ROMA TRE
UNIVERSITÀ DEGLI STUDI

FACOLTÀ DI ECONOMIA

Scuola Dottorale in Economia e Metodi Quantitativi
Dottorato in Metodi Statistici per l'Economia e l'Impresa
XXI ciclo

Le Verosimiglianze Empiriche nel Ricampionamento Bayesiano: Teoria e Indirizzi Applicativi

Relatore: Ch.mo Prof. Brunero Liseo
Coordinatrice: Ch.ma Prof.ssa Julia Mortera

Tesi di:
Flavio Foschi

Anno Accademico 2007/2008

Indice

1 Introduzione **1**

1.1 Ambito di Riferimento 1

1.2 Piano della Tesi 4

2 Verosimiglianze e informazione incompleta **5**

2.1 GMM 5

2.2 Densità Aumentate 8

2.3 Verosimiglianze Semiparametriche 9

2.4 Stime di massima entropia 11

2.5 Verosimiglianze empiriche 14

2.5.1 Misure di divergenza 15

2.5.2 GEL e MD 16

2.6 ETEL 18

2.7 EL in Campo Bayesiano: il Teorema di Sanov 20

3 Ricampionamento Bayesiano **23**

3.1 Il Processo di Dirichlet e il Bootstrap Bayesiano 23

3.2 Distribuzioni Base Vincolate 26

3.2.1 Selezione dei Parametri 31

3.2.2 Un Esempio in Ottica Classica 38

3.2.3 Smoothing del Bootstrap Bayesiano 42

3.3 Impostazione Completamente Predittiva 44

3.4 Teorema di Rappresentazione e MM 46

3.4.1 Ricampionamento Semiparametrico 50

3.4.2 Robustezza delle Stime 51

3.5 Dati Dipendenti 53

4 Applicazioni **57**

4.1 Un VECM per la Domanda di Moneta 57

4.2 Dati Sintetici per la Tutela della Privacy 68

5 Conclusioni **75**

A **84**

A.1 Espansioni Asintotiche 84
A.2 Distribuzione della Discrepanza 87
A.3 Estensione della Consistenza in L-Divergenza 88
A.4 MM e Approssimazione di Sella 91
A.5 Esempio 94

B **97**

B.1 Definizione di Alcune Funzioni 97
B.2 Un Programma 108

Bibliografia **112**

Elenco delle figure

3.1 Densità e relative approssimazioni 29
3.2 Confronto tra stime di una densità 38
3.3 Errori in valore assoluto 39
3.4 Soglie teoriche ed effettive 40
3.5 Estremi superiori per $|\delta_i|$ 41

4.1 Traiettorie da 5000 simulazioni 58
4.2 Medie delle traiettorie 58
4.3 Correlogrammi dei livelli 60
4.4 Correlogrammi delle differenze prime 61
4.5 Test di cointegrazione del massimo autovalore 62
4.6 Test sulle restrizioni per β 64
4.7 Coefficienti e percentili della 1ªequazione del VECM 66
4.8 Coefficienti e percentili della 2ªequazione del VECM 66
4.9 Coefficienti e percentili della 3ªequazione del VECM 67
4.10 Coefficienti e percentili della 4ªequazione del VECM 67
4.11 Intervalli h.p.d. per i momenti misti 71
4.12 I.B.F. 06 (fonte: Banca d'Italia), campione dei redditi da fabbricati . . 74

Elenco delle tabelle

3.1 Quartili della Gamma(20,1) 27
3.2 Stime dei quartili al variare di k 28
3.3 Ampiezze degli intervalli h.p.d. per $\alpha = 0.05$ 28
3.4 Copertura degli intervalli h.p.d. per $\alpha = 0.05$ 28
3.5 Base GEL: stime dei quartili 29
3.6 Base GEL: ampiezze degli intervalli h.p.d. per $\alpha = 0.05$ 30
3.7 Base GEL: coperture degli intervalli h.p.d. per $\alpha = 0.05$ 30
3.8 Base GEL da momenti stimati: stime dei quartili 42
3.9 Base GEL da momenti stimati: ampiezze degli IC h.p.d. per $\alpha = 0.05$ 43
3.10 Base GEL da momenti stimati: coperture degli IC h.p.d. per $\alpha = 0.05$ 43
3.11 Impostazione predittiva: stime dei quartili 50
3.12 Impostazione predittiva: ampiezze degli IC h.p.d. per $\alpha = 0.05$ 50
3.13 Impostazione predittiva: coperture degli IC h.p.d. per $\alpha = 0.05$ 51

4.1 Test del massimo autovalore 61
4.2 Coefficienti dei livelli 64
4.3 Coefficienti delle differenze prime 65
4.4 IC h.p.d. per i β_j, $\alpha = 0.05$ 65
4.5 Dati originali: indici di posizione 69
4.6 Dati originali: indici di dimensione, dispersione, forma 70
4.7 Dati originali: correlazioni 70
4.8 Sintesi per i coefficienti di correlazione 70
4.9 Sintesi per gli indici di posizione 72
4.10 Sintesi per gli indici di dimensione, dispersione, forma 73

A.1 Distribuzione generatrice dei dati 94
A.2 Momenti della distribuzione generatrice dei dati 94
A.3 Momenti delle distribuzioni in W 95
A.4 Distribuzione a posteriori e L-divergenze 96

A.5 Distribuzione di massima entropia 96
A.6 L-divergenza rispetto alla distribuzione di massima entropia 96

Capitolo 1

Introduzione

1.1 Ambito di Riferimento

In prima approssimazione, la statistica non parametrica si prefigge lo studio dei fenomeni oggetto d'interesse senza fare ricorso ad ipotesi sulla legge che genera i dati. Se si ritiene che solo l'evidenza debba essere fonte di discernimento, l'uso del teorema di Bayes sembra, a prima vista, una contraddizione. La statistica bayesiana oggettiva, anche in ambito parametrico, si propone di conservare rigore formale e chiarezza interpretativa offerti dal teorema pur giungendo a conclusioni consistenti rispetto all'approccio classico, caratterizzato dall'assumere che i parametri siano realizzazioni di distribuzioni degeneri, con masse di probabilità concentrate in un solo punto. Estendendo l'idea di priori vaga dal parametro di una distribuzione ad una legge di probabilità è possibile compiere il passaggio logico che conduce alla statistica non parametrica bayesiana. Ciò che cambia è la dimensione del problema da trattare. La stima non riguarda più la distribuzione di un parametro: i "punti" del supporto sono costituiti da leggi di probabilità e l'implementazione di strategie di stima è meno arduo quando siano introdotte restrizione sullo spazio delle leggi da probabilizzare.

Tra le molteplici proposte apparse nella letteratura sull'argomento, quelle riguardanti i processi di Dirichlet (Ferguson, 1973) hanno consolidato nel tempo una crescente popolarità, mentre in anni recenti si è iniziato ad indagare il possibile uso di verosimiglianze costruite sulla base di insiemi incompleti di informazioni (Kim 2002; Schennach 2005). L'oggetto d'interesse solitamente non è la legge che genera i dati ma la posteriori di una grandezza significativa per il ricercatore e l'ottenimento di quest'ultima segue percorsi diversi nei due casi.

- ▷ Facendo uso dei processi di Dirichlet per modellare la legge di probabilità sui dati, la priori non contiene esplicitamente informazioni sul parametro e la po-

steriori per quest'ultimo viene ricavate empiricamente da stime ripetute, ognuna delle quali espressione di una realizzazione della legge sui dati; per tale aspetto le stime non parametriche ottenute sono idealmente ascrivibili al novero delle tecniche di ricampionamento, circostanza singolare dal momento che il concetto di prove ripetute è estraneo al principio di verosimiglianza.

▷ Le verosimiglianze definite da condizioni di ortogonalità esprimono il tipo di conoscenza che il ricercatore intende scontare nella definizione della legge per i dati e la posteriori per il parametro beneficia di una costruzione formalmente coerente al teorema di Bayes. L'uso delle condizioni rende evidente i nessi con il metodo generalizzato dei momenti (Hansen, 1982) che, fondato sulla minimizzazione di una funzione di danno, prescinde dalla esplicitazione di una verosimiglianza.

I processi di Dirichlet permettono di ottenere delle posteriori il cui valore atteso è mistura delle frequenze relative sulle intensità osservate e di una distribuzione iniziale; se al divergere della numerosità campionaria il peso della seconda tende a zero le proprietà di consistenza sono assicurate. L'utilizzo delle verosimiglianze empiriche in ottica bayesiana ha alimentato un ampio dibattito ed il problema della consistenza delle stime, dapprima affrontato sul piano empirico (Lazar, 2003), solo recentemente ha ricevuto una parziale sistemazione teorica (Grendar e Judge, 2007) in riferimento alle verosimiglianze di Owen (1988).

Il bootstrap di Rubin (1981) rappresenta idealmente il gold standard nel campo della statistica bayesiana non parametrica e molta parte della ricerca in questo campo ha preso le mosse dagli elementi di criticità ad esso connaturati. Può essere visto come caso particolare del processo di Dirichlet ovvero come suo limite ovvero essere autonomamente derivato da un modello coniugato per una verosimiglianza multinomiale quando si usi la priori di Haldane ossia una Dirichlet con parametri identicamente nulli. La circostanza che risultino pari a zero le probabilità a posteriori per le intensità assenti dal campione lascia intuire gli stretti legami con il bootstrap di Efron (1979). Tuttavia una importante diversità li distingue: mentre il primo pondera le osservazioni con realizzazioni di un processo aleatorio, il secondo estrae campioni dai dati con qualche prezzo in termini di efficienza. Pur nella consapevolezza dell'abuso terminologico, in entrambi i casi, per brevità, è conveniente usare il termine ricampionamento. Le proposte di Muliere e Secchi (1994) ed Ericson (1969) per il superamento dell'assunzione che i valori osservati esauriscano le modalità del carattere si sostanziano nell'uso di priori informative per il processo di Dirichlet. In questo modo, attribuendo probabilità non nulle ad intensità non rilevate, si può dare conto

di massimi e minimi diversi da quelli campionari e della ragionevole attesa che alla prossimità rispetto ai valori osservati corrisponda quella delle probabilità; la natura non parametrica dell'approccio viene sostanzialmente conservata se il peso riservato alla priori è un infinitesimo della numerosità campionaria o se la priori è vaga. I benefici ai fini delle posteriori per i parametri d'interesse dovrebbero tradursi in intervali di credibilità più corti a causa dell'implicita regolarizzazione del bootstrap di Rubin. L'approccio di Schennach persegue, con strumenti diversi, le medesime finalità, ma ai fini dell'implementazione è necessario poter esplicitare le condizioni di ortogonalità ossia la relazione (a volte molto complessa) tra osservazioni e parametro.

Entrambe le direttrici di ricerca hanno, a ben vedere, elementi di contiguità con un diverso punto di vista dal quale la statistica bayesiana può essere apprezzata, dovuto a de Finetti; il teorema di rappresentazione (de Finetti, 1937) fornisce un'alternativa all'ipotesi di indipendenza e somiglianza delle unità statistiche incluse nel campione introducendo il concetto di scambiabilità; essa concerne l'indipendenza e somiglianza condizionata ad un parametro la cui legge di probabilità è indotta dalla legge di probabilità sulle osservazioni. L'importanza del teorema riguarda la possibilità di giustificare secondo una costruzione bayesiana molti risultati della statistica classica. Un aspetto rilevante concerne il proposito di fondare le inferenze soltanto su quantità che siano realmente osservabili, da cui la necessità di postulare che la legge sui parametri sia implicitamente definita dalla legge sulle osservazioni. L'esigenza nasce ammettendo che sia lecito assegnare probabilità solo ad eventi dei quali si possa affermare la verità o falsità. Piccinato (1996) definisce tale impostazione completamente predittiva perché grazie al concetto di scambiabilità gli unici elementi da probabilizzare sono le osservazioni; il risultato ultimo consiste nella distribuzione predittiva finale che assegna probabilità a "future" osservazioni (o funzioni di esse) condizionatamente al campione osservato, definendo implicitamente come popolazione l'insieme di unità statistiche i cui risultati siano scambiabili. Da un punto di vista concettuale, la predittiva viene costruita secondo l'usuale meccanismo bayesiano di aggiornamento dell'informazione una volta note le intensità manifestate nel campione ed i profondi legami con il bootstrap sono evidenziati dall'assenza, all'interno dell'insieme rispetto al quale lo stimatore della legge di probabilità è condizionato, di qualunque variabile casuale non osservabile. Nel caso più semplice di applicazione del teorema di rappresentazione, la variabile casuale d'interesse è una Bernoulli ed il parametro indotto dalla legge sulle osservazioni è il limite della statistica sufficiente frequenza relativa dei successi, ossia la media della variabile casuale dicotomica.

1.2 Piano della Tesi

La proposta di Muliere e Secchi è formalmente ascrivibile al campo della statistica bayesiana soggettiva: la *priori* è una legge di probabilità sulle osservazioni. Alcune questioni sembrano meritevoli di approfondimenti:

- ▷ posto che il ricercatore abbia conoscenze a priori parziali, ossia tali da riflettere solo specifici aspetti del processo generatore dei dati, come utilizzarle per costruire un processo di Dirichlet?
- ▷ Ove sia definita per esse una legge di probabilità non degenere, come si può scontarla nelle stime?
- ▷ Esistono alternative significativamente diverse rispetto al bootstrap di Rubin se si intende fare ricampionamento in ottica bayesiana oggettiva?
- ▷ In quale modo il ricampionamento bayesiano è in grado di mantenere le caratteristiche di dipendenza, univariate e/o multivariate, presenti nei dati?

Tali quesiti esprimono l'articolazione del percorso di ricerca di cui si intende dare conto in queste pagine. Il secondo capitolo è dedicato ad una breve rassegna volta a delineare significato, ruolo, stima delle verosimiglianze empiriche generalizzate e relazioni con il metodo generalizzato dei momenti; il terzo ne indaga l'utilità ai fini dei processi di Dirichlet e delle stime non parametriche di densità prendendo le mosse da un esercizio di simulazione di Muliere e Secchi (1996) del quale, gradualmente, vengono rilassate le ipotesi fino ad integrare una proposta di ricampionamento in ambito bayesiano oggettivo, fondata sul teorema di rappresentazione e coerente con risultati noti relativi alle statistiche sufficienti a fini predittivi; il quarto capitolo mostra due esercizi intesi a vagliare le possibilità offerte dalla procedura di ricampionamento in presenza di dati economici multivariati dipendenti in serie storica e, rispettivamente, caratterizzati da sparsità tale da ingenerare elevati valori di asimmetria e curtosi; il quinto concerne un'esposizione complessiva delle conclusioni emergenti dalle analisi svolte; nell'appendice A, le dimostrazioni che necessitano di maggiore spazio, inclusa l'estensione del teorema con il quale Grendar e Judge dimostrano la liceità d'uso delle verosimiglianze empiriche in ottica bayesiana; infine, nell'appendice B, i codici R riguardanti funzioni definite allo scopo di condurre le elaborazioni ed un programma di esempio finalizzato alla stima delle distribuzioni predittive e all'approssimazione di quelle relative a sezioni di processi aleatori.

Capitolo 2

Verosimiglianze e informazione incompleta

Nei problemi di stima parametrici, un modo di evitare l'esplicita modellazione della legge generatrice dei dati risiede nel ricorso alla minimizzazione di una funzione di danno. In questo contesto, occupa un ruolo di primo piano il metodo generalizzato dei momenti (*GMM*). Negli anni più recenti, molti studi hanno puntato l'attenzione sulle verosimiglianze semiparametriche o verosimiglianze empiriche generalizzate (*GEL*), costruite anch'esse sulla base di condizioni di ortogonalità esprimenti le conoscenze del ricercatore in ordine a relazioni vincolari tra dati e parametri. In questo capitolo saranno brevemente trattati alcuni argomenti funzionali alle successive analisi:

- ▷ distribuzioni asintotiche di *GMM* e *GEL*,
- ▷ interpretazione di alcune forme funzionali delle *GEL*,
- ▷ definizione delle *GEL* a partire da misure di divergenza tra leggi di probabilità e modalità di calcolo,
- ▷ uso delle verosimiglianze empiriche di Owen in ambito bayesiano.

2.1 GMM

L'idea sottesa dal metodo dei momenti consiste nel fondare la stima dei parametri di un modello sulla relazione funzionale che li lega ai momenti teorici, ricavando un sistema di k equazioni, le condizioni di ortogonalità, in k incognite, il numero di parametri. Facendo uso dei momenti campionari è possibile risolvere il sistema ottenendo, in genere, stime consistenti e asintoticamente normali. Se il numero di parametri da

stimare è inferiore al numero di equazioni il sistema è sovraidentificato e non tutte le condizioni possono essere simultaneamente soddisfatte. Il metodo generalizzato dei momenti permette di trattare un numero q di condizioni sui momenti maggiore del numero k di parametri da stimare, tramite la costruzione di una funzione di danno quadratica. L'ipotesi che esista un unico θ_0 tale da soddisfare le equazioni $E[\mathbf{g}(x,\theta)] = \mathbf{0}$ rappresenta la condizione di identificabilità. Diversi stimatori possono essere ottenuti precisando opportunamente le condizioni sui momenti; un notevole esempio è rappresentato dallo stimatore di massima verosimiglianza, quando si coinsiderino le k equazioni $\mathbf{g}(x_i,\theta) = \partial \ln f(x_i,\theta)/\partial\theta$, mentre l'identificabilità è in questo caso rappresentata dall'essere $\mathbf{I}(\theta)$ di rango pieno. Le proprietà dello stimatore in piccoli campioni sono ignote e le procedure di inferenza sono basate sull'analisi asintotica (Cappuccio e Orsi, 2005). La prima condizione che deve risultare verificata concerne l'uniforme[1] convergenza dei momenti campionari:

$$\overline{\mathbf{g}}(\theta) \equiv \frac{1}{N}\sum_i \mathbf{g}(x_i,\theta) \xrightarrow{p} E[\mathbf{g}(x,\theta)] = \mathbf{0}$$

La convergenza uniforme in probabilità di $\overline{\mathbf{g}}(\theta)$implica la convergenza uniforme della funzione criterio $Q(\theta)$:

$$Q(\theta) \equiv \overline{\mathbf{g}}(\theta)^t \mathbf{V}\overline{\mathbf{g}}(\theta)$$

Mentre ai fini della consistenza è sufficiente richiedere che la matrice $\mathbf{V}$ converga ad una matrice finita, la scelta è rilevante per la varianza asintotica dello stimatore. Per ricavare la legge asintotica dello stimatore si considerino le condizioni del primo ordine:

$$\frac{\partial Q}{\partial\theta} = 2\left[\frac{\partial\overline{\mathbf{g}}(\theta)}{\partial\theta^t}\right]^t \mathbf{V}\overline{\mathbf{g}}(\theta) = \mathbf{0}$$

Per il metodo δ, con θ_+ compreso tra θ e θ_0:

[1] La convergenza deve verificarsi per l'estremo superiore, rispetto al parametro, della funzione aleatoria in modo che sia annullata l'influenza che esso può esercitare sulla successione delle probabilità. Dunque,

- ▷ la successione $\{\mathbf{g}_N(x,\theta)\}$ converge uniformemente in probabilità al limite $\mathbf{g}(\theta)$quando

$$sup_\theta |\mathbf{g}_N(x,\theta) - \mathbf{g}(\theta)| \xrightarrow{p} 0$$

- ▷ se $\{X_N\}$è una successione di variabili casuali *i.i.d.* e se $E[sup_\theta |\mathbf{g}_N(x,\theta)|] < \infty$,
  - ▷ $E[\mathbf{g}(x,\theta)]$ è continua in θ
  - ▷ $\frac{1}{N}\sum_i \mathbf{g}(x_i,\theta) \xrightarrow{p} E[\mathbf{g}(x,\theta)]$

$$\overline{\mathbf{g}}(\theta) = \overline{\mathbf{g}}(\theta_0) + \left.\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right|_{\theta_+} (\theta - \theta_0)$$

Sostituendo nelle condizioni del primo ordine:

$$2\left[\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right]^t \mathbf{V}\left[\overline{\mathbf{g}}(\theta_0) + \left.\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right|_{\theta_+} (\theta - \theta_0)\right] = \mathbf{0}$$

$$-\left[\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right]^t \mathbf{V}\overline{\mathbf{g}}(\theta_0) = \left[\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right]^t \mathbf{V}\left.\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right|_{\theta_+} (\theta - \theta_0)$$

$$-\left[\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right]^t \mathbf{V}\sqrt{N}\overline{\mathbf{g}}(\theta_0) = \left[\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right]^t \mathbf{V}\left.\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right|_{\theta_+} \sqrt{N}(\theta - \theta_0) \tag{2.1}$$

Date le relazioni

$$\begin{cases} \sqrt{N}\overline{\mathbf{g}}(\theta_0) & \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S}_0) \\ \frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t} & = \frac{1}{N}\sum_i \frac{\partial \mathbf{g}(x_i,\theta)}{\partial \theta^t} \quad \xrightarrow{p} \quad E\left[\left.\frac{\partial \overline{\mathbf{g}}(\theta)}{\partial \theta^t}\right|_{\theta_0}\right] = \mathbf{G}_0 \end{cases}$$

dalla (2.1) si ottiene

$$-\mathbf{G}_0^t\mathbf{V}\sqrt{N}\overline{\mathbf{g}}(\theta_0) = \mathbf{G}_0^t\mathbf{V}\mathbf{G}_0\sqrt{N}(\theta - \theta_0)$$

$$-\left(\mathbf{G}_0^t\mathbf{V}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^t\mathbf{V}\sqrt{N}\overline{\mathbf{g}}(\theta_0) = \sqrt{N}(\theta - \theta_0)$$

ed è possibile concludere

$$\sqrt{N}(\theta - \theta_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\left(\mathbf{G}_0^t\mathbf{V}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^t\mathbf{V}\right]\mathbf{S}_0\left[\mathbf{V}\mathbf{G}_0\left(\mathbf{G}_0^t\mathbf{V}\mathbf{G}_0\right)^{-1}\right]\right)$$

Hansen (1982) dimostra che lo stimatore a varianza minima si ottiene ponendo $\mathbf{V} = \mathbf{S}_0^{-1}$, da cui si ricava:

$$\sqrt{N}(\theta - \theta_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left[\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right]^{-1}\right) \tag{2.2}$$

Ai fini della stima, essendo $\mathbf{S}_0$ ignota e dipendente da θ_0, si ricorre ad una procedura iterativa (in genere arrestata alla prima iterazione):

1. ponendo $\mathbf{V} = \mathbf{I}$ si ottiene una stima consistente $\widehat{\theta}$,
2. si stima $\widehat{\mathbf{S}} = \frac{1}{N}\sum_i \left[\mathbf{g}\left(x_i, \widehat{\theta}\right)\mathbf{g}\left(x_i, \widehat{\theta}\right)^t\right]$,
3. la stima efficiente di θ si ottiene minimizzando la funzione criterio $Q(\theta) \equiv \overline{\mathbf{g}}(\theta)^t\widehat{\mathbf{S}}^{-1}\overline{\mathbf{g}}(\theta)$

Per verificare le condizioni sui q momenti, con $q > k$ e sotto l'ipotesi nulla di corretta specificazione, Hansen propone la statistica:
$NQ(\theta) \equiv N\overline{\mathbf{g}}\left(\widehat{\theta}\right)^t \widehat{\mathbf{S}}^{-1}\overline{\mathbf{g}}\left(\widehat{\theta}\right) \sim \chi^2_{q-k}$.

2.2 Densità Aumentate

Chesher e Smith (1997) propongono un test *LR* per valutare la coerenza di un modello parametrico $f(\mathbf{x};\theta)$ rispetto a condizioni sui momenti. Esso viene costruito aumentando la verosimiglianza tramite una funzione η, veicolatrice dei vincoli, tale che risultino verificate le seguenti proprietà:

$$\begin{cases} \forall w \in \mathbb{R}: \ \eta(w), \ \frac{\partial \eta}{\partial w} > 0 \\ \eta(0) = \left.\frac{\partial \eta}{\partial w}\right|_{w=0} = 1 \end{cases}$$

Formalmente la verosimiglianza aumentata si può esprimere come

$$r(\mathbf{x};\theta,\lambda) = [q(\theta,\lambda)]^{-1} f(\mathbf{x};\theta)\eta\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right]$$

essendo il fattore di normalizzazione pari a $q(\theta,\lambda) = \int f(\mathbf{x};\theta)\eta\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right] d\mathbf{x}$. Vale il seguente teorema:

$$E[\mathbf{g}(\mathbf{x},\theta)] = \mathbf{0} \Leftrightarrow \lambda = \mathbf{0}$$

Intuitivamente, poiché λ rappresenta la sensibilità della verosimiglianaza ai vincoli $\mathbf{g}(\mathbf{x},\theta)$, se essi sono soddisfatti da $f(\mathbf{x};\theta)$ ossia se $E_f[\mathbf{g}(\mathbf{x},\theta)] = 0$ deve anche essere $\lambda = 0$. Una semplice verifica consiste nell'osservare che se $\eta(0) = \left.\frac{\partial \eta}{\partial w}\right|_{w=0} = 1$, la verosimiglianza aumentata si riduce a $f(\mathbf{x};\theta)$ e la costante di normalizzazione $E_f\eta\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right] = 1$; sotto blande condizioni di regolarità, derivando ambo i membri di quest'ultima relazione rispetto a λ e $\mathbf{g}(\mathbf{x},\theta)$ si verifica facilmente

$$E[\mathbf{g}(\mathbf{x},\theta)] = E_f \left.\frac{\partial \eta}{\partial\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right]} \frac{\partial\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right]}{\partial \lambda}\right|_{\lambda=0} = \mathbf{0}$$

$$\lambda = E_f \left.\frac{\partial \eta}{\partial\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right]} \frac{\partial\left[\lambda^t \mathbf{g}(\mathbf{x},\theta)\right]}{\partial \mathbf{g}(\mathbf{x},\theta)}\right|_{\mathbf{g}(\mathbf{x},\theta)=0} = \mathbf{0}$$

Il test delle ipotesi in ordine alle condizioni sui momenti viene condotto confrontando il modello vincolato $f(\mathbf{x};\theta)$ e quello libero, ossia $r(\mathbf{x};\theta,\lambda)$; il test per $H_0 : \lambda = \mathbf{0}$ equivale al test su $E[\mathbf{g}(\mathbf{x},\theta)] = \mathbf{0}$. Sotto l'ipotesi nulla è facile caratterizzare score e informazione attesa, constatandone l'invarianza rispetto alla concreta forma funziona-

le assunta da η. In particolare, indicando con y_θ e $y_{\theta\theta}$ le derivate prima e seconda di y rispetto a θ si possono facilmente verificare le seguenti relazioni:

$$\begin{cases} \mathbf{s}(\theta,\lambda)_{\lambda=\mathbf{0}} = \begin{bmatrix} -\frac{q_\theta}{q} + \frac{f_\theta}{f} + \frac{\lambda}{\eta}\mathbf{g}_\theta \\ -\frac{q_\lambda}{q} + \frac{\eta_w}{\eta}\mathbf{g} \end{bmatrix}_{\lambda=\mathbf{0}} = \begin{bmatrix} \frac{f_\theta}{f} \\ \mathbf{g} \end{bmatrix} \\ \mathbf{I}(\theta,\lambda)_{\lambda=\mathbf{0}} = \begin{bmatrix} E\left(\frac{f_\theta}{f}\frac{f_\theta}{f}^t\right) & E\left(\frac{f_\theta}{f}\mathbf{g}^t\right) \\ E\left(\mathbf{g}\frac{f_\theta}{f}^t\right) & E\left(\mathbf{g}\mathbf{g}^t\right) \end{bmatrix} \end{cases}$$

2.3 Verosimiglianze Semiparametriche

Smith (1997) e Newey e Smith (2004) discutono il caso di osservazioni eventualmente dipendenti in serie storica quando non si abbia una specifica conoscenza della densità $f(\mathbf{x};\theta)$; adottando una distribuzione multinomiale, ossia attribuendo ad ogni osservazione nel campione una massa di probabilità pari a $\frac{1}{N}$ e definendo $\mathbf{g}(\theta) \equiv \sum_{j=-m}^{m} \omega(j)\mathbf{g}\left(x_j,\theta\right)$, la quasi-verosimiglianza aumentata risulta:

$$r(\mathbf{x};\theta,\lambda) = \prod N^{-1}\zeta\left[\lambda^t\mathbf{g}(\theta)\right]$$

ossia, passando al logaritmo

$$ln\, r(\mathbf{x};\theta,\lambda) = -NlnN + \sum \eta\left[\lambda^t\mathbf{g}(\theta)\right]$$

Se le $\mathbf{g}\left(x_j,\theta\right)$ sono incorrelate, $\omega(0) = 1$ e $\omega(j) = 0$ per $|j| > 0$. Poiché θ compare soltanto come argomento di η, affinché il problema di ottimo sia ben definito occorre qualificare la soluzione estremante. Indicando con η_w e η_{ww} le derivate prima e seconda di η rispetto all'argomento w, sia η concava, strettamente monotona, definita su un aperto convesso A contentente 0, tale che

$$\begin{cases} \forall w \in A:\ \eta(w) > 0,\ \eta(0) = 1 \\ \eta_w|_{w=0} = \eta_{ww}|_{w=0} = -1 \end{cases}$$

allora:

$$\begin{cases} \widehat{\lambda}(\theta) = \arg\max_\lambda\left[ln\, r(\mathbf{x};\theta,\lambda)\right] \\ \widehat{\theta} = \arg\min_\theta\left[ln\, r\left(\mathbf{x};\theta,\widehat{\lambda}\right)\right] \end{cases}$$

In altri termini, le soluzioni estremanti individuano un punto di sella. Rinviando al paragrafo 2.5.2 gli approfondimenti sulla circostanza, può risultare interessante valutare alcune proprietà asintotiche. Sotto l'ipotesi che le condizioni sui momenti siano

verificate, indicando con θ_0 il "vero" valore del parametro, è possibile mostrare la seguente relazione:

$$\sqrt{S}\begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Lambda_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}(\theta_0)^{-1} \end{bmatrix}\right) \tag{2.3}$$

ove

$$\mathbf{I}(\theta_0)^{-1} = \left(\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right)^{-1}$$
$$\Lambda_0 = \mathbf{S}_0^{-1} - \mathbf{S}_0^{-1}\mathbf{G}_0\mathbf{I}(\theta_0)^{-1}\mathbf{G}_0^t\mathbf{S}_0^{-1}$$

In dettaglio, la (2.3) si può ottenere con semplici calcoli; siano:

$$\mathbf{s}(\lambda,\theta) = \begin{bmatrix} \sum \eta_{w_i}(w)\,\mathbf{g}_i \\ \sum \eta_{w_i}(w)\,\mathbf{g}_{\theta_i}^t\lambda \end{bmatrix}$$

$$\mathbf{H}(\lambda,\theta) = \begin{bmatrix} \sum \eta_{ww_i}\mathbf{g}_i\mathbf{g}_i^t & \sum(\eta_{ww_i}\mathbf{g}_{\theta_i}^t\lambda\mathbf{g}_i + \eta_{w_i}\mathbf{g}_{\theta_i}) \\ \sum(\eta_{ww_i}\mathbf{g}_i^t\lambda\mathbf{g}_{\theta_i} + \eta_{w_i}\mathbf{g}_{\theta_i}^t) & \sum(\eta_{ww_i}\mathbf{g}_{\theta_i}^t\lambda\lambda^t\mathbf{g}_{\theta_i} + \eta_{w_i}\mathbf{g}_{\theta\theta_i^t}^t\lambda) \end{bmatrix}$$

per il metodo δ,

$$\mathbf{s}\left(\widehat{\lambda},\widehat{\theta}\right) = \mathbf{s}(\mathbf{0},\theta_0) + \mathbf{H}(\lambda_+,\theta_+)\begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix}$$

da quest'ultima relazione si trae

$$-\left[\frac{H(\lambda_+,\theta_+)}{N}\right]^{-1}\mathbf{s}(\mathbf{0},\theta_0)\frac{1}{\sqrt{N}} = \begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix}\sqrt{N} \tag{2.4}$$

Concentrando l'attenzione sui fattori a sinistra, per quanto riguarda il primo termine,

$$\left[-\frac{H(\lambda_+,\theta_+)}{N}\right]^{-1} \xrightarrow{p} \begin{bmatrix} E(\mathbf{g}\mathbf{g}^t) & E(\mathbf{g}_\theta) \\ E(\mathbf{g}_\theta^t) & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_0 & \mathbf{G}_0 \\ \mathbf{G}_0^t & \mathbf{0} \end{bmatrix}^{-1} =$$

$$\begin{bmatrix} \mathbf{S}_0^{-1} - \mathbf{S}_0^{-1}\mathbf{G}_0\left(\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^t\mathbf{S}_0^{-1} & \mathbf{S}_0^{-1}\mathbf{G}_0\left(\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right)^{-1} \\ \left(\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^t\mathbf{S}_0^{-1} & -\left(\mathbf{G}_0^t\mathbf{S}_0^{-1}\mathbf{G}_0\right)^{-1} \end{bmatrix} \equiv \mathbf{J}_0$$

mentre per il secondo, ponendo $\Omega_0 \equiv \begin{bmatrix} \mathbf{S}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$,

$$\frac{1}{\sqrt{N}}\mathbf{s}(\mathbf{0},\theta_0) \xrightarrow{d} N(\mathbf{0},\Omega_0)$$

quindi

$$\left[-\frac{H(\lambda_+,\theta_+)}{N}\right]^{-1}\mathbf{s}(\mathbf{0},\theta_0)\frac{1}{\sqrt{N}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\mathbf{J}_0\Omega_0\mathbf{J}_0^t\right)$$

ossia, ricordando la (2.4):

$$\begin{bmatrix} \widehat{\lambda} \\ \widehat{\theta}-\theta_0 \end{bmatrix}\sqrt{N} \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\begin{bmatrix} \Lambda_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}(\theta_0)^{-1} \end{bmatrix}\right) \tag{2.5}$$

Dunque, la legge delle stime *MLE* per θ coincide con quella asintotica ottenuta per le stime *GMM*. Le relazioni tra gli stimatori possono essere apprezzate anche prescindendo dalla considerazione delle rispettive leggi asintotiche; le soluzioni estremanti sono individuate dalle equazioni dello *score*; il primo blocco ha tante equazioni quante sono le condizioni sui momenti, ossia $\dim(\mathbf{g}) = \dim(\lambda)$, mentre il secondo blocco ha un numero di equazioni pari a $\dim(\theta)$:

$$\begin{cases} \sum \eta_{w_i}(w)\,\mathbf{g}_i & = 0 \\ \sum \eta_{w_i}(w)\,\mathbf{g}_{\theta_i}^t\lambda & = 0 \end{cases}$$

Il primo blocco di equazioni coincide con le condizioni di ortogonalità di *GMM*, avendo sostituito $\frac{1}{N}\sum_i \mathbf{g}_i = \mathbf{0}$ con una somma pesata.

2.4 Stime di massima entropia

Considerando, per semplicità, una variabile casuale discreta X che assume valori secondo una distribuzione ignota $p(X)$, si voglia risalire alle probabilità $p(x_i)$ delle singole realizzazioni avendo a disposizione, quale unica conoscenza, il valore atteso di alcune funzioni $\psi(X)$. In mancanza di un numero di vincoli almeno pari al numero di parametri da stimare, il principio di ragione insufficiente di Laplace (considerare equiprobabili gli eventi che sottendono ciascuna realizzazione) parrebbe offrire una soluzione, ma la scelta sarebbe arbitraria: l'assenza di informazione non permette di ritenere una qualsiasi distribuzione più plausibile di un'altra. La teoria dell'informa-

zione suggerisce una soluzione ragionevole: l'ammontare di incertezza implicito in una legge di probabilità può essere misurato in modo univoco dall'entropia:

$$H(\mathbf{p}) = -\sum_i p(x_i)\, ln\left[p(x_i)\right]$$

La distribuzione che massimizza l'entropia nel rispetto dei vincoli noti è, tra le diverse possibili, quella che implica la minore quantità di assunzioni sul processo generatore delle osservazioni. In altri termini, adottare la distribuzione che massimizza l'entropia è coerente con l'ammissione di mancanza di informazioni ulteriori rispetto a quelle rappresentate dai vincoli noti (Jaynes, 1957). Il problema può essere rappresentato c.s.:

$$\mathbf{p} = argmax_{\mathbf{p}}\left\{-\sum p_i ln(p_i)\right\} \text{ sub } \begin{cases} E_{\mathbf{p}}\left[\psi(x)\right] = \mathbf{0} \\ \sum_i p_i = 1 \end{cases}$$

Una volta formata la lagrangiana

$$\mathscr{L} = -\sum p_i ln(p_i) - \rho\left(\sum p_i - 1\right) - \sum p_i \lambda^t \psi_i$$

le condizioni del primo ordine sono:

$$\begin{cases} \partial\mathscr{L}/\partial p_i \equiv -ln(p_i) - 1 - \rho - \lambda^t \psi_i = 0 \\ \partial\mathscr{L}/\partial\rho \equiv (1 - \sum p_i) = 0 \\ \partial\mathscr{L}/\partial\lambda \equiv \sum p_i \psi_i^t = \mathbf{0} \end{cases} \tag{2.6}$$

La prima di esse implica:

$$p_i = e^{-\rho-1-\lambda^t\psi_i}$$

A causa del vincolo di somma unitaria si ottiene:

$$p_i = \frac{e^{-\lambda^t\psi_i}}{\sum e^{-\lambda^t\psi_i}}$$

Inoltre, combinando la prima e la terza delle (2.6) è possibile esplicitare i moltiplicatori di Lagrange:

$$\begin{gathered} \sum p_i\left[-ln(p_i) - 1 - \rho - \lambda^t\psi_i\right]\psi_i^t = 0 \\ \sum p_i ln(p_i)\,\psi_i^t = \lambda^t \sum p_i \psi_i \psi_i^t \\ \left[\sum p_i \psi_i \psi_i^t\right]^{-1}\left[\sum p_i ln(p_i)\,\psi_i^t\right] = \lambda^t \end{gathered}$$

Si può osservare come la massimizzazione dell'entropia conduca, con differenti motivazioni, ad una soluzione simile a quella ispirata dal principio di ragione insufficiente.

Si consideri l'espressione backward[2] della divergenza di Kullback-Leibler di una data distribuzione rispetto alla uniforme:

$$D_{KL}^{B}\left(\mathbf{p},\mathbf{1}\frac{1}{N}\right)=\sum p_i ln\left(Np_i\right)=E_p\left[ln\left(p\right)\right]+ln\left(N\right)=-H\left(p\right)+ln\left(N\right)$$

La massimizzazione dell'entropia dunque equivale alla minimizzazione della divergenza di Kullback-Leibler backward rispetto ad una distribuzione uniforme. Rimane da sottolineare un aspetto finora non evidenziato: al divergere della numerosità campionaria la scelta di massimizzare l'entropia implica il riferimento alla distribuzione che rende massima la probabilità di ottenere le realizzazioni effettivamente osservate. Si deve a G. Wallis (Jaynes e Bretthorst, 2003) una dimostrazione ottenuta senza invocare le consuete tecniche di identificazione per soluzioni estremanti, posto che la distribuzione sia modellata come una multinomiale; indicando con n_i la frequenza assoluta riferita alla realizzazione della modalità i^{ma} e con $p\left(\mathbf{n}\right)$ la probabilità di aver osservato un vettore di realizzazioni, sia

$$p\left(\mathbf{n}\right)=\frac{N!}{\prod\left(n_j!\right)}\prod p_i^{n_i},\ \ \sum n_j=N,\quad E\left(n_i\right)=p_iN$$

Notando che $p\left(\mathbf{n}\right)\leq 1\ \Leftrightarrow\ \frac{N!}{\prod(n_j!)}\leq\frac{1}{\prod p_i^{n_i}}=\prod p_i^{-n_i}$ si evince che

$$p\left(\mathbf{n}\right)\rightarrow 1\Leftrightarrow\frac{N!}{\prod\left(n_j!\right)}\rightarrow\prod p_i^{-n_i}$$

Analoga relazione vale se si considerano funzioni dei due termini:

$$p\left(\mathbf{n}\right)\rightarrow 1\ \Leftrightarrow\left[\frac{1}{N}ln\frac{N!}{\prod\left(n_j!\right)}\right]\rightarrow\left[\frac{1}{N}ln\left(\prod p_i^{-n_i}\right)\right]$$

Dunque, $max\left[p\left(\mathbf{n}\right)\right]\Rightarrow max\left[\left(1/N\right)ln\left(N!\right)-\left(1/N\right)\sum ln\,n_i!\right]$. Tale quantità può essere approssimata utilizzando la formula di Stirling e sostituendo le frequenze assolute con

[2] Corcoran (2000) introduce la notazione backward e forward come alternativa allo scambio degli argomenti nei simboli:

- ▷ $D_{KL}^{B}(\mathbf{p},\mathbf{q})\equiv\sum p_i ln\left(\frac{p_i}{q_i}\right)$
- ▷ $D_{KL}^{F}(\mathbf{p},\mathbf{q})\equiv\sum q_i ln\left(\frac{q_i}{p_i}\right)$

i rispettivi valori attesi:

$$\left[\frac{1}{N}ln\frac{N!}{\prod(n_j!)}\right] \approx \frac{1}{N}ln\left[(2\pi N)^{\frac{1}{2}}\frac{N^N}{e^N}\right] - \frac{1}{N}\sum ln\left[(2\pi p_i N)^{\frac{1}{2}}\frac{(p_i N)^{p_i N}}{e^{p_i N}}\right]$$

Svolgendo e semplificando il secondo membro si ottiene

$$\left[\frac{1}{N}ln\frac{N!}{\prod(n_j!)}\right] \approx \frac{1}{2N}ln(2\pi N) - \frac{1}{2N}\sum ln(2\pi p_i N) - \sum p_i ln(p_i)$$

e si verifica immediatamente che:

$$\lim_{N\to\infty}\left[\frac{1}{N}ln\frac{N!}{\prod(n_j!)}\right] \approx H(p)$$

Resta così dimostrato che al divergere della numerosità campionaria, massimizzare la probabilità di ottenere le realizzazioni osservate equivale a massimizzare l'entropia della distribuzione.

2.5 Verosimiglianze empiriche

Date N realizzazioni $i.i.d.$ di una variabile casuale discreta, i "pesi" p_i che massimizzano la propria produttoria nel rispetto del vincolo di somma unitaria sono facilmente individuabili. Passando ai logaritmi, la lagrangiana è espressa da

$$\mathscr{L} = \sum ln(p_i) - \rho\left(\sum p_i - 1\right)$$

mentre le condizioni del primo ordine sono

$$\begin{cases} \partial L/\partial p_i \equiv \frac{1}{p_i} - \rho = 0 \\ \partial L/\partial \rho \equiv (1 - \sum_i p_i) = 0 \end{cases}$$

Dalla prima si ricava $p_i = 1/\rho$, per la seconda $1 = N/\rho$ ed è immediato verificare $p_i = 1/N$. Se n_i unità statistiche manifestano la medesima modalità del carattere, la relativa probabilità empirica risulta pari a n_i/N ed emerge quale distribuzione la multinomiale. Le proprietà asintotiche derivano dal teorema di Glivenko Cantelli: al divergere di N la funzione di ripartizione empirica converge uniformemente alla (ignota) funzione di ripartizione teorica. Lo scopo di individuare la distribuzione che massimizza la verosimiglianza empirica della variabile casuale N^{pla} descritta dal campione, nel rispetto di vincoli noti (circa il valore atteso di date funzioni delle osservazioni),

può essere perseguito minimizzandone la divergenza rispetto alla distribuzione uniforme. Se si adotta quale misura di diversità la D^B_{KL} si ottengono i risultati descritti nel precedente paragrafo e la verosimiglianza empirica massimizza l'entropia. L'adozione della divergenza di Kullback-Leibler forward, ossia $D^F_{KL} \equiv \sum (1/N)\, ln\,[(1/N)\,/\,(p_i)]$, conduce alla soluzione proposta da Owen (1988). La lagrangiana si può scrivere

$$\mathscr{L} = -\sum \frac{1}{N} ln\,(Np_i) - \rho\left(\sum p_i - 1\right) - \sum p_i \lambda^t \psi_i$$

e le condizioni del primo ordine risultano:

$$\begin{cases} \partial\mathscr{L}/\partial p_i \equiv -\frac{1}{Np_i} - \rho - \lambda^t \psi_i = 0 \\ \partial\mathscr{L}/\partial \rho \equiv (1 - \sum p_i) = 0 \\ \partial\mathscr{L}/\partial \lambda \equiv \sum p_i \psi_i^t = \mathbf{0} \end{cases} \tag{2.7}$$

Dalla prima equazione si ottiene

$$p_i = \frac{1}{N} \frac{1}{-\rho - \lambda^t \psi_i}$$

Poiché deve valere $\sum \left[-\frac{1}{Np_i} - \rho - \lambda^t \psi_i\right] p_i = 0$, a causa dei vincoli imposti dalla seconda e dalla terza equazione si ricava $-\rho = 1$. Le verosimiglianze di massima entropia e le verosimiglianze empiriche sono dedotte dalla minimizzazione delle D^B_{KL} e D^F_{KL}; queste ultime costituiscono gli estremi di misure di divergenza generalizzate espresse dalle discrepanze di Cressie e Read (1984).

2.5.1 Misure di divergenza

Il grado di diversità tra due distribuzioni **p** e **q** è suscettibile di misurazioni alternative, secondo statistiche riconducibili a famiglie di divergenze. Una di esse è costituita dalle discrepanze empiriche di Cressie-Read che possono essere espresse in diversi modi. Due tra i possibili sono:

$$CR_\gamma(\mathbf{p},\mathbf{q}) \equiv \frac{2N}{\gamma(\gamma+1)} \sum p_j \left[\left(\frac{p_j}{q_j}\right)^\gamma - 1\right] = \frac{2N}{\gamma(\gamma+1)} \sum q_j \left[\left(\frac{p_j}{q_j}\right)^{\gamma+1} - 1\right]$$

Si tratta di generalizzazioni della statistica χ^2 (corrispondente al caso di $\gamma = 1$). Infatti, per $\gamma = 1$ il primo tipo di espressione si ottiene come:

$$\chi^2 = N \sum \frac{(p_j - q_j)^2}{q_j} = N \sum \frac{p_j^2 - 2p_j q_j + q_j^2}{q_j} = N \sum \frac{p_j^2}{q_j} - 2N + N =$$

$$N\left[\sum\frac{p_j^2}{q_j}-1\right]=N\sum p_j\left(\frac{p_j}{q_j}-1\right)$$

Le D_{KL}^B e D_{KL}^F sono proporzionali alle soluzioni per le forme indeterminate dei passaggi a limite $\gamma\to 0$ e, rispettivamente, $\gamma\to -1$:

$$\lim_{\gamma\to 0}\frac{2N\sum p_j\frac{\partial}{\partial\gamma}\left[\left(\frac{p_j}{q_j}\right)^{\gamma}-1\right]}{\frac{\partial}{\partial\gamma}\left[\gamma(\gamma+1)\right]}=\lim_{\gamma\to 0}\frac{2N\sum p_j\left(\frac{p_j}{q_j}\right)^{\gamma}ln\frac{p_j}{q_j}}{(\gamma+1)+\gamma}=2N\sum p_j ln\frac{p_j}{q_j}$$

$$\lim_{\gamma\to -1}\frac{2N\sum q_j\frac{\partial}{\partial\gamma}\left[\left(\frac{p_j}{q_j}\right)^{\gamma+1}-1\right]}{\frac{\partial}{\partial\gamma}\left[\gamma(\gamma+1)\right]}=\lim_{\gamma\to -1}\frac{2N\sum q_j\left(\frac{p_j}{q_j}\right)^{\gamma+1}ln\frac{p_j}{q_j}}{(\gamma+1)+\gamma}=2N\sum q_j ln\frac{q_j}{p_j}$$

Se si conviene che $\forall j\ \ q_j=1/N$, il limite per $\gamma\to -1$ coincide con il logaritmo di un rapporto di verosimiglianza:

$$LR=\prod\left(\frac{1/N}{p_j}\right)\Rightarrow lnLR=-\sum ln\left(Np_j\right)=N\sum\frac{1}{N}ln\frac{1/N}{p_j}$$

2.5.2 GEL e MD

È possibile mostrare che le stime ottenute dalla minimizzazione delle discrepanze di Cressie Read (*MD*) sono punti estremanti della verosimiglianza empirica generalizzata (*GEL*). Ponendo $I_\gamma\left(\mathbf{p},\mathbf{1}\frac{1}{N}\right)\equiv\frac{1}{2N}CR_\gamma$ i due approcci possono essere sintetizzati, rispettivamente, dalle equazioni

$$MD:\quad\begin{cases}\widehat{\mathbf{p}}(\theta)=\arg\min_{\mathbf{p}}\left[I_\gamma\left(\mathbf{p},\mathbf{1}\frac{1}{N}\right)-\rho\left(\sum p_i-1\right)-\sum p_i\lambda^t\psi\left(x_i,\theta\right)\right]\\ \widehat{\theta}=\arg\min_\theta I_\gamma\left[\widehat{p}(\theta),\frac{1}{N}\right]\end{cases}$$

$$GEL:\quad\begin{cases}\eta\left[\upsilon^t\psi\left(x_i,\theta\right)\right]\equiv-\frac{1}{\gamma+1}\left[1+\upsilon^t\psi\left(x_i,\theta\right)\right]^{\frac{\gamma+1}{\gamma}}\\ \widehat{\upsilon}(\theta)=\arg\max_\upsilon\frac{1}{N}\sum_i\eta\left[\upsilon^t\psi\left(x_i,\theta\right)\right]\\ \widehat{\theta}=\arg\min_\theta\frac{1}{N}\sum_i\eta\left[\widehat{\upsilon}(\theta)^t\psi\left(x_i,\theta\right)\right]\end{cases}$$

Ponendo, per semplicità di notazione, $\eta\left(\xi_i\right)\equiv\eta\left[\upsilon^t\psi\left(x_i,\theta\right)\right]$, la coerenza tra *MD* e *GEL* è assicurata dal risultato $p_i=\left[\partial\eta\left(\xi_i\right)/\partial\xi_i\right]/\sum_i\left[\partial\eta\left(\xi_i\right)/\partial\xi_i\right]$ (Newey e Smith, 2004). Si noti che le q condizioni sui momenti sono espresse come funzioni di un parametro (k-dimensionale) θ anch'esso oggetto di stima. Mentre nel primo problema (minimimizzazione delle discrepanze empiriche) la soluzione estremante viene individuata da una doppia minimizzazione rispetto a $\mathbf{p}$ e a θ, nel secondo (massimizzazione della verosimiglianza empirica generalizzata), la soluzione estremante è un punto di sella in corrispondenza del quale la funzione obiettivo è minimizzata rispetto a θ e

massimizzata rispetto a λ. Il vantaggio di quest'ultima rappresentazione consiste nel minor numero di parametri da stimare: si passa dai $k+N$ del primo problema ai $k+q$ del secondo. La natura di punto di sella della soluzione può essere dimostrata calcolando $\mathbf{p}(\theta)$ ed effettuandone la sostituzione nella funzione obiettivo. Ciò può essere realizzato in termini assai generali (Kitamura, 2006) ricavando la rappresentazione duale del problema primale (minimizzazione delle discrepanze empiriche).

$$v(\mathbf{p}) \equiv I_\gamma(\mathbf{p}) - \rho\left(\sum p_i - 1\right) - \sum p_i\left[\lambda^t \psi(x_i,\theta)\right]$$

Dalle condizioni del primo ordine si ricava:

$$I'_\gamma(p_i) = \rho + \lambda^t \psi_i \Rightarrow p_i = \left(I'_\gamma\right)^{-1}\left(\rho + \lambda^t \psi_i\right) \tag{2.8}$$

$v(\mathbf{p})$ è una funzione convessa di $\mathbf{p}$ e la sua minimizzazione identifica il punto di massimo della funzione duale $v(\mathbf{z}) \equiv \mathbf{z}^t\mathbf{p} - I_\gamma(\mathbf{p})$. Dalle condizioni del primo ordine per quest'ultima è immediato verificare che $\mathbf{z} = I'_\gamma(\mathbf{p})$, ossia $z_i = \rho + \lambda^t \psi_i$. A causa dei vincoli del *problema primale* $\mathbf{z}^t\mathbf{p} = \rho$ e indicando la funzione duale massimizzata come $v^*(\mathbf{z})$ risulta:

$$v^*(\mathbf{z}) \equiv \rho - I_\gamma\left[\left(I'_\gamma\right)^{-1}(\mathbf{z})\right] \tag{2.9}$$

Condizionatamente a θ, $\mathbf{z}$ dipende soltanto da ρ e λ; da $min_{\mathbf{p}}\{v(\mathbf{p})\} = max_{(\rho,\,\lambda)}\{v(\mathbf{z})\}$ segue immediatamente:

$$min_\theta min_{\mathbf{p}}\{v(\mathbf{p},\theta)\} = min_\theta max_{(\rho,\,\lambda)}\{v(\mathbf{z},\theta)\} \tag{2.10}$$

Per esplicitare $v^*(\mathbf{z})$ occorre ricavare $\mathbf{p}$ come funzione di ρ, λ, γ. Considerando $I_\gamma \equiv (1/N)\{1/[\gamma(\gamma+1)]\}\sum_j\left[(Np_j)^{\gamma+1} - 1\right]$ e derivando rispetto a p_i si ottiene:

$$I'_\gamma(p_i) = \frac{1}{\gamma}(Np_i)^\gamma \Rightarrow \frac{1}{N}\left[\gamma I'_\gamma(p_i)\right]^{\frac{1}{\gamma}} = p_i$$

Sostituendo in ambo i membri dell'ultima relazione la (2.8) si giunge a:

$$\frac{1}{N}\left\{\gamma I'_\gamma\left[\left(I'_\gamma\right)^{-1}\left(\rho + \lambda^t \psi_i\right)\right]\right\}^{\frac{1}{\gamma}} = \left(I'_\gamma\right)^{-1}\left(\rho + \lambda^t \psi_i\right)$$

ossia

$$\frac{1}{N}\left[\gamma\left(\rho + \lambda^t \psi_i\right)\right]^{\frac{1}{\gamma}} = \left(I'_\gamma\right)^{-1}\left(\rho + \lambda^t \psi_i\right) \tag{2.11}$$

Sostituendo il primo membro della (2.11) nella (2.9) risulta

$$v^*(z_i) \equiv \rho - \frac{1}{N}\frac{1}{\gamma(\gamma+1)}\left[\left(\gamma\rho + \gamma\lambda^t\psi_i\right)^{\frac{\gamma+1}{\gamma}} - 1\right] \tag{2.12}$$

La funzione che definisce gli stimatori *GEL* è una trasformata lineare della precedente relazione. Si verifica immediatamente:

$$v^*(z_i) \equiv \rho - \frac{(\gamma\rho)^{\frac{\gamma+1}{\gamma}}}{\gamma}\frac{1}{N}\frac{1}{\gamma+1}\left[1+\gamma\left(\frac{\lambda}{\rho\gamma}\right)^t\psi_i\right]^{\frac{\gamma+1}{\gamma}} - \frac{1}{N}\frac{1}{\gamma(\gamma+1)}$$

$$v^*(z_i) \equiv \underbrace{\rho - \frac{1}{N}\frac{1}{\gamma(\gamma+1)}}_{a} - \underbrace{\frac{(\gamma\rho)^{\frac{\gamma+1}{\gamma}}}{\gamma}}_{b}\frac{1}{N}\frac{1}{\gamma+1}\left[1+\upsilon^t\psi_i\right]^{\frac{\gamma+1}{\gamma}}$$

$$\frac{v^*(z_i)-a}{b} = -\frac{1}{N}\frac{1}{\gamma+1}\left[1+\upsilon^t\psi_i\right]^{\frac{\gamma+1}{\gamma}} \tag{2.13}$$

2.6 ETEL

Schennach (2007) pone a confronto le proprietà asintotiche di *EL* ed *ET* mostrando che sotto l'hp.di corretta specificazione del modello

- ▷ *EL* e *ET* sono asintoticamente equivalenti per la stima di θ,
- ▷ in campioni finiti *EL* procura una minore distorsione.

Premesso che nel caso di errata specificazione del modello, diviene rilevante la stima dello pseudo valore di θ, ossia del parametro θ^* associato alla legge più vicina (nel senso della D_{KL}) al vero processo generatore dei dati[3], poiché due stimatori possono convergere a pseudo-valori diversi, un termine di confronto può essere individuato

[3] Se $\mathbf{p}$ è la distribuzione che ha generato i dati e $\mathbf{q}(\theta)$ è quella per essi ipotizzata, la divergenza di Kullback-Leibler può essere scritta come:

$$\sum p_i ln\frac{p_i}{q_i(\theta)} = \sum p_i ln\, p_i - \sum p_i ln\, q_i(\theta)$$

Il primo addendo è l'opposto dell'entropia, mentre il secondo, $-\sum p_i ln\, q_i(\theta)$, definisce la *L-divergenza* di $\mathbf{q}$ rispetto a $\mathbf{p}$. Derivando rispetto a θ si ricava:

$$\frac{\partial D_{KL}}{\partial\theta} = -\sum p_i\frac{\partial ln\, q_i(\theta)}{\partial\theta}$$

Le condizioni del primo ordine per il massimo della *logverosimiglianza* $ln\, q_i(\theta)$sono espresse dall'equazione

$$\frac{1}{N}\sum\frac{\partial ln\, q_i(\theta)}{\partial\theta} = 0$$

Quest'ultima relazione è l'omologo campionario della condizione del primo ordine per la minimizzazione della $D_{KL}(\mathbf{p},\mathbf{q})$.

nelle rispettive velocità di convergenza. Gli studi condotti da Schennach mostrano che le proprietà asintotiche di *EL* peggiorano rapidamente rispetto a quelle di *ET* nei casi di errata specificazione. In particolare, vale il seguente teorema:

- ▷ sia $\mathbf{g}(x,\theta) \in C_2$, $\sup_{\theta\in\Theta} E\left\|\mathbf{g}(x,\theta)\right\|^2 < \infty$,
- ▷ se $\inf_{\theta\in\Theta}\left\|E\left[\mathbf{g}(x,\theta)\right]\right\| \neq 0$ e $\sup_x \mathbf{u}^t\mathbf{g}(x,\theta) = \infty \quad \forall\theta,\mathbf{u} \quad$ con $\quad \|\mathbf{u}\| = 1$
- ▷ allora, $\nexists\ \theta^*_{EL} : \left\|\widehat{\theta}_{EL} - \theta^*_{EL}\right\| = O_p\left(N^{-1/2}\right)$

Dunque, $\inf_{\theta\in\Theta}\left\|E\left[\mathbf{g}(x,\theta)\right]\right\| \neq 0 \Rightarrow \widehat{\theta}_{EL} \xrightarrow{p} \theta^*_{EL}$ ad una velocità inferiore a $N^{-1/2}$. Nel caso in cui $\sup_x \mathbf{u}^t\mathbf{g}(x,\theta) = \infty$ solo per alcune "direzioni" individuate dal vettore $\mathbf{u}$, il teorema non si applica a quelle ad esse ortogonali. Intuitivamente, se $E\left[\mathbf{g}(x,\theta^*)\right] \neq \mathbf{0}$, le condizioni del primo ordine per il vettore dei moltiplicatori non si annullano in θ^*

$$\frac{1}{N}\sum \frac{1}{1-\lambda^t\mathbf{g}(x_i,\theta^*)}\mathbf{g}(x_i,\theta^*) \neq \mathbf{0}$$

e non è garantito che $\lambda^* \xrightarrow{p} \mathbf{0}$; tale circostanza implica che alcune probabilità empiriche possano risultare negative. Si possono combinare le proprietà di *ET* ed *EL* in modo da mantenere la robustezza di *ET* (che al divergere di N fornisce "pesi" comunque positivi per le osservazioni) e la proprietà di *EL* di fornire una stima di massima verosimiglianza per $\theta^* : \widehat{\theta} = \arg\max_\theta\left[\frac{1}{N}\sum_i \ln\widehat{p}_i(\theta)\right]$. Anche in questo caso lo stimatore ammette una duplice rappresentazione, basata sulla minimizzazione delle discrepanze ovvero sulla natura di punto di sella della soluzione estremante. Formalmente,

- ▷ *MD*:

  $\widehat{\mathbf{p}}(\theta) = \arg\min_{\mathbf{p}}\left\{I_\gamma\left(\mathbf{p},\mathbf{1}\frac{1}{N}\right)\Big|_{\gamma\to 0} + \rho\left(1-\sum_i p_i\right) - \sum_i p_i\left[\lambda^t\mathbf{g}(x_i,\theta)\right]\right\}$

  $\widehat{\theta} = \arg\min_{\mathbf{p}} I_\gamma\left(\widehat{\mathbf{p}}(\theta),\mathbf{1}\frac{1}{N}\right)\Big|_{\gamma\to -1}$
- ▷ *GEL*:

  $\widehat{\lambda}(\theta) = \arg\max_\lambda \frac{1}{N}\sum_i \eta\left[\lambda^t\mathbf{g}(x_i,\theta)\right]_{\gamma\to 0}$

  $\widehat{\theta} = \arg\min_\theta I_\gamma\left(\widehat{\mathbf{p}}(\theta),\mathbf{1}\frac{1}{N}\right)\Big|_{\gamma\to -1}$

  con:

  $\widehat{p}_i(\theta) = \frac{\eta_{\left[\widehat{\lambda}(\theta)^t\mathbf{g}(x_i,\theta)\right]_{\gamma\to 0}}}{\sum_j \eta_{\left[\widehat{\lambda}(\theta)^t\mathbf{g}\left(x_j,\theta\right)\right]_{\gamma\to 0}}}$

In parole, condizionatamente ai valori di θ si ricavano i moltiplicatori $\widehat{\lambda}(\theta)$ che conducono ad un sistema di pesi $\widehat{\mathbf{p}}(\theta)$ tale da massimizzare l'entropia; condizionatamente

a $\widehat{\lambda}(\theta)$ si possono calcolare le stime di massima verosimiglianza empirica per θ, mediante minimizzazione della D_{KL}^{F} tra $\widehat{\mathbf{p}}(\theta)$ e la distribuzione uniforme; la procedura richiede che i passi di ottimizzazione siano ripetuti fino a convergenza delle stime concernenti moltiplicatori e parametri.

2.7 EL in Campo Bayesiano: il Teorema di Sanov

Grendar e Judge (2007) studiano le proprietà delle verosimiglianze empiriche Owen (1988) in un contesto bayesiano. Oggetto d'interesse è la consistenza delle posteriori sulle leggi di probabilità. Limitando per comodità l'attenzione al caso discreto, siano:

- ▷ X, una v.c. con legge discreta $\mathbf{p}$,
- ▷ $W \equiv \left\{\mathbf{w} : \sum w_i g_j(x_i;\theta) = 0,\ \ j = 1,\ldots,\ J,\ \pi(\mathbf{w}) > 0\right\}$, l'insieme delle distribuzioni che soddisfano date condizioni di ortogonalità e hanno probabilità a priori strettamente positiva,
- ▷ $\widehat{\mathbf{w}}(\theta) = argsup_{\mathbf{w}}\ \pi(\mathbf{w}|\theta)$.

Allora:

- ▷ $\hat{\theta} = argsup_{\theta}\ \pi(\widehat{\mathbf{w}}|\theta)$ è lo stimatore MAP di θ,
- ▷ $\widehat{\mathbf{w}}(\hat{\theta}) = arg\,sup_{\theta} sup_{\mathbf{w}}\ \pi(\mathbf{w}|\theta)$ è la distribuzione MAP che stima $\mathbf{p}$.

In parole, per $N \to \infty$ la $\pi(\mathbf{w}|\theta)$ converge quasi certamente alla $\widehat{\mathbf{w}}(\hat{\theta})$, *L-proiezione* di $\mathbf{p}$ su W che minimizza la *L-divergenza* $L(W\|\mathbf{p}) \equiv -\sum p_i ln(w_i)$.
Tali conclusioni possono essere spiegate definendo le seguenti grandezze:

- ▷ $\ell(\mathbf{w}) \equiv e^{\sum ln(w_i)}$, il valore della logverosimiglianza sulle leggi $\mathbf{w}$,
- ▷ $\pi(\mathbf{w}|\mathbf{x}) \propto \pi(\mathbf{w})\, e^{\sum ln(w_i)}$, il valore della posteriori,
- ▷ $\pi(V \subset W|\mathbf{x}) = \sum_{\mathbf{w}\in V} \pi(\mathbf{w}|\mathbf{x}) / \sum_{\mathbf{w}\in W} \pi(\mathbf{w}|\mathbf{x})$,
- ▷ $\hat{\ell}(A) \equiv sup_{\mathbf{w}\in A}\ \ell(\mathbf{w}) \equiv e^{\sum ln(\widehat{w}_{i,A})}$
- ▷ $\hat{\pi}(A|\mathbf{x}) \equiv sup_{\mathbf{w}\in A}\ \pi(\mathbf{w}|\mathbf{x}) \propto \pi(\widehat{\mathbf{w}}_A)\, e^{\sum ln(\widehat{w}_{i,A})}$
- ▷ $L(A\|\mathbf{p}) = L(\mathbf{w}_A\|\mathbf{p}) \equiv -\sum p_i ln(w_i)$ con $\mathbf{w} \equiv argmin_{\mathbf{w}\in A} L(A\|\mathbf{p})$

Secondo il teorema di Sanov, al divergere di N:

$$P\left\{\frac{1}{N}\ln\left[\pi\left(V\subset W\,|\mathbf{x}\right)\right]\rightarrow-\left[L\left(V\,\|\mathbf{p}\right)-L\left(W\,\|\mathbf{p}\right)\right]\right\}=1$$

La consistenza in *L-divergenza* della moda a posteriori si ottiene come corollario del teorema. In particolare, se si considera

$$V^C\equiv\left\{\mathbf{w}:\left[L\left(\mathbf{w}\,\|\mathbf{p}\right)-L\left(W\,\|\mathbf{p}\right)\right]>\varepsilon,\varepsilon>0,\mathbf{w}\in W\right\}$$

segue:

$$P\left\{\pi\left(\mathbf{w}\in V^C\,|\mathbf{x}\right)\rightarrow 0\right\}=1$$

Grendar e Judge illustrano il teorema considerando la relazione

$$\frac{1}{N}ln\frac{\hat{\pi}\left(V\,|\mathbf{x}\right)}{\hat{\ell}\left(W\right)}\leq\frac{1}{N}ln\frac{\pi\left(V\,|\mathbf{x}\right)}{\pi\left(W\,|\mathbf{x}\right)}\leq\frac{1}{N}ln\frac{\hat{\ell}\left(V\right)}{\hat{\pi}\left(W\,|\mathbf{x}\right)}$$

e notando che per gli estremi si può scrivere:

$$\frac{1}{N}ln\frac{\hat{\ell}\left(V\right)}{\hat{\pi}\left(W\,|\mathbf{x}\right)}=\frac{1}{N}\sum ln\left(\hat{w}_{i,V}\right)-\frac{1}{N}\sum ln\left(\hat{w}_{i,W}\right)-\frac{1}{N}ln\left[\pi\left(\hat{\mathbf{w}}_W\right)\right]$$

$$\frac{1}{N}ln\frac{\hat{\pi}\left(V\,|\mathbf{x}\right)}{\hat{\ell}\left(W\right)}=\frac{1}{N}\sum ln\left(\hat{w}_{i,V}\right)+\frac{1}{N}ln\left[\pi\left(\hat{\mathbf{w}}_V\right)\right]-\frac{1}{N}\sum ln\left(\hat{w}_{i,W}\right)$$

Dalla legge forte dei grandi numeri discende l'identità asintotica degli estremi, e insieme ad essa l'enunciato:

$$\frac{1}{N}ln\frac{\hat{\ell}\left(V\right)}{\hat{\pi}\left(W\,|\mathbf{x}\right)}\rightarrow E_{\mathbf{p}}\left[ln\left(\hat{\mathbf{w}}_V\right)\right]-E_{\mathbf{p}}\left[ln\left(\hat{\mathbf{w}}_W\right)\right]\equiv-L\left(V\,\|\mathbf{p}\right)+L\left(W\,\|\mathbf{p}\right)$$

$$\frac{1}{N}ln\frac{\hat{\pi}\left(V\,|\mathbf{x}\right)}{\hat{\ell}\left(W\right)}\rightarrow E_{\mathbf{p}}\left[ln\left(\hat{\mathbf{w}}_V\right)\right]-E_{\mathbf{p}}\left[ln\left(\hat{\mathbf{w}}_W\right)\right]\equiv-L\left(V\,\|\mathbf{p}\right)+L\left(W\,\|\mathbf{p}\right)$$

In altri termini la posteriori, al divergere di N, si concentra in un intorno arbitrariamente piccolo della *L-proiezione* di $\mathbf{p}$ su W. Si noti che, in assenza di ulteriori qualificazioni, le L-divergenze circoscrivono il novero delle distribuzioni che forniscono una posteriori consistente:

- ▷ il risultato di consistenza non è valido per distribuzioni che non minimizzino *L-divergenze* o loro trasformate banali,
- ▷ i risultati di consistenza basati sulle *L-divergenze* vengono meno se si adotta un campionamento non *i.i.d.*.

L'importanza del risultato e la possibilità di una sua estensione costituiscono temi di approfondimento del prossimo capitolo.

Capitolo 3

Ricampionamento Bayesiano

In riferimento all'esposizione di alcune idee circa l'uso delle *GEL* nel ricampionamento bayesiano, sembra opportuno ricordare, pur rinunciando ad ambizioni di esaustività:

- ▷ alcune nozioni concernenti il processo di Dirichlet; quest'ultimo permette di probabilizzare gli elementi di generici spazi misurabili e costituisce l'impianto teorico utile a trattare due tra i più popolari strumenti bayesiani di ricampionamento, ossia
- ▷ il bootstrap con priori impropria,
- ▷ il bootstrap con priori propria.

Successivamente saranno discussi:

- ▷ l'introduzione di informazioni a priori sui parametri d'interesse quando si faccia uso del bootstrap bayesiano,
- ▷ l'implementazione semiparametrica dell'impostazione bayesiana completamente predittiva.

3.1 Il Processo di Dirichlet e il Bootstrap Bayesiano

Sia $(\mathscr{P},A)$ uno spazio misurabile e sia α una misura finita non nulla ($\alpha(\mathscr{P}) < \infty$) su $(\mathscr{P},A)$. F è un processo di Dirichlet di parametro α se per qualunque partizione di $\mathscr{P}$ la distribuzione casuale $(F(A_1),\cdots,F(A_N))$ è una $\mathscr{D}(\alpha(A_1),\cdots,\alpha(A_N))$:

$$\forall \{A_i\}_{i=1}^N (F(A_1),F(A_2),\cdots,F(A_N)) \sim \mathscr{D}(\alpha(A_1),\alpha(A_2),\cdots,\alpha(A_N))$$

Ferguson (1973) mostra che se $\mathbf{x}$ è un campione di N realizzazioni di $F \sim DP(\alpha)$, allora a posteriori $F\,|\,\mathbf{x} \sim DP(\alpha + \sum \delta_x)$:

$$(F(A_1)\cdots,F(A_N)\,|\,\mathbf{x}) \sim \mathscr{D}\left(\alpha(A_1)+\sum\delta_x(A_1),\cdots,\alpha(A_N)+\sum\delta_x(A_N)\right)$$

ove $\delta_x(A_i)$ indica una delta di Dirac e $\sum\delta_x(A_i)$ la frequenza assoluta di realizzazioni che hanno assunto valori appartenenti all'insieme A_i. I valori attesi marginali a posteriori esprimono le $\pi(x^+\in A_i\,|\mathbf{x},\alpha)$ e sono combinazione convessa delle medie a priori e delle frequenze relative osservate:

$$E_F\left(F(A_i)\,|\,\mathbf{x},\alpha\right)=\frac{\alpha(A_i)+\sum\delta_x(A_i)}{\alpha\left(\mathscr{P}\right)+N}=\frac{\alpha(A_i)}{\alpha\left(\mathscr{P}\right)}\frac{\alpha\left(\mathscr{P}\right)}{\alpha\left(\mathscr{P}\right)+N}+F_N(A_i)\frac{N}{\alpha\left(\mathscr{P}\right)+N}$$

Analogamente, la predittiva coincide con la media a posteriori del processo:

$$\pi(x^+\,|\mathbf{x},\alpha)=E\left(F\,|\,\mathbf{x},\alpha\right)=\frac{\alpha+F_N}{\alpha\left(\mathscr{P}\right)+N}$$

Il secondo membro della precedente relazione evidenzia come la predittiva sia somma delle probabilità che la modalità del carattere manifestata da una "futura" osservazione x^+ sia già inclusa nel campione $\mathbf{x}$ ovvero non lo sia:

$$\pi(x^+=x_i\in\mathbf{x}\,|\mathbf{x},\alpha)=\frac{F_N}{\alpha\left(\mathscr{P}\right)+N},\qquad \pi(x^+=x_i\notin\mathbf{x}\,|\mathbf{x},\alpha)=\frac{\alpha}{\alpha\left(\mathscr{P}\right)+N}$$

In termini generali, si usa specificare $\alpha(\cdot)=kF_0(\cdot)$ ove k viene detto parametro di concentrazione e F_0 legge base. k è una misura di precisione e descrive la concentrazione di F attorno a F_0. Per $\frac{k}{N}\to 0$ risulta trascurabile il ruolo di F_0. Si noti che la distribuzione predittiva $\pi(x^+\,|\mathbf{x},\alpha)$ non dipende dall'ordine con il quale si sono manifestate le realizzazioni della sequenza $\mathbf{x}$. Per il teorema di rappresentazione (de Finetti, 1937) la scambiabilità implica che condizionatamente a $F\sim DP\left(kF_0\right)$ gli elementi di $\{X_N\}$ siano *i.i.d.* con funzione di ripartizione F. In ambito predittivo, $\forall\mathscr{B}\in\mathbb{R}$, è solitamente oggetto di interesse la $\pi(x_i\in\mathscr{B}\,|\mathbf{x})=E\left[F\left(\mathscr{B}\right)|\,\mathbf{x}\right]$. Tale quantità, in ottica frequentista viene approssimata con la funzione di ripartizione empirica $F_N\left(\mathscr{B}\right)$, ossia ponendo $E\left[F\left(\mathscr{B}\right)|\,\mathbf{x}\right]=\frac{1}{N}Bin\left(N,F_N\left(\mathscr{B}\right)\right)$. Secondo un'impostazione Bayesiana $F\left(\mathscr{B}\right)$ è una funzione di ripartizione a posteriori modellata secondo un processo di Dirichlet $DP\left(kF_0\right)$. A questo proposito, vale il teorema (Regazzini 1978):

$$\forall N,\mathscr{B}\quad \pi\left[x_{N+1}\in\mathscr{B}\,|\mathbf{x}\right]=\frac{k}{N+k}F_0\left(\mathscr{B}\right)+\frac{N}{N+k}F_N\left(\mathscr{B}\right)\Leftrightarrow F\sim DP\left(kF_0\right)$$

Per k=0, la posteriori coincide con quella ricavata nel bootstrap di Rubin (1981); se il campione osservato, costituito di n osservazioni, contiene m distinti valori, e si adotta per la verosimiglianza riferita alle rispettive probabilità di realizzazione un modello multinomiale si ottiene:

$$L\left(\mathbf{n}\,|\mathbf{p}\right)=\frac{N!}{\prod\left(n_j!\right)}\prod p_j^{n_j},\ j=1,\cdots,m\ ,\ \sum n_j=N$$

La legge a priori coniugata è la Dirichlet

$$\pi\left(\mathbf{p};\alpha\right)=\frac{\Gamma\left(\sum\alpha_j\right)}{\prod\left[\Gamma\left(\alpha_j\right)\right]}\prod p_j^{\alpha_j-1},\ \alpha_j>0\,,\,p_j>0\,,\,\sum p_j=1$$

La distribuzione a posteriori sarà quindi:

$$\pi(\mathbf{p}\,|\mathbf{n}\,;\nu) \propto \prod p_j^{\nu_j-1}, \nu_j > 0\,, p_j > 0\,, \sum p_j = 1\,, \nu_j = n_j + \alpha_j$$

Decidendo di adottare una priori impropria[1] con $\alpha_j = 0 \;\; \forall j$ si ricava:

$$\pi(\mathbf{p}\,|\mathbf{n}\,;\nu) \propto \prod p_j^{\nu_j-1}\,, \nu_j = n_j$$

Le marginali sono delle $Beta\left(n_j, N-n_j\right)$ il cui valore atteso coincide con la frequenza relativa osservata $\frac{n_j}{N}$. Muliere e Secchi (1994, 1996) propongono una tecnica per l'implementazione del bootstrap con priori propria quando si voglia approssimare la condizionata $L(T(F)\,|\mathbf{x})$ ove $F \sim DP\,(kF_0)$, posto che sia

$$T(F) = \begin{cases} inf\,\{t \in \mathbb{R} : F(t) \geq q,\, q \in [0,1]\} \\ \int hdF \end{cases}$$

con le funzioni h a valori reali, limitate e continue q.c. rispetto alla misura di probabilità indotta da F_0. Se F_0 non è discreta e con supporto finito, la posteriori F si può discretizzare costruendo la funzione di ripartizione empirica F^* e, ricavate le probabilità p_j^* associate ai $j = 1, 2, \cdots, m$ punti del supporto di F^*, calcolare in corrispondenza di essi i valori assunti dal funzionale $T(F^*)$. $T(F^*)$ converge debolmente (Davidson, 1994) a $T\,(F)$ al divergere di m. Indicando con $d[A,B]$ la distanza di Lévy[2] tra due funzioni di ripartizione A e B, si dimostra che:

$$\lim_{m\to\infty} d\,[L(T(F^*)\,|\mathbf{x}), L(T(F)\,|\mathbf{x})] = 0$$

Operativamente, la procedura proposta da Muliere e Secchi può essere riepilogata nei seguenti passi:

[1] Si tratta della priori di Haldane. Una priori propria uniforme è stata proposta da Ericson:

$$\begin{cases} \forall j:\; a_j = \varepsilon_j \\ \sum \varepsilon_j = o(1) \end{cases}$$

Previa costruzione di una griglia di valori lungo il range osservato, tali parametri permettono di definire una Dirichlet utile a probabilizzare le intensità assenti dal campione. Simulazioni condotte da Aitkin (2008) la dimostrano dominata dalla priori di Haldane sotto il profilo della copertura effettiva degli intervalli di credibilità per la media.

[2] Una sequenza di misure $\{A_n\}$converge debolmente al limite A se $\exists N$ e un intorno $\Im\,(A)$ tale che $\forall n > N\; A_n \in \Im\,(A)$.

Date due funzioni di ripartizione A e B aventi supporto in $\mathbb{R}$, sia:

$$d[A,B] \equiv inf\,\{\varepsilon > 0\,|\forall x \in \mathbb{R}, A\,(x-\varepsilon) - \varepsilon \leq B\,(x) \leq A\,(x+\varepsilon) + \varepsilon\,\}$$

- ▷ $d[A,B]$ definisce la metrica di Lévy,
- ▷ se $\{A_n\}$ è una sequenza di misure, essa converge debolmente ad A se e solo se $d[A_n, A] \to 0$

- ▷ da $F \sim DP(kF_0)$ si generano $j = 1, 2, \cdots, m$ punti (idealmente) appartenenti al supporto di F^*, approssimazione discreta di F,
- ▷ si generano le $V_j \equiv Gamma((N+k)/m, 1)$ e il vettore $f = \left(\frac{V_1}{\sum V_j}, \cdots, \frac{V_r}{\sum V_j}\right)^t$ approssima una realizzazione di F,
- ▷ si calcola il valore assunto dal funzionale $T(F)$ ossia $t = T(f)$,
- ▷ iterando la procedura un elevato numero di volte si ricava la distribuzione empirica idonea ad approssimare la $L(T(F)|\mathbf{x})$.

La sostanziale differenza rispetto al boostrap di Rubin risiede nell'utilizzo di una priori informativa F_0. In particolare, volendo ottenere la posteriori di una legge di probabilità, la priori è definita su un generico spazio misurabile $(\mathscr{P}, A)$.

3.2 Distribuzioni Base Vincolate

L'approccio bayesiano non parametrico fondato sui processi di Dirichlet è giustificabile alla luce delle seguenti osservazioni:

- ▷ l'approssimazione della distribuzione dei dati con una multinomiale esprime la necessità di ridurre il più possibile l'impatto delle informazioni sulle quali il ricercatore è incerto,
- ▷ l'uso di priori informative consente di superare il limite rappresentato dall'attribuzione di probabilità nulle alle intensità non osservate nel campione; gli strumenti utili a veicolare le informazioni sono la legge base F_0 ed il grado di fiducia k su di essa, costituenti il parametro del processo di Dirichlet $F \sim DP(kF_0)$.

Sembra interessante indagare una diversa possibilità di implementazione rispetto a quella presentata da Muliere e Secchi, consistente nell'imporre alla legge base il rispetto di prefissate condizioni di ortogonalità. In generale, l'interpretazione delle condizioni tipiche del *GMM* quali informazioni ausiliarie giustifica la selezione della legge, tra quelle coerenti con i vincoli $\mathbf{g}(x_i, \theta)$, meno discordante rispetto alla F_0 prescelta; condizionatamente a θ, essa permette di ottenere l'equivalente bayesiano del bootstrap intenzionalmente distorto (Hall e Presnell, 1999).

Sia $E[\mathbf{g}(x, \theta)] = \mathbf{0}$; noti i parametri θ caratterizzanti le condizioni $\mathbf{g}(x_i, \theta)$, si vogliano ponderare le osservazioni campionarie tramite pesi $p_i \sim DP(kF_0)$ tali che

$\sum p_i \mathbf{g}(x_i, \theta) = \mathbf{0}$. Se $V_i \sim Gamma(\nu_i, 1)$, condizionatamente a θ, $\mathbf{p}(\theta) = \left(\frac{V_1}{\sum V_i}, \ldots, \frac{V_N}{\sum V_i}\right)$. Poiché $E(V_i) = \nu_i$, sembrano significative le seguenti scelte:

$$\nu_i = \begin{cases} f_0^*(x_i) = f_0(x_i)\dfrac{1}{(1-\lambda^t \mathbf{g}_i)} \\ f_0^*(x_i) = f_0(x_i)e^{\lambda^t \mathbf{g}_i} \end{cases}$$

Sono così recuperati i risultati ottenuti nell'impostazione classica minimizzando le D_{KL}^F e D_{KL}^B, corrispondenti rispettivamente ad EL ed ET. La consistenza delle stime è garantita per $k = o(N)$ dato che per la posteriori vale:

$$F = \frac{k}{N+k} F_0^*(\mathscr{B}) + \frac{N}{N+k} F_N(\mathscr{B})$$

Se $f_0(\mathbf{x})$ è uniforme e si adottano condizioni $\mathbf{g}(x_i, \mu)$ sui momenti (ammettendo per ipotesi che essi siano noti), al crescere del numero di vincoli si approssima la densità o distribuzione sulle osservazioni e l'esito dell'esperimento tende, per ogni fissato k, a quello che si manifesterebbe scegliendo come priori la "vera" legge di probabilità.

Muliere e Secchi (1996) confrontano il bootstrap di Rubin e quello con priori propria eseguendo un esperimento: generano N osservazioni da una Gamma(20, 1), date diverse distribuzioni base calcolano dalle posteriori i tre quartili, e mediando i valori ottenuti nelle replicazioni studiano il comportamento delle stime al variare del parametro di precisione k. L'esercizio viene qui riproposto in riferimento ad un campione osservato di ampiezza N=25, usando come priori la Gamma(20, 1) ed effettuando 400 replicazioni. Dalla F in ogni replicazione vengono generati 500 punti del supporto ai quali sono associati per 400 cicli di bootstrap i pesi della Dirichlet, in modo da ottenere ad ogni ciclo la stima di ciascun quartile. I risultati delle simulazioni sono riepilogati nella tabella (3.2) e sembrano riprodurre gli esiti illustrati nell'articolo, ossia un graduale avvicinamento delle stime ai quartili della legge di probabilità, esposti nella tabella (3.1), all'aumentare del peso attribuito alla densità base (coincidente proprio con la legge che ha generato le osservazioni).

q_{25}	q_{50}	q_{75}
16.83	19.67	22.81

Tabella 3.1: Quartili della Gamma(20,1)

A completamento dell'esercizio, per ciascun quartile, sono stati calcolati l'ampiezza dell'intervallo *h.p.d.* - il codice R è stato mutuato da Liseo (2006) - corrispondente ad un livello nominale di significatività pari ad $\alpha = 0.05$ e la relativa copertura ef-

k	q_{25}	q_{50}	q_{75}
0.000	16.885	19.606	22.721
1.000	16.882	19.611	22.729
5.000	16.874	19.614	22.740
10.000	16.867	19.623	22.756
25.000	16.848	19.633	22.768

Tabella 3.2: Stime dei quartili al variare di k

fettiva; dalla tabella (3.3) si può notare che al crescere del peso attribuito alla priori, come è lecito attendersi a causa della maggiore quantità di informazione introdotta nelle stime, l'estensione degli intervalli di credibilità si riduca.

k	q_{25}	q_{50}	q_{75}
0.000	4.751	4.895	5.995
1.000	4.721	4.819	5.896
5.000	4.377	4.480	5.485
10.000	3.995	4.096	5.036
25.000	3.316	3.389	4.091

Tabella 3.3: Ampiezze degli intervalli h.p.d. per $\alpha = 0.05$

La copertura effettiva, riportata nella tabella (3.4), eccede quella nominale e mostra un andamento di segno opposto, crescendo all'aumentare del parametro di concentrazione k.

k	q_{25}	q_{50}	q_{75}
0.000	0.952	0.962	0.948
1.000	0.960	0.972	0.962
5.000	0.970	0.972	0.962
10.000	0.978	0.988	0.985
25.000	0.995	0.998	0.998

Tabella 3.4: Copertura degli intervalli h.p.d. per $\alpha = 0.05$

Sulla base di quanto anticipato, un esercizio analogo può essere condotto sostituendo alla priori Gamma(20, 1) le *GEL* dedotte a partire da condizioni sui momenti della legge che ha generato i dati. Mantenendo i parametri dell'esercizio fin qui svolto, sono state considerate due distribuzioni base, riferite rispettivamente all'inclusione dei primi j=2, 4 momenti. I diagrammi della figura 3.1 permettono un confronto im-

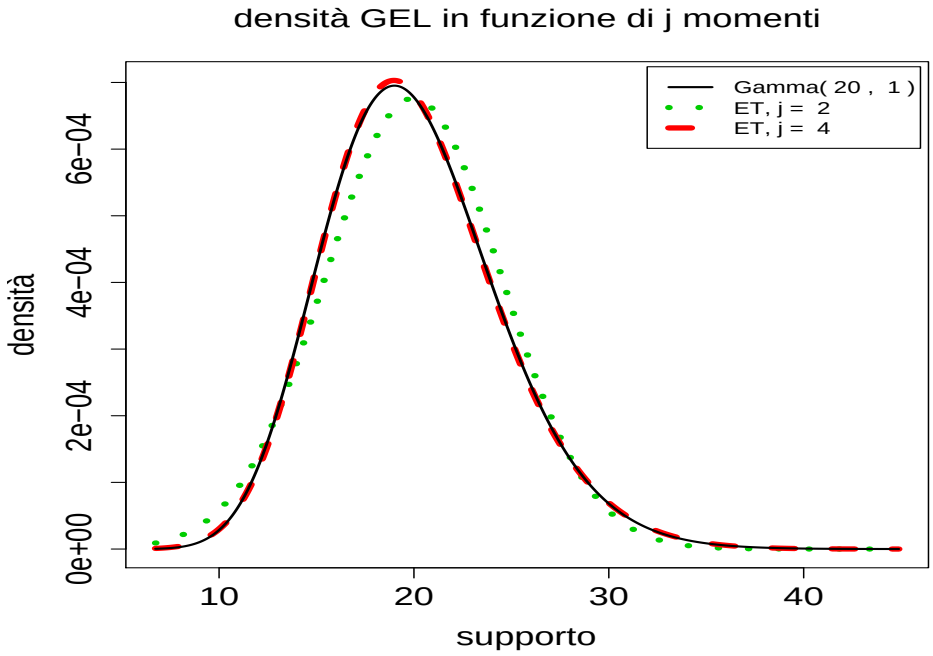


Figura 3.1: Densità e relative approssimazioni

mediato tra la densità Gamma e le approssimazioni ottenute per questa via tramite l'ampliamento esponenziale (ossia la densità di massima entropia). La conoscenza dei momenti può essere considerata in molte applicazioni poco realistica e verrà tra breve rilassata, ma è utile per apprezzare l'impatto del diverso volume di conoscenze a disposizione del ricercatore: esse non riguarderebbero più la legge che ha generato i dati, ma soltanto alcune sue caratteristiche. I risultati, in conformità alle premesse, sono qualitativamente comparabili ai precedenti e mostrano un tendenziale avvicinamento delle stime ai valori dei quartili, all'aumentare dei momenti inclusi nelle condizioni di ortogonalità.

	j=2			**j=4**		
k	q_{25}	q_{50}	q_{75}	q_{25}	q_{50}	q_{75}
1.000	16.930	19.722	22.831	16.928	19.703	22.823
5.000	16.923	19.747	22.855	16.912	19.696	22.814
10.000	16.927	19.775	22.865	16.901	19.688	22.807
25.000	16.924	19.828	22.902	16.875	19.672	22.792

Tabella 3.5: Base GEL: stime dei quartili

L'ampiezza degli intervalli *h.p.d.* sembra testimoniare, secondo le attese, che ad un aumento del peso della priori corrisponda una diminuzione nell'incertezza concer-

	j=2			j=4		
k	q_{25}	q_{50}	q_{75}	q_{25}	q_{50}	q_{75}
1.000	4.737	4.849	6.210	4.723	4.797	6.176
5.000	4.489	4.510	5.610	4.371	4.468	5.669
10.000	4.165	4.113	5.002	4.018	4.118	5.200
25.000	3.508	3.439	3.981	3.295	3.362	4.158

Tabella 3.6: Base GEL: ampiezze degli intervalli h.p.d. per $\alpha = 0.05$

nente i valori stimati. Le coperture degli intervalli, tabella (3.7), sono sostanzialmente analoghe a quelle ottenute usando come priori la legge che ha generato i campioni.

	j=2			j=4		
k	q_{25}	q_{50}	q_{75}	q_{25}	q_{50}	q_{75}
1.000	0.965	0.975	0.942	0.970	0.960	0.952
5.000	0.975	0.970	0.975	0.970	0.970	0.958
10.000	0.980	0.975	0.978	0.978	0.978	0.972
25.000	0.990	0.995	0.995	0.992	0.995	0.990

Tabella 3.7: Base GEL: coperture degli intervalli h.p.d. per $\alpha = 0.05$

Se i momenti non sono noti, possono essere pensati come punti dello spazio parametrico ed essere stimati (impostazione classica), ovvero essere considerati anch'essi delle variabili casuali descritte da una opportuna legge multivariata (ambito bayesiano). Prima di procedere è opportuno sintetizzare gli elementi emersi dalle precedenti considerazioni :

- ▷ per ogni θ di Θ il valore calcolato per λ estrae dall'universo delle distribuzioni (discrete) che soddisfano $E[\mathbf{g}(x,\theta)] = \mathbf{0}$ quella che minimizza la discrepanza di Cressie-Read di parametro γ: al variare di θ i valori della *GEL* e della funzione di danno $\lambda^t \mathbf{g}$ sono quelli corrispondenti all'inviluppo delle distribuzioni ottime,

- ▷ l'uso delle *GEL* quali F_0 è stato discusso condizionatamente a θ. In questo ambito risulta mitigato l'inconveniente, talvolta imputato al bootstrap bayesiano, di non scontare nel processo di stima le eventuali conoscenze a priori (circa i parametri oggetto di indagine) di cui il ricercatore potrebbe disporre,

- ▷ al variare di θ la F_0 ottenuta come *GEL* descrive una famiglia di distribuzioni base; benché sia possibile pensarne l'utilizzo in funzione dei processi di Dirichlet, sembra altrettanto interessante l'alternativa rappresentata dal ricorso a

modelli semiparametrici; prima di affrontare l'argomento possono risultare utili alcune considerazioni introduttive.

3.2.1 Selezione dei Parametri

Ai fini delle applicazioni, è necessario precisare

- ▷ il parametro γ che definisce la discrepanza empirica da minimizzare,
- ▷ il numero di momenti da includere nell'approssimazione della legge di probabilità che genera le osservazioni.

Per quanto riguarda il primo punto, una scelta ragionevole - nell'ambito dell'impostazione bayesiana - sarebbe $\gamma = -1$. Essa definisce le verosimiglianze empiriche che, è stato evidenziato nel cap. 2, forniscono mode a posteriori consistenti in *L-divergenza*, essendo le posteriori costruite a partire dalle logverosimiglianze massimizzate in modo da eliminare i parametri di disturbo rappresentati dai moltiplicatori. Tuttavia, tra le discrepanze empiriche di Cressie-Read, quella ottenuta da $\gamma = 0$, sembra essere maggiormente significativa. Essa è proporzionale alla divergenza di Kullback-Leibler e la sua minimizzazione conduce a leggi di massima entropia. L'utilizzo di queste ultime appare giustificabile sotto diversi punti di vista.

- ▷ La semplificazione computazionale risiede nell'ottenimento di pesi sempre non negativi, circostanza che salvo accorgimenti ad hoc *(Kitamura, 2006)*, non è garantita in piccoli campioni quando si faccia uso delle verosimiglianze empiriche. Le stime ottenute non sono più di massima verosimiglianza ma per costruzione limitano l'impatto, nel processo di stima, degli elementi estranei al set informativo del quale il ricercatore sia ragionevolmente certo.
- ▷ Un motivo che può incoraggiare la scelta di $\gamma = 0$, consiste nella possibilità di derivare le distribuzioni di massima entropia da considerazioni simili a quelle che conducono all'approssimazione del punto di sella (Reid, 1988). Una congrua trattazione delle espansioni asintotiche riferite alle funzioni di verosimiglianza eccede il contenuto delle riflessioni in discorso e sarà sufficiente, per comodità di esposizione, solo qualche sommario riferimento ai termini del problema. L'approssimazione viene solitamente impiegata per ottenere stime più accurate di quelle fondate, via *TCL*, sulla normalità asintotica circa la densità di somme (o funzioni di somme) di osservazioni *i.i.d.* Nel caso di $\bar{x}$, indicando con $K(\upsilon)$ la funzione generatrice dei cumulanti:

$$L(\upsilon) = e^{\upsilon N\bar{x} - NK(\upsilon)} p(\bar{x})$$

da cui si trae

$$\hat{p}(\bar{x}) = e^{NK(\upsilon) - \upsilon N\bar{x}} \hat{L}(\upsilon)$$

$\hat{L}(\upsilon)$ viene ottenuto da uno sviluppo di Edgeworth utilizzando υ per "centrarlo" in corrispondenza del valore $\bar{x}$ rispetto al quale si desidera calcolare la densità; in tal modo, per qualsiasi valore di $\bar{x}$, vale l'accuratezza dello sviluppo per regioni prossime alla media. Imponendo $\int \hat{p}(\bar{x}) d\bar{x} = 1$ e arrestando lo sviluppo al terzo ordine,

$$\hat{p}(\bar{x}) = p(\bar{x}) \left[1 + O\left(N^{-3/2} \right) \right]$$

Sono possibili alcune osservazioni:

- ▷ l'ampliamento esponenziale utilizza la vera $K(\upsilon)$; spesso, nelle applicazioni, la legge seguita dal processo generatore dei dati è ignota e la funzione generatrice dei cumulanti può essere solo stimata;
- ▷ la densità approssimata concerne funzioni di somme di osservazioni, circostanza che riduce l'eterogeneità dei fenomeni indagati.
- ▷ Usando i momenti campionari, ossia la funzione generatrice dei cumulanti empirica, e considerando la relazione tra $\sigma(x)$ e $\sigma(\bar{x})$ per dati *i.i.d.*, sarebbe lecito attendersi

  $$\hat{p}(x) = p(x) \left[1 + O_p\left(N^{-1/2} \right) \right] \tag{3.1}$$

  Rinviando i dettagli della dimostrazione all'appendice, ed indicando le condizioni di ortogonalità concernenti i momenti con il vettore $\mathbf{g}(x)$, nel caso di osservazioni *i.i.d.* è proprio quanto accade definendo:

  $$\widehat{p}(x) = e^{\lambda^t \mathbf{g}(x) - \widehat{K}(\lambda)} \frac{1}{N}$$

  In altri termini, gli errori di approssimazione dipendenti dall'uso dei momenti campionari sono espressi in termini relativi. Nel caso bayesiano, per l'ottenimento della predittiva $p(x^+ | x)$ i momenti verrebbero selezionati da una posteriori ricavata numericamente dalla posteriori sulle unità statistiche incluse nel campione (una Dirichlet con priori non informativa, nel caso del bootstra*p* di Rubin, ossia tale da implicare pesi con un valore atteso a posteriori pari ad N^{-1} per ciascuna delle N unità statistiche incluse nel campione).

▷ Si può dimostrare (all'argomento è dedicato il paragrafo A.3 in appendice)

che definendo opportunamente il supporto delle leggi di probabilità a priori, causa la convergenza delle verosimiglianze empiriche generalizzate ad un comune limite, la consistenza in *L-divergenza* delle posterior*i* si realizza anche minimizzando le discrepanze di Cressie-Read di parametro $\gamma \neq -1$.

▷ Bernardo e Smith (2000) sottolineano come la modellazione di una verosimiglianza empirica secondo la famiglia esponenziale sia connaturata all'esigenza di ottenere la rappresentazione matematica della "vera" legge p e non della sua approssimazione f; come si è osservato nel capitolo 2, la discrepanza di parametro $\gamma = 0$ è proporzionale alla $D_{KL}(p, f)$:

$$D_{KL}(p, f) \equiv \int p(x) ln \frac{p(x)}{f(x)} dx$$

Il caso limite di approssimazione f vaga è rappresentato da una legge uniforme, e da essa discende quale stima di p la legge di massima entropia.

▷ L'ampliamento esponenziale è coerente con il risultato di un importante teorema di Cifarelli e Regazzini (1982). Rinviando al paragrafo 3.3 una breve sintesi di significati ed implicazioni del teorema di rappresentazione per successioni di variabili casuali scambiabili, i termini essenziali del ragionamento possono essere esplicitati secondo l'esposizione fornita da Muliere e Secchi (1992).

  ▷ Sia $T_N : \mathbb{R}^N \to \mathbb{R}$; la successione di statistiche $\{T_N(\mathbf{x})\}$ è sufficiente a fini predittivi se per $r \geq 1$ e $\{i_1, i_2, \cdots, i_r\} \cap \{1, 2, \cdots, N\} = \varnothing$

$$P(x_{i_1}, \cdots, x_{i_r} | x_1, \cdots, x_N) = P(x_{i_1}, \cdots, x_{i_r} | T_N)$$

  ▷ Sia $\{X_N\}$ una sequenza scambiabile di variabili casuali per la quale $\{T_N\}$ costituisca una sequenza di statistiche sufficienti a fini predittivi:

$$T_N(\mathbf{x}) = \frac{1}{N} \sum a(x_i), \quad a : \mathbb{R} \to \mathbb{R}$$

  Se il supporto della legge di probabilità non dipende da T_N, allora esiste una variabile casuale reale Υ condizionatamente alla quale le osservazioni sono indipendenti e somiglianti con legge esponenziale:

$$p(x | \Upsilon = \upsilon) = e^{\upsilon a(x) - K(\upsilon) + b(x)}$$

  ove $K(\upsilon)$ è la funzione generatrice dei cumulanti condizionata a $\Upsilon = \upsilon$. Ponendo $K'(\Upsilon) \equiv \Theta$, da $K'(\upsilon) = E[a(x) | \Upsilon = \upsilon] = \theta$ si ricava $\upsilon =$

$(K')^{-1}(\theta) \equiv d'(\theta)$, da cui

$$p(x\,|\Theta = \theta\,) = e^{d'(\theta)a(x) - \theta d'(\theta) + d(\theta) + b(x)} \tag{3.2}$$

Il vantaggio della riparametrizzazione sta nella possibilità di applicare immediatamente la legge forte dei grandi numeri per sequenze di variabili casuali scambiabili dovuta a de Finetti:

$$P\left\{ \lim_{N\to\infty} T_N(\mathbf{x}) = \Theta \right\} = 1$$

In parole, se il contenuto informativo di una successione di variabili casuali scambiabili è riassumibile per mezzo di una statistica sufficiente a fini predittivi, il limite al quale essa converge quasi certamente è il parametro condizionatamente al quale le osservazioni sono indipendenti e somiglianti con legge di probabilità appartenente alla famiglia esponenziale. Cambiando quel che si deve cambiare in riferimento al framework in discussione, siano:

$$\begin{cases} \mathbf{a}(x) = \left[x,\, x^2,\, \cdots,\, x^q\right] \\ T_N(\mathbf{x}) = \left[\frac{1}{N}\sum x_i,\, \frac{1}{N}\sum x_i^2,\, \cdots,\, \frac{1}{N}\sum x_i^q\right] \\ \theta = \left[\mu_1,\, \mu_2,\, \cdots,\, \mu_q\right] \\ d(\theta) = -ln \int_A \left[e^{\sum_j \upsilon_j (x^j - \mu_j) + b(x)}\right] \\ A = \left\{\mathbf{x} :\, \forall j \in [1, \cdots, q],\, \frac{1}{N}\mathbf{1}^t \mathbf{x}^j = m_j\right\} \end{cases}$$

Allora, si ricava $d'(\theta) = \left[\upsilon_1,\, \upsilon_2,\, \cdots,\, \upsilon_q\right]$ e per $\forall x \in A$ (insieme delle N^{ple}che forniscono i momenti empirici osservati), la (3.2) può essere espressa come:

$$p(x\,|\Theta = \mu\,) = e^{\sum \upsilon_j x^j - \sum \upsilon_j \mu_j + b(x) - ln \int_A \left[e^{\sum_j \upsilon_j (x^j - \mu_j) + b(x)}\right] dx}$$

In riferimento alla variabile casuale N^{pla} descritta dal campione la legge diviene

$$p(\mathbf{x}\,|\Theta = \mu\,) = e^{N\sum \upsilon_j m_j - N\sum \upsilon_j \mu_j + \sum b(x) - Nln \int_A \left[e^{\sum_j \upsilon_j (x^j - \mu_j) + b(x)}\right] dx}$$

Da essa è immediato ricavare la parametrizzazione equivalente

$$p(\mathbf{x}\,|\Upsilon = \upsilon\,) = e^{N\sum \upsilon_j m_j - Nln \int_A \left[e^{\sum_j \upsilon_j x^j + b(x)}\right] dx} e^{\sum b(x)}$$

Ricordando il criterio di fattorizzazione di Neyman, quest'ultima relazione evidenzia il ruolo dei momenti campionari quali statistiche sufficienti.

Circa il secondo problema, la distanza tra due distribuzioni di probabilità che hanno in comune i primi k momenti, misurata in corrispondenza di un determinato punto del supporto, ammette un estremo superiore esprimibile come reciproco di una opportuna forma quadratica (Lindsay e Basak, 2000) $|\mathscr{H}_k(z)|^2$.

In particolare, valgono le seguenti relazioni:

$$\begin{cases} |\mathscr{H}_k(x_i)|^2 = \left(1, x_i, \cdots, x_i^k\right) \begin{pmatrix} 1 & \mu_1 & \cdots & \mu_k \\ \mu_1 & \mu_2 & \cdots & \mu_{k+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_k & \mu_{k+1} & \cdots & \mu_{2k} \end{pmatrix} \begin{pmatrix} 1 \\ x_i \\ \vdots \\ x_i^k \end{pmatrix} \\ sup\,|\delta_i| \equiv sup\,|p_i - \widetilde{p}_i| = \left(|\mathscr{H}_k(x_i)|^2\right)^{-1} \end{cases}$$

Dunque, se $\widetilde{\mathbf{p}}$ e $\mathbf{p}$ sono due distribuzioni caratterizzate dall'avere in comune i primi k momenti, la (3.1) diviene $\widehat{p}_i = \widetilde{p}_i \left[1 + O_p(N^{-1/2})\right]$ e da $\widetilde{p}_i = p_i + \delta_i$ è immediato definire l'errore di approssimazione ε_i:

$$\varepsilon_i \equiv \widehat{p}_i - p_i = p_i O_p\left(N^{-1/2}\right) + \delta_i \left[1 + O_p\left(N^{-1/2}\right)\right]$$

Lindsay e Basak mostrano che l'estremo superiore di $|\delta_i|$ è informativo soltanto per valori di x_i distanti dalla parte centrale della distribuzione e decresce tanto più velocemente quanto maggiore è il numero j di momenti. Questi risultati evidenziano che il contributo dei momenti di ordine più elevato concerne la definizione delle code della legge di probabilità, secondo una risoluzione sempre più precisa via via che ne vengono aggiunti, ma con "guadagni" decrescenti, circostanza che a ben vedere costituisce una giustificazione all'uso del metodo dei momenti. Dunque risulta garantita l'accuratezza delle stime in corrispondenza delle code della "vera" (ed ignota) distribuzione.

Qualitativamente, in riferimento a leggi che ammettano l'esistenza di momenti finiti di ordine arbitrariamente elevato, all'aumentare della numerosità campionaria sarebbe consigliabile incrementare il numero di condizioni di ortogonalità in modo che esso si mantenga $o(N)$: la dimensione crescente del campione permetterebbe di avere un numero di osservazioni divergente rispetto al numero dei momenti rendendone possibili stime consistenti; in presenza di campioni con ampiezza finita, benchè sia numericamente possibile calcolarne di qualunque ordine, occorre tener conto della quantità

limitata di informazione che essi contengono; inoltre, gli incrementi dell'informazione introdotta nelle stime sono decrescenti anche a causa della maggiore instabilità dei momenti empirici di ordine più elevato, suscettibili di notevoli variazioni a fronte di minimi cambiamenti nelle intensità manifestate da poche unità statistiche. La soluzione più ovvia risiede nel ricorso a test delle ipotesi volti a vagliare l'identità di distribuzioni "approssimanti" che ne incorporino un diverso numero. Benché non sia dato sapere quanti momenti finiti ammetta il processo generatore dei dati, se le differenze tra due approssimazioni non sono significative è ragionevole ritenere che sia trascurabile il contributo informativo dei momenti eccedenti la rappresentazione più parsimoniosa. Per fissare le idee si considerino due distribuzioni che, in dipendenza del numero di momenti inclusi, $q-c$ e q, differiscano dalla "vera" legge di probabilità, in ogni punto del comune supporto, al più delle quantità δ_i e, rispettivamente, δ_i'. È facile verificare che i test di confronto tra distribuzioni empiriche si risolvono, date le premesse, in funzioni degli errori di approssimazione δ_i e δ_i'. Considerando $\widehat{p_i} = (p_i + \delta_i)\left[1 + O_p\left(N^{-1/2}\right)\right]$ si possono ad esempio ricordare

- ▷ il *LR*; di immediata applicabilità nel caso di osservazioni *i.i.d.*, si può esprimere come

$$\widehat{LR} = \frac{\prod \widehat{p_i'}}{\prod \widehat{p_i}} = \frac{\prod\left[1 + O_p\left(N^{-1/2}\right)\right]\prod(p_i + \delta_i')}{\prod\left[1 + O_p\left(N^{-1/2}\right)\right]\prod(p_i + \delta_i)}$$

  Si noti come la distribuzione al denominatore sia quella che include $q-c$ condizioni: via via che se ne aggiungono ulteriori ci si allontana, nella metrica indotta dalla discrepanza empirica di parametro γ, dalla distribuzione uniforme (o nel caso siano presenti delle intensità ripetute, dalla multinomiale), ossia dalla distribuzione che massimizza la produttoria dei pesi. Per $1 \leq h \leq c$, la $(q-c+h)^{ma}$ condizione di ortogonalità nella *GEL* che ne incorpora $q-c$ è rappresentata dall'assunzione che sia $\mu_{q-c+h} = \sum p(x_i; \mu_1, \ldots, \mu_{q-c}) x_i^{q-c+h}$ e per costruzione il relativo moltiplicatore υ_{q-c+h} è pari a zero; alla luce della precedente espressione, il *LR* può essere scritto come

$$\widehat{LR} = \left[1 + O_p\left(N^{-1/2}\right)\right]\left[1 + O_p\left(N^{-1/2}\right)\right]^{-1}\frac{\prod(p_i + \delta_i')}{\prod(p_i + \delta_i)}$$

  Poiché $\frac{1}{1+x} = 1 - x + x^2 - \cdots$, segue

$$\widehat{LR} = \left[1 + O_p\left(N^{-1/2}\right)\right]\frac{\prod(p_i + \delta_i')}{\prod(p_i + \delta_i)}$$

e passando ai logaritmi

$$ln(\widehat{LR}) = ln\left[1 + O_p\left(N^{-1/2}\right)\right] + \sum\left[ln(p_i + \delta_i') - ln(p_i + \delta_i)\right]$$

Poiché $ln(1+y) = y - \frac{1}{2}y^2 + \frac{1}{3}y^3 - \cdots$ è lecito scrivere

$$ln(\widehat{LR}) = O_p\left(N^{-1/2}\right) + \sum\left[ln(p_i + \delta_i') - ln(p_i + \delta_i)\right]$$

Dunque,

$$O_p\left(N^{-1/2}\right) - 2\sum\left[ln(p_i + \delta_i') - ln(p_i + \delta_i)\right] \xrightarrow{d} O_p\left(N^{-1/2}\right) + \chi_c^2$$

▷ il test *KS*; a fronte di caratteri continui, indicando con $p_{(i)}$ e $\delta_{(i)}$ rispettivamente $p\left[x_{(i)}\right]$ e $\delta\left[x_{(i)}\right]$, ossia i pesi e gli errori di approssimazione (legati al numero di momenti considerati) riferiti alle unità statistiche che occupano il rango i^{mo} nella graduatoria non decrescente delle intensità,

$$\sum \widehat{p}_{(i)}' - \sum \widehat{p}_{(i)} = \sum(\delta_{(i)}' - \delta_{(i)})\left[1 + O_p\left(N^{-1/2}\right)\right]$$

da cui è immediato ricavare la relazione

$$\max_i \left|\sum \widehat{p}_{(i)}' - \sum \widehat{p}_{(i)}\right| = \max_i \left|\sum(\delta_{(i)}' - \delta_{(i)})\right| \left|1 + O_p\left(N^{-1/2}\right)\right|$$

▷ per caratteri discreti o continui, una valida alternativa è costituita dal calcolo della discrepanza empirica di parametro $\gamma = 0$ tra le due leggi approssimanti, distribuita asintoticamente, sotto l'ipotesi nulla, come un χ_c^2. Rinviando all'appendice i calcoli utili a ricavare la legge di CR_0 per grandi campioni, essa è proporzionale alla divergenza di Kullback-Leibler e nel caso in esame risulta funzione di δ_i e δ_i':

$$2ND_{KL}\left(\widehat{\mathbf{p}}', \widehat{\mathbf{p}}\right) \equiv 2N\sum \widehat{p}_i' ln\frac{\widehat{p}_i'}{\widehat{p}_i} = 2N\sum\left(p_i + \delta_i'\right) ln\frac{p_i + \delta_i'}{p_i + \delta_i}\left[1 + O_p\left(N^{-1/2}\right)\right]$$

Strutturalmente, la considerazione del valore atteso per il logaritmo del rapporto tra probabilità rende (ai fini interpretativi) irrilevante l'ipotesi di indipendenza delle unità nel campione.

Incidentalmente, l'inclusione progressiva delle condizioni sui momenti offre l'ulteriore opportunità di governare la forma della predittiva qualora il ricercatore, sulla base di considerazioni circa la natura del fenomeno indagato, ritenga che i dati non possano provenire da una legge plurimodale; il vincolo sulla forma è perseguibile escludendo l'influenza dei momenti di ordine pari o superiore a quello responsabile della violazione; il programma di esempio riportato nell'appendice B prevede la possibilità di espletare questo controllo.

3.2.2 Un Esempio in Ottica Classica

Riprendendo l'esercizio di Muliere e Secchi , i parametri di forma e velocità (reciproco della scala) di una densità Gamma sono stati posti pari a 20 e, rispettivamente, 1. È parso interessante valutare la bontà dell'approssimazione ottenuta usando la verosimiglianza empirica generalizzata di parametro $\gamma = 0$ (densità di massima entropia o di ampliamento esponenziale, che nei diagrammi è stata sinteticamente indicata con l'acronimo inglese di Exponential Tilting), effettuando un confronto con l'omologa stima non parametrica di nucleo gaussiano. Sono state condotte quattrocento simulazioni, per numerosità campionarie comprese nel range [25, 250] con passi di ampiezza pari a venticinque unità, utilizzando il software R (R Development Core Team, 2008).

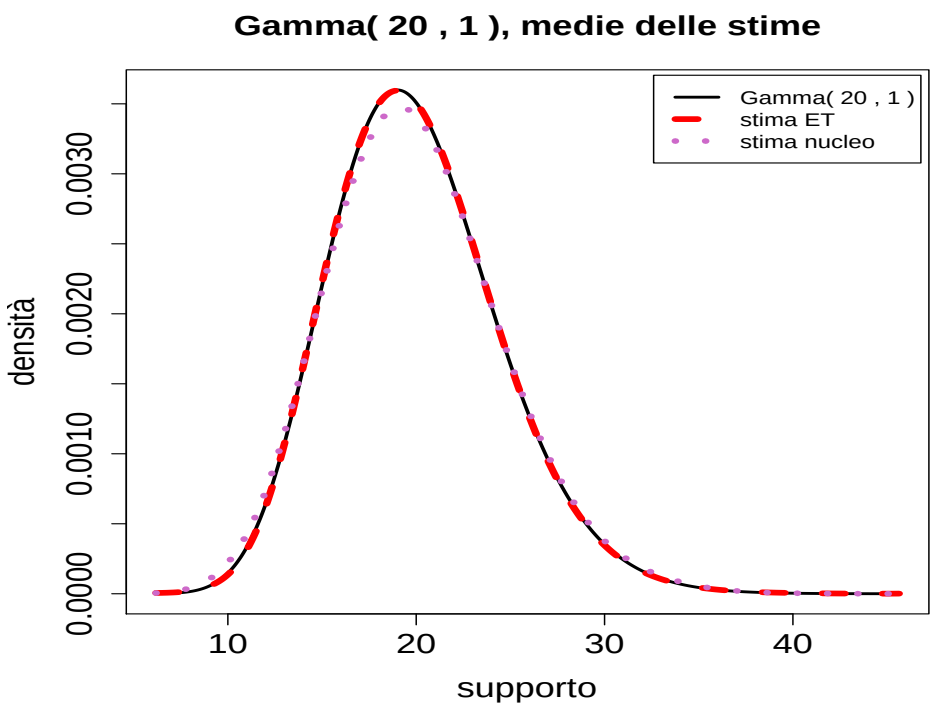


Figura 3.2: Confronto tra stime di una densità

Per entrambi gli stimatori, il supporto è stato confinato ai valori positivi e, in riferimento a quello nucleo, si è scelto di usare l'ampiezza di banda selezionata per default dalla funzione density del package stats. Ai fini della comparazione delle stime, si è ritenuto opportuno elidere l'influenza che sull'approssimazione di una funzione continua esercita il numero di punti del dominio per i quali viene calcolata, utilizzando il medesimo, tenuto fisso in tutte le replicazioni, per entrambe; la figura (3.2) è stata ricavata considerando le medie delle stime ottenute nelle quattrocento replicazioni, relative a campioni di duecentocinquanta unità, usando un supporto di mille punti e, per l'ampliamento esponenziale, almeno i primi quattro momenti campionari (l'inclusione di quelli di ordine superiore è stata decisa in ogni replicazione secondo il test della discrepanza di Cressie-Read di parametro $\gamma = 0$). La figura (3.3) mostra diverse misure riguardanti gli errori di approssimazione, espressi in valore assoluto, al variare della numerosità campionaria. Sono state considerate le medie su quattrocento replicazioni di un indice di dimensione e un indice di dispersione nonché il valore massimo. Per apprezzare l'andamento delle funzioni riferite ai due stimatori è stata inserita la curva di livello corrispondente ad N^{-2}.

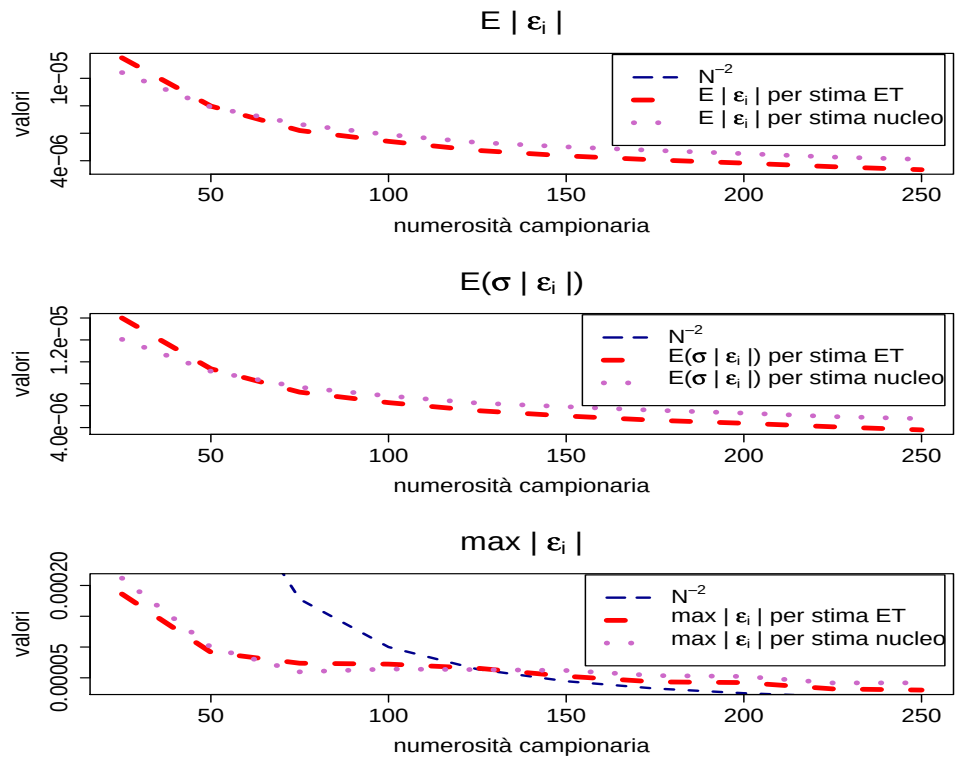


Figura 3.3: Errori in valore assoluto

Il caso in esame sembra suggerire che al divergere dell'ampiezza campionaria, causa

la crescente vicinanza dei momenti empirici alle omologhe quantità che caratterizzano la legge di probabilità, la qualità dell'informazione utilizzata per la stima della densità di massima entropia aumenti così come accade per il contenuto informativo, insito nella prossimità delle osservazioni, catturato dallo stimatore nucleo. Benché il confronto tra gli stimatori ecceda i propositi della ricerca, può essere interessante osservare che - almeno nel caso particolare oggetto dell'esemplificazione - anche per modeste numerosità campionarie l'informazione fornita dai momenti permette stime semiparametriche affatto deteriori rispetto a quelle offerte dallo stimatore nucleo. In altri termini, è conforme alle attese il comune andamento delle curve. Le conclusioni teoriche circa il valore assoluto del massimo errore di approssimazione implicato dall'uso della densità di massima entropia, sono state vagliate empiricamente approfittando della conoscenza concernente il vero processo generatore dei dati. Al variare dell'ampiezza campionaria, è stato individuato il segmento del supporto caratterizzato dal massimo errore ed è stato calcolato il prodotto tra la rispettiva probabilità, secondo la densità Gamma che ha generato i campioni, e $N^{-1/2}$. Le indicazioni che emergono, sintetizzate nella figura (3.4) hanno ovviamente solo valenza qualitativa, dato che le formule di riferimento riguardano genericamente gli ordini di grandezza degli errori. In questo senso, sembra rilevante l'andamento comune delle due curve piuttosto che la rispettiva posizione.

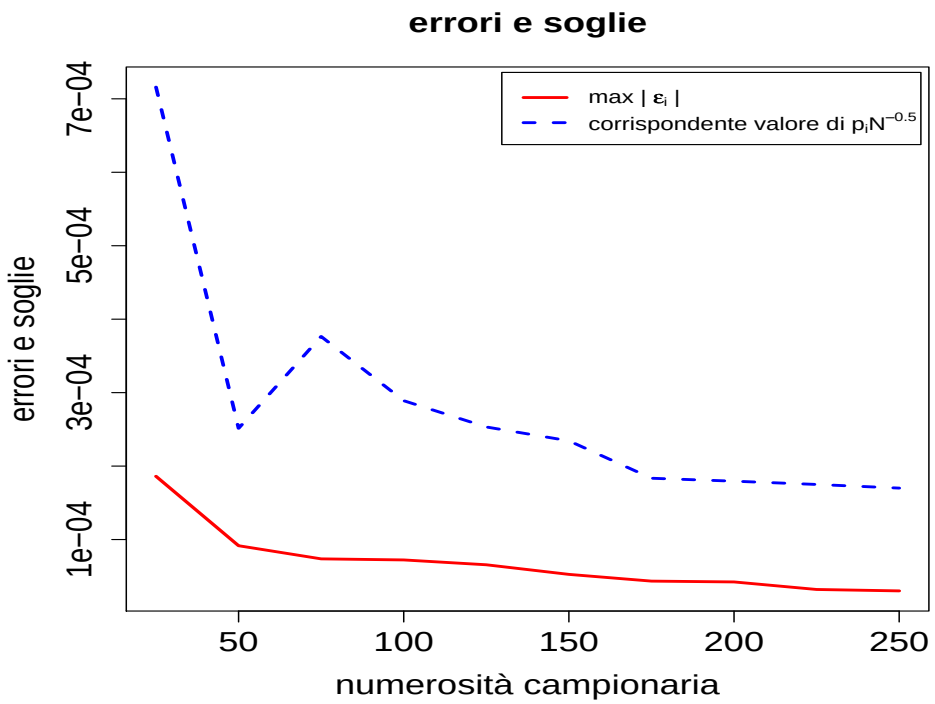


Figura 3.4: Soglie teoriche ed effettive

Dalla conoscenza della legge di probabilità che ha generato le osservazioni è stato possibile calcolare, tramite la funzione generatrice, i primi undici momenti che la caratterizzano, utili alla quantificazione - per ogni punto del supporto discretizzato - del limite superiore concernente la massima distanza (in valore assoluto) tra due distribuzioni che abbiano in comune i primi $j = 1, \cdots, 11$ momenti. La figura (3.5), se da un lato conferma che le diversità si attenuano al crescere del numero di momenti comuni e quanto più ci si avvicina alle code, dall'altro rende evidente che la quantità di informazione introdotta aumenta con tassi di variazione decrescenti; ad esempio, considerando i decrementi delle massime distanze che si notano passando da un certo numero di momenti comuni a quello immediatamente superiore, la diminuzione conseguita con j=2 è visibilmente superiore a quella ingenerata da j=11. Tale circostanza, se da un lato non è sorprendente, esprimendo il legame di dipendenza dei momenti di ordine superiore da quelli che li precedono, dall'altro è consolante in riferimento alla possibilità di ottenere approssimazioni affidabili della legge di probabilità sulle osservazioni facendo uso di un numero limitato di condizioni di ortogonalità.

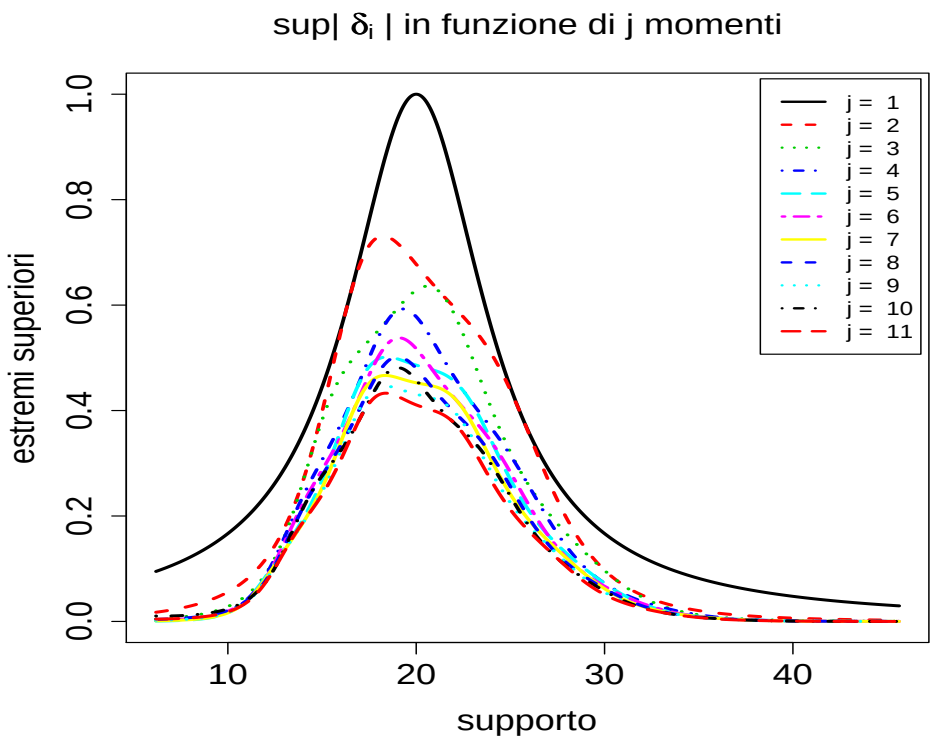


Figura 3.5: Estremi superiori per $|\delta_i|$

3.2.3 Smoothing del Bootstrap Bayesiano

Sulla base delle considerazioni svolte, sembra interessante vagliare la possibilità di ottenere una regolarizzazione delle stime bootstrap facendo uso dei momenti empirici. Pur mantenendo la costruzione proposta da Muliere e Secchi, in questo modo viene meno il ruolo delle conoscenze a priori (salvo che se ne vogliano introdurre di nuove facendo uso di opportune condizioni di ortogonalità): i momenti stimati consentono la costruzione di verosimiglianze empiriche generalizzate ma è assente informazione esterna a quella contenuta nel campione; la questione da indagare riguarda dunque il suo migliore sfruttamento (in termini relativi). In ciascuna delle 400 replicazioni, dalla Gamma(20,1) è stato generato un campione, definito il supporto (mille punti lungo il range ricavato aumentando gli estremi osservati di un ammontare pari a due volte la stima robusta della deviazione standard), calcolati i momenti e costruita la corrispondente densità di massima entropia; il numero di condizioni da utilizzare è stato selezionato tramite il test delle ipotesi fondato sulla distribuzione asintotica della discrepanza di Cressie-Read con parametro $\gamma = 0$. La risultante funzione di ripartizione ha costituito il termine di mistura con la funzione di ripartizione empirica. Invertendo la funzione di ripartizione mistura sono state generate 500 realizzazioni; esse hanno costituito il supporto per 400 cicli di bootstrap i quali hanno generato, in ciascuna replicazione, 400 stime dei tre quartili. Dunque per ciascuna replicazione è stato possibile ottenere una stima puntuale dei tre quartili, individuata dalla media dei valori ottenuti con il bootstrap, e gli intervalli di credibilità con maggior densità a posteriori. Nel caso in esame, al contrario di quanto accade in riferimento all'esercizio del paragrafo 3.2, i momenti utilizzati sono affetti da errore di campionamento e l'alea dovrebbe essere scontata nel bootstrap. A causa della convergenza dei momenti stimati a quelli della legge che ha generato i dati, la mistura con la funzione di ripartizione empirica non avrebbe ragion d'essere ai fini della consistenza, ma è utile per enucleare l'effetto del ricorso ai momenti campionari quando si vogliano effettuare confronti con gli esiti ottenuti usando le informazioni a priori.

Le medie dei quantili stimati sono esposte nella tabella (3.8).

k	q_{25}	q_{50}	q_{75}
1.000	16.951	19.792	23.123
5.000	16.905	19.724	23.019
10.000	16.920	19.714	22.973
25.000	16.920	19.688	22.939

Tabella 3.8: Base GEL da momenti stimati: stime dei quartili

Le ampiezze degli intervalli *h.p.d.*, tabella (3.9), si contraggono in misura tanto maggiore quanto più elevato è il peso assunto nella mistura dalla funzione di ripartizione costruita sulla base dei momenti empirici

k	q_{25}	q_{50}	q_{75}
1.000	4.574	4.958	6.402
5.000	4.138	4.564	5.752
10.000	3.781	4.099	5.155
25.000	3.033	3.317	4.090

Tabella 3.9: Base GEL da momenti stimati: ampiezze degli IC h.p.d. per $\alpha = 0.05$

Le coperture effettive degli intervalli *h.p.d.*, tabella (3.10), diminuiscono al crescere del peso relativo della componente *GEL*. Si evidenzia in questa circostanza un andamento opposto a quello riscontrato facendo uso dei momenti della Gamma(20, 1). Esso è ascrivibile al congiunto agire della bassa numerosità campionaria, che rende le stime dei momenti inaccurate, e dell'insufficiente recupero di dispersione per le posteriori dei quartili ricavate mediante i passi di bootstrap bayesiano, all'aumentare del peso della priori.

k	q_{25}	q_{50}	q_{75}
1.000	0.935	0.960	0.945
5.000	0.925	0.938	0.928
10.000	0.920	0.925	0.895
25.000	0.858	0.858	0.838

Tabella 3.10: Base GEL da momenti stimati: coperture degli IC h.p.d. per $\alpha = 0.05$

I risultati sembrano corroborare l'idea che i momenti empirici potrebbero essere usati per migliorare, almeno in termini di minore ampiezza media degli intervalli *h.p.d.*, le caratteristiche del ricampionamento bayesiano con priori impropria. A questo proposito si sono constatate alcune similitudini con l'uso, quale priori, della legge che ha generato i dati. Rimane aperta una questione di fondo: l'incertezza sui momenti, come già sottolineato, solo parzialmente è stata scontata nei cicli di bootstrap afferenti ogni singola replicazione e secondo una prospettiva bayesiana sarebbe più naturale considerare i parametri in discorso quali realizzazioni di una variabile casuale multidimensionale. Gli effetti di un simile cambiamento sono, come vedremo tra breve, non trascurabili.

3.3 Impostazione Completamente Predittiva

L'impostazione bayesiana definita completamente predittiva (Piccinato, 1996) trova la propria ragione nel tentativo di fondare unicamente su grandezze osservabili l'apprendimento delle leggi che governano i fenomeni oggetto di indagine ovvero la possibilità di prevedere ulteriori osservazioni. Un modello predittivo per una sequenza di variabili casuali è costituito da una misura di probabilità che ne specifica la legge congiunta (Bernardo e Smith, 2000). Nell'impostazione classica, l'inferenza in presenza di campioni costituiti da unità statistiche indipendenti è assicurata dall'ulteriore ipotesi di somiglianza. Quest'ultima rappresenta la soluzione al problema di apprendere dall'esperienza, dato che a causa dell'indipendenza $p(x^+|x) = p(x^+)$. In altri termini, grazie all'ipotesi di somiglianza è possibile l'accumulazione di informazioni riguardanti il fenomeno oggetto di indagine. Una condizione più debole dell'indipendenza, ma dall'indipendenza implicata, è la scambiabilità. Variabili casuali giudicate scambiabili sono caratterizzate da una distribuzione o densità congiunta rispetto alla quale è irrilevante l'ordine secondo il quale si manifestano. Un modo alternativo di esprimere lo stesso concetto consiste nel postulare la non informatività delle etichette associate alle unità statistiche. La scambiabilità riferibile soltanto a sottoinsiemi della sequenza osservata viene detta parziale ed è significativa della presenza di contenuto informativo nelle etichette - addizionali rispetto a quelle che permetterebbero l'individuazione di ciascuna unità statistica - caratterizzanti diversi gruppi di osservazioni. Il teorema di rappresentazione (de Finetti, 1937) stabilisce che una successione di variabili casuali $\{X_n\}$ è scambiabile se e solo se esiste una funzione di ripartizione aleatoria F condizionatamente alla quale le variabili della successione siano indipendenti e somiglianti. In altre parole, se le variabili casuali $\{X_n\}$ sono scambiabili, tutta l'informazione rilevante ai fini inferenziali è contenuta nella funzione di ripartizione. L'esistenza di una legge sulle funzioni di ripartizione, indotta da quella sulle osservazioni, permette di rappresentare quest'ultima senza doverla esplicitare direttamente. Da un punto di vista interpretativo è utile evidenziare due aspetti dello schema concettuale:

- ▷ la legge di probabilità sulla base della quale le osservazioni sono giudicate in qualche senso scambiabili induce una densità di probabilità su un "parametro",
- ▷ condizionatamente al "parametro", le unità statistiche incluse nella sequenza sono indipendenti e somiglianti (almeno all'interno di opportuni sottoinsiemi se, ad esempio, la scambiabilità è parziale).

Mentre oggetto del secondo punto sono densità o distribuzioni del campione caratterizzate in modo classico, il primo punto permette di evidenziare che:

- ▷ i valori assunti dal "parametro" non sono delle costanti, ma rappresentano realizzazioni di una legge di probabilità,
- ▷ la legge di probabilità sul "parametro" è indotta dalla legge di probabilità sulle osservazioni.

Tale costruzione si colloca naturalmente nell'alveo dell'impostazione bayesiana arricchendola di nuovi significati: la legge sul "parametro" viene derivata da una legge sulle osservazioni e ciò indica la capacità, almeno in linea di principio, di poter fondare i processi inferenziali soltanto su grandezze effettivamente osservate. Questa circostanza giustifica l'appellativo di impostazione bayesiana completamente predittiva. Per fissare le idee è conveniente ricordare il caso di variabili dicotomiche $x \in \{0,1\}$:

$$\begin{cases} P(x) = \int_\theta \prod_i \theta^{x_i}(1-\theta)^{1-x_i} dF(\theta) \\ \hat{\theta}(x) = \frac{\sum x_i}{N} \xrightarrow{p} \theta \\ F(\theta) = \lim_{N\to\infty} p\left(\hat{\theta} \leq \theta\right) \end{cases} \tag{3.3}$$

La $F(\theta)$ è dedotta sotto l'ipotesi di scambiabilità e descrive la funzione di ripartizione del parametro θ ottenuta come passaggio a limite della corrispondente funzione di ripartizione empirica nonché, allo stesso tempo, il grado di fiducia rispetto ai valori che la frequenza relativa dei successi, al divergere della numerosità campionaria, può assumere. Quest'ultima quantità è suscettibile, in linea di principio, di apprendimento dall'evidenza sperimentale. La circostanza che le unità statistiche siano indipendenti e somiglianti condizionatamente a θ permette di ricavare la legge sulle osservazioni, ossia la predittiva iniziale, evitando di elicitare direttamente $P(x)$.

Nel caso di variabili casuali continue le caratteristiche dell'approccio possono essere mantenute, ma è meno semplice conservare le possibilità applicative. Indicando con F_p, F_n e $\mathscr{F}$ la funzione di ripartizione congiunta delle osservazioni, la funzione di ripartizione empirica e, rispettivamente, lo spazio di tutte le funzioni di ripartizione definite su $\mathbb{R}$, l'estensione dello schema concettuale prima accennato può essere sintetizzata dalle equazioni:

$$\begin{cases} F_P(x) = \int_{\mathscr{F}} \prod_i F(x_i) dQ(F) \\ Q(F) = \lim_{N\to\infty} p(F_n) \end{cases}$$

L'oggetto dell'apprendimento dall'esperienza, ossia quel che prima era un parametro di dimensioni finite, nel caso in discussione diviene una funzione di ripartizione (parametro infinito-dimensionale) e la legge P sulle osservazioni induce una legge di probabilità sulle funzioni di ripartizione aventi supporto in $\mathbb{R}$. Condizionatamente ad

F, realizzazione ottenuta da $Q(F)$, le unità incluse nella variabile casuale N^{pla} descritta dal campione sono indipendenti e somiglianti e ancora una volta è possibile risalire alla funzione di ripartizione predittiva senza conoscere la legge P. Purtroppo, in assenza di restrizioni che permettano di circoscrivere lo spazio $\mathscr{F}$ supporto di $Q(F)$, la valenza operativa di questo modo di procedere sembra essere limitata. Allora può rivelarsi di qualche utilità l'individuazione di criteri idonei ad enucleare sottoinsiemi di $\mathscr{F}$ che, pur trattabili, siano abbastanza ampi da includere le distribuzioni presumibilmente rilevanti ai fini applicativi.

3.4 Teorema di Rappresentazione e MM

La seconda equazione della (3.3) è facilmente interpretabile alla luce del metodo dei momenti e suggerisce, anche in riferimento a variabili aleatorie più complesse di una Bernoulli, una possibile strategia volta a rendere possibili inferenze completamente predittive. La soluzione al problema in discorso risiede nella possibilità di scrivere, quando siano adottati quali parametri i momenti, le funzioni di verosimiglianza che dovrebbero comparire nella prima equazione del teorema di rappresentazione. Così facendo l'insieme $\mathscr{F}$ sarebbe, in prima approssimazione, ristretto alle sole leggi di probabilità che ammettono almeno un certo numero di momenti finiti; la parametrizzazione di una funzione di verosimiglianza rispetto ad essi è strettamente legata, dal punto di vista concettuale, alle idee sottese nei lavori di K. Pearson; Bera e Bilias (2002) illustrano efficacemente il punto. Sia f la densità di un carattere y oggetto di studio. L'espansione in serie di potenze di f attorno a $y = 0$ fornisce l'approssimazione

$$f = \sum_{j=0}^{q} \frac{y^j}{j!} \left. \frac{d^j f}{dy^j} \right|_{y=0} + R$$

ossia

$$f(y; \upsilon) = \sum_{j=0}^{q} \frac{y^j}{j!} \upsilon_j + R(\upsilon)$$

Si consideri il problema di individuare i parametri υ che minimizzano la distanza rispetto alle frequenze relative osservate $\widehat{f}$, in modo che :

$$f = argmin \int \left[f - \widehat{f} \right]^2 dy$$

Dalle condizioni del primo ordine, sotto le opportune condizioni di regolarità si ricava

$$2\int\left[f-\widehat{f}\right]f'dy=2\sum_{j=0}^{q}\int\left[f-\widehat{f}\right]\left[\frac{y^j}{j!}+\frac{\partial R}{\partial \upsilon_j}\right]dy=0$$

La precedente relazione è implicata (moltiplicando i membri per $j!$) dalle seguenti:

$$\mu_j-m_j+j!\int\left[f-\widehat{f}\right]\left[\frac{\partial R}{\partial \upsilon_j}\right]dy=0\,\forall j$$

Se il resto dell'espansione in serie di potenze è trascurabile si ricava immediatamente $\mu_j=m_j$. Alla luce di tutto questo si giustifica l'idea di adattare una legge nota alla distribuzione empirica quantificandone i parametri in modo da verificare le condizioni sui momenti.

Sembra tuttavia maggiormente interessante il ruolo dei momenti nella stima semiparametrica di una legge di probabilità: il sistema di pesi che verifica le condizioni minimizzando la distanza euclidea (ossia la discrepanza di Cressie-Read di parametro $\gamma=1$) rispetto ad una distribuzione multinomiale è, a tutti gli effetti, una verosimiglianza empirica generalizzata. La possibilità di stimare densità di probabilità facendo uso di verosimiglianze di massima entropia che utilizzino condizioni di ortogonalità riferite ai momenti è stata indagata da Wu (2003).

Per quanto visto, l'uso dei momenti permette lo smoothing dello stimatore istogramma; ciò equivale ad affermare che i moltiplicatori che caricano i vincoli nelle verosimiglianze empiriche generalizzate selezionano tra le infinite leggi appartenenti allo spazio $\mathscr{F}$ quella che soddisfacendoli minimizza la discrepanza nei confronti della distribuzione o densità di riferimento. Al variare dei parametri (momenti) generati dalla posteriori $F(\mu)$, le verosimiglianze massimizzate rispetto ai moltiplicatori descrivono un inviluppo di distribuzioni "ottimali". Momenti e moltiplicatori nell'impostazione predittiva sono parametri di disturbo. I primi possono essere eliminati attraverso una marginalizzazione, in coerenza con il modo di procedere usuale in ambito bayesiano. I secondi sono eliminati mediante ottimizzazione. È tutt'ora aperto il dibattito sulla liceità d'uso delle verosimiglianze profilo nelle inferenze basate su procedure di tipo bayesiano. Grendar e Judge (2007) ricorrendo al teorema di Sanov dimostrano la consistenza in *L-divergenza* della moda a posteriori alla verosimiglianza empirica di Owen *(1988)*. In appendice viene formalizzata l'estensione del teorema al caso delle *GEL*. In termini intuitivi, nel contesto esaminato, l'ottimizzazione è lo strumento che permette di selezionare una legge di probabilità tra le infinite che sono caratterizzate dalle medesime condizioni sui momenti. Formalmente, la seguente espressione del teorema di rappresentazione

$$\begin{cases} p(x) = \int_{\mu} \prod_i p(x_i | \mu, \widehat{\upsilon}) \, dF(\mu) \\ \forall \mu_k < \infty : \frac{1}{N} \sum_i x_i^k \xrightarrow{qc} \mu_k \\ F(\mu) = \lim_{N \to \infty} p\left[\cap_k \left(\frac{1}{N} \sum_i x_i^k < \mu_k\right)\right] \end{cases} \tag{3.4}$$

costituisce un modo alternativo di esprimere l'operare della restrizione su $\mathscr{F}$:

$$\begin{cases} F_P(x) = \int_{\mathscr{F}^*} \prod_i F^*(x_i) dQ(F^*) \\ Q(F^*) = \lim_{N \to \infty} p(F_n) \\ \mathscr{F}^* = \left\{F : \forall \mu_k < \infty, \; F = argmin I_\gamma(F, F_n) \; sub \int x^k dF = \mu_k\right\} \end{cases} \tag{3.5}$$

In parole, $\mathscr{F}^* \subset \mathscr{F}$ include solo i "punti" F, supporto di Q, che per qualunque vettore di momenti μ rende minima la divergenza I_γ tra F e la funzione di ripartizione empirica F_n. Dunque è garantito che Q attribuisca probabilità non nulle solo alle funzioni di ripartizione F maggiormente congruenti con l'evidenza empirica. Le equazioni seconda e terza della (3.4) sottolineano che al divergere della numerosità campionaria, sotto l'ipotesi di ergodicità, i momenti stimati convergono ai momenti della legge di probabilità sulle osservazioni e la funzione di ripartizione empirica converge uniformemente, circostanze che in termini diversi sono espressi dalla seconda e terza equazione della (3.5) le quali evidenziano come Q assegni agli elementi di $\mathscr{F}^*$ gradi di fiducia tanto più elevati quanto maggiore è la prossimità alla F_n. La consistenza è assicurata dal fatto che la restrizione di $\mathscr{F}$ significa l'attribuzione di probabilità nulle solo alle F che, nella classe di quelle caratterizzate da momenti μ, risultino dominate dal punto di vista della prossimità a F_n secondo la metrica indotta da I_γ. Questa radicale semplificazione dischiude interessanti prospettive per inferenze model free completamente predittive, secondo una modellizzazione semiparametrica, con notevoli vantaggi in termini di semplicità e rende possibili confronti tra l'impostazione bayesiana completamente predittiva ed il ricampionamento parametrico classico.
In termini molto generali, ci si è ricondotti alla situazione in cui la sequenza di eventi scambiabili dipende da una legge funzione del parametro θ di dimensione finita. Posto che sulla base di un campione $\mathbf{x}$ sia possibile ricavare la posteriori del parametro θ, la legge per x^+ dato $\mathbf{x}$ viene precisata dalle equazioni:

$$\begin{cases} p(x^+ | \mathbf{x}) = \int_{\theta} \prod_i p\left(x_i^+ | \theta\right) dF(\theta | \mathbf{x}) \\ \forall k : \hat{\theta}_k(\mathbf{x}) \xrightarrow{p} \theta_k \\ F(\theta | \mathbf{x}) = \lim_{N \to \infty} p\left[\cap_k \left(\hat{\theta}_k < \theta_k\right) | \mathbf{x}\right] \end{cases}$$

La prima evidenzia che il campione $\mathbf{x}$ viene utilizzato per apprendere la $F(\theta | \mathbf{x})$ e che

nota la $p\left(x_i^+ \,|\theta\right)$ resta definita la $p\left(x^+ \,|\mathbf{x}\right)$. Generare da $p\left(x^+ \,|\mathbf{x}\right)$ è paragonabile all'operazione che si compie in ambito classico campionando da $p\left(x^+ \middle| \hat{\theta}\right)$. La diversità risiede nell'argomento del condizionamento: usando la $p\left(x^+ \,|\mathbf{x}\right)$ le realizzazioni x^+ dipendono da $\mathbf{x}$ direttamente e non in via mediata tramite $\hat{\theta}$.
In definitiva, la posteriori $F\left(\theta \,|\mathbf{x}\right)$, facendo ricorso al bootstrap bayesiano, può essere simulata numericamente senza dover precisare la legge a priori sul parametro θ; le verosimiglianze empiriche generalizzate offrono una soluzione al problema di stimare le $p\left(x_i \,|\theta\,\right) \in [0,1]$; convenendo che $\theta \equiv \mu$, è utile assumere:

- ▷ l'esistenza di momenti μ finiti di ordine $k = 1, 2, \ldots$
- ▷ quale verosimiglianza per l'i^{ma} $(i = 1, 2, \cdots, N)$ unità statistica inclusa nel campione, una funzione $p\left(x_i \,|\mu\,\right) \in [0,1]$ tale che $\sum_i p\left(x_i \,|\mu\,\right) x_i^k = \mu_k$ $(k = 1, 2, \ldots)$.

Sotto la condizione di scambiabilità si ricava la predittiva iniziale per x:

$$\begin{cases} p\left(x\right) = \int_\mu \prod_i p\left(x_i \,|\mu\,\right) dF\left(\mu\right) \\ \forall \mu_k < \infty : \frac{1}{N} \sum_i x_i^k \overset{qc}{\to} \mu_k \\ F\left(\mu\right) = \lim\limits_{N \to \infty} p\left[\cap_k \left(\frac{1}{N} \sum_i x_i^k < \mu_k\right)\right] \end{cases}$$

e, analogamente, la predittiva per x^+ dato il campione osservato $\mathbf{x}$:

$$\begin{cases} p\left(x^+ \,|\mathbf{x}\right) = \int_\mu \prod_i p\left(x_i^+ \,|\mu\,\right) dF\left(\mu \,|\mathbf{x}\right) \\ \forall \mu_k < \infty : \frac{1}{N} \sum_i \left(x_i^+\right)^k \overset{qc}{\to} \mu_k \\ F\left(\mu \,|\mathbf{x}\right) = \lim\limits_{N \to \infty} p\left\{\cap_k \left[\frac{1}{N} \sum_i \left(x_i^+\right)^k < \mu_k\right] |\mathbf{x}\right\} \end{cases}$$

Utilizzando le distribuzioni predittive a scopo di ricampionamento, viene superato il limite costituito dall'attribuzione di probabilità pari a zero alle intensità non osservate. Tale risultato è dovuto alla modellizzazione semiparametrica. Un aspetto critico è rappresentato dal fatto che difficilmente è possibile considerare tante condizioni di ortogonalità quanti sono i momenti finiti che la variabile casuale ammette, ne è dato sapere quanti siano. Tuttavia, come accennato nel paragrafo 3.2.1, la conseguenza dell'inclusione di un numero limitato di condizioni si può apprezzare sulla scorta dei risultati noti circa l'estremo superiore della divergenza tra leggi che hanno in parte uguali momenti e dal punto di vita applicativo le scelte possono essere demandate agli esiti di opportuni test di ipotesi.

3.4.1 Ricampionamento Semiparametrico

L'esercizio riguardante la stima dei tre quartili della variabile casuale descritta da un campione di venticinque unità è stato ripetuto abbandonando la modellizzazione della legge di probabilità come processo di Dirichlet e adottando in luogo di essa la strategia precedentemente accennata. Per ciascuna delle 400 replicazioni sono state generate 400 realizzazioni della posteriori sui momenti ed in corrispondenza stimate le leggi di massima entropia. La marginalizzazione rispetto ai parametri di disturbo (ossia i momenti) permette di determinare numericamente, in ogni replicazione, una predittiva per le osservazioni dalla quale si possono ricavare stime puntuali circa i tre quartili. Gli intervalli *h.p.d.* sono calcolati per ciascuna replicazione usando le stime puntuali afferenti le 400 verosimiglianze di massima entropia. In questo modo si può dare conto dell'ampiezza e della copertura effettiva degli intervalli. Le medie dei quartili stimati sono riportate nella tabella (3.11).

q_{25}	q_{50}	q_{75}
16.753	19.730	22.687

Tabella 3.11: Impostazione predittiva: stime dei quartili

Le ampiezze degli intervalli *h.p.d.*, tabella (3.12), e le coperture effettive, tabella (3.13), appaiono più congruamente dimensionate rispetto a quelle ottenute nella sezione 3.2.3. Tale evidenza fornisce supporto all'idea che lo smoothing del bootstrap bayesiano, posto in essere tramite le *GEL* ricavate dai momenti empirici, potesse soffrire di un problema di sottovalutazione della variabilità delle stime, non sufficientemente catturata dal ricampionamento sulle unità generate dalla mistura delle funzioni di ripartizione.

q_{25}	q_{50}	q_{75}
4.374	4.020	5.105

Tabella 3.12: Impostazione predittiva: ampiezze degli IC h.p.d. per $\alpha = 0.05$

Nonostante la numerosità campionaria sia esigua, ogni realizzazione della posteriori sui momenti fornisce stime consistenti dei parametri; il fatto che in ciascuna replicazione le 400 verosimiglianze di massima entropia permettano di ricavare gli intervalli *h.p.d.* dei quartili si traduce nell'esecuzione di un bootstrap semiparametrico e

l'avvicinamento delle coperture effettive al livello nominale prescelto ($1-\alpha = 0.95$) ne costituisce il risultato.

q_{25}	q_{50}	q_{75}
0.945	0.948	0.945

Tabella 3.13: Impostazione predittiva: coperture degli IC h.p.d. per $\alpha = 0.05$

3.4.2 Robustezza delle Stime

L'esercizio presentato nella sezione 3.2 illustra l'uso di *GEL* costruite usando, nelle condizioni di ortogonalità, i momenti della legge di probabilità che genera i dati. Ad eventuali osservazioni "devianti" vengono assegnati pesi assai modesti ed il relativo contributo alla morfologia della verosimiglianza risulta irrilevante. Tale esito è dovuto al fatto che i dati non servono alla quantificazione dei momenti (per ipotesi conosciuti dal ricercatore) e recano nelle stime solo informazioni su distanze. Se i momenti sono ignoti e vengono stimati a partire dalle osservazioni, la ponderazione operata dalle *GEL* in piccoli campioni può essere pesantemente affetta dall'influenza di poche unità statistiche. Una utile introduzione ai problemi di robustezza è contenuta in Piccolo (2000). In questa sede è opportuno ricordare la distinzione tra errori locali e valori anomali. I primi, originati essenzialmente da troncamenti e arrotondamenti, si manifestano come modificazioni di lieve entità nei valori. I secondi sono identificati, in prima approssimazione, da disomogeneità tali da renderli remoti rispetto ai rimanenti; facendo salvi i casi originati da errori materiali, il giudizio in ordine a cosa sia anomalo dipende in realtà dai fini della ricerca: si pensi al caso in cui si vogliano spiegare o riprodurre le caratteristiche di dati estremamante dispersi. Posto che il contenuto informativo di osservazioni molto lontane dalla parte centrale di una distribuzione o densità sia ritenuto non funzionale all'analisi, valutazioni in ordine alla robustezza delle stime per i momenti possono essere effettuate ricorrendo alla funzione di influenza *(Hampel, 1974)*. Essa ha la natura di derivata della variazione che subisce la stima m a fronte di una contaminazione di F; indicando con δ_x la funzione di ripartizione di una delta di Dirac, fissato un $\varepsilon > 0$,

$$IF(x,m,F) = \lim_{\varepsilon \to 0} \frac{m[(1-\varepsilon)F - \varepsilon\delta_x] - m(F)}{\varepsilon}$$

Grazie ad essa possono essere calcolati due indici, $\gamma(m,F)$ e $\lambda(m,F)$ esprimenti la sensibilità delle stime m ai valori anomali e, rispettivamente, agli errori locali:

$$\gamma(m,F) = \sup_x |IF(x,m,F)|$$

$$\lambda(m,F) = \sup_{x \neq y} \frac{|IF(x,m,F) - IF(y,m,F)|}{|x-y|}$$

- ▷ Se $\gamma(m,F) < +\infty$ lo stimatore è robusto rispetto ai valori anomali,
- ▷ se $\lambda(m,F) < +\infty$ lo stimatore è robusto rispetto agli errori locali.

Mentre la funzione di influenza fornisce indicazioni sul comportamento di $m(F)$ in un intorno di F, il punto di rottura $\varepsilon^*(m)$ individua la massima frazione di dati che possono essere contaminati prima che $IF(x,m,F) = +\infty$ ossia delimita l'intorno nel quale le stime non perdono completamente significato. Ai nostri fini è utile notare che per le medie campionarie è $\varepsilon^*(m) = 0$ dato che un solo dato anomalo, se sufficientemente grande in valore assoluto, può ingenerare un valore arbitrariamente diverso delle stime.

L'ottenimento di condizioni di ortogonalità robuste sembra essere perseguibile adottando stime robuste per le medie, piuttosto che intervenendo direttamente sulle equazioni che le definiscono. Benché sia possibile pensare all'utilizzo di stimatori *M (Huber, 1964)*, il sistema di pesi definito dalle verosimiglianze empiriche generalizzate sarebbe condizionato ai parametri scelti per la ponderazione strumentale alla robustezza. L'uso di medie troncate o winsorizzate, essendo basato sul rango occupato dalle unità statistiche ordinate in successione non decrescente delle intensità, permette di evitare l'esplicita fissazione di parametri, fatta salva la decisione in ordine al punto di rottura contro il quale si intende difendere le stime. Il calcolo delle medie troncate richiede l'attribuzione di un peso pari a zero alle unità individuate dai ranghi estremi, mentre la winsorizzazione ne sostituisce i valori con quelli corrispondenti ai ranghi adiacenti. Indicando i momenti campionari k^{mi}, troncati e winsorizzati con $m_{k,t}$ e $m_{k,w}$, posto $\varepsilon^*(m)$=αN si possono esplicitare le relazioni che rendono evidenti i legami tra i due approcci:

$$m_{k,t} = \frac{1}{N - \alpha N} \sum_{i=1+\frac{1}{2}\alpha N}^{N-\frac{1}{2}\alpha N} x_{(i)}^k$$

$$m_{k,w} = \frac{1}{N} \left[\sum_{i=1+\frac{1}{2}\alpha N}^{N-\frac{1}{2}\alpha N} x_{(i)}^k + \frac{\alpha N}{2} x_{(1+\frac{1}{2}\alpha N)} + \frac{\alpha N}{2} x_{(N-\frac{1}{2}\alpha N)} \right]$$

Il sacrificio delle osservazioni estreme, se da una parte consente di innalzare il punto di rottura da zero ad αN , dall'altra produce perdita di efficienza nelle stime

e rende indispensabile una valutazione accurata del trade-off efficienza-robustezza. L'implementazione all'interno dei passi di bootstrap utili a generare realizzazioni della posteriori sui momenti è immediata e può essere ritenuta riconducibile al bootstrap con priori propria: ambedue le soluzioni sono ottenute assumendo una distribuzione base derivata dalla uniforme, tale da allocare in corrispondenza delle αN intensità estremali probabilità nulla e, limitatamente alla winsorizzazione, massa di $\alpha N/2$ alle intensità immediatamente adiacenti. Con $\alpha = o(N)$, le proprietà di consistenza sono conservate.

3.5 Dati Dipendenti

In linea di principio, il ricampionamento di dati affetti da qualche forma di dipendenza è realizzabile a condizione che la sequenza con la quale - nel campione - vengono osservate C classi di scambiabilità parziale sia sufficiente a riprodurre i legami tra le osservazioni; individuando ciascuna classe con l'ordine in cui si è manifestata:

$$\begin{cases} \mathbf{x}^{+} = \left\{ x^{+}_{(c)} \right\},\ c = 1,2,\ldots,C \\ p\left(\mathbf{x}^{+}_{c} \,|\mathbf{x}_{c}\right) = \int_{\mu} \prod_{i} p\left(x^{+}_{i,c} \,|\mu_{c}\right) dF\left(\mu_{c} \,|\mathbf{x}_{c}\right) \\ \forall \mu_{k} < \infty : \frac{1}{N_{c}} \sum_{i} \left(x^{+}_{i,c}\right)^{k} \to \mu_{k} \in \mu_{c} \\ F\left(\mu_{c} \,|\mathbf{x}_{c}\right) = \lim_{N\to\infty} p\left\{ \cap_{k} \left[\frac{1}{N_{c}} \sum_{i} \left(x^{+}_{i,c}\right)^{k} < \mu_{k} \in \mu_{c} \right] |\mathbf{x}_{c} \right\} \end{cases} \tag{3.6}$$

La prima equazione esprime l'ordinamento delle realizzazioni, utile a mimare la dipendenza, secondo la successione nella quale si sono manifestate le classi di appartenenza. Tale modo di procedere è comune alle strategie di ricampionamento di impostazione classica, quali il block bootstrap o il subsampling (Politis, 2003). Il primo ricampiona blocchi di osservazioni, mentre il secondo estrae sottocampioni costituiti da unità temporalmente contigue. In entrambi i casi si persegue il mantenimento delle relazioni di dipendenza conservando l'ordinamento delle osservazioni.
In realtà, una esplicita definizione delle classi può risultare non indispensabile: riducendone progressivamente l'ampiezza esse vengono collassate fino a coincidere con i ranghi delle osservazioni. Indicando con c_t il rango associato alla t^{ma} osservazione, la (3.6) può essere riformulata come segue:

$$\begin{cases} \mathbf{x}^{+} = \left\{ x^{+}_{(c_t)} \right\},\ t = 1,2,\ldots,N \\ p\left(\mathbf{x}^{+} \,|\mathbf{x}\right) = \int_{\mu} p\left(\mathbf{x}^{+} \,|\mu\right) dF\left(\mu \,|\mathbf{x}\right) \\ \forall \mu_{k} < \infty : \frac{1}{N} \sum_{i} \left(x^{+}_{i}\right)^{k} \to \mu_{k} \\ F\left(\mu \,|\mathbf{x}\right) = \lim_{N\to\infty} p\left\{ \cap_{k} \left[\frac{1}{N} \sum_{i} \left(x^{+}_{i}\right)^{k} < \mu_{k} \right] |\mathbf{x} \right\} \end{cases} \tag{3.7}$$

In questo caso la prima equazione definisce il vettore $\mathbf{x}^+$ come insieme di realizzazioni delle N classi di scambiabilità parziale, ordinato in base alla successione osservata per esse. Campionare dalla predittiva come se le unità fossero globalmente scambiabili, salvo ordinare le intensità secondo la sequenza in cui si sono manifestati i ranghi, costituirebbe l'interpretazione bayesiana del bootstrap di massima entropia proposto da Vinod (2004) per variabili aleatorie continue, pur in presenza di sostanziali diversità costruttive che ne permetterebbero l'applicazione nei casi discreto e misto; il modo di procedere appena descritto sarebbe soltanto un'approssimazione del framework implicito nella (3.7): l'idea di scambiabilità parziale richiede che ciascuna delle unità componenti la collezione di realizzazioni volta a mimare le proprietà della serie osservata, sia estratta dalla classe di appartenenza; le simulazioni dalla predittiva con ordinamento delle intensità secondo i ranghi osservati, ripetute S volte, sono strumentali alla definizione delle distribuzioni empiriche di ogni sezione; in definitiva, la procedura descritta nella (3.7) può essere così schematizzata:

1. dalla predittiva, per S volte, si estraggono N realizzazioni che vengono ordinate secondo la sequenza temporale dei ranghi nel campione osservato,

2. dalla distribuzione empirica di ogni variabile casuale sezione si estrae uno degli S elementi che la costituiscono, ottenendo una collezione di N realizzazioni tale da simulare una traiettoria del processo aleatorio

3. iterando il passo precedente un elevato numero di volte si ottengono altrettante serie, input per il calcolo delle stime di interesse.

Logica conseguenza è l'attenuazione del vincolo di identità tra la sequenza temporale dei ranghi osservati e ciascuna delle sequenze di ranghi per le traiettorie simulate. La procedura di campionamento-ricampionamento permette, in altri termini, di prescindere dalla esplicita (e problematica) definizione di intervalli di fungibilità tra ranghi, altrimenti necessaria ove si persegua il rilassamento del vincolo direttamente nella fase di ordinamento del campione generato dalla predittiva.

Andando oltre gli aspetti operativi dell'estensione sin qui accennata, sono opportune alcune considerazioni in ordine a significati e conseguenze implicite nell'uso dei momenti:

- ▷ la circostanza che le traiettorie dei processi siano ricavate da condizioni sui momenti relativi all'intero campione stabilisce, in riferimento ai modelli nello spazio degli stati, analogie con lo smoothing,

- ▷ nella misura in cui le distribuzioni empiriche delle sezioni approssimano le corrispondenti leggi di probabilità, la procedura di campionamento - ricampionamento emula implicitamente una catena di Markov possibilmente non omogenea e di ordine imprecisato $\tau - 1$ tale che:

$$P(X_t = x_t) = \int_{\mathbb{R}^{\tau-1}} P(X_t = x_t \,|X_{t-\tau:t-1})dP(X_{t-\tau:t-1})$$

- ▷ se nei dati è presente una struttura di dipendenza, la possibilità di mantenere l'interpretabilità, in termini inferenziali, dei risultati ottenuti tramite simulazione è legata alla proprietà di consistenza dei momenti campionari (ergodicità). La *GEL* ottenuta dalle condizioni sui momenti continua anch'essa ad essere consistente, ma la convergenza si realizza a velocità inferiore di quella garantita dal *TCL* per dati *i.i.d.*: $\widehat{p}_i = \widetilde{p}_i \left\{1 + O_p\left(N^{-\varphi}\right)\right\}$ con $\varphi < 0.5$,

- ▷ in assenza di stazionarietà, la predittiva calcolata per una serie storica ne descrive l'approssimazione, all'interno del campione, con la legge stazionaria avente "localmente" identici momenti campionari. Le variabili casuali sezione ottenute per simulazione non sono sufficienti alla rappresentazione esaustiva del vero processo aleatorio; esse permettono soltanto di circoscrivere le traiettorie caratterizzate da un ristretto insieme di sequenze di ranghi all'interno del segmento temporale osservato.

Il campionamento dalla predittiva, con ordinamento delle realizzazioni secondo la sequenza in cui si sono manifestati i ranghi e ricampionamento dalle sezioni, è spontaneamente idoneo a gestire il trattamento di dati multivariati. Naturalmente la validità della strategia di ricampionamento non è inficiata quando manchi una struttura di dipendenza tra le unità statistiche. La generalizzazione della (3.7), indicando con j l'indice delle variabili, è immediata:

$$\begin{cases} x_j^+ = \left\{x_{j(c_t)}^+\right\},\ t = 1,2,\ldots,N \quad j = 1,2,\cdots,p \\ p\left(x_j^+ \,|\mathbf{x}_j\right) = \int_{\mu_j} p\left(x_j^+ \,\middle|\mu_j\right) dF\left(\mu_j \,|\mathbf{x}_j\right) \\ \forall \mu_k < \infty : \frac{1}{N}\sum_i \left(x_{ji}^+\right)^k \to \mu_{jk} \\ F\left(\mu_j \,|\mathbf{x}_j\right) = \lim_{N\to\infty} p\left\{\cap_k \left[\frac{1}{N}\sum_i \left(x_{ji}^+\right)^k < \mu_{jk}\right] |\mathbf{x}_j\right\} \end{cases} \tag{3.8}$$

Fatta salva la maggiore complessità della notazione, i significati - per un fissato j - non si discostano da quelli illustrati in precedenza. Grazie ai ranghi sono mantenute le relazioni di dipendenza multivariata così come nel caso univariato venivano conserva-

ti, se esistenti, i legami tra le unità statistiche di una serie storica. Il modo di operare sotteso dalla (3.8) può essere descritto nei seguenti passi:

1. dalle p predittive ricavate ciascuna indipendentemente dalle altre, per S volte, si estraggono N realizzazioni che vengono ordinate secondo le sequenze temporali dei rispettivi ranghi nel campione osservato,

2. costruita la distribuzione empirica p-variata di ciascuna delle N variabili casuali sezione, da ognuna si estrae uno degli S elementi che la costituiscono, ottenendo così N realizzazioni la cui collezione simula una traiettoria p-variata del processo aleatorio

3. ripetendo il campionamento dalle sezioni un elevato numero di volte si ottengono le basi dati che formano altrettanti input per il calcolo delle stime oggetto di indagine.

Dunque, la semplificazione introdotta dall'assunzione di scambiabilità parziale, nel caso multivariato è ancor più evidente: scrivere *GEL* tali da incorporare condizioni sui momenti misti sarebbe possibile, ma implicherebbe costi computazionali rapidamente insostenibili all'aumentare del numero di variabili considerate, a causa della crescita esponenziale delle permutazioni sulle potenze. L'ordinamento originato dalla matrice dei ranghi osservati permette di:

- ▷ contenere la crescita degli oneri di calcolo in proporzione lineare rispetto al numero di variabili,

- ▷ assicurare il mantenimento delle relazioni di dipendenza più spesso oggetto di studio, legate a momenti misti moderatamente elevati.

La metodologia di simulazione proposta, utile ad assicurare la conservazione dei nessi di dipendenza sia tra le osservazioni afferenti ad una stessa variabile, sia tra variabili distinte, può trovare applicazione in campi quali ad esempio

- ▷ l'econometria delle serie storiche; la simulazione dalle predittive non richiede una preliminare scomposizione delle serie ovvero l'applicazione di filtri volti ad agevolarne la trattabilità ai fini inferenziali, e per questa via risultano compressi i rischi legati a problemi di errata specificazione,

- ▷ la tutela della privacy; generare artificialmente microdati dalle distribuzioni predittive di interesse sembra offrire una premessa utile alla formulazione di approcci per il controllo del trade-off tra attendibilità nei risultati inferenziali e rischi di intrusione in informazioni non pubbliche,

Alcuni esempi in proposito formano oggetto del capitolo seguente.

Capitolo 4

Applicazioni

4.1 Un VECM per la Domanda di Moneta

In un lavoro del (1990), Johansen e Juselius stimano un modello vettoriale autoregressivo a correzione del divario per indagare le relazioni tra domanda di moneta, reddito, tassi d'interesse ed inflazione. Uno dei paesi studiati è la Finlandia ed i relativi dati, liberamente disponibili all'indirizzo `http://www.math.ku.dk/~sjo/data/finnish_data.html`, sono costituiti da quattro serie storiche in parte non stazionarie, osservate dal secondo trimestre del 1958 al terzo trimestre del 1984, ossia

- ▷ Lrm1: *ln* dello stock di moneta reale,
- ▷ Lny: *ln* del reddito reale,
- ▷ Lnmr: *ln* del tasso ufficiale di sconto praticato dalla Banca Centrale,
- ▷ Difp: tasso di inflazione.

In questa sede sembra opportuno verificare se la metodologia di simulazione proposta permetta, ad esempio, di giungere a conclusioni confrontabili con quelle illustrate nell'articolo per quanto concerne la dimensione dello spazio di cointegrazione e l'interpretazione delle relazioni che lo caratterizzano. La circostanza sarebbe indizio dell'efficace mantenimento sia delle caratteristiche individuali delle serie, sia dei legami di mutua dipendenza che le avvincono. Nella figura (4.1) sono rappresentate le serie storiche osservate e gli estremi, inferiore e superiore delle variabili casuali sezione ottenute generando 5000 traiettorie dalla predittiva. I valori attesi calcolati di ciascuna sezione, figura (4.2), coincidono sostanzialmente con il tracciato delle serie, fatto che assicura il mantenimento in media delle caratteristiche di dipendenza univariata e multivariata.

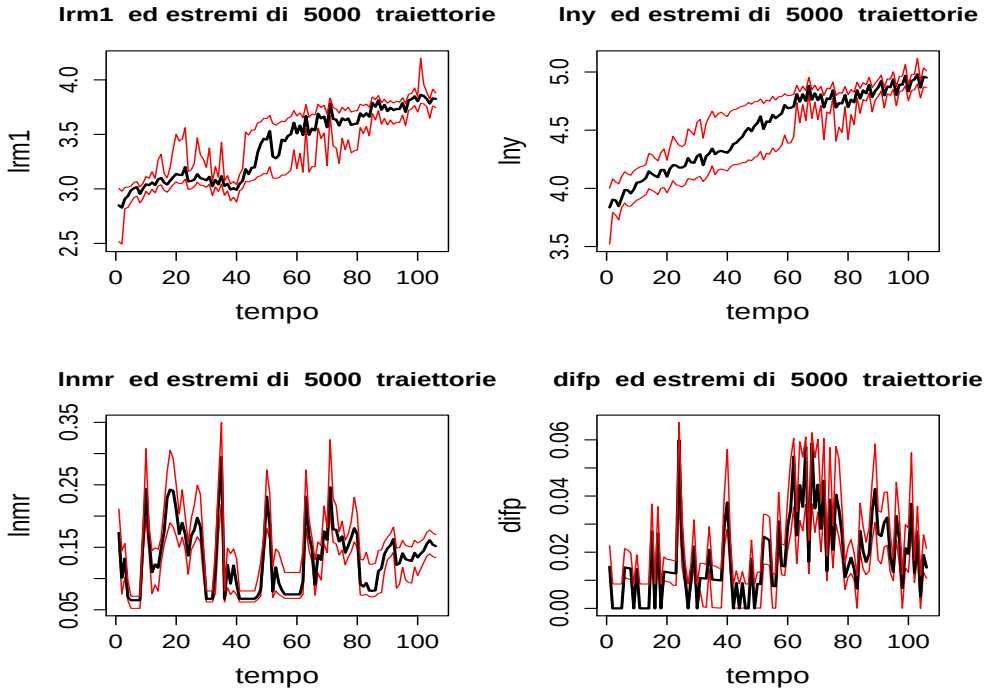


Figura 4.1: Traiettorie da 5000 simulazioni

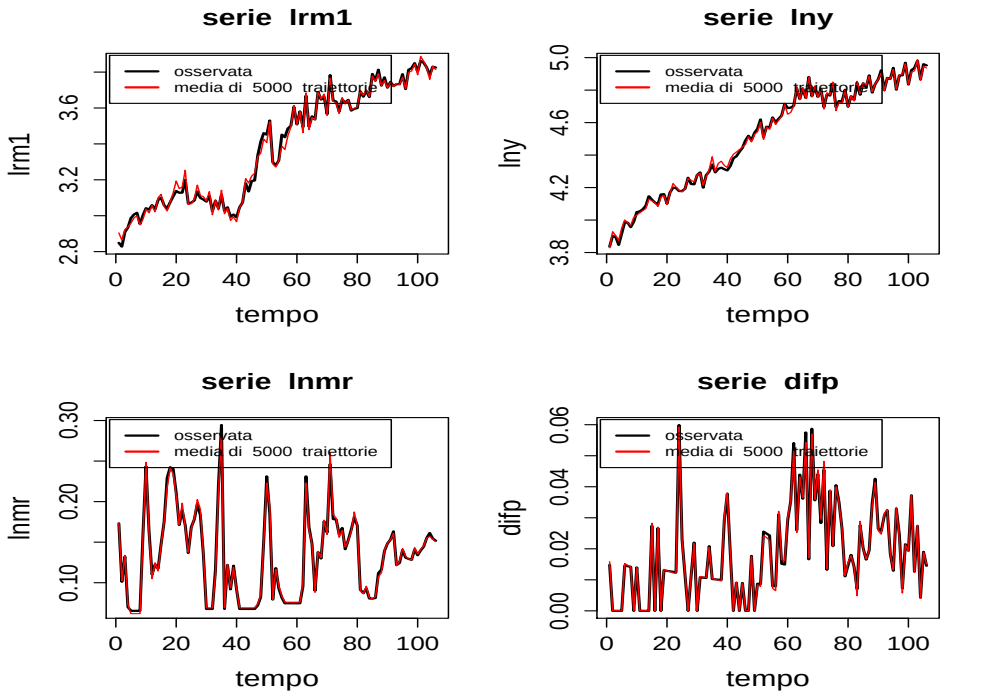


Figura 4.2: Medie delle traiettorie

Prima di procedere nell'analisi, sembra opportuno sottolineare diversi aspetti non privi di interesse ai fini delle simulazioni, per meglio valutare le implicazioni del framework descritto nella sezione 3.5:

- ▷ il campionamento dalla predittiva e l'ordinamento delle realizzazioni secondo i ranghi è strumentale alla definizione delle distribuzioni empiriche per ciascuna sezione del processo aleatorio; le scelta del tipo di campionamento casuale semplice (con o senza ripetizione) dipende in larga misura dalle caratteristiche delle serie storiche osservate; ad esempio, nel caso di *difp* è evidente in molte unità di tempo la presenza di intensità esattamente pari a zero ed il campionamento senza ripetizione non potrebbe riprodurle,

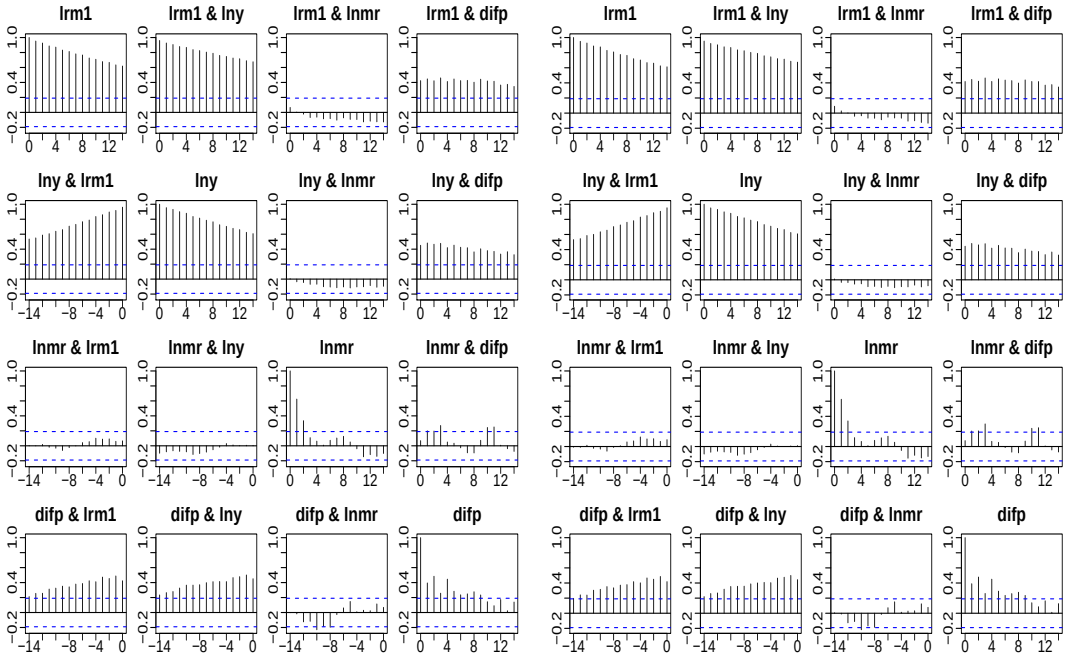
- ▷ i dati artificiali generati in ogni replicazione sono ottenuti come collezione delle realizzazioni ricavate campionando ciascuna sezione indipendentemente dalle altre; tale modo di procedere è coerente con l'idea di ritenere le unità statistiche nel campione parzialmente scambiabili, ossia scambiabili all'interno delle classi individuate dai ranghi; una utile conseguenza è l'attenuazione del vincolo di fedeltà alla sequenza di ranghi manifestata dal campione osservato,

- ▷ non è necessario condurre alcuna analisi preliminare volta ad identificare le componenti ciclo, trend, stagionalità delle serie; tali caratteristiche sono state riprodotte semplicemente accodando realizzazioni delle distribuzioni empiriche delle sezioni.

Proseguendo nell'indagine in discorso, è anche utile osservare l'andamento dei correlogrammi nei livelli, figura (4.3), e nelle differenze prime, figura (4.4). Per completezza sono stati inclusi, accanto ai diagrammi riferiti alle serie osservate, quelli relativi alle medie delle sezioni ricavate dalle traiettorie simulate. La circostanza che i secondi siano di fatto indistinguibili dai primi appare scontata alla luce della sovrapponibilità tra serie originarie e medie delle sezioni. È invece opportuna qualche considerazione di massima sulle possibili dimensioni dello spazio di cointegrazione. Una preliminare ispezione dei correlogrammi non sarebbe richiesta in quanto il metodo di Johansen basa le inferenze sulla considerazione delle complessive caratteristiche delle serie, a prescindere dalla loro natura individuale, ma è utile per fissare le idee circa alcuni elementi qualificanti dell'analisi. In questo senso, ignorando le correlazioni incrociate ai diversi ritardi, si nota come i diagrammi per *lrm1* e *lny* nei livelli siano caratterizzati da un lento decremento esponenziale mentre il comportamento nelle differenze prime è dominato dalla stagionalità trimestrale, così denunciando la probabile presenza di un trend stocastico. I correlogrammi per *lnmr* nelle differenze seguono un

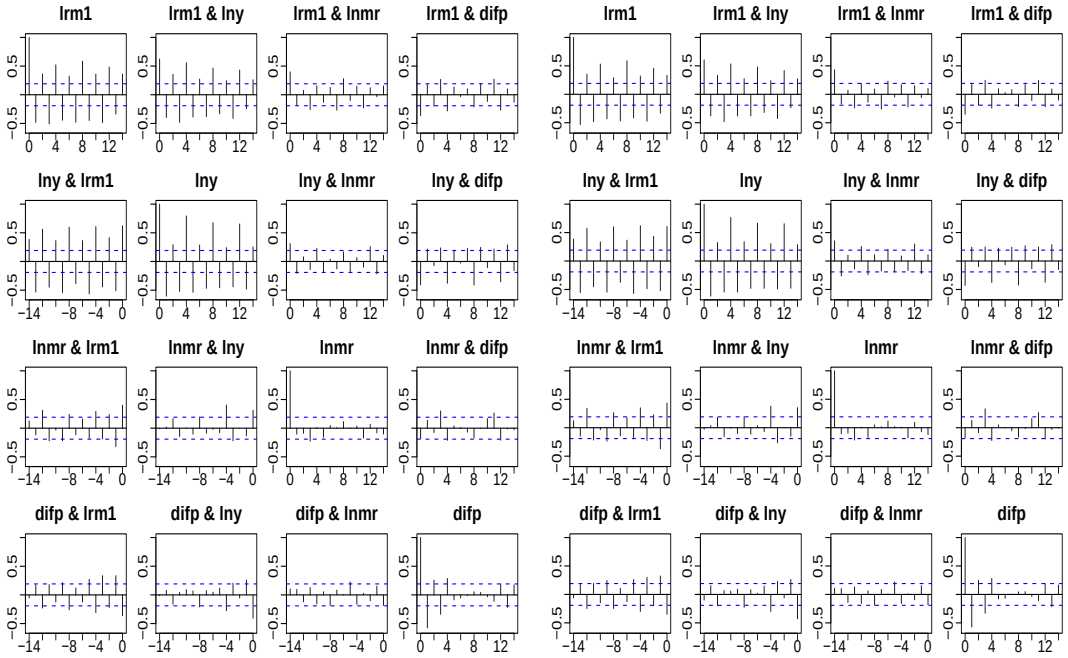
comportamento assimilabile ad un rumore bianco lasciando intuire problemi di sovra-differenziazione e dunque, stazionarietà nei livelli. Pare meno scontato il giudizio su *difp*: i picchi nei livelli non diminuiscono esponenzialmente ma sono apparentemente significativi fino al 9° ritardo, mentre le differenze prime parrebbero possedere una residua struttura autoregressiva.
La presumibile presenza di una serie stazionaria (*lnmr*) implica una dimensione *r* dello spazio di cointegrazione almeno pari ad uno. Se non esistessero combinazioni lineari stazionarie di *difp*, *lrm1* e *lny*, la dimensione rimarrebbe pari ad uno, mentre in caso contrario potrebbe risultare pari a due o tre. Questi ultimi due casi, ipoteticamente, possono essere generati da diverse circostanze:

- ▷ *difp* stazionaria e assenza di cointegrazione tra *lrm1* e *lny* (r=2)
- ▷ *difp* stazionaria e cointegrazione tra *lrm1* e *lny* (r=3)
- ▷ *difp* non stazionaria e cointegrata con *lrm1* e *lny* in modo da formare una (r=2) o due (r=3) combinazioni lineari stazionarie e linearmente indipendenti.



(a) Livelli osservati (b) Medie delle sezioni simulate

Figura 4.3: Correlogrammi dei livelli



(a) Livelli osservati (b) Medie delle sezioni simulate

Figura 4.4: Correlogrammi delle differenze prime

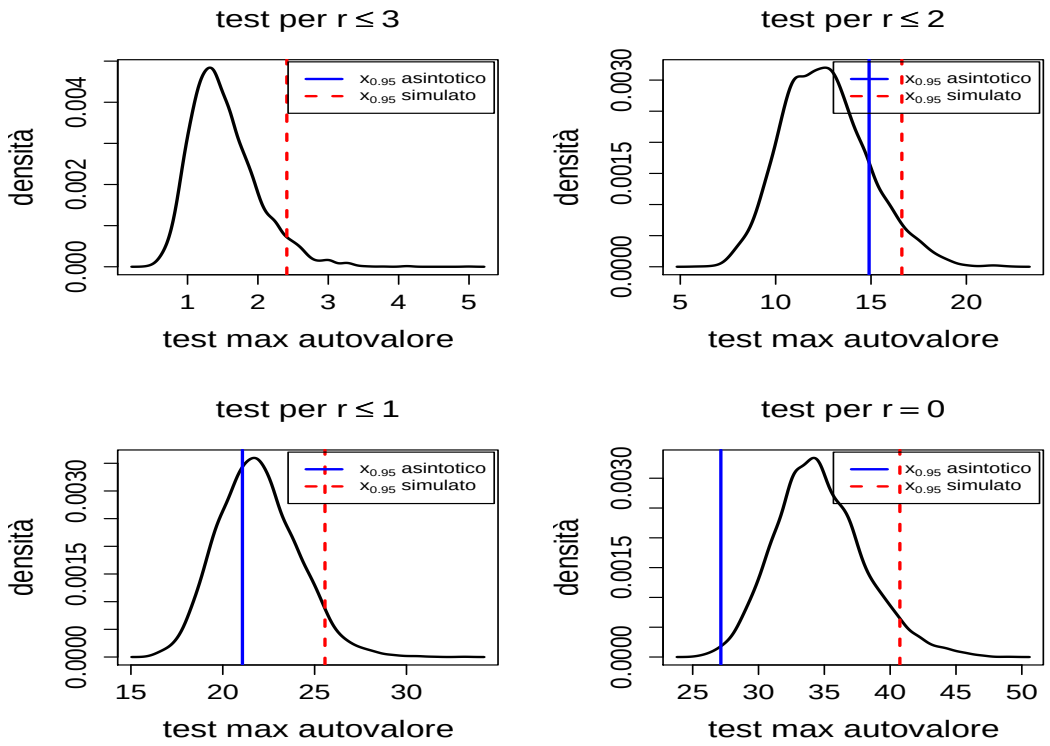
Le elaborazioni sono state eseguite facendo uso del package URCA (Pfaff, 2006). I valori critici per il test del massimo autovalore sui dati osservati sono riportati nella tabella (4.1) :

	test	q_{90}	q_{95}	q_{99}
$r \leq 3$	3.11	6.50	8.18	11.65
$r \leq 2$	7.89	12.91	14.90	19.19
$r \leq 1$	26.64	18.90	21.07	25.75
$r = 0$	38.49	24.78	27.14	32.14

Tabella 4.1: Test del massimo autovalore

Si nota che l'ipotesi $H_0 : r \leq 1$ viene rifiutata al livello di significatività $\alpha = 0.05$ mentre altrettanto non accade per $H_0 : r \leq 2$. L'esito del test della traccia non è diverso e per brevità viene omesso. In definitiva, parrebbe che sia $r = 2$. Ad analogo risultato giungono gli Autori. Tuttavia, Johansen e Juselius sulla scorta di considerazioni di natura economica e statistica, con particolare riferimento ai segni ed ai valori assoluti dai loadings che veicolano i residui di cointegrazione all'interno delle equazioni *VECM,* concludono per una dimensione $r = 3$. Generando 1000 campioni dalle sezioni delle quattro variabili coinvolte sono state simulate le distribuzioni per il test

del massimo autovalore. La figura (4.5) mostra i rispettivi diagrammi, nei quali sono evidenziati il 95^{mo}percentile asintotico e simulato (la figura relativa al test per $r \leq 3$ evidenzia solo quello simulato poiché il valore di 8.18 eccede l'ampiezza disponibile per la rappresentazione grafica).



(a) Medie delle sezioni simulate

Figura 4.5: Test di cointegrazione del massimo autovalore

Nel paragrafo 3.5 si è sottolineata la valenza interpretativa della procedura di ricampionamento in esame, quando siano coinvolte serie storiche non stazionarie. Per le considerazioni svolte ispezionando i correlogrammi, pur non potendo giungere a conclusioni certe in ordine alla dimensione r dello spazio generato dalle combinazioni lineari stazionarie delle serie, sembra di poter escludere che sia $r = 4$; tuttavia, questa sarebbe l'evidenza suggerita dal primo diagramma quando si confrontasse il valore della statistica test con il 95mo percentile della legge simulata. In realtà il condizionamento ai dati osservati, implicato dalla modellazione di ciascuna densità secondo i momenti sottesi dall'intero campione, implica il condizionamento delle distribuzioni simulate alla "vera" dimensione r^* dello spazio di cointegrazione. Per tale ragione i termini del confronto non possono essere rappresentati dal binomio statistica test e distribuzione simulata, ma debbono essere riconosciuti nella coppia formata dalle leggi

di probabilità simulata ed asintotica; la dimensione r^* implicita nei dati non costituisce una restrizione per la forma della distribuzione simulata solo quando quest'ultima è riferita all'ipotesi che r^* sia la dimensione dello spazio di cointegrazione. Le tavole per i test di massima verosimiglianza proposti da Johansen individuano i valori afferenti alle zone critiche usualmente d'interesse e dunque, per un confronto immediato, sono stati considerati i soli percentili corrispondenti ad $\alpha = 0.05$. In prima approssimazione, le distribuzioni più prossime appaiono essere quelle rilevanti per il rifiuto di $H_0 = r \leq 2$: la statistica test del massimo autovalore è funzione del terzo autovalore più grande, ossia quello legato alla circostanza che sia $r = 3$. Si era notato in precedenza che una delle serie, il logaritmo naturale del tasso ufficiale di sconto praticato dall'istituto di emissione, fosse molto probabilmente stazionaria. Sotto tale ipotesi, $r = 3$ significherebbe che le restanti tre variabili generano uno spazio di due dimensioni. Per compatibilità con gli andamenti dei correlogrammi, quest'ultimo potrebbe essere frutto di due stati di natura possibili: tre serie non stazionarie che si cointegrano ovvero una serie stazionaria assieme a due non stazionarie che combinandosi linearmente "perdono" la radice unitaria. Dunque, la questione da chiarire riguarda la natura del tasso d'inflazione, la serie a proposito della quale i correlogrammi non avevano permesso di formulare giudizi netti. Il test delle ipotesi adatto allo scopo concerne le restrizioni alle quali i coefficienti β che individuano il vettore di cointegrazione dovrebbero soggiacere se *difp* fosse stazionaria. Usando la notazione di Johansen, posto che φ sia una matrice di coefficienti con un numero di colonne pari ad r ed un numero di righe pari a s (in questo caso è s=3), la relazione

$$\beta = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \varphi$$

rappresenta la condizione che *lrm1* e *lny* si cointegrino avendo coefficienti uguali in valore assoluto ma di segno opposto, con *lnmr* e *difp* stazionarie. La densità asintotica rilevante per la prova dell'ipotesi è un $\chi^2_{r(p-s)}$. Il valore del test, pari a 3.82 non permette di rifiutare la nulla. Come per il test del massimo autovalore, con 1000 replicazioni ottenute campionando le sezioni ricavate dalle traiettorie generate dalla predittiva, è stata simulata la densità del test in modo da rendere possibile il confronto tra i percentili che isolano una zona critica corrisponendente ad $\alpha = 0.05$. La prossimità evidenziata nella figura (4.6) indica che rispetto ai dati, la matrice esprimente i vincoli su β non impone restrizioni significative, tali da abbattere il valore della logverosimiglianza. Queste conclusioni potrebbero essere derivate anche per altra via; infatti, un

utile prodotto del ricampionamento è costituito dalle distribuzioni dei coefficienti per ciascuna delle equazioni considerate. Le tabelle (4.2) e (4.3) mostrano i coefficienti stimati per le variabili nei livelli e, rispettivamente, per quelle nelle differenze. Questi ultimi riflettono la dinamica di breve periodo delle serie, mentre i primi, ottenuti dal prodotto dei loadings per i β afferiscono agli aggiustamenti di lungo periodo.

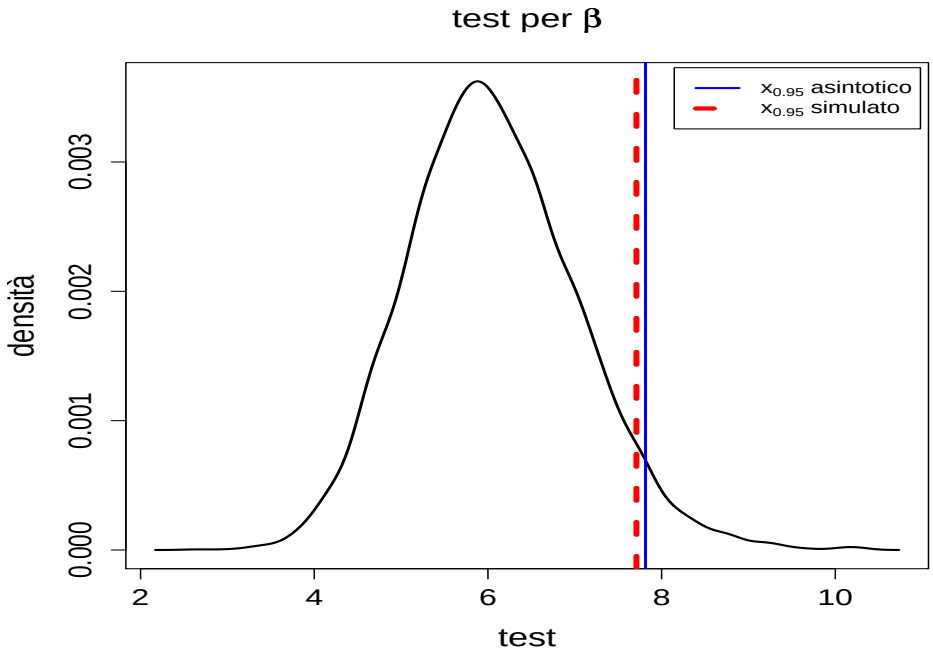


Figura 4.6: Test sulle restrizioni per β

	lrm1.l2	lny.l2	lnmr.l2	difp.l2
lrm1.d	-0.119	0.110	-0.227	-0.418
lny.d	0.024	-0.036	-0.136	-0.214
lnmr.d	0.095	-0.104	-0.479	0.670
difp.d	0.004	0.007	-0.003	-0.484

Tabella 4.2: Coefficienti dei livelli

Le argomentazioni esposte da Johansen e Juselius circa il significato economico delle relazioni individuate fanno riferimento ai coefficienti, prescindendo dalla considerazione dei relativi intervalli di confidenza. Le stime degli intervalli *h.p.d.* al livello di significatività $\alpha = 0.05$ sono esposte nella tabella (4.4).

La variabile $difp_{t-2}$ (ultima riga della tabella (4.4)) è, come sospettato, non significativa nelle prime due equazioni. Nelle figure (4.7), (4.8), (4.9) e (4.10) sono

	lrm1.dl1	lny.dl1	lnmr.dl1	difp.dl1
lrm1.d	-0.209	-0.217	-0.099	-0.012
lny.d	0.044	-0.689	-0.059	-0.158
lnmr.d	-0.161	-0.002	-0.234	0.415
difp.d	0.009	0.024	0.023	-0.822

Tabella 4.3: Coefficienti delle differenze prime

	lrm1.d		**lny.d**		**lnmr.d**		**difp.d**	
	inf	sup	inf	sup	inf	sup	inf	sup
constant	-0.203	0.029	0.056	0.238	0.122	0.242	-0.048	-0.019
sd1	0.015	0.073	0.033	0.072	-0.013	0.013	-0.004	0.003
sd2	0.024	0.061	0.066	0.095	0.010	0.020	-0.009	-0.006
sd3	0.083	0.120	0.081	0.120	0.011	0.025	-0.011	-0.006
lrm1.dl1	-0.635	-0.230	-0.083	0.143	-0.197	-0.070	-0.008	0.029
lny.dl1	-0.217	0.105	-0.726	-0.422	-0.177	0.043	-0.021	0.031
lnmr.dl1	-0.118	0.175	-0.250	0.030	-0.265	-0.095	0.004	0.034
difp.dl1	-0.801	0.480	-0.407	0.459	0.198	0.706	-0.882	-0.795
lrm1.l2	-0.245	-0.055	-0.031	0.131	0.036	0.123	-0.006	0.018
lny.l2	0.047	0.233	-0.143	0.015	-0.131	-0.047	-0.006	0.017
lnmr.l2	-0.241	-0.053	-0.223	-0.026	-0.518	-0.429	-0.010	0.009
difp.l2	-1.103	0.020	-0.432	0.451	0.396	0.901	-0.567	-0.441

Tabella 4.4: IC h.p.d. per i β_j, $\alpha = 0.05$

sintetizzate le distribuzioni empiriche per i parametri di ciascuna equazione. Nonostante siano difficilmente suscettibili di interpretazione economica, per completezza sono state incluse anche quelle relative al nucleo deterministico (intercetta e dummies stagionali) ed alla dinamica di breve periodo (ritardi delle differenze prime). Circa le variabili espresse nei livelli, maggiormente interessanti ai fini dell'esercizio, ispezionando i diagrammi per le prime due equazioni risulta confermato che:

- ▷ $difp_{t-2}$ non offre contributi alla cointegrazione tra *lrm1* e *lny* figurandovi con coefficienti non significativamente diversi da zero,
- ▷ appaiono simili ma speculari le distribuzioni di $lrm1_{t-2}$ e lny_{t-2} (righe terz'ultima e quart'ultima della tabella (4.4).

In definitiva, come sarebbe lecito aspettarsi, gli esiti del test concernente le restrizioni su β sono riflessi qualitativamente dalle distribuzioni empiriche ricavate per i coefficienti delle equazioni.

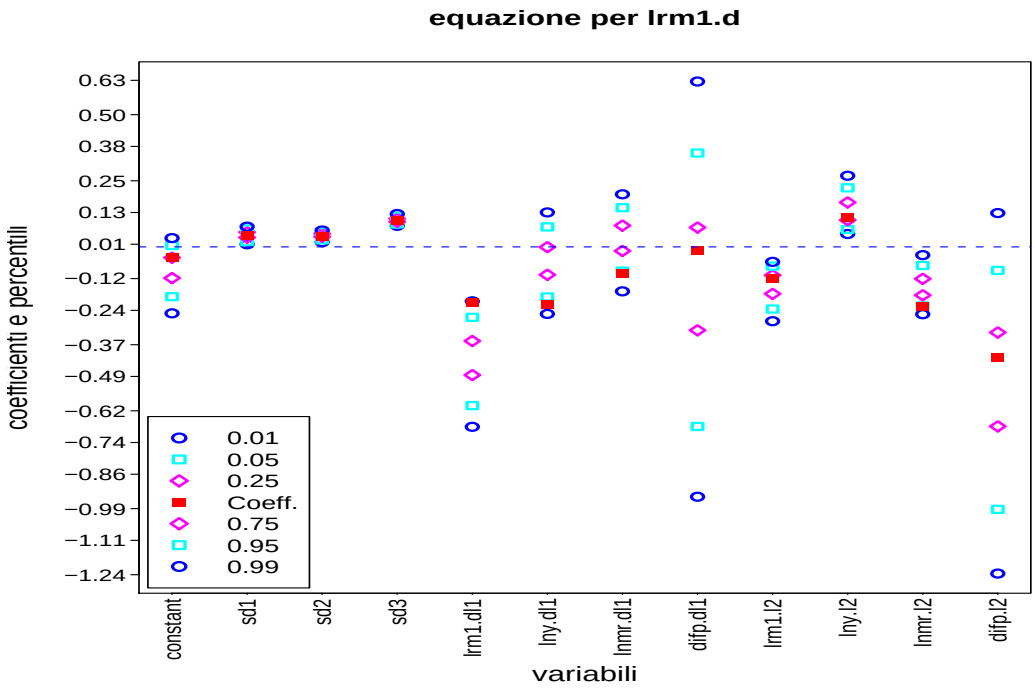


Figura 4.7: Coefficienti e percentili della 1ªequazione del VECM

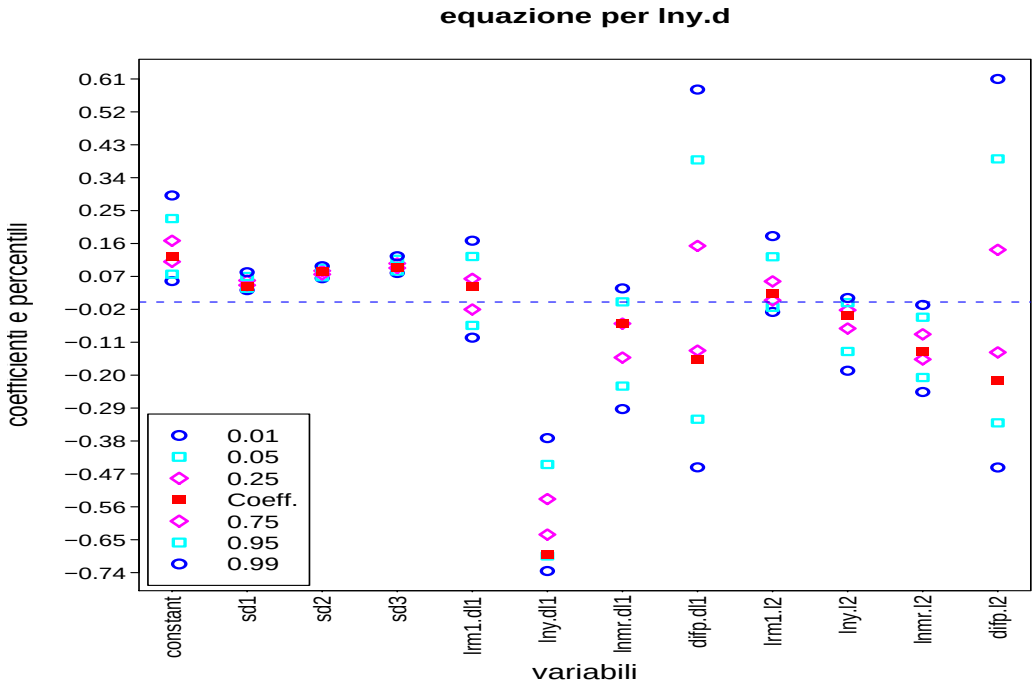


Figura 4.8: Coefficienti e percentili della 2ªequazione del VECM

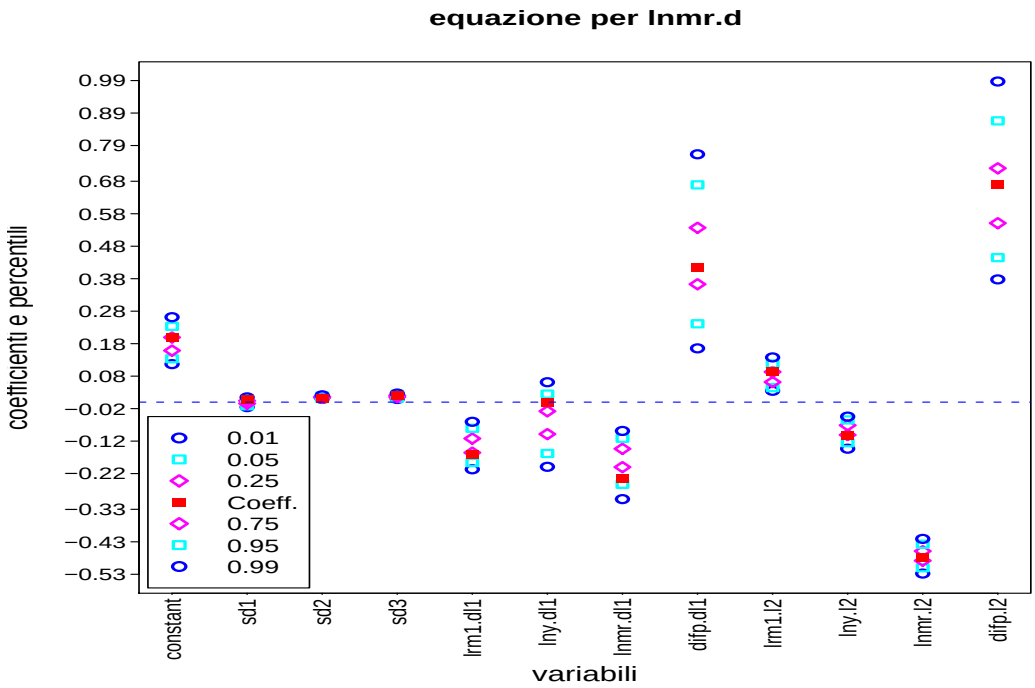


Figura 4.9: Coefficienti e percentili della 3ªequazione del VECM

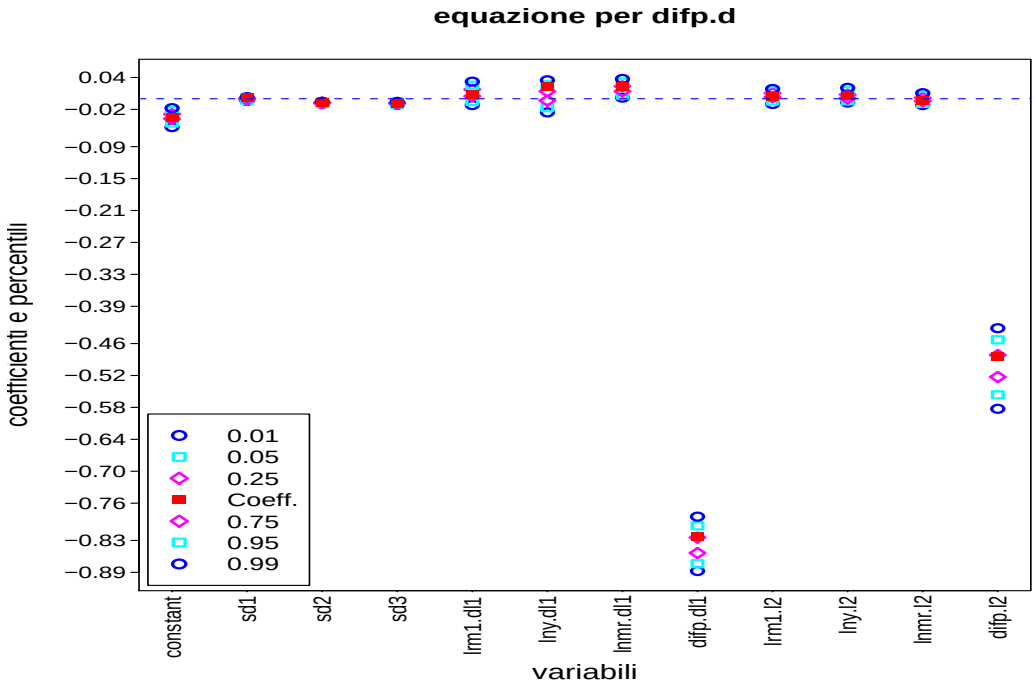


Figura 4.10: Coefficienti e percentili della 4ªequazione del VECM

4.2 Dati Sintetici per la Tutela della Privacy

Le informazioni raccolte dagli istituti di statistica rappresentano un patrimonio informativo della collettività. Accanto ai risultati delle indagini, si percepisce sempre di più la necessità di mettere a disposizione degli utenti i dati raccolti. La diffusione di essi tuttavia non può violare i vincoli di riservatezza imposti da fonti normative nazionali e sovranazionali e da obblighi contrattuali. Seguendo l'Hanbook on Statistical Disclosure Control (AA.VV., 2007), le intrusioni in informazioni non pubbliche (non necessariamente sensibili) si concretizzano quando tramite i dati sia possibile acquisire notizie in merito a specifiche unità statistiche e possono essere di due tipi: identificazione del rispondente o associazione ad una persona o organismo di dati noti dalle indagini. Limitando l'attenzione al rilascio di microdati, l'intrusione avviene quando il singolo record è correttamente assegnato ad un record contenuto in un file esterno disponibile all'intrusore. Uno scenario particolarmente critico si manifesta per dati di natura economica, a causa della sparsità che li caratterizza, ed oggetto dell'identificazione sono frequentemente le imprese; benché molte notizie possano essere acquisite dai bilanci ordinari d'esercizio (ossia da quella informazione esterna d'azienda sulla quale i potenziali investitori fondano le proprie scelte di destinazione dei capitali finanziari) che la riservatezza - anche in questi casi - debba essere tutelata non deve stupire: spesso le indagini possono contenere dettagli superiori a quelli disponibili grazie ai documenti pubblici e occorre comunque ottemperare ai vincoli riconosciuti dagli istituti di statistica verso i rispondenti. La sparsità, intesa come scarsa fungibilità delle unità statistiche, conduce a due ordini di problemi. Il primo riguarda la facilità di identificazione sulla base di poche mutabili e/o variabili: si pensi ad esempio ad imprese che operano sul territorio in condizioni di sostanziale monopolio, ovvero al numero di addetti o al fatturato di quelle che occupano una posizione dominante rispetto alle altre dello stesso settore. Il secondo problema concerne la possibilità di fornire dati sintetici tali da conservare le proprietà statistiche dei dati originali (una volta che sia stato convenuto in cosa consistano) ed è facile notare come le distribuzioni volte a mimare le leggi di probabilità dovrebbero essere profondamente diverse da quelle usualmente considerate nella statistica parametrica, ad esempio in presenza di supporti misti. La capacità di generare dati artificiali non offre di per sé una soluzione ai problemi di tutela della riservatezza dato che le unità statistiche simulate potrebbero in molti casi essere facilmente associate ai veri rispondenti, ma può costituire un punto di partenza per l'implementazione di strategie protettive maggiormente articolate. Un vantaggio appare tuttavia evidente e risiede nel rendere indipendenti le numerosità del campione simulato e del dataset originale: in linea di principio all'aumentare del numero di unità statistiche generate, le operazioni di record linkage

(Liseo *e altri*, 2006) possono divenire computazionalmente onerose.

La procedura di ricampionamento dalla predittiva, illustrata nel cap. 3, è stata applicata ai redditi delle famiglie italiane rilevati nell'Indagine sui Bilanci delle Famiglie 2006 (per brevità I.B.F. 06), resa disponibile dal Servizio delle Statistiche Economiche e Finanziarie della Banca d'Italia all'indirizzo `http://www.bancaditalia.it/statistiche/indcamp/bilfait/dismicro`. L'esercizio è parso interessante per vagliare la bontà del metodo nel riprodurre le caratteristiche di variabili estremamante sparse, quando si convenga che l'indicatore di qualità in ordine alle proprietà da conservare sia rappresentato dal mantenimento dei momenti puri e misti entro un certo ordine. L'assenza di finalità interpretative socioeconomiche, ha indotto a ritenere sufficiente la considerazione di un limitato sottoinsieme delle informazioni disponibili, ottenuto per aggregazione secondo la tipologia di reddito; parimenti, non si è tenuto conto dei pesi associati alle unità incluse nel campione. Ciò premesso, costituiscono oggetto dell'esercizio 7768 osservazioni riferite a cinque variabili espresse in unità di euro:

- ▷ Ycf: redditi da capitali finanziari,
- ▷ Yl: redditi da lavoro dipendente,
- ▷ Yt: pensioni e trasferimenti netti,
- ▷ Ym: redditi netti da lavoro autonomo,
- ▷ Yca: redditi da fabbricati.

Le tabelle (4.5), (4.6) e (4.7) forniscono sintesi delle principali caratteristiche empiriche dei dati considerati, mostrandone rispettivamente indici di posizione, dimensione, dispersione, forma e matrice delle correlazioni. Una volta stimate le predittive delle variabili, ricorrendo a 1000 replicazioni sono state approssimare le distribuzioni delle misure di sintesi.

	Ycf	Yl	Yt	Ym	Yca
min	-25432.36	0.00	0.00	-20000.00	0.00
0.25	3.42	0.00	0.00	0.00	2400.00
0.5	67.25	6105.00	6518.00	0.00	6000.00
0.75	298.20	20000.00	14690.00	0.00	8400.00
max	99789.76	251000.00	429770.00	800000.00	152000.00

Tabella 4.5: Dati originali: indici di posizione

	Ycf	Yl	Yt	Ym	Yca
media	280.54	12054.03	8781.26	4327.05	6564.37
dev. std.	2742.24	15355.02	11194.32	19483.57	7457.06
asimmetria	11.44	2.11	7.88	18.21	5.66
curtosi	322.74	17.07	261.62	549.23	72.47

Tabella 4.6: Dati originali: indici di dimensione, dispersione, forma

	Ycf	Yl	Yt	Ym	Yca
Ycf	1.00	-0.02	0.15	0.12	0.25
Yl	-0.02	1.00	-0.37	-0.07	0.11
Yt	0.15	-0.37	1.00	-0.09	0.16
Ym	0.12	-0.07	-0.09	1.00	0.24
Yca	0.25	0.11	0.16	0.24	1.00

Tabella 4.7: Dati originali: correlazioni

Delle risultanti distribuzioni empiriche, nelle tabelle (4.8), (4.9) e (4.10) sono indicati minimi, massimi, percentili che isolano in ciascuna coda una probabilità del 2.5% e valori attesi. Confrontando le tabelle (4.8) e (4.7) appaiono riprodotte in modo soddisfacente le correlazioni tra le variabili, circostanza importante ai fini della validazione in ambito multivariato della procedura delineata nel paragrafo 3.5.

ρ	min	$q_{2.5}$	m_1	$q_{97.5}$	max
Ycf, Yl	-0.04	-0.03	-0.02	-0.01	0.00
Ycf, Yt	0.13	0.14	0.17	0.20	0.21
Ycf, Ym	0.09	0.11	0.13	0.16	0.17
Ycf, Yca	0.19	0.22	0.25	0.28	0.30
Yl, Yt	-0.43	-0.42	-0.38	-0.31	-0.27
Yl, Ym	-0.09	-0.09	-0.07	-0.06	-0.05
Yl, Yca	0.10	0.11	0.12	0.13	0.14
Yt, Ym	-0.14	-0.13	-0.10	-0.07	-0.06
Yt, Yca	0.14	0.15	0.17	0.18	0.18
Ym, Yca	0.20	0.21	0.23	0.25	0.25

Tabella 4.8: Sintesi per i coefficienti di correlazione

Qualche indicazione in ordine al mantenimento di legami di dipendenza più complessi si può trarre dalla figura (4.11) (in ascissa figurano gli ordini dei momenti misti,

convenendo che le potenze siano riferite rispettivamente a Ycf, Yl, Yt, Ym, Yca). L'ipotesi sottoposta a test può essere espressa come:

$$H_0 : \frac{\widehat{\mu}_{a,b,c,d,e}}{\mu_{a,b,c,d,e}} - 1 = 0 \quad a,b,c,d,e \in \{1,2,3,4\}$$

Ammettendo a fini di semplicità un abuso di notazione, con $\mu_{a,b,c,d,e}$ e $\widehat{\mu}_{a,b,c,d,e}$ sono stati indicati i momenti misti calcolati dal campione osservato e, rispettivamente, dalle distribuzioni simulate. Quelli sottoposti a verifica sono stati selezionati assegnando casualmente a ciascuna delle cinque variabili una potenza espressa da un numero intero compreso tra uno a quattro, estremi inclusi. Gli intervalli *h.p.d.* sono stati ricavati numericamente dalla distribuzione empirica ottenuta calcolando in ognuna delle 1000 replicazioni la differenza percentuale d'interesse. Ad un livello di significatività $\alpha = 0.05$ gli intervalli di credibilità coprono un range di valori molto esteso, mentre risultano sensibilmente più contenuti per $\alpha = 0.1$.

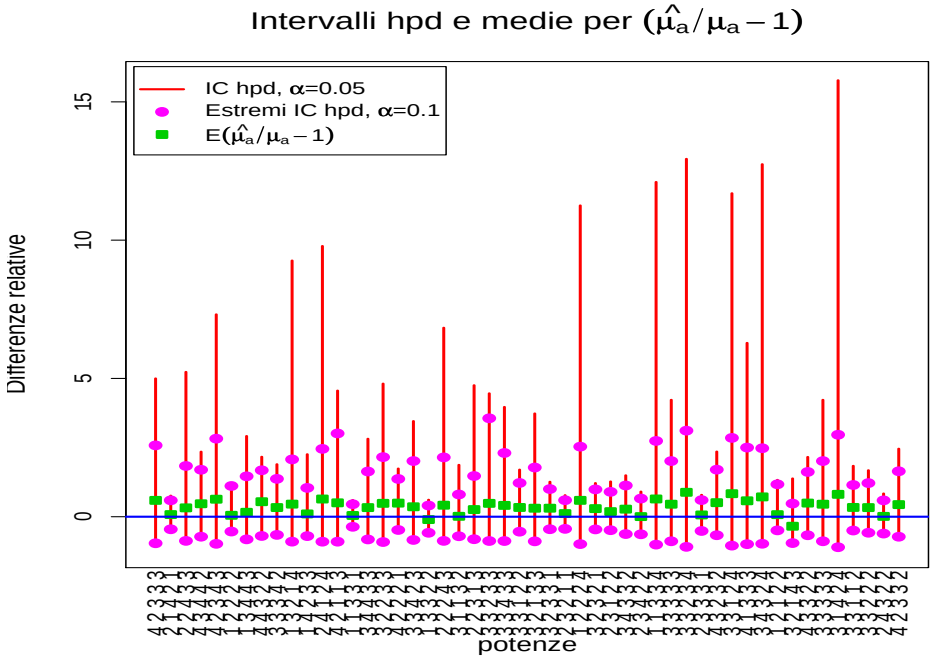


Figura 4.11: Intervalli h.p.d. per i momenti misti

Ispezionando il diagramma (4.11) si può notare che gli intervalli di maggiore ampiezza coincidono con esponenti superiori a due per la variabile Yca. Questa circostanza potrebbe indicare la presenza di pochi valori estremi che, influenzando gli indici di forma, portano ad una distribuzione simulata tale da generare valori elevati più spesso di quanto non accada nel campione osservato. Tale ipotesi, come vedremo

tra breve, appare corroborata dalla congiunta considerazione delle distribuzioni empiriche per indici di posizione e di forma. Concentrando l'attenzione su un contesto univariato, le peculiarità dei dati di indagine sembrano essere imitate dalle unità statistiche simulate in modo complessivamente soddisfacente. Qualche scostamento si può notare per alcuni indici di posizione. In particolare, dal raffronto con la tabella (4.5) si nota che il range del massimo per Yca non contiene il massimo osservato e lo stesso accade, sia pure con scarti meno ampi, per le variabili Yl e Ym.

	min	q_{25}	q_{50}	q_{75}	max
Ycf					
min	-28257.36	0.00	50.24	229.20	27771.46
$q_{2.5}$	-23158.84	0.00	51.54	239.50	42553.21
m_1	-17948.15	0.00	54.92	262.42	81234.94
$q_{97.5}$	-13384.18	0.00	57.82	289.42	112427.36
max	-10848.21	0.00	59.11	308.15	112427.36
Yl					
min	0.00	0.00	0.00	19619.82	98560.53
$q_{2.5}$	0.00	0.00	2020.13	19962.03	110944.37
m_1	0.00	0.00	5390.09	20528.70	221795.21
$q_{97.5}$	0.00	0.00	7583.97	21258.38	230806.53
max	0.00	0.00	8405.62	21726.96	230806.53
Yt					
min	0.00	0.00	5591.32	14470.53	62721.80
$q_{2.5}$	0.00	0.00	5948.45	14594.90	67222.48
m_1	0.00	0.00	6347.69	14953.50	280930.61
$q_{97.5}$	0.00	0.00	6801.58	15333.19	503341.67
max	0.00	0.00	7148.55	15558.07	503341.67
Ym					
min	-22100.65	0.00	0.00	0.00	252094.41
$q_{2.5}$	-22100.65	0.00	0.00	0.00	329144.88
m_1	-17085.07	0.00	0.00	0.00	703471.58
$q_{97.5}$	0.00	0.00	0.00	0.00	794532.02
max	0.00	0.00	0.00	0.00	794532.02
Yca					
min	0.00	0.00	5952.19	8458.65	90919.33
$q_{2.5}$	0.00	956.75	6016.52	8589.31	131798.61
m_1	0.00	1451.85	6137.06	8787.73	138119.63
$q_{97.5}$	0.00	1906.86	6249.53	8966.17	138433.38
max	0.00	2110.72	6297.85	9073.17	138433.38

Tabella 4.9: Sintesi per gli indici di posizione

Inoltre, il range per il terzo quartile di Yca ha un minimo che eccede la stima nel

campione così come i valori simulati in riferimento al primo quartile ed alla mediana di Ycf sono carenti nel riprodurre le grandezze omologhe calcolate secondo i dati.

Tuttavia, asimmetria e curtosi relative alle distribuzioni sintetiche non sembrano risentire dei disallineamenti negli indici di posizione e i valori attesi sono prossimi a quelli misurabili a partire dalle unità osservate, come si evince confrontando le tabelle (4.10) e (4.6).

	media	dev. std.	asimmetria	curtosi
Ycf				
min	197.84	2086.22	3.48	43.14
$q_{2.5}$	238.09	2272.87	4.96	75.34
m_1	290.82	2715.05	10.37	270.19
$q_{97.5}$	352.69	3269.33	17.39	577.95
max	397.12	3704.09	19.72	705.77
Yl				
min	11456.17	14492.54	1.16	4.23
$q_{2.5}$	11709.68	14739.83	1.33	5.25
m_1	12046.54	15333.80	2.07	16.61
$q_{97.5}$	12386.28	16011.83	2.88	28.58
max	12590.03	16664.67	3.35	34.25
Yt				
min	8424.46	9934.32	1.16	4.29
$q_{2.5}$	8557.74	10056.04	1.21	4.59
m_1	8785.83	11179.43	6.40	219.03
$q_{97.5}$	9031.53	14052.15	16.86	572.38
max	9168.29	16190.40	19.16	615.84
Ym				
min	3767.36	13050.51	5.95	60.98
$q_{2.5}$	3928.63	14638.73	7.82	98.44
m_1	4320.22	18895.99	15.80	454.75
$q_{97.5}$	4779.64	24988.68	22.50	748.44
max	5381.89	28783.73	25.06	863.44
Yca				
min	6267.34	6463.97	2.36	16.14
$q_{2.5}$	6401.54	6748.52	3.69	43.59
m_1	6567.86	7443.60	5.56	77.10
$q_{97.5}$	6717.94	8186.09	6.92	99.10
max	6830.95	8696.43	7.47	106.74

Tabella 4.10: Sintesi per gli indici di dimensione, dispersione, forma

Qualche ulteriore riflessione può essere proposta ordinando in successione non decrescente le intensità di Yca e calcolando gli incrementi di ciascuna rispetto a quella

che immediatamente la precede. Così facendo vengono evidenziate diverse unità statistiche sensibilmente "distanti" dalle altre. In particolare, la figura (4.12) mostra che quella identificante il valore massimo per Yca è anche caratterizzata dal più grande incremento di intensità rispetto all'unità più prossima della successione ordinata. In definitiva, i dati sembrano caratterizzati da una sparsità che in parte non viene catturata dalle distribuzioni simulate; le circostanze osservate a proposito degli indici di posizione sembrano dipendere dall'aver ripartito lungo segmenti di supporto delle masse che nei dati sono concentrate su singoli punti.

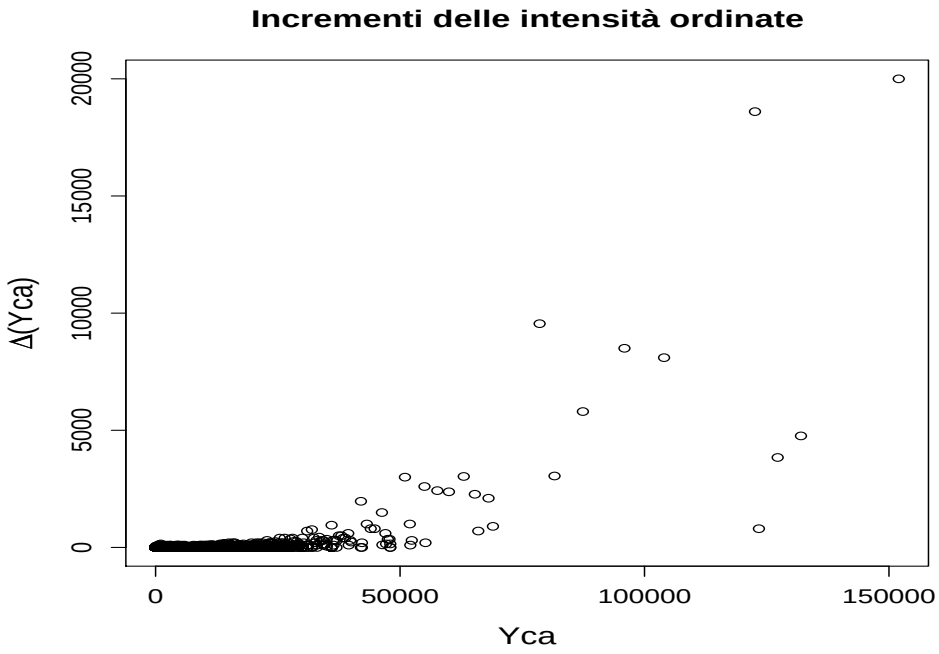


Figura 4.12: I.B.F. 06 (fonte: Banca d'Italia), campione dei redditi da fabbricati

Capitolo 5

Conclusioni

Il concetto di ricampionamento appare estraneo all'ambito bayesiano: i processi di stima sono fondati sul principio di verosimiglianza, in conflitto con la considerazione di ipotetiche prove ripetute, tipica dell'impostazione frequentista. Il bootstrap bayesiano (Rubin, 1981) non ricampiona unità statistiche ma modella la probabilità di averle osservate secondo una procedura pienamente conforme al principio di verosimiglianza: attraverso l'utilizzo di una legge multinomiale e di una priori Dirichlet (non informativa) si ricava una posteriori Dirichlet tale che il valore atteso della probabilità associata a ciascuna osservazione sia pari alla rispettiva frequenza relativa. Le unità statistiche e le intensità osservate su di esse non sono coinvolte in tale modo di agire. Muliere e Secchi (1994) presentano una generalizzazione del bootstrap di Rubin ricorrendo alla modellazione delle posteriori mediante un processo di Dirichlet; in questo modo è possibile tener conto di una distribuzione base idonea ad assegnare probabilità non nulle alle intensità non rilevate. Tuttavia, entrambi gli approcci non spiegano come capitalizzare eventuali conoscenze a priori di rango inferiore ossia tali da esprimere informazioni parziali in ordine al fenomeno indagato. Una possibilità consiste nel definire i valori attesi a posteriori delle probabilità associate alle osservazioni incluse nel campione tramite sistemi di pesi individuati dalle verosimiglianze empiriche generalizzate (Newey e Smith, 2004), così ottenendo l'equivalente bayesiano del bootstrap intenzionalmente distorto (Hall e Presnell, 1999). Questa opportunità è stata verificata rilassando l'ipotesi di priori sulla legge di probabilità in priori su alcuni momenti. Le simulazioni effettuate riprendendo l'esercizio di Muliere e Secchi circa la stima dei tre quartili di una Gamma(20,1) evidenziano l'accorciamento degli intervalli di credibilità all'aumentare del peso attribuito alla distribuzione di massima entropia assunta come base. Benché il controllo del parametro di concentrazione metta al riparo da problemi di inconsistenza quando le conoscenze a priori siano errate, si è indagata la possibilità di inferenze completamente oggettive tramite il solo contenuto informativo

del campione al fine di ottenere una regolarizzazione del bootstrap di Rubin. A questo scopo, le funzioni di ripartizione ricavate da distribuzioni di massima entropia costruite sui momenti campionari sono state usate quali termini di mistura con la funzione di ripartizione empirica. Si è osservato che al crescere del peso della base l'ampiezza degli intervalli diminuisce, ma peggiora la copertura effettiva rispetto al livello nominale; ciò è dovuto alla sottostima della dispersione per le posteriori degli indici di posizione, ossia al fatto che gli errori nelle stime dei momenti non vengono scontati dai passi di bootstrap: questi ultimi permettono soltanto di tener conto dell'errore di campionamento implicito nell'uso - ai fini delle stime - delle unità selezionate invertendo la funzione di ripartizione mistura. Dalle evidenze acquisite nel paragrafo 3.2 si possono trarre due indicazioni:

- ▷ il ricampionamento effettuato per mezzo dei processi di Dirichlet può essere esteso a situazioni più generali di quelle postulate da Muliere e Secchi, scontando le conoscenze a priori - non necessariamete definendo la legge che ha generato i dati - tramite verosimiglianze empiriche generalizzate che assumano il ruolo di legge base,

- ▷ il perseguimento congiunto di inferenze completamente oggettive e regolarizzazione del bootstrap non sembra poter beneficiare della costruzione di processi di Dirichlet basati su *GEL*: salvo che le priori siano delta di Dirac, per tenere conto dell'alea sui parametri delle verosimiglianze empiriche è nessario ridurre la concentrazione del processo, circostanza che rappresenta una contraddizione dello scopo.

In ottica bayesiana l'incertezza attorno ai parametri di una legge di probabilità viene rappresentata dalla priori ed è trasferita nelle stime per mezzo della posteriori; la possibilità di fare ricampionamento superando i limiti del bootstrap di Rubin senza utilizzare priori informative, sembra possa essere ricercata nell'alveo dell'impostazione bayesiana detta (Piccinato, 1996) completamente predittiva. Quest'ultima, se si prescinde dalla suggestione temporale insita nella parola previsione è assimilabile all'idea di ricampionamento in senso classico. Tuttavia, esiste una diversità concettuale. L'inferenza classica, in presenza di campioni costituiti da unità statistiche indipendenti, è resa possibile dall'ulteriore ipotesi di somiglianza. Nell'impostazione completamente predittiva, la distinzione tra grandezze osservabili e non osservabili ed il tentativo di fondare unicamente attraverso la probabilizzazione delle prime l'apprendimento delle leggi che governano i fenomeni oggetto di indagine portano all'introduzione del concetto di scambiabilità: la legge di probabilità sulle osservazioni induce una legge di probabilità su un parametro condizionatamente al quale le osservazioni sono

indipendenti e somiglianti. Nel caso più semplice, relativo ad una variabile casuale dicotomica, al divergere della numerosità campionaria è possibile apprendere il parametro (probabilità di successo) come frequenza relativa; la funzione di ripartizione del parametro esprime il grado di fiducia rispetto ai valori che la frequenza relativa può assumere; per la scambiabilità le osservazioni condizionatamente al parametro sono indipendenti e somiglianti, dunque è possibile risalire alla legge sulle osservazioni integrando la congiunta rispetto al parametro.

Nel caso di variabili continue il parametro non coincide più con uno scalare ma è, a sua volta, una funzione di ripartizione; limitando l'attenzione al caso univariato, la legge sulle osservazioni induce una legge di probabilità sulle funzioni di ripartizione in $\mathbb{R}$. Condizionatamente alle realizzazioni di tale legge, le unità incluse nel campione sono indipendenti e somiglianti ed è possibile, in linea di principio costruire la funzione di ripartizione predittiva. Per rendere operativo il teorema è necessario imporre delle restrizioni al supporto (i cui elementi sono funzioni di ripartizione). Gli esempi in proposito sono diversi, in genere accomunati dall'aggiungere alla scambiabilità altre ipotesi di invarianza quali ad esempio la simmetria sferica, ossia l'invarianza rispetto a rotazioni del sistema di riferimento (Bernardo e Smith, 2000).

Nel paragrafo 3.4 è stata proposta una soluzione diversa per la qualificazione del supporto. Essa estende direttamente la versione del teorema per il caso dicotomico alle variabili continue, interpretando il metodo dei momenti di K. Pearson alla stregua di un metodo di stima semiparametrico per leggi di probabilità, volto a minimizzare la distanza tra le frequenze relative osservate ed una distribuzione interpolante. Si è potuto osservare che

- ▷ al pari delle frequenze relative per una variabile dicotomica i momenti possono essere "appresi" al divergere della numerosità campionaria, circostanza significativa del fatto che sia legittimo interpretarli come parametro indotto dalla legge sulle osservazioni,

- ▷ potendo modellare verosimiglianze empiriche sulla base dei momenti, ci si riconduce alla situazione in cui il parametro condizionatamente al quale gli eventi sono scambiabili ha dimensione finita,

- ▷ la restrizione del supporto consiste nell'attribuzione di probabilità nulle solo alle funzioni di ripartizione che, nella classe di quelle consistenti con i momenti, risultino dominate - dal punto di vista della prossimità alla funzione di ripartizione empirica - secondo la metrica indotta dalla discrepanza di Cressie-Read di parametro γ.

In campo bayesiano, le ricerche volte all'uso di verosimiglianze modellate sulla base di informazioni incomplete hanno ricevuto nuovo impulso in tempi relativamente recenti e costituiscono tutt'ora materia di riflessione. Schennach (2005) individua una possibile soluzione nell'uso di verosimiglianze di massima entropia (Jaynes, 1957). Grendar e Judge (2007) illustrano la consistenza in *L-divergenza* delle mode a posteriori alle verosimiglianze empiriche (Owen, 1988). L'approccio delineato impone la soluzione di diversi problemi, sotto l'ipotesi che la legge (ignota) di probabilità ammetta momenti finiti. Senza rispecchiare un ordine necessariamente cronologico appaiono rilevanti alcune questioni:

1. quanti momenti includere nelle condizioni di ortogonalità ?
2. Come scegliere il parametro γ, ossia la metrica secondo la quale misurare la divergenza tra distribuzione multinomiale e verosimiglianza empirica generalizzata?
3. Scelto γ, quali sono le proprietà delle *GEL* costruite sulla base di stime dei momenti?
4. Come si può giustificarne l'uso in ambito bayesiano?
5. Come deve avvenire l'eliminazione dei parametri di disturbo ai fini del calcolo della predittiva?

▷ Circa la prima domanda, Lindsay e Basak (2000) forniscono un'espressione per l'estremo superiore della distanza tra leggi di probabilità discrete aventi in comune i primi k momenti; fissato un punto del supporto essa diminuisce all'aumentare del numero di momenti condivisi mentre, a parità di questi, decresce nelle zone via via più prossime alle code. Nel paragrafo 3.2.1 si è osservato che i test di uguaglianza tra distribuzioni, quando i termini del confronto siano due *GEL* che incorporano un diverso numero di condizioni sui momenti, si traducono in test sulla significativa diminuzione dell'errore di approssimazione rispetto alla vera (ed ignota) legge di probabilità. Ai test classici si aggiunge una ulteriore possibilità rappresentata dall'uso della discrepanza di Cressie-Read di parametro $\gamma = 0$, proporzionale alla D_{KL}. Nei paragrafi A.1 e A.2 dell'appendice si mostra che essa, sotto l'ipotesi nulla (identità delle due distribuzioni), segue asintoticamente una legge χ^2_c, ove c è il numero di condizioni di ortogonalità suppletive: con $1 \leq h \leq c$, la $(q-c+h)^{ma}$ condizione per la *GEL* "vincolata" implica $\mu_{q-c+h} \neq \sum p(x_i;\mu,\ldots,\mu_{q-c})x_i^{q-c+h}$ e la restrizione rende

il corrispondente moltiplicatore $\upsilon_{q-c+h} \neq 0$; si tratta di un test di generale applicabilità in quanto la struttura della D_{KL} prescinde, ai fini interpretativi, dall'ipotesi di indipendenza e somiglianza così come dall'essere il carattere discreto o continuo.

- ▷ In merito alla seconda, in considerazione dei lavori di Grendar e Judge, la scelta ovvia sarebbe $\gamma = -1$ identificante le verosimiglianze empiriche. Di esse in piccoli campioni non è assicurata la non negatività e Schennach (2007) ne discute alcuni problemi di robustezza (brevemente ricordati nel paragrafo 2.6). Nel paragrafo 3.2.1 si è discussa l'adozione di verosimiglianze di massima entropia ($\gamma = 0$) motivandone l'uso sulla base di diverse considerazioni:

  - ▷ maggiore robustezza in piccoli campioni e limitazione dell'impatto di elementi estranei al set informativo che il ricercatore ritiene idoneo ad orientare le stime,
  - ▷ possibilità di derivarne forma funzionale (e proprietà) per analogia all'approssimazione del punto di sella, utile a stimare la legge di funzioni aventi quale argomento somme di variabili casuali *i.i.d.*,
  - ▷ esigenza di modellare la verosimiglianza empirica secondo la famiglia esponenziale in modo da poterla interpretare come rappresentazione matematica della "vera" legge p e non della sua approssimazione,
  - ▷ coerenza con il teorema di Cifarelli e Regazzini (1982) secondo il quale, condizionatamente al limite di una statistica sufficiente a fini predittivi per successioni di variabili casuali scambiabili, le osservazioni sono indipendenti e somiglianti con legge di probabilità appartenente alla famiglia esponenziale; la procedura proposta riassume l'informazione campionaria nella posteriori sui momenti (ricavata tramite bootstrap di Rubin) e solo le relative realizzazioni sono utilizzate per modellare la verosimiglianza, circostanza che rende evidente il ruolo di esse quali statistiche sufficienti a fini predittivi.

- ▷ Per il terzo punto, nella sezione A.4 dell'appendice, sulla scorta dei risultati di White (1982) e Gasparini (1995), sono state dimostrate:

  - ▷ la normalità asintotica dei moltiplicatori, derivata direttamente dall'ampliamento esponenziale di una legge "iniziale"; il risultato è coerente con quanto accennato nel paragrafo 2.3 ed è ricavato per altra via rispetto al lavoro di Newey e Smith (2004),

- ▷ la natura relativa e l'ordine di grandezza dell'errore che si commette sostituendo ai momenti le loro stime, pari al prodotto tra la probabilità dedotta dall'ampliamento esponenziale, usando i momenti, e $O_p(N^{-1/2})$.

▷ In ordine al quarto punto, l'uso delle *GEL* in ottica bayesiana è stato discusso nel paragrafo A.3 dell'appendice e le proprietà di consistenza *MAP* sono state mostrate per i casi $\gamma \neq -1$ estendendo il risultato di Grendar e Judge; in particolare si è sottolineato che, imponendo opportune qualificazioni all'insieme delle distribuzioni a priori, continuano a valere le conclusioni in ordine alla concentrazione della posteriori attorno alla *L-proiezione* della legge p sulle osservazioni,

▷ La congiunta considerazione dei risultati illustrati nei paragrafi A.3 ed A.4 permette di rispondere all'ultimo quesito:

- ▷ le *GEL* sono funzioni dei momenti e dei moltiplicatori; questi ultimi rappresentano parametri di disturbo e alla luce del teorema di Sanov debbono essere eliminati tramite ottimizzazione; ciò significa selezionare tra le distribuzioni che soddisfano le condizioni sui momenti, quella che minimizza la discrepanza di parametro γ , in coerenza con quanto esposto nel paragrafo 3.4 a proposito del teorema di rappresentazione,
- ▷ le realizzazioni dalla posteriori sui momenti forniscono stime consistenti (Gasparini, 1995),
- ▷ se le *GEL* $\widetilde{p}$ calcolate facendo uso dei momenti della legge che ha generato i dati sono consistenti in *L-divergenza* (circostanza appurata nel paragrafo A.3) per mostrare che la proprietà vale per le *GEL* $\widehat{p}$ calcolate da stime dei momenti è possibile fare ricorso all'equivalenza asintotica tra le prime e le seconde. Nel caso $\gamma = 0$ l'equivalenza discende dal risultato sull'ordine di grandezza dell'errore di approssimazione, illustrato nel paragrafo A.4 dell'appendice,
- ▷ al variare delle realizzazioni della posteriori sui momenti, la minimizzazione delle discrepanze di parametro γ definisce un inviluppo di distribuzioni ottime; ciascuna di queste ultime per costruzione è *L-proiezione* della "vera" (ed ignota) legge di probabilità, circostanza che giustifica l'eliminazione del parametro di disturbo rappresentato dai momenti, mediante marginalizzazione; in altro modo, ricordando che ciascun elemento dell'inviluppo è, nei fatti, realizzazione di un bootstrap semiparametrico, la sintesi costituita dalla media delle realizzazioni appare naturale.

Se le distribuzioni predittive vengono utilizzate a scopo di ricampionamento, utile conseguenza è il superamento di un limite implicito nel bootstrap bayesiano: l'attribuzione di probabilità nulle alle intensità non rilevate. Tale risultato è ovvio dal momento che le nuove realizzazioni sono generate da una propria legge di probabilità. Si noti che il risultato viene conseguito in ambito pienamente oggettivo poiché l'unica informazione veicolata nelle stime è appresa dai dati. Una notevole opportunità è legata alla possibilità di estendere l'applicazione del ricampionamento a situazioni caratterizzate da qualche forma di dipendenza. Questo tema è stato discusso nel paragrafo 3.5 trattando il passaggio alla (meno restrittiva) condizione di scambiabilità parziale sotto il vincolo di ordinamento delle realizzazioni (della predittiva di ciascuna classe) secondo la sequenza con la quale - nel campione - vengono osservate le C classi di scambiabilità. Sotto un profilo operativo, una esplicita definizione delle classi può essere difficoltosa o risultare non indispensabile. Ad esempio, se ci si riferisce a dati ordinati in serie storica, stimando la distribuzione o densità congiunta del campione e collassando le classi in modo da farle coincidere con i ranghi delle osservazioni, si giunge ad una rappresentazione semplificata che - si è notato - fornisce l'interpretazione bayesiana del bootstrap di massima entropia proposto da Vinod (2004).
La procedura proposta in riferimento a dati dipendenti in serie storica e/o multivariati tiene conto dell'ipotesi di scambiabilità parziale usando le realizzazioni ordinate della predittiva per definire le distribuzioni empiriche inerenti le variabili casuali sezione; in questo modo:

- ▷ le serie simulate sono ottenute per accodamento delle realizzazioni ricavate dalle distribuzioni empiriche di ogni classe,
- ▷ si consegue l'obiettivo di rilassare il vincolo di identità nelle sequenze dei ranghi osservate e simulate, evitando l'esplicita definizione di intervalli di fungibilità all'atto dell'ordinamento delle realizzazioni dalla predittiva.
- ▷ i legami di rango all'interno delle variabili casuali sezione multivariate permettono di riprodurre le relazioni di dipendenza espresse da momenti misti di ordine moderatamente elevato, implicando una crescita soltanto lineare degli oneri di calcolo.

Nel capitolo 4 sono state proposte due applicazioni volte a verificare la bontà della procedura di ricampionamento bayesiano. È parso ragionevole ritenere che le valutazioni debbano riguardare la conservazione delle caratteristiche dei dati osservati ed il contributo inferenziale offerto.

Il primo esercizio, paragrafo 4.1, rivisita il lavoro di Johansen e Juselius (1990) sulla domanda di moneta in Finlandia. Il fatto che siano coinvolte quattro serie storiche trimestrali, in parte non stazionarie, costituisce un interessante banco di prova sia dal punto di vista del mantenimento delle dipendenze, sia dal lato inferenziale per misurare il supporto che i dati offrono al modello. Gli esiti rilevanti riguardano tre punti:

- ▷ la dimensione dello spazio di cointegrazione; nell'articolo, il test del massimo autovalore, calcolato in corrispondenza del campione osservato, porta alla conclusione (errata) che la dimensione r sia pari a due e gli Autori concludono (correttamente) per $r = 3$ sulla scorta di valutazioni riguardanti sia i coefficienti relativi ai vettori di cointegrazione sia i connessi significati economici; posto che nei dati sia implicita una certa dimensione r^*, quest'ultima non costituisce una restrizione significativa solo per la distribuzione della statistica che sottende $r = r^*$; coerentemente, il percentile simulato (al prefissato livello di significatività $\alpha = 0.05$) più prossimo a quello asintotico per la statistica del massimo autovalore induce a ritenere $r = 3$,

- ▷ la struttura dello spazio di cointegrazione; generato da due serie stazionarie e dalla combinazione lineare di due non stazionarie, essa viene vagliata per mezzo di un test delle ipotesi sui coefficienti di cointegrazione; condizionatamente ai dati, la distribuzione simulata approssima i percentili della asintotica e permette di giungere alle stesse conclusioni,

- ▷ le distribuzioni empiriche e gli intervalli di credibilità per i coefficienti; il metodo di Johansen non permette di ricavare tali misure di significatività e quelle ottenute per simulazione offrono un modo alternativo per discernere la struttura dello spazio di cointegrazione.

Il secondo esercizio, paragrafo 4.2, è orientato alla generazione di dati sintetici potenzialmente utili per le finalità di statistica ufficiale inerenti la tutela della riservatezza. La disponibilità di una metodologia di simulazione vincolata a condizioni sui momenti, idonea a garantire la conservazione dei nessi di dipendenza tra variabili, sembra costituire un accettabile punto di partenza per l'individuazione di strategie mirate al controllo del trade-off tra fedeltà nei risultati inferenziali e rischio di intrusione in informazioni non pubbliche. Per le simulazioni sono state utilizzate le informazioni campionarie contenute nell'Indagine della Banca d'Italia sui bilanci delle famiglie. Relative ai redditi del 2006, le distribuzioni sono molto sparse, con elevati indici di

asimmetria e curtosi. Queste peculiarità sono difficilmente riproducibili con modelli parametrici e, a tale proposito, è sembrato interessante verificare l'efficacia della procedura di ricampionamento bayesiano:

- ▷ le distribuzioni empiriche concernenti la matrice di correlazione ed il test di identità dei momenti misti mostrano il mantenimento delle relazioni di dipendenza tra variabili, con intervalli di credibilità più ampi per potenze elevate della variabile redditi da fabbricati, caratterizzata da poche osservazioni influenti nella coda destra,

- ▷ le distribuzioni simulate per indici di posizione, dimensione, dispersione e forma evidenziano la sostanziale fedeltà dei dati artificiali alle caratteristiche di quelli osservati. Alcuni disallineamenti concernenti gli indici di posizione, in special modo per i redditi da fabbricati, sono ascrivibili ad osservazioni isolate che orientando la forma della predittiva determinano una sovrastima della probabilità di osservare valori elevati.

Appendice A

A.1 Espansioni Asintotiche

In questa sezione, saranno illustrati alcuni risultati concernenti i parametri ρ e υ dell'ampliamento esponenziale, e la discrepanza di Cressie Read di parametro $\gamma = 0$, utili al fine di ricavare la distribuzione asintotica di quest'ultima. Assumendo che $j = 1, \ldots, q$ siano gli indici che identificano per ciascuna osservazione $i = 1, \cdots, N$ le componenti del vettore ψ_i, valgano le condizioni:

$$\begin{cases} E(\psi) = \mathbf{0} \\ \nu^t E(\psi\psi^t)\nu > 0 \quad \forall \nu \in \mathbb{R}^q \setminus \mathbf{0} \end{cases} \tag{A.1}$$

Riprendendo la notazione del capitolo 3, sia inoltre $\widetilde{p}_i = N^{-1} e^{\rho - 1 + \upsilon^t \psi_i}$. Senza perdita di generalità è utile assumere $\upsilon = \phi\,\upsilon_0$ con $\phi \equiv \|\upsilon\|$ e $\upsilon_0 = \upsilon / \|\upsilon\|$ in modo che $\|\upsilon_0\| = 1$. Un'espansione in serie di McLaurin arrestata al primo ordine, ponendo $\kappa \equiv \rho - 1 + \upsilon^t \psi_i$ permette di scrivere:

$$\frac{1}{N} e^{\rho - 1 + \upsilon^t \psi_i} = \frac{1}{N}\left[1 + (\rho - 1) + \phi\,\upsilon_0^t \psi_i + O(\kappa^2)\right]$$

Da $\sum \widetilde{p}_i \psi_i^t = 0$ segue

$$\sum \frac{1}{N}\left[1 + (\rho - 1) + \phi\,\upsilon_0^t \psi_i + O(\kappa^2)\right] \psi_i^t = 0$$

$$\frac{1}{N}\left[\rho \sum \psi_i^t + \phi\,\upsilon_0^t \sum \psi_i \psi_i^t\right] \approx 0$$

$$\phi\,\upsilon_0^t \approx -\rho \sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1}$$

Poiché $\sum \widetilde{p}_i = 1$, deve valere

$$1 \approx 1 + (\rho - 1) + \phi \upsilon_0^t \frac{1}{N} \sum \psi_i$$

Sostituendo l'espressione per $\phi \upsilon_0^t$ si ricava $1 \approx \rho \left(1 - N^{-1} \sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1} \sum \psi_i\right)$. La precedente relazione implica che $\rho = 1 + o_p(1)$. Ne deriva

$$\phi \upsilon_0^t \approx -\sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1} = -\frac{1}{N} \sum \psi_i^t \left(\frac{1}{N} \sum \psi_i \psi_i^t\right)^{-1}$$

Poiché $\frac{1}{N} \sum \psi_i^t \xrightarrow{p} E(\psi)$ e $\frac{1}{N} \sum \psi_i \psi_i^t \xrightarrow{p} E(\psi \psi^t)$, a causa della (A.1) si ottiene $\phi = o_p(1)$. L'espansione di $\sum \widetilde{p}_i$ arrestata al secondo ordine

$$\frac{1}{N} \sum \left[1 + (\rho - 1) + \upsilon^t \psi_i + \frac{1}{2}(\rho - 1)^2 + \frac{1}{2} \upsilon^t \psi_i \psi_i^t \upsilon + (\rho - 1) \upsilon^t \psi_i + O(|\kappa|^3)\right]$$

permette di esplicitare l'espressione di $\rho - 1$; indicando $A \equiv \sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1} \sum \psi_i$ e trascurando i termini di ordine $O_p(|\rho - 1|^2)$ si ricava

$$1 = 1 + (\rho - 1) - \frac{1}{N} A + \frac{1}{2N} A - (\rho - 1) \frac{1}{N} A$$

ed è immediato verificare che

$$(\rho - 1) = \frac{1}{2N} A \left[1 - \frac{1}{N} A\right]^{-1}$$

Ricordando che $\frac{1}{1+x} = 1 - x + x^2 - \cdots$, e omettendo i termini $o\left(N^{-1}\right)$ si ottiene

$$\rho - 1 \approx \frac{1}{2N} \sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1} \sum \psi_i = o_p(1)$$

La discrepanza di Cressie Read è proporzionale alla $D_{KL}(\widetilde{\mathbf{p}}, 1N^{-1})$ ossia

$$CR_0 = 2N \sum \widetilde{p}_i Ln(N \widetilde{p}_i) = 2 \sum e^{\rho - 1 + \upsilon^t \psi_i} \left(\rho - 1 + \upsilon^t \psi_i\right)$$

La relativa espansione si risolve in

$$2 \sum \left[1 + (\rho - 1) + \upsilon^t \psi_i + O(\kappa^2)\right] \left(\rho - 1 + \psi_i^t \upsilon\right)$$

trascurando i termini $o_p\left(N^{-1}\right)$, gli addendi rimanenti conducono alla quantità $A - 2A + 2A$ da cui

$$CR_0 \approx \sum \psi_i^t \left(\sum \psi_i \psi_i^t\right)^{-1} \sum \psi_i \tag{A.2}$$

A risultati simili si giunge considerando successivi ampliamenti esponenziali della legge di massima entropia, come accadrebbe includendo ulteriori condizioni sui momenti. Sotto le ipotesi

$$\begin{cases} E(\varphi) = \mathbf{0} \\ v^t E(\varphi\varphi^t) v > 0 \quad \forall v \in \mathbb{R}^q \setminus \mathbf{0} \end{cases} \tag{A.3}$$

sia

$$\widetilde{\widetilde{p}}_i = \widetilde{p}_i e^{\tau - 1 + \omega^t \varphi_i}$$

Pertanto,

$$D_{KL}(\widetilde{\widetilde{\mathbf{p}}}, \widetilde{\mathbf{p}}) = \sum \widetilde{p}_i e^{\tau - 1 + \omega^t \varphi_i} \left(\tau - 1 + \omega^t \varphi_i\right)$$

L'espansione di CR_0 permette di giungere alla seguente espressione:

$$CR_0 = 2N \sum \widetilde{p}_i \left[1 + (\tau - 1) + \omega^t \varphi_i + O(\kappa^2)\right] \left(\tau - 1 + \varphi_i^t \omega\right)$$

Dalle condizioni

$$\begin{cases} \sum \widetilde{\widetilde{p}}_i = 1 \\ \sum \widetilde{\widetilde{p}}_i \varphi_i^t = 0 \end{cases}$$

si evincono, con calcoli analoghi ai precedenti, le relazioni

$$\begin{cases} \omega^t \approx -\sum \widetilde{p}_i \varphi_i^t \left(\sum \widetilde{p}_i \varphi_i \varphi_i^t\right)^{-1} \\ \tau - 1 \approx \frac{1}{2} \sum \widetilde{p}_i \varphi_i^t \left(\sum \widetilde{p}_i \varphi_i \varphi_i^t\right)^{-1} \sum \widetilde{p}_i \varphi_i \end{cases}$$

Ponendo $\omega \equiv \varepsilon \omega_0$ con $\|\omega_0\| = 1$, per le (A.3), ε e $\tau - 1$ sono $o_p(1)$; indicando $A \equiv \sum \widetilde{p}_i \varphi_i^t \left(\sum \widetilde{p}_i \varphi_i \varphi_i^t\right)^{-1} \sum \widetilde{p}_i \varphi_i$ e trascurando le quantità $o_p(|\tau - 1|)$ nell'espansione per CR_0 si ottiene

$$CR_0 \approx 2N \left[\frac{1}{2} A - A + A\right] = N \sum \widetilde{p}_i \varphi_i^t \left(\sum \widetilde{p}_i \varphi_i \varphi_i^t\right)^{-1} \sum \widetilde{p}_i \varphi_i \tag{A.4}$$

Ricordando che $\widetilde{p}_i = \frac{1}{N}\left[1 + (\rho - 1) + (\upsilon^t \psi_i) + O(\kappa^2)\right]$, omettendo gli infinitesimi $(\rho - 1)$, $(\upsilon^t \psi_i)$ e $O(\kappa^2)$ la (A.4) si può approssimare come

$$CR_0 \approx \frac{1}{\sqrt{N}} \sum \varphi_i^t \left(\frac{1}{N} \sum \varphi_i \varphi_i^t\right)^{-1} \frac{1}{\sqrt{N}} \sum \varphi_i$$

Quest'ultima espressione consente di riconoscere immediatamente la distribuzione

asintotica di CR_0 sotto le ipotesi di indipendenza e somiglianza.

A.2 Distribuzione della Discrepanza

La (A.2) può essere significativamente espressa come

$$CR_0 \approx \frac{1}{\sqrt{N}} \sum \psi_i^t \left(\frac{1}{N} \sum \psi_i \psi_i^t \right)^{-1} \frac{1}{\sqrt{N}} \sum \psi_i$$

Poiché $N^{-1} \sum \psi_i \xrightarrow{p} E(\psi_i) = 0$, se le unità statistiche incluse nel campione osservato sono *i.i.d.*, per il *TCL* $N^{-1/2} \sum \psi_i^t \xrightarrow{d} \mathcal{N}(0, N^{-1} \sum \psi_i \psi_i^t)$ e $CR_0 \overset{a}{\sim} \chi_q^2$. Opportuni *TCL* permettono di rilassare le condizioni di indipendenza e somiglianza, una volta introdotte ulteriori qualificazioni in merito alla natura dei processi aleatori. Per semplificare i calcoli, in presenza di q condizioni di ortogonalità si può fare uso del Teorema di Cramér-Wold e mostrare che per qualunque vettore ς tale che $\|\varsigma\| = 1$, posto $(\frac{1}{N} \sum \psi_i \psi_i^t)^{-1} \equiv \Psi \Psi^t$, risulta verificata la relazione $\sum X_{i,N} \equiv N^{-1/2} \varsigma^t \Psi^t \sum \psi_i \xrightarrow{d} \mathcal{N}(0,1)$. Una possibile generalizzazione del *TCL* si ottiene, ad esempio, se $X_{i,N}$ è uniformemente L_2-limitato[1], ossia se $\sup_i \|X_{i,N}\|_2 < \infty$:

$$\sup_i \|X_{i,N}\|_2 < \infty \Rightarrow \frac{1}{N} \sum \psi_i^t \Psi \Psi^t \sum \psi_i \overset{a}{\sim} \chi_q^2$$

A questo proposito si può ricordare il seguente teorema (McLeish, 1974):

- ▷ sia $X_{i,N}$ una differenza di martingale [2] tale che
  - ▷ $\max_i \|X_{i,N}\|_2 < \infty$
  - ▷ $\max_i |X_{i,N}| \xrightarrow{p} 0$
  - ▷ $\sum X_{i,N}^2 \xrightarrow{p} 1$

[1] L'uniforme $Lp-$limitatezza di una collezione $\{Z_i\}$ implica

- ▷ l'uniforme $Lr-$limitatezza per $r < p$
- ▷ l'uniforme limitatezza in probabilità: $Z_i = O_p(1)$
- ▷ l'uniforme integrabililità: $\sup_i \left\{ \lim_{M \to \infty} E\left[|Z_i| \, 1_{(|Z_i| > M)} \right] \right\} = 0$

[2] $X_{i,N}$ è una differenza di martingale rispetto alla filtrazione $\mathscr{F}_{i,N}$ se $\forall i, N$

- ▷ $X_{i,N}$ è $\mathscr{F}_{i,N}$-misurabile
- ▷ $E|X_{i,N}| < \infty$
- ▷ $E(X_{i,N} | \mathscr{F}_{i-1,N}) = 0$

allora, $S \equiv \sum X_{i,N} \xrightarrow{d} \mathcal{N}(0,1)$

Poiché la limitatezza viene assunta come condizione (affatto restrittiva se si richiede l'esistenza di momenti finiti di ordine $k > 2$), rimangono da verificare le rimanenti; circa la prima,

$$N^{-1/2} \max_i \left| \varsigma^t \Psi^t \psi_i \right| = N^{-1/2} \varsigma^t \Psi^t O_p(1) \xrightarrow{p} 0$$

mentre per la seconda

$$\sum X_{i,N}^2 = N^{-1} \sum \left(\varsigma^t \Psi^t \psi_i \psi_i^t \Psi \varsigma \right)$$

$$\sum X_{i,N}^2 = \varsigma^t N^{-1} \sum \left(\Psi^t \psi_i \psi_i^t \Psi \right) \varsigma = \varsigma^t I_q \varsigma = 1$$

Rimane così dimostrato che $N^{-1/2} \varsigma^t \Psi^t \sum \psi_i \xrightarrow{d} \mathcal{N}(0,1)$ e dunque

$$N^{-1/2} \Psi^t \sum \psi_i \xrightarrow{d} \mathcal{N}(0, I_q)$$

Segue immediatamente l'enunciato: $N^{-1} \sum \psi_i^t \Psi \Psi^t \sum \psi_i \xrightarrow{d} \chi_q^2$.

A.3 Estensione della Consistenza in L-Divergenza

È possibile mostrare che la minimizzazione delle discrepanze empiriche fornisce posteriori consistenti in *L-divergenza* (ossia tali da concentrarsi - al divergere di N - in un intorno arbitrariamente piccolo della distribuzione approssimante $\mathbf{w} \in W$ che minimizza la *L-divergenza* tra W e la "vera" distribuzione $\mathbf{p}$ generatrice dei dati), con argomenti simili a quelli ricordati nel capitolo 2, a condizione che in aggiunta alle condizioni ivi specificate venga opportunamente definito il supporto costituito dalle leggi di probabilità a priori. Per comprenderne il motivo, si osservi che il teorema può essere espresso in una forma alternativa che ne lascia immutate le conclusioni. Infatti, posto che w sia tale da minimizzare la *L-divergenza* $-\sum p_i ln(w_i)$ rispetto a $\mathbf{p}$, il soddisfacimento delle condizioni di ortogonalità conduce a sistemi di pesi "ottimi" $\mathbf{w}^*$ individuati dalle verosimiglianze empiriche:

$$\mathbf{w}^* = \underset{\mathbf{w}}{argmin} \left(\sum \tfrac{1}{N} ln \tfrac{1/N}{w_i} \right) \quad sub \begin{cases} \sum w_i = 1 \\ \sum w_i \psi_i = 0 \end{cases}$$

La soluzione $w_i^* = N^{-1} \left(-\rho - \upsilon^t \psi_i \right)^{-1}$ garantisce la minimizzazione della *L-divergenza*. Ponendo $\xi \equiv (\rho, \upsilon)$, se la distribuzione uniforme $1N^{-1}$ fosse ponderata con una diver-

sa funzione $h(\psi,\xi)$ asintoticamente equivalente, le conclusioni di consistenza MAP non muterebbero. Dunque sarebbe possibile riformulare il teorema; limitando per semplicità l'attenzione al caso di variabili casuali discrete, siano:

- ▷ X, una v.c. con legge discreta $\mathbf{p}$,
- ▷ $W = \left\{ \frac{1}{N} h(\psi,\xi) = \underset{\mathbf{w}}{argmin}\left(I_\gamma\right), \sum w_i \psi_i = 0, \sum w_i = 1, \pi(\mathbf{w}) > 0 \right\}$, il supporto della priori sulle distribuzioni,
- ▷ $\ell(\mathbf{w}) \equiv e^{\sum ln(w_i)}$, il valore della logverosimiglianza,
- ▷ $\pi(\mathbf{w}|\mathbf{x}) \propto \pi(\mathbf{w}) e^{\sum ln(w_i)}$, il valore della posteriori,
- ▷ $\pi(V \subset W|\mathbf{x}) = \sum_{\mathbf{w}\in V} \pi(\mathbf{w}|\mathbf{x}) / \sum_{\mathbf{w}\in W} \pi(\mathbf{w}|\mathbf{x})$,
- ▷ $\hat{\ell}(A) \equiv sup_{\mathbf{w}\in A}\, \ell(\mathbf{w}) \equiv e^{\sum ln\left(\widehat{w}_{i,A}\right)}$
- ▷ $\hat{\pi}(A|\mathbf{x}) \equiv sup_{\mathbf{w}\in A}\, \pi(\mathbf{w}|\mathbf{x}) \propto \pi(\widehat{\mathbf{w}}_A) e^{\sum ln\left(\widehat{w}_{i,A}\right)}$
- ▷ $L(A\|\mathbf{p}) = L(\mathbf{w}_A\|\mathbf{p}) \equiv -\sum p_i ln(w_i)$ con $\mathbf{w} \equiv argmin_{\mathbf{w}\in A} L(A\|\mathbf{p})$

Al divergere di N,

$$P\left\{ \frac{1}{N} ln\left[\pi(\mathbf{w}\in V|\mathbf{x})\right] \to -\left[L(V\|\mathbf{p}) - L(W\|\mathbf{p})\right] \right\} = 1$$

e definendo

$$V^C \equiv \{\mathbf{w} : [L(\mathbf{w}\|\mathbf{p}) - L(W\|\mathbf{p})] > \varepsilon, \varepsilon > 0, \mathbf{w}\in W\}$$

segue

$$P\left\{ \pi\left(\mathbf{w}\in V^C|\mathbf{x}\right) \to 0 \right\} = 1$$

In altri termini per $N \to \infty$, la $\pi(\mathbf{w}|\mathbf{x})$ si concentra sulla distribuzione $\widehat{w}(\mathbf{x})$ che minimizza la *L-divergenza* $L(\mathbf{w}\|\mathbf{p}) = -\sum p_i \ln(w_i)$. Nel caso in esame, le $\widehat{\mathbf{w}}$ minimizzano la discrepanza di Cressie Read di parametro γ rispetto alla distribuzione uniforme e, nella classe delle funzioni $\mathbf{w} = \frac{1}{N} h(\psi,\xi)$ alle quali il supporto è stato ristretto, minimizzano asintoticamente la *L-divergenza* rispetto ad $\mathbf{p}$. Per convincersi della circostanza, occorre mostrare che tutte le *GEL* convergono ad un comune valore. Espandendo attorno a 1 i termini che "ponderano" $\mathbf{1}N^{-1}$:

- ▷ caso $\gamma = -1$:

$$\frac{1}{N}\frac{1}{-\rho - \upsilon^t \psi_i} = \frac{1}{N}\left[1 + \left(\rho + \upsilon^t \psi_i - 1\right) + O\left(\kappa^2\right)\right]$$

▷ caso $\gamma = 0$:

$$\frac{1}{N} e^{\upsilon^t \psi_i + \rho - 1} = \frac{1}{N} \left[1 + \left(\rho + \upsilon^t \psi_i - 1 \right) + O\left(\kappa^2 \right) \right]$$

▷ caso $-1 < \gamma < 0$[3]:

$$\frac{1}{N} \left[(-\gamma - 1) \left(\rho + \upsilon^t \psi_i \right) \right]^{-1/(\gamma+1)} \propto \frac{1}{N} \left[1 - \frac{1}{\gamma + 1} \left(\rho + \upsilon^t \psi_i - 1 \right) + O\left(\kappa^2 \right) \right]$$

Ponendo $\upsilon = \phi \upsilon_0$, con $\|\upsilon_0\| = 1$ si è notato nel primo paragrafo, relativamente a $\gamma = 0$, che:

$$\begin{cases} \rho = 1 + o_p(1) \\ \phi = o_p(1) \end{cases}$$

In modo analogo (i calcoli si basano sulle medesime espansioni asintotiche e per brevità vengono omessi) tali relazioni valgono per $\gamma \neq 0$ e resta dimostrato che al divergere della numerosità campionaria i diversi sistemi di pesi convergono.

Incidentalmente, si può notare che scegliendo $h(\psi_i, \xi) = e^{\rho - 1 + \psi_i^t \upsilon}$, in ambito classico si otterrebbe di fatto la verosimiglianza empirica ampliata esponenzialmente (Schennach, 2007) e le stime $\widehat{\theta}$ necessarie a determinare il valore assunto dalle condizioni di ortogonalità verrebbero ottenute mediante ottimizzazioni alternate:

$$\begin{cases} \widehat{\xi} = \underset{\xi}{argmax} \left\{ \sum \frac{1}{N} h \left[\psi \left(x_i, \widehat{\theta} \right), \xi \right] \right\} \\ \widehat{\theta} = \underset{\theta}{argmin} \left\{ \sum \frac{1}{N} ln \frac{1/N}{\frac{1}{N} h \left[\psi(x_i, \theta), \widehat{\xi} \right]} \right\} \end{cases}$$

Dalle precedenti considerazioni, emerge che il tratto qualificante circa l'impiego delle *GEL* in ottica bayesiana risiede nella necessità di eliminare il parametro di di-

[3] I calcoli sono immediati e, per comodità, vengono di seguito ricordati:

$$\mathscr{L} = I_\gamma(w_i, 1/N) - \rho \left(\sum w_i - 1 \right) - \sum w_i \left[\upsilon^t \psi_i \right]$$

$$\partial \mathscr{L} / \partial w_i = I'_\gamma - \rho - \upsilon^t \psi_i = 0 \Rightarrow w_i = (I'_\gamma)^{-1} \left(\rho + \upsilon^t \psi_i \right)$$
$$w_i = (I'_\gamma)^{-1} I'_\gamma(w_i) \Rightarrow I'_\gamma(w_i) = \rho + \upsilon^t \psi_i$$

Poiché $I_\gamma(w_i, 1/N) = \sum \left[(N w_i)^{-\gamma} - 1 \right] / [\gamma(\gamma+1)]$, segue: $I'_\gamma = \left[-\gamma (N w_i)^{-\gamma - 1} \right] / [\gamma(\gamma+1)]$, da cui

$$w_i = \frac{1}{N} \left[(-\gamma - 1) \left(\rho + \upsilon^t \psi_i \right) \right]^{-1/(\gamma+1)} \propto \frac{1}{N} \left(\rho + \upsilon^t \psi_i \right)^{-1/(\gamma+1)}$$

sturbo rappresentato dai moltiplicatori ξ mediante ottimizzazione; le condizioni di ortogonalità $\psi(x_i,\theta)$ vengono "alimentate" dalla legge a priori sui parametri θ e stime consistenti per θ sono ricavate dalla moda della posteriori. Per ragionare della distribuzione predittiva finale occorre mostrare che le *GEL* ottenute da stime consistenti dei parametri continuano ad essere consistenti in *L-divergenza*; allo scopo è utile far riferimento al teorema di equivalenza asintotica (Cappuccio e Orsi, 2005): due successioni sono asintoticamente equivalenti se la loro differenza converge in probabilità a zero; nel caso in esame interessa che la successione di funzioni aleatorie $\frac{1}{N}h_N[\psi(x_i,\theta),\xi(\theta)]$ converga uniformememte in probabilità alla funzione limite $\frac{1}{N}h[\psi(x_i,\theta),\xi(\theta)]$, ossia

$$\sup_{\theta}\left|\frac{1}{N}h_N\left[\psi(x_i,\theta),\xi(\theta)\right]-\frac{1}{N}h\left[\psi(x_i,\theta),\xi(\theta)\right]\right|\xrightarrow{p}0 \tag{A.5}$$

Alla dimostrazione di quest'ultima relazione nel caso di $\gamma=0$, è dedicato il paragrafo successivo; poiché ogni realizzazione della posteriori per θ, modellata come processo di Dirichlet, ne fornisce stime consistenti, nel calcolo della predittiva finale l'eliminazione del parametro di disturbo θ può avvenire, secondo consuetudine, mediante marginalizzazione.

A.4 MM e Approssimazione di Sella

In questa sezione, dopo aver usato i nessi formali tra distribuzioni di massima entropia ed approssimazione del punto di sella per verificare la normalità asintotica dei moltiplicatori, viene mostrata la validità della (A.5) impiegando il risultato per ricavare la distribuzione in grandi campioni della CR_0.

L'uso del metodo dei momenti nella stima semiparametrica di una legge di probabilità quando le condizioni vengano veicolate mediante ampliamento esponenziale, minimizzando la divergenza di Kullback-Leibler rispetto ad una distribuzione di riferimento *w,* produce risultati che per costruzione hanno caratteristiche comuni all'approssimazione del punto di sella per una verosimiglianza. Definendo $K(\upsilon)=ln\left[E_w\left(e^{\upsilon^t\psi(x_i)}\right)\right]$, si consideri l'ampliamento esponenziale di $w(x)$

$$L(\upsilon)=e^{\upsilon^t\sum\psi(x_i)-NK(\upsilon)}w(x)$$

Derivandone il logaritmo naturale $\ell(\upsilon)=[\upsilon^t\sum\psi(x_i)-NK(\upsilon)]+ln[w(\mathbf{x})]$, risulta:

$$\ell'(\upsilon)=\sum\psi(x_i)-NK'(\upsilon)\Rightarrow\hat{\upsilon}=(K')^{-1}(\tfrac{1}{N}\sum\psi(x_i))$$
$$\ell''(\upsilon)=-NK''(\upsilon)\Rightarrow\mathbf{I}(\upsilon)=NK''(\upsilon)$$

A causa delle proprietà degli stimatori di massima verosimiglianza

▷ $E\left[\ell'(\upsilon)\right] = 0 \Rightarrow E\left\{\left[\ell'(\upsilon)\right]\left[\ell'(\upsilon)\right]^t\right\} = N\,cov\left[\psi(x)\right]$,

▷ con $\upsilon \neq \mathbf{0}$, $K''(\upsilon) \neq cov\left[\psi(x)\right]$ e $\hat{\upsilon}$ ha dispersione[4]

$$\frac{1}{N}\left[C(\upsilon)\right]^{-1} = \left[K''(\upsilon)\right]^{-1} cov\left[\psi(x)\right]\left[K''(\upsilon)\right]^{-1} \quad \text{(White, 1982)},$$

▷ la velocità di convergenza di $\hat{\upsilon}$ è $N^{-1/2}$: poiché $\frac{1}{N}\sum\psi(x_i) \overset{q.c.}{\to} E\left[\psi(x)\right]$,

$$\sqrt{N}\hat{\upsilon} \overset{A}{\sim} \mathscr{N}\left(\sqrt{N}(K')^{-1}(E\left[\psi(x)\right]), \left[C(\upsilon)\right]^{-1}\right) \tag{A.6}$$

Il calcolo dei valori assunti dalle condizioni di ortogonalità, qualora si ignorino i momenti caratterizzanti la legge di probabilità sulle osservazioni, necessita del ricorso a quantità stimate. In ambito classico, facendo uso dei momenti campionari, il *TCL* per osservazioni *i.i.d.* consente di affermare che l'errore che si commette è $O_p\left(N^{-1/2}\right)$; se le stime dei momenti provengono dalla posteriori ottenuta tramite bootstrap bayesiano, sotto l'ipotesi di scambiabilità vale il Teorema 5[5] in Gasparini (1995) e anche in questo caso l'ordine di grandezza dell'errore è $O_p\left(N^{-1/2}\right)$.
Tenendo conto della (A.6), siano:

$$\begin{cases} \forall j\ \psi_{ij} \equiv \left(x_i^j - \mu_j\right) \\ \mathbf{g}_i = \left[\psi_{i1} + O_p\left(N^{-1/2}\right), \psi_{i2} + O_p\left(N^{-1/2}\right), \cdots\right]^t \\ \lambda^t = \left[\upsilon_1 + O_p\left(N^{-1/2}\right), \upsilon_2 + O_p\left(N^{-1/2}\right), \cdots\right] \end{cases}$$

Data la distribuzione iniziale $\mathbf{w} = \mathbf{1}N^{-1}$, minimizzando la $D_{KL}\left(\mathbf{p}, \mathbf{w}\right)$ sub

[4]Il risultato può essere facilmente derivato sotto le usuali ipotesi: indicando con υ^* lo pseudo valore di υ, per il teorema del valor medio esiste un $\bar{\upsilon}$ compreso tra $\widehat{\upsilon}$ e υ^* tale che:

$$\ell'(\upsilon^*) = \ell'(\widehat{\upsilon}) + \ell''(\bar{\upsilon})(\widehat{\upsilon} - \upsilon^*)$$

Poiché $\ell'(\widehat{\upsilon}) = 0$ e $\widehat{\upsilon} \to \upsilon^*$, $\bar{\upsilon} \to \upsilon^*$, $\ell''(\bar{\upsilon}) \to -\mathbf{I}(\upsilon^*)$ e si ricava

$$\sqrt{N}(\widehat{\upsilon} - \upsilon^*) = \sqrt{N}\left[\mathbf{I}(\upsilon^*)\right]^{-1}\ell'(\upsilon^*) + o_p(1)$$

Ricordando che υ^* minimizza la *L-divergenza* $-E\,\ell(\upsilon)$ deve anche essere

$$\sqrt{N}\ell'(\upsilon^*) \to \mathscr{N}(\mathbf{0}, E\left[\ell'(\upsilon^*)\right]\left[\ell'(\upsilon^*)\right]^t)$$

Segue:

$$\sqrt{N}(\widehat{\upsilon} - \upsilon^*) \to \mathscr{N}(\mathbf{0}, \left[\mathbf{I}(\upsilon^*)\right]^{-1} E\left[\ell'(\upsilon^*)\right]\left[\ell'(\upsilon^*)\right]^t\left[\mathbf{I}(\upsilon^*)\right]^{-1})$$

[5]Se p ammette momenti finiti $\mu_1, \mu_2, \ldots, \mu_{2s}$, indicando con m il valore atteso della *posteriori*, $\sqrt{n}(\mu - m)$ converge debolmente ad una Normale s^{pla}di media nulla:

$$\sqrt{n}(\mu - m) \to \mathscr{N}(\mathbf{0}, \mathbf{C}),\ \forall i, j \leq s:\ c_{ij} = \mu_{i+j} - \mu_i\mu_j$$

$$\begin{cases} \sum_i p_i x_i^j = m_j \ \ (j = 1, 2, \cdots) \\ \sum_i p_i = 1 \end{cases}$$

si ricava

$$\widehat{p}_i = \frac{1}{N} e^{\lambda^t \mathbf{g}_i + \eta - 1}$$

Si noti che $e^{-\eta+1}$ è l'omologo di $e^{K(\upsilon)} = E_w\left(e^{\upsilon^t \psi(x_i)}\right)$, ossia $e^{-\eta+1} = N^{-1}\sum e^{\lambda^t \mathbf{g}_i}$. Con $\upsilon_j = O_p(1)$ e $\psi_{ij} = O_p(1)\ \forall j$ si ha

$$e^{1-\eta} = \sum \frac{1}{N} e^{\left[\upsilon + O_p\left(N^{-1/2}\right)\right]^t \left[\psi_i + O_p\left(N^{-1/2}\right)\right]} = \sum \frac{1}{N} e^{\upsilon^t \psi_i + O_p\left(N^{-1/2}\right)}$$

L'ultimo membro può essero espresso come

$$\sum \frac{1}{N} e^{\upsilon^t \psi_i + O_p\left(N^{-1/2}\right)} \equiv e^{1-\rho+O_p\left(N^{-1/2}\right)}$$

e tenendo conto di quest'ultimo termine è immediato verificare

$$\widehat{p}_i = \frac{1}{N} e^{\left[\upsilon + O_p\left(N^{-1/2}\right)\right]^t \left[\psi_i + O_p\left(N^{-1/2}\right)\right] + \left[(\rho-1) + O_p\left(N^{-1/2}\right)\right]}$$

Ponendo $\widetilde{p}_i \equiv \frac{1}{N} e^{\upsilon^t \psi_i + \rho - 1}$, si può scrivere in modo compatto

$$\widehat{p}_i = \widetilde{p}_i e^{O_p\left(N^{-1/2}\right)}$$

Quest'ultima relazione è suscettibile di una rappresentazione maggiormente significativa; ricordando lo sviluppo in serie di McLaurin di una funzione esponenziale: $e^y = 1 + y + \frac{1}{2}y^2 + \frac{1}{6}y^3 + \cdots$, si ottiene la relazione:

$$e^{O_p\left(N^{-1/2}\right)} = \left[1 + O_p\left(N^{-1/2}\right) + \frac{1}{2} O_p\left(N^{-1}\right) + \cdots\right]$$

che sostituita nella precedente consente di scrivere

$$\widehat{p}_i = \widetilde{p}_i \left[1 + O_p\left(N^{-1/2}\right)\right] \tag{A.7}$$

In parole, l'errore che si commette approssimando $\widetilde{p}_i$ con $\widehat{p}_i$ è $\widetilde{p}_i O_p\left(N^{-1/2}\right)$, ossia per ogni numerosità campionaria è tanto più piccolo quanto più ci si avvicina alle code della distribuzione.

La (A.5), ricordando che $\widehat{p}_i \equiv N^{-1} h[\psi(x_i, m), \xi(m)]$ e $\widetilde{p}_i \equiv N^{-1} h[\psi(x_i, \mu), \xi(\mu)]$, è assicurata dalla (A.7).

Circa la $CR_0(\widehat{\mathbf{p}}, \mathbf{1}N^{-1})$, ripetendo quanto illustrato a proposito della $CR_0(\widetilde{\mathbf{p}}, \mathbf{1}N^{-1})$,

$$CR_0 = 2\sum e^{\rho-1+\upsilon^t \psi_i} \left[1 + O_p(N^{-1/2})\right] \left[\rho - 1 + \upsilon^t \psi_i + O_p(N^{-1/2})\right]$$

$$CR_0 = 2\sum e^{\rho-1+\upsilon^t\psi_i}\left(\rho-1+\upsilon^t\psi_i\right)\left[1+O_p(N^{-1/2})\right]$$

espandendo $\widetilde{p}_i$ si ottiene

$$CR_0 = 2\sum\left[1+(\rho-1)+\upsilon^t\psi_i+O(\kappa^2)\right]\left(\rho-1+\psi_i^t\upsilon\right)\left[1+O_p(N^{-1/2})\right]$$

In analogia a quanto illustrato nel primo paragrafo, sostituendo le espressioni asintotiche di υ e ρ, posto $A \equiv \sum\psi_i^t\left(\sum\psi_i\psi_i^t\right)^{-1}\sum\psi_i$ e trascurando gli elementi $o\left(N^{-1}\right)$, i termini rilevanti che residuano portano a:

$$CR_0 = [A-2A+2A]\left[1+O_p(N^{-1/2})\right]$$

Se ne ricava, come era lecito attendersi,

$$CR_0 \approx \sum\psi_i^t\left(\sum\psi_i\psi_i^t\right)^{-1}\sum\psi_i$$

Restano ovviamente valide le conclusioni in ordine alla distribuzione asintotica.

A.5 Esempio

La simulazione di un caso molto semplice può essere utile a fissare le idee. Sia W un insieme di distribuzioni definite sul supporto $\{1, 2, 3, 4\}$ e tali da massimizzare l'entropia sotto il vincolo delle condizioni sui momenti che, in ipotesi, rappresentano le conoscenze a priori sul fenomeno indagato. La legge **p** generatrice dei dati (ignota al ricercatore) è descritta nella tabella (A.1) .

$p(X=1)$	$p(X=2)$	$p(X=3)$	$p(X=4)$
0.20	0.45	0.35	0.00

Tabella A.1: Distribuzione generatrice dei dati

I primi quattro momenti che la caratterizzano sono esposti nella tabella (A.2).

μ_1	μ_2	μ_3	μ_4
2.15	5.15	13.25	35.75

Tabella A.2: Momenti della distribuzione generatrice dei dati

Da **p** si possono simulare campioni di ampiezza crescente, pari rispettivamente a N=25, 50, 100 in modo da poter calcolare, per ciascuna distribuzione in W, le probabilità a posteriori e le *L-divergenze*. Le quindici condizioni utili a definire le leggi incluse in W sono mostrate nella tabella (A.3) :

i	μ_1	μ_2	μ_3	μ_4
1	1.5	3.3	9.9	34.5
2	1.5	3.1	8.7	29.5
3	1.5	2.9	6.9	18.5
4	1.5	2.7	5.7	13.5
5	1.5	2.5	4.5	8.5
6	3.2	11.8	45.8	181.0
7	3.2	11.6	44.0	170.0
8	3.2	11.4	42.8	165.0
9	3.2	11.2	41.6	160.0
10	3.2	11.2	41.0	154.0
11	3.2	11.0	39.8	149.0
12	3.2	11.0	39.2	143.0
13	3.2	10.8	38.0	138.0
14	3.2	10.6	36.2	127.0
15	3.2	10.4	34.4	116.0

Tabella A.3: Momenti delle distribuzioni in W

Ricavate le distribuzioni, per ciascuna si assuma una probabilità a priori pari a $\pi(\mathbf{w}_i) = 1/15$. Le masse di probabilità che le identificano coincidono con quelle esposte nella tabella 1 in Grendar e Judge (2007); si noti che i momenti di **p**, tabella (A.2), non coincidono con le alternative di cui il ricercatore ha tenuto conto nella costruzione di W; ciò significa che la moda della posteriori individua, nella fattispecie, la legge in W che approssima **p** nel miglior modo possibile (ossia minimizzando la *L-divergenza* tra **p** e W). Sostanzialmente già per N=25 la posteriori si concentra sulla distribuzione alla quarta riga della tabella (A.4) che assicura anche la minimizzazione della *L-divergenza*. Che ciò debba avvenire è intuitivo: indicando con **q** una qualsiasi distribuzione, il ridursi della *L-divergenza* quando **q** si approssima a **p** è banale conseguenza della non negatività della $D_{KL}(\mathbf{p}, \mathbf{q})$:

$$\sum p_i ln p_i - \sum p_i ln q_i \geq 0 \Leftrightarrow -\sum p_i ln q_i \geq -\sum p_i ln p_i$$

Fin qui, l'esemplificazione della consistenza in *L-divergenza*. Ai fini della procedura di ricampionamento proposta, non è necessario specificare le singole leggi costituenti W ma è sufficiente calcolare la verosimiglianza empirica che minimizza la

$p(X=1)$	$p(X=2)$	$p(X=3)$	$p(X=4)$	N=25	N=50	N=100	L-div
0.80	0.00	0.10	0.10	0.00	0.00	0.00	8.83
0.70	0.20	0.00	0.10	0.00	0.00	0.00	7.23
0.70	0.10	0.20	0.00	0.00	0.00	0.00	1.67
0.60	0.30	0.10	0.00	1.00	1.00	1.00	1.45
0.50	0.50	0.00	0.00	0.00	0.00	0.00	7.59
0.20	0.10	0.00	0.70	0.00	0.00	0.00	8.16
0.20	0.00	0.20	0.60	0.00	0.00	0.00	7.50
0.10	0.20	0.10	0.60	0.00	0.00	0.00	1.99
0.00	0.40	0.00	0.60	0.00	0.00	0.00	24.08
0.10	0.10	0.30	0.50	0.00	0.00	0.00	1.92
0.00	0.30	0.20	0.50	0.00	0.00	0.00	4.34
0.10	0.00	0.50	0.40	0.00	0.00	0.00	9.24
0.00	0.20	0.40	0.40	0.00	0.00	0.00	4.82
0.00	0.10	0.60	0.30	0.00	0.00	0.00	4.97
0.00	0.00	0.80	0.20	0.00	0.00	0.00	17.59

Tabella A.4: Distribuzione a posteriori e L-divergenze

discrepanza di Cressie-Read del prescelto parametro γ rispetto alla legge uniforme: al divergere della numerosità campionaria la posteriori sui momenti veicola informazioni sempre più accurate per la stima della predittiva. In altri termini, alla priori sulle leggi viene sostituita la posteriori sui momenti e per costruzione è garantita la coerenza rispetto ai dati. Nel caso in discorso, generati campioni di ampiezza crescente, la tabella (A.5) mostra le stime ottenute per la distribuzione generatrice dei dati.

	$p(X=1)$	$p(X=2)$	$p(X=3)$	$p(X=4)$
N=25	0.362	0.239	0.399	0.000
N=50	0.261	0.380	0.359	0.000
N=100	0.221	0.420	0.359	0.000

Tabella A.5: Distribuzione di massima entropia

I corrispondenti valori della *L-divergenza* sono esposti nella tabella (A.6).

N=25	N=50	N=100
1.168	1.063	1.051

Tabella A.6: L-divergenza rispetto alla distribuzione di massima entropia

Appendice B

Sono di seguito esposti alcuni dei codici R (R Development Core Team, 2008) usati nelle elaborazioni. Per brevità, insieme alle principali funzioni si include un programma utile a richiamarle, finalizzato alla stima della predittiva e, nel caso di processi aleatori, delle leggi concernenti le classi di scambiabilità. Tutti i codici che non hanno trovato spazio in questa appendice sono disponibili dietro richiesta.

B.1 Definizione di Alcune Funzioni

```
#inizio MatDen#######################################################################
#la funzione definisce la distribuzione iniziale per il calcolo delle GEL

#argomenti:
#x:        il vettore dei dati campionari
#tipo:     0=variabile continua, 1=variabile discreta, 2=supporto misto
#segno:    1=supporto non negativo, -1=supporto non positivo, qualsiasi altro
#          valore=supporto non limitato ad un semiasse
#alfa:     costante che espande (se > 1) o contrae (se < 1) il supporto
#          (se tipo=0) o i sottoinsiemi di supporto (se tipo=2)
#steps:    numero di punti nei quali il supporto continuo viene discretizzato
#inf:      estremo inferiore del supporto, se noto
#sup:      estremo superiore del supporto, se noto

#output:
#S:        matrice di 2 colonne per punti del supporto e distribuzione uniforme
#######################################################################################
MatDen=function(x,tipo,segno,alfa,steps,inf,sup) {

if (tipo==0) {

mn=min(x)-alfa*mad(x)
mx=max(x)+alfa*mad(x)

if (segno==1)  {
mn=max(mn,0)
}
```

```
if (segno==-1) {
mx=min(mx,0)
}

if (is.null(inf)==FALSE) {mn=inf}
if (is.null(sup)==FALSE) {mx=sup}

sup=seq(mn,mx,(mx-mn)/(steps-1))
}

if (tipo==1) {
o=(as.numeric(levels(factor(x))))
d.o=diff(o)
sup=seq(min(x),max(x),min(d.o))
}

if (tipo==2) {
basi=sort(x)
sup=sort(rnorm(steps,mean=basi,sd=abs(basi)*alfa))
}

psup=rep(1/length(sup),length(sup))
S=cbind(sup,psup)

S
}
#fine Matden#########################################################################


#inizio bb#############################################################################
#la funzione produce realizzazioni della posteriori sui momenti tramite
#bootstrap di Rubin(1981)

#argomenti:
#y:        il vettore dei dati campionari
#nrep:     numero di realizzazioni da generare
#maxp:     ordine massimo dei momenti da calcolare
#trim:     numero di unità da sopprimere nelle due code
#wins:     1: winsorizzazione, 0:nessuna sostituzione delle unità soppresse

#output sotto forma di lista:
#MM:       elemento (nrep,maxp) della lista, contenente in ogni riga una
#          realizzazione della posteriori sui momenti
#######################################################################################
bb=function(y,nrep,maxp,trim,wins) {
if(trim==0) {nt=0;x=y}
if(trim>0) {
nt=max(1,round(length(y)*trim))
x=(sort(y))[(1+nt):(length(y)-nt)]
if(wins==1) {
cs=rep((sort(y))[1+nt],nt)
cd=rep((sort(y))[(1+nt):(length(y)-nt)],nt)
x=c(cs,x,cd)
}
}
MM=matrix(0,nrep,maxp) ;
fu1=function(q) {rgamma(1,q)}    ;
```

```
for (rep in 1:nrep)                 {
Dirichlet=apply(matrix(1,length(x),1),1,fu1)
Dirichlet=Dirichlet/sum(Dirichlet)
Mb=outer(x,seq(1,maxp),"^")
fu2=function(q) {sum(Dirichlet*q)}
MM[rep,]=t(apply(Mb,2,fu2))
}
list(MM=MM)
}
#fine bb##########################################################################


#inizio funzione g.mom#############################################################
#la funzione genera le matrici esprimenti le condizioni di ortogonalità per
#ciascun record

#argomenti:
#x:       il vettore dei dati campionari
#maxp:    ordine massimo delle potenze considerate
#par:     momenti di ordine 1:maxp

#output:
#g:       matrice (n x maxp) contenente in ogni riga il vettore delle
#         condizioni sui momenti
##################################################################################
g.mom=function(x,maxp,par) {
X=outer(x,seq(1,maxp),"^")
fun=function(y) {t(y)-par} #la recycling agisce lungo le colonne
g=t(apply(X,1,fun))
g
}
#fine g.mom#######################################################################


#inizio funzione Inner#############################################################
#la funzione stima le verosimiglianze empiriche (EL) e le verosimiglianze
#di massima entropia (ET) quando sono presenti almeno due condizioni sui momenti

#argomenti:
#hlim:     numero massimo di iterazioni
#criterio: ordine di approssimazione al di sotto del quale si intende raggiunta
#          la convergenza
#ww:       vettore delle probabilità iniziali
#gm:       matrice delle condizioni di ortogonalità
#GEL:      "EL"=empirical likelihood, "ET"=exponential tilting
#ini:      valori iniziali per i moltiplicatori; 1=valore atteso dei vincoli,
#          2=valore atteso dei vincoli riscalato, qualsiasi altro valore=
#          vettore nullo
#mom:      matrice dei momenti che debbono essere ottenuti
#lim:      numero di condizioni di ortogonalità da considerare (minimo=2)

#output sotto forma di lista:
#hh:       numero di iterazioni effettuate
#crit:     massimo errore percentuale in valore assoluto nella ricostruzione
#          dei momenti
#PG:       errori percentuali sui momenti
#lambda:   vettore dei moltiplicatori all'ultima iterazione
```

```
#mlambda:  matrice dei moltiplicatori calcolati per ogni iterazione e via via
#          che vengono incluse le condizioni di ortogonalità
#ww:       vettore delle "probabilità" associate ad ogni osservazione
##############################################################################
Inner=function(hlim,criterio,ww,gm,GEL,ini,mom,lim) {

g=gm[,1:lim]
mlambda=matrix(0,ncol(gm),hlim)
if (length(ww)==0) {ww=matrix(1/nrow(g),nrow(g),1)}
diff=1000
hh=0

#inizio loop#

while (diff>criterio && hh<hlim)     {
 hh=hh+1
 w=ww

#trasformazioni#

 ss=svd(t(g)%*%g)
 gg=g%*%ss$u%*%diag(ss$d^(-0.5))
 gs=matrix(0,ncol(g),1)
 gi=matrix(0,ncol(g),1)
 for (jj in 1:ncol(g))       {
     gs[jj]=sum(w*(gg[,jj]))
     gi[jj]=sum(w*(g[,jj]))
     }
 if (ini==1) {init=gs}
 if (ini==2) {init=ginv(t(g)%*%g)%*%gi}
 if (ini!=1 || ini!=2) {init=matrix(0,ncol(g),1)}

#EL#

   if (GEL=="EL") {

         gr=function(p) {
         grad=matrix(0,ncol(g),1);
         for (z in 1:ncol(g)) {
         grad[z]=-sum(w*((gg[,z])/(1-(gg)%*%p)))
         }
         return(grad)
         }
         el=function(p) {
         y=(1-(gg)%*%p);
         ly=log(y);
         sum(w*ly)
         }

         outel=constrOptim(init,el,gr,ui=-(gg),ci=-1,control=list(fnscale=-1))
         ww=w*1/(1-(gg)%*%outel$par)
         ww=ifelse(ww>10^(-300),ww,10^(-300))
         ww=ww/sum(ww)
         mlambda[1:ncol(g),hh]=outel$par

         pg=matrix(0,1,ncol(g));
         for (z in 1:ncol(g)) {
```

```
         pg[z]=sum(g[,z]*ww)
         };
      }

#ET#

   if (GEL=="ET") {

         gr=function(p) {
         grad=matrix(0,ncol(g),1);
         for (z in 1:ncol(g)) {
         grad[z]=-sum(w*((gg[,z])*exp(gg%*%p)))
         }
         return(grad)
         }
         et=function(p) {
         ey=-exp(gg%*%p);
         sum(w*ey)
         }

         outel=optim(init,et,gr,control=list(fnscale=-1))
         ww=w*exp(gg%*%outel$par)
         ww=ifelse(ww>10^(-300),ww,10^(-300))
         ww=ww/sum(ww)
         mlambda[1:ncol(g),hh]=outel$par

         pg=matrix(0,1,ncol(g));
         for (z in 1:ncol(g)) {
         pg[z]=sum(g[,z]*ww)
         };
      }

#fine loop#

PG=pg/mom[1:ncol(g)]
diff=max(abs(PG))
if (diff<=criterio || hh==hlim)
{list(hh=hh,crit=diff,PG=PG,lambda=outel$par,mlambda=mlambda,ww=ww)}
}

}
#fine Inner#######################################################################


#inizio funzione ET#################################################################
#la funzione stima la verosimiglianze di massima entropia (ET) quando è presente
#una sola condizione sui momenti

#argomenti:
#hlim:      numero massimo di iterazioni
#criterio:  ordine di approssimazione al di sotto del quale si intende raggiunta
#           la convergenza
#ww:        vettore delle probabilità iniziali
#gm:        matrice delle condizioni di ortogonalità
#ini:       valori iniziali per i moltiplicatori; 1=valore atteso dei vincoli,
#           2=valore atteso dei vincoli riscalato, qualsiasi altro valore=
#           vettore nullo
```

```
#mom:       matrice dei momenti che debbono essere ottenuti

#output sotto forma di lista:
#hh:        numero di iterazioni effettuate
#crit:      massimo errore percentuale in valore assoluto nella ricostruzione
#           dei momenti
#PG:        errori percentuali sui momenti
#lambda:    vettore dei moltiplicatori all'ultima iterazione
#ww:        vettore delle "probabilità" associate ad ogni osservazione
##############################################################################
ET=function(hlim,criterio,ww,gm,ini,mom) {

g=gm
diff=1000
hh=0

#inizio loop#

while (diff>criterio && hh<hlim)     {
 hh=hh+1
 w=ww

#trasformazioni#

 gi=sum(w*g)
 if (ini==1) {init=gi}
 if (ini==2) {init=gi/crossprod(g)}
 if (ini!=1 || ini!=2) {init=0}

#ET#

 gr=function(p) {
    grad=sum(w*(g*exp(g*p)))
    return(grad)
    }
 et=function(p) {
    ey=-exp(g*p);
    -sum(w*ey)
    }

 outel=nlminb(init,et,gradient=gr)
       ww=w*exp(g*outel$par)
       ww=ifelse(ww>10^(-300),ww,10^(-300))
       ww=ww/sum(ww)

       pg=sum(g*ww)

#fine loop#
PG=pg
diff=max(abs(PG))
if (diff<=criterio || hh==hlim)
{list(hh=hh,crit=diff,PG=PG,lambda=outel$par,ww=ww)}
}

}
#fine ET#############################################################################
```

```
#inizio funzione EL##############################################################
#la funzione stima la verosimiglianze empirica (EL) quando è presente
#una sola condizione sui momenti

#argomenti:
#hlim:      numero massimo di iterazioni
#criterio: ordine di approssimazione al di sotto del quale si intende raggiunta
#           la convergenza
#ww:        vettore delle probabilità iniziali
#gm:        matrice delle condizioni di ortogonalità
#ini:       valori iniziali per i moltiplicatori; 1=valore atteso dei vincoli,
#           2=valore atteso dei vincoli riscalato, qualsiasi altro valore=
#           vettore nullo
#mom:       matrice dei momenti che debbono essere ottenuti

#output sotto forma di lista:
#hh:        numero di iterazioni effettuate
#crit:      massimo errore percentuale in valore assoluto nella ricostruzione
#           dei momenti
#PG:        errori percentuali sui momenti
#lambda:    vettore dei moltiplicatori all'ultima iterazione
#ww:        vettore delle "probabilità" associate ad ogni osservazione
##############################################################################
EL=function(hlim,criterio,ww,gm,ini) {

g=gm
estremi=1/g
if (sum(estremi[estremi>0])>0) {upper=min(estremi[estremi>0])}
if (sum(estremi[estremi>0])==0){upper=NULL}
if (sum(estremi[estremi<0])<0) {lower=max(estremi[estremi<0])}
if (sum(estremi[estremi<0])==0){lower=NULL}
diff=1000
hh=0

#inizio loop#

while (diff>criterio && hh<hlim)         {
 hh=hh+1
 w=ww

#trasformazioni#

  gi=sum(w*g)

  if (ini==1) {init=gi}
  if (ini==2) {init=gi/crossprod(g)}
  if (ini!=1 || ini!=2) {init=0}

#EL#

  gr=function(p) {
     grad=sum(w*(g/(1-g*p)))
     return(grad)
     }
  el=function(p) {
     y=(1-g*p);
```

```
      ly=log(y);
      -sum(w*ly)
      }

   outel=nlminb(init,el,gradient=gr,lower=lower,upper=upper)
   ww=w*1/(1-g*outel$par)
   ww=ifelse(ww>10^(-300),ww,10^(-300))
   ww=ww/sum(ww)

   pg=sum(g*ww)

#fine loop#

PG=pg
diff=max(abs(PG))
if (diff<=criterio || hh==hlim)
{list(hh=hh,crit=diff,PG=PG,lambda=outel$par,ww=ww)}
}

}
#fine EL##############################################################################


#inizio mom#############################################################################
#la funzione calcola i momenti pesati

#argomenti:
#ww:       il vettore dei pesi; se ww = NULL viene adottata una uniforme
#x:        vettore di osservazioni
#maxp:     ordine massimo dei momenti da calcolare

#output:
#M:        vettore dei momenti calcolati
##############################################################################################

mom=function(ww,x,maxp) {
if (length(ww)==0) {ww=matrix(1/length(x),length(x),1)}
X=outer(x,seq(1,maxp),"^")
fun=function(y) {sum(ww*y)}
M=apply(X,2,fun)
}
#fine mom##############################################################################


#inizio test############################################################################
#la funzione calcola il test di identità tra leggi che hanno in comune i primi
#q momenti

#argomenti:
#x:        vettore di osservazioni
#w0:       vettore di pesi ottenuto dalle prime npar-1 condizioni
#w1:       vettore di pesi ottenuto dalle prime npar condizioni
#npar:     numero di condizioni sui momenti
#digits:   numero di digit per la suddivisione in classi di un carattere
#          semicontinuo
#ttest:    test da effettuare; "LR"=rapporto di verosimiglianza,
#          "KS"=Kolmogorov-Smirnov, "DKL"=divergenza di Kullback-Leibler,
```

```
#          "BIC"=bayesian information criterion

#output:  il p-value per i primi tre test ovvero il BIC
##############################################################################
test=function(x,w0,w1,npar,digits,ttest)        {
if(ttest=="LR") {
LRatio=-2*log(prod(w1/w0))
t=pchisq(LRatio,npar,lower.tail=FALSE)
}

#Kolmogorov-Smirnov#

if(ttest=="KS") {
t=ks.test(w0,w1)$p.value
}

#DKL#

if(ttest=="DKL") {
dkl=2*length(w1)*sum(w1*log(w1/w0))
t=pchisq(dkl,npar,lower.tail=FALSE)
}

#BIC#

if(ttest=="BIC") {
t=-2*sum(log(w1))+npar*log(length(w1))
}

round(t,2)
}
#fine test#############################################################################


#inizio SamplingR#######################################################################
#la funzione genera campioni dalla predittiva di ampiezza pari o superiore a
#quella osservata, riproducendo l'ordine in cui si manifestano i ranghi

#argomenti:
#v:       vettore delle osservazioni
#S:       matrice di due colonne: supporto e probabilità della predittiva
#ds:      numerosità richiesta; 1=stessa numerosità, z>1=più piccola numerosità
#         eccedente z, ottenuta ricorsivamente
#replace: TRUE=con ripetizione, FALSE=senza ripetizione

#output:
#vvv:     il campione generato dalla predittiva
##############################################################################
SamplingR=function(v,S,ds,replace) {

R=apply(v,2,rank)

#sampling#

vv=matrix(0,nrow(R),ncol(R))
ff=function(k)
{if(nrow(v)>length(S[[k]][,1])) {replace=TRUE}
```

```
 c=sample(S[[k]][,1],nrow(R),replace=replace,prob=S[[k]][,2]);
 o=order(c);
 uu=c[o];
 uu[R[,k]]
}
vv=sapply(k,ff)

#ricorsione#

nvolte=ceiling((log(ds)-log(nrow(R)*2))/log(2))+1
RR=apply(vv,2,rank)
vvv=vv
if (ds!=1) {
for (volte in 1:nvolte){
vv=matrix(0,nrow(RR),ncol(RR))
ff=function(k)
{if(nrow(v)>length(S[[k]][,1])) {replace=TRUE}
 c=sample(S[[k]][,1],nrow(RR),replace=replace,prob=S[[k]][,2]);
 o=order(c);
 uu=c[o];
 uu[RR[,k]]
}
vv=sapply(k,ff)
vvv=rbind(vvv,vv)
RR=apply(vvv,2,rank)
}
}
vvv
}
#fine SamplingR#######################################################################


#inizio ICM#############################################################################
#la funzione stima la legge di probabilità della  variazione percentuale tra
#momenti misti ottenuti da campioni estratti dalla  predittiva e momenti misti
#calcolati sui dati, fornendo intervalli hpd per un prefissato livello di
#significatività

#funzioni richiamate: SamplingR, hpd

#argomenti:
#v:       vettore delle osservazioni
#S:       predittiva (argomento di SamplingR)
#nrep:    numero di replicazioni
#eee:     numero di momenti misti da testare
#maxp:    ordine massimo delle potenze per ciascuna variabile
#replace: TRUE=con ripetizione, FALSE=senza ripetizione (argomento di SamplingR)
#alfa:    livello di significatività (argomento di hpd)
#steps:   numero di punti per i quali stimare la densità (argomento di hpd)

#output sotto forma di lista:
#IC:      intervalli di credibilità
#ptz:     potenze usate per il calcolo dei momenti misti
#######################################################################################
ICM=function(v,S,nrep,eee,maxp,replace,alfa,steps) {
ptz=matrix(0,eee,ncol(v))
ff=function(x) {round(runif(eee,1,maxp))}
```

```
ptz=apply(ptz,2,ff)

#misti osservati#

MMo=matrix(0,eee,1)
for (kk in 1:eee)                    {
Ma=t(t(v)^ptz[kk,])
Mb=matrix(1,nrow(v),1)
for (jj in 1:ncol(Ma))               {
Mb=Mb*Ma[,jj]
}
MMo[kk]=mean(Mb)
}

#replicazioni#

MMs=array(0,c(nrep,eee))
for(rep in 1:nrep) {
vvv=SamplingR(v,S,ds=1,replace=replace)
for (kk in 1:eee)                    {
Ma=t(t(vvv)^ptz[kk,])
Mb=matrix(1,nrow(vvv),1)
for (jj in 1:ncol(Ma))               {
Mb=Mb*Ma[,jj]
}
MMs[rep,kk]=(mean(Mb)-MMo[kk])/MMo[kk]
}
}
hpd.e=function(x) {hpd(x,alfa,steps=steps)$hpdr}
IC=apply(MMs,2,hpd.e)
list(IC=IC,ptz=ptz)
}
#fine ICM##########################################################################


#inizio hpd#########################################################################
#la funzione, mutuata da Liseo (2006), stima gli IC hpd

#argomenti:
#x:       vettore delle osservazioni
#alfa:    livello di significatività
#steps:   numero di punti per i quali stimare la densità
#from:    estremo inferiore del supporto, se noto
#to:      estremo superiore del supporto, se noto

#output sotto forma di lista:
#hpdr:    estremi dell'IC
#moda:    moda della densità
##################################################################################
hpd=function(x,alfa,steps,from,to){
dx=density(x,n=steps,from=from,to=to)
md=dx$x[dx$y==max(dx$y)]
px=dx$y/sum(dx$y)
pxs=-sort(-px)
ct=min(pxs[cumsum(pxs)< alfa])
list(hpdr=range(dx$x[px>=ct]),moda=md)
}
```

```
#fine hpd##########################################################################


#inizio concor#######################################################################
#la funzione stima basi e aree dell'istogramma ottenibile a
#partire da una densità discretizzata

#argomenti:
#x:      punti del supporto
#p:      corrispondenti masse di probabilità

#output sotto forma di lista:
#px:     aree dei bin
#h:      basi dei bin
##################################################################################
concor=function(x,p) {
o=order(x)
x=x[o]
p=p[o]
z=matrix(0,(length(x)+2),1)
z[2:(nrow(z)-1)]=x
z[1]=2*x[1]-x[2]
z[nrow(z)]=2*x[length(x)]-x[length(x)-1]
y=embed(z,2)
h05=rowMeans(y)
h=diff(h05)
px=p*h
list(px=px,h=h)
}
#fine concor#########################################################################
```

B.2 Un Programma

```
#programma di esempio per la stima della predittiva#############################

#input:
#v:     un dataset sotto forma di matrice
#...:   i parametri specifici delle singole funzioni richiamate

#output:
#S:     distribuzione predittiva
#PMat:  realizzazioni per ciascuna classe individuata dai ranghi
#Psint: medie delle sezioni
##################################################################################

k=seq(1,ncol(v))

#posteriori dei momenti#

fm=function(x) {bb(x,nrep=nrep,maxp=maxp,trim=trim,wins=wins)}
M=apply(v,2,fm)

TT=array(0,c(nrep,maxp))
```

```
cntrmoda=0 #parametro che posto=1 vincola all'unimodalità#

fu=function(k)  {

#definizione supporto: i parametri in questo esempio sono vettori con un numero#
#                      di elementi pari al numero di variabili (colonne di v)  #

S=MatDen(v[,k],tipo=tipo[k],segno=segno[k],alfa=alfa[k],steps=steps[k],
         inf=inf[k],sup=sup[k])

LL=array(0,c(length(S[,1]),3))

#inizio ciclo campionamento dalla posterior#

for (spo in 1:nrow(M[[k]]$MM)) {

#condizioni sui momenti#

G=g.mom(S[,1],maxp=maxp,par=M[[k]]$MM[spo,])

#inizio ciclo sui momenti#

lim=2
for (mm in lim:(maxp-1)) {

#stima della verosimiglianza#

L=Inner(hlim=hlim,criterio=criterio,S[,2],G,GEL=GEL,ini=ini,
  mom=M[[k]]$MM[spo,],lim=mm)

#controllo unimodalità#

if (cntrmoda==1) {
zeri=ifelse(diff(L$ww)<=0,-1,1)
zeri=which(rowSums(embed(zeri,2))==0)
if (length(zeri)>1) {break}
}

LL[,2]=S[,2]=L$ww

#test di arresto "DKL", livello di significatività=0.1#
TT[spo,mm]=test(S[,1],w0=LL[,1],w1=LL[,2],npar=1,digits=digits,ttest="DKL")

if (mm<=minp) {LL[,1]=LL[,2]}
if (mm>minp) {
if (TT[spo,mm]>0.1)  {TT[spo,maxp]=mm;break}
{LL[,1]=LL[,2]}
}

#fine ciclo sui momenti#
}

LL[,3]=LL[,3]+LL[,2]/nrow(M[[k]]$MM)

#fine ciclo campionamento dalla posteriori#
}
```

```
S[,2]=LL[,3]
list(S=S,LL=LL,TT=TT)
}

H=lapply(k,fu)

fu=function(k)
{S=H[[k]]$S;
 sup=unique(S[,1]);
 f=factor(S[,1]);
 prob=tapply(S[,2],f,sum);
 dimnames(prob)=NULL;
 cbind(sup,prob)}
S=lapply(k,fu)

#sezioni dei processi: in questo esempio sono costruite con 5000 elementi#

realiz=5000
time=as.matrix(seq(1,nrow(v)))
PMat=array(0,c(nrow(v),realiz,ncol(v)))
for (xx in 1:realiz) {
PMat[,xx,]=SamplingR(v,S,ds=1,replace=replace)
}

#estremi delle sezioni#

mini=apply(PMat,c(1,3),min)
maxi=apply(PMat,c(1,3),max)

#medie delle sezioni#
Psint=v*0
for (rp in 1:nrow(v)) {
for (rc in 1:ncol(v)) {
Psint[rp,rc]=mean(PMat[rp,,rc])
}
}

#grafici#

windows()
dimplot=ceiling(sqrt(ncol(v)))
par(mfrow=c(dimplot,dimplot))
xx=seq(1,ncol(v))
for (j in xx) {
plot(mini[,j],type="l",col="red1", lwd=1, ylim=c(min(mini[,j]),max(maxi[,j])),
main=paste(names(v)[j]," ed estremi di ",realiz," traiettorie"),sub=NULL,
xlab="indice delle sezioni", ylab=names(v)[j], cex.main=1.3,cex.lab=1.5,
cex.axis=1.3)
lines(v[,j],col="black",lwd=2)
lines(maxi[,j],type="l",col="red1",lwd=1)
}

windows()
dimplot=ceiling(sqrt(ncol(v)))
par(mfrow=c(dimplot,dimplot))
leg.txt=c("osservata",paste("media di ",realiz," traiettorie"))
for (dp in 1:ncol(v)) {
```

```
plot(v[,dp],type="l", lwd=2,
main=paste("serie ",names(v)[dp]),sub=NULL, xlab="tempo", ylab=names(v)[dp],
cex.main=1.5,cex.lab=1.5,cex.axis=1.3)
lines(Psint[,dp],col="red")
legend("topleft", legend = leg.txt, col=c("black","red"), lty=1, lwd=2, cex=0.9,
inset=0.01)
}


#test su momenti misti selezionati casualmente#

r.test=ICM(v,S,nrep=nrep,eee=eee,maxp=maxp,replace=replace,alfa=alfa,
       steps=steps)

#termine del programma#############################################################
```

Bibliografia

AA.VV. (2007). *Handbook on Statistical Disclosure Control*. CENtre of EXcellence for Statistical Disclosure Control.

Aitkin M. (2008). Applications of the bayesian bootstrap in finite population inference. *Journal of Official Statistics*, **24**, 21–51.

Bera A.; Bilias Y. (2002). The mm, me, ml, el, ef, and gmm approaches to estimation: a synthesis. *Journal of Econometrics*, **107**, 51–86.

Bernardo J. M.; Smith A. F. (2000). *Bayesian Theory*. John Wiley & Sons Ltd.

Cappuccio N.; Orsi R. (2005). *Econometria*. Il Mulino, Bologna.

Chesher A.; Smith R. J. (1997). Likelihood ratio specification tests. *Econometrica*, **65**, 627–646.

Cifarelli D. M.; Regazzini E. (1982). *Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability*, pp. 185–205. North-Holland Publishing Company.

Corcoran S. A. (2000). Empirical exponential family likelihood using several moment condition. *Statistica Sinica*, **10**, 545–557.

Cressie N.; Read T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440–464.

Davidson J. (1994). *Stochastic Limit Theory*. Oxford University Press.

de Finetti B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Ann. de l'Inst. H. Poincarè*, **7**, 1–68.

Efron B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Ericson W. A. (1969). Subjective bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B*, **31**, 195233.

Ferguson T. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Gasparini M. (1995). Exact multivariate bayesian bootstrap distributions of moments. *The Annals of Statistics*, **23**(3), 762–768.

Grendar M.; Judge G. (2007). A bayesian large deviations probabilistic interpretation and justification of empirical likelihood. Working paper, University of California, Berkeley, Department of Agricultural & Resource Economics.

Hall P.; Presnell B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **61**, 143–158.

Hampel F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**(346), 383–393.

Hansen L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Huber P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73–101.

Jaynes E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, **106**, 620–630.

Jaynes E. T.; Bretthorst G. L. (2003). *Probability theory: the logic of science*. Cambridge University Press.

Johansen S.; Juselius K. (1990). Maximum likelihood estimation and inference on cointegration - with application to the demand for money. *Oxford Bulletin of Economics and Statistics*, **52**(2), 169–210.

Kim J. Y. (2002). Limited information likelihood and bayesian analysis. *Journal of Econometrics*, **107**, 175–193.

Kitamura Y. (2006). Empirical likelihood methods in econometrics: Theory and practice. Discussion Paper 1569, Cowles Foundation.

Lazar N. A. (2003). Bayesian empirical likelihood. *Biometrika*, **90**, 319–326.

Lindsay B.; Basak P. (2000). Moments determine the tail of a distribution (but not much else). *The American Statistician*, **54**(4), 248–251.

Liseo B. (2006). *Introduzione alla statistica bayesiana*. Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale - Università degli studi di Roma La Sapienza.

Liseo B.; Montanari G. E.; Torelli N. (2006). *Metodi Statisitici per l'integrazione di Dati da Fonti Diverse*. Franco Angeli, Milano, 1 edizione.

McLeish D. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability*, **2**(4), 620–628.

Muliere P.; Secchi P. (1992). Exchangeability, predictive sufficiency and bayesian bootstrap. *Journal of the Italian Statistical Society*, **3**, 377–404.

Muliere P.; Secchi P. (1994). Una generalizzazione del bootstrap bayesiano. In *Atti XXXVII Riunione Scientifica SIS*, volume 2, pp. 527–534.

Muliere P.; Secchi P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics*, **48**, 663–673.

Newey W. K.; Smith R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, **72**, 219–255.

Owen A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Pfaff B. (2006). *Analysis of Integrated and Cointegrated Time Series with R*. Springer, New York, prima edizione. ISBN 0-387-27960-1.

Piccinato L. (1996). *Metodi per le decisioni statistiche*. Springer - Collana di Statistica.

Piccolo D. (2000). *Statistica*. Il Mulino, Bologna.

Politis D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, **18**, 219–230.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Regazzini E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giornale dell'Istituto Italiano degli Attuari*, pp. 77–89.

Reid N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, **3**(2), 213–238.

Rubin D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, **9**, 130–134.

Schennach S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, **92**, 31–46.

Schennach S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, **35**, 634–672.

Smith R. J. (1997). Alternative semi-parametric likelihood approaches to generalized method of moments estimation. *Economic Journal*, **107**, 503–519.

Vinod H. D. (2004). Ranking mutual funds using unconventional utility theory and stochastic dominance. *Journal of Empirical Finance*, **11**, 353–377.

White H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Wu X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, **115**, 347–354.