5.3 Misure di divergenza e funzioni RAF

5.3.1 Funzioni di adeguamento dei residui

Sia F_θ una distribuzione discreta con supporto dato da S. Alla luce di quanto mostrato dal teorema *3.3.4*, funzionali del tipo:

$$\Phi(F_\theta, F_n) = \sum_{x \in S} \phi\left(\frac{f_n(x)}{f_\theta(x)}\right) f_\theta(x),$$

rappresentano misure di divergenza con proprietà ottimali, se ϕ è strettamente convessa.

Si definisca la *funzione dei residui di Pearson* (Lindsay, 1994) con:

$$\delta(x) = \frac{f_n(x) - f_\theta(x)}{f_\theta(x)},$$

si osservi che $\delta \in [-1, \infty]$.

Allora, una misura di divergenza può avere la seguente forma:

$$\rho(F_\theta, F_n) = \sum_{x \in S} G\big(\delta(x)\big) f_\theta(x),$$

dove G è una funzione definita sull'insieme [-1, ∞], strettamente convessa, derivabile tre volte e tale che $G(0) = 0$. La divergenza $\rho(F_\theta, F_n)$ è anche detta misura di *disparità* (Lindsay, 1994).

Si supponga di voler minimizzare la divergenza $\rho(F_\theta, F_n)$ rispetto a θ. Se la densità f_θ è derivabile (rispetto a θ) il problema di riduce alla soluzione di un'equazione di stima di questo tipo:

$$\nabla\rho(F_\theta, F_n) = \sum_{x \in S}\left[G'\big(\delta(x)\big)\frac{f_n(x)}{f_\theta(x)} - G\big(\delta(x)\big)\right]\nabla f_\theta(x) = 0,$$

dove il simbolo ∇ indica la derivata rispetto a θ.

Ponendo

$$A(\delta) = (1+\delta)G'(\delta) - G(\delta),$$

la precedente equazione di stima può essere scritta con la seguante forma:

$$\sum_{x \in S} A(\delta)\nabla f_\theta(x) = 0.$$

$A'(\delta) = (1+\delta)G''(\delta)$, dove G è derivabile tre volte e strettamente convessa. Allora $A(\delta)$ è strettamente crescente in [-1, ∞].

Si osservi che se $\sum_{x\in S} f_\theta(x) = 1$ allora $\sum_{x\in S} \nabla f_\theta(x) = 0$. Per questo motivo si verifica semplicemente che:

$$\sum_{x\in S} A(\delta)\nabla f_\theta(x) = 0$$

implica

$$\sum_{x\in S} (A(\delta) - A(0))\nabla f_\theta(x) = 0 .$$

Inoltre $A'(0) = G''(0) > 0$.

Allora è possibile ridefinire $A(\delta)$ con $A(\delta) := (A(\delta) - A(0))/A'(0)$ senza cambiare le soluzioni dell'equazione di stima. In tal modo $A(\delta)$ è una funzione strettamente crescente in [-1, ∞], tale che $A(0) = 0$ e $A'(0) = 1$.

$A(\delta)$ è detta *funzione di adeguamento dei residui* (Lindsay, 1994), o *residual adjustment function* (RAF).

Nel caso $\rho(F_\theta, F_n)$ appartenga alla classe delle divergenze potenziate, allora la funzione RAF ha la forma:

$$A_\lambda(\delta) = \frac{(1+\delta)^\lambda - 1}{\lambda + 1} .$$

L'equazione di stima può essere riscritta come un'equazione di verosimiglianza pesata, infatti:

$$\sum_{x\in S} A(\delta(x))\nabla f_\theta(x) = \sum_{x\in S} [A(\delta(x)) - A(-1)]\nabla f_\theta(x) =$$

$$\sum_{x\in S} w^*(x)(1+\delta(x))\nabla f_\theta(x) = \sum_{x\in S} w^*(x) f_n \nabla \ln f_\theta(x) ,$$

dove i pesi sono dati da $w^*(x) = [A(\delta(x)) - A(-1)]/(1+\delta(x))$.

Si osservi che

$$AE(w^* \nabla \ln f_\theta) = 0 ,$$

infatti, poiché $\lim_{n\to\infty} \delta(x) = 0$, si ha:

$$AE(w^* \nabla \ln f_\theta) = \lim_{n\to\infty} \sum_{x\in S} w^*(x) \nabla f_\theta(x) = -A(-1) \sum_{x\in S} \nabla f_\theta(x) = 0\,,$$

tali stimatori-*M* appartengono quindi alla classe GEE. Per comodità, questi stimatori sono detti *MDE*, ossia "*minimum disparity estimator*" (Lindsay, 1994).

5.3.2 Proprietà di efficienza e robustezza

I concetti di efficienza e di robustezza risultano essere piuttosto speculari, in inferenza statistica. La teoria dell'efficienza si basa sulla supposizione che la distribuzione "vera" G a cui appartiene il campione sia di tipo parametrico, cioè $G = F_\theta \in \mathcal{F}_\theta$. Si cerca quindi lo stimatore efficiente, cioè quello che individua il valore "vero" del parametro con precisione maggiore. Tuttavia nella pratica può accadere che la vera distribuzione non sia esattamente del tipo indicato da F_θ (per ciò si parla di *modello* quando ci si riferisce a F_θ), quindi $G \notin \mathcal{F}_\theta$. In tal caso una certa deviazione dal modello, anche lieve, può avere effetti negativi sulla precisione dello stimatore adottato. Per questo motivo si cerca uno stimatore con proprietà di robustezza, cioè che presenta stime ragionevoli anche quando l'ipotesi $G \in \mathcal{F}_\theta$ non è esatta.

E' noto che il metodo della massima verosimiglianza fornisce stimatori efficienti, ma, solitamente, non robusti (con funzione di influenza potenzialmente illimitata). Tuttavia, come mostrato da Beran (1977), lo stimatore della minima distanza di Hellinger ha buone proprietà di robustezza, oltre ad essere asintoticamente efficiente, se le assunzioni del modello sono esatte.

Lindsay (1994) generalizza questo risultato nella classe degli stimatori MDE, studiando la forma della funzione $A(\delta)$.

In particolare, il comportamento di $A(\delta)$ vicino a $\delta = 0$ determina da un lato l'efficienza, dall'altro la robustezza dello stimatore.

Come visto nel paragrafo *2.1.4*, solitamente si ritiene che la proprietà di efficienza di uno stimatore-*M* dipenda dalla funzione di influenza, per via dell'approssimazione:

$$\sqrt{n}\left[T(F_n) - T(F)\right] \cong \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_F(X_i) .$$

Tuttavia, nella classe degli stimatori MDE, la funzione di influenza si dimostra poco efficace come misura di robustezza.

Teorema 5.3.A

Si indichi con T lo stimatore associato all'equazione $\sum_{x \in S} A(\delta) \nabla f_\theta(x) = 0$. La curva di influenza del funzionale statistico T, nel "punto" $F=F_\theta$, è data da:

$$\frac{\nabla \ln f_\theta}{I(\theta)} .$$

(Lindsay, 1994)

Quindi ogni stimatore MDE ha la stessa funzione di influenza di uno stimatore di massima verosimiglianza. Questo risultato suggerisce che gli stimatori MDE sono asintoticamente efficienti, qualora la vera distribuzione appartenga alla famiglia parametrica $\mathcal{F}_\theta$.

Sempre nel paragrafo *2.1.4*, si è visto che spesso la funzione di influenza approssima la distorsione asintotica dello stimatore di θ, causata da una certa contaminazione, cioè $\Delta T(\varepsilon) \approx \varepsilon \cdot I_F(y)$. Per questo motivo, solitamente, la funzione di influenza è ritenuta un buon indicatore di robustezza. Tuttavia, in questo caso, significherebbe che ogni stimatore MDE sia da ritenersi non robusto, avendo funzione di influenza uguale alla stimatore di massima verosimiglianza. Come dimostrato da Lindsay (1994) quest'ultima approssimazione, basata sullo sviluppo di Taylor del primo ordine, può risultare inaccurata per molti stimatori di tipo RAF. In tali casi, è necessaria un'approssimazione del secondo ordine.

In particolare, è il comportamento di $A(\delta)$ vicino a $\delta = 0$ a determinare il comportamento della funzione $\Delta T(\varepsilon)$ vicino ad $\varepsilon = 0$. Il fatto che gli stimatori RAF abbiano la stessa funzione di influenza è strettamente legato al fatto che le rispettive funzioni $A(\delta)$ hanno tutte la stessa approssimazione al primo ordine, cioè $A(\delta) \approx \delta$, il che li fa sembrare tutti simili allo stimatore di massima

verosimiglianza. E' stato dimostrato (Lindsay, 1994) che l'approssimazione al secondo ordine, data da:

$$A(\delta) \approx \delta + A''(0)\delta^2 / 2 ,$$

determina la proprietà di efficienza al secondo ordine e di robustezza dello stimatore corrispondente. Si noti che, nella classe delle divergenze potenziate, $A''(0) = \lambda$, quindi, ad esempio, per la distanza di Kullback-Leibler (stima di massima verosimiglianza) $\lambda = 0$, per chi-quadrato $\lambda = 1$, per Hellinger $\lambda = -1/2$.

Si indichi con $\hat{\theta}$ lo stimatore di massima verosimiglianza e con $\widetilde{\theta}$ lo stimatore MDE.

Teorema 5.3.B

Si supponga che lo spazio campionario si finito. Allora l'efficienza del secondo ordine di uno stimatore MDE, con funzione RAF $A(\delta)$ è:

$$E_2(\widetilde{\theta}) = E_2(\hat{\theta}) + [A''(0)]^2 D ,$$

dove D è un numero non negativo che dipende dalla distribuzione ma non da $A(\delta)$. (Lindsay, 1994)

Il valore di $A''(0)$, che esprime la curvatura della funzione RAF, è rilevante anche per quanto riguarda la robustezza. In particolare è stato dimostrato (Lindsay, 1994) che per valori di $A''(0)$ diversi da zero (Kullback-Leibler), la funzione di influenza non fornisce una buona approssimazione della $\Delta T(\varepsilon)$ vicino ad $\varepsilon = 0$. Inoltre, ci si aspetta di avere un comportamento più robusto, per valori di $A''(0)$ negativi (distanza di Hellinger, chi-quadrato modificato….).

5.4 Minima divergenza: estensione al caso continuo

5.4.1 Spacing

Si è visto che il metodo della minima distanza rispetto alle varie misure di divergenza non è propriamente applicabile nel caso continuo, dato che non si può confrontare una distribuzione discreta con una continua. In particolare si consideri la divergenza di Kullback-Leibler. Nel caso continuo essa è riconducibile al metodo della massima verosimiglianza grazie ad approssimazioni che, in certi

casi, possono non essere valide. Questo suggerisce l'utilizzo di metodi in grado di funzionare in condizioni più generali possibili, svincolati dalle approssimazioni in questione. Un metodo valido, ad esempio, è quello visto precedentemente basato sulla discretizzazione di F_θ. Oltre a questo approccio, in alternativa al metodo della massima verosimiglianza, esiste un metodo basato sul massimizzare un prodotto di particolari probabilità (dette "spacing") anziché un prodotto di densità (Cheng, Amin 1983, Ranneby 1984). Un difetto, da un punto di vista interpretativo, del metodo della massima verosimiglianza nel caso continuo, infatti, può essere quello di consistere in un confronto tra le densità di F_θ e quelle di F_n, le quali però rappresentano delle probabilità vere e proprie.

Il metodo di stima proposto da Cheng e Amin è detto stima del *massimo prodotto di spacing*, ed è applicabile in condizioni più generali rispetto alla massima verosimiglianza.

Si consideri la distribuzione continua F_θ con densità f_θ, la quale assume valori strettamente positivi nell'intervallo (α_1, α_2) e valore zero all'esterno. L'incognito parametro θ (può essere anche un vettore) appartiene allo spazio parametrico Θ. Inoltre, anche valori di α_1 e α_2 possono essere incogniti ed essere quindi componenti dell'eventuale vettore θ.

Sia θ_0 il vero valore del parametro e $x_1, x_2, ..., x_n$ un campione ordinato di dimensione n estratto dalla distribuzione "vera" F_{θ_0}. Si ponga $x_0 = \alpha_1$, $x_{n+1} = \alpha_2$ e si definiscano i valori:

$$D_i(\theta) = \int_{x_{i-1}}^{x_i} f_\theta(x)dx = F_\theta(x_i) - F_\theta(x_{i-1}), \quad i = 1,2,....,n+1,$$

detti *spacing* del campione. Si osservi che $\sum_{i=1}^{n+1} D_i(\theta) = 1$.

Il metodo del massimo prodotto di spacing (*maximum spacing product, MPS*) consiste nell'individuazione del valore θ che massimizza la media geometrica degli spacing:

$$G(\theta) = \left\{ \prod_{i=1}^{n+1} D_i(\theta) \right\}^{1/n+1},$$

o, equivalentemente, il suo logaritmo $H(\theta) = \ln G(\theta)$.

$$H(\theta) = \frac{1}{n+1} \sum_{i=1}^{n+1} \ln D_i(\theta).$$

Si osservi che il massimo di *G* (o di *H*) è limitato da sopra, dato che $\sum D_i = 1$.

Dati i vettori $D = (D_1, D_2, ..., D_{n+1})$ e $D' = (D'_1, D'_2, ..., D'_{n+1})$, ricordando che $\sum D_i = \sum D'_i = 1$, si ha che, se $D \prec D'$ allora:

$$\sum_{i=1}^{n+1} \phi(D_i) > \sum_{i=1}^{n+1} \phi(D'_i),$$

con ϕ strettamente concava.

Si osservi quindi che la funzione

$$H = \ln \left\{ \prod_{i=1}^{n+1} D_i \right\}^{1/n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} \ln D_i$$

è una funzione Schur-concava rispetto al vettore $(D_1, D_2, ..., D_{n+1})$.

Questo significa che il massimo di $G(\theta)$ o di $H(\theta)$ si ottiene quando i valori $D_i(\theta)$ sono uguali tra loro. Quindi si ha che $D_i(\theta_0) = D_j(\theta_0)$, per $i, j = 1,2,....,n+1$. In altri termini θ è tanto più vicino a θ_0 tanto più rende "omogenei" i valori degli spacing.

Si osservi che, siccome $\sum D_i(\theta) = 1$, $\forall \theta$, ogni vettore di $n+1$ spacing $(D_1(\theta), D_2(\theta), ..., D_{n+1}(\theta))$ può essere visto come una distribuzione di probabilità discreta, indicata ad esempio con $D(\theta)$. Ranneby (1984) descrive il metodo del massimo prodotto di spacing come il metodo che minimizza la distanza di Kullback-Leibler tra gli spacing provenienti da F_θ e quelli provenienti da F_{θ_0} (la vera distribuzione), ossia tra le due distribuzioni discrete $D(\theta)$ e $D(\theta_0)$.

Infatti, considerando che $D_i(\theta_0) = D_j(\theta_0)$, per $i, j = 1,2,....,n+1$, si ha che:

$$\sum_{i=1}^{n+1} D_i(\theta) = (n+1)D_i(\theta) = 1 \;\Rightarrow\; D_i(\theta_0) = \frac{1}{n+1}, \quad \text{per } i = 1,2,....,n+1.$$

Ranneby propone quindi di massimizzare, anziché $H(\theta)$, la seguente quantità:

$$-\rho_1(D(\theta_0), D(\theta)) = \frac{1}{n+1}\sum_{i=1}^{n+1} \ln\big((n+1)D_i(\theta)\big).$$

Il metodo è equivalente a quello proposto da Cheng e Amin, dato che il risultato è uguale. La diversa interpretazione però suggerisce la possibilità di individuare stimatori che minimizzino altri tipi di divergenze tra $D(\theta)$ e $D(\theta_0)$, come ad esempio la distanza di χ^2 o quella di Hellinger.

Si parla in questo caso di metodo del massimo prodotto di spacing generalizzato ("*Generalized maximu spacing product", GMSP*) (Ranneby, Ekstrom, 1997). Ossia si cerca il valore di θ che minimizzi una qualsiasi funzione Schur-convessa di $D(\theta)$, data da:

$$\Phi\big(D(\theta_0), D(\theta)\big) = \int \phi\left(\frac{dD(\theta)/d\nu}{dD(\theta_0)/d\nu}\right) dD(\theta_0) = \frac{1}{n+1}\sum_{i=1}^{n+1} \phi\big((n+1)D_i(\theta)\big),$$

dove ν è la misura del conteggio e ϕ una funzione strettamente convessa.

Ricordando le proprietà delle funzioni Schur-convesse si ha che:

$$D(\theta_1) \prec D(\theta_2) \Rightarrow \Phi\big(D(\theta_0), D(\theta_1)\big) < \Phi\big(D(\theta_0), D(\theta_2)\big).$$

Quindi, anche in questo caso più generale, lo stimatore GMSP è il valore di θ che rende più omogenei e uniformi possibili i valori degli spacing, rispetto alle differenti distanze o divergenze, espresse da funzioni di tipo $\Phi\big(D(\theta_0), D(\theta)\big)$. Tali divergenze possono sempre calcolarsi senza particolari problemi, dato che si ha a che fare con distribuzioni discrete.

Ad esempio ponendo $\phi(x) = (x-1)^2$ si ottiene lo stimatore GMSP rispetto alla distanza χ^2, cioè lo stimatore che minimizza la funzione:

$$\rho_2(D(\theta_0), D(\theta)) = \int \left(\frac{dD(\theta)/d\nu}{dD(\theta_0)/d\nu} - 1\right)^2 dD(\theta_0) = (n+1)\sum_{i=1}^{n+1}\left(D_i(\theta) - 1/(n+1)\right)^2.$$

Ponendo invece $\phi(x) = (\sqrt{x} - 1)^2$ si ottiene lo stimatore GMSP rispetto alla distanza di Hellinger, cioè lo stimatore che minimizza:

$$\rho_3(D(\theta_0), D(\theta)) = \int \left(\sqrt{\frac{dD(\theta)/d\nu}{dD(\theta_0)/d\nu}} - 1 \right)^2 dD(\theta_0) =$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} \left(\sqrt{D_i(\theta)} - \sqrt{1/(n+1)} \right)^2 .$$

Può essere utile confrontare, da un punto di vista intuitivo, il metodo della massima verosimiglianza e del massimo prodotto di spacing.
Si consideri la log-verosimiglianza:

$$\int \ln f_\theta dF_n = \frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(x_i) .$$

I termini in *H* corrispondenti a $\ln f_\theta(x_i)$ sono dati dai $\ln D_i$, i quali si possono anche esprimere come:

$$\ln D_i = \ln \int_{x_{i-1}}^{x_i} f_\theta dx = \ln\left[f_\theta(x_i)(x_i - x_{i-1}) \right] + R_\theta(x_i, x_{i-1}) =$$

$$= \ln f_\theta(x_i) + \ln(x_i - x_{i-1}) + R_\theta(x_i, x_{i-1}) .$$

In situazioni standard, in cui cioè i punti α_1, α_2 delimitanti il supporto di F_θ sono noti, il "resto" R_θ è infinitesimo di ordine $O\left(\left|x_i - x_{i-1}\right|\right)$, seppur dipenda da θ. Ma $x_i - x_{i-1}$ tende a zero in probabilità al crescere di *n*, tranne che per un trascurabile numero di termini. Quindi, dato che $\ln(x_i - x_{i-1})$ non dipende da θ, il comportamento di $\ln D_i$, rispetto a θ, è sostanzialmente lo stesso di $\ln f_\theta(x_i)$. In conclusione, nelle situazioni standard, la massima verosimiglianza e il massimo prodotto di spacing sono asintoticamente equivalenti (Cheng, Amin 1983). Si osservi che anche l'approssimazione introdotta ora è sostanzialmente analoga alle approssimazioni utilizzate per la definizione di funzione di verosimiglianza e per l'interpretazione del metodo della minima distanza di Kulllback-Leibler, nel caso continuo.

5.4.2 Stimatori di densità

Si è visto che, qualora la distribuzione oggetto di studio sia continua, applicare il metodo della minima distanza risulta problematico. Infatti non è possibile calcolare la distanza tra le densità (rispetto alla misura di Lebesgue) e le frequenze delle osservazioni.

L'approccio studiato nel precedente paragrafo, basato sugli spacing, consiste nel "discretizzare" la distribuzione F_θ per rendere possibile il confronto con F_n.

L'approccio alternativo è concettualmente opposto: consiste nel creare uno stimatore non-parametrico della funzione di densità delle osservazioni. Tale stimatore è detto *densità empirica* (Shao, 2003).

Una classe molto utilizzata di stimatori di densità è quella degli *stimatori Kernel* dati da:

$$f^*(x) = \int_{S_n} k(x;t,h)dF_n(t) = \sum_{t \in S_n} k(x;t,h) f_n(t),$$

dove k, detta *funzione Kernel*, rappresenta una famiglia di densità, di parametri (*t, h*), come, per esempio, le densità di una distribuzione Normale, con valore atteso *t* e varianza *h*. Il valore di *h*, così come il tipo di distribuzione *k*, possono essere scelti a seconda delle esigenze.

$f^*(x)$ quindi è una media di funzioni di densità al variare di *t*, pesate rispetto alle frequenze delle osservazioni. Il parametro *h* controlla la forma della funzione di densità $f^*(x)$. Al diminuire di *h* la curva avrà una forma sempre più frastagliata, al crescere di *h* si ottiene una curva sempre più "liscia". Questa tecnica infatti prende il nome di "smoothing": sostituendo le frequenze dei singoli punti con curve più o meno "ampie" (a seconda di *h*), si crea una funzione di densità empirica (continua), più o meno liscia.

A questo punto, è possibile costruire una distanza (divergenza) tra la densità empirica $f^*(x)$ e la densità del modello $f_\theta(x)$, data dalla formula generale:

$$\int_S \phi\left(\frac{f_\theta(x)}{f^*(x)}\right) f^*(x) d\lambda,$$

dove λ è la misura di Lebesgue e ϕ è una funzione strettamente convessa.

Ad esempio ponendo $\phi(x) = (\sqrt{x} - 1)^2$ si ottiene la distanza di Hellinger:

$$\int_S \left[\sqrt{f^*(x)} - \sqrt{f_\theta(x)}\right]^2 f^*(x) d\lambda ,$$

studiata da Beran (1977).

Un approccio più moderno (Basu, Lindsay, 1994) è quello di applicare a $f_\theta(x)$ lo stesso "smoothing" utilizzato per calcolare $f^*(x)$, ottenendo:

$$f_\theta{}^*(x) = \int_S k(x;t,h) dF_\theta(t) ,$$

per poi calcolare la stima di minima distanza rispetto a funzionali della seguente forma:

$$\int_S \phi\left(\frac{f_\theta{}^*(x)}{f^*(x)}\right) f^*(x) d\lambda ,$$

dove λ è la misura di Lebesgue e ϕ è una funzione strettamente convessa, oppure nella forma delle funzioni RAF:

$$\int_S G\big(\delta^*(x)\big) f^*(x) d\lambda ,$$

dove $\delta^*(x) = \dfrac{f_n{}^*(x) - f_\theta{}^*(x)}{f_\theta{}^*(x)}$ e G è una funzione definita su [-1, ∞], strettamente convessa, derivabile tre volte e tale che $G(0) = 0$.

E' stato mostrato (Basu, Lindsay, 1994) che applicare lo stesso "smoothing" a entrambe F_θ e F_n rende il metodo della minima distanza consistente e asintoticamente normale, indipendentemente dal parametro *h*. La consistenza di $f^*(x)$ non è necessaria. In maniera simile a quanto visto nel paragrafo *5.3.2*, scegliendo in modo appropriato la funzione kernel, si ottengono stimatori efficienti al primo ordine. Le differenze tra efficienza e robustezza degli stimatori considerati dipendono ancora dalle funzioni RAF. La derivata seconda della funzione RAF nel punto 0 ($A''(0)$) determina infatti le proprietà di efficienza e di robustezza dello stimatore corrispondente.

Capitolo 6

Misure di divergenza modificate

6.1 Divergenza di Hellinger "penalizzata"

Si osservi che le distanze di Hellinger e di Chi quadrato possono essere entrambe scritte nella seguente forma:

$$\sum_{x\in S_n}\phi\left(\frac{f_n(x)}{f_\theta(x)}\right)+\sum_{x\notin S_n}f_\theta(x),$$

in particolare:

$$\rho_3(F_\theta,F_n)=\sum_{x\in S_n}\left(\sqrt{f_n(x)}-\sqrt{f_\theta(x)}\right)^2+\sum_{x\notin S_n}f_\theta(x),$$

$$\rho_2(F_\theta,F_n)=\sum_{x\in S_n}\frac{\left(f_n(x)-f_\theta(x)\right)^2}{f_\theta(x)}+\sum_{x\notin S_n}f_\theta(x).$$

Quindi, tanto più diminuisce $P_\theta(x\notin S_n)$ (o tanto più cresce $P_\theta(x\in S_n)$) tanto più diminuisce la distanza. In altri termini, la probabilità di quei punti non presenti nel campione risulta determinante. Tuttavia, in certe situazioni, il peso attribuito ai punti $x\notin S_n$ può essere sproporzionato rispetto all'addendo $\sum_{x\in S_n}\phi\left(\frac{f_n}{f_\theta}\right)$. Questo può causare stime poco convincenti, rispetto ad esempio alla massima verosimiglianza, soprattutto quando il campione è di dimensioni modeste e, quindi, $P_\theta(x\notin S_n)$ è maggiore.

Questo possibile inconveniente è stato affrontato da Basu e Harris (1997) definendo la classe delle *distanze di Hellniger penalizzate*, date da:

$$\sum_{x\in S_n}\left(\sqrt{f_n(x)}-\sqrt{f_\theta(x)}\right)^2+h\sum_{x\notin S_n}f_\theta(x),$$

dove l'adendo $P_\theta(x\notin S_n)$ è definito *penalità* e il peso della penalità è regolato dal coefficiente $h\in[0,1]$.

Lo stimatore corrispondente si dirà *MHPD* "*minimum penalized Hellinger distance estimator*" (Basu, Harris, Basu, 1997).

Si osservi che si è posto $h \in [0,1]$. Secondo questa definizione, quindi, il peso di $P_\theta(x \notin S_n)$, per piccoli campioni, può essere troppo grande, e si preferisce attenuarne l'effetto servendosi del coefficiente *h*. Però non si può sapere con esattezza a quale valore di *h* corrisponde la stima *MHPD* "migliore". Basu e Harris, sulla base di una simulazione, ipotizzano che i risultati più convincenti si ottengano per h=0.5, cosa che in generale non corrisponde ai risultati forniti dalle simulazioni che verranno mostrate in seguito. Pare infatti che il valore ottimale di *h* vari a seconda della forma della distribuzione (quindi del vero valore di θ) e anche a seconda della numerosità del campione.

Nelle tabelle dell'appendice *B.1* sono riportati medie e M.S.E. (men square error) degli stimatori MHPD per distribuzioni Binomiali con diversi valori di *p* e diverse ampiezze campionarie.

Osservando le tabelle si nota che diminuendo il valore di *h* si ottengono stime migliori. Bisogna però fare attenzione, perché si nota che valori troppo piccoli di *h* possono portare ad un peggioramento nelle stime. In particolare, esiste sempre un valore "ottimale" di *h* al di la del quale le stime cominciano a peggiorare. Questo valore dipende principalmente dalla distribuzione e dall'ampiezza campionaria. Essendo la distribuzione incognita, non si può mai sapere quale sia il miglior valore di *h*, quindi, in generale, porre *h* troppo piccolo può essere decisamente rischioso.

Nonostante i miglioramenti che esso porta, il metodo delle stime MHPD non risulta, in generale, coerente con l'ordinamento di maggiorazione. Questo può essere dimostrato con un controesempio.

Un campione casuale semplice di dimensione $n = 4$ assume i valori (ordinati) $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$, con $x_{(2)} = x_{(3)}$. Si ponga $x_1 = x_{(1)}$, $x_2 = x_{(2)} = x_{(3)}$ e $x_3 = x_{(4)}$, quindi $S_n = \{x_1, x_2, x_3\}$. I valori di $f_n(x_i)$, $i = 1,2,3$, si possono allora rappresentare in forma vettoriale con $f_n(x) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

Si considerino due distribuzioni F_{θ_1} e F_{θ_2}. Si rappresentino coi vettori $f_{\theta_1}(x)$ e $f_{\theta_2}(x)$ i valori rispettivamente di $f_{\theta_1}(x_i)$ e $f_{\theta_2}(x_i)$, $i = 1,2,3$ e si supponga che:

$$f_{\theta_1}(x) = (0.1,\ 0.3,\ 0.3) \quad \text{e} \quad f_{\theta_2}(x) = (0.1,\ 0.3,\ 0.4)\ .$$

Si possono ottenere senza difficoltà i riordini crescenti di $\frac{f_{\theta_1}}{f_n}(x)$ e $\frac{f_{\theta_2}}{f_n}(x)$ rispetto a μ_{F_n}, dati da:

$$\frac{f_{\theta_1}}{f_n}(x)^{(\mu_{F_n})\uparrow} = \begin{cases} 0.4 & se \quad 0 \le x < 0.25 \\ 0.6 & se \quad 0.25 \le x < 0.75 \\ 1.2 & se \quad 0.75 \le x \le 1 \end{cases};$$

$$\frac{f_{\theta_2}}{f_n}(x)^{(\mu_{F_n})\uparrow} = \begin{cases} 0.4 & se \quad 0 \le x < 0.25 \\ 0.6 & se \quad 0.25 \le x < 0.75 \\ 1.6 & se \quad 0.75 \le x \le 1 \end{cases}$$

Si verifica agevolmente che:

$$\int_0^x \frac{f_{\theta_2}}{f_n}^{(\mu_{F_n})\uparrow} dt \ge \int_0^x \frac{f_{\theta_1}}{f_n}^{(\mu_{F_n})\uparrow} dt, \quad \forall x \in [0,1],$$

quindi $\frac{f_{\theta_2}}{f_n} \prec_{(\mu_{F_n})}^{w} \frac{f_{\theta_1}}{f_n}$.

Allora si sa che $\rho_3(F_{\theta_2}, F_n) < \rho_3(F_{\theta_1}, F_n)$. Tuttavia non è detto questo valga anche per le distanze di Hellinger penalizzate. Infatti è possibile individuare i valori di *h* per i quali vale:

$$\sum_{x \in S_n} \left(\sqrt{f_n(x)} - \sqrt{f_{\theta_2}(x)}\right)^2 + h \sum_{x \notin S_n} f_{\theta_2}(x) > \sum_{x \in S_n} \left(\sqrt{f_n(x)} - \sqrt{f_{\theta_1}(x)}\right)^2 + h \sum_{x \notin S_n} f_{\theta_1}(x) \text{ In}$$

particolare, si ha che:

$$0.07671 + h \cdot 0.2 > 0.06144 + h \cdot 0.3,$$

per $h < 0.1527$.

Quindi, se il peso che si attribuisce a $P_\theta(x \notin S_n)$ è troppo piccolo, lo stimatore può fornire risultati poco ragionevoli, cosa che a quanto pare, è confermata anche dalle simulazioni.

In conclusione il metodo degli stimatori MHPD, pur portando dei miglioramenti piuttosto evidenti (da quanto si vede nell'appendice *B.1*) è da utilizzare con cautela, dato che le distanze utilizzate non sono in generale coerenti rispetto al principio fondamentale della maggiorazione (debole).

6.2 Divergenza di Chi quadrato "depenalizzata"

Si è visto che anche la distanza di Chi quadrato può essere scritta come:

$$\rho_2(F_\theta, F_n) = \sum_{x \in S_n} \frac{\left(f_n(x) - f_\theta(x)\right)^2}{f_\theta(x)} + \sum_{x \notin S_n} f_\theta(x).$$

Anche in questo caso il peso attribuito ai punti $x \notin S_n$ può essere sproporzionato rispetto all'addendo $\sum_{x \in S_n} \left(\frac{f_n}{f_\theta} - 1 \right)^2$, soprattutto quando la numerosità campionaria non è elevata. In particolare, risulta evidente, osservando casi pratici, che, a differenza di ciò che avviene per la distanza di Hellinger, l'addendo $\sum_{x \notin S_n} f_\theta(x)$ ha un peso piccolo rispetto all'addendo $\sum_{x \in S_n} \left(\frac{f_n}{f_\theta} - 1 \right)^2$.

Di conseguenza, è possibile attenuare l'effetto del peso sproporzionato, all'interno della distanza, di $\sum_{x \in S_n} \left(\frac{f_n}{f_\theta} - 1 \right)^2$, servendosi di un coefficiente $h \in [0,1]$.

Si può quindi definire la seguente misura di divergenza:

$$h \sum_{x \in S_n} \frac{\left(f_n(x) - f_\theta(x)\right)^2}{f_\theta(x)} + \sum_{x \notin S_n} f_\theta(x),$$

dato che il secondo addendo dell'espressione è detto "penalità", questa particolare divergenza può essere definita *divergenza di Chi^2 "depenalizzata"*. Si osservi la differenza con la divergenza di Hellinger "penalizzata" (paragrafo *6.1*).

Lo stimatore corrispondente si dirà stimatore *MCDD* ("minimum Chi-square depenalized divergence").

Si osservi che, a differenza di degli stimatori MHPD, gli stimatori MCDD sono coerenti rispetto alla maggiorazione debole.

Siano F_{θ_1} e F_{θ_2} due distribuzioni tali che $\frac{f_{\theta_1}}{f_n} \prec_{(\mu_{F_n})}^{w} \frac{f_{\theta_2}}{f_n}$.

Allora si sa che $\rho_2(F_{\theta_1}, F_n) < \rho_2(F_{\theta_2}, F_n)$ e che, senz'altro:

$$\sum_{x_i \notin S_n} f_{\theta_1}(x_i) \leq \sum_{x_i \notin S_n} f_{\theta_2}(x_i).$$

Si supponga, per assurdo, che esista il valore h' tale che

$$h'\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_1}(x)\right)^2}{f_{\theta_1}(x)}+\sum_{x\notin S_n}f_{\theta_1}(x)>h'\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_2}(x)\right)^2}{f_{\theta_2}(x)}+\sum_{x\notin S_n}f_{\theta_2}(x),$$

ciò è possibile solo se:

$$\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_1}(x)\right)^2}{f_{\theta_1}(x)}-\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_2}(x)\right)^2}{f_{\theta_2}(x)}>0.$$

Dato che, per ipotesi:

$$\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_1}(x)\right)^2}{f_{\theta_1}(x)}-\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_2}(x)\right)^2}{f_{\theta_2}(x)}<\sum_{x\notin S_n}f_{\theta_2}(x)-\sum_{x\notin S_n}f_{\theta_1}(x),$$

allora si ottiene che h' è sicuramente maggiore di 1, il che è assurdo, dato che, per definizione, h dev'essere compreso tra i valori 0 e 1.

Quindi se F_{θ_1} e F_{θ_2} sono due distribuzioni tali che $\frac{f_{\theta_1}}{f_n}\prec^{w}_{(\mu_{F_n})}\frac{f_{\theta_2}}{f_n}$, si ha:

$$h\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_1}(x)\right)^2}{f_{\theta_1}(x)}+\sum_{x\notin S_n}f_{\theta_1}(x)<h\sum_{x\in S_n}\frac{\left(f_n(x)-f_{\theta_2}(x)\right)^2}{f_{\theta_2}(x)}+\sum_{x\notin S_n}f_{\theta_2}(x),$$

per $h\in[0,1]$.

Nell'appendice *B.2* sono riportati medie e M.S.E. (men square error) degli stimatori studiati in questo paragrafo, per distribuzioni Binomiali con diversi valori di p e diverse ampiezze campionarie. Da quanto si osserva, pare che, solitamente, si ottiene il massimo miglioramento delle stima per valori di h intorno a 0.3-0.4. Anche in questo caso comunque la scelta di h può dipendere dal valore di p e dall'ampiezza campionaria.

6.3 Verosimiglianza "amplificata"

Quando il supporto S_n della funzione di distribuzione empirica F_n è contenuto nel supporto S di F_θ, non è possibile confrontare F_θ e F_n in base al concetto di maggiorazione forte.

Su S_n infatti si ha:

$$\sum_{x\in S_n} f_n(x) = \sum_{i=1}^{n} f_n(x_i) = \sum_{i=1}^{n} \frac{1}{n} = 1;$$

$$\sum_{x\in S_n} f_\theta(x) = \sum_{i=1}^{n} f_\theta(x_i) \le 1,$$

il che impedisce di stabilire un ordinamento di maggiorazione forte tra i valori f_θ / f_n, concetto grazie al quale si ritiene coerente, per le proprietà delle funzioni Schur-convesse, il metodo della minima distanza (o divergenza). Per confrontare due distribuzioni sulla base di tale concetto, è possibile "riadattare" sul supporto S_n la distribuzione F_θ, in modo che dia anch'essa somma uno. Per questo si consideri, al posto di F_θ, la seguente distribuzione condizionata:

$$R_\theta(y) = P_\theta(x \le y | x \in S_n) = \frac{P_\theta(x \le y)}{P_\theta(x \in S_n)} =$$

$$= \frac{P_\theta(x \le y)}{P_\theta\{(x = x_1) \cup \ldots \cup (x = x_n)\}} = \frac{P_\theta(x \le y)}{\sum_{j=1}^{n} P_\theta(x = x_j)}.$$

Si osservi che R_θ è una distribuzione discreta sul supporto S_n, con densità data da $\frac{dR_\theta}{d\nu} = r_\theta$, dove ν è la misura del conteggio. Da un punto di vista teorico, quindi, è possibile confrontare R_θ e F_n con le varie misure di divergenza esistenti.

Si sa infatti che, per $\theta_1, \theta_2 \in \Theta$:

$$\frac{r_{\theta_1}}{f_n} \prec \frac{r_{\theta_2}}{f_n} \Rightarrow . \Phi(F_n, R_{\theta_1}) < \Phi(F_n, R_{\theta_2})$$

Applicare il metodo della minima distanza tra R_θ e F_n significa quindi trovare il valore θ che minimizza la funzione:

$$\Phi(F_n, R_\theta) = \int_{S_n} \phi\left(\frac{r_\theta}{f_n}\right) dF_n = \sum_{x \in S_n} \phi\left(\frac{r_\theta(x)}{f_n(x)}\right) f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \phi\left(n(r_\theta(x_i))\right),$$

dove ϕ è una qualsiasi funzione strettamente convessa.

Questo tipo di procedura può essere interpretata, da un punto di vista intuitivo e logico, nella maniera seguente.

Si ponga ad esempio $\phi(x) = -\ln x$. Minimizzare $\Phi(F_n, R_\theta)$ equivale a minimizzare:

$$\sum_{x \in S_n} -\ln(f_\theta(x)) f_n + n \ln\left(P_\theta(S_n)\right),$$

oppure a massimizzare:

$$\prod_{i=1}^{n} \frac{P_\theta(x = x_i)}{\left(\sum_{i=1}^{n} P_\theta(x = x_j)\right)^n},$$

espressione in cui il numeratore coincide con la funzione di verosimiglianza.

Si cerca quindi il valore θ che, allo stesso tempo, massimizzi la funzione di verosimiglianza e minimizzi la misura di probabilità dell'insieme S_n, il che è senz'altro un controsenso rispetto al concetto di maggiorazione debole (da sopra) e quindi rispetto agli stimatori visti fin'ora. Infatti, la misura di probabilità di S_n è tanto più piccola tanto più la distribuzione scelta (rispetto a θ) attribuisce pesi piccoli ai punti $(x_1, ..., x_n)$, il che avviene in una situazione concettualmente opposta a quella in cui la verosimiglianza è massima. Quindi si cerca un compromesso, si massimizza la verosimiglianza del campione, mettendosi nella situazione più sfavorevole possibile ai dati estratti.

Si osservi comunque che nel precedente rapporto, il denominatore (la verosimiglianza) pesa più del denominatore (dato che $\sum_{i=1}^{n} P_\theta(x = x_j) \leq 1$). Quindi, soluzioni poco "verosimili", caratterizzate cioè da valori troppo piccoli di $P_\theta(x = x_i)$, $i = 1, ..., n$, non sono prese in considerazione, dato che minimizzano maggiormente la verosimiglianza piuttosto che $P_\theta(x \in S_n)$. Ad esempio, se θ

fosse un parametro di locazione, per soluzioni "lontane" ($\theta \to \pm\infty$), la verosimiglianza tenderebbe a zero più velocemente di $P_\theta(x \in S_n)$.

L'ideale motivazione per l'utilizzo di questo metodo sarebbe il caso in cui il campione presenta dei "buchi" imputabili ad un'inefficienza dei metodi di campionamento. In questo caso può essere ragionevole considerare con maggior riguardo quelle soluzioni che concentrano proprio in quei "buchi" gran parte della loro mole, in modo da non favorire ulteriormente quelle soluzioni già favorite dal campione. Si osservi poi che, per $n \to \infty$, $\sum_{i=1}^{n} P_\theta(x = x_j) \to 1$. Per questo motivo il metodo appena introdotto può esser considerato asintoticamente equivalente alla massima verosimiglianza.

Tuttavia, come verrà mostrato da alcune simulazioni, questa procedura produce stime spesso distorte ed errate. Questo dipende, logicamente, dal troppo peso attribuito, all'interno della formula, alla minimizzazione di $P_\theta(x \in S_n)$. Prendendo spunto da quanto visto per le divergenze "penalizzate" si può allora "smorzare" il peso di $P_\theta(x \in S_n)$ all'interno della minimizzazione servendosi di un coefficiente. Si minimizza quindi la funzione:

$$\sum_{x \in S_n} -\ln(f_\theta(x)) f_n + h \cdot n \ln\left(P_\theta(S_n)\right),$$

dove $h \in [0,1]$, il che equivale a minimizzare:

$$\sum_{x \in S_n} -\ln\left(\frac{f_\theta / [P_\theta(S_n)]^{nh}}{f_n}\right) f_n .$$

Dato che i valori di f_θ vengono "amplificati" al crescere di *h*, allora tale funzione può prendere il nome di *divergenza di verosimiglianza con probabilità amplificate*, o, per semplicità, *divergenza di verosimiglianza "amplificata"*.

Gli stimatori corrispondenti possono quindi essere chiamati *MALD* ("minimum amplified likelihood divergence"), per mantenere una notazione uniforme agli stimatori divergenze studiati nei paragrafi *6.1* e *6.2*.

Ovviamente, per $h = 0$ si ottiene la stima di massima verosimiglianza. Dalle simulazioni svolte (appendice *B.3*), invece, si osserva che "amplificando" troppo c'è il rischio di distorsione (come per un impianto audio).

Si può dimostrare che questo metodo di stima non risulta coerente con l'ordinamento di maggiorazione debole (da sopra) . Questo può essere dimostrato con un controesempio.

Un campione casuale semplice di dimensione $n = 4$ assume i valori (ordinati) $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$, con $x_{(2)} = x_{(3)}$. Si ponga $x_1 = x_{(1)}$, $x_2 = x_{(2)} = x_{(3)}$ e $x_3 = x_{(4)}$, quindi $S_n = \{x_1, x_2, x_3\}$. I valori di $f_n(x_i)$, $i = 1,2,3$, si possono allora rappresentare in forma vettoriale con $f_n(x) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

Si considerino due distribuzioni F_{θ_1} e F_{θ_2}. Si rappresentino coi vettori $f_{\theta_1}(x)$ e $f_{\theta_2}(x)$ i valori rispettivamente di $f_{\theta_1}(x_i)$ e $f_{\theta_2}(x_i)$, $i = 1,2,3$ e si supponga che:

$$f_{\theta_1}(x) = (0.2,\ 0.3,\ 0.2) \quad \text{e} \quad f_{\theta_2}(x) = (0.2,\ 0.3,\ 0.3)\ .$$

Si verifica agevolmente che:

$$\int_0^x \frac{f_{\theta_2}}{f_n} \overset{(\mu_{F_n})}{\uparrow} dt \geq \int_0^x \frac{f_{\theta_1}}{f_n} \overset{(\mu_{F_n})}{\uparrow} dt, \quad \forall x \in [0,1],$$

quindi $\frac{f_{\theta_2}}{f_n} \prec^{w}_{(\mu_{F_n})} \frac{f_{\theta_1}}{f_n}$.

Tuttavia è possibile trovare i valori di *h* tali che:

$$\sum_{x \in S_n} -\ln(f_{\theta_2}(x)) f_n + h \cdot n \ln\left(P_{\theta_2}(S_n)\right) > \sum_{x \in S_n} -\ln(f_{\theta_1}(x)) f_n + h \cdot n \ln\left(P_{\theta_1}(S_n)\right).$$

In particolare la precedente disequazione vale per $h > 0.3045$, dato che:

$$\sum_{x \in S_n} -\ln(f_{\theta_2}(x)) f_n = 4.017\ ,\quad \sum_{x \in S_n} -\ln(f_{\theta_1}(x)) f_n = 4.422\ ;$$

$$n \ln(P_{\theta_2}(S_n)) = -0.223\ ,\quad n \ln(P_{\theta_1}(S_n)) = -0.356\ .$$

Nell'appendice *B.3* sono riportati medie e M.S.E. (men square error) degli stimatori studiati in questo paragrafo, per distribuzioni Binomiali con diversi valori di *p* e diverse ampiezze campionarie. Dalle tabelle si possono osservare dei miglioramenti rispetto al metodo della massima verosimiglianza, in termini di M.S.E.. Per valori di *h* piuttosto "piccoli", cioè tra 0.1 e 0.4, la distorsione delle stime è comunque ad un livello accettabile, mentre il M.S.E. è leggermente inferiore.

Capitolo 7

Stima dei parametri nel modello di Rasch

7.1 Minima divergenza applicata al modello di Rasch

Il metodo della minima divergenza, che corrisponde ad un criterio logico piuttosto semplice, può essere applicato in generale a qualsiasi tipo di modello. In ogni situazione sarebbe auspicabile studiare ed individuare la misura di divergenza "migliore" per la stima dei parametri.

Come caso di studio, è stato preso in considerazione il *modello di Rasch*, in cui il problema della stima è senz'altro più complesso rispetto alle situazioni standard.

Il modello Rasch è utilizzato per l'analisi e lo studio di test, questionari, o altri strumenti di valutazione adatti a misurare l'abilità (o caratteristiche simili) di alcuni soggetti. Si sottopone un gruppo di n individui ad un identico questionario e si cerca di determinare dei valori numerici per descrivere e collocare su una scala (ovviamente convenzionale) le diverse abilità dei soggetti e le difficoltà delle m domande presenti nel questionario.

I risultati vengono raccolti in una matrice $n \times m$ $\{x_{ij}\}$. Il valore di x_{ij} all'interno di ogni cella può essere 0 o 1 a seconda, rispettivamente, che il soggetto i dia una risposta giusta oppure sbagliata alla domanda j. Le probabilità di successo (o insuccesso) dipendono dai parametri (incogniti) β_j (j = 1, 2, ..., m) di "difficoltà" delle domande e θ_i (i = 1, 2, ..., n) di "abilità" degli individui e seguono la seguente formula:

$$p_{ij}(x_{ij}) = P(X_{ij} = x_{ij} \,|\, \theta_i, \beta_j) = \frac{\exp[x_{ij}(\theta_i - \beta_j)]}{1 + \exp(\theta_i - \beta_j)}.$$

I parametri di difficoltà delle domande β_j sono tali da dare somma zero, ossia

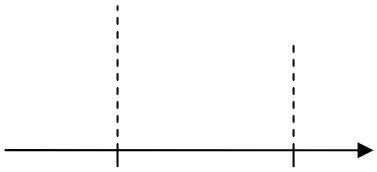
$$\sum_{J=1}^{m} \beta_j = 0.$$

Contando il numero di risposte esatte ottenuto da ciascun individuo si ottengono i *punteggi totali* di riga dati da $\sum_{j=1}^{m} x_{ij}$.

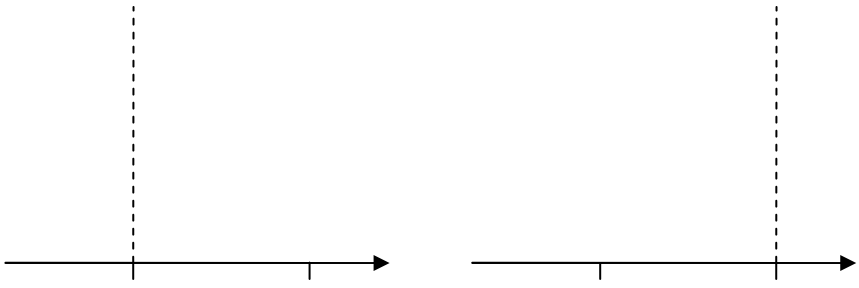
Il modello di Rasch si basa sull'assunzione che tali punteggi totali siano statistiche sufficienti per i parametri θ_i. Quindi il punteggio totale di un soggetto contiene tutta l'informazione disponibile riguardo la sua abilità. E' interessante, dunque, sapere a quante domande l'individuo ha risposto, piuttosto che a quali. Allo stesso modo si assume che i punteggi totali di colonna siano statistiche sufficienti per i parametri β_j di difficoltà delle domande.

Queste ipotesi di sufficienza suggeriscono che la stima del parametro di abilità di un soggetto non debba dipendere dalla difficoltà delle domande a cui ha risposto e, viceversa, che la stima del parametro di difficoltà di una domanda non debba dipendere dalle abilità dei soggetti che hanno risposto a tale domanda.

Le stima degli $m+n$ parametri possono essere calcolate con il metodo della minima divergenza. Infatti, per ogni cella della matrice, si pone un problema di minima divergenza tra una distribuzione "teorica" data da $p_{ij}(x_{ij})$ nei punti 0 e 1 e una distribuzione empirica degenere nel punto 0 o 1 (a seconda che la risposta rappresentata da x_{ij} sia rispettivamente errata o corretta).



distribuzione teorica



distribuzione empirica (x_{ij}=0) *distribuzione empirica* (x_{ij}=1)

Minimizzando congiuntamente tutte le $n \times m$ divergenze si ottengono ovviamente diverse espressioni, a seconda del tipo di divergenza utilizzato. Nel seguito verranno confrontate le stime di massima verosimiglianza (min. Kullback-Leibler) e minimo Chi^2.

Utilizzando Kullback-Leibler, l'espressione da minimizzare risulta:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} -\ln p_{ij}(x_{ij}),$$

le soluzioni della minimizzazione sono le stime di massima verosimiglianza dei parametri.

Utilizzando la divergenza di Chi^2, l'espressione da minimizzare è:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{\left(p_{ij}(x_{ij})-1\right)^2}{p_{ij}(x_{ij})}+\frac{\left(1-p_{ij}(x_{ij})\right)^2}{1-p_{ij}(x_{ij})}\right]=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(p_{ij}(x_{ij})-1\right)^2}{p_{ij}(x_{ij})\left(1-p_{ij}(x_{ij})\right)}$$

le soluzioni della minimizzazione sono le stime di minimo Chi-quadrato dei parametri.

Per calcolare la bontà di adattamento delle stime, si utilizza solitamente un indice basato sui *residui standardizzati*, dati da:

$$z_{ij}=\frac{x_{ij}-E[X_{ij}]}{\sqrt{V[X_{ij}]}}=\frac{x_{ij}-p_{ij}(1)}{\sqrt{p_{ij}(1)\left(1-p_{ij}(1)\right)}}.$$

Le stime "migliori" in termini di adattamento sono quelle che minimizzano la somma dei quadrati dei residui, quindi:

$$\sum_{i=1}^{n}\sum_{J=1}^{m}(z_{ij})^2=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(p_{ij}(x_{ij})-1\right)^2}{p_{ij}(x_{ij})\left(1-p_{ij}(x_{ij})\right)}.$$

Allora si nota subito che, per definizione, la stima di minimo Chi^2 è senz'altro quella che si adatta meglio.

Dato che, per motivazioni che si vedranno in seguito, la stima di minimo Chi^2 non è solitamente presa in considerazione per il modello Rasch, è lecito sollevare dei dubbi sulla correttezza dell'utilizzo di un indice di adattamento che praticamente coincide con la statistica Chi^2.

7.2 Caratteristiche della stima di minimo Chi^2

Da alcuni esempi si può subito osservare che, a differenza delle stime di massima verosimiglianza, le stime di minimo Chi-quadrato non sono conformi rispetto alla proprietà di sufficienza dei punteggi di riga (o di colonna). Infatti è possibile che due soggetti con uguale punteggio (cioè numero di risposte corrette) abbiano stime differenti per quanto riguarda il parametro di abilità. Analogamente, due domande cui hanno risposto lo stesso numero di soggetti, possono avere stime diverse dei parametri di difficoltà. Ciò è senz'altro in contraddizione con le più consuete ipotesi del modello Rasch. Per questo motivo i risultati forniti dalla stima di minimo Chi-quadrato non sono solitamente considerati accettabili (Linacre, 2004).

Le stime, in tutti i prossimi esempi, sono state calcolate con *Mathematica 5.0* (Wolfram, 2003).

Esempio 7.2.A

Si consideri la seguente tabella (3×3) delle osservazioni:

0	1	0	**1**
0	0	1	**1**
1	0	1	**2**
1	**1**	**2**	

I primi due soggetti rispondono solamente ad una domanda e, dall'altro lato, le prime due domande ottengono risposta da un solo soggetto. In questa situazione le stime MLE per i primi due soggetti e per le prime due domande coincidono, per la proprietà di sufficienza dei punteggi totali.

I valori delle stime di minimo Chi^2, invece, risultano:

θ_1	-0.3463
θ_2	-0.4938
θ_3	0.45542
β_1	0.3655
β_2	0.2181
β_3	-0.5836

Minimo Chi^2

θ_1	-0.7684
θ_2	-0.7684
θ_3	0.79747
β_1	0.5219
β_2	0.5219
β_3	-1.0439

Massima verosimiglianza

Da quanto si osserva, il primo e il secondo soggetto rispondono correttamente ad una domanda su tre. La stima di massima verosimiglianza delle abilità di tali soggetti è identica, mentre stimando i parametri con Chi^2 il primo individuo risulta più abile del secondo. Questo risultato può avere una giustificazione logica sebbene, come si è detto, non sia solitamente accettabile. Infatti il secondo individuo risponde alla terza domanda, che è la più "facile" (dato che due su tre la conoscono), quindi si può considerare meno abile del primo, che risponde ad una domanda più "difficile".

Si può intuire anche da altri simili esempi che le stime di minimo Chi^2, pur non essendo coerenti con le statistiche sufficienti, corrispondono spesso a logiche (più o meno complicate) che derivano dalla composizione della tabella stessa. Comunque non sempre è facile comprendere tali logiche e trovare delle giustificazioni ragionevoli alle stime, anche con tabelle di piccole dimensioni.

Esempio 7.2.B

Si consideri la seguente tabella (3×3) delle osservazioni:

0	0	1	**1**
0	1	1	**2**
1	1	0	**2**
1	**2**	**2**	

I valori delle stime di minimo Chi^2 risultano:

θ_1	-0.4554
θ_2	0.49385
θ_3	0.34637
β_1	0.5836
β_2	-0.3655
β_3	-0.2181

Minimo Chi^2

θ_1	-0.7974
θ_2	0.76845
θ_3	0.76845
β_1	1.04395
β_2	-0.5219
β_3	-0.5219

Massima verosimiglianza

In questa situazione, il terzo individuo è considerato meno abile del secondo, pur essendo l'unico che ha risposto alla prima domanda. Inoltre la terza domanda è considerata più difficile della seconda, pur essendo l'unica cui ha saputo rispondere l'individuo meno abile (il primo).

In questi casi l'analisi deve essere più profonda, i risultati vanno interpretati anche considerando i valori numerici associati ai vari parametri stimati e alle distanze tra di essi. Ad esempio è possibile stimare i parametri di abilità, dopo aver fissato in maniera opportuna i parametri di difficoltà delle domande. In questo caso è possibile individuare, a parità del punteggio totale, quale sia il "pattern" di risposte cui corrisponda la stima di abilità maggiore.
A differenza di quanto avviene per la massima verosimiglianza, è possibile esplicitare la soluzione della stima di minimo Chi^2 dell'abilità di un soggetto, in funzione dei parametri β corrispondenti alle domande cui ha o non ha risposto. Questo rappresenta senza dubbio una notevole comodità.
Se il soggetto *i*-esimo ha risposto a *p* domande su *m*, si indichi con β'_{ik}, $k = 1,..,p$, i parametri associati alle domande che hanno avuto risposta e con β''_{ih}, $h = 1,\ldots,$ *m-p*, i parametri associati alle domande non risposte. Allora si può mostrare che (Bertoli-Barsotti):

$$_{CHI}\hat{\theta}_i = \frac{1}{2}\ln\left\{\frac{\sum_{k=1}^{p}\exp(\beta'_{ik})}{\sum_{h=1}^{m-p}\exp(-\beta''_{ih})}\right\}.$$

A seconda del numero totale di domande e del numero di risposte corrette, è possibile individuare il pattern dell'individuo più abile.

Si considerino i seguenti casi, in cui le domande sono quattro e i soggetti rispondono ad una, due oppure tre domande.

Ad esempio se β_1=3, β_2=2, β_3=-2, β_4=-3, per rendere massimo il rapporto nella parentesi quadra, rispondendo a una domanda, bisogna rispondere alla domanda più difficile. Infatti il rapporto nella parentesi quadra cresce tanto più è alto il numeratore (automaticamente il denominatore decresce). Il soggetto è tanto più abile tanto più la domanda cui ha risposto è difficile.

Si consideri ora il caso in cui il soggetto risponde a due domande su quattro.

Dalla formula si capisce che l'individuo più abile deve rispondere alla domanda più difficile e non può non rispondere alla domanda più facile. Infatti, rispondendo alla prima domanda si massimizza il numeratore con e^3. Non rispondendo però alla domanda più facile, si vanifica il risultato, perché anche il denominatore viene massimizzato (sempre con e^3).

Rispondendo alla domanda più facile e alla più difficile (1,0,0,1) il rapporto risulta $(e^3 + e^{-3})/(e^2 + e^{-2}) = 2.67601$, quindi l'abilità del soggetto è $1/2\ln(2.67601) = 0.49216$. Ottenendo un pattern opposto (0,1,1,0), rispondendo cioè alla domande di difficoltà "intermedia", il rapporto risulta essere il reciproco del precedente, $(e^2 + e^{-2})/(e^3 + e^{-3}) = 0.373691$ e l'abilità del soggetto è opposta rispetto alla precedente, quindi -0.49216.

In caso di pattern (1,1,0,0), ad esempio, il rapporto risulta $(e^3 + e^2)/(e^3 + e^2) = 1$, quindi l'abilità del oggetto è sicuramente zero, come d'altra parte avviene (si può verificare semplicemente) per i pattern (1,0,1,0), (0,1,0,1), (1,1,0,0) e (0,0,1,1). Questo è dovuto alla diposizione "simmetrica", in questo caso particolare, dei parametri di difficoltà (β_1=3, β_2=2, β_3=-2, β_4=-3). Si possono ottenere risultati diversi cambiando i valori dei parametri β_j, come si può osservare dal seguente esempio.

Esempio 7.2.C

Si osservi la seguente tabella 6×4 delle osservazioni:

1	0	0	1
0	1	1	0
1	0	1	0
0	1	0	1
1	1	0	0
0	0	1	1

Cambiando i parametri β_j, e ponendo β_1=3, β_2=1.5, β_3=-2, β_4=-2.5 sarà impossibile ottenere rapporti che diano uno: allora si avranno senz'altro stime diverse da zero. Tuttavia raggruppando a coppie i vari pattern, si osservi come le combinazioni tra i vari β_j diano origine a rapporti che sono l'uno il reciproco dell'altro. Come si è visto, questo da origine a stime opposte. Le stime di minimo Chi^2 sono infatti:

θ_1	0.487164
θ_2	-0.487164
θ_3	0.244283
θ_4	-0.244283
θ_5	0.113668
θ_6	-0.113668

Questo esempio, oltre che a confermare quello che è stato detto precedentemente, fa intuire che, tendenzialmente, rispondere a domande più difficili paga maggiormente.

Esempio 7.2.D

Si considerino pattern in cui i soggetti hanno risposto a tre domande su quattro, dove β_1=3, β_2=2, β_3=-2, β_4=-3. Allora, data la tabella 4×4

1	1	1	0
1	1	0	1
1	0	1	1
0	1	1	1

si ottengono le stime:

θ_1	0.15908
θ_2	0.65753
θ_3	2.50459
θ_4	2.51237

In questo caso risulta più importante non rispondere alla domanda più facile piuttosto che rispondere alla domanda più difficile. Infatti, come si vede anche dalla formula, il rapporto è massimizzato quando il denominatore è minimo, cioè pari a e^{-3} (risposta più difficile errata).

Il caso in cui le domande sono cinque è piuttosto significativo, perché avendo un numero dispari di domande si possono creare pattern strutturalmente diversi.
Ad esempio, si ponga β_1=3, β_2=2, β_3=0, β_4=-2, β_5=-3 .
Per un soggetto che risponde ad una domanda, il principio è identico al caso delle quattro domande, quindi il più abile è colui che risponde alla domanda più difficile. Un soggetto che risponde a due domande è il più abile se segue il pattern (1,0,0,0,1), altrimenti nelle situazioni intermedie è più abile chi risponde a domande più difficili, ad esempio (0,1,1,0,0) è meglio di (0,0,1,1,0).
La situazione di tre risposte su cinque rappresenta una novità rispetto al caso delle quattro domande. Infatti questa situazione funge da spartiacque, tre è più della metà delle domande, quindi chi risponde a tre domande ha parametro di abilità positivo e "non può permettersi" di sbagliare una domanda semplice (cioè con parametro negativo). A tal proposito si consideri la tabella:

Esempio 7.2.E

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
1	1	0	1	0
1	1	0	0	1
1	0	0	1	1
0	1	0	1	1
0	1	1	0	1
1	0	1	1	0
1	0	1	0	1

Le stime di minimo Chi^2 corrispondenti sono:

θ_1	0.01787
θ_2	0.42977
θ_3	0.92829
θ_4	0.13479
θ_5	0.59407
θ_6	1.44112
θ_7	0.988079
θ_8	0.063064
θ_9	0.024134
θ_{10}	0.516398

Quindi i soggetti più abili non possono sbagliare le due domande più semplici. Il più abile di tutti è il soggetto che, oltre alle due domande più semplici, risponde anche alla più difficile. Il meno abile di tutti è il soggetto che sbaglia le due più semplici, pur rispondendo alle tre più difficili.

Infine, per un soggetto che risponde a quattro domande su cinque, il principio è identico al caso delle tre domande su quattro domande.

I risultati ottenuti per quattro o cinque domande possono essere generalizzate a questionari più articolati. Ovviamente bisognerà tener conto di un maggior numero di combinazioni e di una molteplicità di situazioni, che comunque possono essere sintetizzate con situazioni più semplici come quelle degli esempi precedenti.

Il principio generale quindi, è che l'individuo poco abile (ad esempio che risponde ad una domanda) va considerato più abile tanto più la domanda cui risponde è difficile. Tanto più invece, il soggetto è abile (risponde a più domande, ad es. tre su quattro, tre su cinque) tanto più viene penalizzato se non risponde alle domande semplici, che "non può permettersi di sbagliare". Nel caso intermedio (due domande su quattro) l'individuo più abile conosce le domande con i valori più "estremi". Talvolta, inoltre, l'abilità viene stimata pari a zero, questo avviene quando le difficoltà delle domande (risposte e non) si "compensano" tra loro.

Riprendendo l'esempio 7.2.2, ora è possibile spiegare i risultati con una logica: il terzo individuo risulta meno abile del secondo perché entrambi hanno risposto a due domande su tre (quindi hanno parametri di abilità positivi) ma il terzo soggetto non ha risposto alla domanda facile (parametro di difficoltà negativo), quindi viene penalizzato nella stima.

Concludendo, pare che la semplice formula della divergenza di Chi^2 dia origine a stime che rispettano un criterio logico ben preciso il quale, pur non corrispondendo alle ipotesi più consuete del modello Rasch, è senz'altro interessante e utile, data la sua ragionevolezza.

7.3 Soluzione di alcuni problemi di stima

In alcune circostanze, le stime dei parametri del modello di Rasch divergono, ossia i metodi di stima noti forniscono dei valori numerici troppo elevati (in valore assoluto) per essere considerati. Questo avviene, ad esempio, quando un individuo risponde correttamente a tutte le domande, oppure a nessuna. Analogamente, lo stesso problema sorge quando ad una domanda rispondono correttamente tutti i soggetti, oppure nessuno.

Calcolando le stime con *Mathematica 5.0* (Wolfram, 2003) si ottengono numeri sempre più alti, tanto più aumenta la precisione dei calcoli.

Posti β_1=3, β_2=2, β_3=0, β_4=-2, β_5=-3, ad esempio, l'abilità del soggetto che risponde a tutte le domande (1,1,1,1,1) ha una stima di massima verosimiglianza che risulta (con la precisione di calcolo "standard" di *Mathematica* 5.0) pari a 37.2063 e una stima di Chi^2 pari a 35.9791. L'abilità del soggetto che non risponde a nessuna delle domande (0,0,0,0,0), invece, ha una stima di massima verosimiglianza pari a -37.2063 e una stima di Chi^2 pari a -35.9791.

Questi risultati non sono accettabili perché di un ordine di grandezza che non li rende paragonabili e confrontabili coi valori delle stime ottenute con pattern "vicini". Ad esempio, per (0,1,1,1,1) si ottiene una stima di massima verosimiglianza pari a 2.66583 e una stima di Chi^2 pari a 2.57438, per (0,0,0,0,1) si ottiene una stima di massima verosimiglianza pari a - 2.66583 e una stima di Chi^2 pari a - 2.57438.

Il principio alla base di questo inconveniente è molto logico. Nel caso di pattern (1,1,..,1,1,1) si cerca il parametro di abilità tale per cui la distribuzione "teorica" sia più "vicina" possibile a 1, qualunque sia la domanda. Ovviamente, tanto più è elevato il valore del parametro di abilità, tanto più si ha la certezza di rispondere correttamente alle domande, perciò la stima diverge. Questo non avviene per un pattern di tipo (0,1,..,1,1,1) perché la presenza di uno zero "ancora" le stime tra valori ragionevoli. Vale il viceversa per (0,...,0,0,0,0), mentre il discorso è analogo quando i pattern riguardano le domande anziché i soggetti.

Alcuni programmi cancellano automaticamente le righe (o le colonne) di questo tipo, eliminando il problema ma modificando in un certo senso la situazione reale.

Nel programma *Winsteps* (Linacre, 2009) è proposta una soluzione alternativa che consiste nel modificare il valore dei punteggi totali estremi ("extremescores"). Se

un individuo risponde correttamente a tutte le domande, si modifica il suo punteggio sostituendo al valore 1, ottenuto per la domanda che ha avuto meno risposte, un valore in (0,1): solitamente 0.5 (per semplicità). Ci si comporta in maniera opposta in caso di pattern (0,0,…,0).
In tal modo si possono ottenere delle stime accettabili senza modificare drasticamente i dati iniziali.
Ad esempio, posti β_1=3, β_2=2, β_3=0, β_4=-2, β_5=-3, l'abilità del soggetto con un pattern di risposte (0.5,1,1,1,1) viene stimata pari a 3.73975 (massima verosimiglianza) oppure 3.50815 (minimo Chi^2).
Un possibile inconveniente è che, cambiando anche solo un punteggio si può alterare la struttura della matrice dei dati, qualora per esempio ci siano più possibilità per sostituire il valore 0.5.

Esempio 7.3.A
Si consideri la tabella 4×4

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

In questo caso si può sostituire 0.5 nella prima o nella seconda domanda. In ogni caso una delle due domande viene penalizzata. Sostituendo in questo modo:

$$\begin{pmatrix} 0.5 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Alla seconda domanda corrisponde un punteggio "1.5" anziché "2", quindi viene in un certo senso penalizzata.
Si ottengono infatti le stime:

θ_1	2.18665
θ_2	-0.0089
θ_3	-0.0089
θ_4	-0.0089
β_1	1.0892
β_2	0.46424
β_3	-0.7767
β_4	-0.7767

Massima verosimiglianza

θ_1	1.58078
θ_2	-0.0055
θ_3	0.02980
θ_4	-0.0298
β_1	0.54701
β_2	0.24673
β_3	-0.4050
β_4	-0.3887

Minimo Chi^2

In entrambi i casi la stima di θ_1 ha un valore ragionevole. Si osservi però che il parametro di difficoltà della prima domanda è maggiore di quello della seconda, sebbene entrambe hanno avuto due risposte.

Un altro caso in cui le stime divergono è quando la matrice dei dati è "*mal-condizionata*" cioè ha una struttura di questo tipo:

$$\left[\begin{array}{c:c} 1\backslash 0 & 1 \\ \hdashline 0 & 1\backslash 0 \end{array}\right].$$

Una struttura del genere è abbastanza chiara, si riescono a distinguere anche ad occhio i gruppi dei soggetti "più" e "meno" abili ed il gruppo delle domande "più" e "meno" difficili. Naturalmente gli individui più abili tendono a rispondere correttamente alle risposte più facili, mentre quelli meno abili tendono a non rispondere alla domande più difficili. In questo caso ogni soggetto del primo gruppo è più abile di ogni soggetto del secondo gruppo. Infatti non si verifica mai che un individuo del secondo gruppo risponda ad una domanda cui non abbia risposto ogni individuo del primo gruppo. Il divario "netto" e troppo profondo fra i due gruppi fa divergere le stime.

Questo problema di stima è una logica conseguenza del metodo della minima distanza. Si cercano i parametri di θ_i e β_j tali che la distribuzione "teorica" sia più "vicina" possibile a 1, nel secondo quadrante, o a 0, nel terzo quadrante. Per il

metodo della minima distanza, tanto più è elevato il valore θ_i per i soggetti del primo e tanto più è basso il valore β_j per le domande "più facili", tanto più si ha la certezza di rispondere correttamente, nel secondo quadrante. Il discorso è piuttosto analogo per quanto riguarda il terzo quadrante. Per questo motivo, le stime divergono in questo modo: $\theta_i \to \infty$ nel primo gruppo e $\theta_i \to -\infty$ nel secondo; $\beta_j \to \infty$ nel primo gruppo e $\beta_j \to -\infty$ nel secondo.

Non sono note delle soluzioni a questo problema. Prendendo spunto dalla soluzione di Winsteps (Linacre,2009) per i punteggi estremi, è possibile modificare i punteggi in modo da creare un "link" tra i due gruppi. Sostituendo un valore $0<\alpha<1$ (ad esempio "0.5") al punteggio "1" corrispondente al soggetto "più abile" e alla domanda "più facile" si può verificare che almeno un individuo del secondo gruppo ottenga un punteggio superiore, nella domanda "più facile", rispetto all'individuo "più abile". Equivalentemente, è possibile sostituire α al punteggio "0" corrispondente al soggetto "meno abile" e alla domanda "più difficile".

Questo link solitamente ridimensiona notevolmente le distanze tra i due gruppi e riporta le stime su valori accettabili.

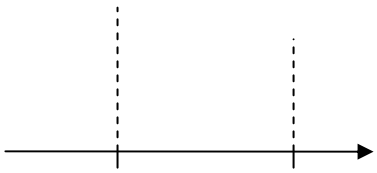
Per sintetizzare, questo metodo verrà chiamato d'ora in avanti "*metodo- α*".

La controindicazione per questo metodo-α è che comunque la struttura della tabella viene alterata, come nell'esempio *7.3.A*. Infatti il metodo va utilizzato solamente nel caso ci sia un effettivo problema di mal-condizionamento della matrice, altrimenti si rischia solamente di distorcere la realtà rappresentata dai dati. Inoltre in ogni situazione bisogna scegliere con attenzione la casella dove sostituire il punteggio, magari dopo alcuni ragionamenti e facendo qualche tentativo: questo può risultare complicato.

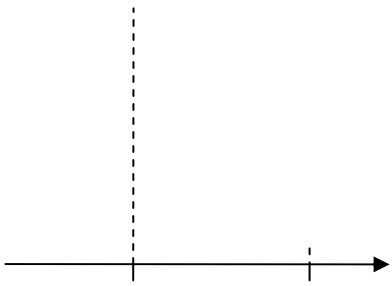
La soluzione alternativa proposta consiste nell'ipotizzare che le distribuzioni empiriche in ogni casella non siano degeneri (0,1), ma frequenze "molto vicine" a 0 e a 1 (con somma 1 ovviamente). Ad esempio in caso di successo si ipotizza di ottenere 1 con frequenza 0.99 (anziché 1) e 0 con frequenza 0.01 (anziché 0). In caso di insuccesso, invece, 0 con frequenza 0.99 (anziché 1) e 1 con frequenza 0.01 (anziché 0). Si può interpretare questa modifica pensando che ogni domanda venga ripetuta 100 volte, e che ogni individuo risponda correttamente 99 volte (oppure 1 volta) su 100.

In questo modo, si cercano i parametri di θ_i e β_j tali che la distribuzione "teorica" sia più "vicina" possibile a 0.99, nel secondo quadrante della matrice mal-condizionata, o a 0.01, nel terzo quadrante. Per raggiungere tali valori, comunque "vicini" a 1 e 0, sono sufficienti stime dei parametri con un ordine di grandezza ragionevole. Si noti inoltre che questa soluzione può essere adottata anche per ottenere le stime in presenza di punteggi estremi (1,1,…,1) o (0,0,..,0), in alternativa alla soluzione di Winsteps. In questo caso è preferibile smorzare ulteriormente le distanze tra le stime, utilizzando ad esempio (0.2, 0.98) oppure (0.3, 0.97).

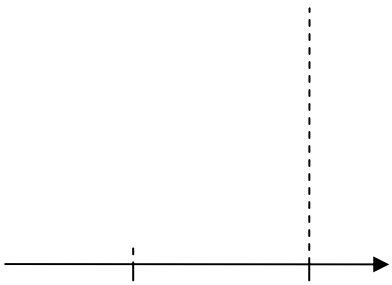
In generale, dato β<1 si può sostituire la distribuzione di frequenze (0,1) con una distribuzione (β, 1-β), questo metodo verrà chiamato d'ora in avanti "*metodo-β*". Il valore percentuale β può essere interpretato come il grado di "incertezza" attribuito alle risposte osservate. β può esser variato a seconda di quanto si vuole attutire l'effetto di una eventuale struttura "estrema" della tabella di dati. In ogni situazione è possibile determinarne il valore ideale: aumentando β si ottengono dei valori delle stime sempre più "vicini" tra loro, a scapito di risultati meno precisi in termini di adattamento. Dagli esempi studiati, pare che per ottenere notevoli miglioramenti nei risultati siano sufficienti valori tra 0.001 e 0.05.



distribuzione teorica



distribuzione empirica (x_{ij}=0)



distribuzione empirica (x_{ij}=1)

Applicando il metodo della minima divergenza rispetto a Kullback-Leibler e a Chi^2, si ottengono le seguenti funzioni da minimizzare.

$$\sum_{i=1}^{n}\sum_{j=1}^{m} -(1-\beta)\ln\ p_{ij}(x_{ij}) - \beta\ \ln(1-\ p_{ij}(x_{ij}))\ ,$$

per quanto riguarda Kullback-Leibler (massima verosimiglianza);

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{\left(p_{ij}(x_{ij})-(1-\beta)\right)^2}{p_{ij}(x_{ij})}+\frac{\left(1-\ p_{ij}(x_{ij})-\beta\right)^2}{1-\ p_{ij}(x_{ij})}\right]=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(p_{ij}(x_{ij})-(1-\beta)\right)^2}{p_{ij}(x_{ij})\left(1-p_{ij}(x_{ij})\right)},$$

per quanto riguarda il minimo Chi^2.

Correggendo (alterando) leggermente le frequenze delle osservazioni si riesce a ridimensionare notevolmente le divergenze, e le stime non divergono.

Questa sostituzione delle frequenze è lieve ed uniforme, quindi, a differenza delle correzioni viste in precedenza, non modifica gli equilibri all'interno della tabella. Lo si nota osservando che il metodo-β può essere applicato anche a matrici che non sono mal-condizionate. In tal caso i risultati ottenuti non vengono alterati, se non da un punto di vista di "scala", la cui dimensione dipende ovviamente dal valore di β utilizzato. Spesso si ottengono stime praticamente equivalenti a quelle ottenute con le frequenze "vere" (0,1). Si può dire allora che il metodo-β sia piuttosto "neutrale".

Per questi motivi, confermati da un discreto numero di esempi (riportati nell'appendice C), il metodo-β pare preferibile rispetto al metodo-α.

Conclusioni

La maggiorazione si è rivelata essere uno strumento matematico molto utile anche in ambito statistico. Nel capitolo *1* si è di definito tale concetto in modo più generale possibile, estendendolo dalla maggiorazione vettoriale alla maggiorazione rispetto ad una misura qualsiasi (Bertoli-Barsotti, 2001). Sotto certe condizioni, infatti, è possibile confrontare la "dissomiglianza" tra coppie di vettori o funzioni, riordinandoli con dei particolari pesi determinati dalla misura. E' stata fornita anche una definizione generale di maggiorazione debole, utile qualora non vi siano le condizioni necessarie per effettuare un confronto tramite la maggiorazione (forte).

Questo lavoro ha fornito una chiave di lettura alternativa per lo studio di alcuni noti metodi di stima, basati sul metodo di minima distanza tra distribuzioni: è stato possibile, nel capitolo *3*, osservare come certe proprietà di tali procedure siano dimostrabili in modo molto semplice, applicando i teoremi del capitolo *2*.

Nel capitolo *4* sono state individuate alcune condizioni sufficienti (ricavate dal teorema *1.3.A*) sotto le quali una misura di divergenza è un funzionale Schur-convesso, ossia rispetta il pre-ordinamento di maggiorazione forte. Grazie a questo risultato, si è definita una classe piuttosto ampia di divergenze con tale proprietà: a tale classe appartiene ad esempio la divergenza di Kullback-Leibler, su cui si basa il metodo di massima verosimiglianza, e le divergenze di Chi^2 e di Hellinger.

Alcune divergenze, le cui formule soddisfano le ipotesi del teorema 1.3.D, risultano coerenti anche rispetto alla maggiorazione debole (ad esempio Kullback-Leibler). Nel paragrafo *4.4* si è dimostrato (teorema *4.3.A*) che anche altri tipi di funzionali, con forma tale da non rendere possibile l'applicazione del teorema *1.3.D*, sono comunque coerenti con la maggiorazione debole: tra queste figurano ad esempio le divergenze di Chi^2 e di Hellinger.

Nel capitolo *5*, dopo aver verificato che tutte le misure di divergenza studiate fanno parte della classe generale delle divergenze potenziate, si è svolto un approfondimento sulle tre misure più importanti, studiandone le proprietà. A tal proposito si è visto che gli stimatori basati sulla minima divergenza possono essere efficienti e\o robusti: ciò dipende dalla forma di una particolare funzione,

legata alla formula della divergenza stessa, detta RAF (funzione di adeguamento dei residui; Lindsay, 1994).

Nel capitolo *6* sono state proposte alcune modifiche alle misure di divergenza più utilizzate, per migliorare le stime qualora il numero di osservazioni campionarie non sia sufficientemente elevato. Infatti, dallo studio effettuato nel capitolo *4*, risulta che il confronto della dissomiglianza tra la distribuzione empirica e le varie distribuzioni teoriche (con supporto S) risulta più complicato qualora non vi sia almeno un'osservazione per ogni punto del supporto S. Indicando con S_n il supporto campionario, cioè l'insieme dei diversi punti osservati, può infatti accadere che $S_n \subset S$. In questa situazione, piuttosto frequente, soprattutto se il campione è piccolo, si può utilizzare, per il confronto, la maggiorazione debole, in cui è determinante il peso della probabilità di S_n. Nelle formule delle più note divergenze, tale probabilità può avere un peso sproporzionato (Hellinger, Chi^2) o non averne affatto (K.L.). Per questo motivo, nel capitolo *6* sono stati proposte delle misure di divergenza alternative, modificate controllando, con un coefficiente h, il peso di $P(S_n)$ all'interno della formula. Dalle simulazioni effettuate nell'appendice *B* si possono osservare i miglioramenti delle stime (in termini di distorsione e M.S.E.) ottenute utilizzando queste divergenze modificate, al variare di h, dell'ampiezza campionaria e della forma della distribuzione. In particolare pare che la divergenza di "verosimiglianza amplificata" (paragrafo *6.3*) sia quella che dia i risultati migliori: per certi valori di h (intorno a 0.1, 0.2) si ottengono stime non distorte e con un M.S.E. inferiore alla stime M.L.E..

Nel capitolo *7* si è cercato di sfruttare le conoscenze acquisite per studiare metodi di stima di minima divergenza nel modello di Rasch. In questo modello, le stime di minimo Chi^2 non vengono solitamente considerate, perché non conformi alla proprietà di sufficienza dei punteggi totali. Tuttavia è stato possibile, esplicitando la formula matematica dello stimatore (min. Chi^2) ed analizzando un buon numero di situazioni, intuire il criterio logico alla base dei risultati. Tale criterio risulta molto interessante e ragionevole: la stima di minimo Chi^2 prende in considerazione alcuni aspetti e alcune logiche, all'interno dei dati, alle quali la stima di massima verosimiglianza non è assolutamente sensibile.

Nel paragrafo *7.3* sono state proposte due differenti soluzioni a problemi di stima nel modello Rasch, relativi ai "punteggi estremi" ed al "mal-condizionamento" delle matrici: questi metodi sono stati chiamati (almeno provvisoriamente)

metodo-α e *metodo-β*. Nell'appendice *C*, i due metodi sono stati confrontati in diverse situazioni (tabelle di varie dimensioni e strutture, con o senza punteggi estremi…) e sotto diversi aspetti (adattamento, valori numerici delle stime e "neutralità"): alla luce di questi esempi pare che il metodo-β, basato proprio sul concetto di minima distanza, dia risultati migliori.

In conclusione, lo studio della maggiorazione e delle misure di divergenza si è dimostrato utile non solo dal punto di vista teorico ma anche dal punto di vista pratico: infatti da un lato è stato possibile classificare gran parte dei metodi di stima (basati su funzionali Schur-convessi), dall'altro sono state proposte soluzioni per la stima in situazioni particolari, modificando le più note misure di divergenza in modo da risolvere eventuali problemi o difetti dei metodi già noti.

Le conoscenze acquisite in questa tesi potranno, ottimisticamente, essere utili per risolvere anche altre problematiche, relative alla stima. Inoltre alcuni argomenti trattati andrebbero ulteriormente approfonditi e sviluppati.

Innanzitutto, le misure di divergenza studiate nei capitoli *4* e *5* hanno una forma simile, riconducibile a quella definita nel teorema *1.3.A*: una prospettiva di grande interesse sarebbe quella di trovare altri funzionali Schur-convessi, al di fuori dalla classe individuata, quindi con una formula che non sia necessariamente del tipo:

$$\Phi(F,G)=\int_S \phi\left(\frac{g}{f}\right)dF \quad (\phi \text{ strettamente convessa}).$$

Sarebbe importate anche approfondire, con ulteriori simulazioni, lo studio delle divergenze "modificate" presentate nel capitolo *6*, con l'obiettivo di identificare più chiaramente i valori "ideali" del coefficiente *h*, al variare dell'ampiezza campionaria e della forma della distribuzione.

Per quanto riguarda il modello di Rasch, potrebbe essere interessante stimare i parametri anche con altre misure di divergenza, oltre che alle due utilizzate. In caso, vi sono diverse soluzioni possibili: ad esempio si potrebbe tentare con la divergenza di Hellinger, oppure con le stesse divergenze "modificate" proposte nel capitolo *6*. La speranza è quella di trovare stime differenti, magari "migliori", oppure spunti di riflessione interessanti, come avvenuto per Chi^2. Lo stesso metodo di minimo Chi^2, alla luce di quanto si è detto, risulta degno di un approfondimento.

Infine i metodi di stima presentati nel paragrafo *7.3* andrebbero confrontati su un numero ancora maggiore di casi, eventualmente ricorrendo a simulazioni di matrici mal condizionate o con punteggi estremi.
Questi argomenti potrebbero quindi essere tenuti in considerazione per eventuali studi futuri.

Appendice A

Misura e integrazione

A.1 Misura

Si dice *misura (positiva)* su uno spazio misurabile $(X,\mathcal{M})$ una funzione $\mu: X \to [0,\infty]$ con la proprietà di additività numerabile (Rudin, 1974, pag.17).

Ossia, data una collezione $\{E_i\}$ di insiemi appartenenti a $\mathcal{M}$, allora:

$$\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i).$$

Si osservi che μ può assumere valore $+\infty$.

Si dice *misura complessa* su uno spazio misurabile $(X,\mathcal{M})$ una funzione complessa $\mu : X \to C$ con la proprietà di additività numerabile (Rudin, 1974, pag. 124). Ossia, dato l'insieme $E \in \mathcal{M}$ e una sua qualunque partizione $\{E_i\}$, $i=1,2,3,...$, μ è tale che:

$$\mu(E) = \mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i).$$

Come si vedrà in seguito, la precedente serie deve necessariamente convergere, differentemente da quanto richiesto per la misura positiva, che può anche assumere valore infinito.

Si cerca dunque una misura positiva λ che domini μ, cioè tale che $\lambda(E) \geq |\mu(E)|$, $\forall E \in \mathcal{M}$, e che assuma valore il più piccolo possibile.

Si tenga presente che $|z|$ indica la *norma* in C del numero complesso $z \in C$..

Dato che $|\mu(E)| \leq \sum_{i=1}^{\infty} |\mu(E_i)|$, $\forall \{E_i\}$, allora dovrà essere:

$$\lambda(E) \geq \sum_{i=1}^{\infty} |\mu(E_i)|, \quad \forall \{E_i\}.$$

Quindi $\lambda(E)$ sarà almeno uguale all'estremo superiore della precedente serie, preso rispetto ad ogni possibile partizione di E. Si definisce quindi la funzione $|\mu|$ su $\mathcal{M}$ con:

$$|\mu|(E) = \sup_{\{E_i\}} \sum_{i=1}^{\infty} |\mu(E_i)| .$$

Si dimostra che $|\mu|$, detta *variazione totale* di μ, è una misura (positiva). Inoltre $|\mu|$ domina μ ed è la più piccola rispetto ad ogni altra misura λ.

Un'ulteriore proprietà di $|\mu|$ è che $|\mu(X)| < \infty$. Allora, dato che $|\mu|(X) \geq |\mu|(E) \geq |\mu(E)|$, si può concludere che ogni misura complessa è limitata. Si osservi quindi che una misura positiva è una misura complessa, solo nel caso in cui essa sia finita, cioè $\mu(X) < \infty$. (Rudin, 1974, pag. 125 in avanti)

Una misura reale μ è un caso particolare di misura complessa (non può valere $\pm\infty$). La definizione infatti coincide con la precedente, se non per il fatto che $\mu : X \to \Re$. Indicando sempre con $|\mu|$ la variazione totale di μ, si considerino:

$$\mu^+ = \tfrac{1}{2}(|\mu| + \mu), \qquad \mu^- = \tfrac{1}{2}(|\mu| - \mu).$$

μ^+ e μ^-, dette rispettivamente variazione positiva e negativa di μ, sono entrambe misure (positive) su $(X,\mathcal{M})$.

Inoltre vale la *decomposizione di Jordan* (Rudin, 1974, pag.128):

$$\mu = \mu^+ - \mu^-, \qquad |\mu| = \mu^+ + \mu^- .$$

Si dimostra che le misure di variazione positiva e negativa sono concentrate su due insiemi disgiunti e complementari.

Teorema della decomposizione di Hahn

Data una misura reale μ sullo spazio misurabile $(X,\mathcal{M})$, esistono due insiemi disgiunti P e N tali che $P \cup N = X$ e tali che μ^+ e μ^- soddisfino:

$$\mu^+(E) = \mu(E \cap P), \qquad \mu^-(E) = \mu(E \cap N) .$$

(Rudin, 1974, pag. 134)

Inoltre, tra tutte le varie rappresentazioni di una misura reale μ come differenza tra due misure positive, la decomposizione di Jordan gode della proprietà di minimo.

Ossia, se $\mu=\lambda_1-\lambda_2$, dove λ_1 e λ_2 sono misure positive, allora $\lambda_1 \geq \mu^+$ e $\lambda_2 \geq \mu^-$.
(Rudin, 1974, pag 135)

A.2 Funzioni a variazione limitata

Data una funzione reale f, si definisce la funzione di *variazione totale* T_f con

$$T_f(x) = \sup_{\{x_i\}} \sum_{j=1}^{n} \left| f(x_i) - f(x_{i-1}) \right| .$$

Dove $\{x_i\}$ rappresenta una partizione delimitata dai punti $x_0, x_1, \ldots . x_n = x$, tali che

$$-\infty < x_0 < x_1 < \ldots . x_n = x.$$

T_f è una funzione monotona non decrescente e non negativa. Cioè:

$$0 \leq T_f(x) \leq T_f(y) \leq \infty \quad \text{se } x<y.$$

Quindi, se T_f è limitata, ammette limite finito per $x \to \infty$. In tal caso si dice che f è una funzione a *variazione limitata (BV)* e si indica con $V(f) = \lim_{x \to \infty} T_f(x)$ la sua *variazione totale*. (Rudin, 1974, pag. 171 in avanti)
Una funzione f a variazione limitata, continua da destra (convenzionalmente) e tale che

$\lim_{x \to -\infty} f(x) = 0$, è detta funzione a *variazione limitata normalizzata (NBV)*.

Si dimostra che è sempre possibile associare una funzione NBV ad una misura reale μ e, viceversa, ad ogni funzione NBV corrisponde una misura reale.

Teorema

- Se μ è una misura reale su $(\Re, \mathcal{B})$ e se $f(x) = \mu((-\infty, x)]$, $\forall x \in \Re$, allora f è NBV.
- Ad ogni $f \in$ NBV corrisponde un'unica misura reale su $(\Re, \mathcal{B})$ tale che: $f(x) = \mu((-\infty, x])$ e $T_f(x) = |\mu|((-\infty, x])$, $\forall x \in \Re$.
- Se $f(x) = \mu((-\infty, x])$, allora f è continua nei punti x in cui $\mu(\{x\})=0$. (Ossia i singoli punti x hanno misura diversa da 0 solo in corrispondenza ad eventuali "salti" di $f(x)$)

Nel caso in cui μ è una misura positiva, la funzione definita da $f(x) = \mu((-\infty, x])$, non decrescente, continua da destra, e tale che:

$$\lim_{x \to -\infty} f(x) = 0 \text{ e } \lim_{x \to \infty} f(x) = \mu(\Re), \qquad (f{:}\Re \to \Re)$$

è detta *distribuzione.* (Billingsley, 1986, pag.176)

Si dimostra che una funzione è a variazione limitata, se e solo se può essere espressa come differenza di due funzioni monotòne non decrescenti e limitate (Billigsley, 1986, pag. 435 e 436). Infatti, la differenza tra due funzioni non decrescenti e limitate è chiaramente a variazione limitata. Ma è vero anche il viceversa, si definiscano:

$$P_f(x) = \sup_{\{x_i\}} \sum_{j=1}^{n} \left(f(x_i) - f(x_{i-1})\right)^+ \text{ e } N_f(x) = \sup_{\{x_i\}} \sum_{j=1}^{n} \left(f(x_i) - f(x_{i-1})\right)^-,$$

dove f^+ e f^- sono tali che $f^+ - f^- = f$ e $f^+ + f^- = |f|$ (parte positiva e negativa di f). $P_f(x)$ e $N_f(x)$ sono rispettivamente la variazione positiva e negativa di *f*.

Si osservi che $P_f(x)$ e $N_f(x)$ sono non decrescenti e limitate, se *f* è BV. Allora potrà sempre scriversi:

$f(x) = P_f(x) - N_f(x)$. Si noti che le funzioni $P_f(x)$ e $N_f(x)$ sono le funzioni NBV associate rispettivamente alla variazione positiva e negativa di μ, cioè μ^+ e μ^-. Un caso particolare quindi dell'ultimo teorema è che è possibile associare in modo univoco una distribuzione NBV ad ogni misura positiva finita.

Tenendo conto della proprietà di minimo della decomposizione di Jordan, si osservi che, se esistono due distribuzioni *g*(*x*) e *h*(*x*) tali che *f*(*x*)=*g*(*x*)-*h*(*x*), allora si avrà $P_f(x) \le g(x)$ e $N_f(x) \le h(x)$.

Ricapitolando, ogni misura reale μ può essere determinata, in modo equivalente, dalla differenza tra due distribuzioni qualsiasi, il cui valore non può essere inferiore a quello delle distribuzioni di μ^+ e μ^-.

A.3 Integrale di Lebesgue-Stieltjes

Si consideri una misura (positiva) μ e la distribuzione $F(x) = \mu((-\infty, x])$, F:$\Re \to \Re$. Se φ è μ-misurabile, l'integrale di φ rispetto a μ può essere scritto equivalentemente come integrale di φ rispetto a F (notazione di Lebesgue-Stieltjes):

$$\int_A \varphi(x) dF(x), \ \forall \mathrm{A} \in B.$$

Si definiscono inoltre:

$$F(a^+) = \lim_{x \to a^+} F(x),$$

$$F(a^-) = \lim_{x \to a^-} F(x), \text{ dove } -\infty \leq a \leq \infty.$$

Si osservi che, per $-\infty \leq a \leq b \leq \infty$:

$F(b)$-$F(a)$=$\mu((a,b])$; $F(b)$-$F(a^-)$=$\mu([a,b])$; $F(b^-)$-$F(a)$=$\mu((a,b))$;$F(b^-)$-$F(\mathrm{a}^-)$=$\mu([a,b))$.

A.4 Formula di integrazione per parti

Siano F e G due distribuzioni e $H(x)$=$F(x)G(x)$. Allora, per $-\infty \leq a \leq b \leq \infty$:

$$\int_a^b F(y) dG(y) + \int_a^b G(x^-) dF(x) = H(b) - H(a) \quad \text{se } a,b \in \Re.$$

Dimostrazione

Sia D=$\{(x,y) \in \Re^2 : a<x \leq y \leq b\}$. Applicando il teorema di Fubini-Tonelli:

$$\iint_D dF(x) dG(y) = \int_a^b \int_a^b I_D(x,y) dF(x) dG(y) = \int_a^b dF(x) \int_a^b I_D(x,y) dG(y) =$$

$$= \int_a^b (G(b) - G(x^-)) dF(x) = \int_a^b dG(y) \int_a^b I_D(x,y) dF(x) = \int_a^b (F(y) - F(a)) dG(y).$$

Considerando che:

$\int_a^b F(a))dG(y) = F(a)(G(b) - G(a))$ e $\int_a^b G(b))dF(x) = G(b)(F(b) - F(a))$,

dall'equazione $\int_a^b (F(y) - F(a))dG(y) = \int_a^b (G(b) - G(x^-))dF(x)$ si ottiene facilmente la formula dell'integrazione per parti.

Si noti che, quando $b=\infty$ e\o $a=-\infty$, $H(b)$-$H(a)$ potrebbe non esistere, ed essere della forma ∞-∞.

In tal caso, la precedente formula rimane valida solo se $H(b)$-$H(a)$ esiste. Allora esiteranno anche i limiti:

$H(\infty) = \lim_{x \to \infty^+} H(x)$ e \o $H(-\infty) = \lim_{x \to -\infty^+} H(x)$. (Hoffmann-Jorgensen, 1994)

Un caso particolare si ottiene quando F e G non hanno punti di discontinuità in comune (Billingsley, 1986, pag. 239). In tal caso l'insieme $D' = \{(x,y) \in \Re^2 : a<x=y \le b\}$ ha misura nulla. Per questo motivo si può scrivere:

$$\int_a^b F(y)dG(y) + \int_a^b G(x)dF(x) = H(b) - H(a).$$

Ad esempio, se $F(x)=x$, distribuzione (continua) associata alla misura di Lebesgue, si ha che:

$$\int_a^b ydG(y) + \int_a^b G(x)dx = bG(b) - aG(a).$$

Il caso ancor più particolare è quando sia F che G sono assolutamente continue rispetto alla misura di Lebesgue, con densità f e g. Allora, poiché $f(x)=F'(x)$ e $g(x)=G'(x)$, la formula di integrazione per parti si può esprimere nella sua forma più nota:

$$\int_a^b F(x)G'(x)dx + \int_a^b G(x)F'(x)dx = H(b) - H(a).$$

A.5 Integrazione per sostituzione

Data la distribuzione F, ponendo $F(\infty)=\alpha$ e $F(-\infty)=\beta$, si definisce la funzione *inversa generalizzata* di F, con:

$$\bar{F}(y)=\inf\{x\in\Re: F(x)\geq y\} \quad \forall\alpha<y<\beta.$$

Si veda Bertoli-Barsotti (2001)

Essendo F non decrescente e delimitata da α e β (che possono anche essere $\pm\infty$), si osservi che $\bar{F}$ è non decrescente e finita su (α,β), $\bar{F}:(\alpha,\beta)\rightarrow\Re$. Inoltre, essendo F continua da destra, $\bar{F}$ è continua da sinistra. Valgono quindi le condizioni:

$$F(x)\geq y \Leftrightarrow x\geq \bar{F}(y);$$

$$F(\bar{F}(y)^{-}) \leq y \leq F(\bar{F}(y)).$$

Sia λ la misura di Lebesgue su (α,β) e si definisca:

$$\nu(B) = \lambda(\bar{F}^{-1}(B)), \text{ dove } B\in B(\Re).$$

Dato che $\bar{F}$ è misurabile, se $B\in\mathcal{B}(\Re)$ allora $\bar{F}^{-1}(B)\in\mathcal{B}((\alpha,\beta))$. Quindi $\lambda(\bar{F}^{-1}(B))$ è ben definita su $\mathcal{B}(\Re)$. Inoltre $\bar{F}^{-1}\{\cup_n B_n\}=\cup_n\{\bar{F}^{-1}(B_n)\}$ e $\bar{F}^{-1}(B_n)$ sono disgiunti in (α,β) se B_n sono disgiunti in $\Re$. Allora la proprietà di additività numerabile di λ implica quella di ν, che quindi è una misura su $\mathcal{B}(\Re)$.

Ponendo $B=(a,b]$ si ha, per le due condizioni ottenute precedentemente:

$$\bar{F}^{-1}(B)=\{y\in(\alpha,\beta): a<\bar{F}(y)\leq b\}=\{y\in(\alpha,\beta): F(a)<y\leq F(b)\}.$$

Allora $\nu((a,b])=F(b)-F(a)$. Questo permette, ponendo $x=\bar{F}(y)$, di scrivere:

$$\int_{\Re} h(x)dF(x) = \int_{\alpha}^{\beta} h(\bar{F}(y))dy,$$

dove $h:\Re\rightarrow\Re$ è una funzione $\mathcal{B}(\Re)$-misurabile. Questa espressione dà la formula di *integrazione per sostituzione* (Hoffmann-Jorgensen, 1994, pag. 205 e 206).

Se F è continua, si può ottenere una seconda formula di integrazione per sostituzione. Si ponga $h(x)=g(F(x))$. Essendo F continua, $y=F(\bar{F}(y))$ e,

conseguentemente, $h(\bar{F}(y))=g(y)$. Allora, per il risultato ottenuto precedentemente:

$$\int_{\Re} g(F(x))dF(x) = \int_{\alpha}^{\beta} g(y)dy \text{ , dove } g:[\alpha,\beta]\rightarrow\Re \text{ è } \mathcal{B}((\alpha,\beta))\text{-misurabile.}$$

(Hoffman-Jorgensen, 1994, pag. 205 e 206)

Ovviamente, come nell'integrazione per parti, se la misura associata a F è assolutamente continua rispetto alla misura di Lebesgue, con densità $f(x)=F'(x)$, si ha il seguente caso particolare:

$$\int_{\Re} g(F(x))F'(x)dx = \int_{\alpha}^{\beta} g(y)dy \,.$$

A.6 Teorema di convergenza monotòna

Sia μ una misura (positiva) su uno spazio $(X,\mathcal{M})$ e una $\{f_n\}$ una successione di funzioni μ-misurabili tale che:

- $0\leq f_1(x)\leq f_2(x)\leq\ldots\leq f_n(x)\leq\ldots$;
- $\lim_{n\rightarrow\infty} f_n(x) = f(x)$.

Allora $f\geq 0$ è μ-misurabile e $\lim_{n\rightarrow\infty}\int_X f_n(x)d\mu = \int_X \lim_{n\rightarrow\infty} f_n(x)d\mu = \int_X f(x)d\mu$.

(Rudin, 1974, pag. 22)

Appendice B

Stime basate su divergenze modificate: simulazioni

B.1 Stime MHPD

Nelle seguenti tabelle sono riportati medie e M.S.E. (men square error) degli stimatori MHPD per distribuzioni Binomiali con diversi valori di p e diverse ampiezze campionarie. Si è utilizzato il software *Mathematica 5.0* (Wolfram, 2003).

B.1.1)

Confronti tra stime MHPD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.5 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.50094 e M.S.E. pari a 0.00249. .Si osservi che la stima migliore si ha intorno al valore h=0.3.

h	*Media*	*M.S.E.*
1.0	0.50072	0.00308
0.9	0.50074	0.00299
0.8	0.50077	0.00291
0.7	0.5008	0.00283
0.6	0.50083	0.00276
0.5	0.50086	0.00270
0.4	0.50090	0.00266
0.3	0.50095	0.00264
0.2	0.50102	0.00267
0.1	0.50112	0.00275
0.0	0.50126	0.00294

B.1.2)

Confronti tra stime MHDP, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.3. In tal caso le stime di massima verosimiglianza presentano una media pari a 0.30003 e un M.S.E. pari a 0.00213. Si osservi che la stima migliore si ha intorno al valore h=0.3.

h	*Media*	*M.S.E.*
1.0	0.2924	0.002646
0.9	0.2933	0.002555
0.8	0.2943	0.002468
0.7	0.2953	0.002389
0.6	0.2965	0.002318
0.5	0.2979	0.002259
0.4	0.2993	0.002218
0.3	0.3010	0.002201
0.2	0.3030	0.002224
0.1	0.3052	0.002303
0.0	0.3079	0.002471

B.1.3)

Confronti tra stime MHPD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.1 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.1003 e M.S.E. pari a 0.000883. .Si osservi che la stima migliore si ha intorno al valore h=0.5.

h	*Media*	*M.S.E.*
1.0	0.0882	0.001026
0.9	0.0897	0.000996
0.8	0.0913	0.000969
0.7	0.0930	0.000947
0.6	0.0949	0.000932
0.5	0.0971	0.000924

0.4	0.0994	0.000931
0.3	0.1021	0.000960
0.2	0.1051	0.001018
0.1	0.1086	0.001116
0.0	0.1128	0.001282

B.1.4)

Confronti tra stime MHPD, per 1000 campioni di numerosità 20 da una distribuzione binomiale di parametri n=10 e p=0.3 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.30072 e M.S.E. pari a 0.0010225. .Si osservi che la stima migliore si ha intorno al valore h=0.1.

h	*Media*	*M.S.E.*
1.0	0.2949	0.001237
0.9	0.2954	0.001208
0.8	0.2959	0.001180
0.7	0.2965	0.001155
0.6	0.2971	0.001131
0.5	0.2976	0.001110
0.4	0.2983	0.001092
0.3	0.2989	0.001078
0.2	0.2996	0.001069
0.1	0.3004	0.001066
0.0	0.3012	0.001070

B.1.5)

Confronti tra stime MHPD, per 1000 campioni di numerosità 20 da una distribuzione binomiale di parametri n=10 e p=0.5. In tal caso le stime di massima verosimiglianza presentano una media pari a 0.50034 e M.S.E. pari a 0.00131487. .Si ottiene la stima migliore intorno ai valori h=0.1 e h=0.

h	*Media*	*M.S.E.*
1.0	0.500124	0.001596
0.9	0.500111	0.001564
0.8	0.500096	0.001532
0.7	0.50008	0.001501
0.6	0.500063	0.001473
0.5	0.500044	0.001446
0.4	0.500023	0.001422
0.3	0.5	0.001402
0.2	0.499974	0.001387
0.1	0.499946	0.001379
0.0	0.499915	0.001378

B.1.4)

Confronti tra stime MHPD, per 1000 campioni di numerosità 20 da una distribuzione binomiale di parametri n=10 e p=0.1 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.100645 e M.S.E. pari a 0.000449925. Si ottiene la stima migliore intorno ai valori h=0.3 e h=0.4.

h	*Media*	*M.S.E.*
1.0	0.09289	0.0005017
0.9	0.09357	0.0004925
0.8	0.09428	0.0004842
0.7	0.09502	0.0004768
0.6	0.09581	0.0004707
0.5	0.09664	0.0004662

0.4	0.09752	0.0004636
0.3	0.09845	0.0004633
0.2	0.09945	0.0004660
0.1	0.10052	0.0004724
0.0	0.10167	0.0004836

Dalle simulazioni svolte, non si riesce a trovare il valore *h* che garantisca una stima migliore. Può dipendere dal valore dal valore di *p* nella distribuzione e dall'ampiezza campionaria utilizzata.

B.2 Stime MCDD

Nelle seguenti tabelle sono riportati medie e M.S.E. degli stimatori basati sulla divergenza di Chi^2 depenalizzata, per distribuzioni Binomiali con diversi valori di *p* e diverse ampiezze campionarie. Si è utilizzato il software *Mathematica 5.0* (Wolfram, 2003).

B.2.1)

Confronti tra stime MCDD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.5 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.50164 e M.S.E. pari a 0.0024662. Si ottiene la stima migliore intorno al valore h=0.4.

h	*Media*	*M.S.E.*
1.0	0.502528	0.002677
0.9	0.502544	0.002668
0.8	0.502562	0.002659
0.7	0.502585	0.002648
0.6	0.502613	0.002638
0.5	0.502649	0.002628
0.4	0.502696	0.002621
0.3	0.502762	0.002625
0.2	0.502861	0.002666
0.1	0.503039	0.002871
0.0	0.505415	0.008356

B.2.2)

Confronti tra stime MCDD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.3 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.30127 e M.S.E. pari a 0.0020107. Si ottiene la stima migliore intorno al valore h=0.4.

h	*Media*	*M.S.E.*
1.0	0.30626	0.002182
0.9	0.30599	0.002173
0.8	0.30566	0.002163
0.7	0.30526	0.002153
0.6	0.30475	0.002142
0.5	0.30408	0.002132
0.4	0.30317	0.002126
0.3	0.30185	0.002133
0.2	0.29972	0.002182
0.1	0.29561	0.002416
0.0	0.22075	0.025212

B.2.3)

Confronti tra stime MCDD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.1 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.0991 e M.S.E. pari a 0.0009168. Si ottiene la stima migliore intorno ai valori h=0.4, h=0.3.

h	*Media*	*M.S.E.*
1.0	0.109034	0.0011328
0.9	0.108392	0.0011176
0.8	0.107621	0.0011007
0.7	0.106678	0.0010820
0.6	0.105495	0.0010614
0.5	0.103968	0.0010397

0.4	0.101917	0.0010187
0.3	0.098998	0.0010043
0.2	0.094446	0.0010163
0.1	0.085909	0.0011450
0.0	0.002633	0.0097490

B.2.4)

Confronti tra stime MCDD, per 1000 campioni di numerosità 8 da una distribuzione binomiale di parametri n=10 e p=0.5 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.499325 e M.S.E. pari a 0.00335156. Si ottiene la stima migliore intorno ai valore h=0.4, h=0.3.

h	*Media*	*M.S.E.*
1.0	0.499808	0.0038221
0.9	0.499772	0.0038036
0.8	0.499731	0.0037835
0.7	0.499684	0.0037621
0.6	0.499628	0.0037401
0.5	0.499563	0.0037191
0.4	0.499487	0.0037029
0.3	0.499397	0.0037020
0.2	0.499293	0.0037494
0.1	0.499176	0.0039912
0.0	0.500315	0.0106346

B.2.5)

Confronti tra stime MCDD, per 1000 campioni di numerosità 8 da una distribuzione binomiale di parametri n=10 e p=0.3 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.298237 e M.S.E. pari a 0.00263422. Si ottiene la stima migliore intorno al valore h=0.8, 0.7.

h	*Media*	*M.S.E.*
1.0	0.303886	0.00271395
0.9	0.303471	0.00271026
0.8	0.302978	0.00270844
0.7	0.302384	0.00270992
0.6	0.30165	0.00271717
0.5	0.300722	0.00273482
0.4	0.299504	0.0027721
0.3	0.297824	0.00284968
0.2	0.295305	0.00302332
0.1	0.290832	0.00350644
0.0	0.229254	0.0226492

B.2.6)

Confronti tra stime MCDD, per 1000 campioni di numerosità 8 da una distribuzione binomiale di parametri n=10 e p=0.1 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.102113 e M.S.E. pari a 0.00128703. Si ottiene la stima migliore intorno ai valori h=0.4, h=0.3.

h	*Media*	*M.S.E.*
1.0	0.112417	0.00160872
0.9	0.111327	0.0015783
0.8	0.110048	0.001546
0.7	0.108525	0.00151221
0.6	0.106679	0.00147785
0.5	0.104389	0.00144492

0.4	0.101454	0.00141787
0.3	0.097510	0.0014074
0.2	0.091779	0.0014440
0.1	0.081983	0.00165123
0.0	0.006536	0.00964285

B.3 Stime MALD

Nelle seguenti tabelle sono riportati medie e M.S.E. degli stimatori basati sulla divergenza di verosimiglianza "amplificata", per distribuzioni Binomiali con diversi valori di p e diverse ampiezze campionarie. Si è utilizzato il software *Mathematica 5.0* (Wolfram, 2003).

B.3.1)

Confronti tra stime MALD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.5 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.499 e M.S.E. pari a 0.0027992. Si ottengono le stime migliori intorno al valore h=0.4.

h	*Media*	*M.S.E.*
1.0	0.4996	0.003285
0.9	0.4995	0.003045
0.8	0.4994	0.002910
0.7	0.49939	0.002828
0.6	0.49932	0.002780
0.5	0.4992	0.002754
0.4	0.49919	0.002744
0.3	0.49912	0.002746
0.2	0.49908	0.002757
0.1	0.49904	0.002775

B.3.2)

Confronti tra stime MALD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.4 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.40107 e M.S.E. pari a 0.0022711. .Si ottengono le stime migliori intorno al valore h=0.1.

h	*Media*	*M.S.E.*
1.0	0.4063	0.003167
0.9	0.4055	0.002854
0.8	0.4048	0.002655
0.7	0.4042	0.002520
0.6	0.4036	0.002426
0.5	0.4031	0.002361
0.4	0.4026	0.002318
0.3	0.4022	0.002290
0.2	0.4018	0.002274
0.1	0.4014	0.002269

B.3.3)

Confronti tra stime MALD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.3 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.3010 e M.S.E. pari a 0.0021335. Si ottengono le stime migliori intorno al valore h=0.2.

h	*Media*	*M.S.E.*
1.0	0.3086	0.002780
0.9	0.3066	0.002512
0.8	0.3053	0.002351
0.7	0.3040	0.002244
0.6	0.3029	0.002173
0.5	0.3018	0.002127

0.4	0.3009	0.002099
0.3	0.3000	0.002084
0.2	0.2992	0.002080
0.1	0.2984	0.002084

B.3.4)

Confronti tra stime MALD, per 1000 campioni di numerosità 10 da una distribuzione binomiale di parametri n=10 e p=0.1 In tal caso le stime di massima verosimiglianza presentano una media pari a 0.09949 e M.S.E. pari a 0.0009051. Si ottengono le stime migliori intorno al valore h=0.1.

h	*Media*	*M.S.E.*
1.0	0.1198	0.0019941
0.9	0.1157	0.0014834
0.8	0.11292	0.0012958
0.7	0.11054	0.0011768
0.6	0.10846	0.0010933
0.5	0.10662	0.0010328
0.4	0.10495	0.0009884
0.3	0.10342	0.0009559
0.2	0.10201	0.0009324
0.1	0.10070	0.0009159

Appendice C

Modello di Rasch: soluzioni ai problemi di stima

C.1) Confronti tra i due metodi studiati (α e β) applicati a tabelle 10×6 mal-condizionate, con strutture interne più varie possibili, create appositamente. Nel primo esempio sono riportai i valori delle stime, nei successivi sono riportati solamente i valori dell'indice di adattamento, dato dalla somma dei residi standardizzati al quadrato $\sum_{i=1}^{n}\sum_{J=1}^{m}(z_{ij})^2$ (paragrafo *7.1*). In questi casi, ovviamente, i confronti tra i due metodi vanno fatti a parità di condizioni (ossia tra stime ottenute con la stessa misura di divergenza: K.L. o Chi^2). Nel secondo esempio si è anche in presenza di un punteggio estremo: in tal caso entrambi i metodi forniscono ugualmente stime accettabili. Si è utilizzato il software *Mathematica 5.0* (Wolfram, 2003)

C.1.1)

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Matrice dati

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 0.5 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Matrice dati modificata (metodo-α)

Stime

θ_i	*Chi^2* α*=0.5*	*Chi^2* β*=0.01*	*K.L.* α*=0.5*	*K.L.* β*=0.01*
θ_1	1.40726	2.95526	2.61062	3.75018

θ_2	2.11293	3.02747	3.41492	3.75018
θ_3	1.21672	2.04965	1.83404	2.15047
θ_4	1.17599	1.9629	1.83404	2.15047
θ_5	-0.0368	-0.0647	-0.1207	-0.1346
θ_6	-1.2526	-2.0858	-1.9435	-2.2726
θ_7	-1.3488	-2.1897	-1.9435	-2.2726
θ_8	-1.2584	-2.1897	-1.9435	-2.2726
θ_9	-1.9342	-2.8887	-3.3181	-3.6532
θ_{10}	-2.0309	-2.8887	-3.3181	-3.6532

β_i	*Chi^2* $\alpha=0.5$	*Chi^2* $\beta=0.01$	*K.L.* $\alpha=0.5$	*K.L.* $\beta=0.01$
β_1	2.11902	3.22371	3.6755	4.12265
β_2	1.55732	2.50853	2.52894	2.95308
β_3	1.04446	1.80619	1.56146	1.87927
β_4	-1.8592	-2.6735	-3.0052	-3.2610
β_5	-1.4079	-2.1913	-2.1789	-2.4328
β_6	-1.4536	-2.6735	-2.5816	-3.2610

Adattamento

Metodo	$\alpha=0.5$	$\beta=0.01$
Chi^2	0.4575	0.4054
K.L.	0.4343	0.4216

C.1.2)

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*
Chi^2	0.4742	0.4054
K.L.	0.4287	0.4171

C.1.3)

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*
Chi^2	0.5021	0.4649
K.L.	0.4927	0.4900

C.1.4)

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*
Chi^2	0.5022	0.4649
K.L.	0.4942	0.4905

C.1.5)

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*
Chi^2	0.5034	0.4698
K.L.	0.5133	0.4994

C.1.6)

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*	*β=0.05*
Chi^2	0.491451	0.438519	0.467874
K.L.	0.50401	0.511431	0.489649

Da quanto osservato nei precedenti esempi, l'errore (indice di adattamento) del metodo β è inferiore a quello del metodo α, sia utilizzando Chi^2 che Kullback-Leibler. Pare quindi che con il metodo β si ottengano risultati migliori.

C.2)

Confronti tra i due metodi applicati a tabelle 15×10 mal-condizionate, con strutture interne più varie possibili. Anche variando le dimensioni delle tabelle si può confermare quanto osservato precedentemente. In certi casi, aumentando leggermente il valore di β, i risultati possono migliorare ulteriormente, se si utilizza Kullback-Leibler. Tuttavia è consigliabile utilizzare valori di β piuttosto piccoli, per alterare il meno possibile la distribuzione "vera".

C.2.1)

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*
Chi^2	0.390815	0.353838
K.L.	0.372199	0.370202

C.2.2)

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*	*β=0.05*
Chi^2	0.395936	0.35694	0.397452
K.L.	0.402531	0.399727	0.387483

C.2.3)

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Metodo	*α=0.5*	*β=0.01*	*β=0.05*
Chi^2	0.440424	0.40486	0.437585
K.L.	0.466399	0.462586	0.439918

C.3)

Dal seguente esempio si osserva che il metodo-β non altera eccessivamente i valori delle stime, quando la tabella non è mal-condizionata. I valori tendono a diminuire (in valore assoluto) al crescere di β, sia con Kullback-Leibler che con Chi^2.

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Nelle seguenti tabelle sono riportate le stime di minimo Chi^2 e le stime ottenute combinando il metodo-β (β*=0.01 e* β*=0.05*) con Chi^2.

θ_i	*Chi^2*	*Chi^2* β*=0.01*	*Chi^2* α*=0.5*
θ_1	0.916889	0.916441	0.904992

θ_2	0.410641	0.410499	0.406834
θ_3	-0.097449	-0.097300	-0.093562
θ_4	0.554393	0.554009	0.544248
θ_5	0.53093	0.53058	0.52169
θ_6	-0.541752	-0.541378	-0.531902
θ_7	-0.045587	-0.045501	-0.043366
θ_8	-0.534399	-0.534046	-0.525074
θ_9	-0.486797	-0.486525	-0.479587
θ_{10}	-0.916998	-0.916549	-0.905075

β_i	*Chi^2*	*Chi^2* $\beta=0.01$	*Chi^2* $\alpha=0.5$
β_1	-0.925794	-0.925234	0.910961
β_2	-0.278515	-0.278359	-0.274391
β_3	-0.340205	-0.339961	-0.333782
β_4	0.555644	0.555396	0.549009
β_5	-0.010736	-0.010726	-0.010447
β_6	0.999607	0.998884	0.980573

Nelle seguenti tabelle sono riportate le stime di minimo K.L. e le stime ottenute combinando il metodo-β ($\beta=0.01$ *e* $\beta=0.05$) con K.L..

θ_i	*K.L.*	*K.L.* $\beta=0.01$	*K.L.* $\alpha=0.5$
θ_1	2.00527	1.92256	1.63427
θ_2	0.879942	0.848583	0.737102
θ_3	-0.013410	-0.011933	-0.007625
θ_4	0.879942	0.848583	0.737102

θ_5	0.879942	0.848583	0.737102
θ_6	-0.893129	-0.861002	-0.746589
θ_7	-0.013410	-0.0119336	-0.007625
θ_8	-0.893129	-0.861002	-0.746589
θ_9	-0.893129	-0.861002	-0.746589
θ_{10}	-1.99197	-1.91177	-1.63038

β_i	*K.L.*	*K.L.* $\beta=0.01$	*K.L.* $\alpha=0.5$
β_1	-1.73381	-1.66639	-1.42861
β_2	-0.525064	-0.506441	-0.44021
β_3	-0.525064	-0.506441	-0.44021
β_4	1.06754	1.0293	0.893123
β_5	-0.006913	-0.006344	-0.004557
β_6	1.72331	1.65632	1.42047

Bibliografia

Adimari G., Ventura L. (2001) *Robust inference for generalized linear models with appplication to logistic regression*, Statistics & Probability letters 55, 413-419.

Ali S. M., Silvey S. D. (1965) *A general class of coefficients of divergence of one distribution from another*, University of Manchester, 131-142

Anatolyev Stanislav, Grigory Kosenok (2005) *An alternative to maximum likelihood based on spacings*, Cambridge university press, 21 (2): 472–476.

Anwar M., Hussain S., Pecari J. (2009) *Some Inequalities for Csiszàr-Divergence Measures*, Int. Journal of Math. Analysis, 3, no. 26, 1295 – 1304.

Bai Z. D., Radhakrishna Rao C., Wu Y. (1992) *M-estimation of multivariate linear regression parameter under a convex discrepancy function*, Statistica Sinica 2, 237-254.

Basu A., Harris I., *Robust predictive distributions for exponential families*, Biometrika, 81, 4, 790-794.

Basu A., Basu S., Chaudhuri G. (1997) *Robust minim divergence procedures for count data models*, The Indian Journal of statistics, 59, B, Pt. 1, 11-27.

Basu A, Harris I., Basu S. (1997) *Minimum distance estimation: the approach using density –based distances*, Handbook of Statistics, 15, 21-48.

Basu A., Lindsay B. G. (1994) *Minimum disparity estimation for continuous models: efficency, distributions and robusteness*, Ann. Inst. Statist. Math. 46, 683-705.

Barlow R. E., Marshall A. W., Proschan F. (1969) *Some inequalities for starshaped and convex functions*, Pacific Journal of Mathematics 29, 19-42.

Beran R. (1977) *Minimum Hellinger distance estimates for parametric models*, The Annals of Statistics, 5(3), 445-463.

Bertoli-Barsotti L. (2001) *Maggiorazione e Schur-convessità: generalizzazioni e applicazioni statistiche*

Bertoli-Barsotti L. (2005) *On the existence and uniqeness of JML estimates for the partial credit model*, Psychometrika, vol. 70, No. 3, 517-531.

Berkson J. (1980) *Minimum Chi-square, not maximum likelihood!*, The annals of Statistics vol.8, n. 3, 457-487.

Billingsley P. (1986) *Probability and measure, 2nd edition*, Wiley and sons.

Borovkov A.A. (1998) *Mathematical statistics*, Gordon and Breach science publishers.

Borovkov A.A. (1998) *Probability theory*, Gordon and Breach science publishers.

Brunk H. D. (1955) *On an inequality for convex function,* Proc. Amer. Math. Soc. 7, 817-824.

Burkill H. *A note on rearrangements of functions*, Amer. Math. Monthly 71, 887-888.

Cheng R. C., Amin N. A. (1982) *Estimating parameters in continuous univariate distributions with a shifted origin* J.R. Statistic. Soc. B 45, No 3, 394-403.

Cheng R. C., Iles T.C. (1987) *Corrected maximum likelihood in non-regular problems*, J.R. Statistic. Soc. B 49, No 1, 95-101.

Cressie N., Pardo L. (2002) *Model checking loglinear models using ϕ-divergences and MLEs*, Journal of statistical planning and inference 103, 437-453.

Cressie N., Read T.C (1984) *Multinomial Goodness-of-fit-tests,* J. R. Statist. Soc. B, 46, No. 3, 440-464.

Csiszàr I., Shields P. (2004) *Information theory and statistics: a tutorial*, Foundations and Trends in Communications and Information Theory, 1, No. 4, 417–528.

Fishburn P. C. (1980) *Stochastic dominance and moments of distributions*, Math, Oper. Res. 5, 94-100.

Giampaglia (2008) *Il modello di Rasch nella ricerca sociale*, ed. Liguori.

Gosh K., Rao Jammalamadaka S. (2001) *A general estimation method using spacings*, Journal of statistical planning and inference 93, 71-82.

Hampel F. (1971) *A general qualitative definition of robustness*, the Annals of Mathematical Statistics 42, No. 6, 1887-1896.

Hampel F. (1974) *The influence curve and its role in robust estimation*, Journal of American Statiisctical Association 346 (Vol. 69), 383-393.

Hardy G. H., Littlewood J. E., Pòlya G. (1929) *Some simple inequalities satisfied by convex functions*, Messenger of Math. 58, 145-152.

Hardy G. H., Littlewood J. E., Pòlya G. (1934) *Inequalities*, Cambridge.

Hoffmann-Jorgensen J. (1994) *Probability with a view toward statistics,* Chapman and Hall.

Hogg R.V, McKean J.W., Craig A.T. (2005) *Introduction to mathematical statistics, 6th edition,* Pearson.

Huber, P. J. (2004) *Robust Statistics*, Wiley.

Kanji H.(1983) *On the use of Minimum Chi-square estimation*, The Statistician 32, 379-394.

Karamata J. (1932) *Sur une inègalitè relative aux functions convexes*, Publ. Math. Univ. Belgrade 1, 145-148.

Karlin S., Novikoff A. (1963) *Generalized convex functions*, Pacific J. Math. 13, 173-180.

Kiefer J., Wolfowitz J. (1956) *Consistency of the maximim likelihood estimator in the presence of infinitely many incidental parameters,* Ann. Math. Statist., 27, 887-906.

Kradelburg Z., Dukic D., Lukic M., Matic I. (2005) *Inequalities of Karamata, Schur and Murihead, and some applications*, The Teaching of Mathematics, Vol. VIII, 1, 31-45.

Kullback S., Leibler R.A. (1951) *On information and sufficiency* The Annals of Mathematical Statistics 22 (1), 79–86.

Linacre J. M. (2004) *Rasch model estimation: further topics* Journal of applied measurement 5(1), 95-110.

Linacre J. M. (2009) WINSTEPS®, *Rasch measurement computer program*, Beaverton, Oregon: Winsteps.com.

Lindsay B. (1983) *The geometry of mixture likelihoods: a general theory*, The Annals of Statistics, Vol. 11, No. 1 , 86-94.

Lindsay B. (1983) *The geometry of mixture likelihoods, part II: the exponential family*, The Annals of Statistics, Vol. 11, No. 3 , 783-792.

Lindsay B. (1994) *Efficency versus robustness: the case for minimm Hellinger distance and related methods*, The Annals of Statistics, 22(2), 1081-1114.

Marshall A. W., Olkin I. (1979) *Inequalities: theory of majorizations and their applications,* Academic press, New York.

Marshall A. W., Proschan F. (1970) *Mean life of series and parallel systems*, J. Appl. Prob. 7, 165-174.

Nakamura T. (1991) *Existence of maximum likelihood estimates for interval-censored data from some three parameter models with a shifted origin*, J.R. Statist. Soc. B 53, No 1, 211-220.

Neyman J. (1949) *Contribution to the theory of* χ^2 *test*, Proceedings of first Berkley Symposium, 239-273.

Parr W., Schcant W. (1980) *Minimum distance and robust estimation*, Journal of the American Statistical Association, 75, No 371, 616-624.

Ranneby B. (1984) *The maximum spacing method. An estimation method related to the maximum likelihood method*, Scandinavian journal of Statistics 11 No 2, 93-112.

Ranneby B., Ekstrom M. (1997) *Maximum spacing estimates based on different metrics.*

Read T.R.C., Cressie N.A.C (1988) *Goodness-of-fit statistics for discrete multivariate data*, Springer series in Statistics.

Reiersol O. (1961) *Linear and non-linear multiple comparisons in logit analysis*, Biometrika, 48, No 3 e 4, 359-365.

Roberts A. W., Varberg D. (1973) *Convex functions*, Academic press Inc.

Rudin W.(1974) *Real and complex Analysis, 2nd edition,* McGraw-Hill.

Sarkar S., Basu A. (1995) *On disparity based robust tests for two discrete populations*, The Indian Journal of Statistics, 57, B, Pt. 3, 353-364.

Shao J. (2003) *Mathematical statistics*, *2nd edition*, Springer texts in Statistics.

Shao Y. (2001) *Consistency of the maximum product of spacings method and estimation of a unimodal distribution*, Statistica sinica 11, 1124-1140.

Simpson D. (1987) *Minimum Hellinger estimation for the analysis of count data*, Journal of the American Statistical Association, 82, No 399, 802-807.

Taneja I.J., Kumar P. (2004) *Relative information of type s, Csizar's f-divergence, and information inequalities*, Information Sciences, 166, 105-125.

Wolfowitz J. (1954) *Estimation by the minimum distance method in nonparametric stochastic difference equations*, Ann. Math . Stat., 25, No.2, 203-217.

Wolfowitz J. (1957) *The minimum distance method*, Ann. Math . Stat., 28, No.1, 75-88.

Wolfram Research (2003), Inc., *Mathematica*, *Version 5.0*, Champaign, IL.

Zeger S., Liang K., Albert P.S. (1988) *Models for Longitudinal Data: A Generalized Estimating Equation Approach,* Biometrics, 44, No. 4., 1049-1060.