Inferenza non parametrica nel contesto di dati dipendenti: Polinomi Locali e Verosimiglianza Empirica

Indice

Capitolo 1

Il contesto operativo

1.1 Introduzione

Il presente lavoro si inserisce nel contesto delle più recenti ricerche sugli strumenti di analisi nonparametrica ed in particolare analizza l'utilizzo dei polinomi locali e della verosimiglianza empirica, nel caso di dati dipendenti.

La principale categoria di dati dipendenti di interesse, ad oggi, è rappresentata dalle serie storiche che sono in grado di rappresentare un qualsiasi fenomeno e la sua tendenza evolutiva, legando la variabile esplicativa Y_t al valore che la stessa variabile ha assunto in un periodo precedente.

E' assolutamente indubbia l'utilità di elaborare una tale applicazione, poichè questo tipo di dati costituisce sicuramente quello più interessante per quanto riguarda ricerche di tipo economomico - finanziario.

Le principali forme di dipendenza che verranno trattate in questo lavoro sono quelle che rispondono alla definizione di $\alpha-mixing$ ed in particolare il nostro si presenta come un tentativo di conciliare, in questo ambito, tecniche non parametriche, rappresentate dai polinomi locali, all'approccio di Empirical Likelihood, cercando di aggregare ed enfatizzare i punti di forza di entrambe le metodologie: i polinomi locali ci forniranno una stima più e accurata da collocare all'interno della definizione di verosimiglianza empirica fornita da Owen (1988).

I vantaggi sono facili da apprezzare in termini di immediatezza ed utilizzo pratico di questa tecnica. I risultati vengono analizzati sia da un punto di vista teorico, sia confermati poi, da un punto di vista empirico, riuscendo a trarre dai dati anche utili informazioni in grado di fornire l'effettiva sensibilità al più cruciale e delicato parametro da stabilire nel caso di stimatori polinomi locali.

Lungo tutto l'elaborato presenteremo, in ordine, dapprima il contesto all' interno del quale andremo ad operare, precisando più nello specifico le forme di dipendenza trattate. All'interno del capitolo secondo, enunceremo le caratteristiche e proprietà dei polinomi locali, successivamente, nel corso del capitolo terzo, analizzeremo nel dettaglio la verosimiglianza empirica, con particolare attenzione, anche in questo caso, alle proprietà teoriche, infine, nel quarto capitolo presenteremo risultati teorici personali, conseguiti a partire dalla trattazione teorica precedente.

Il capitolo conclusivo propone uno studio di simulazione, sulla base delle proprietà teoriche ottenute nel capitolo precedente. Non soltanto, avvalendosi di tali risultati, ne conferma la validità, ma aggiunge anche un'analisi, per i test proposti, alla sensibilità rispetto al parametro di *smoothing* impiegato. Nelle battute conclusive troveranno spazio delucidazioni sui risultati delle simulazioni.

1.2 *Mixing* di processi stocastici

Prima di concentrarci, nello specifico, alle tipologie di stimatori impiegati ci soffermiamo ora sul contesto all'interno del quale andiamo ad operare, fissando con particolare attenzione per le varie forme di dipendenza riscontrabili nel caso di serie storiche di dati.

Con mixing di processi stocastici intendiamo delimitare una particolare forma di dipendenza, presente nei dati generati dal processo.

Diverse sono le categorie di dipendenza che è possibile riscontrare, le principali sono rappresentate dalle forme di $\alpha-mixing$, $\beta-mixing$, $\rho-mixing$ e $\varphi-mixing$.

1.2.1 Definizione e tipologie di *mixing*

Dato uno spazio di probabilità (Ω, A, P), con B e C due sottospazi di A, la correlazione presente tra B e C può essere espressa mediante vari coefficienti, ognuno dei quali specifica una diversa forma di dipendenza.

Tali coefficienti vengono identificati come segue:

- $\alpha-mixing$

$$\alpha = \alpha(B,C) = \sup_{\substack{B\in\mathcal{B}\\C\in\mathcal{C}}} |P(B\cap C) - P(B)\,P(C)|\,, \tag{1.1}$$

- $\beta-mixing$

$$\beta = \beta(B,C) = E \sup_{C\in\mathcal{C}} |P(C) - P(C|\mathcal{B})|\,, \tag{1.2}$$

- $\rho-mixing$

$$\rho = \rho(B,C) = \sup_{\substack{X\in L^2(\mathcal{B})\\Y\in L^2(\mathcal{C})}} |corr(X,Y)|\,, \tag{1.3}$$

- $\varphi-mixing$

$$\varphi = \varphi(B,C) = \sup_{\substack{B\in\mathcal{B},P(B)>0\\C\in\mathcal{C}}} |P(C) - P(C|B)|\,, \tag{1.4}$$

Un processo $(X_t, t\in Z)$ é $\alpha-mixing$ o *strongly mixing* se

$$\alpha_k = \sup_{t\in Z} \alpha(\sigma(X_s, s\le t), \sigma(X_s, s\le t+k)) \xrightarrow[k\to+\infty]{} 0 \tag{1.5}$$

In maniera analoga si determinano le altre forme di *mixing*, utilizzando i coefficienti sopra definiti.

E' possibile verificare che le varie forme di mixing di processi stocastici, in base alle definizioni date, sono legate da un rapporto di implicazione spiegato dallo schema sottostante.

$$\varphi - mixing \Longrightarrow \begin{cases} \beta - mixing\,,\ \rho - mixing \Longrightarrow \{\alpha - mixing \end{cases}$$

La dipendenza di tipo $\phi-mixing$, quindi, implica sia la $\beta-mixing$ che la $\rho-mixing$ e, a loro volta, queste ultime due implicano la $\alpha-mixing$.

Occorre precisare che lo schema non è corretto se viene letta per ogni freccia la doppia implicazione (Bosq,1998).

Capitolo 2

Gli stimatori Polinomi Locali nel caso di dati dipendenti

2.1 Introduzione

L'applicazione di tecniche non parametriche alle serie storiche è una tradizione ormai consolidata, a partire dalla prima e più rudimentale metodologia di analisi rappresentata dal periodogramma.

Sebbene in grado di restituire una stima non distorta della densità spettrale di un sottostante processo stazionario, questo strumento però non si è rivelato capace di fornire una stima che risultasse consistente.

E' nata così la necessità di implementare *finestre spettrali* o di utilizzare un sistema di pesi che consentisse di ovviare a questa limitazione e diverse sono state le tipologie di finestre proposte per ottenere stime consistenti.

Col tempo le tecniche non parametriche applicate ai dati storici hanno suscitato sempre maggiore interesse e sono stati prodotti quindi copiosi lavori, atti ad affinare tecniche e metodologie che rispondessero all'esigenza di perfezionare continuamente i risultati precedenti già ottenuti.

La tecnica non parametrica oggetto d'indagine all'interno del nostro lavoro è quella dei polinomi locali, metodologia che nasce alla fine degli anni '70, a partire dai lavori di Stone (1977) e Cleveland (1979), per suscitare poi grande interesse negli anni immediatamente successivi, grazie alle interessanti proprietà di questi stimatori messe in luce da vari autori, quali Tsybakov (1986), Fan (1993), Hastie e Loader (1993), Ruppert e Wand (1994), Fan e Gijbels (1995).

I primi risultati e le prime applicazioni di polinomi locali hanno riguardato principalmente i dati indipendenti ed identicamente distribuiti.

Soltanto a partire dalla fine degli anni '90 recenti e innovative applicazioni a dati dipendenti sono state rese possibili a seguito dei lavori pionieristici di Masry e Fan (1997), Masry (1996a), Masry (1996b), Härdle e Tsybakov (1997), Opsomer (1997), Vilar-Fernández e Vilar-Fernández (1998), Vilar-Fernández e Vilar-Fernández (2000), i quali hanno messo in luce metodologie di applicazione e proprietà per *mixing* di processi stocastici.

La scelta di adottare la tecnica dei polinomi locali è dovuta alle principali proprietà che essa garantisce. Prima di tutto essa consente una riduzione della distorsione rispetto allo stimatore Nadaraya-Watson e della varianza rispetto

allo stimatore Gass-Müller, come dimostrato da Chu e Marron (1991) e Fan (1992), si adatta automaticamente ai bordi Fan, Gasser, Gijbels, Brockmann e Engel (1997), ed in particolare la sua superiorità viene evidenziata anche nel caso di stima delle derivate. Cruciale per l'utilizzo dei polinomi locali è la scelta del parametro di *bandwidth* al quale dedicheremo particolare attenzione nel corso della trattazione.

2.2 I polinomi locali

Nell'ambito degli studi econometrico-finanziari la strutturazione di modelli in grado di rappresentare la complessità del contesto reale ha condotto alla formulazione di modelli dinamici la cui struttura può essere sintetizzata dal modello:

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t \qquad t \in \mathbb{Z}. \tag{2.1}$$

Nella presente formulazione, viene illustrato il caso generico in cui Y rappresenta la variabile dipendente ed $m(\cdot)$ ne dichiara la relazione esistente con la variabile X, spiegata a meno di un errore, inosservabile e casuale, ε, che generalmente è definito a media nulla e varianza finita.

$\{Y_t, X_t\}$ rappresenta il processo con $X_t \in \mathbb{R}^{\,d}$, $Y_t \in \mathbb{R}$.

Il modello viene comunemente applicato alle serie storiche, le quali rispondono naturalmente all'esigenza di esprimere la dipendenza delle osservazioni con le realizzazioni della stessa variabile nel tempo passato.

In quest'ultima ipotesi, il modello viene schematizzato nella seguente forma:

$$Y_t = m(Y_{t-1}) + \sigma(Y_{t-1})\varepsilon_t \qquad t \in \mathbb{Z} \tag{2.2}$$

dove la variabile esplicativa è la stessa variabile oggetto d'indagine ritardata nel tempo, Y_{t-1}. Diversi sono gli autori che hanno illustrato le proprietà probabilistiche del processo, come Doukhan e Ghindés (1980,1981), Chan e Tong (1985), Mokkadem (1987), Diebolt e Guégan (1990) e Ango Nze (1992).

Interesse dei ricercatori, per poter definire Y_t, in entrambe le formulazioni, è quello di ottenere una stima delle due funzioni $m(\cdot)$ e $\sigma^2(\cdot)$, che rapprentano, rispettivamente, la media e la varianza condizionata del processo.

Poiché il nostro lavoro si inserisce in un contesto di dati dipendenti, nel seguito faremo riferimento alla seconda formulazione, che più propriamente si adatta al caso in esame. Il nostro obiettivo è stimare la funzione di regressione:

$$m(x) = E(Y_t | Y_{t-1} = x), \tag{2.3}$$

quando gli errori sono dipendenti e soddisfano le ipotesi di $\alpha-mixing$ descritte nel capitolo precedente.

Solitamente, per effettuare la stima di tale funzione si ricorre allo sviluppo, tramite l'espansione di Taylor, che ci fornisce la seguente approssimazione:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \cdots + \frac{m^{(p)}(xo)}{p!}(x - x_0)^p \tag{2.4}$$

mediante la quale sarà poi possibile, ottenuta una stima delle derivate della funzione, con il metodo dei polinomi locali, stimare la funzione stessa.

Nel seguito, sotto l'assunzione che esistano, calcolate in x, le $(p+1)$ derivate della funzione di regressione, presenteremo lo stimatore polinomi locali per la funzione di regressione e ne analizzeremo le proprietà teoriche.

2.2.1 Definizione e caratteristiche dello stimatore

Lo stimatore polinomi locali, $\beta(x) = (\beta_0(x), \beta_1(x), \ldots, \beta_p(x))'$, con $\beta_j(x) = m^{(j)}(x)/(j!)$, $j = 0, 1, \ldots, p$, viene calcolato minimizzando la quantità:

$$\sum_{t=1}^{n} \left(Y_t - \sum_{j=0}^{p} \beta_j(x) \, (x_t - x)^j \right)^2 \omega_t, \tag{2.5}$$

ovvero tramite la tecnica dei minimi quadrati pesati. Gli $\omega_t = n^{-1} K_h(x_t - x)$ rappresentano i pesi, mentre K è la funzione *kernel*.

Convenzionalmente $K_h(u) = h^{-1} K(h^{-1}u)$, invece h, più correttamente h_n, rappresenta la cosiddetta *finestra*, parametro che determina l'ampiezza dell'intorno considerato e, di conseguenza, il grado di *smoothing* effettuato.

Per ragioni di semplicità notazionale nel seguito ometteremo l'indice n, considerando implicita la dipendenza di h da n.

La determinazione della finestra, come già accennato, è una delle fasi più delicate e cruciali per la stima effettuata tramite i polinomi locali, perché questo parametro è in grado di influenzare il trade-off *bias-variance* esistente per lo stimatore. Proprio per il notevole rilievo che questa scelta assume, dedicheremo particolare attenzione a questa fase nel seguito della trattazione, per ora esaminiamo più nello specifico la natura dello stimatore e delle stime prodotte.

In termini matriciali la minimizzazione della (2.5) corrisponde al problema di minizzare, rispetto a β , il prodotto matriciale $(Y - X\beta)'W(Y - X\beta)$.

La strutturazione delle matrici coinvolte nel modello espresso dalla (2.1) è data da:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & (x_1 - x) & \ldots & (x_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_n - x) & \ldots & (x_n - x)^p \end{pmatrix}. \tag{2.6}$$

Assumendo l'invertibilità di $X'WX$, con $W = diag(\omega_1, \ldots, \omega_n)$, si ottiene, Fan and Gijbels (1996), lo stimatore polinomi locali:

$$\hat{\beta}(x) = (X'WX)^{-1} X'WY. \tag{2.7}$$

Costruito $\hat{\beta}(x)$, si determina con semplicità lo stimatore polinomi locali per le varie derivate $m^{(j)}(x)$, che ci consentono quindi di stimare $m(x)$.

Lo stimatore di $m^{(j)}(x)$, infatti, sarà dato da $\hat{m}^{(j)(x)} = j!\hat{\beta}_j(x)$, dove $\hat{\beta}_j(x)$ rappresenta il j-esimo componente di $\hat{\beta}(x)$, $j = 0, 1, \ldots, p$.

La media e la varianza condizionata di $\widehat{\beta}(x)$, saranno date da:

$$E(\widehat{\beta}|X) = \beta + (X'WX)^{-1} X'Ws\,; \tag{2.8}$$

$$Var(\widehat{\beta}|X) = (X'WX)^{-1}(X'\Sigma X)(X'WX)^{-1}, \tag{2.9}$$

con $s = (m(X_1), \ldots, m(X_p)) - X\beta$ e $\Sigma = diag\{K_h^2(X_i - x_0)\sigma^2(X_i)\}$.

Altro stimatore a cui siamo interessati, oltre a quello della media condizionale, $m(x)$, è quello del momento secondo, $E(Y_t^2|X_{t-1} = x)$, stimatore che ci consentirà poi di fare inferenza sulla varianza condizionata del processo.

Definito il vettore

$$Y_2 = \begin{pmatrix} Y_1^2 \\ \vdots \\ Y_n^2 \end{pmatrix}, \tag{2.10}$$

lo stimatore della media condizionata del processo è dato da:

$$\hat{m}(x) = (X'WX)^{-1}X'WY; \tag{2.11}$$

mentre lo stimatore del momento secondo corrisponderà a:

$$\hat{m}_2(x) = (X'WX)^{-1}X'WY_2. \tag{2.12}$$

I parametri in grado di influenzare la qualità della stima sono: l'ampiezza della finestra, h, l'ordine di approssimazione, p, infine la funzione kernel, K, che nel nostro caso rappresenta la funzione peso.

Non tutti i parametri elencati influiscono, però, allo stesso modo sui risultati prodotti. E' noto in letteratura che la stessa funzione kernel riesce a modificare sostanzialmente le stime generate in maniera quasi ininfluente rispetto a quanto accade, invece, al variare dell'h.

E' per questo motivo che solitamente si impongo per K condizioni poco restrittive, quali la simmetria e la limitatezza del supporto, necessarie a garantire la validità di tutte le proprietà degli stimatori polinomi locali.

Analogamente, per quanto attiene al grado di approssimazione del polinomio p, si interviene, convenzionalmente, limitando la scelta ad un valore di grado dispari (Fan e Gijbels,1995; Ruppert e Wand, 1994).

Si preferisce, comunque, un ordine di approssimazione non troppo elevato, in quanto, incrementando il valore di p, alla riduzione della distorsione corrisponde una maggiore varianza, fenomeno del tutto indesiderato.

2.2.2 Proprietà teoriche dei Polinomi Locali per dati dipendenti

Numerose sono le proprietà e i vantaggi derivanti dall'applicazione dei polinomi locali alla regressione. Di seguito proponiamo un breve *excursus* in cui esibiamo i risultati noti, ottenuti tramite l'utilizzo dei suddetti stimatori.

E' stato Stone (1977) il primo ad introdurre l'utilizzo di una funzione che costituisse un sistema di peso usato per stimare la probabilità condizionale di una variabile di risposta Y, a partire dai corrispondenti valori di X.

Qualche anno dopo Cleveland (1979) ha esteso questa prima idea di Stone suggerendo un algoritmo capace di ottenere la stima di una curva che fosse robusta rispetto agli outliers. Successivamente, lo stesso autore (Cleveland 1988), si è occupato di descrivere le proprietà dello stimatore dei polinomi locali applicato al problema di regressione.

Fan e Gijbels (1992) hanno stabilito poi alcune proprietà per lo stimatore polinomi locali, a partire dalla formulazione che considera l'ipotesi di finestra variabile. Essi hanno ricavato i valori della distorsione condizionata e della varianza condizionata dello stimatore $\hat{m}(x)$ ottenuto dalla minimizzazione di

$$\sum_{t=1}^{n} \left(Y_t - \beta_0 - \beta_1(x - X_j)\right)^2 \alpha(X_j) K\left(\frac{(x - X_j)}{h_n}\alpha(X_j)\right) \tag{2.13}$$

Il caso generico trattato include chiaramente quello specifico in cui il parametro che stabilisce la variabilità della finestra, $\alpha(X_j)$ è pari ad uno, riconducendo la formulazione all'ipotesi in cui la finestra risulta fissa.

I risultati di Fan e Gijbels, riferiti al caso in cui la variabile X sia univariata, sono stati estesi da Ruppert e Wand (1994) nel caso di dati multivariati.

Anche in questo caso sono state ottenute precise formulazioni per la distorsione e la varianza condizionale dello stimatore, utili ad effettuare valutazioni sull' MSE condizionale. L'attenzione per i valori condizionati è giustificata dal fatto che il condizionamento garantisce che i momenti dello stimatore esistano, con probabilità che tende ad uno.

Tra le varie proprietà dello stimatore polinomi locali, Fan (1993) ha dimostrato che il modello locale lineare che utilizza il kernel di Epanechnikov ottimizza il minimax rischio lineare. Si parla di minimax efficienza, il valore di rischio *minimax* è un criterio utilizzato da benchmark per l'efficienza di uno stimatore che restituisce la numerosità campionaria necessaria ad ottenere una certà qualità di risultato. Fan, Gasser, Gijbels, Brockmann e Engel (1997) hanno poi esteso questi primi risultati allo stimatore polinomi locali di ordine p e nel caso di stime di derivate. Altro vantaggio derivante dall'applicazione dei polinomi locali è rappresentato dalle buone performance di questo stimatore negli estremi, *boundary points*.

E' una caratteristica della maggior parte degli stimatori non parametrici quella di avere un andamento diverso nei pressi degli estremi del supporto dei dati, rispetto a quanto accade inveci per i punti interni a questo intervallo.

Fan e Gijbels (1992), ad esempio, hanno dimostrato che gli stimatori di Nadaraya-Watson e Gasser-Müller convergono più lentamente negli estremi dell' intervallo. Lo stesso accade nel caso di dati multivariati, come attestano Ruppert e Wand (1994).

Altra caratteristica accertata per questi stimatori, sia nel caso univariato che in quello multivariato, è data dal fatto che la varianza condizionale negli estremi risulta maggiore rispetto a quella calcolata per i punti interni.

Fan e Gijbels spiegano questa evidenza con la presenza di un minor numero di punti utilizzati nell'implementare una stima nei pressi dei punti estremi dell'intervallo. Tra le differenze presenti, oltre ad una più corposa varianza, Ruppert e Wand (1994) segnalano anche che le stime per l'intercetta e il coefficiente angolare risultano non asintoticamente ortogonali negli estremi, a dispetto di quanto accade per i punti interni.

Cheng et al. (1997) attestano che non esiste nessuno stimatore lineare capace di superare le performance dello stimatore polinomi locali agli estremi, in una logica *minimax*, per quanto riguarda l' MSE. Gli autori, invece che provare, per via diretta, che qualsiasi altra correzione possibile risulta inferiore rispetto ai risultati ottenuti dai polinomi locali, saggiano l'ottimalità dello stimatore polinomi locali, in una logica *minimax*, dimostrando che nessun altro stimatore può apportare un miglioramento, in termini di efficienza, rispetto a quanto realizzato dai polinomi locali. Lo stimatore polinomi locali gode, dunque, dell'efficienza *minimax* sia per punti interni che negli estremi. Si rimanda a Hastie e Loader (1993) per ulteriori approfondimenti sul tema.

Ulteriore vantaggio conseguente dall'utilizzo dei polinomi locali risiede nel fatto che, lavorando in un contesto locale, non rileva sapere se la varianza condizionale, $var(Y|X = x)$, sia o meno costante nell'intervallo, in quanto in un intorno locale la differenza sarebbe comunque trascurabile.

Tutti questi risultati giustificano la preferenza per lo stimatore polinomi locali quale strumento utilizzato nella nostra analisi. Occupandoci in questa sede dell'applicazione per la stima della funzione di regressione e della varianza condizionale, nel caso di dati dipendenti, analizziamo più tecnicamente le proprietà degli stimatori per questi due funzionali.

All'interno della copiosa letteratura in tema di serie storiche, molti sono i lavori che si sono occupati di stimare la funzione di regressione, ovvero la media condizionata, con modelli di tipo VAR o modelli strutturali (Lütkepohl, 1992).

Per quanto riguarda invece la varianza condizionata si è inizialmente scelto di ipotizzare che essa fosse fissa o che avesse una legge predeterminata fino agli anni '80, quando Engle (1982) e Robinson (1983,1984), nella letteratura econometrica, Collomb (1984) e Vieu (1995), nella letteratura statistica, hanno messo in evidenza il problema della determinazione di tale quantità.

Ciò ha condotto ad un particolare interessamento verso questa problematica che, nei modelli ARCH di Engle, viene risolta assumendo la varianza condizionata come combinazione lineare dei quadrati delle innovazioni del processo.

Nel presente lavoro ci riconduciamo all'impostazione trattata da Härdle e Tsybakov (1997), i quali partendo dal modello di tipo (2.2) forniscono interessanti caratteristiche degli stimatori di $m(\cdot)$ e $\sigma^2(\cdot)$, derivando la loro convergenza in probabilità e la rispettiva distribuzione in legge.

Convergenza in probabilità e normalità asintotica dello stimatore polinomi locali

Härdle e Tsybakov (1997) partono dalla constatazione che il primo caso in cui si discute il problema congiunto di stimare la media e la varianza condizionale è dato dal lavoro di Gouriéroux e Montfort (1992), tramite un modello del tipo:

$$Y_t = \sum_{j=1}^{J} \alpha_j I(Y_{t-1} \in A) + \sum_{j=1}^{J} \gamma_j I(Y_{t-1} \in A)\epsilon_i. \tag{2.14}$$

Si noti come il modello finora introdotto rappresenta una generalizzazione, in chiave asintotica, quindi per $J \to \infty$, con le ϵ_i variabili casuali i.i.d. a media nulla e varianza costante, pari ad uno, $\sigma(y)$ strettamente positiva e Y_0 indipendente dalle $\{\epsilon_i\}$.

Nel presente lavoro supponiamo che la variabile Y_t derivi da un processo stazionario, ipotesi esemplificativa ma non penalizzante poiché, pur non assumendo la stazionarietà del processo iniziale, comunque i risultati riportati in questa sede, sotto le ipotesi considerate, risultano validi asintoticamente per qualunque tipo di processo (Härdle e Tsybakov, 1997).

Il modello viene sintetizzato come segue:

$$Y_t = m(Y_{t-1}) + s(Y_{t-1})\varepsilon_t.$$

Chiamato $m_2(x)$ il momento secondo del processo Y_t condizionato a $Y_{t-1} = x$ e detto $m(x)$ il momento condizionato di Y_t , calcolato in x, entrambi i momenti stimati, di ordine primo e secondo, convergono in probabilità ai momenti veri del processo.

Tecnicamente, sotto le seguenti ipotesi:

- $E(\epsilon_i^2) = 1, E(\epsilon_i) = E(\epsilon_i)^3 = 0, m_{4\epsilon} = E\{(\epsilon_i^2 - 1)^2\} < \infty$
- la densità $p(\cdot)$ di ϵ_1 esiste e verifica che l' $inf_{x \in H} p(x) > 0$ per ogni compatto $H \subset \mathbb{R}^1$
- esistono $C_1 > 0$ e $C_2 > 0$, costanti, tali che $|m(y)| \leq C_1(1 + |y|)$ e $|s(y)| \leq C_2(1 + |y|)$, con $y \in R^1$
- la funzione $s(\cdot)$ è tale che l' $inf_{y \in H} s(y) > 0$ per ogni compatto $H \subset \mathbb{R}^1$
- $C_1 + C_2 E|\epsilon_1| < 1$
- le funzioni m e s sono $(l-1)$ volte derivabili e continue ed esistono le derivate da un lato $m_{\pm}^{(l)}(x)$ e $s_{\pm}^{(l)}(x)$, in $x \in \mathbb{R}^1$
- esiste la densità $\mu(\cdot)$ della distribuzione invariante $\pi(\cdot)$, limitata, continua e strettamente positiva in un intorno di x
- il kernel $K : \mathbb{R}^1 \to \mathbb{R}^+$ è una funzione limitata, su supporto compatto, tale che $K > 0$ su in insieme di misura di Lebesgue positiva
- $h_n = \gamma n^{-1/(2l+1)}$, con $\gamma > 0$
- il valore iniziale Y_0 è un numero fisso in $\mathbb{R}^1$

Härdle e Tsybakov (1997) verificano le seguenti convergenze in probabilità:

$$\hat{m}_2(x) \xrightarrow{P} m_2(x) \ , \tag{2.15}$$

$$\hat{m}(x) \xrightarrow{P} m(x) \ , \tag{2.16}$$

dove $\hat{m}_2(x)$ e $\hat{m}(x)$ rappresentano, rispettivamente, la stima del momento secondo e della media condizionata, sulla base dei dati osservati.

Inoltre vale la convergenza in legge, congiunta, in un punto fisso $x \in \mathbb{R}^1$, alla distribuzione normale, espressa come in Härdle e Tsybakov (1997):

$$n^{l/(2l+1)} \begin{pmatrix} \hat{m}_2(x) - m_2(x) \\ \hat{m}_(x) - m_(x) \end{pmatrix} \xrightarrow{L} N(b(x); \Sigma(x)) \tag{2.17}$$

per $n \to \infty$

$$b(x) = \begin{pmatrix} b_{m_2}(x) \\ b_m(x) \end{pmatrix}, \tag{2.18}$$

$$\Sigma(x) = \frac{\sigma^2(x)}{\beta\mu(x)} \begin{pmatrix} 4m^2(x) + \sigma^2(x)m_{4\epsilon} & 2m(x) \\ 2m(x) & 1 \end{pmatrix} \otimes D. \tag{2.19}$$

La matrice D è il risultato del prodotto matriciale $A^{-1}\Phi A^{-1}$, dove le matrici A e Φ sono definite come segue:

$$A = \int F(u)F(u)'K(u)du, \tag{2.20}$$

$$\Phi = \int F(u)F(u)'K(u)^2du. \tag{2.21}$$

La matrice $b(x)$ è definita mediante le seguenti componenti:

$$b_m(x) = A^{-1}\frac{\gamma^1}{l!}\int F(u)u^lK(u)m^{(l)}(x;u)du, \tag{2.22}$$

$$b_{m_2}(x) = A^{-1}\frac{\gamma^1}{l!}\int F(u)u^lK(u)m_2^{(l)}(x;u)du, \tag{2.23}$$

che corrispondono alle distorsioni asintotiche.

La matrice $F(u)$ è data da

$$F(u) = \begin{pmatrix} 1 \\ u \\ \vdots \\ \frac{u^{l-1}}{(l-1)!} \end{pmatrix}, \tag{2.24}$$

con

$$u_{in} = \frac{Y_{i-1} - x}{h_n}. \tag{2.25}$$

Nella matrice $F(u)$ è stata omessa la dipendenza di u da n ed i, per semplicità notazionale.

2.2.3 Estensioni

Una interessante generalizzazione dei risultati finora presentati viene fornita dallo stesso Masry (1996), che tratta il caso multivariato, nonché il caso in cui la variabile di interesse non sia necessariamente Y, ma anche una sua eventuale trasformazione, $\Psi(Y)$, definita come una qualunque funzione misurabile.

E' da attribuire a Masry e Fan (1997), invece, l'importante estensione, ai dati dipendenti, del risultato prodotto da Tsybakov (1986), ovvero la normalità congiunta asintotica dello stimatore dei polinomi locali, con annessa distribuzione asintotica del vettore $\hat{\beta}(x) = (\hat{\beta}_0(x), \hat{\beta}_1(x), \ldots, \hat{\beta}_p(x))'$, partendo però da ipotesi del tutto diverse rispetto a quelle presentate da Härdle e Tsybakov (1997).

Nel loro lavoro gli autori forniscono la distribuzione asintotica dello stimatore delle derivate della funzione di regressione, $\hat{\beta}_v(x) = \hat{m}^{(v)}(x)/v!$, ovvero delle singole componenti del vettore $\hat{\beta}(x)$, per poi trattare a loro volta, il caso della trasformazione della variabile Y.

Proprio per il differente approccio teorico, il lavoro è da considerarsi una valida alternativa alla proposta di Härdle e Tsybakov (1997), in quanto gli autori riescono a provare gli stessi risultati teorici finora esposti, imponendo però ipotesi del tutto estranee all'assetto teorico precedentemente illustrato.

Il caso multivariato

$\{Y_i, X_i\}, i \in \mathbb{Z}$, nel setup ultizzato da Masry (1996), è un processo congiuntamente stazionario, definito sull'asse dei reali.

Si assume $E|\Psi(Y_1)| < \infty$ e si definisce la funzione multivariata di regressione:

$$m(x_1, \ldots, x_d) = E[\Psi(Y_d)|X_1 = x_1, \ldots, X_d = x_d] \tag{2.26}$$

con $d \geq 1$.

La possibilità di stabilire Ψ arbitrariamente fa sì che i risultati ottenuti siano validi per le diverse definizioni della trasformazione, di cui le principali di maggiore interesse coincidono con il caso in cui $\Psi(Y) = Y$ e la stima ci restituisce la media condizionale di Y_d e delle sue derivate, oppure quando $\Psi(Y) = I\{Y \leq y\}$, che ci fornisce la stima della densità condizionale $m(x) = P[Y \leq y|X_1 = x_1, \ldots, X_d = x_d]$ e delle sue derivate, infine il caso $\Psi(Y) = Y^2$, che corrisponde al momento secondo condizionato.

Il problema viene, dunque, riformulato considerando:

$$\underline{X}_j = (X_{j+1}, \ldots, X_{j+d}) \tag{2.27}$$

ed

$$m(\underline{x}) = E[\Psi(Y_d)|\underline{X}_0 = \underline{x}]. \tag{2.28}$$

Assumendo che esistano finite le derivate di ordine $(p+1)$ di $m(z)$ e che siano continue in x, possiamo approssimare $m(z)$ localmente tramite un polinomio multivariato di ordine p:

$$m(\underline{z}) = \sum_{0 \leq |\underline{k}| \leq p} \frac{1}{\underline{k}!} D^{\underline{k}} m(\underline{y})|_{\underline{y}=\underline{z}} (\underline{z} - \underline{x})^{\underline{k}} \tag{2.29}$$

dove

$$\underline{k} = (k_1, \ldots, k_d) \ , \quad \underline{k}! = (k_1! \times \cdots \times k_d!), \quad |\underline{k}| = \sum_{i=1}^{d} k_i, \tag{2.30}$$

$$\underline{x}^{\underline{k}} = (x_1^{k_1} \times \cdots \times x_d^{k_d}) \tag{2.31}$$

$$\sum_{0 \le |\underline{k}| \le p} = \sum_{j=0}^{p} \sum_{k_1=0}^{j} \cdots \sum_{k_d=0_{k_1+\cdots+k_d=j}}^{j} \tag{2.32}$$

e

$$(D^{\underline{k}} m)(\underline{y}) = \frac{\partial^{\underline{k}} m(\underline{y})}{\partial y_1^{k_1} \ldots \partial y_d^{k_d}}. \tag{2.33}$$

Posto $K(\underline{u})$ la funzione peso su R^d e definito h il parametro di *smoothing*, date le osservazioni $\{Y_i, X_i\}_{i=0}^{n}$, si considerano i minimi quadrati pesati, in termini multivariati:

$$\sum_{i=0}^{n-d} \left[\Psi\left(Y_{d+i}\right) - \sum_{0 \le |\underline{k}| \le p} \beta_{\underline{k}}(\underline{x})(\underline{X}_i - \underline{x})^{\underline{k}} \right]^2 K((\underline{X}_i - \underline{x})/h). \tag{2.34}$$

Minimizzando la (2.34) rispetto a $\beta_{\underline{k}}$, otteniamo $\hat{\beta}_{\underline{k}}(\underline{x})$ e tramite la (2.29) , $\underline{k}!\hat{\beta}_{\underline{k}}(\underline{x})$ ci fornirà una stima di $(D^{\underline{k}} m)(x)$, in questo modo, analogamente al caso univariato, otteniamo la stima delle derivate della $m(\underline{x})$, nonchè della funzione di regressione stessa, con:

$$\widehat{(D^{\underline{k}} m)}(x) = \underline{k}!\hat{\beta}_{\underline{k}}(\underline{x}). \tag{2.35}$$

Sotto le condizioni tecniche specificate da Masry (1996) vale ancora la convergenza in probabilità ed in distribuzione dello stimatore appena presentato, nonché per ognuna delle derivate stimate, fino all'ordine p.

La stima delle derivate della funzione di regressione

La normalità asintotica congiunta dello stimatore delle derivate della funzione di regressione, $m(x)$, ovvero delle sue derivate di ordine v, $m^{(v)}(x)$, con $v = 0, \ldots, p$, viene provata da Masry e Fan (1997), date le seguenti assunzioni:

- K limitato nel supporto [-1,1]
- $f(u, v; l) \leq M_1$ e $E\{Y_1^2 + Y_l^2 | X_1 = u, X_l = v\} \leq M_2, l \geq 1$, per u e v in un intorno di x, dove $f(u, v; l)$ rappresenta la densità congiunta di X_0 e X_1
- per processi *strongly mixing* assumiamo che $\delta > 2$ e $a > 1 - 2/\delta$, $\sum l^a [\alpha(l)]^{1-2/\delta} < \infty$, $E\{|Y_0|^\delta | X = u\} \leq M_3 < \infty$ per u in un intorno di x .

Gli autori provano la convergenza in probabilità dello stimatore della media condizionale del processo alla vera funzione di regressione, quindi:

$$\hat{m}^{(v)}(x) \xrightarrow{P} m^{(v)}(x) \tag{2.36}$$

e

$$\hat{m}_2^{(v)}(x) \xrightarrow{P} m_2^{(v)}(x) \tag{2.37}$$

Inoltre, sotto le ipotesi classiche di $h \to 0$, $nh \to \infty$, per i processi di tipo *strongly mixing* si ipotizza che esista una sequenza di interi positivi che soddisfano la condizione $s_n \to \infty$ e $s_n = o((nh)^{1/2})$ tale che $(nh)^{1/2} \alpha(s_n) \to 0$, per $n \to \infty$.

Considerata la densità condizionale $G(y|x)$ di Y dato $X = x$ continua nel punto x, vale inoltre, per $n \to \infty$, la normalità asintotica congiunta dello stimatore $\hat{\beta}_v(x) = \hat{m}^{(v)}(x)/v!$.

La convergenza in legge ad una normale multivariata viene fornita non soltanto per le singole componenti di $\hat{\beta}_v(x)$, ovvero le stime delle derivate della funzione di regressione, ma anche per le trasformazioni di Y, $\Psi(Y)$, purché per entrambe le distribuzioni asintotiche h verifichi la condizione di convergenza $h_n = O(n^{-1/(2p+3)})$.

2.2.4 Considerazioni sui vari approcci presentati

La principale differenza tra gli approcci appena considerati sta nel fatto che le condizioni imposte da Härdle e Tsybakov (1997) trattano la variabile Y come parte di un processo Markoviano e sfruttano le proprietà di tali modelli per derivare la normalità asintotica dello stimatore utilizzato.

Le prime cinque ipotesi dichiarate nel lavoro, infatti, garantiscono l'ergodicità geometrica del processo Y_i, il quale viene studiato in qualità di catena di Markov.

Nel secondo caso, invece, Masry e Fan (1997) trattano il problema analiticamente, secondo una impostazione più propriamente tecnica, basata cioé su condizioni imposte alle funzioni piuttosto che al processo.

Essi prescindono perciò dall'utilizzo delle proprietà Markoviane.

Tutte le assunzioni, infatti, riguardano la funzione kernel, la densità di X, la distribuzione condizionale di Y posto X e le proprietà dei processi *mixing*.

Questo secondo approccio, insieme a quello multivariato di Masry (1996), può risultare particolarmente utile nel caso in cui l'attenzione sia posta sulla stima, oltre che della media condizionata della variabile Y, di una trasformazione misurabile della variabile Y o sulle derivate della funzione di regressione stessa.

2.3 Determinazione del parametro di *bandwidth*

Durante tutta la trattazione l'utilizzo delle funzioni kernel ha implicitamente posto una problematica legata alla determinazione del parametro che definisce l'ampiezza della finestra ottimale per effettuare lo *smoothing*. Il parametro cioè che defisce l'ampiezza dell'intorno di dati da considerare per effettuare la stima puntuale.

In particolare, è noto in letteratura che stabilire correttamente l'ampiezza della finestra di *smoothing* comporta differenze sostanziali, in termini di risultati molto più evidenti rispetto alla scelta della stessa funzione kernel. Sottostimare o sovrastimare l'h, conduce non soltanto a stime troppo spigolose o al contrario troppo lisciate, ma ha anche importanti ripercussioni sull'entità della distorsione e della varianza dello stimatore.

All'aumentare del parametro di *bandwidth*, h, si ottiene una riduzione della varianza dello stimatore, data dall'inclusione nell'intervallo di un maggior numero di punti , viceversa, però, l'incremento dell'intervallo comporta una maggiore distanza media tra i punti che vengono coinvolti nel calcolo ed il punto in cui la stima viene prodotta. Questo si traduce in un aumento della distorsione dello stimatore, perché con una maggiore distanza si ottiene una minore precisione, in corrispondenza di una maggiore variabilità.

L'intervento di h nella determinazione della varianza e nella distorsione dello stimatore, in maniera inversamente proporzionale nel primo caso e direttamente proporzionale nel secondo, pone il problema di identificare un valore per questo parametro in grado di bilanciare i due fenomeni opposti. Si tratta di riuscire a dare una soluzione al trade-off *bias-variance*, noto in letteratura.

Il più popolare metodo di risoluzione a questo problema prevede la minimizzazione del MSE.

Ha senso minimizzare l'errore quadratico medio, poiché, in quanto misura della qualità della stima della funzione di regressione, minimizzare questa quantità significa riuscire a garantire la migliore stima possibile.

Altro problema che si incontra nel caso in cui si voglia determinare l'ampiezza della finestra di *smoothing* è se scegliere un unico h, globale, oppure effettuare una scelta di h variabile. Nel primo caso si integra l'MSE per ottenere una misura dell'errore al quale fare riferimento che sia la stessa su tutto lo spazio parametrico. La minimizzazione del MISE, Mean Integrated Squares Error, è una tecnica popolare e condivisa per derivare l'h globale ottimale (Ruppert et al., 1995; Xia e Li, 2002; Fan e Gijbels, 1992).

L'ottimizzazione può essere effettuata empiricamente, tramite la *Cross-Validation*, oppure tramite l'utilizzo delle espressioni asintotiche note per la distorsione e la varianza dello stimatore. In quest'ultimo caso si ottiene una formulazione per l'MSE, in cui sono coinvolti, però, termini non noti, in particolare $m_2(x)$, $\sigma(x)$ e $f(x)$, che vengono stimati secondo un approccio detto di tipo *plug-in*. Le stime ottenute, cioé, concorrono alla determinazione del valore del MSE da minimizzare.

2.3.1 h plug-in per la stima delle derivate della funzione di regressione

Esaminiamo ora, più nello specifico come viene determinato l'h ottimale, tramite la procedura *plug-in*, per i due stimatori $m(x)$ ed $m_2(x)$, a partire dal modello per dati dipendenti, rispettivamente prima in presenza di omoschedasticità per la media condizionata e poi di eteroschedasticità per il momento secondo condizionato.

h ottimale *locale* e *globale* per modelli di tipo omoschedastici

Supponiamo che il modello in esame sia del tipo:

$$Y_t = m(Y_{t-1}) + \varepsilon_t. \tag{2.38}$$

Un simile modello ben si presta a stimare $\hat{m}(x)$. Scompare rispetto alla (2.2) la componente legata alla varianza condizionata, data l'ipotesi di omoschedasticità, σ_ε^2 è costante e finita per tutte le osservazioni.

Sotto le ipotesi, precedentemente illustrate, di Masry e Fan (1997), con $h = O(n^{1/(2p+3)})$, per il teorema 5, la distribuzione asintotica dello stimatore $\hat{m}^{(v)}(x)$ è data da:

$$\sqrt{nh_n^{2v+1}}\left(\hat{m}^{(v)}(x) - m^{(v)}(x) - \frac{m^{(p+1)}(x)v!B_v}{(p+1)!}h_n^{p+1-v}\right) \xrightarrow{L} N\left(0, \frac{(v!)^2 V_v \sigma_\varepsilon^2}{\mu(x)}\right) \tag{2.39}$$

per tutti i punti di continuità di $\{\mu\}$, se $\mu(x) > 0$, con B_v e V_v rispettivamente dati dal $v-esimo$ elemento delle matrici $S^{-1}\mu$ ed il $v-esimo$ elemento di $S^{-1}\tilde{S}S^{-1}$, μ definita come nelle assunzioni di Härdle e Tsybakov (1997), riportate precedentemente.

Le matrici μ, S ed $\tilde{S}$ sono definite come segue:

$$\mu_j = \int_{-\infty}^{+\infty} u^j K(u)du, \quad v_j = \int_{-\infty}^{+\infty} u^j K^2(u)du,$$

$$S = \begin{pmatrix} \mu_0 & \dots & \mu_p \\ \vdots & \ddots & \vdots \\ \mu_p & \dots & \mu_{2p} \end{pmatrix}, \quad \tilde{S} = \begin{pmatrix} v_0 & \dots & v_p \\ \vdots & \ddots & \vdots \\ v_p & \dots & v_{2p} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{pmatrix}.$$

La distorsione e la varianza dello stimatore diventano, conseguentemente:

$$distorsione\ di\ \hat{m}^{(v)}(x) = \frac{m^{(p+1)}(x)v!B_v}{(p+1)!}h_n^{p+1-v} \tag{2.40}$$

$$varianza\ di\ \hat{m}^{(v)}(x) = \frac{(v!)^2 V_v \sigma_\varepsilon^2}{nh^{2v+1}\mu(x)}. \tag{2.41}$$

Minimizzare l'MSE corrispondente alle espressioni asintotiche appena riportate, significa minimizzare, secondo la decomposizone *distorsione varianza*, la somma del quadrato della distorsione, più la varianza di $\hat{m}^{(v)}(x)$.

La finestra ottimale per stimare la $v-esima$ derivata, supponendo che $p-v$ sia dispari, Fan e Gijbels (1995), viene data dalla seguente espressione:

$$h_{v,opt}(x) = \left(\frac{[(p+1)!]^2 V_v \sigma_\varepsilon^2 / \mu(x)}{2(p+1-v)[m^{(p+1)}(x)]^2 B_v^2} \right)^{1/(2p+3)} \frac{1}{n^{1/(2p+3)}} \tag{2.42}$$

in cui, come precedentemente accennato, $m^{(p+1)}$, f e σ_ε^2 risultano non note, dunque, occorre per esse una stima, mentre è possibile calcolare V_v e B_v, a seconda del grado p e del kernel utilizzato.

La dipendenza dei tali valori dalla scelta del kernel e del grado del polinomio non viene indicata nella simbologia per pura semplicità notazionale.

Come già specificato, oltre all'h puntuale è possibile ottenere un valore globale per il parametro di *bandwidth* scegliendo di minimizzare invece dell' MSE, il suo valore integrato, ovvero il MISE.

Questo, per la decomposizione *distorsione-varianza*, conduce alla minimizzazione della somma della varianza integrata $Ivar[\hat{m}^{(v)}]$ ed il quadrato della distorsione integrata, $Ibias^2[(\hat{m}^{(v)})]$.

Dato il modello (4.1), avremo:

$$Ibias^2[(\hat{m}^{(v)})] = \left[\frac{v! B_v}{(p+1)!} \right]^2 h_n^{2(p+1-v)} \int_\chi [m^{(p+1)}(x)]^2 \mu(x) dx \tag{2.43}$$

ed

$$Ivar[\hat{m}^{(v)}] = \frac{v! V_v}{n h^{2v+1}} \sigma_\varepsilon^2 \lambda(\chi), \tag{2.44}$$

con $\lambda(\chi)$ misura di Lebesgue riferita al compatto $\chi \subset \mathbb{R}^1$ su cui è definito l'integrale nella (2.43).

Definito $R = \int_\chi [m^{(p+1)}(x)]^2 \mu(x) dx$, l'$h$ globale, che chiameremo h_G, sarà dato da:

$$h_G = \left(\frac{[(p+1)!]^2 V_v \sigma_\varepsilon^2 \lambda(\chi)}{2(p+1-v) R B_v^2} \right)^{(\frac{1}{2p+3})} \left(\frac{1}{n} \right)^{(\frac{1}{2p+3})} \tag{2.45}$$

h ottimale *locale* e *globale* per modelli di tipo eteroschedastici

Esaminiamo ora il caso in cui il modello in esame sia espresso da:

$$Y_t = s(Y_{t-1})\varepsilon_t. \tag{2.46}$$

E' questo il caso di un modello caratterizzato dalla presenza di eteroschedasticità e l'attenzione è rivolta alla stima del momento secondo condizionato.

Definiti:

$$s^2(x) = E(Y_t^2|Y_{t-1} = x)$$

e

$$s^4(x)m_{4\varepsilon} = var(Y_t^2|Y_{t-1} = x),$$

con

$$m_{4\varepsilon} = E\{(\varepsilon_i^2 - 1)^2\} < \infty,$$

otterremo, secondo il teorema 5 di Masry e Fan (1997), la seguente distribuzione asintotica per lo stimatore:

$$\sqrt{nh_n^{2v+1}}\left(\hat{m}_2^{(v)}(x) - m_2^{(v)}(x) - \frac{[s^2(x)]^{(p+1)}v!B_v}{(p+1)!}h_n^{p+1-v}\right) \xrightarrow{L} N\left(0, \frac{(v!)^2V_v s^4(x)m_{4\varepsilon}}{\mu(x)}\right). \tag{2.47}$$

In questo caso, conseguentemente:

$$distorsione\ di\ \hat{m}_2^{(v)}(x) = \frac{[s^2(x)]^{(p+1)}v!B_v}{(p+1)!}h_n^{p+1-v} \tag{2.48}$$

$$varianza\ di\ \hat{m}_2^{(v)}(x) = \frac{(v!)^2V_v s^4(x)m_{4\varepsilon}}{nh_n^{2v+1}\mu(x)}. \tag{2.49}$$

Procedendo per minimizzazione della somma della varianza di $\hat{m}_2^{(v)}(x)$ ed il quadrato della distorsione di $\hat{m}_2^{(v)}(x)$, secondo la decomposizione *distorsione varianza*, l'h ottimale sarà dato da:

$$h_{2v,opt}(x) = \left(\frac{[(p+1)!]^2V_v s^4(x)m_{4\varepsilon}/f(x)}{2(p+1-v)\{[s^2(x)]^{(p+1)}\}^2B_v^2}\right)^{1/(2p+3)}\frac{1}{n^{1/(2p+3)}}. \tag{2.50}$$

Anche in questo caso è possibile procedere per integrazione per ottenere una misura dell' MSE globale ed un conseguente h, ugualmente valido per l'intero spazio χ del processo stocastico $\{Y_t\}$.

Considerato il compatto $\chi \subset \mathbb{R}^1$, otterremo:

$$Ibias^2[(\hat{m}_2^{(v)})] = \left[\frac{v!B_v}{(p+1)!}\right]^2 h_n^{2(p+1-v)} \int_\chi \{[s^2(x)]^{(p+1)}\}^2 \mu(x)dx \tag{2.51}$$

ed

$$Ivar[\hat{m}_2^{(v)}] = \frac{(v!)^2 V_v m_{4\varepsilon}}{nh^{2v+1}} \int_\chi s^4(x)dx, \tag{2.52}$$

Definito poi $R_2 = \int_\chi \{[s^2(x)]^{(p+1)}\}^2 \mu(x)dx$, e $R_3 = \int_\chi s^4(x)dx$, la misura dell'h globale sarà data da:

$$h_{2G} = \left(\frac{[(p+1)!]^2 V_v R_3 m_{4\varepsilon}}{2(p+1-v)R_2 B_v^2}\right)^{(\frac{1}{2p+3})} \left(\frac{1}{n}\right)^{(\frac{1}{2p+3})}. \tag{2.53}$$

2.3.2 Il caso del kernel di Epanechnikov

I valori di B_v e V_v per le espressioni riportate in questa sede possono essere calcolati deterministicamente.

In via del tutto esemplificativa, nell'ipotesi di impiego del kernel di Epanechnikov:

$$K(u) = 0.75(1-u^2)\mathbb{I}\{|u| \leq 1\}. \tag{2.54}$$

con $v = 0$ e $p = 1$, avremo dei valori per B_v pari a1/5 mentre V_v risulterà pari a 3/5.

Capitolo 3

Empirical Likelihood nel caso di dati dipendenti

3.1 Introduzione

Il presente capitolo ha il compito non soltanto di esibire i principali risultati che caratterizzano la funzione di verosimiglianza empirica, con particolare attenzione al caso di dati dipendenti, ma anche di proporre un'evoluzione, dalla presentazione della verosimiglianza empirica alla messa a punto di stimatori che comportano l'impiego della verosimiglianza empirica, congiunta agli stimatori polinomi locali, appena trattati nel capitolo precedente. I primi paragrafi introduttivi illustrano la verosimiglianza empirica cercando di metterne in luce principalmente le peculiarità e le caratteristiche, nonché la sua importanza nel caso non parametrico, proponendo un parallelismo con importanti proprietà che la verosimiglianza empirica riesce a conservare rispetto al caso parametrico. Viene accuratamente analizzato anche il contesto di applicazione di queste tecniche ai dati fino a giungere a valutazioni che verrano poi ampliate nel capitolo successivo, tramite risultati originali, rispetto alla letteratura.

3.2 La verosimiglianza empirica: come nasce e principali caratteristiche

La verosimiglianza empirica viene introdotta da Owen (1988,1990), con due lavori nei quali questa tecnica viene impiegata per costruire intervalli di confidenza. La più importante intuizione di Owen è stata quella di mostrare la facile adattabilità della metodologia a tutta una serie di problematiche, anche diverse da quelle specifiche per cui la verosimiglianza empirica nasceva.

Owen (1990) presenta la verosimiglianza empirirca come un'alternativa ai metodi bootstrap di tipo *likelihood*, come quello proposto da Hall (1987).

I vantaggi derivanti dall'utilizzo della nuova metodologia si sostanziano nel non dover imporre una forma predeterminata all'intervallo di confidenza, in quanto la forma rifletterà in maniera naturale i dati impiegati, donando più peso alle informazioni dov'è maggiore la densità del parametro da stimare.

Inoltre, per l'empirical likelihood è possibile applicare la correzione di Bartlett, non necessita di parametri di scala o di asimmetria, nè tantomento di una statistica pivotale. Le regioni rispettano le trasformazioni e preservano il *range* dei dati. A differenza degli intervalli di confidenza costruiti con le tecniche bootstrap, la verosimiglianza empirica non necessita di simulazioni Monte Carlo per la costruzione delle regioni di confidenza, ma soltanto della risoluzione, tramite metodi matematici, di un problema di ottimizzazione vincolata. Questo chiaramente comporta che per problematiche più complesse il bootstrap diventa più facilmente applicabile mentre l'applicazione della verosimiglianza empirica viene ostacolata da un'imponente complessità computazionale.

Owen parte dalla definizione di funzione di distribuzione empirica quale stima di massima verosimiglianza della distribuzione dalla quale il campione è estratto, per poi analizzare come la funzione di verosimiglianza per le distribuzioni venga utilizzata per la costruzione del rapporto di massima verosimiglianza. Owen (1988) deriva una estensione al caso non parametrico del teorema di Wilks (1938) noto per i rapporti di verosimiglianza di tipo parametrico.

Partendo da osservazioni indipendenti $X_1, \ldots, X_n$ tratte da una funzione di distribuzione F_0, la funzione di distribuzione empirica F_n è spesso considerata come la stima non parametrica di massima verosimiglianza di F_0, perché massimizza

$$L(F) = \prod_{i=1}^{n} \{F(X_i) - F(X_i-)\}$$

rispetto a tutte le possibili funzioni di distribuzione F. Il rapporto di verosimiglianza empirica viene definito, quindi, come:

$$R(F) = L(F)/L(F_n).$$

Supponiamo che l'interesse sia su $T(F_0)$, con $T(\cdot)$ funzionale statistico. La stima non parametrica di massima verosimiglianza di $T(F_0)$ é data da $T(F_n)$.

Owen (1988) dimostra che intervalli del tipo

$$\{T(F)|R(F) \geq c\} \tag{3.1}$$

possono essere utilizzati come intervalli di confidenza per $T(F_0)$.

Gli statistici utilizzano il rapporto di verosimiglianza nel caso parametrico per costruire intervalli di confidenza e valutare test. In alcuni casi però la presenza di parametri di disturbo rende difficile l'utilizzo di tale rapporto. L'estensione del teorema di Wilks al caso non parametrico consente la convergenza asintotica di $-2logR_0$ ad una χ^2_q con R_0 risultato della massimizzazione del rapporto di massima verosimiglianza e p numero di restrizioni imposte.

Non sempre comunque intervalli della forma (3.15) funzionano. Ad esempio, questi falliscono miseramente nel caso in cui F_0 è assolutamente continua e $T(F)$ è rappresentato dai punti in cui la funzione F salta.

Una restrizione naturale è data dal considerare le distribuzioni con un supporto limitato tra $[-M, M]$ per un valore di M positivo. Quindi è possibile restringere il campo alle distribuzioni che hanno un supporto compreso nel campione, dunque, alle distribuzioni per cui vale $F \ll F_n$. In questo modo il vantaggio per gli statistici si sostanzia nel non dover stabilire un valore M, inoltre il problema si riduce al caso di dimensione finita.

Sotto le ipotesi di funzione di distribuzione non degenere F_0, per $X_1, \ldots, X_n$, variabili casuali indipendenti, con $\int |x|^3 dF_0 < \infty$ e c>1 definiti

$$\mathscr{F}_{c,n} = \{F | R(F) \geq c, F \ll F_n\},$$

$$X_{U,n} = sup \int xdF \quad , \quad X_{L,n} = inf \int xdF,$$

per ogni $F \in \mathscr{F}_{c,n}$, per $n \to \infty$, vale:

$$P\{X_{L,n} \leq E(X) \leq X_{U,n}\} \to P(\chi^2_{(1)} \leq -2logc).$$

Il lavoro , qualche anno dopo, dello stesso Owen (1990), rappresenta una generalizzazione al caso multivariato di questo stesso risultato, considerando stimatori più complessi.

Partendo da $X, X_1, X_2, \ldots$ vettori di variabili casuali indipendenti ed identicamente distribuite in $\mathbb{R}^p$, con $E(X) = \mu_0$ e $var(X) = \sum$ di rango $q > 0$, $r < 1$ positivo, si definisce $C_{r,n} = \{\int XdF | F \ll F_n, R(F) \geq r\}$, dove $C_{r,n}$ è un insieme convesso e vale:

$$lim_{n\to\infty} P(\mu_0 \in C_{r,n}) = P(\chi^2_q \leq -2logr). \tag{3.2}$$

Inoltre, se $E(||X||^4) < \infty$:

$$|P(\mu \in C_{r,n}) - P(\chi^2_q \leq -2logr)| = O(n^{-1/2}), \tag{3.3}$$

con $||\cdot||$ indichiamo la distanza euclidea.

La verosimiglianza empirica può essere utilizzata in tutti i casi in cui il nostro parametro può essere espresso come funzione *smooth* di momenti dei dati. Si suppone, cioè, che il parametro di interesse sia $\theta = g\{E(X)\}$, con $E(X)$ vettore della media della popolazione, da cui viene estratto il campione s - variato $X_1, X_2, ..., X_n$ e $g : \mathbb{R}^s \to \mathbb{R}^r$ è la funzione *smooth*.

Avere un parametro espresso in funzione delle medie dei dati ci consente, come abbiamo già visto nel caso dei polinomi locali, di poter indagare non soltanto sui momenti della popolazione ma anche su quantità suscettibili di essere espresse come funzione dei momenti stessi. Questo avviene nel caso della varianza, dov'è possibile operare semplicemente tramite differenza di momenti. Supponiamo di avere, cioé, il set di dati in forma scalare, $Y_1, Y_2, ...Y_n$, possiamo riaccorparli come $X_i = (Y_i, Y_i^2)'$ in modo tale che la X abbia parametri r=1, s=2, e ottenere:

$$\theta = g\{E(X)\} = E(X^{(2)}) - [E(X^{(1)})]^2 = E(Y^2) - [E(Y)]^2; \tag{3.4}$$

dove $X^{(i)}$ rappresenta l' i-esimo componente del vettore bivariato X.

In questo caso, chiaramente, abbiamo operato la trasformazione $g : \mathbb{R}^2 \to \mathbb{R}^1$ dove $g(u, v) = u - v^2$.

La versatilità della verosimiglianza empirica fa sì che essa si presti facilmente ad essere impiegata anche in altri contesti, come per il calcolo dei quantili della distribuzione, per le statistiche U e i relativi quantili.

3.2.1 La verosimiglianza empirica in letteratura

E' lo stesso Owen (1990) ad evidenziare come il fatto che F_n fosse considerato lo stimatore di massima verosimiglianza di F_0 era già noto come dimostra il lavoro di Kiefer e Wolfowitz (1956).

Gli autori avevano effettuato uno studio di consistenza dello stimatore in presenza di varie impostazioni di effetti casuali. Kaplan e Meier (1958), invece, forniscono una derivazione dello stimatore limite del prodotto della funzione di sopravvivenza come uno stimatore di massima verosimiglianza non parametrico e Johansen (1978) mostra che lo stimatore limite del prodotto è uno stimatore di massimaverosimiglianza secondo la logica di Kiefer e Wolfowitz.

Successivamente Bailey (1984) dimostra che, in assenza di tempo di fallimento delle code e di covariate dipendenti dal tempo, la massimaverosimiglianza applicata a modello di Cox restituisce la comune stima della derivata di verosimiglianza per il parametro di regressione β. Vardi (1985) ha, poi, utilizzato la massima verosimiglianza non parametrica per stimare le funzioni di distribuzione in presenza di selezione della distorsione.

Il primo caso di utilizzo del rapporto di verosimiglianza empirica per la costruzione di intervalli di confidenza, invece, è da attribuire a Thomas e Grunkemeier (1975). Gli autori esibivano un'applicazione al caso delle probabilità di sopravvivenza stimate tramite la curva di Kaplan e Meier e dimostravano empiricamente che gli intervalli basati sulla verosimiglianza empirica, per le probabilità di sopravvivenza, basati sulla distribuzione $\chi^2_{(1)}$ godono di un errore di copertura dell'intervallo corretto.

A differenza degli intervalli basati sulla formula di Greenwood, gli intervalli di Thomas e Grunkemeier godono della possibilità di essere asimmetrici e non includono mai valori all'esterno dell'intervallo $[0, 1]$, caratteristica particolarmente allettante per le probabilità di sopravvivenza definite nello stesso intervallo. Successivamente, anche Cox e Oakes (1984) ottengono risultati per lo stesso intervallo, in via del tutto indipendente dal lavoro appena citato.

Una versione univariata della (3.2) e (3.3) compare per la prima volta in un lavoro di Owen (1985), oltre che, qualche anno dopo, nel lavoro di Owen (1988), di cui abbiamo già riportato i risultati. DiCiccio, Hall e Romano (1991) hanno poi dimostrato che l'errore nella (3.3) è da considerare di ordine $O(n^{-1})$ se vengono effettuate le ipotesi che consentono l'espansione di Edgeworth e che la correzione di Bartlett riduce invece l'errore a $O = (n^2)$. Permane, tuttavia, l'ordine di convergenza pari a $O = (n^{1/2})$ per i casi di problemi di tipo unilaterale.

Interessanti sono, infine, i risultati che riguardano la verosimiglianza multinomiale. Hoeffding (1965) dimostra che, per le distribuzioni multinomiali, i testi basati sul rapporto di verosimiglianza sono asintoticamente ottimali secondo la logica di Bahadur.

Tusnady (1977) estende i risultati di Hoeffding considerando, una sequenza di famiglie multinomiali, usando partizioni finite dello spazio campionario e dimostrando, anche in questo caso, che i test basati su queste famiglie di distribuzioni sono asintoticamente ottimali secondo la logica di Bahadur, sotto condizioni di regolarità imposte alla sequenza di partizioni.

Occorre indicare, però, che Tusnady omette di precisare come scegliere il set ottimale di partizioni da effettuare.

Berk e Jones (1979) impiegano la verosimiglianza discreta per testare se è possibile decretare che un campione derivi dalla distribuzione uniforme. La loro legge asintotica non corrisponde ad una χ^2, ma è legata invece alla teoria del valore estremo.

E' degno di nota il fatto che sia Tusnady (1977) che Berk e Jones (1979) hanno tratto i loro risultati avvalendosi della distanza di Kullback-Leibler tra la misura empirica e l'insieme di misure ipotizzate. I lavori che vengono qui presentati rappresentano un excursus logico di anticipazione e preparazione al risultato che verrà stilato nei paragrafi successivi per quest'elaborato.

3.3 Come si definisce la Verosimiglianza Empirica

Supponiamo di avere un campione $X_1, X_2, ..., X_n$, estratto da una popolazione e $p = (p_1, p_2, ..., p_n)$ sia un vettore per cui vale che ogni componente $p_i \geq 0$ e $\Sigma_i p_i = 1$. Chiamiamo $\theta(p)$ il valore assunto dal parametro, quando la popolazione è discreta, con pesi p_i per le rispettive X_i, con $1 \leq i \leq n$.

La verosimiglianza empirica per $\theta \in \mathbb{R}^1$, calcolata in $\theta = \theta_1$, si definisce come:

$$L(\theta_1) = \max_{p:\theta(p)=\theta_1, \sum p_i=1} \prod_{i=1}^{n} p_i, \tag{3.5}$$

dunque, come già accennato, la verosimiglianza empirica può essere considerata la verosimiglianza data da una distribuzione multinomiale, il cui numero di parametri è pari alla numerosità campionaria meno uno.

La caratteristica più interessante di questo stimatore è che riesce a conservare buona parte delle proprietà di cui gode la verosimiglianza nel caso parametrico.

Sotto il vincolo che $\sum p_i = 1$, il prodotto $\prod_{i=1}^{n} p_i$ è massimizzato ponendo $p_i = n^{-1}$ per $1 \leq i \leq n$, tramite l'utilizzo dei moltiplicatori di Lagrange. Per un tale valore di p, $\theta(p) = \hat{\theta}$ e lo stimatore è detto stimatore *bootstrap*. Lo stimatore di *massima verosimiglianza empirica*, dunque, non è altro che un caso particolare di stimatore bootstrap (Hall, La Scala, 1990).

Quando, ad esempio, θ rappresenta la media della popolazione, $\theta = n^{-1} \sum X_i$, ovvero la media campionaria:

$$L(\hat{\theta}) = n^{-n}$$

dunque

$$L(\theta_1)/L(\hat{\theta}) = \max_{p:\theta(p)=\theta_1, \sum p_i=1} \prod_{i=1}^{n} np_i. \tag{3.6}$$

Questa prima definizione della verosimiglianza empirica ci tornerà utile nel mostrare come proprietà valide per il caso parametrico vengano conservate, in buona parte, nel momento in cui andremo a trattare il caso non parametrico di questo stesso stimatore.

3.3.1 Il teorema di Wilks nel caso non parametrico

Una delle più importanti proprietà che la verosimiglianza empirica assorbe dal caso parametrico è l'estensione del teorema di *Wilks*.

Per capire meglio lo spessore di una simile estensione presentiamo il risultato mediante un parallelismo con il caso parametrico. Se indichiamo con $L(\theta)$ la classica verosimiglianza campionaria e $\hat{\theta}$ lo stimatore di massima verosimiglianza, il rapporto di logverosimiglianza sarà dato da:

$$l(\theta) = -2log\{L(\theta)/L(\hat{\theta})\}. \tag{3.7}$$

Supponiamo ora che θ_0 sia il vero valore del parametro da stimare, θ, che non ci siano parametri di disturbo e che t sia il rango della matrice di varianza asintotica di $n^{1/2}\hat{\theta}$.

Il teorema di Wilks afferma che, sotto appropriate condizioni di regolarità, $l(\theta_0)$ segue la distribuzione asintotica χ^2_q.

Conoscere la distribuzione asintotica di $l(\theta_0)$, diventa il primo passo per costruire intervalli di confidenza, basati sulla verosimiglianza parametrica.

Si può, a questo punto, procedere tramite le tavole della variabile χ^2_q, per calcolare la:

$$P(\chi^2_q \leq c) = 1 - \alpha, \tag{3.8}$$

dove $1 - \alpha$ rappresenta il livello di copertura desiderato per l'intervallo.

Quindi, si stabilisce l'intervallo, secondo la regola:

$$\mathscr{R}_c = \{\theta : l(\theta) \leq c\} \tag{3.9}$$

e la copertura asintotica di $\mathscr{R}_c$ pari a $1 - \alpha$, sarà garantita da:

$$P(\theta_0 \in \mathscr{R}_c) = P\{l(\theta_0) \leq c\} \to 1 - \alpha \tag{3.10}$$

per $n \to \infty$, vedi Wilks (1938) e Chernoff (1954).

L'estensione al caso non parametrico avviene definendo dapprima il rapporto di verosimiglinza, l e c, rispettivamente secondo la (3.13), (3.7) e la (3.8), in seguito costruendo la regione di confidenza, grazie alla (3.9). A questo punto è garantita la validità della (3.10), per mezzo della versione del teorema di Wilks valido per la verosimiglianza empirica. E' possibile infatti generalizzare il risultato di Owen (1990), stabilendo la validità del teorema di Wilks per la verosimiglianza empirica, se θ è rappresentato da una funzione della media della popolazione. Lo stimatore bootstrap in questo caso corrisponderà alla stessa funzione applicata alla media campionaria. Più chiaramente, se $\theta = g(\mu)$, lo stimatore sarà dato da $\hat{\theta} = g(\bar{X})$, dove abbiamo indicato con $\bar{X}$ la media campionaria, definita da $n^{-1}\sum X_i$. Supponiamo che θ sia un vettore di lunghezza r, mentre X_i e $\mu = (\mu^{(1)}, \ldots, \mu^{(s)})'$ siano di lunghezza pari ad s.

Indichiamo con $\mu_0 = E(X)$ il vero valore della media della popolazione e $\theta_0 = g(\mu_0)$ il vero valore di θ. Se $g = (g^{(1)}, \ldots, g^{(r)})'$ ha derivata continua in un intorno di μ_0, allora la matrice della varianza asintotica di $n^{1/2}\hat{\theta}$ diventa pari a $V = v_0 \sum v_0'$, dove $v_0 = (v_0^{(ij)})$ rappresenta la matrice di ordine $r \times s$ definita tramite:

$$v_0^{(ij)} = \partial g^{(i)}(\mu)/\partial \mu^{(j)}|_{\mu=\mu_0}$$

e la matrice della varianza della popolazione sarà $\sum = E\{(X-\mu_0)(X-\mu_0)'\}$.

Se X ha varianza finita e g derivata continua in un intorno di μ_0, posto $t \leq min(r,s)$ il rango di V ed l la logverosimiglianza empirica, $l(\theta_0)$ avrà una distribuzione asintotica χ^2_q.

3.3.2 La verosimiglianza empirica per la media della popolazione

Il caso in cui θ è pari a μ, ossia la media della popolazione, è di particolare interesse sotto vari aspetti. Come già riportato, prima di tutto la verosimiglianza empirica in questo caso assume una formulazione semplice, inoltre le regioni di verosimiglianza per la media, sia nel caso univariato che multivariato, sono sempre convesse, infine, moltissime statistiche oggetto di interesse sono funzioni della media.

Tramite i moltiplicatori di Lagrange è possibile dimostrare che i valori delle p_i che massimizzano $\prod p_i$ sotto il vincolo di $\sum p_i X_i = \mu$ e $\sum p_i = 1$ sono date da:

$$p_i(\mu) = n^{-1}\{1 + \lambda'(X_i - \mu)\}^{-1} \tag{3.11}$$

dove il vettore $\lambda = \lambda(\mu)$ di lunghezza s è calcolato tramite la risoluzione della seguente equazione:

$$\sum_{i=1}^{n}\{1 + \lambda'(X_i - \mu)\}^{-1}(X - \mu) = 0. \tag{3.12}$$

Quindi il rapporto di verosimiglianza empirica per la media è dato da:

$$l(\mu) = -2\sum_{i=1}^{n} log\{np_i(\mu)\} = 2\sum_{i=1}^{n} log\{1 + \lambda'(X_i - \mu)\}. \tag{3.13}$$

Per comprendere come è possibile garantire che gli intervalli di verosimiglianza empirica per la media siano convessi, basti notare che le verosimiglianze multinomiali sono funzioni di distribuzioni concave.

Dunque, se $p = (p_1, \ldots, p_n)$ e $q = (q_1, \ldots, q_n)$ sono distribuzioni di probabilità che soddisfano:

$$\prod_{i=1}^{n} p_i \geq C, \prod_{i=1}^{n} q_i \geq C,$$

per qualche $C > 0$, allora:

$$\prod_{i=1}^{n}\{\beta p_i + (1-\beta)q_i\} \geq C,$$

per ogni $0 \leq \beta \leq 1$. Consideriamo, $c > 0$, l'indice della regione di massima verosimiglianza di $\mathscr{R}_c$ definita nella (3.9) e fissiamo $C = n^{-n}e^{-c/2}$.

La regione $\mathscr{R}_c$ diventa:

$$\mathscr{R}_c = \{\mu : l(\mu) \leq c\} = \{\mu : L(\mu) \leq C\} = \left\{\mu : \max_{p:\sum p_i X_i = \mu, \sum p_i = 1} \prod p_i \leq C\right\}. \tag{3.14}$$

Se $\mu, v \in \mathscr{R}_c$, allora devono esistere le distribuzione di probabilità per p e q, tali che :

$$\sum p_i X_i = \mu\,, \quad \sum q_i X_i = v\,, \quad \prod p_i \geq C\,, \quad \prod q_i \geq C\,.$$

Per la concavità delle verosimiglianze multinomiali, la distribuzione $r = \beta p + (1-\beta)q$ soddisfa $\prod r_i \geq C$ e la $\sum r_i X_i = \beta\mu + (1-\beta)v$, dunque, $\beta\mu + (1-\beta)v \in \mathscr{R}_c$. Poiché $\mu, v \in \mathscr{R}_c$ questo implica che, a sua volta, $\beta\mu + (1-\beta)v \in \mathscr{R}_c$ per ogni $0 \leq \beta \leq 1$, dunque $\mathscr{R}_c$ è convesso.

Questo risultato non si estende agli intervalli per le funzioni *smooth* della media, ma implica comunque che questi intervalli siano connessi senza la presenza di punti vuoti.

Se, infatti, i parametri θ e ω siano legati da $\theta = g(\omega)$, tramite una qualche funzione g, allora i rispettivi intervalli $\mathscr{R}_{c,\theta}$ ed $\mathscr{R}_{c,\omega}$ saranno legati da:

$$\mathscr{R}_{c,\theta} = \{g(\omega) : \omega \in \mathscr{R}_{c,\omega}\}. \tag{3.15}$$

Dunque se θ è una funzione continua della media della popolazione, allora il suo intervallo di verosimiglianza empirica è dato dall'applicazione di quella stessa funzione su un convesso.

Supponiamo che ω rappresenti una media s-variata. Se $s \geq 2$, per una numerosità campionaria pari ad n ed un dato $c > 0$, possiamo scegliere una funzione continua, nondegenerativa $g : \mathbb{R}^s \to \mathbb{R}^s$ tale che, con probabilità positiva, $\mathscr{R}_{c,\theta}$, definito come nella (3.15), non sia convesso.

Questo risultato non è valido per $s = 1$ ed inoltre in tal caso la funzione $l(\mu)$ definita come nella (3.13) è convessa, com'è possibile provare a partire dalla (3.12) e verificando che la $l''(\mu) > 0$. Quindi, così come riportato in Hall, La Scala (1990) nel teorema 2.2, quale estensione degli enunciati presenti in Owen (1990), è possibile concludere che un intervallo di confidenza di verosimiglianza empirica costruito per la media della popolazione è sempre convesso ed un intervallo di confidenza di verosimiglianza empirica costruito per una funzione continua della media della popolazione è sempre collegato e senza punti vuoti. Inoltre, tranne nel caso in cui la funzione sia scalare, l'intervallo può essere non convesso con probabilità positiva.

3.3.3 Proprietà del secondo ordine

Tutte le caratteristiche esposte finora, in principal modo la convergenza in distribuzione del rapporto di verosimiglianza empirica, sono proprietà dette del primo ordine. Andremo ora a presentare quelle del secondo ordine, come ad esempio l'ordine di convergenza della distribuzione asintotica. E' possibile verificare come questo risultato può essere migliorato applicando un'ulteriore proprietà della verosimiglianza parametrica, ossia la correzione di Bartlett.

L'accuratezza della copertura dell'intervallo

La formula (3.10) precedentemente illustrata ci indica che gli intervalli di confidenza basati sulla verosimiglianza empirica sono asintoticamente dell'ordine corretto. questo significa che, se scegliamo c secondo la (3.8) e se $\mathscr{R}_c$ viene stabilito secondo la (3.9), allora:

$$P(\theta_0 \in \mathscr{R}_c) \to 1 - \alpha,$$

per $n \to \infty$. L'errore di copertura effettivo, dato da :

$$P(\theta_0 \in \mathscr{R}_c) - (1 - \alpha)$$

è trascurabile, essendo di ordine pari ad n^{-1}, piuttosto che $n^{-1/2}$.

DiCiccio, Hall e Romano (1988), ed in seguito DiCiccio e Romano (1989), riescono infatti a scomporre la $P(\theta_0 \in \mathscr{R}_c)$ e dimostrare, tramite l'espansione di Edgeworth che:

$$P(\theta_0 \in \mathscr{R}_c) = 1 - \alpha + O(n^{-1}).$$

La correzione di Bartlett

Altra importante operazione che è possibile operare per la verosimiglianza empirica è il famoso risultato noto sotto il nome di correzione di Bartlett. La causa dell'inaccuratezza appena esaminata per gli intervalli di confidenza costruiti con la verosimiglianza empirica risiede proprio nell'approssimazione con la χ^2. L'errore di approssimazione è di ordine n^{-1}. La correzione di Bartlett consente di apportare miglioramenti a questo risultato, consentendo di ridurre l'ordine da n^{-1} a n^{-2}, si parla perciò di intervalli del secondo ordine corretti.

La logica che opera all'interno della correzione di Bartlett è davvero semplice. Una parte dell'errore di approssimazione della χ^2_q con la distribuzione di $l(\theta_0)$ è spiegato tramite dall'evidenza che le medie delle due distribuzioni non sono coincidenti, ovvero $E\{l(\theta_0)\} \neq q$. La correzione di Bartlett prevede di riscalare $l(\theta_0)$, cosicché possa conseguire la media corretta. L'approssimazione con la χ^2_q viene applicata cioè non più a $l(\theta_0)$, bensì a $ql(\theta_0)/E\{l(\theta_0)\}$, in questo modo, con a costante:

$$E\{l(\theta_0)\} = q\{1 + n^{-1}a + O(n^{-2})\},$$

quando, invece, consideriamo il caso in cui a viene stimata, poiché $\hat{a} = a + O_p(n^{-1/2})$, allora sarà $1 + n^{-1}\hat{a} = 1 + n^{-1}a + O_p(n^{-3/2})$.

3.4 Polinomi locali e verosimiglianza empirica

Le proprietà appena esaminate per la verosimiglianza empirica valgono, evidentemente, in via asintotica. Questo significa che per n finito si pone un problema di approssimazione che ci si appresta a risolvere amalgamando le proprietà degli stimatori polinomi locali, trattati in precedenza, alla verosiglianza empirica.

Gli autori Chen e Qin (2000) propongono un approccio integrato di queste due tecniche per dati i.i.d. attraverso lo stimatore lineare, caso particolare dei polinomi locali per $p = 1$. Il risultato è un miglioramento dell'errore di copertura dell'intervallo di verosimiglianza empirica poiché, introducendo lo stimatore lineare dei polinomi locali, si migliorano le performance della verosimiglianza empirica rispetto al caso in cui l'intervallo venga calcolato sulla base dell' approssimazione asintotica alla normale dello stesso stimatore lineare.

L'intervallo ottenuto in quest'ultimo caso, infatti, presenta un errore più elevato per i punti esterni, rispetto ai punti interni dell'intervallo, evidenza alquano sorprendente, vista la naturale capacità degli stimatori polinomi locali di adattamento ai bordi. In un report tecnico degli stessi autori viene chiarito che tale circostanza si spiega con un incremento del termine di secondo ordine della varianza dello stimatore $\hat{m}_l(x)$ rispetto ai bordi dell'intervallo.

Il miglioramento ottenuto è spiegabile con la varianza dello stimatore che la verosimiglianza empirica implicitamente adotta per lo stimatore lineare dei polinomi locali. Sfruttando, infatti, la caratteristica, propria dello stimatore della verosimiglianza empirica, di studentizzazione automatica, si ottiene una selezione della varianza dello stimatore $\hat{m}_l(x)$ implicita, che ne garantisce lo stesso ordine di errore lungo ogni punto dell'intervallo, interno o bordo che sia.

Partendo da n coppie di osservazioni indipendenti $\{Y_i, X_i\}_{i=1}^n$, con Y_i variabili di risposta associate ai punti $X_i \in [0, \infty)$ secondo la legge $Y_i = m(X_i) + \varepsilon_i$, m funzione di regressione non nota con supporto $[0, \infty)$ e ε_i errori indipendenti a media nulla e varianza finita, secondo le ipotesi già descritte nel capitolo precedente, definita f la funzione di densità delle X_i e

$$V(x) = E[\{Y - m(x)\}^2 | X = x]$$

la varianza condizionata di Y, dato $X = x$ e definito K il kernel con h parametro di *smoothing* si impongono le seguenti condizioni:

- K è una funzione di densità simmetrica e limitata, definita sul compatto $[-1, 1]$
- per $n \to \infty$, $h \to 0$ ed $nh \to \infty$ esiste un numero reale $s \geq 4$ tale che $E|Y|^s < \infty$, $nh^{2s} \to 0$ e $n^{s-2}h^s \to \infty$
- f, m e V hanno derivate finite fino al secodno ordine in un intorno di x, con $f(x) > 0$ e $V(x) > 0$.

Lo stimatore lineare dei polinomi locali per $m(x)$ per ogni $x \in [0, \infty)$ è dato da:

$$\hat{m}_l(x) = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i} \tag{3.16}$$

dove

$$W_i = K_h(x-X_i)\left\{s_{n,2} - \frac{(x-X_i)s_{n,1}}{h}\right\},\ s_{n,l} = (nh)^{-1}\sum_{i=1}^{n}\frac{K_h(x-X_i)(x-X_i)^l}{h^l}, \tag{3.17}$$

per $l = 0, 1, 2$ e, al solito, $K_h(\cdot) = K(\cdot/h)$.

Vengono scelti $p_1, \ldots, p_n$ numeri non negativi a somma 1 , per qualunque $x \in [0, \infty)$, in questo caso la verosimiglianza empirica di $E\{\hat{m}_l(x)\}$, valutata in θ, è data da:

$$L(\theta) = \max_{\sum p_i W_i(Y_i-\theta)=0} \prod_{i=1}^{n} p_i. \tag{3.18}$$

Se applichiamo il metodo dei moltiplicatori di Lagrange per ottenere le p_i ottimali, il rapporto della logverosimiglianza empirica diventa:

$$l(\theta) = -2log\{L(\theta)n^n\} = 2\sum log\{1 + \lambda(\theta)W_i(Y_i - \theta)\}, \tag{3.19}$$

con $\lambda(\theta)$ che soddisfa

$$\sum_{i=1}^{n} W_i(Y_i - \theta)\{1 + \lambda(\theta)W_i(Y_i - \theta)\}^{-1} = 0. \tag{3.20}$$

Teorema 1 (Chen e Qin, 2000) *Sotto la validità delle tre ipotesi esplicitate in precedenza, se* $nh^5 \to 0$ *e* $m''(x) \neq 0$, $l\{m(x)\}$ *si distribuisce come una* χ^2_1.

Otteniamo così una verosimiglianza empirica che, a tutti gli effetti, risulta essere quella di $E\{\hat{m}_l(x)\} = m(x) + distorsione$, piuttosto che semplicemente $m(x)$.

Per ovviare a questo problema ed ottenere una verosimiglianza empirica riferita esclusivamente ad $m(x)$ abbiamo due opzioni note in letteratura.

La prima consiste, così come esibito da Hall (1991), in una correzione esplicita, dopo aver determinato per via diretta l'entità della distorsione stessa. La seconda alternatica invece pratica una riduzione della distorsione, tramite la sottostima del grado di *smoothing* effettuato, come evidenziato da Hall (1992) e Neumann (1995).

Hall (1992) dimostra che la soluzione migliore in termini di accuratezza di copertura dell'intervallo è ottenuta in questa seconda ipotesi, tramite i cosiddetti methodi di *undersmoothing*.

Con l'aggiunta di una sola condizione utile ad effettuare l'*undersmoothing* è possibile estendere il teorema di Wilks, così come dimostrato da Chen e Qin (2000).

3.5 I principali sviluppi della verosimiglianza empirica nell'ambito delle serie storiche

Una evoluzione, rispetto al lavoro di Chen e Qin (2000), sull'applicazione dei polinomi locali alla verosimiglianza empirica è rappresentata dai risultati ottenuti qualche anno dopo da Chen, Härdle e Li (2003). Gli autori trattano il caso in cui questi stimatori vengono applicati ai processi di α-*mixing*.

Le ipotesi adottate prevedono di partire da una serie storica strettamente stazionaria $\{Y_i, X_i\}_{i=1}^n$, con $Y_i \in \mathbb{R}$ ed $X_i \in \mathbb{R}^d$, considerare $m(x) = E(Y|X = x)$ la media condizionale del processo, f la funzione di densità di X e $\sigma^2(x) = var(Y|X = x)$ la varianza condionata del processo Y, posto $X = x \in S$ e supporre che $\{m_\theta|\theta \in \Theta\}$ sia un modello parametrico per la media m e che $\hat{\theta}$ sia uno stimatore di θ in tale modello parametrico.

In realtà, il problema di testare la funzione di regressione parametrica contro un'alternativa di tipo non parametrico era stato già oggetto di indagine nel caso di dati indipendenti ed identicamente distribuiti.

Härdle e Mammen (1993) dapprima avevano proposto una statistica test basata su una distanza L_2 tra uno stimatore kernel non parametrico della media condizionata e la funzione parametrica ipotizzata.

La stessa tecnica di confronto tra la stima non parametrica e quella parametrica era stata utilizzata da Eubank e Spiegelman (1990) e Hart (1997), mentre Kreiss et al. (1998) avevano esteso il test di Härdle e Mammen al contesto di dati dipendenti, tramite l'implementazione del wild bootstrap.

Hjellvik et al. (1998) avevano poi testato la linearità per mezzo degli stimatori polinomi locali e Koul e Stute (1999) avevano proposto un test per le serie storiche basato sul processo empirico costruito a partire dai residui della stima parametrica sotto l'ipotesi nulla H_0.

Recentemente Horowitz e Sporokoiny (2001) hanno proposto un test per i dati indipendenti che si avvaleva del test di Härdle e Mammen, utilizzandolo simultaneamente per un insieme di valori di *bandwidth*. Il test risultava consistente per H_1 con $c_n = O(n^{1/2}\sqrt{[log\{log(n)\}]})$.

Kitamura (1997) ha affrontato, invece, il problema avvalendosi della verosimiglianza empirica a blocchi, per parametri attinenti a processi debolmente dipendenti. Monti (1997) aveva costruito intervalli di confidenza per un parametro di una serie storica stazionaria tramite il metodo di stima di Whittle. Per dati indipendenti, la verosimiglianza empirica è stata utilizzata infine da Tripatchi e Kitamura (2000) per testare le restrizioni sui momenti condizionati e da Fan e Zhang (2000) per funzioni non parametriche basate su una approssimazione di tipo locale.

3.5.1 La verosimiglianza empirica e i polinomi locali nel caso α - *mixing*

Chen, Härdle e Li (2003) considerano il processo stocastico $\{(X_i, Y_i)\}_{i=1}^n$, strettamente stazionario con $Y_i \in \mathbb{R}$ e $X_i \in \mathbb{R}^d$. Sia $m(x) = E(Y|X = x)$ la media condizionata, f la funzione di densità di X e $\sigma^2(x) = var(Y|X = x)$ la varianza condizionata di Y posto $X = x \in S$.
Gli stessi autori, inoltre, testano l'ipotesi nulla, $H_0 : m(x) = m_\theta(x)$ per ogni $x \in S$, contro una serie di alternative non parametriche descritte dall'ipotesi $H_1 : m(x) = m_\theta(x) + c_n\Delta_n(x)$, dove c_n rappresenta una sequenza di valori non casuali tendenti a 0 per $n \to \infty$ e $\Delta_n(x)$ sequenza di funzioni limitate. Dapprima gli autori introducono lo stimatore kernel non parametrico per m. Dato $S = \{x \in \mathbb{R}^d | f(x) \geq \beta\}$ per $\beta > 0$ insieme compatto, si assume che $S = [0, 1]^d$, consideriamo Λ il kernel univariato di ordine r, definito sul supporto compatto [-1,1], tale che:

$$\int \lambda(t)dt = 1,$$

$$\int t^l \lambda(t)dt = 0, se 1 \leq l \leq r - 1,$$

$$\int t^r \lambda(t)dt = k_r \neq 0$$

per qualsiasi intero $r \geq 2$, e K il prodotto kernel d- dimensionale di Λ, ossia:

$$K(t_1, ;t_d) = \prod_{i=1}^{d} \Lambda(t_i).$$

Detto h il parametro positivo di *bandwidth*, convenzionalmente $K_h(u) = h^{-d}K(h^{-1}u)$. Lo stimatore non parametrico di $m(x)$ considerato è il Nadaraya-Watson dato da:

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i K_h(x - X_i) / \sum_{i=1}^{n} K_h(x - X_i).$$

Il modello parametrico, dunque, interviene nel modo seguente:

$$\tilde{m}_{\tilde{\theta}}(x) = \sum_{i=1}^{n} K_h(x - X_i) m_{\tilde{\theta}}(X_i) / \sum_{i=1}^{n} K_h(x - X_i).$$

Per ovviare al problema di distorsione associato alla stima non parametrica, la statistica test considerata è basata sulle differenze tra $\tilde{m}_{\tilde{\theta}}$ e $\hat{m}$, piuttosto che $\hat{m}$ e $m_{\tilde{\theta}}$.

In un punto arbitrario $x \in S$, considerate le $p_i(x)$ i pesi associati a (X_i, Y_i), la verosimiglianza empirica per $\tilde{m}_{\tilde{\theta}}$ e $\hat{m}$ è rappresentata da:

$$L\{\tilde{m}_{\tilde{\theta}}\} = max \left\{ \prod_{i=1}^{n} p_i(x) \right\},$$

sotto il vincolo di:

$$\sum_{i=1}^{n} p_i(x) = 1$$

e

$$\sum_{i=1}^{n} p_i(x) K(\frac{(x - X_i)}{h})\{Y_i - \tilde{m}_{\tilde{\theta}}\} = 0.$$

Tramite l'utilizzo dei moltiplicatori di Lagrange, i pesi ottimali sono dati da:

$$p_i(x) = n^{-1}\left[1 + \lambda(x) K(\frac{(x - X_i)}{h})\{Y_i - \tilde{m}_{\tilde{\theta}}\}\right]^{-1}$$

dove i $\lambda(x)$ sono calcolati come le radici di:

$$\sum_{i=1}^{n} \frac{K\{\frac{(x-X_i)}{h}\}\{Y_i - \tilde{m}_{\tilde{\theta}}\}}{1 + \lambda(x) K\{\frac{(x-X_i)}{h}\}\{Y_i - \tilde{m}_{\tilde{\theta}}\}} = 0.$$

Lo stimatore di massima verosimiglianza empirica, ottenuto per $p_i(x) = n^{-1}$, corrisponde allo stimatore Nadaraya-Watson di $\hat{m}(x)$. Il rapporto di logverosimiglianza empirica sarà dato da:

$$l\{\tilde{m}_{\tilde{\theta}}\} = -2log[L\{\tilde{m}_{\tilde{\theta}}\} n^n].$$

Chen, Härdle e Li (2003), definiscono le seguenti ipotesi:

A1 K è un kernel univariato di ordine r definito su un supporto compatto in $[-1, 1]$, Lipschitz continuo con $d < 4$ ed il parametro di *smoothing* $h = O(n^{-1/(d+2r)})$

A2 f, m, σ^2 hanno derivate continue fino al secondo ordine in S ed entrambe $fe\sigma^2$ sono limitate inferiormente in S

A3 $\hat{\theta}$ è uno stimatore parametrico di θ all'interno della famiglia del modello parametrico e $sup_{x \in S} |m_{\hat{\theta}}(x) - m_{\theta}(x)| = O_p(n^{-1/2})$

A4 $\Delta_n(x)$ è limitato uniformente rispetto ad x ed n e $c_n = n^{1/2} h^{-d/4}$, che rappresenta l'ordine della differenza tra le ipotesi H_0 e H_1

A5 $E[exp\{a_0 | Y_1 - m(X_1)\}] < \infty$ per $a_0 > 0$; $E(|Y_i|^k | X_i) < \infty$ per $k > 1$; per ogni i, $E\{Y_i - m(X_i) | \Omega_{i-1}\} = 0$, dove $\Omega_{i-1}\}$ rappresenta la σ - algebra generata da $\{(X_{j+1}, Y_j)\}_{j=1}^{i-1}$

A6 la densità condizionata di X dato Y, $f_{X|Y} \leq A_1 < \infty$, la densità condizionata congiunta di (X_1, X_l) dato (Y_1, Y_l) è limitata per ogni $l > 1$ e la densità congiunta di $(X_1, Y_1, X_s, Y_s, X_t, Y_t)$ per $t > s > 1$ è continua e limitata da una costante che è indipendente da s e t

A7 il processo $\{(X_i, Y_i)\}$ è strettamente stazionario ed α - *mixing* con $\alpha(k) = a\rho^k$ per $a > 0$ e $\rho \in (0, 1)$.

Lemma 1 (Chen, Härdle e Li, 2003) *Sotto le ipotesi A1-A7 si ha che:*

$$sup_{x\in S}|\lambda(x)| = o_p\{(nh^d)^{1/2}log(n)\}. \tag{3.21}$$

Inoltre, all'interno dello stesso articolo si dimostra, per le ipotesi sopra riportate, che:

$$l\{\tilde{m}_{\tilde{\theta}}\} = nh^d\frac{\{\hat{m}(x)-\tilde{m}_{\tilde{\theta}}\}^2}{V(x;h)} + \tilde{O}\{(nh^d)^{1/2}log^3(n) + h^2log^2(n)\}$$

con $\tilde{O}$ definito nel modo seguente: $\gamma(x) = \tilde{O}_p(\delta_n)$ corrispondente a

$$\sup_{x\in S}|\gamma(x)| = O_p(\delta_n);$$

il che si traduce con l'implicazione che $l\{\tilde{m}_{\tilde{\theta}}\}$ è asintoticamente equivalente ad una distanza L_2 studentizzata tra $\tilde{m}_{\tilde{\theta}}(x)$ e $\hat{m}(x)$.

Capitolo 4

Inferenza con la verosimiglianza empirica e i polinomi locali nel caso di dati dipendenti: valutazione del parametro di *bandwidth*

4.1 Introduzione

Il capitolo che segue contiene al suo interno risultati innovativi rispetto a quanto presente in letteratura, proponendo un'estensione del teorema di Wilks, valida nel caso univariato e per l'impiego di stimatori di tipo Kernel.

Durante il capitolo precedente sono stati illustrati diversi risultati presenti in letteratura che interessavano dapprima l'ambito dei dati indipendenti ed identicamente distribuiti, considerando anche il caso di utilizzo dei polinomi locali, poi si è trattato il caso α - *mixing*, dove lo stimatore considerato era di tipo kernel.

Partendo dall'articolo di Chen, Härdle e Li (2003), considereremo per semplicità il caso in cui $x \in \mathbb{R}^1$ e stileremo nuovi risultati che dimostrano la convergenza in legge ad una χ^2 degli stimatori $l[m(x)]$ e $l[s^2(x)]$, per due diverse formulazioni di modelli.

Il caso in cui ci occupiamo dello stimatore $l[m(x)]$ concerne il modello che nel seguito verrà indicato nella (4.1), di tipo omoschedastico, mentre, al contrario, $l[s^2(x)]$ viene studiato a partire dal modello (4.2), di tipo eteroschedastico.

Sarebbe stato possibile stimare in maniera congiunta la media e la varianza condizionata del processo, avvalendoci di un unico modello con una formulazione più complessa, ma essendo lo scopo del nostro operato valutare la sensibilità all'h, rispetto alle ipotesi ed i modelli, si è preferito occuparci separatamente dei due valori da stimare, in modo da isolare l'impatto, nell'uno e nell'altro caso, che il parametro di *bandwidth* esercita per la stima.

L'ambito di applicazione del nostro lavoro, così come anticipato nei capitoli precedenti, è quello delle serie storiche ed in particolar modo ci poniamo nel caso di processi per cui valgano le ipotesi di α-*mixing*.

L'impiego della verosimiglianza empirica viene combinato all'utilizzo dei polinomi locali per la strutturazione di due tipologie di test, effettuate sui due stimatori, $l[m(x)]$ e $l[s^2(x)]$, rispettivamente utilizzati in corrispondenza della media condizionata per il modello omoschedastico e della varianza condizionata per il modello eteroschedastico.

4.2 L'impostazione dei test d'ipotesi

Detto θ il valore obiettivo possiamo considerare un primo test dato da:

$$H_0 : m(x) = \theta_m$$

contro l'ipoteso alternativa

$$H_1 : m(x) \neq \theta_m,$$

per il modello $Y_t = m(Y_{t-1}) + \varepsilon_t$. Nel caso, invece, il test venga utilizzato per la varianza condizionata, le impostazioni diventano:

$$H_0 : s^2(x) = \theta_s$$

contro l'ipotesi alternativa

$$H_1 : s^2(x) \neq \theta_s,$$

per il modello di tipo $Y_t = s(Y_{t-1})\varepsilon_t$.

La stima di $\hat{m}(x)$ e $\hat{s}^2(x)$ viene costruita mediante l'utilizzo dei polinomi locali, mentre sarà impiegata la verosimiglianza empirica per consentire l'inferenza, a partire dai dati a disposizione.

Il vantaggio ottenuto si sostanza, senza dubbio, nella mancata necessità di dover determinare esplicitamente la varianza per lo stimatore polinomi locali, ma avvalendosi al contrario della caratteristica della verosimiglianza empirica di stabilire implicitamente una varianza per questo tipo di stimatore.

L'intervallo di confidenza generato, inoltre, è un intervallo del secondo ordine corretto, rispetto alla possibilità di fare inferenza esclusivamente in termini asintotici.

4.3 La logica di valutazione dell'h per le differenti casistiche

Andare a valutare l'impatto di h per le diverse ipotesi significherà nel nostro caso considerare, per le diverse casistiche, come si comporta il test formulato e come varia la potenza del test al variare del valore di h considerato.

In tutti i casi definiti, andremo a verificare come varia la probabilità di non rifiutare l'ipotesi nulla sia nel caso in cui H_0 sia da considerarsi vera, che nel caso inverso. In questo modo sarà facile identificare non soltanto l'attendibilità della copertura dell'intervallo di confidenza, confrontando il valore ottenuto con quello nominale prescelto, mediante la verifica del numero dei casi in cui l'ipotesi nulla vera viene effettivamente accettata, ma sarà anche possibile valutare l'effettivo tasso di rifiuto, nel caso di ipotesi nulla non vera. Questo ci consentirà di avere anche una stima empirica per la potenza del test costruito.

Il vero problema che sarà interessante considerare è dato dalla verifica di cosa avviene nel caso di *undersmoothing* e se le evidenze empiriche ci confermano i risultati teorici in precedenza riportati.

L'ordine di convergenza dell'h ottimale, stabilito tramite la minimizzazione dell'MSE, infatti, così com'è stato sottolineato in precedenza, non può essere impiegato nel caso si costruisca uno stimatore dato dalla combinazione della verosimiglianza empirica con i polinomi locali.

Come riportato da Chen e Van Keilegom (2009), nel caso di stime Kernel $\hat{m}(x)$ non risulta uno stimatore corretto di $m(x)$, ma piuttosto:

$$E\{\hat{m}(x)\} = m(x) + bias(x) + o(h^2),$$

dove il termine $bias(x)$, che rappresenta la distorsione per lo stimatore, è dato da

$$1/2h^2\{m''(x) + 2m'(x)f'(x)/f(x)\},$$

con f che sta ad indicare la densità delle X_i. Quindi la verosimiglianza empirica risulterebbe valutata presso un θ che non corrisponde ad $m(x)$, ma $m(x) + bias(x)$.

Così come riportato in precedenza, Hall (1991) propone due tipologie di opzioni per ovviare a questa problematica, la prima consiste nel determinare l'esatta incidenza della $bias(x)$ e sottrarre quest'ammontare dalla stima kernel prodotta. La seconda prevede, invece, l'adozione di una tecnica cosiddetta di *undersmoothing*, avvalendosi, cioè, di un kernel con un parametro di *bandwidth* dato da $h = (n^{-1/(4+d)})$, dove d rappresenta la dimensione di X.

E' lo stesso Hall a prediligere la seconda soluzione e la nostra attenzione sarà proprio posta ai casi in cui il nostro h realizza l'*undersmoothing*.

L'operato si compone dapprima di uno spazio per i risultati teorici, che vengono di seguito riportati all'interno di questo stesso capitolo. Successivamente, le evidenze empiriche, frutto di diverse casistiche, vengono analizzate grazie al metodo delle simulazioni Monte Carlo, raccolte e commentate nel capitolo successivo, dove verrà effettuata una analisi anche alla luce dei risultati teorici qui conseguiti.

4.4 Risultati teorici presentati

Partendo da un modello omoschedastico, secondo la formulazione:

$$Y_t = m(Y_{t-1}) + \varepsilon_t \tag{4.1}$$

è possibile proporre, analogamente a quanto dimostrato da Chen, Härdle, Li (2003), un'estensione del teorema di Wilks per $l[m(x)]$.

Le ipotesi operate sono le seguenti:

A1 K è un kernel univariato di ordine 2 definito su un supporto compatto in $[-1, 1]$, limitato

A2 f ed m hanno derivate continue fino al secondo ordine nel punto x, con $f(x) > 0$

A3 $E[exp\{a_0|Y_2 - m(Y_1)\}] < \infty$ per $a_0 > 0$; $E(|Y_i|^k|Y_{i-1}) < \infty$ per $k > 1$; per ogni i, $E\{Y_i - m(Y_{i-1})|\Omega_{i-1}\} = 0$, dove $\Omega_{i-1}\}$ rappresenta la σ - algebra generata da $\{Y_j\}_{j=1}^{i-1}$

A4 la densità condizionata di Y_{i-1} dato Y_i, $f_{Y_{i-1}|Y_i} \leq A_1 < \infty$, la densità condizionata congiunta di (Y_0, Y_{l-1}) dato (Y_1, Y_l) è limitata per ogni $l > 1$ e la densità congiunta di $(Y_0, Y_1, Y_{s-1}, Y_s, Y_{t-1}, Y_t)$ per $t > s > 1$ è continua e limitata da una costante che è indipendente da s e t

A5 il processo $\{Y_i\}$ è strettamente stazionario ed α - *mixing* con $\alpha(k) = a\rho^k$ per $a > 0$ e $\rho \in (0, 1)$

Le funzioni f e m sono definite come in Chen, Härdle, Li (2003).

Sia $\hat{m}(x)$ lo stimatore di $m(x)$ con:

$$\hat{m}(x) = \frac{\sum_{i=2}^{n} Y_i K_h(x - Y_{i-1})}{\sum_{i=2}^{n} K_h(x - Y_{i-1})}$$

Proposizione 1. *Sotto le ipotesi A1-A5* $l[m(x)] \xrightarrow{L} \chi^2_{(1)}$, *per* $n \to \infty$, *se* $nh^5 \to 0$ *e questa condizione è anche necessaria se* $m''(x) \neq 0$.

Dimostrazione. Usando la dimostrazione del Lemma 1 in Chen, Härdle, Li (2003) risulta:

$$|\hat{\lambda}(x)| = o_p\{(nh)^{1/2}\}.$$

Di nuovo, seguendo gli stessi ragionamenti presenti in Chen, Härdle, Li (2003) a pagina 666, si ha che:

$$l[m(x)] = nh\frac{(\hat{m}(x) - m(x))^2}{V(x; h)} + O\{(nh)^{-1/2} + h^2\},$$

con

$$V(x;h) = \frac{v(x;h)}{b^2(x;h)},$$

$$v(x;h) = \sigma_\varepsilon^2 h \int K_h^2(x-y)f(y)dy$$

e

$$b(x;h) = h \int K_h(x-y)f(y)dy$$

ma

$$nh(\hat{m}(x) - m(x))^2 = O_p\{(nh^5)[m''(x)]^2\},$$

come dimostrano Härdle e Tsybakov (1997), Masry e Fan (1997). Il risultato segue. □

Ponendoci ora nel caso di modello eteroschedastico, consideriamo:

$$Y_t = s(Y_{t-1})\varepsilon_t, \tag{4.2}$$

è possibile dimostrare che un risultato analogo vige per $l[s(x)]$.
Le condizioni imposte sono le seguenti:

A1 K è un kernel univariato di ordine 2 definito su un supporto compatto in $[-1,1]$, limitato

A2 f ed s^2 hanno derivate continue fino al secondo ordine nel punto x, con $f(x) > 0$

A3 $E[exp\{a_0|Y_2/s(Y_1)\}] < \infty$ per $a_0 > 0$; $E(|Y_i|^k|Y_{i-1}) < \infty$ per $k > 2$; per ogni i, $E\{Y_i/s(Y_{i-1})|\Omega_{i-1}\} = 0$, dove $\Omega_{i-1}\}$ rappresenta la σ - algebra generata da $\{Y_j\}_{j=1}^{i-1}$

A4 la densità condizionata di Y_{i-1} dato Y_i, $f_{Y_{i-1}|Y_i} \leq A_1 < \infty$, la densità condizionata congiunta di (Y_0, Y_{l-1}) dato (Y_1, Y_l) è limitata per ogni $l > 1$ e la densità congiunta di $(Y_0, Y_1, Y_{s-1}, Y_s, Y_{t-1}, Y_t)$ per $t > s > 1$ è continua e limitata da una costante che è indipendente da s e t

A5 il processo $\{Y_i\}$ è strettamente stazionario ed α - *mixing* con $\alpha(k) = a\rho^k$ per $a > 0$ e $\rho \in (0,1)$

Sia $\hat{m}_2(x)$ lo stimatore di $s^2(x) \equiv m_2(x)$ con:

$$\hat{m}_2(x) = \frac{\sum_{i=2}^n Y_i^2 K_h(x - Y_{i-1})}{\sum_{i=2}^n K_h(x - Y_{i-1})}$$

Proposizione 2. *Sotto le assunzioni in Härdle e Tsybakov (1997) e le condizioni A1-A5 appena riportate,* $l[s^2(x)] \xrightarrow{L} \chi^2_{(1)}$, *per* $n \to \infty$, *se* $nh^5 \to 0$ *e questa condizione è anche necessaria se* $[s^2(x)]'' \neq 0$.

Dimostrazione Avvalendoci dello stesso approccio utilizzato per la Proposizione 1, otteniamo:

$$l[s^2(x)] = nh\frac{(\hat{m}_2(x) - m_2(x))^2}{V_2(x;h)} + O\{(nh)^{-1/2} + h^2\},$$

con

$$V_2(x;h) = \frac{v_2(x;h)}{b^2(x;h)},$$

$$v_2(x;h) = m_{4\varepsilon}h\int K_h^2(x-y)f(y)s^4(y)dy,$$

$$b(x;h) = h\int K_h(x-y)f(y)dy,$$

dove

$$s^2(x) = E(Y_t^2|Y_{t-1} = x)$$

e

$$s^4(x)m_{4\varepsilon} = var(Y_t^2|Y_{t-1} = x),$$

$$m_{4\varepsilon} = E\{(\varepsilon_i^2 - 1)^2\} < \infty,$$

ma

$$nh(\hat{m}_2(x) - m_2(x))^2 = O_p\{(nh^5)\{[s^2(x)]''\}^2\},$$

come dimostrano Härdle e Tsybakov (1997), Masry e Fan (1997). Il risultato segue. □

Capitolo 5

I risultati empirici

5.1 Introduzione

Nello scorso capitolo abbiamo trattato, da un punto di vista teorico, la convergenza in legge degli stimatori polinomi locali con la verosimiglianza empirica.

Il capitolo in esame contiene, tramite l'utilizzo delle simulazioni Monte Carlo, non soltanto la conferma, per via empirica, dei risultati teorici precedentemente proposti, ma riesce anche a strutturare la valutazione della sensibilità del parametro di *bandwidth*, raccogliendo per ognuna delle casistiche esaminate, l'affidabilità del test, sia confrontando la copertura nominale per l'intervallo con quella reale, sia considerando la potenza del test, quando viene valutato il caso in cui l'ipotesi nulla non sia vera.

I test saranno effettuati su varie tipologie di processi, calcolati in differenti punti e impostati nel caso di ipotesi nulla vera oppure no.

5.2 Struttura delle Simulazioni

I modelli presi in esame per questo studio sono tre, classificabili pricipalmente in due categorie. Nel primo caso, $Y_t = m(Y_{t-1}) + \varepsilon_t$ rappresenta un modello di tipo omoschedastico che ben si presta alla stima di $m(x)$ e due sono i modelli esaminati appartenenti a tale tipologia.

Al contrario, il terzo ed ultimo modello che risponde alla formulazione di tipo: $Y_t = s(Y_{t-1})\varepsilon_t$, rappresenta un modello eteroschedastico, per il quale verrà stimato $s^2(x)$. E' chiaro che i primi due modelli considerati forniscono una stima della media condizionale del processo, mentre il terzo sarà utile a definire la stima della varianza condizionata del processo.

Per ognuno dei modelli considerati due tipologie di errori verranno esaminati: errori distribuiti normalmente ed errori provenienti da una t di Student.

In entrambi i casi le due v.c. sono state costruite per garantire la varianza unitaria del processo e media nulla. Ancora, per ognuno dei modelli e con entrambe le opzioni di errori viene esaminata una numerosità pari a 500 e poi a 1000, per indagare il comportamento asintotico dei risultati, l'efficienza e l'eventuale consistenza. Gli stimatori polinomi locali per $m(\cdot)$ e $s^2(\cdot)$ sono con $p = 1$ (polinomi locali lineari). Tali stimatori sono una generalizzazione degli stimatori considerati nelle proposizioni 1 e 2.

5.2.1 I modelli

Il primo modello considerato é rappresentato da un AR(1) con la seguente formulazione:

$$Y_t = 0.8Y_{t-1} + \varepsilon_t. \tag{5.1}$$

Due saranno i casi esaminati per le ε_t, ovvero il caso in cui si distribuiscano come una Normale Standard e quello in cui derivino da una variabile del tipo:

$$T_{(10)}\sqrt{\frac{8}{10}}.$$

Le due numerosità esaminate sono pari a 500 e 1000.

Il secondo modello è costituito da un Expar(1), del tipo:

$$Y_t = 0.6Y_{t-1} + 3.5Y_{t-1}e^{-Y_{t-1}^2} + \varepsilon_t. \tag{5.2}$$

Anche in questo caso si strutturano le due tipologie di errori e di numerosità considerata.

Infine, il terzo modello è rappresentato da un modello di tipo ARCH, tipicamente utilizzato per l'analisi della volatilità. La sua formulazione è la seguente:

$$Y_t = (0.1 + 0.3Y_{t-1}^2)^{1/2}\varepsilon_t \tag{5.3}$$

ed anche in questo caso verificheremo la sensibilità all'h, sotto le ipotesi considerate per la legge degli errori e per la numerosità.

Dapprima consideriamo i box-plot corrispondenti alle stime dei tre modelli. Per i primi due modelli, così come specificato, si effettua una stima di $m(x)$, per l'ultimo la quantità stimata è la varianza condizionata del processo, $s^2(x)$.

Nelle tre figure che seguono, per ogni modello viene riportato il caso (a) e (c) in cui la stima viene prodotta per un valore di $x = -0.5$, mentre corrispondentemente nei casi (b) e (d) la stima viene prodotta ad un valore di $x = +0.5$. Viene anche segnalato per ognuno dei grafici il valore vero di x.

E' sicuramente interessante notare la precisione delle stime prodotte.

I box-plot generati ne rivelano la consistenza, in quanto per $n = 1000$, quindi all'aumentare della numerosità, l'errore quadratico medio decresce, com'era auspicabile. Il valore evidenziato nei grafici dalla retta parallela all'asse delle ascisse rappresenta il valore vero, questo ci indica come i box-plot centrino effettivamente il valore reale, in particolare per il primo modello che, del resto, presenta una formulazione più semplice rispetto agli altri.

Ciononostante, anche per gli altri due modelli i risultati si presentano pressoché analoghi a quelli ottenuti. Nel caso del modello Expar(1) si conferma un miglioramento delle prestazioni, come esibito dal relativo box-plot, all'aumentare del valore di n.

Il caso dell'ARCH, invece, diventa significativo perché degno di nota è come il grafico (c), ovvero quello che considera $x = -0.5$, presenti un box-plot davvero concentrato intorno al valore vero ma, di contro, evidenzia anche un problema di outliers, abbastanza distanti dalle rette determinate.

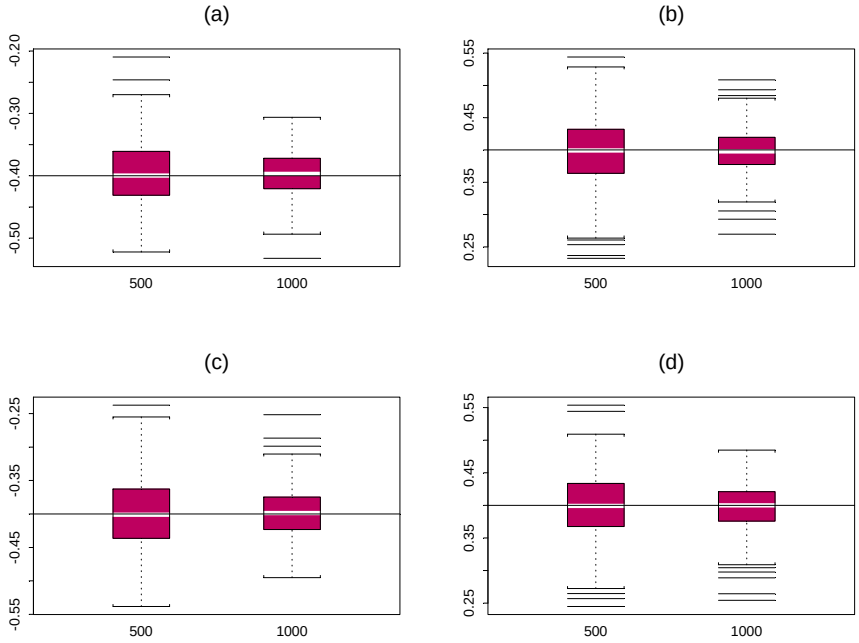

Figura 5.1: La figura rappresenta i box-plot della stima puntuale di $m(x)$ ottenuta tramite le simulazioni del modello AR(1) corrispondente alla (5.1). Si considera un valore di x calcolato pari a -0.5, per i casi (a) e (c) e di $+0.5$, invece, per (b) e (d). Ogni grafico riporta il caso in cui n sia pari a 500, poi 1000.

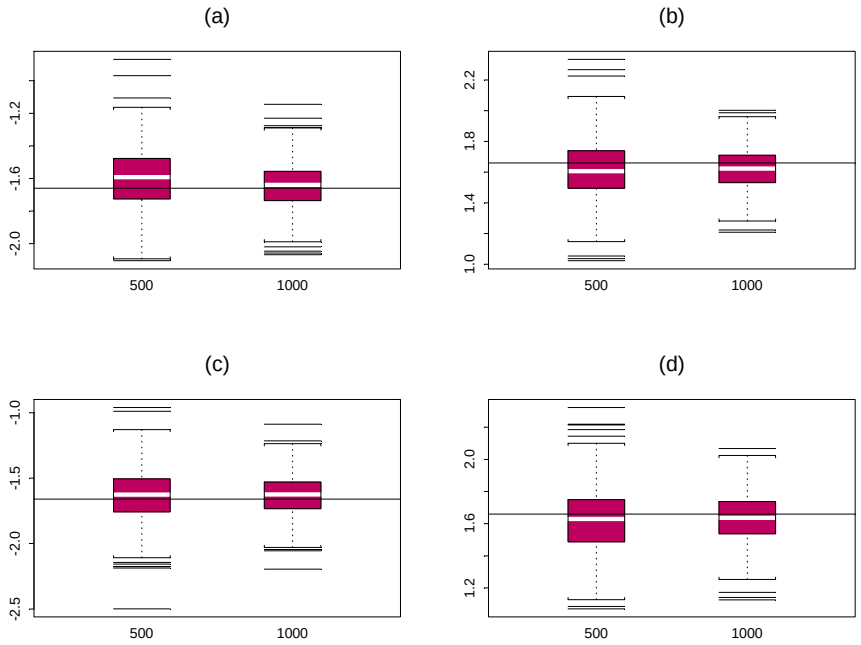

Figura 5.2: La figura rappresenta i box-plot della stima puntuale di $m(x)$ per il modello Expar(1) che risponde alla (5.2), ottenuta tramite simulazioni, per un valore di x calcolato pari a -0.5, nei casi (a) e (c), $+0.5$ invece per (b) e (d). Ogni grafico riporta il caso in cui n sia pari a 500, poi 1000.

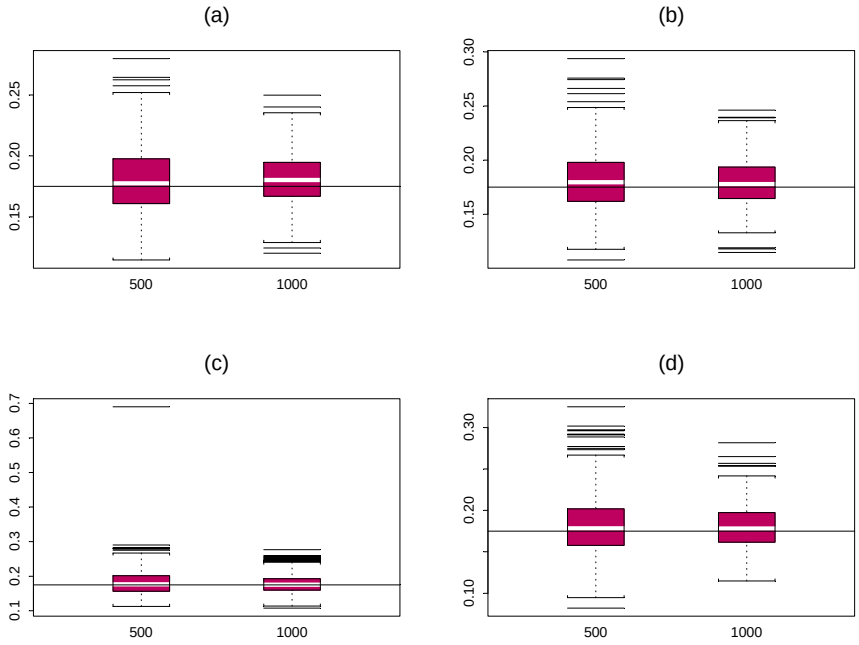

Figura 5.3: La figura rappresenta i box-plot della stima puntuale di $s^2(x)$ ottenuta tramite le simulazioni sul modello ARCH, formulato come nella (5.3), per un valore di x calcolato pari a -0.5, per i casi (a) e (c), $+0.5$ invece per (b) e (d). Ogni grafico riporta il caso in cui n sia pari a 500, poi 1000.

5.3 Analisi del livello di copertura e potenza del test

In questo paragrafo, alla luce delle proprietà teoriche dimostrate nel capitolo precedente, analizziamo la performance delle varie tipologie di test, esaminandone parallelamente la potenza del test e la sua *size*.

Sarà possibile effettuare questa valutazione poiché, per ogni modello e casistica differente, considerando quindi entrambe le tipologie di errori, i test vengono impostati corrispondentemente per tre valori di θ, un valore inferiore, nel caso (a), al θ vero; un valore di θ coincidente, nel caso (b), con il valore vero; infine, un valore di θ, caso (c), maggiore rispetto al vero valore del parametro.

Tutto questo, detto θ il valore obiettivo, si traduce nella seguente impostazione dei test per i primi due modelli che stimano $m(x)$:

test (a)

$$H_0 : m(x) = \theta_m \quad , \quad H_1 : m(x) \neq \theta_m, \tag{5.4}$$

con $\theta_m < \theta$ vero;

test (b)

$$H_0 : m(x) = \theta_m \quad , \quad H_1 : m(x) \neq \theta_m, \tag{5.5}$$

con $\theta_m = \theta$ vero;

test (c)

$$H_0 : m(x) = \theta_m \quad , \quad H_1 : m(x) \neq \theta_m, \tag{5.6}$$

con $\theta_m > \theta$ vero.

Per il modello ARCH, invece, i tre casi corrispondono alle seguenti impostazioni:

test (a)

$$H_0 : s^2(x) = \theta_s \quad , \quad H_1 : s^2(x) \neq \theta_s, \tag{5.7}$$

con $\theta_s < \theta$ vero;

test (b)

$$H_0 : s^2(x) = \theta_s \quad , \quad H_1 : s^2(x) \neq \theta_s, \tag{5.8}$$

con $\theta_s = \theta$ vero;

test (c)

$$H_0 : s^2(x) = \theta_s \quad , \quad H_1 : s^2(x) \neq \theta_s, \tag{5.9}$$

con $\theta_s > \theta$ vero.

I differenti valori di h testati, che indicheremo con h_{AR}, utilizzati per il modello AR(1), definito come nella (5.1), sono i seguenti:

$$\begin{aligned} h_{AR} = (&h_1 = 1, h_2 = 1.5, h_3 = 2, h_4 = 2.5, h_5 = 3, \\ &h_6 = 3.5, h_7 = 4, h_8 = 5, h_9 = 7, h_{10} = 10) \end{aligned}$$

In relazione a questo modello, i valori veri di θ sono: $\theta_m = -0.4$ per $x = -0.5$ e $\theta_m = +0.4$ per $x = +0.5$. In riferimento ai test (5.4) e (5.6) i valori di θ utilizzati sono: $\theta_m = -0.55$ e $\theta_m = -0.25$ per $x = -0.5$, $\theta_m = +0.25$ e $\theta_m = +0.55$ per $x = +0.5$.

Si noti che nei grafici vengono riportati solo i valori di h che vantano una certa significatività, per questo motivo il lettore troverà sull'asse delle ascisse non tutti i 10 valori appena elencati, ma esclusivamente quelli che sono stati valutati utili nella formulazione del grafico.

Per il primo modello si riscontrano risultati definibili quasi impeccabili nel caso $n = 1000$. La potenza del test infatti è molto prossima ad 1, mentre contemporaneamente l'errore risulta molto contenuto, non superando mai la soglia di significatività del 0.05.

Risultati del tutto analoghi valgono nel caso $x = +0.5$, che qui omettiamo per non appesantire la trattazione e lo stesso si verifica nel caso gli errori siano distribuiti come una t di Student.

In questo primo caso sembrerebbe non influente la sensibilità al parametro di *bandwidth*, evidenza che invece viene spiegata semplicemente dal fatto che, calcolando il valore vero di h per il modello (5.1) l'h globale, scopriamo che questo è pari a ∞ sia nel caso di errori normali, che nel caso la distribuzione degli errori sia una t di Student. Questo spiega l'andamento dei tre grafici, replicato per ognuna delle ipotesi considerate e per i diversi livelli di *bandwidth* considerati.

Già nell'ipotesi di modello Expar(1) la sensibilità all'h diventa cruciale. Per bassi valori del parametro otteniamo, così come per i casi precedenti, risultati soddisfacenti. Questo ci conferma il caso reale.

Calcolando, infatti, l'h globale ottimale, questo risulta pari a 0.469 per $n = 500$ ed errori normali; $0,409$ per $n = 1000$, stesse impostazioni del caso precedente, figura 5.5 e 5.7; 0.473 per $n = 500$ ed errori distribuiti come una t di Student; $0,412$ per le stesse impostazioni ed $n = 1000$, figura 5.6 e 5.8.

E' interessante notare come allontanandoci dal valore vero, nel caso (c) della figura 5.5 e 5.6, il test tenda a perdere di potenza quando l'h si allontana dal valore ottimale. Lo stesso accade nel caso (a), se le stime vengono effetuate in $x = +0.5$, figura 5.7 e 5.8.

I dieci valori di h impiegati per il modello Expar(1), vengono elencati di seguito:

$$h_{Expar} = (h_1 = 0.3, h_2 = 0.35, h_3 = 0.4, h_4 = 0.45, h_5 = 0.47,$$
$$h_6 = 0.5, h_7 = 0.55, h_8 = 0.6, h_9 = 0.7, h_{10} = 0.8)$$

In relazione a questo modello, i valori veri di θ sono: $\theta_m = -1.66$ per $x = -0.5$ e $\theta_m = +1.66$ per $x = +0.5$. In riferimento ai test (5.4) e (5.6) i valori di θ utilizzati sono: $\theta_m = -2.0$ e $\theta_m = -1.3$ per $x = -0.5$, $\theta_m = +1.3$ e $\theta_m = +2.0$ per $x = +0.5$.

Il caso dell'ARCH è indubbiamente quello in cui è più visibile la sensibilità all'h. Essendo il modello più complesso, ai differenti valori di h impiegati corrispondono differenze sostanziali in termini di *size*, caso (b), e potenza del test, caso (a) e (c). Questo appare chiaro guardando le corrispondenti figure. Soltanto per valori bassissimi di h l'errore risulta contenuto e infatti questo ci viene confermato dai dati reali, in quanto, calcolando l'h globale ottimale, esso risulta

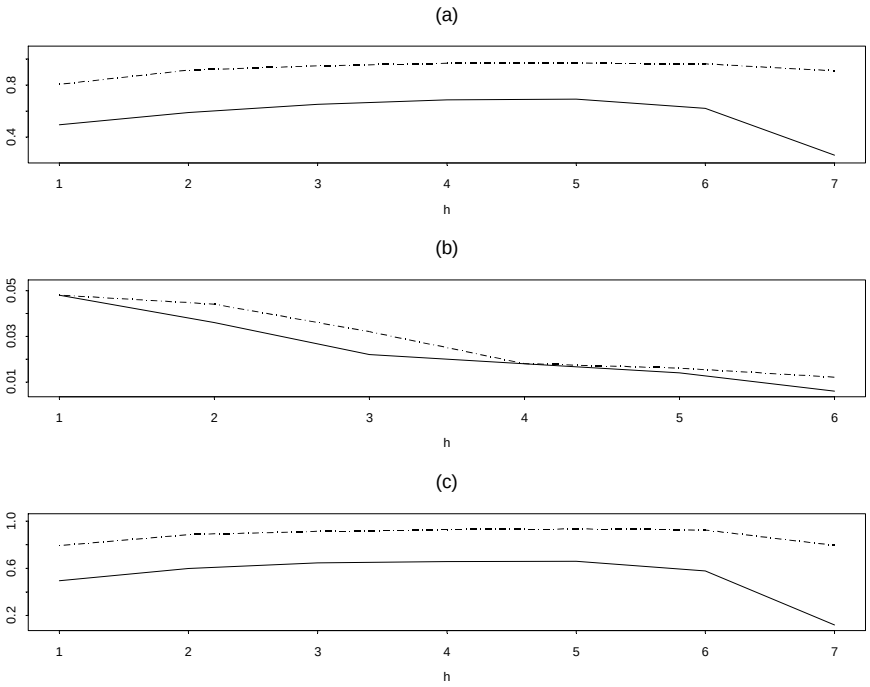

Figura 5.4: Modello AR(1), errori normali, $x = -0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

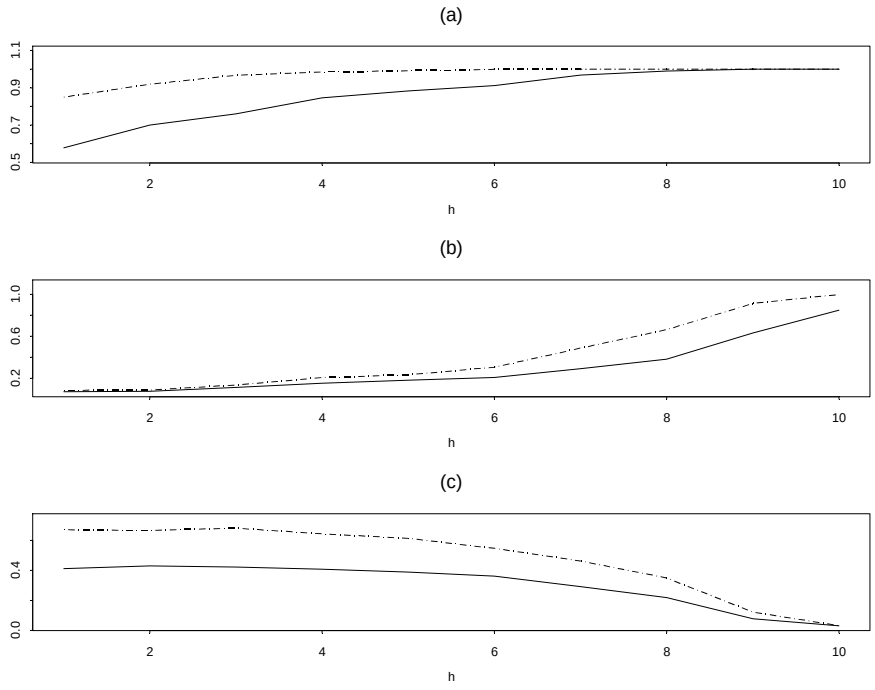

Figura 5.5: Modello Expar(1), errori normali, $x = -0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

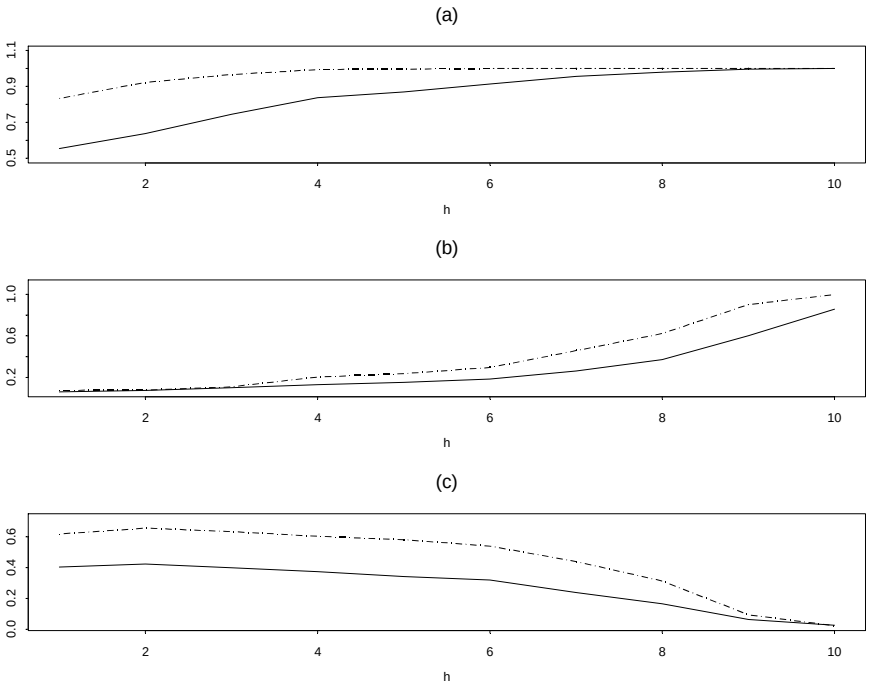

Figura 5.6: Modello Expar(1), errori t di student, $x = -0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

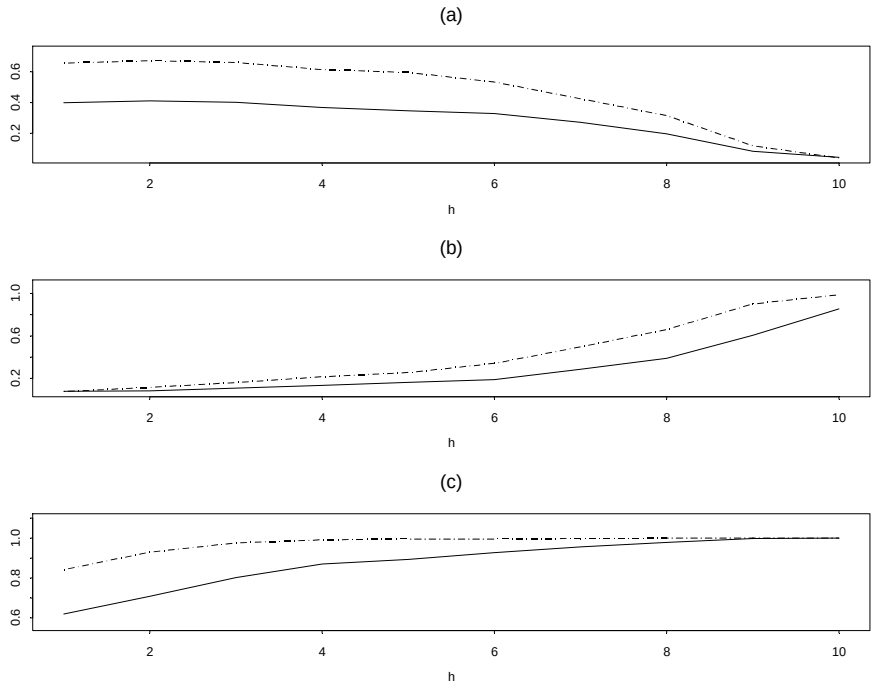

Figura 5.7: Modello Expar(1), errori normali, $x = +0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

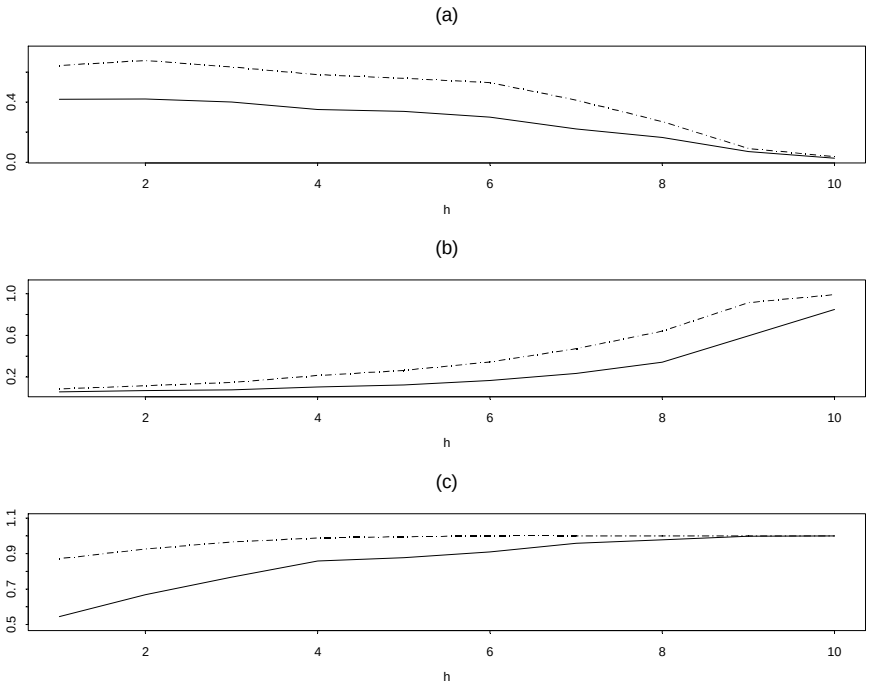

Figura 5.8: Modello Expar(1), errori t di student, $x = +0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

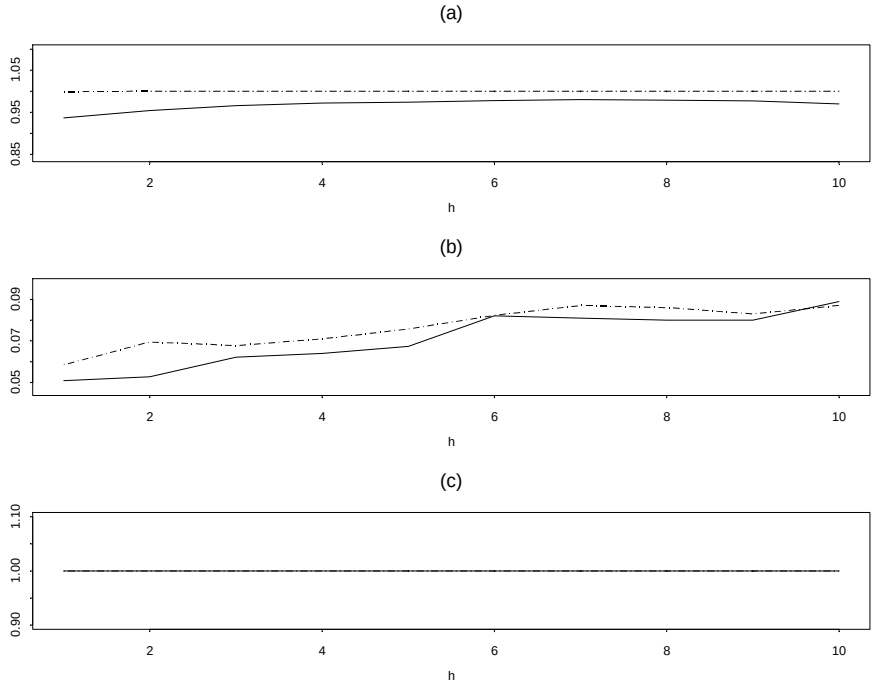

Figura 5.9: Modello ARCH, errori normali, $x = -0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

pari a 0.383 per $n = 500$ ed errori normali; $0,333$ per $n = 1000$, con le stesse impostazioni espresse per il caso precedente, figura 5.9 e 5.10; 0.39 per $n = 500$ ed errori distribuiti come una t di Student; $0,339$ con le stesse impostazioni espresse per il caso precedente ed $n = 1000$, figura 5.11 e 5.12.

Gli h impiegati per l'ARCH sono:

$$h_{ARCH} = (h_1 = 0.3, h_2 = 0.35, h_3 = 0.4, h_4 = 0.45, h_5 = 0.5, \\ h_6 = 0.6, h_7 = 0.7, h_8 = 0.8, h_9 = 0.9, h_{10} = 1)$$

In relazione a questo modello, i valori veri di θ sono: $\theta_s = -0.175$ per $x = -0.5$ e $\theta_s = +0.175$ per $x = +0.5$. In riferimento ai test (5.7) e (5.9) i valori di θ utilizzati sono: $\theta_s = +0.05$ e $\theta_s = +0.1$ per $x = -0.5$, $\theta_s = +0.05$ e $\theta_s = +0.1$ per $x = +0.5$.

Nelle figure 5.9 e 5.10, il grafico (c) presenta una potenza identicamente pari ad 1, in quanto, il valore di θ_s considerato è pari a 0.05. Infatti, basta considerare il valore di $\theta_s = 0.1$ del grafico (a) per avere già una potenza molto vicina ad 1.

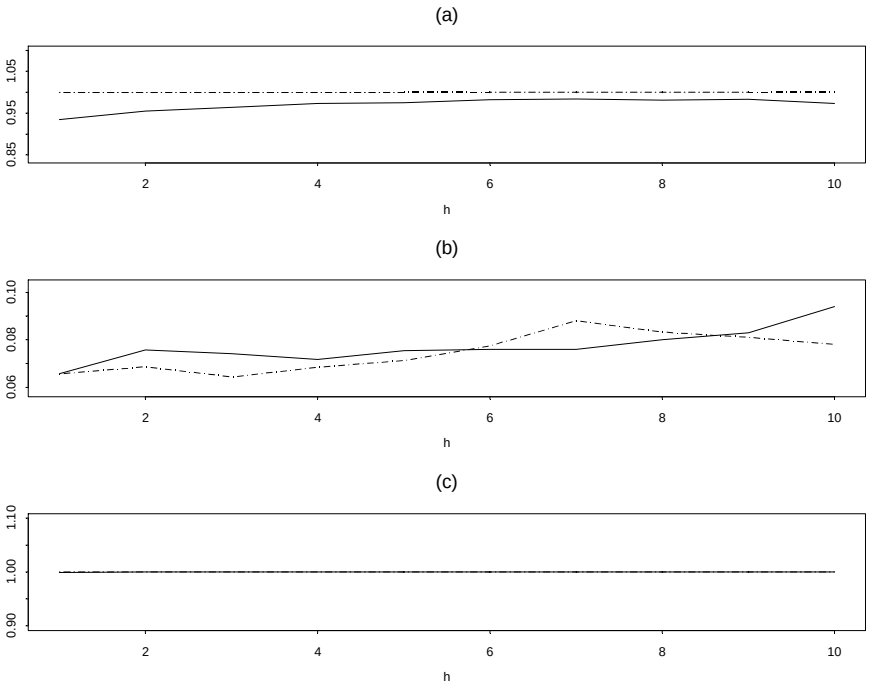

Figura 5.10: Modello ARCH, errori normali, $x = +0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

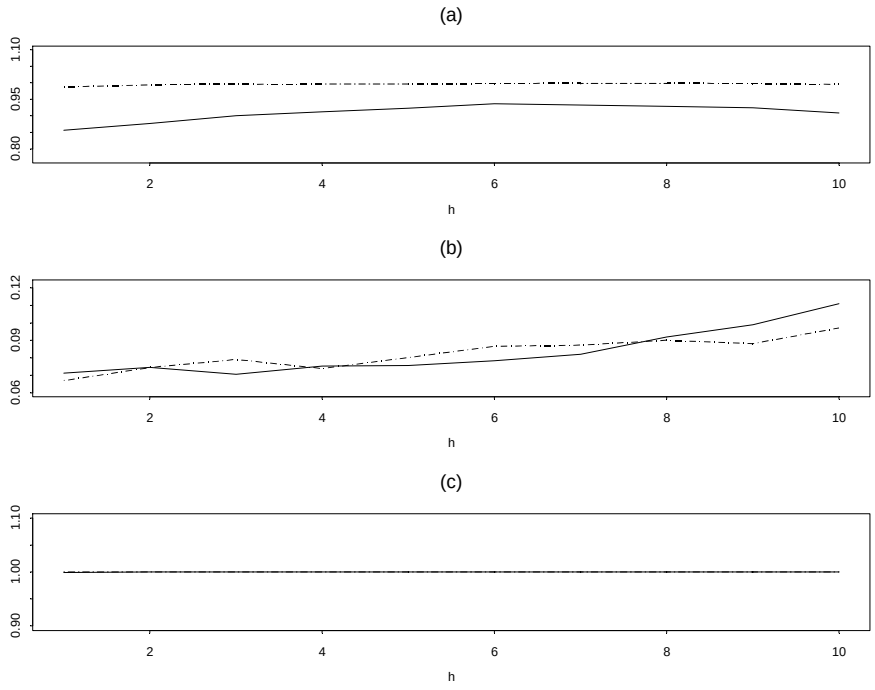

Figura 5.11: Modello ARCH, errori t di Student, $x = -0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

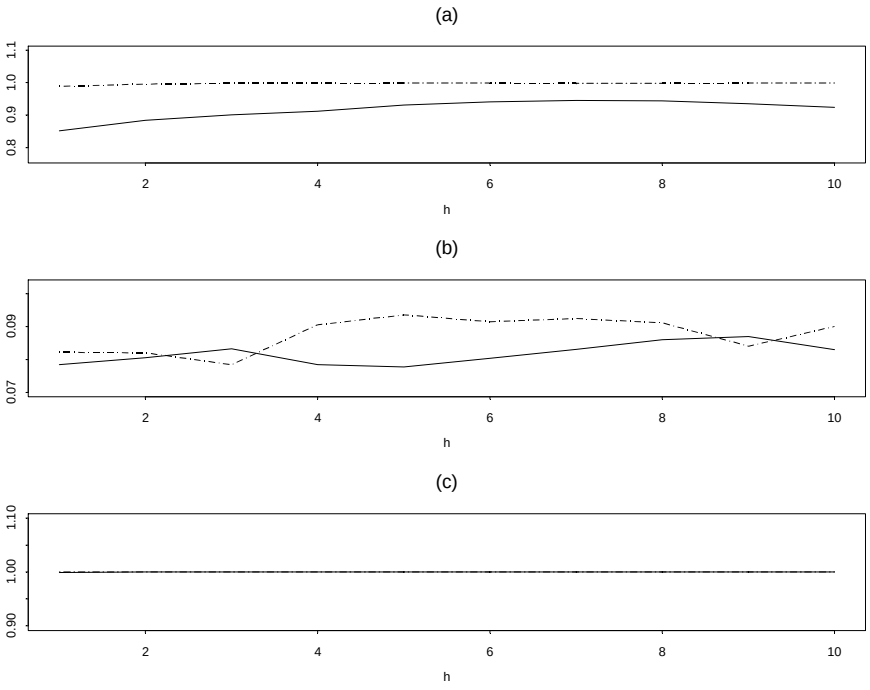

Figura 5.12: Modello ARCH, errori t di Student, $x = +0.5$. I grafici rappresentano rispettivamente, nel caso (a) e (c), la potenza dei test (5.4) e (5.6). Il grafico (b), rappresenta il livello di significatività del test (5.5) per $\alpha = 0.05$. La linea continua rappresenta $n = 500$, quella tratteggiata $n = 1000$.

5.4 Conclusioni

Nelle battute conclusive del nostro lavoro si presenta un problema di trade-off per ognuno dei modelli considerati, che si risolve nello scegliere il giusto compromesso tra una misura del livello di significatività del test che risulti affidabile, senza perdere la possibilità di ottenere una potenza del test che possa tendere al valore ottimale 1.

Vengono riportati numerosi casi di test elaborati tenendo conto di vari aspetti da valutare. Viene preso in considerazione il valore di x in prossimità del suo valore centrato per valori superiori o inferiori ad esso, vengono impostati vari test a seconda dell'ipotesi che veda θ il reale valore espresso dall'ipotesi nulla, un valore inferiore o uno superiore. Infine, viene considerato l'impatto di h, proponendo stime per un numero di 10 valori differenti e verificando cosa succede ai principali parametri di valutazione del nostro test, allontandoci man mano dal valore ottimale dell'h globale.

Per il modello AR(1) gli ottimi risultati conseguiti dal test non risentono effettivamente della scelta dell'h, effetto che potrebbe farci ipotizzare ad una non incidenza del parametro in discussione, pur confermando la validità del test effettuato tramite lo stimatore esaminato durante tutta la trattazione. La vera ragione di qeusto risultato si cela dietro al fatto che, calcolando l'h ottimale per il modello (5.1) ci si accorge che il suo valore vero è ∞.

Questo giustifica la non rilevanza della scelta dell'h rispetto alle performances ottenute.

I casi Expar(1) ed ARCH(1) non soltanto conseguono i risultati attesi dal test, ma mostrano anche in maniera evidente come questi cambiano al variare della determinazione del parametro di *bandwidth*. Nel caso in cui ci si allontani dalla stima ottimale infatti lo stesso test tende a perdere di potenza.

I risultati illustrati confermano empiricamente quanto affermato nel caso teorico e ci forniscono anche evidenze concrete del reale peso del parametro di *bandwidth* per le stime ottenute.

Ringraziamenti

I miei ringraziamenti più sentiti e sinceri vanno a tutti coloro che mi hanno ricordato di credere in me ... proprio quando un pò me ne stavo dimenticando!

Prima di tutto, al prof per avermi trainato con forza quando proprio volevo gettare le redini.

Ad Antonella per l'immancabile, instancabile, insostituibile sostegno pratico e morale.

A Nello, perché quando qualcuno riesce a renderti la vita più bella hai un motivo in più per provarci, per ricordarti di credere in te... per non arrenderti... ma soprattutto perché, in ogni caso, sai che vale la pena andare avanti!

Alle persone che, pure nelle loro difficoltà, mi sono state d'esempio.

Ai miei per la loro persistenza...

Alle mie zie, a quelle che sono nate combattenti ma con un sorriso da pin-up e a quelle dall'affetto sconfinato e materno, capaci di farti sentire che la tua famiglia va ben oltre le mura di casa...

Agli zii 'simpaticoni', quelli che ti fa sempre piacere vedere e che sorridi se sai che a quel compleanno ci sono loro!

Alle nonne, inscindibile legame col passato, che più passa il tempo e meno te ne vuoi separare!

...alle persone che non ci sono più, perché quando qualcuno ti è entrato dentro, può passare una vita intera ma, non è passato un attimo dall'ultimo sorriso d'intesa scambiato!

A Paola, presente in ogni caso, ad ogni ora, per ogni smarrimento o noia, per crucci e capricci tipicamente femminili e per i grattacapi che invece sembrano così insormontabili!

Agli amici che, come lei, anche da lontano sai che ci sono o alle serate divertenti passate in compagnia, quelle che torni a casa e stai bene.

Al contributo di Gelsomina, Giovanna, Kikko e Paolo nelle ultimissime battute, per quelle operazioni 'semplici' che sotto tesi diventano essenziali...una mano, un passaggio, una connessione, un caffè o semplicemente un cornetto!

Alle persone che, anche se non ti conoscono, ti danno il loro aiuto, ti offrono un gesto di supporto, quelle che pensi sempre che dovrebbero essercene di più al mondo, perchè ti danno la speranza di credere nel bene!

...e poi alle persone che con tanta pazienza mi hanno ostacolato, giudicato... perché posso dirvi ancora una volta di *avercela fatta*! =)

L'amore, di qualunque tipo, che sia entusiasmo e passione per le cose che si fanno, sostegno illimitato per qualcuno che si ama più della propria vita, gioia di vivere che inonda ogni pensiero, può solo portare un valore, senza il quale la vita non avrebbe tanto senso di essere vissuta. *Grazie*, insomma, a tutti coloro che questo valore, questo senso, alla mia vita l'hanno portato!

... di cuore, Anna!

Bibliografia

[1] Ango Nze, P. (1992), *Critères d'ergodicité de quelques modèls à représentation markovienne*, Comptes Rendus des Seances de l'Academie des Sciences Paris, 315, 1, 1301-1304.

[2] Bailey, K. R. (1984), *Asymptotic equivalence between the Cox estimator and the general ML estimator of regression and survival parameters in the Cox model*, Annals of Statistics, 12, 730-736.

[3] Berk, R. H. and Jones, D. H. (1979), *Goodness-of-fit statistics that dominate the Kolmogorov statistics*, Z. Wahrsch. Verw. Gebiete, 47, 47-59.

[4] Bosq, D. (1998), *Nonparametric statistics for stochastic progrss: estimation and prediction*, Springer, New York.

[5] Chan, K. S. and Tong, H. (1985), *On the use of deterministic Lyapunov functions for the ergodicity of stochastic difference equations*, Advances in Applied Probability, 17, 666-678.

[6] Chen, S. X. and Qin, Y-S (2000), *Empirical likelihood confidence interval for a local linear smoother*, Biometrika, 87, 946-953.

[7] Chen, S. X. and Härdle, W. and Li, M. (2003), *An empirical likelihood godness-of-fit test for time series*, Journal of Royal Statistical Society, ser. B 65, 663-678.

[8] Chen, S. X. and Van Keilegom, I. (2009), *A goodness-of-fit test for parametric and semi-parametric models in multiresponse regression*, Bernoulli, 4, 955-976.

[9] Cheng, M.-Y. and others (1997), *On automatic boundary corrections*, Annals of Statistics, 25 (4), 1691-1708.

[10] Chernoff, H. (1954), *On the distribution of the likelihood ratio*, Annals of Mathematical Statistics, 25, 573-578.

[11] Chu, C. K. and Marron, J. S. (1991), *Choosing a kernel regression estimator*, Statistical Science, 6, 409-419.

[12] Cleveland, W. S. (1979), *Robust locally weighted regression and smoothing scatterplots*, Journal of American Statistical Association, 74, 829-836.

[13] Cleveland, W. (1988), *Regression by local fitting methods, properties, and computational algorithms*, Journal of Econometrics, 37 (1), 87-114.

[14] Collomb, G. (1984), *Propriétés de convergence presque complète du prédicteur à noyau*, Zeitschrift fur wahrscheinlichkeitstheorie und verwandte gebiete, 66, 441-460.

[15] Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London.

[16] DiCiccio, T. J. and Hall, P. J. and Romano, J. (1988), *Barlett adjustment for empirical likelihood*, Technical Report, No. 298, Dept. Statistics, Stanford Univ.

[17] DiCiccio, T.J. and Hall, P. and Romano, J.P. (1991), *Empirical likelihood is Bartlett-correctable*, Annals of Statistics,19 (2),1053-10.

[18] DiCiccio, T. J. and Romano, J. P. (1989), *On adjustments based on the signed root of the empirical likelihood ratio statistic*, Biometrika, 76, 447-456.

[19] Diebolt, J. and Guégan, D. (1990), *Probabilistic properties of the general nonlinear autoregressive process of the order one*, Technical Report, 128, L.S.T.A Université Paris VI.

[20] Doukhan, P. and Ghindés, M. (1980), *Estimation dans le processus* $X_n+1 = f(X_n) + \varepsilon_n + 1$, Comptes Rendus des Seances de l'Academie des Sciences Paris, ser. A, 297, 61-64.

[21] Doukhan, P. and Ghindés, M. (1981), *Processus autorégressifs nonlinéaires*, Comptes Rendus des Seances de l'Academie des Sciences Paris, ser. A, 290, 921-923.

[22] Engle, R. F. (1982), *Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation*, Econometrica, 50, 987-1008.

[23] Eubank, R. L. and Spiegelman, C. H. (1990), *Testing the goodness of fit of a liner model via non parametric regression techniques*, Journal of the American Statistical Association, 85, 387-392.

[24] Fan, J. (1992), *Design-adaptive nonparametric regression*, Journal of American Statistical Association, 87, 998-1004.

[25] Fan, J. (1993), *Local linear regression smoothers and their minimax efficiency*, Annals of Statistics, 21, 196-216.

[26] Fan, J. and Gijbels, I. (1992), *Variable bandwidth and local linear regression smoothers*, Annals of Statistics, 20 (4), 2008-2036.

[27] Fan, J. and Gijbels, I. (1995a), *Adaptive order polynomial fitting: bandwidth robustification and bias reduction*, Journal of Computational and graphical Statistics, 4 (3), 213-227.

[28] Fan, J. and Gijbels, I. (1995b), *Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation*, Journal of Royal Statistical Society, ser. B, 57, 371-394.

[29] Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall, London.

[30] Fan, J. and Gasser, T. and Gijbels, I. and Brockmann, M. and Engel, J. (1997), *On non-parametric estimation via local polynomial regression*, Annals of the Institute of Statistical Mathematics, 49, 79-99.

[31] Fan, J. and Zhang, J. (2000), *Sieve empirical likelihood ratio tests for non parametric functions*, Research Report, Department of Statistics, Chinese University of Hong Kong, Hong Kong.

[32] Hall, P. (1987), *On the bootstrap and likelihood-based confidence regions*, Biometrika, 74, 481-494.

[33] Hall, P. (1991), *Edgeworth expansions for nonparametric density estimator, with applications*, Statistics, 22, 215-232.

[34] Hall, P. (1992), *On bootstrap confidence intervals in nonparametric regression*, Annals of Statistics, 20, 695-711.

[35] Hall, P. and La Scala, B. (1990), *Methodology and algorithms of empirical likelihood*, International Statistical Review, 58, 109-127.

[36] Härdle, W. and Mammen, E. (1993), *Comparing nonparametric versus parametric versus parametric regression fits*, Annals of Statistics, 21, 1926-1947.

[37] Härdle, W. and Tsybakov, A. (1997), *Local polynomial estimators of the volatility function in nonparametric autoregression*, Journal of Econometrics, 81, 223-242.

[38] Hart, J. (1997), *Nonparametric Smoothing and Lack-of-fit Tests*, Springer, Heidelberg.

[39] Hastie, T. and Loader, C. (1993), *Local regression: automatic kernel carpentry (with discussions)*, Statistical Science, 8 (2), 120-143.

[40] Hjellvik, V. and others (1998), *Linearity testing using local polynomial approximation*, Journal of Statistical Planning and Inference, 68, 295-321.

[41] Hoeffding, W. (1965), *Asymptotically optimal tests for multinomial distributions*, Annals of Mathematical Statistics, 36, 396-401.

[42] Horowitz, J. L. and Spokoiny, V. G. (2001), *An adaotive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative*, Econometrica, 69, 599-631.

[43] Johansen, S. (1978), *The product limit estimator as maximum likelihood estimator*, Scandinavian Juornal of Statistics, 5, 195-199.

[44] Kaplan, E. and Meier, P. (1958), *Nonparametric estimation from incomplete observations*, Journal of American Statistical Association, 53, 457-481.

[45] Kiefer, J. and Wolfowitz, J. (1956), *Consistency of the maximum likelihood estimator in the precence of infinitely many incidental parameters*, Annals of Mathematical Statistics, 27, 887-906.

[46] Kitamura, Y. (1997), *Empirical likelihood methods with weakly dependent processes*, Annals of Statistics, 25, 2084-2102.

[47] Koul, H. L. and Stute, W. (1999), *Nonparametric model checks for time series*, Annals of Statistics, 27, 204-236.

[48] Kreiss, J. and others (1998), *Bootstrap tests for simple structures in nonparametric time series regression*, Discussion Paper, Humboldt-Universität zu Berlin, Berlin.

[49] Lütkepohl, H. (1992), *Introduction to multiple time series analysis*, Springer, Heidelberg.

[50] Neumann, M. H. (1995), *Automatic bandwidth choice and confidence intervals in nonparametric regression*, Annals of Statistics, 23,1937-59.

[51] Masry, E. (1996a), *Multivariate regression estimation: local polynomial fitting for time series*, Stochastic Processes and theri Applications, 65, 81-101.

[52] Masry, E. (1996b), *Multivariate local polynomial regression for time series: uniform strong consistency and rates*, Journal of Time Series Analysis, 17, 571-599.

[53] Masry, E. and Fan, J. (1997), *Local polynomial estimation of regression funtion for mixing processes*, Scandinavian Journal of Statistics, 24, 165-179.

[54] Mokkadem, A. (1987), *Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géometrique*, Journal of time series analysis, 8, 195-204.

[55] Monti, A. C. (1997), *Empirical likelihood confidence regions in time series models*, Biometrika, 84, 395-405.

[56] Opsomer, J. D. (1997), *Nonparametric regression in the presence of correlated errors*, In *Modelling longitudinal and spatially correlated data: methods, applications and future directions*, Sringer, New York, 339-348.

[57] Owen, A. B. (1985), *Nonparametric likelihood ratio confidence intervals*, Tecnical Report LCS 6, Dept. Statistics, Standford Univ.

[58] Owen, A.B. (1988), *Empirical likelihood ratio confidence intervals for a single functional*, Biometrika, 75, 237-249.

[59] Owen, A.B. (1990), *Empirical likelihood confidence regions*, Annals of Statistics, 18, To appear.

[60] Robinson, P. M. (1983), *Nonparametric estimator for time series*, Journal of Time Series Analysis, 4, 185-207.

[61] Robinson, P. M. (1984), *Robust nonparametric autoregression*, In *Robust and nonlinear time series analysis*, Springer, Heidelberg.

[62] Ruppert, D. and Wand, M. P. (1994), *Multivariate weighted least squares regression*, Annals of Statistics, 22, 1346-1370.

[63] Ruppert, D. and others (1995), *An effective bandwidth selector for local least squares regression*, Journal of the American Statistical Association, 90 (432), 1257-1270.

[64] Stone, C. J. (1977), *Consistent nonparametric regression*, Annals of Statistics, 5, 595-645.

[65] Thomas, D. R. and Grunkemeier, G. L., (1975), *Confidence interval estimation of survival probabilities forcensored data*, Journal of American Statistical Association, 70, 865-871.

[66] Tripathi, G. and Kitamura, Y. (2000), *On testing conditional moment restictions: the canonical case*, Research Report, Department of Economics-Madison, Madison.

[67] Tsybakov, A. B. (1986), *Robust reconstruction of functions by the local-approximation method*, Problems of Information Transmission, 22, 133-146.

[68] Tusnady, G. (1977), *On asymptotically optimal tests*, Annals of Statistics, 5, 385-393.

[69] Vardi, Y. (1985), *Empirical distributions in selection bias models*, Annals of Statistics, 13, 178-203.

[70] Vieu, P. (1995), *Order choice in noonlinear autoregressive models*, Discussion Paper, Laboratoire de Statistique et Probabilités, Université Toulouse, Toulouse.

[71] Vilar-Fernández, J. A. and Vilar-Fernández, J. M. (1998), *Recursive estimation of regression functions by local polynomial fitting*, Annals of the Institute of Statistical Mathematics, 50 (4), 729-754.

[72] Vilar-Fernández, J. M. and Vilar-Fernández, J. A. (2000), *Recursive local polynomial regression under dependence conditions*, Test, 9 (1), 209-233.

[73] Wilks, S.S. (1938), *The large-sample distribution of the likelihood ratio for testing composite hypoteses*, Annals of Mathematical Statistics, 9, 60-62.

[74] Xia,Y. and Li, W. K. (2002), *Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting*, Journal of Multivariate Analysis, 83 (2), 265-287.