UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

" METODI BOOTSTRAP PER SERIE STORICHE "

RELATORE: PROF. SSA BISAGLIA LUISA

LAUREANDO: CESA ROBERTO

MATRICOLA N. 1160933

ANNO ACCADEMICO 2019 / 2020

"Il/La candidato/a, sottoponendo il presente lavoro, dichiara, sotto la propria personale responsabilità, che il lavoro è originale è che non stato già sottoposto, in tutto in parte, dal/dalla candidato/a o da altri soggetti, in altre Università italiane o straniere ai fini del conseguimento di un titolo accademico. Il/La candidato/a dichiara altresì che tutti i materiali utilizzati ai fini della predisposizione dell'elaborato sono stati opportunamente citati nel testo e riportati nella sezione finale 'Riferimenti bibliografici' e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo al documento originale"

ABSTRACT

Dall'inizio del mio percorso accademico, la teoria economica affrontata era spesso fondata su forti assunzioni che, alle volte, potevano non rispecchiare il reale comportamento dei processi economici. Da qui nasce il mio interesse verso la grande gamma di metodi non parametrici di cui fanno parte i metodi boostrap.

La prima idea di bootstrap nasce nel 1979 con la pubblicazione da parte di Bradley Efron del paper scientifico "Bootstrap methods: another look at the jackknife". Queste tipologie di procedure vengono incluse nella classe più ampia dei metodi di ricampionamento e vengono utilizzate con lo scopo di effettuare inferenza statistica senza fare affidamento su ipotesi troppo stringenti.
Ideati esclusivamente per un' applicazione su campioni di osservazioni indipendenti, i metodi bootstrap trovano il loro limite davanti a processi di dati correlati fra loro come le serie storiche. Nei decenni successivi alla prima pubblicazione di Efron vengono, perciò, sviluppati ulteriori versioni di bootstrap in grado di far fronte a questo problema e capaci di effettuare inferenza statistica anche laddove i processi osservati fossero caratterizzati da dipendenza.
L'intenzione di questo elaborato è quello di riassumere i metodi bootstrap che hanno ottenuto maggior successo nell'applicazione alle serie storiche, senza addentrarmi in tecnicismi che richiederebbero conoscenze matematiche molto avanzate, con lo scopo di descrivere i principi su cui si fondano e il loro funzionamento.
Inizierò presentando il concetto di bootstrap così come inteso da Efron (1979) per poi presentare le metodologie sviluppate con riferimento alle serie storiche.
Dopo aver presentato l'utilizzo del bootstrap nello studio di modelli parametrici come i modelli autoregressivi (AR), approfondirò il concetto di wild bootstrap e di AR-sieve bootstrap. Infine, mi dedicherò allo studio del metodo bootstrap a blocchi e di una sua generalizzazione chiamata "blocks of blocks bootstrap" seguendo il modello presentato in Politis e Romano (1992a).

INDICE

CAPITOLO 1. INTRODUZIONE AI METODI BOOTSTRAP

1.1 L'origine

Nel Cambridge Dictionary di lingua inglese il termine "bootstrap" è definito come: "To improve your situation or become more succesful, without help from others or without advantages that others have", ovvero migliorare la propria situazione facendo ricorso solo alle proprie risorse. È esattamente questo il principio sui cui si basano i metodi che andrò ad approfondire nel corso di questa ricerca.
La maggior parte dei modelli ideati per fare inferenza statistica è spesso fondata su assunzioni molto forti come l'assunzione di normalità o di varianza costante che, alle volte, possono portare a stime poco accurate laddove il processo sottostante non rispecchi le date ipotesi. Durante il ventesimo secolo, sono molti gli studiosi che si sono cimentati nella ricerca di metodi non-parametrici (ovvero senza ricondurre la popolazione a specifici modelli) allo scopo di fare inferenza statistica. È da questa necessità che nella seconda metà degli anni novanta iniziano gli studi dei cosiddetti metodi bootstrap.
Inclusi tra le tecniche di "statistica non-parametrica", i metodi bootstrap fanno parte di un insieme più ampio di metodi accomunati sotto il nome di "resampling methods". Il primo ad utilizzare la dicitura bootstrap fu Bradley Efron nel 1979, nel suo paper *" Another look at the Jackknife "* pubblicato nella rivista *"Annals of Statistics"* (Efron, 1979). Nonostante queste procedure di ri-campionamento fossero già popolari all'epoca (i.e. *jackknife* e *delta method*), esse venivano utilizzate principalmente per lo studio della varianza e della distorsione delle statistiche d'interesse e spesso si dimostravano inadeguate se applicate a campioni numerosi (Chernick- LaBuddle, 2011). Il primo passo di Efron fu quello di definire un metodo più generale del suo predecessore *jackknife* e di ampliarne gli utilizzi a una gamma più ampia di problemi come la stima degli intervalli di confidenza. Inoltre, in Efron (1979) vengono dimostrati il successo e la maggior efficienza dei metodi bootstrap in campi dove i procedimenti precedenti risultavano inadeguati.

1.2 Il funzionamento

Lo scopo dei metodi bootstrap è quello di fare inferenza statistica sul modello identificato senza effettuare ipotesi a priori e affidarsi a specifiche formule. In particolare queste tecniche permettono di stimare la distribuzione di un determinato parametro della popolazione sulla

base dei dati osservati. È stato dimostrato (si veda, per esempio, Hesterberg, 2011) che l'efficacia di questi metodi può essere alle volte di molto superiore a quella di procedure fondate su forti assunzioni. Il procedimento che li accomuna è il seguente:

- Definiamo con F la distribuzione di una popolazione qualsiasi dalla quale viene estratto un campione di dati indipendenti e identicamente distribuiti di ampiezza *n.* Tramite questo campione viene ricavata una stima del parametro di interesse θ (che chiameremo "T(F)") e una stima della distribuzione della popolazione (che chiameremo " F_n "). In genere la stima della distribuzione di una popolazione viene effettuata tramite la funzione di ripartizione empirica, ovvero la distribuzione cumulativa che si ottiene assegnando ad ogni osservazione del campione la stessa probabilità.
  Nel caso in cui la distribuzione della popolazione fosse nota, non è necessario effettuarne una stima, ma basterà prelevare campioni dalla distribuzione conosciuta (bootstrap completamente parametrico).

- Nel passaggio seguente avviene la simulazione del modello studiato nella dimensione bootstrap. Durante questa fase F_n viene trattata come la vera distribuzione della popolazione. A partire da F_n vengono, dunque, estratti casualmente *n* elementi per formare un nuovo campione. Questa procedura viene ripetuta molteplici volte ottenendo, così, una serie di campioni chiamati "campioni bootstrap". Per ognuno di questi campioni viene calcolato il valore del parametro di interesse (stima bootstrap).
  Il totale delle stime bootstrap effettuate andrà a comporre un istogramma di valori che chiameremo distribuzione bootstrap o "F_n^* ". Quest'ultima sarà un'approssimazione della reale distribuzione campionaria del parametro di interesse. Da F_n^* verrà, quindi, ricavato il valore bootstrap della statistica ($T(F^*_n)$).

È importante sottolineare che il centro della distribuzione bootstrap non è una stima accurata del centro della distribuzione campionaria. Nonostante la distribuzione bootstrap rispecchi l'estensione, la distorsione e l'asimmetria di quest'ultima, esse sono centrate su due valori diversi della statistica. Infatti, mentre la prima è centrata sul valore campionario della stessa, la seconda è centrata sul valore reale. È chiaro, dunque, che il bootstrap non è da considerare un metodo per ottenere una stima migliore del parametro, bensì, come strumento utile a quantificare il comportamento dello stesso (i.e standard error, distorsione, intervalli di

confidenza …) e poterne, di conseguenza, valutare l'accuratezza senza l'utilizzo di formule specifiche fondate su assunzioni forti come quella di normalità della popolazione.
Per rendere più chiaro il concetto, i metodi bootstrap possono essere pensati come una simulazione del nostro modello in una dimensione parallela dove il campione di osservazioni viene considerato come la popolazione.
Il vantaggio che ci forniscono queste procedure è il seguente: estrarre una serie di campioni dalla popolazione reale può essere molto costoso e può richiedere molto tempo, inoltre, essendo il parametro di interesse ignoto, non c'è modo sicuro di valutare l'accuratezza del modello.
Nella dimensione bootstrap, invece, eseguire una campionamento non ha praticamente alcun costo. Oltre a questo, essendo la distribuzione della popolazione stata stimata dal ricercatore, il valore del parametro nella dimensione bootstrap è conosciuto. È possibile, quindi, applicare il modello per diverse volte e avere un'idea molto più precisa del comportamento del parametro.

Ciò nonostante, anche l'estrazione di campioni bootstrap può richiedere molto tempo e potenza computazionale. Consideriamo, per esempio, un campione di *n* elementi da una popolazione qualsiasi. Tramite la procedura del ricampionamento, estraiamo con reinserimento da tale campione *n* elementi per B volte dove B indica il numeri di campioni bootstrap che vogliamo ottenere. Il numero massimo di combinazioni (ovvero quello che ci fornirebbe la distribuzione boostrap più accurata) è dato dalla formula:

$$\left(\frac{2n-1}{n}\right)=\frac{(2n-1)!}{n!(n-1)!}$$

Di conseguenza, basterebbe anche una dimensione campionaria pari a 50 per ottenere un numero enorme di combinazioni, difficile da gestire addirittura da un computer. Per questo motivo, spesso, la distribuzione bootstrap viene approssimata tramite una simulazione Monte Carlo.

1.3 L'accuratezza della distribuzione bootstrap

Una volta compreso il funzionamento della metodologia bootastrap, è necessario comprendere in quali situazioni essa possa essere considerata affidabile. Le condizioni che esporrò in seguito si basano sulla legge dei grandi numeri e spiegano come questi metodi, all'aumentare della numerosità campionaria, possano essere considerati accurati.

Prima di tutto è necessario che l'approssimazione Monte Carlo rispecchi la vera distribuzione bootstrap. È stato dimostrato che nella maggior parte dei casi un ordine di grandezza di 10^4 è sufficiente per considerare l'approssimazione accurata (Hesterberg,2011).

In secondo luogo è richiesto che la distribuzione bootstrap approssimi quello che è il vero comportamento della statistica ricercata, e che, dunque, rispecchi la distribuzione reale della stessa.
Perché ciò avvenga, sono richieste due condizioni:

1) Come prima cosa la distribuzione campionaria, dato il modello stimato, deve essere consistente e, di conseguenza, essere centrata sul valore reale del parametro. È richiesto che il valore campionario della statistica tenda a quello reale, al tendere di n all'infinito ($T(F_n) \rightarrow T(F)$ con $n \rightarrow \infty$). In genere, nel caso in cui la distribuzione stimata sia empirica e le osservazioni siano i.i.d. , possiamo considerare questa condizione come rispettata grazie al teorema Glivenko-Cantelli (vedi Chernick- LaBuddle, 2011).

2) In secondo luogo è necessario che la distribuzione bootstrap sia consistente rispetto alla distribuzione campionaria. Di conseguenza il valore bootstrap del parametro deve tendere a quello campionario, al tendere di n all'infinito ($T(F_n^*) \rightarrow T(F_n)$ con $n \rightarrow \infty$).

Diversi studi sulle proprietà asintotiche del bootstrap di Efron possono essere trovati in Singh (1981) e Bickel e Freedman (1981). La verifica di queste condizioni asintotiche va oltre lo scopo di questa ricerca.

Data la grande utilità e versatilità, i metodi bootstrap trovano applicazione in moltissimi campi. Nel mio lavoro di tesi mi dedicherò principalmente al loro utilizzo nell'analisi di serie storiche.

CAPITOLO 2. I METODI BOOTSTRAP APPLICATI ALLE SERIE STORICHE

2.1 Introduzione

La metodologia bootstrap introdotta da Efron costituisce un ottimo strumento se applicata a un campione di variabili casuali indipendenti e identicamente distribuite. In questo caso il metodo del ricampionamento funziona correttamente ed è in grado di fornire informazioni su ogni aspetto del comportamento del parametro.
D'altro canto, possono sorgere ostacoli nel momento in cui vi sia una dipendenza tra i dati del campione come nel caso delle serie storiche.
I primi approfondimenti a riguardo vennero affrontati da Singh (1981) che, oltre ad eseguire uno studio sulle proprietà asintotiche del bootstrap, dimostra l'invalidità del metodo di Efron in caso di dipendenza nei dati della popolazione.
Il procedimento di stima della popolazione teorizzato da questo ultimo e il consecutivo ricampionamento casuale per la costruzione di campioni bootstrap non possono essere efficaci in processi affetti da correlazione. L'estrazione di singole osservazioni in modo casuale non potrebbe in alcun modo mantenere la correlazione che vi è tra i dati e porterebbe, dunque, a risultati inaffidabili.
Di conseguenza, è stato necessario per gli studiosi sviluppare implementazioni di questi metodi bootstrap, con lo stesso principio, ma con la capacità di conservare durante il procedimento di ricampionamento la relazione di dipendenza fra i dati.

Negli anni sono state molte le versioni del bootstrap studiate per permetterne l'applicazione alle serie storiche. La sezione si aprirà con un approfondimento del bootstrap basato sui modelli, ovvero una metodologia semi-parametrica (in quanto si basa sull'assunzione che il processo rispetti un determinato modello) fondata sull'idea del bootstrap dei residui di una regressione (Wu, 1986).
In seguito descriverò quelli che sono stati due dei metodi di maggior successo tra i metodi bootstrap non parametrici nell'applicazione alle serie storiche, l'autoregressive-sieve bootstrap e il bootstrap a blocchi.

2.2 Bootstrap basato sui modelli

Non appena il bootstrap venne introdotto, gli studiosi iniziarono a interrogarsi su una sua possibile applicazione al di fuori di contesti contenenti dati i.i.d.. Pochi anni dopo la pubblicazione del lavoro di Efron nel 1979, vennero effettuati diversi studi (tra i primi ricordiamo Efron-Tibshirani 1993) riguardanti vari aspetti delle metodologie bootstrap, tra cui versioni adattate allo studio delle serie storiche.

La prima procedura che venne studiata, nonché la più intuitiva, fu quella di applicare il bootstrap (e in particolare il ricampionamento) sui residui di un autoregressione, assumendo che tali residui si distribuissero come una serie di valori indipendenti e identicamente distribuiti. In questo caso, dunque, l'ipotesi iniziale è che la popolazione sottostante si comporti come un processo autoregressivo a media mobile (ARMA), invertibile (quindi esprimibile come un processo AR) e stazionario.
Ipotizziamo per semplicità di studiare un modello AR (p) a media zero:

$$X_t = \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + e_t \ , \ t = 1, 2, ..., n$$

dove i β_i, i = 1, 2, ... , p sono i coefficienti autoregressivi mentre e_t , t = 1, 2, … , n , ovvero i residui, sono una successione di variabili casuali indipendenti e identicamente distribuite con media zero. Affinché il modello rispetti queste caratteristiche, i coefficienti devono seguire certe condizioni. In particolare, le radici dell'equazione caratteristica associata al polinomio autoregressivo $\Psi_p(z)$ devono essere tutte maggiori di 1 in valore assoluto.

$$\Psi_p(z) = 1 - \beta_1 z - \beta_2 z^2 - ... - \beta_p z^p$$

Nel caso in cui una o più delle radici dell'equazione $\Psi_p(z) = 0$ siano uguali a uno il processo non è più definito stazionario. Laddove, invece, vi sia la presenza di radici minori di uno il modello è definito esplosivo. Per il momento assumiamo che le radici rispettino tali ipotesi. Lo studio dei metodi bootstrap in processi esplosivi o instabili (ovvero contenenti radici unitarie) richiede conoscenze matematico-statistiche elevate e non verrà affrontato in questo elaborato.

Definiamo

- $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$ una serie storica di dimensioni *n*.
- $\hat{m}_n(X_{t-1},\ldots,X_{t-p})$ lo stimatore parametrico di $E[X_t \mid X_{t-1}, X_{t-2}, \ldots, X_{t-p}]$ con *p* ordine del modello autoregressivo.

Per semplicità considereremo lo stimatore $\hat{m}_n$ come totalmente parametrico. Di conseguenza esso potrà essere semplificato a $\hat{m}_n(X_{t-1},\ldots,X_{t-p}) = \sum_{i=1}^{p} \hat{\beta}_k X_{t-k}$ dove $\hat{\beta}_k$ sono le stime dei parametri autoregressivi basati sul campione $\mathbf{X}_n$.

Prima di tutto viene stimato il modello autoregressivo. Da quest'ultimo è possibile ricavare *n-p* residui stimati.

$$\hat{e}_t = X_t - \sum_{i=1}^{p} \hat{\beta}_k X_{t-k} \quad ,t=p+1,\ldots,n$$

Come detto in precedenza questi termini di errore sono considerati realizzazioni di variabili casuali indipendenti e identicamente distribuiti con media uguale a zero. Non essendoci alcuna correlazione fra questi ultimi e il valore osservato X_t è possibile applicare le procedure sviluppate da Efron. I residui fungono da campione da cui viene effettuato il ricampionamento.

Mantenendo, dunque, lo stesso modello (lo stesso stimatore $\hat{m}_n$) è possibile generare la serie bootstrap $X^*_1, \ldots, X^*_n$ dove

$$X^*_t = \sum_{i=1}^{p} \hat{\beta}_k X^*_{t-k} + e^*_t \quad ,t=1,\ldots,n$$ con e_t^* estratto casualmente dai residui campionati.

Nonostante la maggior parte degli autori concordino su questi procedimenti, vi è una varietà di interpretazioni sulla scelta dei primi *p* valori $X^*_{1-p}, \ldots, X^*_0$ in quanto non possono essere stimati da valori precedenti. In diverse pubblicazioni la serie bootstrap viene stimata a partire dal valore X^*_{p+1} , mentre i primi *p* valori vengono posti uguali ai dati originali (osservati). D'altra parte diversi studiosi ritengono che il metodo migliore sia quello di porre $X^*_j = 0$ (ovvero la media che è stata assunta) per $j \leq -p$ ed eseguire la stima vista in precedenza per il resto dei valori.

È dimostrato, tuttavia, che l'effetto di questa scelta sulle proprietà asintotiche del bootstrap è trascurabile (Kreiss e Franke, 1989).

2.2.1 L'accuratezza del bootstrap model-based

Nel caso in cui la sequenza $\{X_n\}_{n \geq 1}$ rispecchi le condizioni poste all'inizio della sezione 2.2, il bootstrap applicato in precedenza risulta uno strumento estremamente affidabile.
È stato dimostrato in Bose (1988) che, nel caso in cui il processo sia stazionario, la distribuzione bootstrap approssima la distribuzione campionaria di uno stimatore dei minimi quadrati standardizzato in modo più accurato di una distribuzione normale.
Allo stesso modo, stimando i vari parametri tramite le equazioni di Yule-Walker otterremo sempre modelli stabili e causali nella procedura di bootstrap (Kreiss e Lahiri, 2012).
Laddove, invece, vi sia una condizione di non-stazionarietà il bootstrap tende a essere poco accurato ed è necessario uno studio più approfondito per la sua valutazione.

Nel caso in cui la serie storica sia effettivamente una realizzazione del modello che abbiamo ipotizzato, allora non vi sarà alcun problema e il bootstrap funzionerà correttamente. Ciò nonostante, parliamo di un'assunzione molto vincolante. Non vi è modo di verificare che la struttura del processo rispecchi esattamente quella di un processo AR(p) e, inoltre, non vi è modo di identificare il vero valore dell'ordine p. Il bootstrap basato sui modelli, difatti, porterà a uno stimatore consistente nel caso in cui il processo segua questa forma:

$$X_t = m_\theta(X_{t-1}, \dots, X_{t-p}) + e_t \quad ,t \in \mathbb{Z}$$

Nel caso in cui il modello proposto non rispecchi correttamente il processo osservato, il bootstrap parametrico potrà portare, comunque, a uno stimatore consistente laddove la distribuzione asintotica del parametro di interesse non vari passando dal processo reale a un processo del tipo $X_t = m_\theta(X_{t-1}, \dots, X_{t-p}) + e_t \quad ,t \in \mathbb{Z}$.

2.2.2 Wild bootstrap

L'idea di Efron e dei suoi colleghi nasce dall'assunzione che i residui stimati siano indipendenti e identicamente distribuiti e che, dunque, abbiano varianza costante, ovvero, che

siano caratterizzati da omoschedasticità. Assunta questa ipotesi per vera, un loro ricampionamento casuale non porterà a distorsioni nel modello.
Nel caso in cui questa assunzione cada e vi sia, quindi, presenza di eteroschedasticità, non sarà più possibile estrarre i residui in modo casuale in quanto il modello risulterebbe impreciso. La soluzione a questo problema risiede nel cosiddetto wild bootstrap.

Introdotto inizialmente da Wu (1986), tale metodologia prevede che i termini di errori applicati al modello bootstrap vengono stimati nel seguente modo:

$$e^*_t = \hat{e}_t \cdot \eta^*_t \quad ,t = p+1,...,n$$

dove η^*_t è una variabile casuale con media uguale a zero e varianza unitaria, mentre $\hat{e}_t$ è un valore estratto casualmente dagli *n-p* residui campionati.
Durante gli anni il metodo è stato proposto in diverse varianti ed è stato dimostrato essere un efficace rimedio all'eteroschedasticità.
Ciò che distingue le proposte dei vari autori riguarda la distribuzione che dovrebbe assumere questa variabile.
Una delle soluzioni più accreditate è quella proposta da Mammen (1993) dove, oltre alle due condizioni su media e varianza esposte qui sopra, viene richiesta anche l'unitarietà del momento terzo $E[\eta^{*3}_t] = 1$.
In questo caso la variabile viene posta uguale a $(1+\sqrt{5}) / 2$ con probabilità $p_1 = (\sqrt{5}-1) / 2\sqrt{5}$
e uguale a $(1-\sqrt{5}) / 2$ con probabilità $p_1 = (\sqrt{5}+1) / 2\sqrt{5}$.

2.2.3 Bootstrap residual based per stimatori non-parametrici

Nel caso in cui lo stimatore $\hat{m}_n(X_{t-1},...,X_{t-p})$ fosse non-parametrico, le proprietà della serie bootstrap generata potrebbero risultare complicate da investigare.
Tramite stimatori non-parametrici, infatti, è possibile effettuare una stima accurata della struttura del processo solo nelle regioni di dati non lontane, dal punto di vista temporale, dai dati osservati nel campione estratto. Di conseguenza effettuare uno studio dello stimatore non parametrico in zone di dati lontane da quelle da noi osservate porterebbe probabilmente a stime imprecise e la stabilità del processo bootstrap non potrebbe essere garantita.

Ciò nonostante, affinché la consistenza asintotica della procedura bootstrap possa essere confermata, è necessario che la serie bootstrap generata risulti stabile.
Una soluzione a questo problema (Kreiss e Lahiri, 2012) può essere ottenuta generando le osservazioni bootstrap nel seguente modo:

$X_t^* = \hat{m}_n(X_{t-1}, \ldots, X_{t-p}) + e^*_t$,t=1,..,n con e_t^* estratto casualmente dai residui campionati.

Così facendo non si otterrà più una serie storica nella dimensione bootstrap dal momento che ogni valore verrà stimato tramite le osservazioni del processo reale. Ciò nonostante, questo procedimento permetterà alla serie bootstrap di essere stabile in quanto rispecchierà la distribuzione marginale di *p* dimensioni del processo reale per costruzione.

2.3 Autoregressive-sieve bootstrap

Le procedure viste in precedenza sono molto sensibili a problemi di non corretta specificazione del modello. Nel caso in cui quest'ultimo sia specificato erroneamente (parlando di processi autoregressivi, un ordine *p* diverso da quello reale), anche le stime bootstrap saranno distorte e inaffidabili. Inoltre, l'affidarsi a modelli prestabiliti va contro quello che era la causa di maggior successo del bootstrap: la completa non parametricità.
La teoria dell'autoregressive-sieve bootstrap si basa sul più famoso "method of sieves", un metodo utilizzato, appunto, per applicare approcci parametrici a problemi non parametrici. L'idea di fondo è quella di stimare un processo non parametrico e di comportamento ignoto tramite un modello parametrico che diventa più accurato e affidabile al crescere della dimensione campionaria. Per un approfondimento è possibile fare riferimento a Grenander (1981) e Geman e Hwang (1982).

Il primo ad introdurre la procedura nello specifico fu Bühlmann (1997), proponendo l'autoregressive-sieve bootstrap come alternativa ai metodi proposti fino a quel momento.
Il procedimento si basa sull'applicazione del bootstrap ai residui di un'autoregressione come discusso nella sezione 2.2. Ciò che differenzia questo metodo da quello basato sui modelli è la totale non parametricità delle assunzioni. In questo caso, infatti, il processo sottostante non è considerato essere un AR(p) di dato ordine *p* stabilito a priori, bensì è ipotizzato essere di ordine infinito. Questo rende il procedimento di più generale applicazione e di minor sensibilità a errori di specificazione del modello.

Si ipotizzi che $\{X_i\}_{i\in\mathbb{Z}}$ sia un processo stazionario, ergodico ed invertibile, ovvero esprimibile sotto forma di modello autoregressivo (il procedimento potrebbe essere applicato anche a processi MA(∞), ma risulterebbe più lento e complicato).

$$X_t - \mu = \sum_{i=1}^{\infty} \beta_i (X_{t-i} - \mu) + e_t \quad , \quad t \in \mathbb{Z}$$

Si definisca P come la distribuzione congiunta di tale processo.
Nel caso in cui i dati fossero i.i.d. con distribuzione comune F, sarebbe sufficiente stimare $\hat{P}_n = \hat{F}_n^{\infty}$ dove F viene calcolato a partire dal campione. Nel caso di dipendenza fra i dati, invece, tale procedimento non è applicabile in quanto la distribuzione congiunta P non è definibile come un semplice prodotto delle singole distribuzioni.
Come soluzione al problema appena esposto, Bühlmann propone il seguente procedimento.

È data una serie storica di n osservazioni $\boldsymbol{X}_n$ $(X_{t-n,\dots,}X_t)$ che si assume sia una realizzazione di un processo autoregressivo di ordine infinito. L'idea su cui si fonda il metodo è quella di approssimare un processo ignoto con una sequenza di modelli parametrici.
Per stimare il modello Bühlmann esegue una vera e propria approssimazione del processo tramite un AR(*p*), dove *p* non è scelto a priori, ma è una funzione di *n*, con *n* dimensione campionaria ($p = p(n)$). Ciò che è richiesto è che *p* tenda all'infinito e che $p(n)/n$ tenda a zero al tendere all'infinito di *n* ($p(n) \to \infty$, $p(n) / n \to 0$ con $n \to \infty$).
Qui di seguito descriverò i passi dell'algoritmo del metodo AR-sieve bootstrap seguendo la procedura proposta da Kreiss e Lahiri (2012).

1) STIMA DEL MODELLO
Una volta identificato l'ordine *p*, è il momento di stimare il modello e applicarlo ai dati. I vari studiosi che hanno affrontato l'argomento concordano sul fatto che l'utilizzo delle equazioni di Yule-Walker sia il metodo più conveniente. Difatti, oltre alla semplicità e velocità di esecuzione della tecnica, esse garantiscono la stazionarietà e la causalità del processo bootstrap (Kreiss, Lahiri 2012). Tramite queste ultime, dunque, è possibile ottenere una stima dei coefficienti di autoregressione.

Date le assunzioni viste in precedenza, si può affermare che la matrice quadrata delle autocovarianze $\mathbf{\Gamma}_p$ è definita positiva (sarà, quindi, invertibile).

$$\Gamma_p = (\gamma(i-j))_{i,j=1,2,\dots,p} = \begin{bmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_p \\ \gamma_1 & \gamma_0 & \dots & \gamma_{p-1} \\ \dots & \dots & \dots & \dots \\ \gamma_p & \gamma_{p-1} & \dots & \gamma_0 \end{bmatrix}$$

Di conseguenza, è possibile eseguire una stima del modello tramite la seguente equazione

$$\hat{X}_t = \sum_{j=1}^{p} \beta_j(p) X_{t-j}$$

Il vettore $\beta(p)$ dei coefficienti è definito dalla relazione:

$$\Gamma_p * \beta(p)' = \gamma_p'$$

La dimostrazione di questi risultati va oltre lo scopo di questo studio. Per il lettore interessato è possibile fare riferimento a Brockwell e Davis (1991) sezione 5.1 .

Dati questi risultati è possibile ottenere una stima dei coefficienti di autocovarianza tramite il campione osservato:

$\hat{\gamma}_h = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_t - \hat{u}_n)(X_{t+|h|} - \hat{u}_n)$ dove $\hat{u}_n$ è la media campionaria.

$$\hat{\Gamma}_p = (\hat{\gamma}(i-j))_{i,j=1,2,\dots,p} = \begin{bmatrix} \hat{\gamma}_0 & \hat{\gamma}_1 & \dots & \hat{\gamma}_p \\ \hat{\gamma}_1 & \hat{\gamma}_0 & \dots & \hat{\gamma}_{p-1} \\ \dots & \dots & \dots & \dots \\ \hat{\gamma}_p & \hat{\gamma}_{p-1} & \dots & \hat{\gamma}_0 \end{bmatrix}$$

A partire da questi ultimi verrà, poi, calcolato il vettore dei coefficienti del modello.

$$\hat{\beta}(p)' = \Gamma^{-1}{}_p * \gamma_p'$$

2) STIMA DEI RESIDUI

Una volta stimato il modello è possibile ottenere il vettore dei residui. Quest'ultimo è dato da

$$\tilde{e}_t(p) = (X_t - \hat{\mu}_n) - \sum_{j=1}^{p} \hat{\beta}_j(p)(X_{t-j} - \hat{\mu}_n) \ , \ t = p+1, p+2, \ldots, n$$

Tramite questo vettore viene, successivamente, ricavato il vettore dei residui centrati $\hat{e}_t(p) = \tilde{e}_t(p) - \bar{e}$ dove $\bar{e}$ è inteso come la media dei valori dei residui stimati inizialmente. A questo punto viene definita $\hat{F}_n$ la distribuzione empirica dei residui centrati.

Una volta fatto ciò, si può procedere con il campionamento bootstrap (X_1^* , X_2^* , … , X_n^*) degli n valori osservati nel nostro campione. (Il metodo bootstrap applicato è quello residual based in quanto è possibile notare che i residui e_t^* appartengono e sono estratti casualmente dalla distribuzione $\hat{F}_n$).

Viene, quindi stimato il seguente processo bootstrap secondo la metodologia già vista nel model-based:

$$(X^*_t - \hat{\mu}_n) = \sum_{i=1}^{\hat{p}} \hat{\beta}_i (X^*_{t-i} - \hat{\mu}_n) + e^*_t \ , \ t \in \mathbb{Z} \ , \ e^*_t \sim \hat{F}_n$$

I primi *p* valori non verranno considerati in quanto, come detto in precedenza, non sono stati stimati, bensì stabiliti a priori (posti uguale ai valori osservati o posti uguale alla media campionaria).

3) CALCOLO DELLA DISTRIBUZIONE BOOTSTRAP

Il passo finale tratta quello che è lo scopo di utilizzo dei metodi bootstrap, ovvero, uno studio della distribuzione del parametro di interesse.

Definiamo con:

- $\hat{\theta}_n$ la stima del parametro di interesse basata sul campione estratto dal vero processo
- θ il parametro di interesse del vero processo
- $L_n = L(c_n(\hat{\theta}_n - \theta))$ la distribuzione campionaria del parametro di interesse
- T^*_n la stima (secondo la stessa procedura di $\hat{\theta}_n$) del parametro di interesse basata sul campione estratto dal processo bootstrap

- θ^* il parametro di interesse del processo bootstrap
- $L^*_n = L^*(c_n(T^*_n - \theta^*))$ la distribuzione bootstrap del parametro di interesse

In questo caso c_n è definita come una sequenza infinita di numeri reali non negativi e crescenti. Essa viene aggiunta con lo scopo di ottenere una serie di distribuzioni che non diverga al crescere della dimensione campionaria.
Definendo L^*_n come la distribuzione bootstrap AR-sieve del parametro θ basata sul campione iniziale di dimensione *n,* è, ora, possibile studiarne il comportamento.

2.3.1 La scelta dell'ordine *p*

La maggior parte degli articoli scientifici riguardanti l'AR-sieve bootstrap concorda sul fatto che il miglior metodo per identificare l'ordine *p* del modello sia quello del minimo valore AIC Akaike Information Criteria), ovvero un criterio utilizzato per quantificare la qualità di un modello. Questo metodo è stato dimostrato da Shibata (1980) essere il migliore per la previsione di modelli AR(∞).
Per un approfondimento riguardo a questo argomento è possibile fare riferimento a Shibata (1980).

2.3.2 L'accuratezza dell'AR-sieve bootstrap

Sotto quali assunzioni riguardanti il vero processo e la statistica studiata, la distribuzione L^*_n approssima accuratamente L_n ?
Come già visto in precedenza, quasi tutti gli studiosi concordano sul fatto che il processo X_t debba essere considerato un processo autoregressivo stazionario di ordine infinito

$$X_t = \sum_{i=1}^{\infty} \beta_i (X_{t-i}) + e_t \quad , \; t \in \mathbb{Z}$$

dove i termini di errore e_t sono una sequenza indipendente e identicamente distribuita.
Il problema di applicazione consiste nel fatto che dimostrare la linearità o l'autoregressività infinita di un processo può essere molto complicato.
Ciò nonostante, dopo vari studi di tipo sia pratico che teorico, l'AR-sieve bootstrap si è dimostrato essere un metodo molto efficace e accurato. La consistenza di questo metodo è

stata dimostrata ampiamente da vari autori per una vasta gamma di stimatori. Il lettore interessato alle dimostrazioni pratiche può fare riferimento a Bühlmann (1997).

2.4 Bootstrap a blocchi

Come già accennato all'inizio di questa sezione, i modelli parametrici possono risultare inadeguati qualora il processo di riferimento non rispecchi adeguatamente date assunzioni. È per far fronte a questo problema che nascono i primi studi dei metodi bootstrap basati sulla divisione del campione in blocchi. L'idea di fondo su cui si basano queste procedure è la seguente: dividendo il campione X_n in sezioni di una lunghezza adeguata è possibile preservare la dipendenza temporale fra i dati. Inoltre, le osservazioni nel tempo tendono a essere incorrelate tra loro se la distanza è adeguatamente elevata. Di conseguenza è possibile trattare due blocchi abbastanza distanti come indipendenti e, quindi, interscambiabili.

Nel tempo sono stati molti gli studiosi che hanno approfondito queste procedure. Nel seguito introdurrò quelle che sono state le principali varianti del metodo bootstrap a blocchi come riportato da Lahiri (2003). Successivamente mi soffermerò sulla variante che ha riscontrato il maggior successo ed è considerata la più accurata (Kreiss, Lahiri 2012) tra queste, ovvero il *moving block bootstrap.* Concluderò con una sua generalizzazione che ho ritenuto importante affrontare, il *blocks of blocks bootstrap*.

2.4.1 I principali bootstrap a blocchi

I primi cenni di questa teoria possono ricondursi a Carlstein (1986) che introdusse l'idea di ciò che venne in seguito identificato con il nome di *non-overlapping block bootstrap (NBB)*. Secondo il metodo appena citato, il campione di lunghezza n viene diviso in b blocchi, ognuno di lunghezza uguale ℓ ($n=\ell * b$). È richiesto che la lunghezza cresca al crescere della dimensione campionaria. Il ricampionamento per la creazione di campioni bootstrap viene, dunque, eseguito estraendo casualmente con reinserimento b blocchi tra quelli formati in precedenza e creando campioni di lunghezza n.

Una decina di anni dopo la prima pubblicazione di Efron, il bootstrap a blocchi venne sviluppato nelle pubblicazioni di Kunch (1989) e Liu e Singh (1992) con i primi studi del cosiddetto *moving block bootstrap (MBB)*. La grande novità che lo differenzia rispetto alla

versione precedente, è la possibilità di sovrapposizione dei vari blocchi. In questo caso il secondo elemento di ogni blocco è anche il primo elemento del blocco successivo. Questo processo porta il campione a essere suddiviso in $n - \ell + 1$ blocchi di pari lunghezza. I vantaggi che questa versione apportò furono il maggior numero di blocchi tra cui era possibile eseguire un campionamento e, dunque, uno sfruttamento maggiore delle informazioni a disposizione. Questo era necessario specialmente in caso di dimensioni ridotte del campione.

Un difetto del *MBB* risiedeva nel fatto che, per costruzione, le osservazioni del campione non avevano tutte la stessa rilevanza nella formazione dei campioni bootstrap. Basti pensare che, mentre le osservazioni centrali appartenevano a ℓ blocchi ciascuna, le prime e le ultime $\ell - 1$ erano incluse in modo minore. È per questa ragione che venne introdotto il *circular block bootstrap* (Politis e Romano, 1992b). In questa versione il campione iniziale viene considerato in modo circolare, "attaccando" la parte finale a quella iniziale. Così facendo ogni osservazione riceve la stessa importanza ed è inclusa nello stesso numero di blocchi.

Un ulteriore generalizzazione del *MBB* viene proposta in Politis e Romano, 1992a, chiamata *blocks of blocks bootstrap*. Il metodo consiste nel raggruppamento delle osservazioni in vettori di dimensione *p*, per poi suddividere ulteriormente il campione in blocchi. Seguendo questa procedura il bootstrap può essere applicato anche allo studio di parametri di distribuzioni congiunte di dimensione *p* facenti riferimento a processi stazionari.

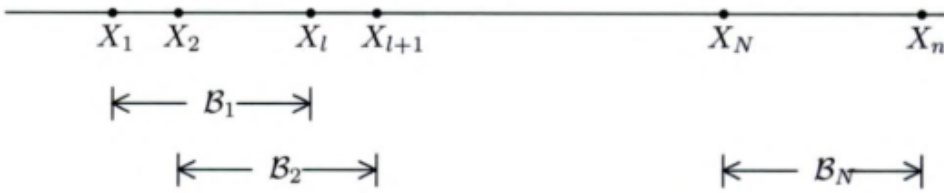
Pochi anni dopo nasce il cosidetto *stationary bootstrap (SB)* (Politis e Romano, 1994b). Basato sul *blocks of blocks bootstrap*, in questo procedimento la lunghezza dei blocchi non viene stabilita a priori, ma è, invece, una variabile casuale con una propria distribuzione. Grazie a questa innovazione, le pseudo-serie storiche generate nella dimensione bootstrap possono essere considerate stazionarie.

2.4.2 Il moving block bootstrap

Sia $\{X_t\}=(X_1, X_2, \dots, X_n)$ una serie storica stazionaria. Prima di tutto viene scelta (secondo procedure che affronterò in seguito) la lunghezza ℓ dei blocchi come numero intero positivo e minore di *n* (per semplificazione assumiamo *n* essere un multiplo di ℓ) .
Dopodiché, il campione viene suddiviso in $N=n-\ell+1$ blocchi di uguale lunghezza ℓ, dove il secondo elemento di ogni blocco corrisponde al primo del successivo.

Generalmente è richiesto che la lunghezza ℓ aumenti all' aumentare di n ($\ell \to \infty$, $n \to \infty$), e che cresca proporzionalmente in modo minore ($\ell/n \to 0$, $n \to \infty$).
Si otterrà, dunque, la seguente suddivisione:

$B_1 = (X_1 , X_2 , X_3 , \dots , X_\ell)$
$B_2 = (X_2 , X_3 , X_4 , \dots , X_{\ell+1})$
… …
$B_N = (X_{n-\ell+1} , X_{n-\ell+2} , X_{n-\ell+3} , \dots , X_n)$



Eseguita la suddivisione, è il momento di effettuare il ricampionamento. In questo caso, nella formazione di campioni bootstrap, i blocchi estratti non verranno sovrapposti, bensì posti in sequenza uno all'altro. A questo punto si estraggono da $\{ B_1 , B_2 , \dots , B_N \}$ k blocchi in modo casuale e con reinserimento.
Ciò che si otterrà sarà un vettore di k blocchi di lunghezza ℓ $\{ B_1^* , B_2^* , \dots , B_k^* \}$, dove gli elementi all'interno di ogni blocco sono denominati in questo modo:

$$B^*_i = \{X^*_{(i-1)*\ell+1} , X^*_{(i-1)*\ell+2} , \dots , X^*_{i*\ell}\} \ , \ i = 1 , \dots , k$$

Il risultato sarà, dunque:

$B^*_1 = \{X^*_1 , X^*_2 , \dots , X^*_\ell\}$
…
$B^*_k = \{X^*_{(k-1)*\ell+1} , X^*_{(k-1)*\ell+2} , \dots , X^*_{k*\ell}\}$

Ciò che ne risulterà sarà il campione bootstrap di $m = k * \ell$ elementi: $\{X^*_1 , X^*_2 , \dots , X^*_m\}$.
È possibile notare che, qualora la lunghezza ℓ dovesse essere uguale a 1, i blocchi si ridurrebbero a semplici osservazioni e il metodo risulterebbe identico a quello proposto da Efron nel 1979 per dati i.i.d.. Nel caso in cui, invece, ℓ sia maggiore di uno, ogni blocco avrà

la capacità di preservare la distribuzione congiunta di ℓ elementi del processo. Inoltre, come definito all'inizio della sezione al tendere di *n* all'infinito anche ℓ tenderà all'infinito. Di conseguenza, il ricampionamento sarà in grado di cogliere e preservare qualsiasi distribuzione congiunta di dimensione finita caratterizzante il processo reale $\mathbf{X}_n$.

Definiamo, ora, con F_n la distribuzione empirica di $X_1, X_2, \dots, X_n$. Detto ciò, lo stimatore del parametro ricercato sarà dato da $\hat{\theta}_n = T(F_n)$.
Allo stesso modo chiamiamo $F^*_{n,m}$ la distribuzione empirica basata sul singolo campione bootstrap $\{X^*_1, X^*_2, \dots, X^*_m\}$ dato $\mathbf{X}_n$. La versione bootstrap dello stimatore appena visto sarà definita come

$$\theta^*_{n,m} = T(F^*_{n,m}) \quad .$$

In genere, come nel bootstrap di Efron, è conveniente avere una dimensione campionaria bootstrap uguale a quella iniziale. Dunque, se definiamo *b* come il più piccolo numero intero tale che $b * \ell \geq n$, allora *k* dovrebbe essere posto uguale a *b* e $\theta^*_{n,m} = T(F^*_{n,m})$ dovrebbe essere calcolato usando solo i primi *n* valori del campione a cui fa riferimento.

Nel paragrafo che segue studierò una versione più generale del moving block bootstrap proposto da Künsch (1989). Questo metodo, proposto in Politis e Romano (1992a), ha lo scopo di estenderne gli utilizzi a uno specchio più ampio di statistiche e rendere il modello più generale.

2.4.3 Il blocks of blocks bootstrap

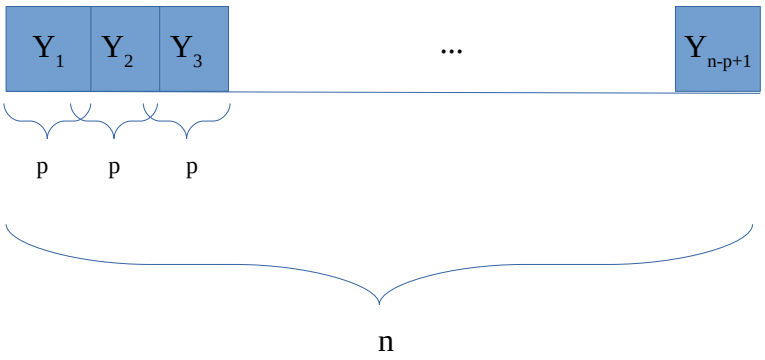
Lo stimatore della forma $\hat{\theta}_n = T(F_n)$ è basato su una distribuzione empirica monodimensionale F_n ed è utilizzabile per lo studio di molti parametri di uso comune come la media campionaria. Ciò nonostante può risultare inadatto in determinati casi. Uno dei difetti che caratterizza il bootstrap a blocchi è la creazione di campioni bootstrap non stazionari. A causa di questo, possono sorgere difficoltà nello studio di statistiche che richiedono distribuzioni a più dimensioni come l'autocorrelazione. Di conseguenza, è necessario introdurre una distribuzione congiunta che comprenda anche la dipendenza fra le variabili.

Ipotizziamo di dividere X_n in n-p+1 blocchi di una data lunghezza p in modo tale che ognuno di essi sia in grado di preservare la dipendenza tra i dati al suo interno. In questo caso p rappresenta il numero di dimensioni a noi necessario per studiare la statistica desiderata (per esempio nel caso in cui volessimo studiare l'autocovarianza di ordine uno, Cov (X_t , X_{t+1}), p sarà uguale a 2). Esso può essere definito a priori o tendere all'infinito con $n \to \infty$.
Successivamente, denominiamo ogni blocco come Y_j (X_j , X_{j+1} , … , X_{j+p-1}) , j=1,…, n-p+1, vettore di p osservazioni.
Questo processo porterà il campione a essere così suddiviso:

SUDDIVISIONE A)

$Y_1 = (X_1 , \dots , X_p)$
$Y_2 = (X_2 , \dots , X_{p+1})$
…
$Y_{n-p+1} = (X_{n-p+1} , \dots , X_n)$

Per un totale di *n-p+1* gruppi di lunghezza *p*.



Definiamo con $F_{p,n}$ la funzione di distribuzione cumulativa empirica di p dimensioni basata sul nostro campione di dati $\mathbf{X}_n$:

$$F_{p,n} = (n-p+1)^{-1} \sum_{j=1}^{n-p+1} \delta_{Y_j}$$

Per ogni $y \in \mathbb{R}^p$, δ_y rappresenta la probabilità in $\mathbb{R}^p$ (quindi un vettore di probabilità di lunghezza p).

$F_{p,n}$ sarà, dunque, relativa alla probabilità congiunta di p variabili casuali insieme (e non più la distribuzione di un singolo istante temporale). Essa è definita da vettori di probabilità e non singoli valori. Questo ci permetterà di studiare parametri che richiedono l'osservazione di più istanti temporali nello stesso momento (mentre, per esempio, per il calcolo della media è sufficiente prendere le singole osservazioni nel tempo e osservarne il valore, per statistiche come la correlazione sarà necessario considerare più osservazioni allo stesso momento e, perciò, avremo bisogno che la nostra distribuzione empirica sia la distribuzione congiunta di più variabili insieme).
Una volta ottenuta la nuova versione del campione, possiamo eseguire la divisione in blocchi di lunghezza prestabilita ℓ.

SUDDIVISIONE B)

$(Y_1, \dots, Y_\ell)$
$(Y_2, \dots, Y_{\ell+1})$
…
$(Y_{n-p-\ell+2}, \dots, Y_{n-p+1})$

Per un totale di $N = n-p-\ell+2$ blocchi di lunghezza ℓ.

Come spiegato in precedenza, una volta eseguito il ricampionamento dei blocchi, quest'ultimi vengono posti in sequenza senza sovrapposizione. Per semplicità scegliamo k numero di blocchi estratti come $k \in \mathbb{R}$ tale che $k^* \ell = n - p+1$. Grazie a questa assunzione i campioni bootstrap e il nostro campione originale avranno la stessa dimensione. (Come ho spiegato nel paragrafo precedente non sempre è possibile trovare il numero intero che rispetti l'equazione e a volte è necessario scegliere il minor k che rispetti $k^* \ell \geq$ *lunghezza campione*. In questo caso assumiamo per semplicità che quell'intero esista.
Estraendo casualmente e con reinserimento, ogni campione bootstrap assumerà la seguente forma:

$B^*_1 = \{Y^*_{S1+1}, Y^*_{S1+2}, \dots, Y^*_{S1+\ell}\}$
$B^*_2 = \{Y^*_{S2+1}, Y^*_{S2+2}, \dots, Y^*_{S2+\ell}\}$
…
$B^*_k = \{Y^*_{Sk+1}, Y^*_{Sk+2}, \dots, Y^*_{Sk+\ell}\}$

Per un totale di k blocchi.

In questa rappresentazione (S_1 , … , S_k) sono una serie di variabili casuali che vengono estratte con reinserimento dall'insieme di possibili punti di partenza (0 , … , $n - p - \ell + 1$) (vedi SUDDIVISIONE A).

QUAL' È LO SCOPO DI QUESTA ULTERIORE SUDDIVISIONE?

Per spiegare meglio il perché di questa vettorializzazione del nostro campione, ipotizziamo di voler studiare la correlazione di primo ordine Corr (X_t, X_{t+1}) su un campione X_1 , … , X_n . Nell'analizzare il campione originale ogni elemento verrà associato a quello temporalmente precedente.
Ipotizziamo ora di eseguire un blocks of blocks bootstrap con lunghezza vettoriale $p = 1$, ovvero il classico *moving block bootstrap* presentato all'inizio della sezione. Una volta diviso il campione in blocchi, averne ricampionati k e averli posti in sequenza per ottenere il campione bootstrap di lunghezza $k * \ell$ è possibile procedere con l'analisi della correlazione. L'avere utilizzato una distribuzione monodimensionale per lo studio di una statistica che richiede, invece, $p = 2$ porterà al seguente problema: nel momento in cui verrà analizzata la correlazione tra il valore alla fine di un blocco e il valore all'inizio del blocco successivo, il risultato sarà una correlazione mai riscontrata nello studio del campione originale in quanto le due variabili casuali molto probabilmente non saranno distanziate temporalmente di una sola unità.
Ipotizziamo, invece, di eseguire un MBB con una lunghezza vettoriale $p = 2$, ovvero la lunghezza corretta per lo studio di una statistica come la correlazione di primo ordine. In questo caso i componenti dei blocchi che formano il campione bootstrap saranno vettori bidimensionali contenenti osservazioni che per costruzione della nostra procedura sono sequenziali temporalmente. Lo stimatore sarà dunque calcolato con vettori di valori appartenenti al campione originale.

Per il corretto funzionamento delle procedure è necessario che la statistica da noi ricercata sia frutto di una funzione simmetrica nelle variabili casuali (Y_1 , … , Y_{n-p+1}). Una funzione di questo tipo, infatti, rimane uguale anche invertendo l'ordine dei fattori che la definiscono (per esempio nello studio della media campionaria l'inversione dell'ordine degli addendi non modifica in alcun modo il risultato). Se ciò non fosse, estrarre blocchi casualmente

modificherebbe chiaramente l'ordine dei nostri dati e lo studio della statistica risulterebbe impraticabile.
Lo studio dei parametri in questo moving block bootstrap generalizzato avviene nel modo seguente.
Nella dimensione reale, ovvero durante lo studio del vero campione, lo stimatore del nostro parametro assumerà la seguente forma:

$$\hat{\theta}_n = T(F_{p,n})$$

In questo caso $F_{p,n}$ è la funzione di distribuzione cumulativa empirica esplicitata all'inizio della sezione e $T(\cdot)$ è una funzione liscia, ovvero differenziabile infinite volte (Bülmann, 2002). Questo stimatore sarà, dunque, basato su una funzione di probabilità definita in $\mathbb{R}^p$ (e non più in $\mathbb{R}$ come studiato in precedenza).
Il corrispondente stimatore nella dimensione bootstrap, sarà, di conseguenza:

$$\hat{\theta}^*_{m,n} = T(\widetilde{F}^*_{m,n})$$

In questo caso *m* è un valore utilizzato per alleggerire la notazione. Esso equivale, infatti, alla lunghezza del nostro campione bootstrap ovvero: $m = k * \ell = n - p + 1$.

$\widetilde{F}^*_{m,n}$, invece, rappresenta la funzione di distribuzione cumulativa empirica del campione bootstrap ($Y^*_1, \ldots, Y^*_m$) ed è definita nel seguente modo:

$$\widetilde{F}^*_{m,n} = m^{-1} \sum_{j=1}^{m} \delta_{Y^*_j}$$

In sintesi, il principio di questa generalizzazione del moving block bootstrap è quello di ricampionare blocchi di vettori Y piuttosto che di singole osservazioni di modo da rendere più ampio lo specchio di statistiche analizzabili dal metodo. Lo stimatore sarà dunque calcolato con vettori di valori appartenente al campione originale.

2.4.4 La scelta della lunghezza dei blocchi

La scelta della lunghezza influenza pesantemente l'accuratezza dei metodi bootstrap a blocchi. Ciò nonostante ℓ non ha un'interpretazione pratica rilevante (Bühlmann 2002, sezione 2.3) e non esiste un metodo stabilito in modo univoco come il migliore da adottare.

È in Lahiri (2003) che vengono raggruppati quelli che sono considerati i due metodi di maggior successo nella scelta della lunghezza dei blocchi.
Il primo dei due è chiamato *plug-in principle.* Secondo questo principio, viene derivata un'espressione di ottimizzazione di ℓ secondo criteri prestabiliti. Dopodiché, basandosi sulle caratteristiche della popolazione, vengono inserite in essa (appunto il cosiddetto "plug-in") tutti i parametri mancanti e viene calcolata la lunghezza. Il difetto principale di questo metodo è la richiesta di un grande lavoro analitico nello sviluppo dell'espressione di ottimizzazione la quale dipende spesso su svariati parametri della popolazione.

Il secondo metodo presentato, più generale e di maggior applicabilità, è il cosiddetto *cross validation method.* In questa procedura viene creato uno stimatore della lunghezza basato su un criterio pre-determinato. Lo stimatore della lunghezza ottimale verrà dato dalla minimizzazione di questo criterio. Rispetto al *plug-in principle* la mole di lavoro analitico è decisamente inferiore in quanto non vi è la necessità di derivare una espressione in grado di teorizzare la lunghezza ottima per il metodo. Dall'altra parte il lavoro computazionale è molto più intenso.

La ricerca di un metodo accurato per la scelta ottimale della lunghezza è stato un argomento di grande dibattito. Molti ricercatori hanno effettuato studi su di essi e conoscenze maggiori sono richieste per la comprensione dell'argomento. Per un approfondimento è possibile fare riferimento a Lahiri (2003).

2.4.5 L'applicabilità del bootstrap a blocchi

Il bootstrap a blocchi è stato ideato con lo scopo di essere applicato a un processo stazionario $(X_t)_{t\in Z}$, dove $X_t \in R^p$ $(p \geq 1)$. Per il corretto funzionamento di questi metodi è, innanzitutto, richiesto (come specificato all'inizio di questa sezione) che lunghezza ℓ cresca al crescere di n e che il rapporto ℓ / n tenda a 0 al tendere di n all'infinito.
Inoltre, perché il metodo bootstrap sia applicabile sono richieste alcune caratteristiche asintotiche che ne garantiscano l'affidabilità (vedi sezione 1.2). Lo studio di queste proprietà è, però, relativo alla scelta del metodo bootstrap e alla scelta della statistica da studiare. Negli anni la validità e l'accuratezza di questi modelli sono state dimostrate da diversi studiosi in moltissime delle loro applicazioni. Un approfondimento di queste dimostrazioni va oltre lo

scopo di questo elaborato. Per informazioni più dettagliate riguardo l'applicabilità del bootstrap a blocchi il lettore interessato può fare riferimento a Bühlmann (2002).

2.5 Un confronto tra Ar-sieve e block bootstrap

Nel corso degli anni sia l'autoregressive-sieve bootstrap, che il bootstrap a blocchi si sono dimostrate ottime soluzioni ai problemi di ricampionamento su serie storiche. Nonostante non vi sia modo di stabilire a priori ed in maniera assoluta quale dei due metodi sia il migliore, molti ricercatori si sono dedicati al confronto tra questi procedure tramite studi teorici ed esperimenti pratici. È possibile, di conseguenza, effettuare un confronto tra le due metodologie bootstrap seguendo le valutazioni di Bülmann (1999).

Prima di tutto l'AR-sieve bootstrap impone una struttura più rigida sul processo sottostante rispetto al bootstrap a blocchi. Nel caso in cui questa struttura rispecchi la reale natura del processo, l'autoregressive-sieve bootstrap offrirà delle performance migliori. In aggiunta, i campioni bootstrap generati da questa tecnica saranno stazionari e non presenteranno modifiche nella struttura che, invece, possono occorre tramite il ricampionamento casuale dei blocchi (Bülmann ,1997).
In opposizione a questo, il bootstrap a blocchi non dipende da alcun modello specifico nella formazione dei campioni bootstrap. Questo rende il procedimento di più generale applicazione.

È possibile, inoltre, notare una somiglianza nella scelta dell'ordine ottimale p del modello nell'AR-sieve bootstrap, con la scelta della lunghezza ottimale ℓ relativa al bootstrap a blocchi. Difatti, entrambe i valori ottimali dovrebbero dipendere in qualche modo dal processo sottostante, dalla statistica studiata e dallo scopo dell'esperimento. Secondo Bülmann (1997), però, il criterio AIC di selezione di p tiene conto solamente del processo sottostante. Nella scelta di ℓ, invece, è necessario considerare anche il parametro di studio. Questo rende il procedimento del bootstrap a blocchi più complicato.

Il valore p è considerato avere un vantaggio in quanto di facile interpretazione. Essendo l'ordine del modello autoregressivo, è possibile effettuare controlli diagnostici per verificarne l'adeguatezza. Questo, invece, non è possibile con la lunghezza ℓ in quanto non vi è nessuna interpretazione relativa a questo valore. Di conseguenza, la sua verifica risulta essere più

complicata. Oltretutto, la scelta dell'ordine p è considerata essere di poca rilevanza rispetto alle prestazioni dell'AR-sieve bootstrap a patto che la stima sia ragionevolmente accurata. La scelta della lunghezza ℓ nel bootstrap a blocchi, invece, risulta essere molto influente sulle prestazioni dello stesso.

CAPITOLO 3. CONCLUSIONI

Negli anni successivi alla prima pubblicazione di Efron(1979), il metodo boostrap si è dimostrato essere uno strumento di grande affidabilità ed efficacia se comparato ai metodi di ricampionamento preesistenti.
Nel mio elaborato ho deciso di presentare le metodologie bootstrap che hanno ottenuto maggior successo nell'applicazione alle serie storiche.
Il bootstrap basato sul modello autoregressivo di ordine p è affidabile e di facile e immediata esecuzione. Ciò nonostante, non essendo possibile verificare l'effettiva struttura del processo reale, questo metodo risulta di limitata applicazione.
Per questo motivo sono nate, in seguito, tecniche bootstrap completamente non-parametriche. Tra queste l'autoregressive-sieve bootstrap è considerato da molti uno dei metodi più efficaci ed accurati. In questo caso il processo reale è assunto comportarsi come un modello autoregressivo di ordine infinito. Non richiedendo alcuna assunzione di tipo parametrico, esso risulta essere di vasta applicazione.
Un'altra metodologia bootstrap non-parametrica e di applicabilità ancora maggiore è il boostrap a blocchi. La suddivisione del campione in blocchi di lunghezza uguale permette al ricercatore di preservare la struttura di dipendenza originale durante il processo del ricampionamento. Seguendo lo stesso principio, la sua applicazione viene generalizzata a un numero maggiore di stimatori tramite il blocks of blocks bootstrap.
Tuttavia, i vari studiosi concordano sul fatto che non esista un metodo bootstrap per eccellenza, ma al contrario è necessario effettuare una valutazione del processo osservato e stabilire, di conseguenza, il metodo più adeguato.

Per concludere, i metodi bootstrap stanno diventando sempre più imprescindibili negli studi di inferenza statistica. Nel mio elaborato ho deciso di soffermarmi principalmente sulla loro applicazione alle serie storiche in quanto ritengo queste ultime essere una tipologia di processi di fondamentale importanza nella ricerca economica e finanziaria. Inoltre, l'utilizzo dei computer e delle simulazioni di Monte Carlo è diventato di grande rilevanza negli studi di carattere econometrico in quanto le condizioni di affidabilità del bootstrap vengono rispettate in gran parte di queste applicazioni.
Ritengo, dunque, che i metodi bootstrap siano strumenti di grandissima utilità laddove lo scopo sia effettuare inferenza statistica e, data la loro grande applicabilità, risulteranno essere sempre più fondamentali nello studio di processi di carattere economico.

BIBLIOGRAFIA

Bergström F. (2018), Bootstrap Methods in Time Series Analysis, Bachelor Thesis in Mathematical Statistics, Stockholm University.

Bickel Peter J., Freedmann David A. (1981), Some asymptotic theory for the bootstrap, University of California, Berkeley: The Annals of Statistics, Vol. 9, No. 6, 1196-1217.

Bose A. (1988, Edgeworth correction by bootstrap in autoregression, Indian Statistical Institute and Purdue University: The Annals of Statistics, Vol. 16, No. 4, 1709-1722.

Brockwell P. J. and Davis R. A. (1987), Time Series: Theory and Methods, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, USA: by Springer-Verlag New York Inc.

Bühlmann P. (1997), Sieve bootstrap for time series, Department of Statistics, University of California, Berkeley CA 94720-3860, USA: Bernoulli 3(2), 1997, 123–148.

Bühlmann P. (1999), Bootstraps for Time Series, Seminar fur Statistik Eidgenossische Technische Hochschule (ETH), CH-8092 Zurich, Switzerland: Research Report No. 87.

Bühlmann P. , Kunsch Hans R. (1999), Block length selection in the bootstrap for time series, Seminar fur Statistik, ETH Zurich, 8092 Zurich, Switzerland : Computational Statistics & Data Analysis 31 (1999), 295-310.

Bühlmann P. (2002), Bootstraps for Time Series, Zürich, Switzerland: Statistical Science , Vol. 17, No. 1, 52–72.

Chernick Michael R., LaBudde Robert A. (2011), An introduction to bootstrap methods with application to R, Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Cap. 1-5-6.

Dimitris N. Politis (2003), The Impact of Bootstrap Methods on Time Series Analysis, Institute of Mathematical Statistics: Statistical Science, Vol. 18, No. 2, Silver Anniversary of the Bootstrap , 219-230.

Efron, B. (1979), Bootstrap methods: another look at the jackknife, Stanford University: Ann. Statist. 7, 1-26.

Efron, B.; Tibshirani, R. (1986), Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist. Sci. 1 (1986), Vol.. 1, 54–77.

Geman, S. and Hwang, C.-R. (1982) Nonparametric maximum likelihood estimation by the method of sieves, Annals of statistics, Vol. 10, 401–414.

Grenander, U. (1981) Abstract Inference. New York: Wiley.

Härdle W., Horowitz J. and Kreiss J. P. (2003), Bootstrap Methods for Time Series, [1]CASE-Centerf or AppliedS tatisticsa nd Economics,H umboldt-Universitditzu Berlin, Berlin, Germany.[2] Departmento f Economics,N orthwesternU niversity,E vanston,I L, USA. [3]lnstitutefor Mathematical Stochastics, Technical University of Braunschweig, Braunschweig, Germany: International Statistical Review (2003), 71, 2, 435-459, Printed in The Netherlands.

Hesterberg Tim (2011), Bootstrap, Google, Seattle, USA: JohnWiley & Sons, Inc. WIREs Comp Stat 2011 3 497–526 DOI: 10.1002/wics.182, Vol. 3, 497-526.

Horowitz Joel L., Bootstrap methods in econometrics, Department of Economics Northwestern University Evanston, IL 60208 U.S.A..

Kreiss, J.P. , FRANKE J. (1989), Bootstrapping stationary autoregressive moving-average model, Technical University Braunsch weig and University of Kaiserslautern: journal of time series analysis Vol. 13, No.4, 197-317.

Kreiss, J.P., 1997. Asymptotic properties of residual bootstrap for autoregressions. Manuscript, Institute forMathematical Stochastics, Technical University of Braunschweig, Germany.

Kreiss J. P., Lahiri S. N. (2012), Bootstrap Methods for Time Series, Technische Universit¨at Braunschweig, Institut f¨ur Mathematische Stochastik, Pockelsstrasse 14, D-38106 Braunschweig, Germany, Department of Statistics, TAMU-3143 Texas A & M University, College Station, TX 77843, USA: Time Series Analysis: Methods and Applications, Vol. 30, 3-26.

Künsch H. R. (1989), The jackknife and the bootstrap for general stationary observations, ETH Zurich: The annals of statistic, Vol. 17, No. 3, 1217-1241.

Lahiri S.N. (2003), Resampling Methods for Dependent Data, Department of Statistics Iowa State University Ames, IA 50011-1212 USA, Originally published by Springer-Verlag New York, Inc. in 2003.

Mammen E. (1993), Bootstrap and Wild Bootstrap for High Dimensional Linear Models, Universität Heidelberg , The annals of statistics, Vol. 21, No. 1, 255-285.

Politis Dimitris N. and Romano Joseph P. (1994), The Stationary Bootstrap, Journal of the American Statistical Association, Vol. 89, No. 428 (Dec., 1994), 1303-1313.

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, Annals of statistics, Vol. 8, 147-164

Singh K. (1981), On the asymptotic accuracy of Efron's bootstrap, Stanford University: The annals of statics, Vol. 9, No. 6, 1187-1195.

Wu C. F. J. (1986), Jackknife, Bootstrap and other resampling methods in regression analysis, University of Winsconsin-Madison, The annals of statistics, Vol. 14, No. 4, 1261-1295.

SITOGRAFIA

An introduction to the bootstrap methods, disponibile su https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60

Chen Yen-Chi (2017), Lecture 6: Bootstrap for Regression, STAT/Q SCI 403: Introduction to Resampling Methods, disponibile su http://faculty.washington.edu/yenchic/17Sp_403/Lec6-bootstrap_reg.pdf .

Lahiri S.N., Selecting optimal block lengths for block bootstrap methods, Department of Statistics, Iowa State University, Ames, IA 50011. Disponibile su https://www.interfacesymposia.org/I03/I2003Proceedings/LahiriSoumendra/LahiriSoumendra.paper.pdf

The Glivenko-Cantelli Theorem, disponibile su http://home.uchicago.edu/~amshaikh/webfiles/glivenko-cantelli_topics.pdf .

https://slideplayer.com/slide/13513676/