Modelli Statistici per Dati Economici

Daniele Zago

20 giugno 2021

Indice

Lezione 1: Introduzione al corso

2020-09-28

L'esame consiste di quattro domande in 2h30', suddivise in tre teoriche ed una da risolvere in R; eventualmente, la domanda pratica sarà l'applicazione a un dataset di una delle domande teoriche.

Novità sul corso

- Rivisitazione con nuovi argomenti: *Bayesian VAR* e modelli *state-space*, oltre alla rivisitazione di temi classici come quelli frequentisti.
- Esercitazioni con R e focus sull'approccio computazionale.
- Studio della rappresentazione *state-space* e del filtro di Kalman & Bucy.
- Temi di discussione e di approfondimento, metodologici, computazionali e applicati.

Questo corso presenterà modelli e strumenti per analizzare dati che *evolvono nel tempo*, per qualsiasi ambito applicativo in cui si presentino queste problematiche.

1.1 Modellazione dinamica

I dati economici sono per loro natura osservazionali, in quanto il processo generatore non è pianificato dallo statistico e dunque non hanno un'origine sperimentale. Spesso il nostro ruolo sarà di stimare i parametri di un modello formulato da esperti nel settore, senza necessariamente sapere quali siano i dettagli sottostanti ad esso.

Esempio (Capital Asset Pricing Model (CAPM))
Si prezza un titolo di mercati finanziari, per cui si valuta il rischio finanziario $r_{i,t}$ al tempo t con un modello lineare

$$r_{i,t} = \alpha_i + \beta_i r_{M,t} + \varepsilon_{i,t},$$

dove $\mathbb{E}\left[\varepsilon_{i,t}|r_{M,t}\right] = 0$ e $\mathbb{V}\left[\varepsilon_{i,t}|r_{M,t}\right] = (\sigma_i^{\varepsilon})^2$. Il parametro β rappresenta il profilo di rischio associato al titolo, che può amplificare (> 1) o smorzare (< 1) il segnale di rischio rispetto al mercato.

Le assunzioni del modello dicono che il profilo di rischio, cioè il suo *andamento*, sarà sempre lo stesso dall'inizio alla fine della sua vita e pari a quello di una start-up. Ovviamente non è realistico e sarebbe opportuno farlo variare nel tempo: un modello *statico*, in questo senso, introduce una distorsione se i dati sono il risultato di un processo stocastico che evolve con il passare del tempo.

Esempio (CAPM (cont.))
Si sceglie di utilizzare invece un modello *dinamico*, in cui i coefficienti $\beta = \beta_{i,t}$ evolvono nel tempo con una struttura di dipendenza. Questa è specificata tramite una *regola di evoluzione*

stocastica e permette l'identificabilità del modello:

$$r_{i,t} = \alpha_i + \beta_{i,t} + r_{M,t} + \sigma_i^{\varepsilon}\varepsilon_{i,t}$$

$$\beta_{i,t+1} = \beta_{i,t} + \sigma_i^{\eta}\eta_{i,t}$$

$$\eta_0 = \mathcal{N}(0, \nu_0^2)$$

dove $\varepsilon_{i,t} \perp\!\!\!\perp \eta_{i,s}$ e con $\varepsilon_{i,t}, \eta_{i,t} \sim \mathcal{N}(0,1)$.

Questo è un esempio di *modello lineare a parametri variabili*, per il quale si vuole stimare in modo *non distorto il modello*.

L'uso di un modello dinamico è fondamentale per catturare l'evoluzione di un sistema stocastico.

Se si usassero tutti i dati relativi al COVID per stimare un modello lineare statico, si sarebbe ottenuta una stima intermedia tra le due rette di regressione pre-lockdown e post-lockdown, per cui *distorta*.

Il modello lineare statico si può rappresentare e stimare anche in forma dinamica, semplicemente imponendo che la regola di evoluzione sia costante: $\beta_{i,t+1} = \beta_{i,t} \implies \hat{\beta}_{T|\mathcal{T}} = \hat{\beta}_{\text{ols}}$.

La stima di un modello dinamico implica la necessità di stimare $\hat{\beta}_{T|\mathcal{T}} = \mathbb{E}\left[\beta_T | \mathscr{F}_T\right]$, cioè la media a posteriori sulla base dell'informazione $\mathscr{F}_T$ conosciuta al tempo T. Dal momento che ci interessa il valore atteso a posteriori, è necessario conoscere bene le idee e i metodi computazionali della statistica bayesiana.

Riferimenti

Modelli VAR	Lütkepohl, 2005
	Shumway e Stoffer, 2017
	Hamilton, 1994 (cons.)
State-Space	Durbin e Koopman, 2012
Panel Data	Diggle et al., 2013
Computazionale	Bishop, 2006
	Kroese e Chan, 2013
	Robert e Casella, 2004
Approfondimenti	Fahrmeir e Tutz, 2010
Laboratori	Kwon, 2016

Lezione 2: Introduzione alle serie storiche multivariate

2020-09-29

Variabili economiche e aziendali sono

- *Autocorrelate*, dunque è necessario usare modelli con dipendenza temporale.
- *Cross-correlate*, per cui si possono modellare congiuntamente con un modello vettoriale, per migliorare la capacità interpretativa e previsiva.

Un modello VAR (*Vector AutoRegressive*) è un processo stocastico in cui la risposta è un vettore y_t. Nel caso più semplice hanno una struttura del tipo

$$y_t = \Phi y_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots$$

dove $\Phi \in R^{d \times d}$ è una matrice di transizione e $\varepsilon_t \sim \mathcal{N}_d(0, \Sigma)$.

Osservazioni

- Una delle ipotesi fondamentali è la normalità dell'errore stocastico, anche se non va sempre bene e ci sono estensioni in caso fossero necessarie diverse ipotesi.
- La matrice Σ tipicamente è piena, anche se potremmo lavorare con $\Sigma^{-1} = \Omega$ e ottenere quello che si chiama *graphical VAR*. Se $\sigma_{ij} = 0$, sto ipotizzando che non vi sia alcuna dipendenza tra y_i e y_j in generale, mentre se $\omega_{ij} = 0$ significa che $y_i \perp\!\!\!\perp y_j |$resto

Questi due modi di vedere il processo, con il secondo utilizzato in applicazioni biologiche, sono fondamentalmente diversi e ci concentreremo soprattutto sul primo.

2.1 Quantità di base

Una serie storica multipla $\boldsymbol{y}_t$ è un vettore k-dimensionale, $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{k,t})^\top$ in cui ciascuna componente è una serie storica univariata.

Definiamo i momenti

$$\mathbb{E}\left[\boldsymbol{y}_t\right] = \mu = (\mu_1, \mu_2, \dots, \mu_n)^\top$$

$$\mathbb{V}\left[\boldsymbol{y}_t\right] = \Sigma = (\sigma_{ij})_{i,j=1,\dots,k}$$

Indichiamo $\Sigma \in S^k_{++}$ per dire che Σ è una matrice simmetrica e definita positiva. Con $[\Sigma]_{(i,j)}$ si indica l'elemento σ_{ij} della matrice.

L'interdipendenza è associata a correlazioni non nulle, che possono essere *contemporanee* o *non contemporanee* (in presenza di ritardi). Definiamo la *matrice di autocovarianza* $\Gamma(l)$ come la matrice delle funzioni

$$\mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{t-l)} - \mu)^\top\right] = \mathbb{C}(\boldsymbol{y}_t, \boldsymbol{y}_{t-l}) = \Gamma(l)$$

$$\mathbb{E}\left[(y_{i,t} - \mu_i)(y_{j,t-l} - \mu_j)\right] = \mathbb{C}(y_{i,t}, y_{j,t-l}) = \gamma_{ij}(l)$$

Nelle diagonali sono presenti le autocorrelazioni dei singoli processi che compongono la serie storica multivariata. Implicitamente si assume che la media al tempo $\dots, t-1, t, t+1, \dots$ sia sempre la stessa (parte dell'assunzione di *stazionarietà*).

Esempio (Correlazione nei contagi del COVID)
La correlazione tra diversi processi è utile per modellare, ad esempio, contagi avvenuti in diverse regioni. Se il singolo processo è il contagio in ciascuna regione, $\gamma_{ij}(l)$ potrebbe modellare la correlazione tra l'aumento di contagi in Veneto e l'aumento di contagi in Campania, ad un lag l passato.

Dalla funzione di autocovarianza si può ottenere quella di *autocorrelazione* tramite

$$\rho(l) = D^{-1/2}\Gamma(l)D^{-1/2},$$

con $D = \text{diag}\left(\gamma_{1,1}(0), \gamma_{2,2}(0), \ldots, \gamma_{k,k}(0)\right)$, ovvero

$$[\rho(l)]_{(i,j)} = \rho_{i,j}(l) = \frac{\gamma_{i,j}(l)}{\sqrt{\gamma_{i,i}(0)\gamma_{j,j}(0)}}.$$

La funzione $\rho_{ij}(l)$ misura la dipendenza lineare tra $y_{i,l}$ e $y_{j,t-l}$: se $\rho_{ij}(l) \neq 0$, si può affermare che $y_{j,t}$ *anticipa* la serie $y_{i,t}$ al tempo $t - l$.
La cross-correlazione non è simmetrica contemporaneamente rispetto alle variabili ed al tempo. In generale, si ha che $\rho_{ij}(l) \neq \rho_{ji}(l)$, in quanto si sta invertendo la serie che anticipa dalla serie che segue. La simmetria si ha in generale solo in un caso, quando $l = 0$.

Generalizzando, vale $\Gamma(l) = \Gamma(-l)^\top$

Si possono fare i grafici delle autocorrelazioni, sulle diagonali di $\Gamma(l)$, con le funzioni `acf` e `acf2`.

Lag positivo: Italia sul passato della Francia. Lag negativo: Francia sul passato dell'Italia.

Esercizi: calcolare i momenti di ...

Lezione 3: Stimatori della varianza e modelli VAR(1)

2020-10-06

3.1 Stimatore con il metodo dei momenti

La funzione di cross-covarianza si può stimare con il metodo dei momenti. Se $\boldsymbol{y}_t$ sono T osservazioni, allora $\overline{\boldsymbol{y}}$ è la media empirica e

$$\hat{\Gamma}(l) = \frac{1}{T}\sum_{t=l+1}^{T}(\boldsymbol{y}_t - \overline{\boldsymbol{y}})(\boldsymbol{y}_{t-l} - \overline{\boldsymbol{y}})^\top, \quad l \in \mathbb{N}$$

e usando la proprietà $\hat{\Gamma}(l) = \hat{\Gamma}(-l)$ per $l < 0$.

Il singolo elemento della matrice si può scrivere come

$$\hat{\gamma}_{ij}(l) = \sum_{t=l+1}^{T}(y_{i,t} - \bar{y}_i)(y_{j,t-l} - \bar{y}_j),$$

e lo stimatore della matrice soddisfa

$$\hat{\Gamma}(l) \xrightarrow{P} \Gamma(l).$$

Lo stimatore delle cross-correlazioni è

$$\hat{\rho}_{i,j}(l) = \frac{\hat{\gamma}_{i,j}(l)}{\sqrt{\hat{\gamma}_{i,i}(0)\hat{\gamma}_{j,j}(0)}}.$$

Proprietà

Le proprietà dello stimatore campionario $\hat{\rho}(l)$ sono piuttosto complicate e dipendono dagli ignoti $\rho_{i,j}(l)$, ma Hannan (1970) ha dimostrato che $\hat{\rho}_{i,j}(l) \sim \mathcal{N}(\rho_{i,j}(l), \sigma_\rho)$, con espressioni delle varianze abbastanza complicate.

Per T grande, in totale assenza di cross-correlazione e autocorrelazione, si ha un processo White Noise multivariato e ci sono approssimazioni del tipo

$$\mathbb{V}\left[\hat{\rho}_{i,j}(l)\right] \approx \frac{1}{T} \qquad l \neq 0,$$

$$\mathbb{V}\left[\hat{\rho}_{i,j}(0)\right] \approx \frac{\left(1 - \rho_{i,j}(0)^2\right)^2}{T} \qquad i \neq j.$$

Se le serie sono mutualmente indipendenti, $\rho_{i,j}(l) = 0$ per ogni i, j, le serie possono esser autocorrelate e

$$\mathbb{V}\left[\hat{\rho}_{i,j}(l)\right] \approx \frac{1}{T}\sum_{-\infty}^{\infty}\rho_{i,i}(l)\rho_{j,j}(l), \quad i \neq 0.$$

Questo dice che se ho un white noise questa quantità va approssimativamente a $1/T$, ed è possibile testare un'ipotesi nulla di assenza di cross-correlazione tra coppie di componenti di $\boldsymbol{y}_t$.

Osservazioni

La distribuzione approssimata ha varianze che dipende da ρ, per cui se si vuole studiare un'ipotesi $H_0 : \rho_{i,j}(l) = 0$ per $l > 0$, bisogna per forza considerare l'autocorrelazione delle singole serie.

In tal caso, si effettua un *pre-whitening*, cioè

1. Si stima un modello ARIMA a ciascuna serie marginalmente.
2. Si calcola la cross-correlazione tra le serie dei residui stimati.

A cambiare è la sola autocorrelazione marginale, per cui non si va a modificare la cross-dipendenza. In questo modo, le varianze delle cross-correlazioni tra i residui $\hat{\varepsilon}_{i,t}$ e $\hat{\varepsilon}_{j,t-l}$ sono approssimativamente $1/T$.

Non è interessante la significatività dei coefficienti, quello che interessa è avere covarianze completamente all'interno delle bande di Bartlett date da $\pm 1.96 \frac{1}{\sqrt{T}}$.

Test di Ljung-Box

Test importante per l'analisi della dipendenza, perché va a considerare l'ipotesi che ci sia almeno un elemento di $\Gamma(l)$ sia diverso da 0.
Il test univariato $\rho_{j,j}(l) = 0$ per $l = 1, 2, \dots, p$ corrisponde a

$$\mathcal{Q}_{\mathrm{LB}}(p) = T(T+1) \sum_{l=1}^{p} \frac{1}{T-l} \hat{\rho}_{j,j}(l)^2 \xrightarrow{H_0} \chi^2(p).$$

Nel caso multivariato, le cose si complicano notevolmente: l'ipotesi diventa

$$\begin{cases} H_0 : \rho_{i,j}(l) = 0 \quad i, j = 1, 2, \dots, k, \quad l = 1, 2, \dots, p \\ H_1 : \rho_{i,j}(l) \neq 0 \quad \text{per qualche } i, j, l \end{cases}$$

e la statistica test diventa

$$\mathcal{Q}_{\mathrm{LB},k}(p) = T^2 \sum_{l=1}^{p} \frac{1}{T-l} \operatorname{tr}\left(\hat{\Gamma}(l)^\top \hat{\Gamma}(0)^{-1} \hat{\Gamma}(l) \hat{\Gamma}(0)^{-1} \right),$$

con distribuzione asintotica 106G
Questa statistica test è simile al caso univariato, cioè sta effettuando una media pesata delle covarianze elevate al quadrato fino al lag p.

Attenzione: Per ogni lag p si hanno k^2 ipotesi, per cui la distribuzione ha $k^2 p$ gradi di libertà, che aumentano molto rapidamente anche per valori piccoli. Si potrebbero facilmente avere più gradi di libertà che osservazioni, per cui ci sono i soliti problemi di test di ipotesi e maledizione della dimensionalità.

3.2 Modelli VAR(1)

› Si studia il VAR(1), perché qualunque modello VAR(p) si può scrivere in forma equivalente come un VAR(1).

› Si potrebbe saltare direttamente sulla stima anche senza studiare le proprietà formali, perché quelle si lasciano agli economisti.

Def. (White noise k-dimensionale)
Un processo stocastico si dice *White Noise k-dimensionale*, e si indica $\varepsilon \sim \mathrm{WN}(0, \Sigma)$, se ε_t è tale che

$$\mathbb{E}\left[\varepsilon_t\right] = 0,$$

$$\mathbb{V}\left[\varepsilon_t\right] = \Sigma,$$

$$\mathbb{E}\left[\varepsilon_{t-j}\varepsilon_{t-l}^{\top}\right] = 0,$$

con Σ definita positiva ma non necessariamente diagonale. Se ε_t è gaussiano per ogni t, allora si indica con $\varepsilon_t \sim \mathrm{WNN}(0, \Sigma)$.

Osservazione

Questo vettore è utile per fare in modo che tutta la struttura di dipendenza sia catturata da ciò che viene aggiunto a questo processo base, in modo simile a come nel modello lineare si utilizza

$$y = \underbrace{X\beta}_{\text{struttura}} + \underbrace{\varepsilon}_{\text{base}}.$$

Def. (Processo VAR(1))
Il vettore k-dimensionale $\boldsymbol{y}_t$ segue un processo VAR(1) se

$$\boldsymbol{y}_t = \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{WN}(0, \Sigma),$$

dove $\Phi_0 \in \mathbb{R}^k$ è il vettore delle intercette, $\Phi_1 \in \mathbb{R}^{k\times k}$ è la *matrice di transizione* e ε_t è il *termine di errore* o *innovazione del modello.*

Osservazioni

› Il modello VAR(1) è un sistema di equazioni che contempla l'interdipendenza a livello

- *contemporaneo* tramite Σ, che in generale è piena
- *ritardato* tramite Φ_1, in generale non simmetrica: $\varphi_{1,12}$ è l'impatto della prima variabile sulla seconda, mentre $\varphi_{1,21}$ è l'impatto della seconda sulla prima.

Il processo VAR(1) prende una serie $\boldsymbol{y}_t$ e ci impone due fonti separate di dipendenza:

1. $\Phi_0 + \Phi_1 \boldsymbol{y}_{t-1}$: dipendenza al lag $l = 1, 2, \ldots$, in quanto si può espandere ricorsivamente.
2. ε_t : dipendenza contemporanea.

Queste due fonti sono *ortogonali*, cioè la dipendenza totale è la loro somma.

3.3 Rappresentazione triangolare del VAR(1)

Per esplicitare la dipendenza contemporanea, si può usare la *fattorizzazione di Cholesky modificata*, definita come

$$\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^\top = LD^{1/2}D^{1/2}L^\top,$$

dove L triangolare con elementi diagonali pari a 1 e D matrice diagonale:

$$L = \begin{pmatrix} 1 & l_{12} & \ldots & l_{1k} \\ 0 & 1 & \ldots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \qquad D = \operatorname{diag}(d_1, \ldots, d_k).$$

Osservazioni

› D non ha come elementi diagonali le varianze, verificare per casa.

› Su R si può invertire una matrice triangolare con il comando per la `forward substitution` con costo $O(p^2)$ invece di $O(p^3)$.

Moltiplicando a destra e sinistra per L^{-1}, si ottiene equivalentemente

$$\begin{aligned} L^{-1}\boldsymbol{y}_t &= L^{-1}\Phi_0 + L^{-1}\Phi_1 \boldsymbol{y}_{t-1} + L^{-1}\varepsilon_t \\ &= \Phi_0^* + \Phi_1^* \boldsymbol{y}_{t-1} + \eta_t, \end{aligned}$$

dove η_t è un vettore di componenti tra loro linearmente indipendenti, poiché la matrice è di rango pieno, e varianza $\mathbb{V}\left[\eta_t\right] = L^{-1}\Sigma\,(L^{-1})^\top = L^{-1}LDL^\top(L^\top)^{-1} = D$, dunque incorrelati.

Caratteristiche

Siccome L^{-1} è triangolare inferiore, si ha che per ciascun $y_{j,t}$, la dipendenza contemporanea è relativa solamente alle componenti $y_{i,t}$ con $1 \leq i < j$:

$$y_{j,t} = \Phi_{0,t}^* - \sum_{i=1}^{j-1} l_{j,i} y_{i,t} + \sum_{i=1}^{k} \Phi_{1,j,i}^* y_{i,t-1} + \eta_{j,t}.$$

Con la scrittura del modello in forma *triangolare*, le innovazioni $\eta_{j,t}$ sono incorrelate e la struttura di dipendenza si è trasferita interamente nella dipendenza dagli y_i ai tempi precedenti e contemporanei.

Questo in realtà nasconde un problema di identificazione del modello, in quanto se si scelgono ordini diversi per le variabili, si ottengono modelli con variabili differenti come regressori e quindi con struttura diversa.

Lezione 4: Momenti del VAR(1)

2020-10-07

4.1 Rappresentazione VMA(∞) del VAR(1)

Si studiano ora le proprietà stocastiche del processo, per le quali è necessario introdurre un'altra rappresentazione del VAR(1).

L'equazione che definisce il VAR(1) è per sua natura ricorsiva e permette di ottenere una nuova rappresentazione, che viene definita implicitamente:

$$\begin{aligned}
\boldsymbol{y}_t &= \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \varepsilon_t \\
&= \Phi_0 + \Phi_1\Phi_0 + \Phi_1^2 \boldsymbol{y}_{t-2}\Phi_1\varepsilon_{t-1} + \varepsilon_t + \\
&= \Phi_0 + \Phi_1\Phi_0 + \Phi_1^2\Phi_0 + \Phi_1^3 \boldsymbol{y}_{t-3} + \Phi_1^2\varepsilon_{t-2} + \Phi_1\varepsilon_{t-1} + \varepsilon_t \\
&\vdots \\
&= \underbrace{\left(I_k + \sum_{i=1}^{j-1} \Phi_1^i\right)\Phi_0}_{\text{intercetta}} + \underbrace{\Phi_1^j \boldsymbol{y}_{t-j}}_{\text{passato}} + \underbrace{\sum_{i=0}^{j-1} \Phi_1^i \varepsilon_{t-i}}_{\text{stocastico}}
\end{aligned}$$

Osservazioni

› Nella pratica, per la previsione ci si può fermare al primo tempo osservato $\boldsymbol{y}_0$ e, assumendolo noto, ne deriva che

$$\boldsymbol{y}_t = \left(I_k + \sum_{i=1}^{t} \Phi_1^i\right)\Phi_0 + \Phi_1^t \boldsymbol{y}_0 + \sum_{i=0}^{t} \Phi_1^i \varepsilon_{t-i}.$$

› Poiché il processo è definito per tutta la storia, spesso si ipotizza la presenza di un infinito passato considerando il limite per $j \to \infty$.

Assunzione forte: gli autovalori $\lambda_1, \ldots, \lambda_k$ di Φ_1 sono tali che $|\lambda_j| < 1$ per ogni j. L'ipotesi è equivalente alla condizione di stazionarietà $|\varphi| < 1$ per un processo AR(1).

Sotto questa ipotesi, si può dimostrare che

$$\lim_{j\to\infty} \Phi_1^j = 0 \tag{1}$$

per cui in presenza di un infinito passato il termine $\Phi_1^j \boldsymbol{y}_{t-j}$ converge a zero. Questo si può vedere applicando la decomposizione spettrale a Φ_1, e osservando che la potenza j-esima è

$$\Phi^j = \Gamma\Lambda^j\Gamma^\top,$$

con Λ matrice degli autovalori.

Come conseguenza, il sistema si dimentica molto velocemente di $\boldsymbol{y}_0$ all'aumentare di t, in quanto l'unica relazione tra $\boldsymbol{y}_t$ e $\boldsymbol{y}_0$ avviene attraverso Φ_1^t.

Quanto più violata è questa ipotesi, tanto più vi è dipendenza dal punto iniziale del processo (*persistenza*). Per processi molto autocorrelati, vi è una violazione molto profonda di quest'assunzione.

Inoltre, sotto l'assunzione si ha un risultato di convergenza analogo alle serie geometriche

$$I_j + \sum_{i=1}^{j-1} \Phi_1^j \xrightarrow{j\to\infty} (I_k - \Phi_1)^{-1} \tag{2}$$

per cui, siccome $|\lambda_1|, \dots, |\lambda_k| < 1$ implica l'assoluta sommabilità $\sum_{i=0}^{\infty} |\Phi_1^i| < \infty$, si ha una rappresentazione del processo nella forma

$$\boldsymbol{y}_t = \mu + \sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i},$$

dove $\mu = (I_k - \Phi_1)^{-1}$. Questa rappresentazione è detta a media mobile vettoriale di ordine infinito VMA(∞).

Con questa rappresentazione, si può utilizzare il fatto che i ε_j sono incorrelati per calcolare valore atteso e varianza del processo. Per ottenere la varianza si osserva che $\mathbb{V}\left[\varepsilon_t\right] = \mathbb{E}\left[\varepsilon_t \varepsilon_t^\top\right] = \Sigma$ e $\mathbb{E}\left[\varepsilon_t \varepsilon_s^\top\right] = 0$ per $t \neq s$:

$$\mathbb{E}\left[\boldsymbol{y}_t\right] = \mathbb{E}\left[\mu + \sum_{i=1}^{\infty} \Phi_1^i \varepsilon_{t-i}\right] = (I_k - \Phi_1)^{-1} \Phi_0.$$

$$\mathbb{V}\left[\boldsymbol{y}_t\right] = \mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_t - \mu)^\top\right] = \sum_{i=0}^{\infty} \Phi_1^i \Sigma (\Phi_1^i)^\top.$$

Osservazione

La varianza è una sommatoria infinita di termini, per cui bisogna verificare che sia un valore finito. Poiché si ha la condizione $\sum_{i=0}^{\infty} |\Phi_1^i| < \infty$, si ha un analogo matriciale della proprietà scalare

$$|a| < 1 \implies a^2 < |a| < 1,$$

dunque anche $\mathbb{V}\left[\boldsymbol{y}_t\right] < \infty$.

Sfruttando nuovamente la rappresentazione come VMA(∞), si può ricavare la matrice di covarianza

al lag l :

$$\begin{aligned}\Gamma(l) &= \mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{t-l} - \mu)^\top\right] \\ &= \mathbb{E}\left[\Big(\sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i}\Big)\Big(\sum_{j=0}^{\infty} \Phi_1^j \varepsilon_{t-j-l}\Big)^\top\right] \\ &= \mathbb{E}\left[\Big(\sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i}\Big)\Big(\sum_{j=i+l} \Phi^j \varepsilon_{t-j-l}\Big)^\top\right] \qquad (\neq 0 \iff j = i+l) \\ &= \sum_{i=0}^{\infty} \Phi_1^{i+l} \Sigma \left(\Phi_1^i\right)^\top\end{aligned}$$

4.2 Equazione dinamica per Γ

Si osserva inoltre che, per associatività della moltiplicazione tra matrici, si può raccogliere Φ_1 e scrivere

$$\Gamma(l) = \Phi_1 \sum_{i=0}^{\infty} \Phi_1^{i+l-1} \Sigma \left(\Phi_1^i\right)^\top = \Phi_1 \Gamma(l-1).$$

Importante: Questa è un'*equazione lineare dinamica* per Γ, nel senso che è un'equazione che mette in relazione una quantità con se stessa al tempo precedente:

$$x(t) = f\big(x(t-1)\big),$$

In particolare, osservando che Φ_1 guida l'evoluzione della matrice di covarianza, si ha che

$$\boxed{\Gamma(l) \text{ ha la stessa evoluzione di } \boldsymbol{y}_t}$$

Nei processi dinamici lineari, l'equazione che determina il processo determina anche l'evoluzione dell'autocovarianza. Esiste un'altra equazione che ha le stesse proprietà di evoluzione del processo, che è la *funzione di previsione*.

Gli autovalori di Φ_1 regolano anche la persistenza dell'autocovarianza e cross-covarianza, per cui se $|\Phi_1| < 1$ si avrà un decadimento esponenziale dell'autocovarianza per lag crescenti.

Infine, ripetendo questa operazione l volte, si osserva un comportamento di *exponential decay* per $\Gamma(l)$ analogo a quello del processo AR(1):

$$\Gamma(l) = \Phi_1^l \sum_{i=0}^{\infty} \Phi_1^i \Sigma \left(\Phi_1^i\right)^\top = \Phi_1^l \Gamma(0).$$

Lezione 5: Stazionarietà e Stabilità

2020-10-12

Il processo VAR(1) si può anche scrivere in termini dell'operatore di ritardo L, definito come $L^p \boldsymbol{y}_t = \boldsymbol{y}_{t-p}$. Con questo operatore, si può scrivere

$$\boldsymbol{y}_t = \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \varepsilon_t$$

$$(I_k - \Phi_1 L)\boldsymbol{y}_t = \Phi_0 + \varepsilon_t$$

$$\Phi(L)\boldsymbol{y}_t = \Phi_0 + \varepsilon_t,$$

con $\Phi(L) = (I_k - \Phi_1 L)$ che prende il nome di *polinomio caratteristico del VAR(1).*

La precedente assunzione $|\lambda|_1, |\lambda_2|, \ldots, |\lambda_k| < 1$ prende il nome di *condizione di stabilità*, che si dimostra essere verificata se l'equazione caratteristica inversa del VAR(1)

$$\boxed{\det(I_k - \Phi z) = 0 \text{ ha radici in modulo } > 1}$$

Def. (Stazionarietà)
Sia $(\boldsymbol{y}_t)_{t\in\mathbb{Z}}$ un processo stocastico. Si dice che $(\boldsymbol{y}_t)_{t\in\mathbb{Z}}$ è *stazionario* se il momento primo e il momento secondo sono indipendenti dal tempo:

1. $\mathbb{E}\left[\boldsymbol{y}_t\right] = \mu$ per ogni $t \in \mathbb{Z}$;
2. $\mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{t-h} - \mu)^\top\right] = \Gamma(h) = \Gamma(-h)^\top$ per ogni $t \in \mathbb{Z}$ e $h \in \mathbb{Z}$.

Il calcolo del valore atteso e covarianza della lezione precedente mostra che, per un VAR(1),

$$\text{Stabilità} \implies \text{Stazionarietà}.$$

Prop. (Autovalori ed eq. caratteristica del VAR(1))
Per il processo VAR(1), le seguenti condizioni sono equivalenti:

1. $\det(\Phi_1 - \lambda I_k) = 0$ *per* $|\lambda| < 1$.
2. $\det \Phi(L) = \det(I_k - \Phi_1 z) = 0$ *per* $|\lambda| \geq 1$.

Dim.
No.

□

5.1 Cross-covarianze contemporanee del VAR(1)

Per calcolare le cross-covarianze, si può scrivere il VAR(1) sottraendo la media

$$\begin{aligned}
\boldsymbol{y}_t &= \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \varepsilon_t \\
\boldsymbol{y}_t - \mu &= \Phi_0 - \mu + \Phi_1(\boldsymbol{y}_{t-1} - \mu + \mu) + \varepsilon_t && (-\mu \text{ a sx e dx, } \pm\mu \text{ a dx}) \\
\tilde{\boldsymbol{y}}_t &= \Phi_0 - (I_k - \Phi_1)\mu + \tilde{\boldsymbol{y}}_{t-1} + \varepsilon_t \\
\tilde{\boldsymbol{y}}_t &= \Phi_0 - \cancel{(I_k - \Phi_1)}\cancel{(I_k - \Phi_1)^{-1}}\Phi_0 + \Phi_1 \tilde{\boldsymbol{y}}_{t-1} + \varepsilon_t && (\mu = (I_k - \Phi_1)^{-1}\Phi_0) \\
\tilde{\boldsymbol{y}}_t &= \Phi_1 \tilde{\boldsymbol{y}}_{t-1} + \varepsilon_t.
\end{aligned}$$

Con questa scrittura, la matrice di cross-covarianza $\Gamma(h)$ è

$$\begin{aligned}
\mathbb{E}\left[\tilde{\boldsymbol{y}}_t \tilde{\boldsymbol{y}}_{t-h}^\top\right] &= \mathbb{E}\left[(\Phi_1 \tilde{\boldsymbol{y}}_{t-1} + \varepsilon_t)\tilde{\boldsymbol{y}}_{t-h}^\top\right] \\
&= \Phi\, \mathbb{E}\left[\tilde{\boldsymbol{y}}_{t-1} \tilde{\boldsymbol{y}}_{t-h}^\top\right] + \mathbb{E}\left[\varepsilon_t \tilde{\boldsymbol{y}}_{t-h}^\top\right]
\end{aligned}$$

Per $h \neq 0$ si ottiene l' espressione già trovata dell'equazione dinamica per $\Gamma(l)$. Per $h = 0$, invece, si ha

$$\Gamma(0) = \Phi\Gamma(-1) + \Sigma = \Phi\Gamma(1)^\top + \Sigma.$$

Dal momento che $\Gamma(1) = \Phi_1\Gamma(0)$, sostituendolo nell'espressione si ha

$$\Gamma(0) = \Phi_1\Gamma(0)\Phi_1^\top + \Sigma,$$

equazione dalla quale adesso si può ricavare (con un po' di fatica) $\Gamma(0)$.

Def. (Operatore vec)
Si definisce l'*operatore* vec come la trasformazione che vettorizza una matrice colonna per colonna, ovvero l'isomorfismo $\mathcal{M}_{m\times n}(\mathbb{R}) \cong \mathbb{R}^{mn}$. Ad esempio, nel caso 2×2 vale

$$\operatorname{vec}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a \\ c \\ b \\ d \end{pmatrix}.$$

Proprietà

› vec è un operatore lineare

$$\operatorname{vec}(A + B) = \operatorname{vec} A + \operatorname{vec} B$$

› Per A, B, C matrici conformabili si ha

$$\operatorname{vec}(ABC) = (C^\top \otimes A)\operatorname{vec} B,$$

$$\operatorname{vec}(AB) = (I \otimes A)\operatorname{vec} B,$$

$$\operatorname{vec}(BC) = (C^\top \otimes I)\operatorname{vec} B,$$

dove $\otimes : \mathbb{R}^{m\times n} \times \mathbb{R}^{p\times q} \to \mathbb{R}^{mp\times nq}$ è il prodotto di Kronecker

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}$$

Applicando l'operatore vec all'espressione ottenuta per la matrice $\Gamma(0)$, si ottiene

$$\begin{aligned} \operatorname{vec}\Gamma(0) &= \operatorname{vec}\left(\Phi_1\Gamma(0)\Phi_1^\top + \Sigma\right) \\ &= \operatorname{vec}\left(\Phi_1\Gamma(0)\Phi_1^\top\right) + \operatorname{vec}(\Sigma) && \text{(linearità)} \\ &= (\Phi_1 \otimes \Phi_1)\operatorname{vec}(\Gamma(0)) + \operatorname{vec}\Sigma && \text{(prop. 2)} \end{aligned}$$

dunque ricavando $\operatorname{vec}\Gamma(0)$ si ha

$$\operatorname{vec}\Gamma(0) = (I_{k^2} - \Phi_1 \otimes \Phi_1)^{-1}\operatorname{vec}\Sigma. \tag{3}$$

Risolvendo questo sistema di equazioni, si può calcolare $\operatorname{vec}\Gamma(0)$ e poi ricostruire la matrice $\Gamma(0)$.

Infine, si può ottenere la matrice di cross-correlazione del VAR(1) come

$$R(l) = D^{-1}\Gamma(l)D^{-1},$$

dove $D = \operatorname{diag}(\sqrt{\gamma_{1,1}(0)}, \dots, \sqrt{\gamma_{k,k}(0)})$.

Lezione 6: Processo VAR(p)

2020-10-13

Un processo VAR(p) è definito come

$$\boldsymbol{y}_t = \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \Phi_2 \boldsymbol{y}_{t-2} + \dots \Phi_p \boldsymbol{y}_{t-p} + \varepsilon_t,$$

riscrivibile in forma compatta tramite l'operatore ritardo

$$\begin{aligned} &\Phi(L)\boldsymbol{y}_t = \Phi_0 + \varepsilon_t, \\ &\Phi(L) = I_k - \Phi_1 L - \Phi_2 L^2 - \dots - \Phi_p L^p. \end{aligned} \tag{4}$$

6.1 Momenti di un VAR(p)

Se il processo è stazionario, si può calcolare il valore atteso osservando che

$$\mathbb{E}\left[\boldsymbol{y}_t\right] = \Phi_0 + \Phi_1 \mathbb{E}\left[\boldsymbol{y}_{t-1}\right] + \dots + \Phi_p \mathbb{E}\left[\boldsymbol{y}_{t-p}\right] + \mathbb{E}\left[\varepsilon_t\right],$$

ma poiché $\mathbb{E}\left[\boldsymbol{y}_t\right] = \mathbb{E}\left[\boldsymbol{y}_s\right]$ per ogni t, s, si ha

$$\mathbb{E}\left[\boldsymbol{y}_t\right] = \Phi(1)^{-1}\Phi_0.$$

Anche per il VAR(p) si può cercare una rappresentazione VMA(∞) del tipo

$$\boldsymbol{y}_t = \mu + \sum_{j=0}^{\infty} \Psi^j \varepsilon_{t-j},$$

che si può ottenere dall'equazione (4) invertendo l'operatore $\Phi(L)$:

$$\begin{aligned} \boldsymbol{y}_t &= \Phi(L)^{-1}\Phi_0 + \Phi(L)^{-1}\varepsilon_t \\ &= \Phi(1)^{-1}\Phi_0 + \Phi(L)^{-1}\varepsilon_t \end{aligned}$$

In generale, per un VAR(p) si ha una rappresentazione VMA(∞) dove

$$\Psi_k = \sum_{j=1}^{\min(k,p)} \Psi_{k-j}\Phi_j, \quad \Psi_0 = I_k,$$

che viene usata per calcolare gli standard error per gli intervalli di confidenza.

La matrice di varianza contemporanea si può calcolare dalla rappresentazione VMA(∞) come

$$\Gamma(0) = \mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_y - \mu)^\top\right] = \sum_{i=0}^{\infty} \Psi_i \Sigma \Psi_i^\top,$$

che è sicuramente finita se il processo è stazionario.

Si ha che

$$(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{y-l} - \mu)^\top = \sum_{j=1}^{p} \Phi_j \boldsymbol{y}_{t-j} \boldsymbol{y}_{t-l}^\top + \varepsilon_t \boldsymbol{y}_{t-l}^\top$$

$$\implies \mathbb{E}\left[(\boldsymbol{y}_t - \mu)(\boldsymbol{y}_{t-l} - \mu)^\top\right] = \sum_{j=1}^{p} \Phi_j \, \mathbb{E}\left[\boldsymbol{y}_{t-j} \boldsymbol{y}_{t-l}^\top\right] + \underbrace{\mathbb{E}\left[\varepsilon_t \boldsymbol{y}_{t-l}^\top\right]}_{=0 \text{ per } l>0}.$$

Dunque si ha una forma di $\Gamma(l)$ simile al VAR(1):

$$\Gamma(l) = \begin{cases} \Phi_1\Gamma(-1) + \ldots + \Phi_p\Gamma(-p) + \Sigma & \text{se } l = 0 \\ \Phi_1\Gamma(l-1) + \ldots + \Phi_p\Gamma(l-p) & \text{se } l \neq 0 \end{cases}$$

$$= \begin{cases} \Phi_1\Gamma(1)^\top + \ldots + \Phi_p\Gamma(p)^\top + \Sigma & \text{se } l = 0 \\ \Phi_1\Gamma(l-1) + \ldots + \Phi_p\Gamma(l-p) & \text{se } l \neq 0 \end{cases}$$

Anche in questo caso è la stessa equazione del processo e sono necessari p "valori iniziali" dati dalle matrici ai lag precedenti.

Per un processo VAR(2) la formula ricorsiva diventa

$$\Gamma(0) = \Phi_1\Gamma(1)^\top + \Phi_2\Gamma(2)^\top + \Sigma$$

$$\Gamma(1) = \Phi_1\Gamma(0) + \Phi_2\Gamma(1)^\top$$

$$\Gamma(k) = \Phi_1\Gamma(k-1) + \Phi_2\Gamma(k-2)$$

con la necessità di calcolare i valori iniziali $\Gamma(0)$ e $\Gamma(1)$.

6.2 Forma compagna del VAR(p)

Per il calcolo dei valori iniziali del VAR(p) si può utilizzare la *forma canonica* o *forma compagna* del modello, che rappresenta il VAR(p) come un processo VAR(1):

$$\underbrace{\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{pmatrix}}_{(k\cdot p)\times 1} = \underbrace{\begin{pmatrix} \Phi_0 \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}}_{A_0} + \underbrace{\begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ I_k & 0 & \ldots & 0 & 0 \\ 0 & I_k & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_k & 0 \end{pmatrix}}_{A_1} + \underbrace{\begin{pmatrix} \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \boldsymbol{y}_{t-3} \\ \vdots \\ \boldsymbol{y}_{t-p} \end{pmatrix}}_{X_{t-1}} + \underbrace{\begin{pmatrix} \varepsilon_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{u_t}$$

La matrice Σ_x deriva dalla forma del termine stocastico u_t e corrisponde a

$$\Sigma_x = \begin{pmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Stabilità e stazionarietà

La forma canonica scritta in modo compatto permette di rappresentare il VAR(p) come un VAR(1) di dimensione kp

$$X_t = A_0 + A_1 X_{t-1} + u_t,$$

a cui si possono applicare le formule del processo VAR(1). In particolare, stabilità e stazionarietà si possono tradurre in termini di stazionarietà del VAR(1) corrispondente, ovvero

$$\boxed{\text{VAR}(p) \text{ stazionario} \iff |I_{kp} - A_1 z| = 0 \text{ per } z > 1}$$

che corrisponde alla condizione

$$\boxed{|I_k - \Phi_1 z - \Phi_2 z^2 - \ldots - \Phi_p z^p| = 0 \quad \text{per } z > 1}$$

Inoltre, ricordando la formula (3) per la varianza di un AR(1), si ha

$$\operatorname{vec} \Gamma_X(0) = \left(I_{(kp)^2} - A_1 \otimes A_1\right)^{-1} \operatorname{vec} \Sigma_x.$$

Considero la matrice di selezione del primo elemento $\boldsymbol{y}_t$

$$J = [I_k \; 0_k \; \ldots \; 0_k] \implies \boldsymbol{y}_t = J X_t,$$

per calcolare i momenti di $\boldsymbol{y}_t$ a partire da X_t. Infatti,

$$\mathbb{E}\left[\boldsymbol{y}_t\right] = J\,\mathbb{E}\left[X_t\right] = J\mu_x$$

$$\Gamma(l) = J\Gamma_X(l)J^\top.$$

Sempre usando J, si può estrarre la componente Σ di Σ_x che è definita positiva, lasciando stare il resto.

6.3 Introduzione alla stima dei parametri

Il modo classico di stimare i parametri è il metodo dei minimi quadrati, ma usando i metodi bayesiani si ottengono cose più interessanti.

Si consideri un campione di $T + p$ osservazioni di un processo $\boldsymbol{y}_t$, dove le prime p si definiscono "pre-campionarie" e sono usate per condizionare ai valori iniziali. Queste osservazioni iniziali possono anche essere fittizie, in realtà.

Se si definiscono le quantità

$$\begin{aligned}
Y &= (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_T) \in \mathbb{R}^{k\times T} \\
B &= (\Phi_0, \Phi_1, \Phi_2, \dots, \Phi_p) \in \mathbb{R}^{k\times(kp+1)} \\
\varepsilon &= (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T) \in \mathbb{R}^{k\times T} \\
z_t &= (1, \boldsymbol{y}^\top, \boldsymbol{y}_{t-1}^\top, \dots, \boldsymbol{y}_{t-p+1})^\top \in \mathbb{R}^{(kp+1)\times 1} \\
Z &= (z_1, z_2, \dots, z_{T-1}) \in \mathbb{R}^{(kp+1)\times T}
\end{aligned}$$

Tutto questo serve a scrivere il processo VAR(p) nella forma di un modello di regressione lineare, in quanto per $t = 2, 3, \dots, T$ si ha

$$\begin{aligned}
\boldsymbol{y}_t &= \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \Phi \boldsymbol{y}_{t-2} + \dots + \Phi_p \boldsymbol{y}_{t-p} + \varepsilon_t \\
&= B z_{t-1} + \varepsilon_t,
\end{aligned}$$

per cui si può scrivere in forma compatta come un modello lineare vettoriale

$$Y = BZ + \varepsilon.$$

In questo caso, l'ipotesi standard $\mathbb{V}\left[\varepsilon_t\right] = \sigma_\varepsilon^2 I$ risulta troppo restrittiva, per cui si utilizza una struttura di cross-correlazione:

$$\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \Sigma.$$

Questa scrittura è utile, ma non permette di stimare facilmente i parametri di regressione dentro B. Per ricavare lo stimatore di B conviene riscrivere il modello con l'operatore vec, in particolare

$$Y \in \mathbb{R}^{k\times T} \to \operatorname{vec}(Y) \in \mathbb{R}^{kT\times 1},$$

quindi poiché il vec è lineare e

$$\operatorname{vec}(BZ) = (Z^\top \otimes I_k)\operatorname{vec}(B),$$

si ha che il modello si può scrivere nella forma di regressione lineare

$$\begin{aligned}
\operatorname{vec}(Y) &= (Z^\top \otimes I_k)\operatorname{vec}(B) + \operatorname{vec}(\varepsilon) \\
\boldsymbol{y} &= (Z^\top \otimes I_k)\beta + \boldsymbol{u}.
\end{aligned} \tag{5}$$

Il vettore dei parametri è

$$\beta = (\Phi_0^\top, \operatorname{vec}(\Phi_1)^\top, \operatorname{vec}(\Phi_2)^\top, \dots, \operatorname{vec}(\Phi_n)^\top)^\top \in \mathbb{R}^{k(kp+1)\times 1},$$

mentre il termine di errore ha momenti pari a

$$\mathbb{E}\left[\boldsymbol{u}\right] = 0$$

$$\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\top\right] = \mathbb{E}\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix}(\varepsilon_1^\top, \varepsilon_2^\top, \ldots, \varepsilon_n^\top) = \begin{pmatrix}\Sigma & 0 & \ldots & 0\\ 0 & \Sigma & \ldots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \ldots & \Sigma\end{pmatrix} = I_T \otimes \Sigma.$$

Nel caso classico di errori incorrelati $\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \sigma^2 I_T$ (*sfericità*), le soluzioni ai minimi quadrati sono le classiche

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top \boldsymbol{y}, \qquad \text{con } X = Z^\top \otimes I_k,\\ \hat{\sigma}^2 &= \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{T-k}.\end{aligned} \tag{6}$$

Lezione 7: Stima dei parametri

2020-10-14

Il VAR(p) scritto come modello lineare (5) è di dimensione più grande di quelli usuali e ha una matrice di varianza e covarianza ignota, che ne rende complicata la stima.

7.1 Feasible Generalized Least Squares (FGLS)

Sia $\mathbb{E}\left[\varepsilon\varepsilon^\top\right] = \Sigma$, con $\Sigma^{-1} = PP^\top$, allora

$$P^\top Y = P^\top X\beta + P^\top\varepsilon$$

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}.$$

In questa forma, si ha la varianza pari a

$$\mathbb{E}\left[\tilde{\varepsilon}\tilde{\varepsilon}^\top\right] = P^\top\Sigma P = P^\top(PP^\top)^{-1}P = I.$$

Applicando i minimi quadrati al modello così trasformato, si ottiene lo stimatore dei *minimi quadrati generalizzati*:

$$\begin{aligned}\hat{\beta} &= (\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\tilde{Y}\\ &= \left(X^\top\Sigma^{-1}X\right)^{-1}X^\top\Sigma^{-1}Y.\end{aligned}$$

Problema Questo si può fare solo nel caso in cui Σ sia una matrice nota.

Siccome Σ non è nota, si può invece iterare una procedura per calcolare uno stimatore "vicino" al GLS, chiamato FGLS (*Feasible Generalized Least Squares*).

Algoritmo 1 Feasible Generalized Least Squares

1: **Init:**
   Stima iniziale $\hat{\beta} = \beta_{\text{ols}}$
2: **while** *not converged* **do**
3: $\quad\hat{\varepsilon} \leftarrow y - X\hat{\beta}$ ▷ Calcolo residui
4: $\quad\hat{\Sigma} \leftarrow \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i\hat{\varepsilon}_i^\top$ ▷ Stima covarianza
5: $\quad\hat{\beta} = (X^\top\hat{\Sigma}^{-1}X)^{-1}X^\top\hat{\Sigma}^{-1}Y$ ▷ Applico GLS
6: **end while**

Stimatore FGLS esatto

Tornando alla forma compatta del VAR(p), sostituendo le quantità $X = (Z^\top \otimes I_k)$ e $\Sigma = I_T \otimes \Sigma$, si ottiene lo stimatore GLS

$$\begin{aligned}\hat{\beta} &= (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}\boldsymbol{y}\\ &= \left((Z^\top \otimes I_k)^\top(I_T \otimes \Sigma)^{-1}(Z^\top \otimes I_k)\right)^{-1}(Z^\top \otimes I_k)^\top(I_T \otimes \Sigma)^{-1}\boldsymbol{y}.\end{aligned} \tag{7}$$

Teo. (Proprietà del prodotto di Kronecker)

1. *Siano A, C matrici conformabili e B, D matrici conformabili, allora*

$$(A \otimes B)(C \otimes D) = (AC \otimes BD).$$

2. *Per qualunque A, B si ha che*

$$(A \otimes B)^\top = A^\top \otimes B^\top$$

3. *Date A e B matrici invertibili, allora*

$$(A \otimes B)^{-1} = \big(A^{-1} \otimes B^{-1}\big).$$

Dim.
No.

□

Usando queste due proprietà, si può riscrivere la stima dei minimi quadrati (7) come

$$\begin{aligned}
\hat{\beta} &= \big(\underbrace{(Z^\top \otimes I_k)^\top}_{(2)}\underbrace{(I_T \otimes \Sigma)^{-1}}_{(3)}(Z^\top \otimes I_k)\big)^{-1}\underbrace{(Z^\top \otimes I_k)^\top}_{(2)}\underbrace{(I_T \otimes \Sigma)^{-1}}_{(3)}Y \\
&= \big(\underbrace{(Z \otimes I_k)(I_T \otimes \Sigma^{-1})(Z^\top \otimes I_k)}_{(1)}\big)^{-1}\underbrace{(Z \otimes I_k)(I_T \otimes \Sigma^{-1})}_{(1)}Y \\
&= \big(ZZ^\top \otimes \Sigma^{-1}\big)^{-1}\big(Z \otimes \Sigma^{-1}\big)\boldsymbol{y} \\
&= \big((ZZ^\top)^{-1} \otimes \Sigma\big)\big(Z \otimes \Sigma^{-1}\big)\boldsymbol{y} \\
&= \big((ZZ^\top)^{-1}Z \otimes I_k\big)\boldsymbol{y}
\end{aligned}$$

Osservazioni

› Lo stimatore GLS non dipende dalla matrice di varianze e covarianze e risulta identico allo stimatore OLS. Infatti, prendendo i minimi quadrati (6) e usando le proprietà del prodotto di Kronecker,

$$\begin{aligned}
\hat{\beta}_{\text{ols}} &= (X^\top X)^{-1}X^\top \boldsymbol{y} \\
&= \big((Z^\top \otimes I_k)^\top (Z^\top \otimes I_k)\big)^{-1}(Z^\top \otimes I_k)^\top \boldsymbol{y} \\
&= \big(ZZ^\top \otimes I_k\big)^{-1}(Z \otimes I_k)\boldsymbol{y} \\
&= \big((ZZ^\top)^{-1} \otimes I_k\big)(Z \otimes I_k)\boldsymbol{y} \\
&= \big((ZZ^\top)^{-1}Z \otimes I_k\big)\boldsymbol{y}.
\end{aligned}$$

› Si può anche scrivere lo stimatore per il modello non vettorizzato, ricordando che $\boldsymbol{y} = \operatorname{vec} Y$:

$$\begin{aligned}\hat{\beta} &= \left((ZZ^\top)^{-1}Z \otimes I_k\right) \operatorname{vec} Y \\ &= \operatorname{vec}\left(Y\left(Z^\top(ZZ^\top)^{-1}\right)\right) \\ \hat{B} &= YZ^\top(ZZ^\top)^{-1}\end{aligned}$$

Da qui, siccome $B = (\Phi_0, \Phi_1, \Phi_2, \ldots, \Phi_p)$, si può ricavare immediatamente la stima delle matrici del processo.

7.2 Metodo dei momenti

Un'alternativa apprezzata dagli economisti è il metodo dei momenti, che per il VAR(p) porta alla stessa soluzione $\hat{\beta}$ e $\hat{B}$ ricavata dai GLS.

Ipotizzando incorrelazione tra innovazioni ed esplicative al tempo precedente, $\mathbb{E}\left[\varepsilon_t z_{t-1}^\top\right] = \mathbf{0}_{k\times k}$, si può scrivere

$$\begin{aligned}\mathbb{E}\left[\varepsilon_t z_{t-1}^\top\right] &= \mathbf{0}_{k\times k} \\ \mathbb{E}\left[(\boldsymbol{y}_t - Bz_{t-1})z_{t-1}^\top\right] &= \mathbf{0}_{k\times k} \qquad \text{(def.)} \\ \mathbb{E}\left[\boldsymbol{y}_t z_{t-1}^\top - Bz_{t-1}z_{t-1}^\top\right] &= \mathbf{0}_{k\times k},\end{aligned}$$

da cui si ottiene per linearità che

$$\begin{aligned}B\,\mathbb{E}\left[z_{t-1}z_{t-1}^\top\right] &= \mathbb{E}\left[\boldsymbol{y}_t z_{t-1}^\top\right] \\ B &= \mathbb{E}\left[\boldsymbol{y}_t z_{t-1}^\top\right]\mathbb{E}\left[z_{t-1}z_{t-1}^\top\right]^{-1}.\end{aligned}$$

Sostituendo i momenti empirici a quelli teorici, si ottiene nuovamente lo stimatore GLS

$$\hat{B} = \frac{1}{T}YZ^\top\left(\frac{1}{T}ZZ^\top\right)^{-1} = YZ^\top(ZZ^\top)^{-1}.$$

7.3 Stimatore equazione per equazione

Si possono vettorizzare anche entrambi i lati dell'equazione che definisce il modello lineare, per determinare uno stimatore che coincide con una regressione lineare per ciascuna equazione del VAR(p). Questo metodo viene usato per la stima dal pacchetto `vars` di R.

Il VAR k-dimensionale si può vedere come una sequenza di k regressioni sullo stesso vettore ritardato.

$$\begin{aligned}\overbrace{\operatorname{vec}(Y^\top)}^{\text{riga per riga}} &= \operatorname{vec}(Z^\top B^\top + \varepsilon^\top) \\ &= (I_k \otimes Z^\top)\operatorname{vec}(B^\top) + \operatorname{vec}(\varepsilon^\top) \\ \tilde{\boldsymbol{y}} &= (I_k \otimes Z^\top)\tilde{\beta} + \tilde{\boldsymbol{u}}.\end{aligned} \tag{8}$$

Il termine di errore è tale che $\mathbb{E}\left[\tilde{\boldsymbol{u}}\right] = \mathbf{0}$ e la matrice di varianze e covarianze è

$$\mathbb{E}\left[\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top\right] = \Sigma \otimes I_T = \begin{pmatrix} \sigma_{11}I_T & \sigma_{12}I_T & \dots & \sigma_{1k}I_T \\ \sigma_{21}I_T & \sigma_{22}I_T & \dots & \sigma_{2k}I_T \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1}I_T & \sigma_{k2}I_T & \dots & \sigma_{kk}I_T \end{pmatrix}.$$

Questa scrittura porta ad avere nel sotto-vettore $\tilde{\beta}_{1:(kp+1)}$ i coefficienti della prima equazione del VAR(p), in $\tilde{\beta}_{(kp+1):(2kp+2)}$ i coefficienti della seconda equazione, ...

Lo *stimatore equazione per equazione* risulta essere, dopo aver usato di nuovo le proprietà di $\otimes$ per semplificare, pari a

$$\begin{aligned} \hat{\tilde{\beta}} &= \left(I_k \otimes (ZZ^\top)^{-1}Z\right)\tilde{\boldsymbol{y}}. \\ &= \begin{pmatrix} (ZZ^\top)^{-1}Z & 0 & \dots & 0 \\ 0 & (ZZ^\top)^{-1}Z & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (ZZ^\top)^{-1}Z \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{y}}_1 \\ \tilde{\boldsymbol{y}}_2 \\ \vdots \\ \tilde{\boldsymbol{y}}_n \end{pmatrix} \\ &= \left[(ZZ^\top)^{-1}Z\tilde{\boldsymbol{y}}_j\right]_{j=1,2,\dots,k}. \end{aligned}$$

Una volta stimati i parametri, si può trovare lo stimatore di Σ attraverso i residui di regressione

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t^\top = \frac{1}{T}\hat{\varepsilon}\hat{\varepsilon}^\top,$$

dove la matrice dei residui è

$$\hat{\varepsilon} = Y - \hat{B}Z = Y - YZ^\top(ZZ^\top)Z = Y(I - \mathcal{P}_Z).$$

Il modello scritto nella forma (8) fa parte della classe dei modelli SURE (*Seemingly Unrelated Regression Equations*), sono modelli multivariati con struttura

$$y = X\beta + u$$

$$X = \begin{pmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & X_k \end{pmatrix}$$

Diverso dai modelli lineari classici, in quanto la matrice del disegno può essere diversa per ogni risposta e il termine di disturbo può avere una struttura di correlazione più complicata. Nel caso del VAR(p) si ha una struttura di correlazione a blocchi e matrice di disegno uguale per tutte le equazioni, dunque

$$X = I \otimes Z,$$

$$\mathbb{E}\left[uu^\top\right] = \Sigma \otimes I_k.$$

Il modello SURE si usa per fare test congiunti su parametri in equazioni diverse del sistema, ad

esempio per verificare se un parametro entra in una equazione o meno.

Lo stimatore per il SURE è il metodo GLS, a causa della struttura della matrice X.

Lezione 8: Proprietà dello stimatore LS

2020-10-19

Per poter descrivere le proprietà asintotiche, è necessario che valgano le seguenti ipotesi:

$$\begin{aligned}
&\textbf{Innovazione}: \varepsilon_t \sim \mathrm{WN}(0, \Sigma) \\
&\textbf{Incorrelazione}: \mathbb{E}\left[\varepsilon_t \varepsilon_s^\top\right] = \begin{cases} \Sigma & \text{se } s = t \\ 0 & \text{altrimenti} \end{cases} \\
&\textbf{Curtosi limitata}: \mathbb{E}\left[\varepsilon_{i,t}\varepsilon_{j,t}\varepsilon_{l,t}\varepsilon_{m,t}\right] \leq c \qquad \forall i,j,l,m,t.
\end{aligned} \tag{9}$$

Se si assume $\varepsilon_t \sim \mathrm{WNN}(0, \Sigma)$, allora le assunzioni (9) sono automaticamente soddisfatte.

Teo. (Distribuzione asintotiche)
Sotto le ipotesi (9)*, se il processo* $\boldsymbol{y}_t$ *è stabile e stazionario, allora*

(*i*) *Esiste una matrice* Γ *non singolare tale che* $\Gamma = \mathrm{plim}_{T\to\infty} \frac{ZZ^\top}{T}$ *(varianza* $\to$ *0).*

(*ii*) $\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathrm{vec}(\varepsilon_t z_{t-1}^\top) = \frac{1}{\sqrt{T}} \mathrm{vec}(\varepsilon Z^\top) = \frac{1}{\sqrt{T}}(Z \otimes I_k)\boldsymbol{u}$.

(*iii*) $\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathrm{vec}(\varepsilon_t z_{t-1}^\top) \xrightarrow{d} \mathcal{N}(0, \Gamma \otimes \Sigma)$.

Dim.
No.

□

8.1 Consistenza di $\hat{\boldsymbol{\beta}}$

Guardando alla consistenza di $\hat{\beta}$, si ha che

$$\begin{aligned}
&\operatorname*{plim}_{T\to\infty}(\hat{\beta} - \beta) \overset{\text{cont.}}{=} \mathrm{vec}\Big(\operatorname*{plim}_{T\to\infty}(\hat{B} - B)\Big), \\
&\operatorname*{plim}_{T\to\infty} \hat{B} \overset{\text{def.}}{=} \operatorname*{plim}_{T\to\infty} YZ^\top(ZZ^\top)^{-1}.
\end{aligned}$$

Si osserva inoltre che, come conseguenza del teorema,

$$\operatorname*{plim}_{T\to\infty} \frac{1}{T}\varepsilon Z^\top \overset{(2)}{=} 0 \implies \operatorname*{plim}_{T\to\infty}(\varepsilon Z^\top(ZZ^\top)^{-1}) = \underbrace{\operatorname*{plim}_{T\to\infty} \frac{1}{T}\varepsilon Z^\top}_{\longrightarrow 0} \underbrace{\operatorname*{plim}_{T\to\infty} \left(\frac{1}{T} ZZ^\top\right)^{-1}}_{\perp\!\!\!\perp \text{ da } T} = 0.$$

Dunque, mettendo assieme i due risultati, si ha che

$$\operatorname*{plim}_{T\to\infty} YZ^\top(ZZ^\top)^{-1} = B + \underbrace{\operatorname*{plim}_{T\to\infty} \left(\varepsilon Z^\top(ZZ^\top)^{-1}\right)}_{=\,0} = B.$$

Osservazioni

1. Lo stimatore LS di B per il processo VAR(p) risulta consistente.
2. Siccome vale se $\mathbb{E}\left[\varepsilon z_{t-1}^\top\right] = 0$, per avere stimatori consistenti è necessario che non ci sia correlazione tra il termine di errore e la matrice del disegno.

Teo. (Continuous mapping theorem)
Sia $g : \mathbb{R}^p \longrightarrow \mathbb{R}^m$ una funzione continua, allora se $(X_T)_{T\in\mathcal{T}}$ è una sequenza di variabili casuali sullo stesso probabilità tale che $X_T \xrightarrow{P} X$, allora

$$g(X_T) \xrightarrow{P} g(X).$$

Osservazione

Applicando il teorema si ha che

$$\operatorname*{plim}_{T\to\infty}\left(\frac{1}{T}ZZ^\top\right)^{-1} = \left(\operatorname*{plim}_{T\to\infty}\frac{1}{T}ZZ^\top\right)^{-1} = \Gamma^{-1}.$$

8.2 Normalità asintotica di $\hat{\beta}$

Per la normalità asintotica, conviene considerare $\hat{\beta}$ in forma vettoriale

$$\begin{aligned}\hat{\beta} &= \big((ZZ^\top)^{-1}Z \otimes I_k\big)\boldsymbol{y}\\ &= \big((ZZ^\top)^{-1}Z \otimes I_k\big)\big((Z^\top \otimes I_k)\beta + \boldsymbol{u}\big)\\ &= \beta + \big((ZZ^\top)^{-1}Z \otimes I_k\big)\boldsymbol{u},\end{aligned}$$

per cui

$$\begin{aligned}\sqrt{T}(\hat{\beta}-\beta) &= \sqrt{T}\big((ZZ^\top)^{-1}Z \otimes I_k\big)\boldsymbol{u}\\ &= \Big(\underbrace{\Big(\frac{1}{T}ZZ^\top\Big)^{-1} \otimes I_k}_{\xrightarrow{P}\,\Sigma^{-1}\otimes I_k}\Big)\quad \underbrace{\frac{1}{\sqrt{T}}(Z \otimes I_k)\boldsymbol{u}}_{(ii)+(iii)\ \xrightarrow{P}\ \mathcal{N}(0,\Gamma\otimes\Sigma)}.\end{aligned}$$

Ora, applicando le proprietà del prodotto di Kronecker,

$$(\Gamma^{-1} \otimes I_k)(\Gamma \otimes \Sigma)(\Gamma^{-1} \otimes I_k) = \Gamma^{-1} \otimes \Sigma,$$

per cui

$$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma).$$

Ora, nella convergenza dello stimatore si hanno due matrici ignote, date da Γ^{-1} e Σ. Esattamente come nel modello lineare, dove si effettua il plug-in

$$\sigma^2(X^\top X)^{-1} \rightsquigarrow \hat{\sigma}^2(X^\top X)^{-1},$$

anche in questo caso si utilizzano delle stime campionarie

$$\hat{\Gamma} = \frac{1}{T} ZZ^\top, \qquad \hat{\Sigma} = \frac{1}{T} \hat{\varepsilon}\hat{\varepsilon}^\top.$$

Dai risultati precedenti, $\hat{\Gamma} = \frac{1}{T} ZZ^\top$ è uno stimatore consistente per Γ, mentre per Σ si ha il seguente teorema.

8.3 Consistenza di $\hat{\Sigma}$

Teo. (Consistenza di $\hat{\Sigma}$)
Sia $\boldsymbol{y}_t$ un processo VAR(p) k-dimensionale stabile, con $\varepsilon_t \sim WN(0, \Sigma)$, e sia $\overline{B}$ uno stimatore dei coefficienti B tale che $\sqrt{T}(\overline{B} - B)$ converge in probabilità. Se si considera per $c \in \mathbb{R}^+$ l'analogo della somma dei quadrati dei residui

$$\hat{\Sigma}_c = \frac{(Y - \overline{B}Z)(Y - \overline{B}Z)^\top}{T - c},$$

allora

$$\operatorname*{plim}_{T\to\infty} \sqrt{T}(\hat{\Sigma}_c - \varepsilon\varepsilon^\top/T) = 0.$$

Osservazioni

1. Una volta trovato $\overline{B}$ che converge in distribuzione, con il plug-in si ottiene uno stimatore consistente di $\hat{\Sigma}$
2. Si considera $c \in \mathbb{R}^+$ per avere l'analogo dello stimatore corretto di σ^2.

Dim.
Slide 29 da leggere, si sviluppa il quadrato dei residui. La dimostrazione per la forma quadratica è elaborata.

□

Lo stimatore non distorto di Σ definito sulla base di questo risultato è

$$\tilde{\Sigma} = \frac{1}{T - k(p+1)} \hat{\varepsilon}\hat{\varepsilon}^\top.$$

Anche se è una bella proprietà, non è molto utile: lo stimatore bayesiano del VAR è distorto, ma con performance migliori.

La seguente proposizione chiarisce meglio il risultato precedente.

Proposizione 2.1

Date $\{x_T\}_{T\in\mathbb{T}}$ e $\{y_T\}_{T\in\mathbb{T}}$ é sequenze di variabili casuali appartenenti allo stesso spazio di probabilità , e y é una variabile casuale appartenente al medesimo spazio di probabilità, tali che $y_T \xrightarrow[T\to\infty]{\mathcal{D}} y$ ed inoltre $\text{plim}_{T\to\infty} y_T - x_T = 0$ allora anche $x_T \xrightarrow[T\to\infty]{\mathcal{D}} y$.

Osservazione 2.3

Quindi se $\sqrt{T}\text{vec}(\varepsilon\varepsilon^\intercal/T - \Sigma)$ ed inoltre il Teorema precedente stabilisce che

$$\text{plim}_{T\to\infty}\sqrt{T}\left(\frac{(Y-\overline{B}Z)(Y-\overline{B}Z)^\intercal}{T-c} - \frac{\varepsilon\varepsilon^\intercal}{T}\right) = 0,$$

ne deriva che $\sqrt{T}\text{vec}(\hat{\Sigma}-\Sigma)$ e $\sqrt{T}\text{vec}(\tilde{\Sigma}-\Sigma)$ hanno entrambe la stessa distribuzione limite e che questa distribuzione risulta indipendente dalla distribuzione limite di $\hat{B}$.

Questo risultato é particolarmente utile al fine di dimostrare la consistenza di entrambi gli stimatori $\tilde{\Sigma}$ e $\hat{\Sigma}$ di Σ.

Figura 1: propProdotto

Teo. (Consistenza degli stimatori di $\boldsymbol{\Sigma}$)

Sia $\boldsymbol{y}_t$ un processo VAR(p) k-dimensionale stabile, con $\varepsilon_t \sim WN(0,\Sigma)$, allora

$$\operatorname*{plim}_{T\to\infty}\hat{\Sigma} = \operatorname*{plim}_{T\to\infty}\frac{1}{T}\varepsilon\varepsilon^\top.$$

Dim.

Si ha che

$$\mathbb{E}\left[\frac{1}{T}\varepsilon_t\varepsilon_t^\top\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\varepsilon_t\varepsilon_t^\top\right] = \Sigma$$

$$\mathbb{V}\left[\frac{1}{T}\operatorname{vec}\varepsilon\varepsilon^\top\right] = \frac{1}{T^2}\sum_{t=1}^{T}\underbrace{\mathbb{V}\left[\operatorname{vec}\varepsilon_t\varepsilon_t^\top\right]}_{\text{momento quarto}} \leq \frac{T\nu}{T^2}\xrightarrow{t\to\infty} 0,$$

dove l'ultima diseguaglianza vale per ν costante, che esiste sempre se il momento quarto è limitato (per Hp.).

□

Lezione 9: Inferenza nei processi VAR(p)

2020-10-20

9.1 Test di ipotesi

Si osserva che

$$\frac{\hat{\beta}_i - \beta_i}{\hat{s}_i} \xrightarrow{d} \mathcal{N}(0,1),$$

dove $\hat{s}_i^2 = \left[(ZZ^\top)^{-1} \otimes \hat{\Sigma}\right]_{(i,i)}$.

Con l'approccio di verosimiglianza si può ricavare un test rapporto di verosimiglianza (LRT) per verificare ipotesi sui parametri. Sia C matrice $N \times (k^2p+k)$, si vuole verificare un insieme di ipotesi *lineari* di tipo

$$\begin{cases} H_0 : C\beta = c \\ H_1 : \overline{H}_0 \end{cases}$$

Si può effettuare la procedura standard per verificare il test, stimando i parametri $\hat{\beta}$ del modello non vincolato e $\hat{\beta}_r$ del modello vincolato e poi calcolando

$$\hat{\lambda}_{\text{LR}} = 2\left\{\ell(\hat{\beta}) - \ell(\hat{\beta}_r)\right\} \xrightarrow{d} \chi^2_N,$$

con N numero di restrizioni sui parametri.

Alternativamente si può usare il test di Wald, che è più veloce da effettuare e si basa sul *metodo Delta*:

Prop.
Sia $\hat{\beta}$ tale che

$$\sqrt{T}\big(\hat{\beta} - \beta\big) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma),$$

allora se $g : \mathbb{R}^{k(kp+1)} \to \mathbb{R}^N$, $g \in C^1$, vale che

$$\sqrt{T}\big(g(\hat{\beta}) - g(\beta)\big) \xrightarrow{d} \mathcal{N}\Big(0, \frac{\partial g(\beta)}{\partial \beta}(\Gamma^{-1} \otimes \Sigma)\frac{\partial g(\beta)}{\partial \beta}^\top\Big)$$

Conseguenza

Nel caso del test di ipotesi, la funzione è $g(\beta) = C\beta - c$, per cui

$$\sqrt{T}(C\hat{\beta} - C\beta) \xrightarrow{d} \mathcal{N}\big(0, C(\Gamma^{-1} \otimes \Sigma)C^\top\big),$$

e di conseguenza, poiché C applica N restrizioni,

$$T(C\hat{\beta} - C\hat{\beta})^\top \big(C(\Gamma^{-1} \otimes \Sigma)C^\top\big)^{-1}(C\hat{\beta} - C\beta) \xrightarrow{d} \chi^2_N.$$

Dunque, sotto H_0

$$\hat{\lambda}_W = (C\hat{\beta} - c)^\top \big(C((ZZ^\top)^{-1} \otimes \hat{\Sigma})C^\top\big)^{-1}(C\hat{\beta} - c) \xrightarrow{d} \chi^2_N.$$

Osservazioni

› Non richiede una doppia stima del modello come il LRT.

› Vale solo per g differenziabili.

9.2 Selezione dell'ordine p del modello

Si può stimare un VAR(p) solo se si fissa a priori il valore di p. Nei processi VAR(p) le funzioni di cross-correlazione tante e più complicate di quelle degli ARMA(p, q), per cui vengono usate nell'analisi diagnostica. La selezione dell'ordine è affidata invece a test statistici o criteri di selezione.

Test sequenziali

Si definisce un range $0, 1, \dots, M$ e si effettua una sequenza di ipotesi nulle

$$\begin{aligned}
&H_0 : \Phi_M = 0 && \text{vs} \quad H_1 : \Phi_M \neq 0 \\
&H_0 : \Phi_{M-1} = 0 && \text{vs} \quad H_1 : \Phi_{M-1} \neq 0 | \Phi_M = 0 \\
&H_0 : \Phi_{M-2} = 0 && \text{vs} \quad H_1 : \Phi_{M-1} \neq 0 | \Phi_M = 0, \Phi_{M-1} = 0 \\
&\vdots \\
&H_0 : \Phi_1 = 0 && \text{vs} \quad H_1 : \Phi_1 \neq 0 | \Phi_M = 0, \dots, \Phi_2 = 0
\end{aligned}$$

L'ordine viene individuato al primo rifiuto dell'ipotesi nulla, con il valore di M che si sceglie in base alla dimensione campionaria e alla frequenza dei dati.

Criteri di informazione

Si possono usare i criteri di informazione per scegliere l'ordine del processo. Data la verosimiglianza per $\boldsymbol{u} = \boldsymbol{y} - (Z^\top \otimes I_k)\beta$,

$$\ell_T(\beta, \Sigma) = -\frac{kT}{2} \log 2\pi - \frac{T}{2} \log |\Sigma| - \frac{1}{2} \boldsymbol{u}^\top (I_T \otimes \Sigma)^{-1} \boldsymbol{u},$$

con il termine quadratico che è

$$\big(\boldsymbol{y} - (Z^\top \otimes I_k)\beta\big)^\top (I_T \otimes \Sigma)^{-1} \big(\boldsymbol{y} - (Z^\top \otimes I_k)\beta\big) \tag{10}$$

Si ha che

$$\begin{aligned}
(10) &= \big(\operatorname{vec}(Y - \hat{B}Z)\big)^\top \big(I_T \otimes \hat{\Sigma}\big)^{-1} \operatorname{vec}(Y - \hat{B}Z) \\
&= \operatorname{tr}\big[\big(\operatorname{vec}(Y - \hat{B}Z)\big)^\top (I_T \otimes \hat{\Sigma})^{-1} \operatorname{vec}(Y - \hat{B}Z)\big)\big] \\
&= \operatorname{tr}\big[(Y - \hat{B}Z)^\top \hat{\Sigma}^{-1} (Y - \hat{B}Z)\big] \\
&= \operatorname{tr}\big[\hat{\Sigma}^{-1} (Y - \hat{B}Z)(Y - \hat{B}Z)^\top\big].
\end{aligned} \tag{11}$$

Usando l'espressione dello stimatore ML di Σ (uguale a LS)

$$\hat{\Sigma} = \frac{\hat{\varepsilon}\hat{\varepsilon}^\top}{T} = \frac{1}{T}(Y - \hat{B}Z)(Y - \hat{B}Z)^\top,$$

e sostituendo in (11), si ottiene

$$\begin{aligned}\operatorname{tr}\left(\hat{\Sigma}^{-1}(Y-\hat{B}Z)(Y-\hat{B}Z)^\top\right) &= \operatorname{tr}\left(T((Y-\hat{B}Z)^\top)^{-1}(Y-\hat{B}Z)^{-1}(Y-\hat{B}Z)(Y-\hat{B}Z)^\top\right)\\ &= T\operatorname{tr}(I_k)\\ &= T\cdot k.\end{aligned}$$

Di conseguenza, nel punto di massimo vale

$$\ell_T(\hat{\beta},\hat{\Sigma}) = \underbrace{-\frac{kT}{2}(1+\log 2\pi)}_{\text{indip. da } \hat{\beta},\hat{\Sigma}} - \frac{T}{2}\log|\hat{\Sigma}|.$$

In conclusione, rimuovendo il termine indipendente dai parametri, il criterio di Akaike è definito da

$$\mathrm{AIC}(m) = \log|\hat{\Sigma}_m| + \underbrace{\frac{2mk^2}{T}}_{\text{penalità}}$$

e l'ordine ottimo $\hat{p}$ viene scelto per minimizzazione

$$\hat{p} = \operatorname*{argmin}_m AIC(m).$$

Lo stimatore di $\hat{p}$ con il criterio AIC non gode però di buone proprietà. Si dice che uno stimatore di p è *consistente* se

$$\hat{p} \xrightarrow{P} p.$$

Si consideri una famiglia di stimatori basati su una struttura simile all'AIC,

$$\hat{p} = \log|\hat{\Sigma}_m| + \frac{m}{T}c_m,$$

con c_m funzione non decrescente di T. Allora, lo stimatore $\hat{p}$ è consistente se e solo se

$$\begin{cases} c_T \xrightarrow{T\to\infty} \infty \\ \frac{c_T}{T} \xrightarrow{T\to\infty} 0 \end{cases} \tag{12}$$

inoltre si dice che $\hat{p}$ è *fortemente consistente* se vale anche

$$\frac{c_T}{2\log\log T} > 1. \tag{13}$$

Prop.
AIC non è consistente.

Dim.
Per l'AIC, si ha che la funzione $c_T = 2T^2$ è tale che

$$\lim_{T\to\infty}\frac{2T^2}{T} = +\infty,$$

per cui non vale (12) e dunque non è consistente.

□

Due criteri di selezione consistenti sono il criterio di Hannah-Quinn e il Bayesian Information Criterion:

$$\mathrm{HQ}(m) = \log|\hat{\Sigma}_m| + mk^2 \frac{2\log\log T}{T},$$

$$\mathrm{BIC}(m) = \log|\hat{\Sigma}_m| + mk^2 \frac{2\log T}{T}.$$

Il criterio HQ è fortemente consistente se $k > 1$, mentre BIC lo è sempre. La gente usa l'AIC perché deriva dalla divergenza di Kullback-Leibler, mentre HQ non ha giustificazione teorica. BIC è consistente e ha giustificazione teorica in chiave bayesiana.

Lezione 10: Previsione

2020-10-26

Si costruisce un modello sia per catturare evidenze empiriche, sia per effettuare previsioni dell'evoluzione futura dei fenomeni di interesse.

Def. (Definizioni preliminari)

- $\mathcal{F}_t$: *set informativo* contenente i dati storici di $\boldsymbol{y}_t$ fino all'istante t. È una filtrazione, ovvero una σ-algebra tale che $\mathcal{F}_j \subset \mathcal{F}_{j+1}$ per ogni j.
- h : *orizzonte di previsione*
- $\boldsymbol{y}_{t+h|t}$: *previsione* per $\boldsymbol{y}_{t-h}$ fatta sulla base di $\mathcal{F}_t$.
- $\hat{\boldsymbol{e}}_t(h) = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}$: *errore di previsione* nel prevedere $\boldsymbol{y}_{t+h}$ usando le informazioni al tempo t.

10.1 Previsione puntuale

Le previsioni al tempo t sono basate su $\mathbb{E}\left[\,\cdot\,|\mathcal{F}_t\right]$, operatore che calcola il valore atteso condizionato all'informazione al tempo t:

$$\boldsymbol{y}_{j|t} = \begin{cases} \mathbb{E}\left[\boldsymbol{y}_j|\mathcal{F}_t\right] & \text{se } j > t \\ \boldsymbol{y}_j & \text{se } j \leq t \end{cases},$$

$$\mathbb{E}\left[\varepsilon_{t+j}|\mathcal{F}_t\right] = 0 \qquad \text{se j} > 0.$$

Per esempio, al tempo $t+2$ si ha

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{y}_{t+2}|\mathcal{F}_t\right] &= \mathbb{E}\left[\Phi_0 + \Phi_1\boldsymbol{y}_{t+1} + \Phi_2\boldsymbol{y}_t + \dots\right] \\ &= \Phi_0 + \Phi_1\,\mathbb{E}\left[\boldsymbol{y}_{t+1}|\mathcal{F}_t\right] + \Phi_2\boldsymbol{y}_t + \dots \qquad \text{(linearità)} \\ &= \Phi_0 + \Phi_1\hat{\boldsymbol{y}}_{t+1|t} + \Phi_2\boldsymbol{y}_t + \dots \end{aligned}$$

Osservazioni

- In un processo VAR(p) stabile, la funzione di previsione soddisfa l'equazione alle differenze prime
$$\hat{\boldsymbol{y}}_{t+h|t} = \Phi_0 + \Phi_1\hat{\boldsymbol{y}}_{t+h-1|t} + \Phi_2\hat{\boldsymbol{y}}_{t+h-2|t} + \dots + \Phi_p\hat{\boldsymbol{y}}_{t+h-p|t}, \quad \text{per } h > p,$$
e il problema di Cauchy associato risulta definito con le condizioni iniziali $\hat{\boldsymbol{y}}_{t+j|t}$ per $j = 0, 1, \dots, p-1$. In questo caso, le condizioni iniziali vengono date dalle prime iterazioni della funzione di previsione, per cui è equivalente dire che le condizioni iniziali sono $\boldsymbol{y}_{t-j}$ per $j = 0, 1, \dots, p-1$.
- Per un VAR(1) stabile, si ha l'equazione alle differenze prime
$$\hat{\boldsymbol{y}}_{t+h|t} = \Phi_0 + \Phi_1\hat{\boldsymbol{y}}_{t+h-1|t}, \quad \text{per } h > 1,$$

con il problema di Cauchy definito dalla condizione iniziale $\boldsymbol{y}_t$.

$$\begin{aligned}
\hat{\boldsymbol{y}}_{t+1|t} &= \Phi_0 + \Phi_1 \boldsymbol{y}_t \\
\hat{\boldsymbol{y}}_{t+2|t} &= \Phi_0 + \Phi_1 \hat{\boldsymbol{y}}_{t+1|t} \\
&= (I_k + \Phi_1)\Phi_0 + \Phi_1^2 \hat{y}_t \\
&\vdots \\
\hat{\boldsymbol{y}}_{t+h|t} &= \Phi_0 + \Phi_1 \hat{\boldsymbol{y}}_{t+h-1|t} \\
&= (I_k + \Phi_1 + \Phi_1^2 + \ldots + \Phi_1^{h-1})\Phi_0 + \Phi_1^h \boldsymbol{y}_t \\
&\xrightarrow{h\to\infty} \boldsymbol{\mu} \quad \text{se stazionario, poiché valgono (1) e (2).}
\end{aligned} \tag{14}$$

10.2 Errore di previsione

VAR(1)

Effettuando $h-1$ sostituzioni ricorsive, il VAR(1) diventa

$$\boldsymbol{y}_{t+h} = \underbrace{(I_k + \Phi_1 + \Phi_2^2 + \ldots + \Phi_1^{h-1})\Phi_0 + \Phi_1^h \boldsymbol{y}_t}_{\text{deterministico}} + \underbrace{\varepsilon_{t+h} + \Phi_1 \varepsilon_{t+h-1} + \Phi^2 \varepsilon_{t+h-2} + \ldots + \Phi_1^{h-1} \varepsilon_{t+1}}_{\text{valore atteso nullo}}.$$

L'errore di previsione è allora

$$\begin{aligned}
\hat{\boldsymbol{e}}_t(h) &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} \\
&= \sum_{j=0}^{h-1} \Phi_1^j \varepsilon_{t+h-j},
\end{aligned}$$

ovvero una media mobile di ordine h delle innovazioni, da cui si può calcolare la varianza di $\hat{\boldsymbol{e}}_t(h)$ dall'incorrelazione tra ε_j e ε_k. Notare che $\Phi_1^j = \Psi_j$ della rappresentazione VMA(∞) per il VAR(1).

VAR(p)

Per il VAR(p) generico le cose si complicano molto e conviene invece usare diverse strade

a) Rappresentazione VMA(∞).

b) Forma canonica.

a) Rappresentazione VMA(∞)

La rappresentazione VMA(∞) è

$$\begin{aligned}
\boldsymbol{y}_t &= \mu + \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j} \\
&= \mu + \Psi(L)\varepsilon_t,
\end{aligned}$$

con (come visto nella Lezione 6)

$$\Psi_k = \sum_{j=1}^{\min\{k,p\}} \Psi_{k-j}\Phi_j, \quad \Psi_0 = I_k,$$

$$\mu = \Phi(1)^{-1}\Phi_0 = (I_k - \Phi_1 - \ldots - \Phi_p)^{-1}\Phi_0.$$

Le previsioni sono

$$\begin{aligned}\hat{\boldsymbol{y}}_{t+1|t} &= \mathbb{E}\left[\boldsymbol{y}_{t+1}|\mathcal{F}_t\right] \\ &= \mathbb{E}\left[\mu + \cancel{\varepsilon_{t+1}} + \Psi_1\varepsilon_t + \Psi_2\varepsilon_{t-1} + \ldots|\mathcal{F}_t\right] \\ &= \mu + \sum_{j=1}^{\infty}\Psi_j\varepsilon_{t+1-j}.\end{aligned}$$

Ripetendo il ragionamento per un orizzonte h generico, si ha

$$\hat{\boldsymbol{y}}_{t+h|t} = \mu + \sum_{j=h}^{\infty}\Psi_j\varepsilon_{t+h-j}.$$

L'errore di previsione è quindi

$$\begin{aligned}\hat{\boldsymbol{e}}_t(1) &= \boldsymbol{y}_{t+1} - \hat{\boldsymbol{y}}_{t+1|t} = \varepsilon_{t+1} \\ \hat{\boldsymbol{e}}_t(2) &= \varepsilon_{t+2} + \Psi_1\varepsilon_{t+1} \\ &\vdots \\ \hat{\boldsymbol{e}}_t(h) &= \sum_{j=0}^{h-1}\Psi_j\varepsilon_{t+h-j},\end{aligned} \tag{15}$$

cioè ancora una volta è una media mobile degli errori $\{\varepsilon_{t+h-j}\}_{j=1}^{h}$. Come nell'AR(1), le matrici dei pesi sono le Ψ_j.

b) Forma canonica

Data la forma canonica del VAR(p)

$$X_t = A_0 + A_1X_{t-1} + u_t,$$

sfruttandone la rappresentazione VAR(1), si può applicare la ricorsione (14) sostituendo A_0 e A_1 al posto di Φ_0 e Φ_1, per ottenere

$$\begin{aligned}\hat{X}_{t+h|t} &= A_0 + A_1\hat{X}_{t+h-1|t} \\ &= (I_{kp} + A_1 + A_1^2 + \ldots + A_1^{h-1})A_0 + A_1^hX_t,\end{aligned}$$

da cui

$$\begin{aligned}\boldsymbol{e}_t(h) &= X_{t+h} - \hat{X}_{t+h|t} \\ &= \sum_{j=0}^{h-1}A_1^ju_{t+h-j}.\end{aligned}$$

I valori previsti si ricavano usando la matrice di selezione $J = [I_k \;\; \mathbf{0}_{k\times k} \;\; \mathbf{0}_{k\times k} \;\; \dots \;\; \mathbf{0}_{k\times k}]$

$$\begin{aligned}\hat{\boldsymbol{y}}_{t+h|t} = J\hat{X}_{t+h|t} \implies \boldsymbol{e}_t &= \sum_{j=0}^{h-1} JA_1^j u_{t+h-j} \\ &= \sum_{j=0}^{h-1} JA_1^j J^\top J u_{t+h-j} \qquad (J^\top J u_t = u_t).\end{aligned}$$

Proprietà dell'errore di previsione

› Il *valore atteso* è $\mathbb{E}\big[\hat{\boldsymbol{e}}_t(h)\big] = 0$, in quanto $\hat{\boldsymbol{e}}_t(h)$ è media mobile di ε_{t-j} a media nulla.

› La *varianza dell'errore di previsione* corrisponde a

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{e}}_t(h)\hat{\boldsymbol{e}}_t(h)^\top\right] &= \mathbb{E}\left[\big(\sum_{j=0}^{h-1}\Psi_j\varepsilon_{t+h-j}\big)\big(\sum_{j=0}^{h-1}\Psi_j\varepsilon_{t+h-j}\big)^\top\middle|\mathcal{F}_t\right] \\ &= \sum_{j=0}^{h-1}\Psi_j\Sigma\Psi_j^\top \\ &= \mathrm{MSE}(h).\end{aligned} \tag{16}$$

Il MSE soddisfa la relazione

$$\mathrm{MSE}(h) = \Psi_{h-1}\Sigma\Psi_{h-1}^\top + \mathrm{MSE}(h-1).$$

Con la forma canonica, si sarebbe ottenuto

$$\begin{aligned}\mathrm{MSE}_Y(h) &= J\ \mathrm{MSE}_X(h)\ J^\top \\ &= \sum_{j=0}^{h-1} JA_1^j J^\top J\Sigma_u J^\top (A_1^j)^\top J^\top, \qquad \Sigma_u = \mathrm{diag}(\Sigma, \mathbf{0}_{k\times k}, \dots, \mathbf{0}_{k\times k}).\end{aligned}$$

› Sotto ipotesi di stazionarietà, prendendo il limite in (16) per $h \to \infty$, si ha

$$\lim_{h\to\infty}\hat{\boldsymbol{y}}_{t+h|t} = \mu,$$

$$\lim_{h\to\infty}\mathrm{MSE}(h) = \sum_{j=0}^{\infty}\Psi_j\Sigma\Psi_j^\top = \Gamma(0).$$

› Sotto ipotesi di gaussianità, siccome dalla (15) si ha che $\boldsymbol{e}_t$ è combinazione lineare degli ε_j, allora

$$\hat{\boldsymbol{e}}_t(h) \sim \mathcal{N}_k\big(0, \mathrm{MSE}(h)\big),$$

e di conseguenza anche le singole previsioni

$$\hat{y}_{i,t+h|t} \sim \mathcal{N}\big(y_{i,t+h}, \hat{\sigma}^2_{i,t+h|t}\big)$$

e l'intervallo di previsione per la singola componente è

$$\hat{y}_{i,t+h|t} \pm z_{1-\frac{\alpha}{2}} \hat{\sigma}_{i,t+h|t},$$

con $\hat{\sigma}_{i,t+h|t}$ elemento sulla diagonale i-esimo di $\mathrm{MSE}(h)$. Alternativamente, si possono costruire regioni di previsioni congiunte per tutte le osservazioni

$$(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})^\top \mathrm{MSE}(h)^{-1} (\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}) \sim \chi^2(k),$$

o per un sottoinsieme di dimensione m, tramite la matrice di selezione $F = \begin{bmatrix} I_m & \mathbf{0}_{m\times(k-m)} \end{bmatrix}$

$$(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})^\top F^\top \mathrm{MSE}(h)^{-1} F (\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t}) \sim \chi^2(m).$$

In generale si corregge la regione per bande di confidenza multiple usando la correzione di Bonferroni. Allo stesso modo, ci si può concentrare su una singola componente e correggere la previsione su *più passi*.

Lezione 11: Previsioni (cont.)

2020-10-27

Prop. (Ottimalità delle previsioni)
Le previsioni definite come

$$\hat{\boldsymbol{y}}_{t+h|t} = \mathbb{E}\left[\boldsymbol{y}_{t+h}|\mathcal{F}_t\right]$$

minimizzano la varianza (MSE) di ciascuna componente di $\boldsymbol{y}_t$.

Dim.
Dato $\overline{\boldsymbol{y}}_{t+h}$ un generico previsore, si ottiene

$$\begin{aligned}\mathrm{MSE}(\overline{\boldsymbol{y}}_{t+h}) &= \mathbb{E}\left[(\boldsymbol{y}_{t+h} - \overline{\boldsymbol{y}}_{t+h})(\boldsymbol{y}_{t+h} - \overline{\boldsymbol{y}}_{t+h})^\top|\mathcal{F}_t\right] \\ &= \mathbb{E}\left[(\boldsymbol{y}_{t+h} \pm \hat{\boldsymbol{y}}_{t+h|t} - \overline{\boldsymbol{y}}_{t+h})(\boldsymbol{y}_{t+h} \pm \hat{\boldsymbol{y}}_{t+h|t} - \overline{\boldsymbol{y}}_{t+h})^\top|\mathcal{F}_t\right] \\ &= \mathrm{MSE}(\hat{\boldsymbol{y}}_{t+h|t}) + \begin{cases}\mathbb{E}\left[(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})(\hat{\boldsymbol{y}}_{t+h|t} - \overline{\boldsymbol{y}}_{t+h})^\top|\mathcal{F}_t\right] & (2) \\ \mathbb{E}\left[(\boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t})(\boldsymbol{y}_{t+h} - \overline{\boldsymbol{y}}_{t+h})^\top|\mathcal{F}_t\right] & (3) \\ \mathbb{E}\left[(\hat{\boldsymbol{y}}_{t+h|t} - \overline{\boldsymbol{y}}_{t+h})(\hat{\boldsymbol{y}}_{t+h|t} - \overline{\boldsymbol{y}}_{t+h})^\top|\mathcal{F}_t\right] & (4)\end{cases}\end{aligned}$$

Dal momento che $(2) = 0,\ (3) = 0,\ (4) \geq 0$, si conclude che

$$\mathrm{MSE}(\overline{\boldsymbol{y}}_{t+h}) \geq \mathrm{MSE}(\hat{\boldsymbol{y}}_{y+h|t}).$$

□

Rimuovendo l'ipotesi di innovazioni i.i.d, come nel caso dei VARMA (che non vedremo), si ha comunque un risultato di ottimalità per $\hat{\boldsymbol{y}}_{t+h|t}$.

Prop. (Ottimalità tra i previsori lineari)
Il previsore $\hat{\boldsymbol{y}}_{t+h|t}$ *è quello con minimo MSE per* $\boldsymbol{y}_{t+h}$ *nella classe dei previsori lineari.*

Osservazione

Nel machine learning si vuole violare la linearità (ad es. con metodi ensemble, reti neurali, ...), perché se ci si limitasse alla linearità si avrebbe già il previsore ottimo.

11.1 Previsione con errori di stima

Fin'ora si assumeva di conoscere i veri $\Phi_0, \Phi_1, \ldots$, mentre ora si vuole valutare cosa succede aggiungendo l'incertezza sui parametri ignoti stimati:

$$\hat{\boldsymbol{y}}^*_{t+h|t} = \hat{\Phi}_0 + \hat{\Phi}_1\hat{\boldsymbol{y}}^*_{t+h-1|t} + \hat{\Phi}_2\hat{\boldsymbol{y}}^*_{t+h-2|t} + \ldots + \hat{\Phi}_p\hat{\boldsymbol{y}}^*_{t+h-p|t},$$

con la rappresentazione VMA(∞)

$$\hat{\boldsymbol{y}}^*_{t+h|t} = \hat{\mu} + \sum_{j=h}^{\infty} \hat{\Psi}_j \varepsilon_{t+h-j}.$$

Gli errori di previsione sono ora

$$\begin{aligned}\hat{\boldsymbol{e}}_t(h) &= \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}^*_{t+h|t} \\ &= \boldsymbol{y}_{t+h} \pm \hat{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}^*_{t+h|t} \\ &= \underbrace{\sum_{j=0}^{h-1} \Psi_j \varepsilon_{t+h-j}}_{(1)} + \underbrace{\hat{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}^*_{t+h|t}}_{(2)}\end{aligned}$$

Prop.

Sotto ipotesi piuttosto generali, si ha che:

- *I termini* (1) *e* (2) *sono indipendenti.*
- $E\big[\hat{\boldsymbol{e}}_t(h)\big] = 0$ *e la varianza è*

$$\mathbb{E}\left[\hat{\boldsymbol{e}}_t(h)\hat{\boldsymbol{e}}_t(h)^\top\right] = MSE(h) + \mathbb{E}\left[\big(\hat{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}^*_{t+h|t}\big)\big(\hat{\boldsymbol{y}}_{t+h|t} - \hat{\boldsymbol{y}}^*_{t+h|t}\big)^\top\right].$$

Dim.

No.

□

Osservazione

Il valore atteso non è facile da calcolare, se non con un'approssimazione asintotica.

Prop. (Varianza asintotica per $h = 1$)

Se l'orizzonte è $h = 1$, si ha che

$$\mathbb{E}\left[\hat{\boldsymbol{e}}_t(h)\hat{\boldsymbol{e}}_t(h)^\top\right] = \Sigma + \frac{kp+1}{T}\Sigma = \frac{T+kp+1}{T}\Sigma.$$

Dim.

No.

□

Osservazione

Akaike ha suggerito di usare questa varianza asintotica come criterio di selezione dell'ordine p del processo. Dopo aver corretto lo stimatore di Σ con

$$\hat{\Sigma}_m = \frac{T}{T - km - 1}\left(\frac{\hat{\varepsilon}_m \hat{\varepsilon}_m^\top}{T}\right),$$

si può usare il criterio del *Final Prediction Error*

$$\mathrm{FPE}(m) = \left(\frac{T + km + 1}{T - km - m}\right)^k \left|\frac{\hat{\varepsilon}_m \hat{\varepsilon}_m^\top}{T}\right|.$$

Si dimostra che l'FPE non è consistente, in quanto

$$\log \mathrm{FPE}(m) = \mathrm{AIC}(m) + \frac{2k}{T} + O(T^{-2}).$$

Le previsioni calcolate con un numero fisso di passi h al variare del set informativo si dicono *previsioni dinamiche*

$$\hat{\boldsymbol{y}}_{t+j+1|t+j} = \hat{\mu}_{(t)} + \sum_{l=1}^{\infty} \hat{\Psi}_{l,(t)} \varepsilon_{t+j+1-l}, \quad j = 1, 2, \ldots, M.$$

Di solito si usano delle *rolling window* di ampiezza fissata.

11.2 Analisi strutturali

Le analisi viste fin'ora non permettono di apprezzare l'effettiva dinamicità del sistema, ad esempio se un'altra variabile subisce uno *shock*.

Lezione 12: Analisi strutturali

2020-11-06

Il VAR(p) contiene molti parametri e interpretarli spesso è difficili, a causa delle varie interazioni. Le proprietà dinamiche si sintetizzano allora tramite diversi tipi di *analisi strutturali* complementari:

1. Decomposizione della varianza dell'errore di previsione (FEVD).
2. Analisi di causalità secondo Granger.
3. Analisi delle *impulse response functions* (IRF).

Tutte queste analisi si basano sulla rappresentazione VMA(∞).

12.1 Forecast Error Variance Decomposition (FEVD)

Si misura il contributo dell'innovazione di y_j alla varianza dell'errore di previsione h passi in avanti sulla variabile y_i. In particolare, si misura quanto effetto può avere uno shock su y_i sulla varianza dell'errore di previsione di y_j:

$$\text{shock } y_{i+h} \xrightarrow{??} \mathbb{V}\left[\hat{y}_{j,t+h|t}\right].$$

Per questa analisi è particolarmente conveniente utilizzare la rappresentazione VMA(∞), perché è la somma di shock incorrelati temporalmente.

Supposta $\mu = 0$, nella rappresentazione VMA(∞)

$$\boldsymbol{y}_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j}, \quad \varepsilon_s \sim \text{WN}(0, \Sigma),$$

le innovazioni sono correlate a causa della varianza Σ. Per questo, si *ortogonalizzano le innovazioni* con la decomposizione di Cholesky $\Sigma = PP^\top$, da cui

$$\nu_t = P^{-1}\varepsilon_t \implies \mathbb{E}\left[\nu_t \nu_t^\top\right] = I_k.$$

Moltiplicando per PP^{-1}, si può scrivere la VMA(∞) in termini di innovazioni ortogonali

$$\boldsymbol{y}_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j} = \sum_{j=0}^{\infty} \Psi_j P P^{-1} \varepsilon_{t-j} = \sum_{j=0}^{\infty} \Theta_j \underbrace{\nu_{t-j}}_{\text{ortogonali}}, \tag{17}$$

da cui segue che, se $\vartheta_{j;i} = (\vartheta_{j;i,1} \quad \vartheta_{j;i,2} \quad \dots \quad \vartheta_{j;i,k})$ è la i-esima riga di Θ_j,

$$y_{i,t} = \sum_{j=0}^{\infty} \sum_{l=1}^{k} \boldsymbol{\vartheta}_{j;i,l} \cdot \nu_{l,t-j}.$$

Dunque, ha forma di combinazione di shock ortogonali sia nel tempo sia tra loro. Con questa rappresentazione, gli errori di previsione diventano

$$\hat{\boldsymbol{e}}_t(h) = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h|t} \overset{(15)}{=} \sum_{j=0}^{h-1} \Psi_j \varepsilon_{t+h-j} \overset{(PP^{-1})}{=} \sum_{j=0}^{h-1} \Theta_j \nu_{t+h-j}.$$

Prop. (FEVD)

Poiché le innovazioni sono tali che $\mathbb{E}\left[\nu_j \nu_j^\top\right] = I_k$*, per la i-esima osservazione si ha che*

$$\mathbb{E}\left[e_{i,t}^2(h)\right] = \sum_{j=0}^{h-1}\sum_{l=1}^{k} \vartheta_{j;i,l}^2$$

$$= \underbrace{\sum_{j=0}^{h-1} \vartheta_{j;i,1}^2}_{\text{contrib. di } \nu_{1,t+h-j}} + \underbrace{\sum_{j=0}^{h-1} \vartheta_{j;i,2}^2}_{\text{contrib. di } \nu_{2,t+h-j}} + \ldots + \underbrace{\sum_{j=0}^{h-1} \vartheta_{j;i,k}^2}_{\text{contrib. di } \nu_{k,t+h-j}}$$

Def. (FEVD)

Normalizzando rispetto al totale, si ottiene il contributo di FEVD della variabile l sulla variabile i

$$\text{FEVD}_{i,l}(h) = \frac{\sum_{j=0}^{h-1} \vartheta_{j;i,l}^2}{\sum_{j=0}^{h-1}\sum_{l=1}^{k} \vartheta_{j;i,l}^2}$$

Osservazione

La FEVD indica quanto una variabile y_j è determinante o meno nell'errore di previsione futuro per y_i. In generale si rappresenta sotto forma di matrice, dove l'elemento (i, j) è la proporzione di variabilità di y_j su y_i

	y_1	y_2	y_3	$\cdots$	y_k
y_1	$\text{FEVD}_{1,1}$	$\text{FEVD}_{1,2}$	$\text{FEVD}_{1,3}$	$\ldots$	$\text{FEVD}_{1,k}$
y_2	$\text{FEVD}_{2,1}$	$\text{FEVD}_{2,2}$	$\text{FEVD}_{2,3}$	$\ldots$	$\text{FEVD}_{2,k}$
y_3	$\text{FEVD}_{3,1}$	$\text{FEVD}_{3,2}$	$\text{FEVD}_{3,3}$	$\ldots$	$\text{FEVD}_{3,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_k	$\text{FEVD}_{k,1}$	$\text{FEVD}_{k,2}$	$\text{FEVD}_{k,3}$	$\ldots$	$\text{FEVD}_{k,k}$

12.2 Analisi di causalità

Un altro strumento per valutare l'influenza tra variabili è lo studio della ***causalità***. Ci sono diversi tipi di causalità, qui ne introduciamo una usata in particolare in ambito econometrico.

Def. (Causalità secondo Granger)

Se una variabile, o un gruppo di variabili Z è di ausilio nel migliorare le previsioni di un'altra variabile, o gruppo di variabili, X, allora Z causa X secondo Granger:

$$Z \xrightarrow{\mathcal{G}} X.$$

Osservazioni

› Poiché si invoca il principio che un effetto non può precedere una causa, a livello, temporale, si può parlare di "causalità".

› Lo studio della causalità secondo Granger permette di utilizzare processi VAR con più variabili, invece di sovraparametrizzare aumentando troppo l'ordine del processo.

Prop. (Causalità secondo Granger)
Sia $\boldsymbol{y}_t = (\boldsymbol{y}_{1,t} \;\; \boldsymbol{y}_{2,t})^\top$ *il vettore partizionato, con* $\mathcal{F}_t = \sigma(\boldsymbol{y}_t)$, $\mathcal{F}_{1,t} = \sigma(\boldsymbol{y}_{1,t})$, $\mathcal{F}_{2,t} = \sigma(\boldsymbol{y}_{2,t})$. *Allora, si ha causalità secondo Granger,* $\boldsymbol{y}_{1,t} \xrightarrow{\mathcal{G}} \boldsymbol{y}_{2,t}$, *se la stima includendo l'informazione su* $\boldsymbol{y}_{1,t}$ *migliora il MSE su* $\hat{\boldsymbol{y}}_{2,t+h|t}$. *In termini matriciali,*

$$MSE(\hat{\boldsymbol{y}}_{2,t+h|t}|\mathcal{F}_t) < MSE(\hat{\boldsymbol{y}}_{2,t+h|t}|\mathcal{F}_{2,t}).$$

Osservazione

Se vale che $\boldsymbol{y}_{1,t} \xrightarrow{\mathcal{G}} \boldsymbol{y}_{2,t}$ e $\boldsymbol{y}_{2,t} \xrightarrow{\mathcal{G}} \boldsymbol{y}_{1,t}$, si dice che il processo presenta *feedback*.

A volte è utile considerare una causalità data dal misurare variabili contemporanee, ma in istanti di tempo diversi. Nelle analisi reali, alcune variabili che sarebbero misurate contemporaneamente vengono ritardate nella loro pubblicazione (e.g. produzione industriale del trimestre si conosce prima del PIL).

Def. (Causalità istantanea)
Se si estende il set informativo con elementi contemporanei alla previsione stessa e si considera un orizzonte di un solo periodo, si ha *causalità istantanea* da $\boldsymbol{y}_{1,t}$ verso $\boldsymbol{y}_{2,t}$ se

$$\mathrm{MSE}\big(\hat{\boldsymbol{y}}_{2,t+1|t}|\mathcal{F}_t \cup \boldsymbol{y}_{1,t+1}\big) < \mathrm{MSE}\big(\hat{\boldsymbol{y}}_{2,t+1|t}|\mathcal{F}_t\big)$$

In particolare, partizionando Ψ_j nello stesso modo di $\boldsymbol{y}_t$,

$$\Psi_j = \begin{pmatrix} \Psi_{j,1,1} & \Psi_{j,1,2} \\ \Psi_{j,2,1} & \Psi_{j,2,2} \end{pmatrix},$$

si può scrivere la funzione di previsione come

$$\hat{\boldsymbol{y}}_{1,t+h|t} = \underbrace{\sum_{j=h}^{\infty} \Psi_{j,1,1}\varepsilon_{1,t+h-j}}_{\text{componente } 1^{\text{a}} \text{ var.}} + \underbrace{\sum_{j=h}^{\infty} \Psi_{j,1,2}\varepsilon_{2,t+h-j}}_{\text{componente } 2^{\text{a}} \text{ var.}}.$$

Usando invece il solo set informativo della prima variabile $\mathcal{F}_{1,t} \subset \mathcal{F}_t$, si ha

$$\hat{\boldsymbol{y}}_{1,t+h|t} = \sum_{j=h}^{\infty} F_{j,1,1}\nu_{1,t+h-j}.$$

Dunque, i predittori sono equivalenti se

1. $F_{j,1,1} = \Psi_{j,1,1}$,
2. $\nu_{1,t} = \varepsilon_{1,t}$,
3. $\Psi_{j,1,2} = \mathbf{0}$.

Si usa allora l'ultima condizione per verificare l'assenza di causalità.

Prop. (Assenza di causalità (VMA))
Dato un generico VAR(p), le ipotesi di assenza di causalità si possono verificare come un test di ipotesi sulle matrici Ψ_j della forma VMA(∞):

(*i*) $\boldsymbol{y}_{1,t} \overset{\mathcal{G}}{\not\to} \boldsymbol{y}_{2,t}$ *se* $\Psi_{j,2,1} = 0 \quad j = 1, 2, \ldots$

(*ii*) $\boldsymbol{y}_{2,t} \overset{\mathcal{G}}{\not\to} \boldsymbol{y}_{1,t}$ *se* $\Psi_{j,1,2} = 0 \quad j = 1, 2, \ldots$

Osservazione

Siccome l'ipotesi coinvolgerebbe un numero infinito di parametri posti a 0, per testare l'assenza di causalità si può utilizzare la rappresentazione VAR(p).

Osservando che $\Psi_j = \Phi^j$, nel caso di una matrice 2×2 si può effettuare il seguente test:

Prop. (Assenza di causalità (VAR))
Dato un generico VAR(p), le ipotesi di assenza di causalità si possono verificare con i test di ipotesi:

(*i*) $\boldsymbol{y}_{1,t} \overset{\mathcal{G}}{\not\to} \boldsymbol{y}_{2,t}$ *se* $\Phi_{j,2,1} = 0 \quad j = 1, 2, \ldots, p$

(*ii*) $\boldsymbol{y}_{2,t} \overset{\mathcal{G}}{\not\to} \boldsymbol{y}_{1,t}$ *se* $\Phi_{j,1,2} = 0 \quad j = 1, 2, \ldots, p$

Osservazioni

› Per fare questo, si può usare un qualunque test sui parametri, ad esempio quello di Wald a partire dalla stima GLS

$$\lambda_w = (C\hat{\beta} - c)^\top \left[C\big((ZZ^\top)^{-1} \otimes \hat{\Sigma}\big) C^\top \right]^{-1} (C\hat{\beta} - c) \xrightarrow{d} \chi^2_N.$$

› Tutto questo vale solo nel caso 2×2, perché Φ^j è triangolare se e solo se Φ è triangolare. Nel caso di più blocchi vale nel caso $h = 1$, mentre la generalizzazione al caso con più di due blocchi deve porre a zero più componenti.

Lezione 13: Impulse Response Function

2020-11-07

La ***forecast impulse response*** o *analisi dei moltiplicatori* studia il comportamento delle variabili del modello a seguito di uno shock ad una di esse.

Caso VAR(1)

Supposto un modello VAR(1) per $\boldsymbol{y}_t = (y_{1,t} \;\; y_{2,t} \;\; y_{3,t})^\top$, si vuole studiare l'evoluzione del sistema per $t = 1, 2, \dots$ in seguito ad uno shock unitario $\varepsilon_{1,0} = 1$ e assumendo che non ci siano altri shock:

$$\begin{cases} \varepsilon_{1,0} = 1 \\ \varepsilon_{2,0} = \varepsilon_{3,0} = 0 \\ \varepsilon_t = \mathbf{0}_k & \text{se } t > 0 \end{cases}$$

Allora, il sistema dopo j tempi diventa

$$\begin{pmatrix} y_{1,j} \\ y_{2,j} \\ y_{3,j} \end{pmatrix} = \Phi^j \begin{pmatrix} y_{1,0} \\ y_{2,0} \\ y_{3,0} \end{pmatrix},$$

per cui se lo shock colpisce la variabile l-esima, la risposta impulsiva del sistema dopo j tempi è contenuta nella l-esima colonna di Φ^j.

Inoltre, siccome $\Psi_j = \Phi^j$, si ha la seguente osservazione che vale anche nel caso generale.

Prop. (IRF per un VAR(p))
Dato un generico processo VAR(p) con rappresentazione VMA(∞) data da

$$\boldsymbol{y}_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j},$$

il coefficiente $\psi_{j;i,l}$ è la risposta dopo j periodi per la variabile i a seguito di uno shock della variabile l.

Osservazione

I coefficienti $\psi_{j;i,l}$ sono detti anche *moltiplicatori dinamici*, perché sono il fattore moltiplicativo dopo j periodi a seguito di uno shock unitario a $t = 0$. Sono la *reazione attesa* della variabile a seguito dello shock ed hanno una connessione con la causalità di Granger.

Prop. (IRF e causalità)

Dato un generico processo VAR(p) la cui rappresentazione VMA(∞) è

$$\boldsymbol{y}_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j},$$

se $\psi_{j;i,l} = 0$ per $j = 0, 1, 2, \ldots$, allora la variabile l non causa i nel senso di Granger.

Infine, si osserva che i moltiplicatori sono la variazione di una funzione lineare (la VMA) rispetto a un cambiamento unitario di ε_{t-j}. Dunque, si possono interpretare in termini di derivate.

Def. (Moltiplicatori dinamici)

Dato un generico processo VAR(p) la cui rappresentazione VMA(∞) è

$$\boldsymbol{y}_t = \sum_{j=0}^{\infty} \Psi_j \varepsilon_{t-j},$$

allora i *moltiplicatori dinamici* sono

$$\frac{\partial \boldsymbol{y}_t}{\partial \varepsilon_{t-j}^{\top}} = \Psi_j \implies \frac{\partial y_{i,t}}{\partial \varepsilon_{l,t-j}} = \psi_{j;i,l}.$$

Def. (Impulse Response Function)

La *impulse response function* (IRF) è la sequenza $\psi_{j;i,l}$ osservata al variare di $j = 0, 1, 2, \ldots$

Def. (Impulse Response Function cumulata)

La *impulse response function cumulata* (oppure *moltiplicatori intermedi*) è la sequenza osservata al variare di $j = 0, 1, 2, \ldots$ di

$$s_{m;i,l} = \sum_{j=0}^{m} \psi_{j;i,l}, \qquad S_m = \sum_{j=0}^{m} \Psi_j.$$

Def. (Moltiplicatori totali)

Si definiscono i *moltiplicatori totali* come

$$\lim_{m\to\infty} S_m = \sum_{j=0}^{\infty} \Psi_j = (I_k - \Phi_1 - \ldots - \Phi_p)^{-1} = \big(I_k - \Phi(1)\big)^{-1}.$$

Osservazione

Il risultato sui moltiplicatori totali dice che, se il VAR è stabile, il valore limite esiste finito e si chiama *effetto di lungo periodo* dovuto allo shock.

Per isolare l'effetto che ha una variabile su un'altra, si possono usare le *funzioni di risposta impulsiva ortogonalizzate* usando la solita decomposizione di Cholesky (17).

Def. (IRF ortogonalizzate)
Le IRF ortogonalizzate corrispondono agli elementi della matrice Θ_j per $j = 0, 1, 2, \ldots$

In generale, si ricavano degli intervalli di confidenza per le IRF sulla base di metodi di simulazione (bootstrap).

1. Ricampionare le innovazioni assumendo una distribuzione (bootstrap parametrico).
2. Ricampionamento tenendo conto della dipendenza (*block bootstrap*, non parametrico).

Lezione 14: Processi integrati e radici unitarie

2020-11-09

Questo argomento estende l'analisi a processi non stazionari, tramite l'utilizzo di serie storiche integrate. Utilizzare le differenze prime di solito comprende la distruzione di variabilità come conseguenza della stazionarizzazione.

$$\textsc{Trend} \longrightarrow \begin{cases} \textsc{deterministico} & \sum_{i=1}^{t} \mu_i : \quad \mathbb{E}\left[y_t\right] = t\mu \\ \textsc{stocastico} & \sum_{i=1}^{t} \varepsilon_i : \quad \mathbb{V}\left[y_t\right] = t\sigma^2 \end{cases}$$

In particolare, un processo non stazionario con trend stocastico è tale che

$$\lim_{t\to\infty} \mathbb{V}\left[y_t\right] = \lim_{t\to\infty} t\sigma^2 = \infty,$$

per cui il processo diventa imprevedibile per t grande. Allora, il processo differenziato è

$$y_t - y_{t-1} = \varepsilon_t \implies \mathbb{V}\left[(1-L)y_t\right] = \sigma^2,$$

per cui differenziando si perde quasi tutta la variabilità del fenomeno originale:

$$\frac{\mathbb{V}\left[y_t\right]}{\mathbb{V}\left[(1-L)y_t\right]} = t \tag{18}$$

Infine, il processo con trend stocastico ha un effetto permanente su y_t, perché si cumulano tutti gli shock con lo stesso peso.

Aggiungendo un'intercetta, si osserva

$$y_t = \mu + y_{t-1} + \varepsilon_t \implies y_t = y_0 + \underbrace{t\cdot\mu}_{\textsc{drift}} + \underbrace{\sum_{i=1}^{t}\varepsilon_i}_{\textsc{stocastico}} \quad .$$

Def. (Processo integrato di ordine zero)
Si dice che y_t è *integrato di ordine zero*, $y_t \sim I(0)$, se è un processo con rappresentazione

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \Psi(L)\varepsilon_t,$$

tale che $\Psi(1) = \sum_{j=0}^{\infty} \psi_j = c < \infty$, $c \neq 0$.

Def. (Processo integrato di ordine d)
Un processo y_t viene detto *integrato di ordine d*, $y_t \sim I(d)$, se

$$(1-L)^d y_t = \Delta^d y_t = \eta_t,$$

con $\eta_t \sim I(0)$ processo integrato di ordine 0.

Osservazione

- $\text{WN}(0, \sigma^2) \sim I(0)$.
- $\text{AR}(1) \sim I(1)$ se non ha radici unitarie.
- $\text{MA}(1)$ definito come $x_t = \varepsilon_t - \varepsilon_{t-1}$ è tale che

$$\sum_{j=0}^{\infty} \psi_j = 1 - 1 = 0,$$

  dunque non è $I(0)$. Risulta I(0) se il coefficiente a media mobile è $|\psi| < 1$.
- Un processo $\text{ARIMA}(p, d, q)$ privo di radici unitarie in AR e MA è integrato di ordine d, poiché

$$\Delta^d y_t = x_t \sim I(0)$$

  segue un processo $\text{ARMA}(p, q)$ stazionario e invertibile. Il processo ha d radici unitarie.

Proprietà importanti dei processi integrati

Valgono le seguenti relazioni tra processi integrati e le loro combinazioni lineari:

- Se $x_t \sim I(0)$, allora $a + bx_t \sim I(0)$
- Se $x_t \sim I(1)$, allora $a + bx_t \sim I(1)$
- Se $x_t \sim I(0)$ e $y_t \sim I(0)$, allora $ax_t + by_t \sim I(0)$
- Se $x_t \sim I(0)$ e $y_t \sim I(1)$, allora $ax_t + by_t \sim I(1)$
- Se $x_t \sim I(1)$ e $y_t \sim I(1)$, allora di norma $ax_t + by_t \sim I(1)$, ma non è sempre detto. Potrebbe essere anche $I(0)$, e si dice che i processi sono *cointegrati*.

Def. (Processo stazionario in differenza)

Un processo $y_t \sim I(d)$ è chiamato *stazionario in differenza*, perché

$$\Delta^d y_t \sim I(0)$$

Def. (Processo attorno a un trend deterministico)

Un processo *stazionario attorno ad un trend* è definito come

$$y_t = \beta_0 + \beta_1 t + x_t, \quad x_t \sim I(0).$$

Osservazione

- Con una regressione lineare si può rimuovere il trend deterministico e ottenere i residui $x_t \sim I(0)$.

› Invece, se si differenzia il processo si ottiene

$$\Delta y_t = \beta_1 + \Delta x_t,$$

che contiene una componente MA(1) stazionaria ma non invertibile (Δx_t), quindi $\Delta y_t \not\sim I(0)$.

È dunque di particolare rilevanza essere in grado di discriminare tra processi a trend deterministico e processi con trend stocastico. Questo si traduce nella verifica della presenza di **radici unitarie** nel processo.

Esempio (ARIMA$(p, 1, q)$)
Considerato un processo ARIMA$(p, 1, q)$ con intercetta

$$y_t = \beta_0 + y_{t-1} + x_t, \quad x_t \sim \text{ARMA}(p, q),$$

x_t stazionario e invertibile. Allora, $\Delta y_t = \beta_0 + x_t$ è un processo I(0) e, sostituendo, vale che

$$y_t = \beta_0 t + y_0 + \sum_{j=0}^{t-1} x_j.$$

Dunque il processo ARIMA$(p, 1, q)$ con intercetta contiene sia un trend deterministico sia un trend stocastico.

Prop. (ARIMA)
Un modello ARIMA(p, d, q) con costante, oltre ai trend stocastici determinati dalle radici unitarie, ha un trend deterministico polinomiale di grado d.

14.1 Test per radici unitarie

Si vuole studiare la presenza di test stocastici e/o deterministici. Si consideri un processo con errori autocorrelati secondo un AR(1):

$$y_t = \gamma_0 + \gamma_1 t + u_t$$

$$u_t = \rho u_{t-1} + \varepsilon_t.$$

Tramite sostituzioni, si ottiene

$$\begin{aligned}
y_t &= \gamma_0 + \gamma_1 t + u_t \\
&= \gamma_0 + \gamma_1 t + \rho u_{t-1} + \varepsilon_t \\
&= \gamma_0 + \gamma_1 t + \rho(y_{t-1} - \gamma_0 - \gamma_1(t-1)) + \varepsilon_t \\
&= \underbrace{\gamma_0(1-\rho) + \gamma_1\rho}_{\beta_0} + \underbrace{\gamma_1(1-\rho)}_{\beta_1} t + \rho y_{t-1} + \varepsilon_t \\
&= \beta_0 + \beta_1 t + \rho y_{t-1} + \varepsilon_t
\end{aligned}$$

Osservazione

Se $\rho = 1$, il processo è un random walk con drift, mentre se $|\rho| < 1$ il processo diventa stazionario attorno a un trend deterministico.

Applicando la differenza prima, si ottiene

$$\Delta y_t = \beta_0 + \beta_1 + \underbrace{(\rho - 1)}_{\delta} y_{t-1} + \varepsilon_t, \tag{19}$$

dunque se $\delta = \rho - 1 = 0$ si ha un processo con trend stocastico, mentre se $\delta < 0$ si ha invece un processo stazionario attorno a un trend. Quindi, il **test di radice unitaria** (*unit root test*) diventa un test sul modello lineare (19)

$$\begin{cases} H_0 : \delta = 0 & \text{TREND STOCASTICO} \\ H_1 : \delta < 0 & \text{TREND DETERMINISTICO} \end{cases}$$

Si potrebbe pensare di utilizzare la regressione lineare per effettuare questo test, tuttavia poiché il processo sotto H_0 è non stazionario, la statistica test è

1. Con distribuzione *non standard* (non è una t di Student).
2. Cambia in base al tipo di trend che si vuole testare (lineare, quadratico, ...).

Per questo motivo, i p-value si ottengono di solito via simulazione.

14.2 Test di Dickey-Fuller

Si basa sulla regressione senza trend deterministico

$$\Delta y_t = \delta y_{t-1} + \varepsilon_t,$$

che verifica l'ipotesi nulla con la statistica

$$\frac{\hat{\delta}}{\sqrt{\mathbb{V}\left[\hat{\delta}\right]}} \xrightarrow{T\to\infty} \mathrm{DF}(\cdot),$$

che converge alla *distribuzione di Dickey-Fuller*, che si ricava come particolare funzione dei moti Browniani.

Aggiungendo una variazione del trend, ad esempio

$$\Delta y_t = \beta_0 + \delta y_{t-1} + \varepsilon_t,$$

la distribuzione limite è **diversa**, per cui il test va ripetuto per ciascuna componente di trend.

Tuttavia, per fortuna la distribuzione non dipende dal valore di β_0, ma solo dalla sua presenza o meno. Sotto l'ipotesi nulla, il processo prevede la presenza di un trend stocastico e deterministico

lineare. Dunque, è importante calibrare bene le ipotesi del modello.

Il limite dei test di Dickey-Fuller è il fatto di poter verificare solo modelli molto semplici, per cui sono stati introdotti i test *Augmented Dickey-Fuller* (ADF), che si basano sul modello

$$\Delta y_t = \underbrace{\alpha + \gamma t + \delta y_{t-1} + \varepsilon_t}_{\text{Dickey-Fuller}} + \underbrace{\sum_{j=1}^{p} \beta_j \Delta y_{t-j}}_{\text{Augmented}} . \tag{20}$$

Sotto ipotesi nulla, la componente $\sum_{j=1}^{p} \beta_j \Delta y_{t-j}$ che aumenta il test è stazionaria, in quanto $\Delta y_{t-j} \sim \text{I}(0)$. In questo modo, si includono dei termini che tengono conto delle autocorrelazione e sono stazionari sotto l'ipotesi nulla, per cui le distribuzioni dei test non sono molto diverse da quelle del test DF standard.

Il test dipende da p, per cui si possono usare i criteri di informazione per scegliere l'ordine migliore.

In alternativa ai criteri di informazione, si ha la **procedura di Ng e Perron**, che alterna l'identificazione di radici unitarie alla verifica della presenza di trend.

Procedura di Ng e Perron

1. Scegliere $p_{\max}$
2. Stimare i parametri nell'equazione (20) e valutare la significatività di β_p, in caso si rimuove e si riparte da 1.
3. Terminato il passo 2), con $p = p^*$ scelto si testa l'ipotesi di radice unitaria $H_0 : \delta = 0$.
   - Se si rifiuta l'ipotesi, si termina e il processo **non ha trend stocastico**.
   - Se si accetta l'ipotesi e si conclude che ci sia trend stocastico, è necessario verificare la correttezza del trend deterministico.
4. Il processo (20) sotto Hp. nulla è AR(p) sulle differenze Δy_t con trend deterministico e costante. Si stimano i parametri sotto il vincolo $\delta = 0$ e si testa l'**assenza di trend deterministico** $H_0 : \gamma = 0$ con un test standard gaussiano.
   - Se γ significativo, si testa al passo 3) con un test gaussiano.
   - Se γ non significativo, si prosegue.
5. Si stima l'equazione priva di trend testata allo step precedente

$$\Delta y_t = \alpha + \delta y_{t-1} + \sum_{j=1}^{p^*} \beta_j \Delta y_{t-j} + \varepsilon_t$$

   e si valuta l'ipotesi nulla su δ con test DF. Se si rifiuta, non c'è radice unitaria, altrimenti si continua.
6. Si ripete la procedura, studiando sotto $\delta = 0$ l'ipotesi $H_0 : \alpha = 0$ con un test gaussiano. Se anche questa procedura non è significativa, si specifica finalmente

$$\Delta y_t = \delta y_{t-1} + \sum_{j=1}^{p^*} \beta_j \Delta y_{t-j} + \varepsilon_t$$

   e si fa un ultimo test di radice unitaria.

Osservazioni

› Il test verifica l'ipotesi I(1) vs. I(0), ma non altri ordini di integrazione.

Lezione 15: Cointegrazione

2020-11-10

Def. (Processi cointegrati)
Se due processi sono $x_t \sim \mathrm{I}(1)$ e $y_t \sim \mathrm{I}(1)$ e tali che

$$z_t = ax_t + by_t \sim \mathrm{I}(0),$$

allora si dicono *cointegrati* e si indicano con $\mathrm{CI}(1,1)$.

Osservazione

- Si possono pensare come due random walk, il cui processo combinato è stazionario, e.g. White Noise.
- La cointegrazione distrugge la dinamica evolutiva dei processi, esattamente come nel considerare Δx_t e Δy_t, tuttavia senza perdere informazione sulla variabilità.
- Ci sono legami con l'*analisi fattoriale*, perché si considera z_t come somma di processi stazionari latenti in modo da spiegare l'intera variabilità, non solo quella residuale.

Esempio (Processi cointegrati)
Si consideri il processo

$$x_t = x_{t-1} + \alpha(x_{t-1} - \beta y_{t-1}) + \eta_t$$

$$y_t = y_{t-1} + \varepsilon_t$$

con $\alpha \in (-2, 0)$ analogo dei *loading factors* e $\eta_t \sim \mathrm{I}(0)$, $\varepsilon_t \sim \mathrm{I}(0)$, $\eta_t \perp\!\!\!\perp \varepsilon_t$. Si ha che x_t è non stazionario per $\beta \neq 0$, poiché si può scrivere come combinazione lineare di I(0) e I(1):

$$x_t = \underbrace{(1+\alpha)x_{t-1}}_{\text{stazion.}} - \underbrace{\alpha\beta y_{t-1}}_{\text{non stazion.}} + \nu_t. \tag{21}$$

È importante osservare che x_t di per sé sarebbe stazionario, ma diventa non stazionario perché condivide il trend stocastico di y_t.

Invece, se si considera la combinazione lineare $z_t = x_t - \beta y_t$, si osserva che

$$\begin{aligned} z_t &= x_t - \beta y_t \\ &= x_{t-1} + \alpha(x_{t-1} - \beta y_{t-1}) + \eta_t - \beta y_{t-1} - \beta\varepsilon_t \\ &= (1+\alpha)x_{t-1} - (1+\alpha)\beta y_{t-1} + \eta_t - \beta\varepsilon_t \\ &= (1+\alpha)\underbrace{(x_{t-1} - \beta y_{t-1})}_{z_{t-1}} + \underbrace{\eta_t - \beta\varepsilon_t}_{\nu_t \sim \mathrm{I}(0)} \\ &= (1+\alpha)z_{t-1} + \nu_t \\ &\sim \mathrm{AR}(1) \text{ stazionario se } \alpha \in (-2, 0). \end{aligned}$$

Con la notazione di prima,

$$\begin{cases} x_t \sim \mathrm{I}(1) \\ y_t \sim \mathrm{I}(1) \\ x_t - \beta y_t \sim \mathrm{I}(0) \end{cases} \implies z_t \sim \mathrm{CI}(1,1) \tag{22}$$

Osservazione

Se non ci fosse una relazione che lega il trend stocastico di y_t a quello di x_t ($\beta = 0$), si può dimostrare che non è possibile trovare una relazione di cointegrazione (22) tra i due processi.

In particolare, vale il seguente risultato fondamentale:

Conclusione fondamentale della cointegrazione
Per avere cointegrazione, x_t e y_t devono ***condividere lo stesso trend stocastico***.

Regressione spuria

Dati due processi con trend stocastico I(1) indipendenti ($\implies$ **non conintegrati**), ad esempio random walk, se si costruisce il modello di regressione

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t,$$

$$y_t \sim \mathrm{AR}(1) \text{ non staz.}$$

$$x_t \sim \mathrm{AR}(1) \text{ non staz.}$$

ci si aspetta che β_1 non sia significativamente diverso da zero. Invece, tipicamente la regressione lineare fallisce e si osserva una **relazione spuria** tra i processi:

1. β_0, β_1 significativi e molto grandi.
2. R^2 e statistica F elevati.

3. Statistica DW (autocorrelazione dei residui) quasi nulla, per cui residui vengono classificati come White Noise.

Se invece si stima il modello sulle differenze prime, che sono I(0),

$$\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + \varepsilon_t,$$

si ottiene $\gamma_1 \approx 0$ come ci si attenderebbe dal fatto che sono indipendenti.

Philips ha dimostrato quanto segue:

Prop. (Regressione spuria)
Dati x_t, y_t due processi I(1) indipendenti, la regressione $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ è tale che

- *$\hat{\beta}_0, \hat{\beta}_1$ non convergono in probabilità a costanti.*
- *$\hat{\beta}_1$ non converge a una gaussiana centrata in 0.*
- *La distribuzione di $\hat{\beta}_0$ diverge.*
- *Le statistiche test di significatività non hanno distribuzione limite e divergono.*
- *R^2 ha distribuzione limite non degenere e la statistica DW converge a zero.*

Conclusioni

Dunque, quando si vogliono mettere in relazione due serie storiche non stazionarie, si hanno due possibilità:

1. **Le serie non condividono trend stocastico**: non c'è cointegrazione e le stime dei parametri non funzionano. Bisogna per forza differenziare le serie.
2. **Le serie sono cointegrate**: allora la regressione ha senso e si ottengono risultati attendibili.

Questo vale sia in econometria quanto in qualunque altra modellazione dinamica, che siano dati biologici o aziendali.

15.1 Analisi di cointegrazione

Ci si occupa di combinazioni lineari stazionarie di variabili non stazionarie, accomunate da un trend stocastico.

Def. (Cointegrazione)
Un processo $\boldsymbol{y}_t$ con componenti $y_{i,t} \sim \mathrm{I}(1)$ si dice *cointegrato* se esiste $\beta \in \mathbb{R}^k$ tale che $\beta^\top \boldsymbol{y}_t \sim \mathrm{I}(0)$, e si indica con $\boldsymbol{y}_t \sim \mathrm{CI}(1,1)$.

Def. (Cointegrazione di ordine b)
Un processo $\boldsymbol{y}_t$ con componenti $y_{i,t} \sim \mathrm{I}(d)$ si dice *cointegrato di ordine b* se esiste $\beta \in \mathbb{R}^k$ tale che $\beta^\top \boldsymbol{y}_t \sim \mathrm{I}(d-b)$, e si indica con $\boldsymbol{y}_t \sim \mathrm{CI}(d,b)$.

Osservazioni

- β si chiama *vettore di cointegrazione* e $\beta^\top \boldsymbol{y}_t$ si chiama *relazione di cointegrazione*.
- Nel campo dell'econometria, la relazione di cointegrazione fornisce la relazione di equilibrio di lungo periodo tra le variabili.
- Se si definisce $\beta^* = c\beta$, per $c \neq 0$, allora β^* è ancora un vettore di cointegrazione. Per questo, di solito si normalizza per convenzione con

$$\beta = (1, \beta_2, \beta_3, \ldots, \beta_k)^\top \tag{23}$$

Con la normalizzazione (23), si può scrivere

$$\begin{aligned} \overbrace{\beta^\top \boldsymbol{y}_t}^{\varepsilon_t \sim \mathrm{I}(0)} &= y_{1,t} - \beta_2 y_{2,t} - \beta_3 y_{3,t} - \ldots - \beta_k y_{k,t} \\ \implies y_{1,t} &= \beta_2 y_{2,t} + \beta_3 y_{3,t} + \ldots + \beta_k y_{k,t} + \varepsilon_t. \end{aligned}$$

Se $\boldsymbol{y}_t \in \mathbb{R}^2$, si ha al massimo una sola relazione di cointegrazione possibile, mentre se $\boldsymbol{y}_t \in \mathbb{R}^k$ si hanno $0 < r < k$ relazioni di cointegrazione. Infatti, ci possono essere relazioni di cointegrazione tra blocchi di processi:

$$\beta_1 = (1, 0, 0, 0.6)$$

$$\beta_2 = (1, 0.5, 0.7, 0)$$

Questi r vettori di cointegrazione sono in generale linearmente indipendenti e si può scrivere la **matrice di cointegrazione**

$$B = (\beta_1 \quad \beta_2 \quad \ldots \quad \beta_r),$$

con r **rango di cointegrazione**. Anche in questo caso di solito si fa una normalizzazione per identificare i vettori di cointegrazione.

Esempio (Domanda di moneta)

Gli individui detengono una determinata quantità di moneta definita in termini reali, per cui la quantità di moneta in termini nominali deve essere proporzionale al livello dei prezzi.

Inoltre, la domanda di moneta dipende dal reddito reale e aumenti del reddito incrementano la domanda di moneta.

Infine, il tasso di interesse è un costo-opportunità di tenere la moneta in banca. Un aumento degli interessi dipende dalla domanda di moneta.

Riassumendo, se $m_t = \log \text{MONETA}_t$ è un processo che risulta non stazionario, $p_t = \log \text{PREZZO}_t$, $y_t = \log \text{REDDITO}$, $i_t = \log \text{INTERESSI}_t$, allora si può scrivere che

$$m_t = \alpha + \beta_1 p_t + \beta_2 y_t + \beta_3 i_t + \varepsilon_t.$$

Postulando da considerazioni di teoria economica che $\beta = 1$, $\beta_2 > 0$, $\beta_3 < 0$ e che $\varepsilon_t \sim \mathrm{I}(0)$

sia stazionario, ne deriva che (p_t, y_t, i_t) sono in relazione di cointegrazione.

Domanda

Statisticamente, da che fondamento deriva questa cointegrazione? Dal fatto che p_t, y_t, i_t siano tutti processi I(1) e nessuna di esse sia stazionaria. Se questa proprietà viene a mancare, allora anche la cointegrazione viene meno.

Esempio (Reddito permanente)

I consumi c_t possono essere permanenti, c_t^P, e transitori, c_t^T. La componente permanente è proporzionale al reddito permanente, ovvero $c_t^P = \beta y_t^P$, e dunque

$$c_t = c_t^P + c_t^T \implies \underbrace{c_t - \beta y_t^P}_{\beta = (1, -\beta)} = c_t^T .$$

Per costruzione c_t^T deve essere stazionaria.

Esempio (Prezzi spot e prezzi futuri)

L'efficienza dei mercati richiede che i prezzi futuri f_t siano l'aspettativa ad oggi del prezzo futuro. Postulando $f_t = \mathbb{E}\left[s_{t+1}\right]$, dove s_t è il prezzo spot al tempo t, se la relazione è soddisfatta si ha che il prezzo spot futuro $\mathbb{E}\left[s_{t+1}\right]$ è tale che

$$s_{t+1} - \mathbb{E}\left[s_{t+1}\right] = \varepsilon_{t+1} \sim \mathrm{I}(0).$$

Dunque, l'ipotesi di efficienza dei mercati porta alla *unbiased forward market hypothesis*, cioè la cointegrazione tra f_t e $\mathbb{E}\left[s_{t+1}\right]$. Di nuovo, è necessario che tutte le variabili siano I(1), e tipicamente i prezzi sono almeno cointegrate di ordine 1.

Esempio (Parità del potere di acquisto)

La teoria della parità del potere di acquisto postula che il livello dei prezzi di due economie siano proporzionali al tasso di cambio.

Dunque, $e_t^{\mathrm{E/U}} = p_t^{\mathrm{E}} \ldots$

Lezione 16: Vector Error Correcting Models

2020-11-16

I modelli ARIMA consistono in modelli ARMA applicati alle differenze del processo, ma come abbiamo visto da (18), la maggior parte della variabilità del processo sta nella non stazionarietà.

Si può allora evitare di applicare la differenziazione, in modo da **non perdere** la variabilità del processo, nel caso in cui si abbia a che fare con processi cointegrati.

I processi cointegrati I(1) condividono dei trend comuni, che influenzano la dinamica delle variabili osservate.

16.1 Error Correcting Model

Per analizzare la cointegrazione, si usano **modelli a correzione di errore** (ECM): assumendo x_t, y_t che condividono un trend comune, e quindi una relazione di lungo periodo, supponiamo che valga la relazione

$$y_t = \beta x_t + \varepsilon_t$$

Supponiamo anche che $H_0 : \beta = 1$, ovvero che sotto ipotesi non ci siano distorsioni di scala tra le due variabili. Per studiare l'effetto di uno shock nel modello, bisogna assumere che ci sia una dinamica autoregressiva:

$$y_t = \mu + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t \tag{24}$$

da cui attraverso qualche passaggio, ponendo $\tilde{\alpha} = \alpha - 1$, $\tilde{\beta} = \frac{\beta_0+\beta_1}{1-\alpha}$, si ricava **TODO** sistemare

$$y_t = \mu + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

$$y_t - y_{t-1} = \mu + (\alpha_1 - 1) y_{t-1} + \beta_0 x_t \pm \beta_0 x_{t-1} + \beta_1 x_{t-1} + \varepsilon_t$$

$$\Delta y_t = \mu + \beta_0 \Delta x_t + (\alpha_1 - 1) y_{t-1} + (\beta_0 + \beta_1) x_{t-1} + \varepsilon_t$$

$$\Delta y_t = \mu + \beta_0 \Delta x_t + \tilde{\alpha}\big(y_{t-1} - \tilde{\beta} x_{t-1}\big) + \varepsilon_t$$

Dunque, la dinamica evolutiva di y_t è della forma

$$\boxed{\Delta y_t = f(\Delta x_t; y - \beta x)}$$

In particolare, il sistema può trovarsi in due stati:

1. **Equilibrio**: $y_t - \tilde{\beta} x_t = 0$ e di conseguenza la dinamica segue quella basilare (24),

$$\Delta y_t = \beta_0 \Delta x_t + \varepsilon_t.$$

2. **Disequilibrio**: $y_t - \tilde{\beta} x_t \neq 0$, e la deviazione dall'equilibrio si ripercuote sui tassi di crescita di y_t a causa di $\tilde{\alpha}(y_{t-1} - \tilde{\beta} x_{t-1})$.

Prop.
Se y_t e x_t sono I(1), allora per def. Δy_t e Δx_t sono I(0), dunque $y_t - \tilde{\beta} x_t$ è la relazione di cointegrazione tra y_t e x_t.

Dim.
Poiché sappiamo che $\Delta y_t, \Delta x_t, \varepsilon_t \sim \text{I}(0)$, allora a sinistra e a destra dell'uguaglianza si hanno processi I(0) e per le relazioni notevoli vale

$$\underbrace{\Delta y_t}_{\text{I}(0)} = \beta_0 \underbrace{\Delta x_t}_{\text{I}(0)} + \tilde{\alpha} \underbrace{\left(y_{t-1} - \tilde{\beta} x_{t-1}\right)}_{\text{I}(0)} + \underbrace{\varepsilon_t}_{\text{I}(0)},$$

ma poiché $y_{t-1}, x_{t-1} \sim y_t, x_t \sim \text{I}(1)$, allora $(1, \tilde{\beta})$ è il vettore di cointegrazione.

□

Riassumendo

L'ECM collega il lungo periodo (cointegrazione) con il breve periodo, tramite un **aggiustamento dinamico** proporzionale alla deviazione dall'equilibrio di lungo periodo.

$$\tilde{\alpha} \rightsquigarrow \text{velocità di ritorno al lungo periodo.}$$

La rappresentazione ECM per due variabili (y_t, c_t) con intercetta si può scrivere equivalentemente come

$$\Delta y_t = \gamma_y \Delta c_t + \alpha_y (c_{t-1} - y_{t-1} - \mu) + \varepsilon_{y,t}$$

$$\Delta c_t = \gamma_c \Delta y_t + \alpha_c (c_{t-1} - y_{t-1} - \mu) + \varepsilon_{c,t}$$

con la stessa relazione di integrazione, in quanto il **trend è lo stesso**:

$$u_{t-1} = c_{t-1} - y_{t-1} - \mu,$$

da cui, se si ipotizza che $\alpha_c = 0, \alpha_y = 0.5$, si può allora scrivere

$$\Delta y_t = \gamma_y \Delta c_t + 0.5 u_{t-1} + \varepsilon_{y,t}$$

$$\Delta c_t = \gamma_c \Delta y_t + \varepsilon_{c,t}$$

In tale situazione, si ha che

- c_t è RW con drift, guida la relazione e non è influenzata dal disequilibrio.
- $y_t \sim \text{I}(1)$ ma dipende da c.
- In condizione di equilibrio, $u_{t-1} = 0$ e i valori attesi condizionato sono perfettamente simmetrici,

  $$\mathbb{E}\left[\Delta y_t | c_{t-1}, y_{t-1}\right] = \gamma_y \Delta c_t$$

  $$\mathbb{E}\left[\Delta c_t | c_{t-1}, y_{t-1}\right] = \gamma_c \Delta y_t$$

  mentre in **disequilibrio** si ha

  $$\mathbb{E}\left[\Delta y_t | c_{t-1}, y_{t-1}\right] = \gamma_y \Delta c_t + 0.5(c_{t-1} - y_{t-1} - \mu).$$

Questo si interpreta dicendo che, se ora i consumi c_t sono aumentati (diminuiti), allora negli anni successivi bisogna accelerare (diminuire) il reddito y_t per tornare all'equilibrio.

Come si stimano $\tilde{\alpha}, \tilde{\beta}$ e i parametri della dinamica di breve periodo?

Se le $\boldsymbol{y}_t \in \mathbb{R}^k$ con tutte le componenti I(1) e rango r, allora ci sono r vettori di cointegrazione, equivalentemente esiste $B \in \mathbb{R}^{k \times r}$ tale che

$$B^\top \boldsymbol{y}_t = u_t, \qquad u_{i,t} \sim \mathrm{I}(0) \quad \forall i = 1, \dots, k.$$

Osservazioni

1. La cointegrazione non è nota a priori e va verificata ($r = 0$?).
2. Il rango r va stimato in qualche modo.
3. B va stimata in qualche altro modo.
4. Le variabili devono essere tutte per forza I(1), altrimenti non funziona.

Ci sono due casi rilevanti per $\boldsymbol{y} \in \mathbb{R}^k$

1. $r = 0$ oppure $r = 1$, si usa l'**approccio di Engle & Granger**.
2. $r > 1$, si usa l'**approccio di Johansen** (VECM).

Lezione 17: Stima del VECM

2020-11-17

17.1 Approccio di Engle & Granger

Si suppone che $r = 1$, $y_{i,t} \sim \mathrm{I}(1)$ per ogni $i = 1, \dots, k$ e che ci sia una relazione di cointegrazione di cui sappiamo qual è la y_i causata.

Si stima allora $\beta^\top \boldsymbol{y} = \varepsilon_t$ con vincolo $\beta_1 = 1$, ovvero il modello lineare semplice

$$y_{1,t} = \beta_2 y_{2,t} + \beta_3 y_{3,t} + \ldots + \beta_k y_{k,t} + \varepsilon_t \tag{25}$$

Due approcci diversi, se β è noto o meno, la cui differenza sta nella teoria asintotica (che non vedremo mai).

Se $\boldsymbol{\beta}$ è noto (fisica, economia, a volte in biologia, . . .)

Si possono allora calcolare i residui di cointegrazione $u_t = y - X\beta \sim \mathrm{I}(0)$ e si può fare un test di radice unitaria (ADF) per verificare che ci sia cointegrazione

$$\text{test ADF} \begin{cases} y \rightsquigarrow I(1) \\ x \rightsquigarrow I(1) \\ u \rightsquigarrow I(0) \end{cases} \implies \text{cointegrazione}$$

Se $\boldsymbol{\beta}$ non è noto

Si stima

$$\hat{\varepsilon}_t = y_{1,t} - \hat{\beta}_2 y_{2,t} - \hat{\beta}_3 y_{3,t} - \ldots - \hat{\beta}_k y_{k,t}.$$

Si testano tutti gli $y_{i,t}$ per I(1) e $\hat{\varepsilon}$ per I(0). Tuttavia, i valori critici da utilizzare per β non noto non sono quelli standard, ma sono stati derivati da Engle & Granger.

Valgono le stesse raccomandazioni relative al test ADF, per quanto riguarda l'inclusione di diversi tipi di trend, ecc. . .

Stimando la relazione con `lm(`$y_1 \sim y_{2:k}$`)`, si hanno le seguenti proprietà:

Prop. (Proprietà di $\hat{\boldsymbol{\beta}}$ nell'approccio di Engle & Granger)
Si può usare lo stimatore OLS e, ***sotto esistenza di cointegrazione****, si ha che*

- *$\hat{\beta}$ è uno stimatore superconsistente*
- *$\hat{\beta}$ converge a β con velocità T invece di $\sqrt{T}$.*

Osservazioni

- Purtroppo, $T(\hat{\beta} - \beta)$ converge a una distribuzione non gaussiana.
- Per piccoli campioni, lo stimatore col metodo OLS non è quello a massima efficienza.

Dynamic OLS

La regressione di cointegrazione (25) ignora la dinamica temporale di $y_2, y_3, \ldots, y_k$, per cui $\hat{\varepsilon}_t$ ha ACF persistente.

Si usano allora i "*Dynamic OLS*" (DOLS) per pulire i residui il più possibile dall'autocorrelazione:

$$y_{1,t} = \beta_2 y_{2,t} + \beta_3 y_{3,t} + \ldots + \beta_k y_{k,t} + \sum_{j=-p}^{p} \sum_{i=2}^{k} \underbrace{\delta_{i,j} \Delta y_{i,t-j}}_{\text{I(0)} \implies \text{ok}} + \varepsilon_t.$$

I lag passati e futuri sono inclusi solo per stimare meglio i β_j. Sotto ipotesi di cointegrazione, β_{DOLS} è ancora asintoticamente normale e si possono valutare la significatività sui coefficienti di cointegrazione.

Una volta stimata questa relazione, si possono calcolare i residui

$$\hat{\varepsilon}_t = y_t - \hat{\beta} x_t$$

e procedere poi alla stima della rappresentazione ECM con gli OLS (poiché tutti I(0))

$$\begin{aligned} \Delta y_t &= \beta_0 \Delta x_t + \tilde{\alpha}(y_{t-1} - \tilde{\beta} x_{t-1}) + \eta_t \\ &= \underbrace{\beta_0 \Delta x_t}_{\text{breve periodo}} + \underbrace{\tilde{\alpha} \hat{\varepsilon}_{t-1}}_{\text{lungo periodo}} + \eta_t \end{aligned}$$

Osservazione Come si può vedere, i dati vengono utilizzati due volte per la stima

1. Stima di $\hat{\varepsilon} \implies \tilde{\beta}$
2. Ottenuti $\tilde{\beta}$, stima dell'ECM $\implies \beta_0,\ \tilde{\alpha}$

Questo si può fare perché la stima di $\hat{\varepsilon}$ è superconsistente, dunque permette di utilizzare due volte l'informazione senza invalidare le procedure.

17.2 Approccio di Johansen

Supponiamo ora che $r > 1$ ignoto e sia sempre $\boldsymbol{y}_t \in \mathbb{R}^k$, in questo caso non si può usare la procedura discussa prima ma

1. Si usano i VAR per stimare r e le relazioni.
2. Il modello prende il nome di VECM.

Consideriamo un modello VAR(p) senza costante

$$\boldsymbol{y}_t = \Phi_0 + \Phi_1 \boldsymbol{y}_{t-1} + \Phi_2 \boldsymbol{y}_{t-2} + \ldots + \Phi_p \boldsymbol{y}_{t-p} + \varepsilon_t,$$

con

$$|\Phi(z)| \neq 0 \qquad \text{se } |z| \leq 1.$$

Se $\boldsymbol{y}_t$ è cointegrato, allora tutte le variabili sono I(1) e il modello VAR(p) non è adeguato.

Si costruisce una relazione alternativa aggiungendo e sottraendo delle quantità (analogo di quanto fatto nel caso univariato)

$$\begin{aligned}
\Delta \boldsymbol{y}_t = \quad & -\boldsymbol{y}_{t-1} + \Phi_1 \boldsymbol{y}_{t-1} + \Phi_2 \boldsymbol{y}_{t-2} + \ldots + \Phi_p \boldsymbol{y}_{t-p} + \varepsilon_t \\
\Delta \boldsymbol{y}_t = \quad & -\boldsymbol{y}_{t-1} + \Phi_1 \boldsymbol{y}_{t-1} + \Phi_2 \boldsymbol{y}_{t-2} + \ldots + \Phi_p \boldsymbol{y}_{t-p} \\
& + (\Phi_2 + \Phi_3 + \ldots + \Phi_p)\boldsymbol{y}_{t-1} - (\Phi_2 + \Phi_3 + \ldots + \Phi_p)\boldsymbol{y}_{t-1} \\
& + (\Phi_3 + \Phi_4 + \ldots + \Phi_p)\boldsymbol{y}_{t-2} - (\Phi_3 + \Phi_4 + \ldots + \Phi_p)\boldsymbol{y}_{t-2} \\
& \vdots \\
& + \Phi_p \boldsymbol{y}_{t-p+1} - \Phi_p \boldsymbol{y}_{t-p+1} + \varepsilon_t.
\end{aligned}$$

Riordinando i termini, si ha

$$\Delta \boldsymbol{y}_t = (-I_k + \Phi_1 + \Phi_2 + \ldots + \Phi_p)\boldsymbol{y}_{t-1} + \boldsymbol{TODO}$$

Prop. (Rappresentazione VECM)
Il ***vector error correction model*** *(VECM) è dato da*

$$\Delta \boldsymbol{y}_t = \Pi \boldsymbol{y}_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta \boldsymbol{y}_{t-j} + \varepsilon_t,$$

dove

$$\Pi = (-I_k + \Phi_1 + \Phi_2 + \ldots + \Phi_p) = -\Phi(1)$$

$$\Gamma_j = -(\Phi_{j+1} + \Phi_{j+2} + \ldots + \Phi_p)$$

Osservazione

- La rappresentazione è ancora una volta di tipo $\Delta \boldsymbol{y}_t = f(\Delta \boldsymbol{x}, \mathrm{LR})$.
- Siccome $\Delta \boldsymbol{y}_t, \varepsilon_t \sim \mathrm{I}(0)$, le proprietà dipendono da $\Pi \boldsymbol{y}_{t-1}$, che è combinazione lineare dei processi. Se si stima una matrice Π tale che $\Pi \boldsymbol{y}_{t-1} \sim \mathrm{I}(0)$, allora questa è la matrice che dà la relazione di cointegrazione.
- Si può tornare indietro dal VECM al VAR tramite le relazioni inverse

$$\Phi_1 = \Gamma_1 + \Pi - I_k$$

$$\Phi_p = -\Gamma_{p-1}$$

- $\boldsymbol{y}_t \sim \mathrm{I}(1)$, allora per forza c'è radice unitaria e $\det \Phi(1) = \det \Pi = 0$, di conseguenza Π ha rango $\operatorname{rank} \Pi < k$. In generale, si può dimostrare che $\operatorname{rank} \Pi = r$ è esattamente il numero di

relazioni di cointegrazione.

Prop. (Rango di Π)
Se Π *è a rango pieno allora non c'è cointegrazione, visto che le serie di partenza sono* $I(1)$ *non cointegrate.*

In generale, i possibili sottocasi sono i seguenti:

i. Se $\operatorname{rank}\Pi = k = \dim \boldsymbol{y}_t$, allora vuol dire che sono tutti processi I(1).

ii. Se $\operatorname{rank}\Pi = 0$, non si ha cointegrazione e l'unica cosa che si può fare è il VAR sulle differenze
$$\Delta \boldsymbol{y}_t = \sum_{j=1}^{p-1} \Gamma_j \Delta \boldsymbol{y}_{t-j} + \varepsilon_t.$$

iii. Se $0 < \operatorname{rank}\Pi < k$, ci sono r vettori di cointegrazione, cioè r combinazioni lineari degli $\boldsymbol{y}_t$ che sono I(0).

› Se io ignoro tutto questo, il modello VAR su $\boldsymbol{y}$ presenta relazioni spurie e l'inferenza è falsata.

Nel caso in cui $\operatorname{rank}\Pi = r < k$, si può ottenere una scrittura esplicita della relazione di cointegrazione (ricordiamo essere r vettori di dimensione k), simile a quanto trovato nel caso unidimensionale:

$$\Pi = AB^\top, \qquad A, B \in \mathbb{R}^{k\times r}, \quad \operatorname{rank} A = \operatorname{rank} B = r.$$

La rappresentazione VECM diventa dunque

$$\Delta \boldsymbol{y}_t = AB^\top \boldsymbol{y}_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta \boldsymbol{y}_{t-j} + \varepsilon_t,$$

per cui $B = (\beta_1 \quad \beta_2 \quad \cdots \quad \beta_r) \in \mathbb{R}^{k\times r}$ contiene gli r vettori di cointegrazione.

Perché vale che $\boldsymbol{B^\top y} \sim$ I(0)? Perché la relazione di cointegrazione non è univoca e si preserva moltiplicando per la matrice A.

Normalizzazione di B Nell'ECM abbiamo imposto $(1, -\tilde{\beta})$ come relazione di integrazione. Il vincolo che si impone in questo caso è

$$B^\top = (I_r \quad -B^*),$$

da cui si può scrivere il modello lineare multivariato

$$\boldsymbol{y}_{1,t} = B^* \boldsymbol{y}_{2,t} + \eta_t^*,$$

con η_t errore di cointegrazione.

Lezione 18: Stimatori di cointegrazione

2020-11-23

Esempio (VECM per un VAR(1) bivariato)
Considerando un VAR(1), si ha

$$\boldsymbol{y}_t = \Phi_1 \boldsymbol{y}_{t-1} + \varepsilon_t,$$

con la forma VECM che diventa

$$\Delta \boldsymbol{y}_t = \Pi \boldsymbol{y}_{t-1} + \varepsilon_t$$

e se le serie sono cointegrate si ha una sola possibile relazione di cointegrazione, $\beta^* = (1 \quad -\beta_1^{-1}\beta_2)$

$$y_{1,t} = \beta^* y_{2,t} + u_t.$$

L'ordine delle variabili viene scelto come conseguenza della teoria economica. In questo caso la matrice Π ha rango 1 ed è pari a

$$\Pi = AB^\top = \begin{pmatrix} \alpha_1 & \alpha_1\beta_2^* \\ \alpha_1 & \alpha_1\beta_2^* \end{pmatrix}$$

La rappresentazione VECM diventa allora

$$\begin{aligned} \Delta y_{1,t} &= \alpha_1(y_{1,t-1} - \beta_2^* y_{2,t-1}) + \varepsilon_{1,t} \\ \Delta y_{2,t} &= \alpha_2(y_{1,t-1} - \beta_2^* y_{2,t-1}) + \varepsilon_{2,t} \\ &= \dots \end{aligned}$$

con α_2, β_2^* che hanno lo stesso segno, in modo che a una deviazione dal lungo periodo segua un aggiustamento.

Seguendo la procedura di Johansen, la stima dei parametri di un VAR con cointegrazione segue una procedura in più passi:

Procedura di Johansen per la stima del VECM
Per un VAR(p) cointegrato, si eseguono i seguenti passi:

1. Si stimano i parametri del VAR(p) per $\boldsymbol{y}_t$.
2. Si stima la matrice Π e si studia il rango per determinare le relazioni di cointegrazione.
3. Se necessario si impongono dei vincoli di normalizzazione su B.
4. Si stima il modello VECM.

Esempio

Nel caso bivariato, si deve valutare il rango di Π verificando l'ipotesi

$$\begin{cases} H_0 : r < 2 \\ H_1 : r = 2 \end{cases},$$

che è equivalente a

$$\begin{cases} H_0 : r = 1 \\ H_1 : r = 2 \end{cases} \quad \& \quad \begin{cases} H_0 : r = 0 \\ H_1 : r = 2 \end{cases},$$

In questo caso, H_1 : stazionarietà di $\boldsymbol{y}_t$.

18.1 Stimatori sotto Hp. di cointegrazione

Sotto ipotesi nulla $H_0 : r = 1$, si ha il modello

$$\Delta \boldsymbol{y}_t = AB^\top \boldsymbol{y}_{t-1} + \varepsilon_t,$$

e sotto ipotesi di normalità la verosimiglianza è pari a (conseguenza dei processi gaussiani):

$$\ell(\alpha, B, \Sigma) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{T}(\Delta \boldsymbol{y}_{t-1} - AB^\top \boldsymbol{y}_{t-1})^\top \Sigma^{-1}(\Delta \boldsymbol{y}_{t-1} - AB^\top \boldsymbol{y}_{t-1}),$$

che si può massimizzare rispetto ai parametri. Alternativamente, se si ipotizza che B sia noto, si può scrivere che

$$B^\top \boldsymbol{y}_{t-1} = \boldsymbol{z}_{t-1} \sim \mathrm{I}(0),$$

pertanto

$$\Delta \boldsymbol{y}_t = A \underbrace{B^\top \boldsymbol{y}_{t-1}}_{\boldsymbol{z}_{t-1}} + \varepsilon_t \sim \mathrm{I}(0)$$

e dunque si può stimare un modello di regressione (OLS, FGLS, ...) rispetto a $\boldsymbol{z}_{t-1}$ per ottenere $\hat{\alpha}(B)$ e $\hat{\Sigma}(B)$. Ottenute queste, si può massimizzare la *log-verosimiglianza profilo* (o *log-verosimiglianza concentrata* in econometria) $\ell_c(B)$ rispetto a B.

Prop.

Si può mostrare che massimizzare la verosimiglianza profilo $\ell_c(B)$ è equivalente a trovare l'autovettore associato all'autovalore massimo della matrice Π.

Osservazione

- Nel caso $k = 2$, sotto cointegrazione si ha per forza che $\operatorname{rank}\Pi = 1$, quindi $\hat{\lambda}_2 = 0$ e l'autovettore associato a $\hat{\lambda}_1$ corrisponde a $\hat{B}$.

Prop. (Proprietà degli stimatori nell'approccio di Johansen)
Gli stimatori secondo Johansen hanno le seguenti proprietà:

1. $\hat{B}$ *è superconsistente*
2. *I vettori di* $\hat{B}$ *hanno distribuzione non standard pari a una mistura di normali.*
3. *I test di ipotesi sui vettori di* B *hanno distribuzione asintotica* χ^2.
4. $\hat{\alpha}(B)$ *ha distribuzione asintotica gaussiana.*
5. $\hat{\Sigma}(B)$ *è uno stimatore consistente.*

18.1.1 Caso VAR(1), $k = 2$

Per studiare il numero di relazioni di cointegrazione, si va inizialmente a testare che $\operatorname{rank}\Pi = 1$, dunque $H_0 : r = 1$. Si dimostra che, sotto H_0, la log-verosimiglianza è pari a

$$\ell_C(B)\Big|_{H_0:r=1} \propto -\frac{T}{2}\log(1-\lambda_1),$$

mentre sotto l'ipotesi alternativa

$$\ell_C(B)\Big|_{H_1:r=2} \propto -\frac{T}{2}\left(\log(1-\lambda_1)+\log(1-\lambda_2)\right),$$

per cui si può costruire il test log-rapporto di verosimiglianza

$$\mathrm{LRT}_{\lambda_1,\lambda_2} = -T\log(1-\lambda_2).$$

Passando al secondo sistema di ipotesi, $H_0 : r = 0$ contro $H_1 : r = 2$, la statistica test diventa

$$\mathrm{LRT}_{\lambda_1,\lambda_2} = -T\left(\log(1-\lambda_1)+\log(1-\lambda_2)\right).$$

In entrambi i casi, le distribuzioni sono non standard e si usano valori tabulati per trovarne il p-value.

Osservazione

Le due statistiche si chiamano *"trace tests" di Johansen*, perché sono test che verificano quanti autovalori di Π sono diversi da zero. Inoltre, sempre Johansen suggerisce di partire dall'ipotesi $H_0 : r = 0$ e procedere in maniera sequenziale in avanti.

Un ulteriore test si può scrivere per verificare due ipotesi consecutive, per il caso in cui si debbano verificare valori di rango $0 < r < k$. In questo caso, si verificano le ipotesi

$$\begin{cases} H_0 : r = j \\ H_1 : r = j+1 \end{cases}$$

che nel caso $r = 0$ vs. $r = 1$ diventa

$$\mathrm{LRT} = -T\log(1-\lambda_1).$$

18.1.2 Caso VAR(p) generale

Nel caso generico VAR(p), si ha il modello VECM

$$\Delta \boldsymbol{y}_t = \Pi \boldsymbol{y}_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta \boldsymbol{y}_{t-j} + \varepsilon_t,$$

Rispetto al caso precedente, ci sono anche le differenze $\Delta \boldsymbol{y}_{t-j}$ che influenzano le procedure di stima. L'approccio di Johansen generale va a depurare $\boldsymbol{y}_{t-1}$ e $\Delta \boldsymbol{y}_t$ dall'effetto dei ritardi:

Prop. (Approccio di Johansen nel caso VAR(p))
Per il VAR(p) generico, si usa la seguente procedura:

1. *Si "ortogonalizza" rispetto alle differenze, calcolando $\hat{u}_t$ e $\hat{v}_t$ dalle regressioni*

$$\Delta \boldsymbol{y}_t = \sum_{j=1}^{p-1} W_j \Delta \boldsymbol{y}_{t-j} + u_t$$

$$\boldsymbol{y}_{t-1} = \sum_{j=1}^{p-1} W_j \Delta \boldsymbol{y}_{t-j} + v_t$$

2. *Si utilizzano $\hat{u}_t$ al posto di $\Delta \boldsymbol{y}_t$ e $\hat{v}_t$ al posto di $\boldsymbol{y}_{t-1}$ per calcolare la funzione di log-verosimiglianza profilo per B.*
3. *Ipotizzato un rango r, si stima $\hat{B}$.*
4. *Data la matrice stimata $\hat{B}$, si possono stimare i rimanenti parametri del modello con i minimi quadrati su*

$$\Delta \boldsymbol{y}_t = A \hat{B}^\top \boldsymbol{y}_{t-1} + \sum_{j=1}^{p} \Gamma_j \Delta \boldsymbol{y}_{t-j} + \varepsilon_t.$$

Osservazione

Nel caso VAR(p), si avranno gli autovalori

$$\lambda_1 > \lambda_2 > \ldots > \lambda_k,$$

con gli ultimi $k - r$ autovalori nulli se $\operatorname{rank} \Pi = r < k$. Si costruirà allora un test per verificare che siano nulli, con la log-verosimiglianza concentrata che nel punto di massimo vale

$$\ell_C(B) \propto -\frac{T}{2} \log \Big(\sum_{j=1}^{r} (1 - \lambda_j) \Big)$$

Il test sequenziale diventa allora, per $r = 1, \ldots, k$

$$\begin{cases} H_0 : r = j \\ H_1 : r = j + 1 \end{cases}$$

con i test basati sul rapporto di verosimiglianza che hanno forma pari a

$$\text{LRT} = -T \log(1 - \lambda_{r+1}).$$

Lezione 19: Modelli State-Space

2020-11-30

I modelli state-space comprendono tutti i modelli lineari, statici e dinamici, oltre ai modelli autoregressivi e per risposte non continue.

In questa sezione ci occuperemo della versione gaussiana dei modelli state space.

I modelli state-space sono costituiti da due equazioni, una sulle variabili osservate detta *di misura* (*measurement equation*) e una che rappresenta l'evoluzione dinamica degli stati latenti, detta di *transizione* (*transition equation*),

$$\boldsymbol{y}_t = \boldsymbol{c}_t + Z_t\boldsymbol{\alpha}_t + G_t\boldsymbol{\varepsilon}_t, \qquad \text{EQUAZIONE DI MISURA}$$

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{d}_t + T_t\boldsymbol{\alpha}_t + H_t\boldsymbol{\varepsilon}_t. \qquad \text{EQUAZIONE DI TRANSIZIONE}$$

I vettori e le matrici che compaiono nelle equazioni sono

- $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, \dots, y_{k,t})^\top \in \mathbb{R}^{k\times 1}$ è il vettore di variabili osservate.
- $\boldsymbol{\alpha}_t \in \mathbb{R}^{d\times 1}$ è il vettore degli stati latenti.
- $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, I_{k+m})$ è il vettore delle innovazioni, condiviso tra $\boldsymbol{y}$ e $\boldsymbol{\alpha}$.
- $\boldsymbol{c}_t \in \mathbb{R}^{k\times 1}$, $\boldsymbol{d}_t \in \mathbb{R}^{d\times 1}$.
- Z_t è la matrice dei *loading factors* per gli stati latenti.
- G_t è una matrice quadrata tale che $GG^\top$ è definita positiva.
- T_t è la matrice dell'evoluzione VAR degli stati latenti.
- H_t quadrata tale che $HH^\top$ è definita positiva.

Osservazioni

- È sufficiente un VAR(1) nell'equazione dinamica per descrivere una vasta classe di modelli, perché ogni VAR(p) si può scrivere in forma compagna come VAR(1).
- In generale, $\boldsymbol{\alpha}_t$ non è osservato e si assume sia un *fattore latente*, la cui stima eventualmente si può ricavare con la sua distribuzione a posteriori.
- Di solito si definisce uno stato iniziale Y_0, da cui si ricava la distribuzione iniziale per α_1 come

  $$\alpha_1 \sim \mathcal{N}\big(\mathbb{E}\left[\alpha_1|Y_0\right], \mathbb{V}\left[\alpha_1|Y_0\right]\big).$$

  Quando il processo per α_t contiene componenti non stazionarie, i momenti condizionali non sono finiti e, in tal caso, si può inizializzare in modo diffuso come

  $$P_{1|0} = kI_{d+m}, \qquad k \text{ grande}.$$

- Nel caso generale le quantità $\boldsymbol{c}_t, Z_t, G_t, \boldsymbol{d}_t, T_t, H_t$ possono variare nel tempo ma, tipicamente, vengono assunte come quantità deterministiche nel tempo.

Al massimo possono cambiare deterministicamente, ma in ogni caso non possono essere funzioni di altri processi stocastici.

› ε_t è in comune ai due processi, per cui vi è una correlazione tra y_t e α_t se $G_t \not\perp H_t$:

$$\operatorname{Cov}\left(y_t, \alpha_{t+1} | \alpha_t\right) = G_t H_t^\top.$$

In quasi tutti i libri, si assume $G_t \perp H_t$ per semplificare.

Il seguente teorema è fondamentale per poter scrivere il *filtro di Kalman*, che permette di derivare le previsioni del processo sulla base dei dati osservati.

Teo. (Distribuzione condizionata per le normali)
Se il vettore $(Y, X)^\top$ *ha distribuzione*

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\Big(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \Big),$$

allora

$$Y|X \sim \mathcal{N}(\mu_{Y|X}, \Sigma_{Y|X}),$$

con

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Osservazione

La verosimiglianza completa del processo non è scrivibile, poiché non si osservano i processi latenti al meno di assumere che $\boldsymbol{\alpha}_t$ sia una funzione deterministica, e in tal caso si ottiene semplicemente un modello lineare per $\boldsymbol{y}_t$

$$\boldsymbol{y}_t \sim \mathcal{N}(\boldsymbol{c}_t + Z_t\boldsymbol{\alpha}_t, G_t G_t^\top).$$

L'approccio per la stima e previsione comporta una marginalizzazione degli α_t, attraverso l'applicazione del filtro di Kalman.

Esempio (Random walk con noise)
L'esempio più semplice di modello state-space non triviale è

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t & \varepsilon_t &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \\ \mu_{t+1} &= \mu_t + \eta_t & \eta_t &\sim \mathcal{N}(0, \sigma_\eta^2) \end{aligned} \tag{26}$$

Questo modello generalizza il random walk semplice, aggiungendo una componente aggiuntiva di disturbo nel p.

Osservazione

Se $\sigma_\varepsilon^2 = 0$, allora ε_t può eliminare e (26) diventa un modello autoregressivo completamente osservato:

$$\begin{cases} y_t = \mu_t \\ \mu_t = \mu_{t-1} + \eta_t \end{cases} \implies y_t = y_{t-1} + \eta_{t-1}.$$

Se invece $\sigma_\varepsilon^2 > 0$, allora applicando la differenza prima si ha

$$\Delta y_t = \eta_{t-1} + \varepsilon_t - \varepsilon_{t-1},$$

per cui dalla condizione $\varepsilon_t \perp \eta_t$, si ha che l'autocorrelazione del processo è pari a

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} 1 & \tau = 0 \\ -\frac{\sigma_\varepsilon^2}{2\sigma_\varepsilon^2 + \sigma_\eta^2} & \tau = 1 \\ 0 & \tau > 1 \end{cases}$$

Osservazione

Basandoci sull'esempio precedente, non c'è differenza tra le scritture

$$\Delta y_t \sim \text{IMA}(1,1) \qquad \text{forma strutturale}$$

$$y_t \sim \text{State-Space} \qquad \text{forma ridotta}$$

1. **Forma strutturale**: informazione sulla struttura del processo che ha generato i dati.
2. **Forma ridotta**: informazioni su come potrebbe essere fatto y in seguito all'ipotesi strutturale.

La caratteristica speciale di questi modelli è che il grosso del lavoro sta nello scrivere il processo nella forma state-space, perché applicando il filtro di Kalman si può ottenere tutto quello che serve.

Osservazioni

› Con l'approccio basato sul filtro di Kalman si possono analizzare tranquillamente anche processi non stazionari, senza quindi rinunciare a modellare la variabilità del processo come nella differenziazione.

› L'inferenza *esatta* sui processi ARMA gaussiani avviene attraverso la rappresentazione state-space, che non necessita di condizionarsi alle prime p osservazioni.

› L'inferenza esatta, che sfrutta tutta l'informazione nel singolo campione senza perdere dati, è rilevante quando i processi sono ad alta dimensionalità e con poche osservazioni (PIL, ...).

Lezione 20

2020-12-01

Esempio (Modello di regressione dinamico)
Il modello di regressione dinamico consente di utilizzare parametri variabili nel tempo e verificare di conseguenza se ci sono stati cambiamenti sostanziali nel valore di β_t:

$$y_t = \mu_t + \boldsymbol{x}_t^\top \boldsymbol{\beta}_t + \sigma_\varepsilon \varepsilon_t$$

$$\mu_{t+1} = \varphi \mu_t + \sigma_\eta \eta_t^{(\mu)}$$

$$\boldsymbol{\beta}_{t+1} = \Phi \boldsymbol{\beta}_t + \Sigma^{1/2} \boldsymbol{\eta}_t^{(\beta)}$$

Solitamente, si assume che $\varphi = 1$ e $\Phi = I_p$, da cui la rappresentazione state-space mediante i parametri

$$\begin{cases} \boldsymbol{\alpha}_t = (\mu_t, \beta_{1,t}, \dots, \beta_{p,t})^\top \\ \boldsymbol{\vartheta} = \left(\sigma_\varepsilon, \sigma_\eta, \varphi, \operatorname{vec}(\Phi)^\top, \operatorname{vec}(\Sigma)^\top\right)^\top \end{cases}$$

e inoltre $\varepsilon_t \sim \mathcal{N}(0,1)$, $\eta_t^\mu \sim \mathcal{N}(0,1)$ e $\eta_t^\beta \sim \mathcal{N}(0, I)$.

Esempio (AR(1) con noise)
Il processo AR(1) con noise è tale che

$$\begin{aligned} y_t &= c + Z\alpha_t + G\varepsilon_t \\ &= \mu + \begin{pmatrix} \sigma_\varepsilon & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_t/\sigma_\varepsilon \\ \eta_t/\sigma_\eta \end{pmatrix} \\ &= \mu_t + \varepsilon_t. \end{aligned}$$

L'equazione di transizione diventa

$$\begin{aligned} \alpha_{t+1} &= T_t \alpha_t + H_t \varepsilon_t \\ &= \varphi \mu_t + \begin{pmatrix} 0 & \sigma_\eta \end{pmatrix} \begin{pmatrix} \varepsilon_t/\sigma_\varepsilon \\ \eta_t/\sigma_\eta \end{pmatrix}. \end{aligned}$$

Si nota inoltre che $H_t G_t^\top = \begin{pmatrix} 0 & \sigma_\eta \end{pmatrix} \begin{pmatrix} \sigma_\varepsilon \\ 0 \end{pmatrix} = 0$, per cui $H_t \perp G_t$.

20.1 ARMA(p,q) state-space

Considerato un modello ARMA(p, q)

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \xi_t + \vartheta_1 \xi_{\xi-1} + \vartheta_2 \xi_{t-1} + \ldots + \vartheta_q \xi_{t-q},$$
$$\xi_t \sim \mathcal{N}(0, \sigma_\xi^2),$$

si può costruire la sua forma state-space con $m = \max\{p, q+1\}$ elementi usando la forma compagna per la componente AR e facendo entrare nella matrice di varianza la dinamica MA.

In particolare, per il modello

$$\boldsymbol{y}_t = \boldsymbol{c}_t + Z_t \boldsymbol{\alpha}_t + G_t \boldsymbol{\varepsilon}_t,$$
$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{d}_t + T_t \boldsymbol{\alpha}_t + H_t \boldsymbol{\varepsilon}_t$$

si può imporre $\boldsymbol{\varepsilon}_t = \xi_{t+1}$, $Z_t = Z = (1 \quad 0 \quad 0 \quad \cdots \quad 0)$, $G_t = G = 0$ e dunque $y_t = \alpha_{1,t}$. La matrice di transizione è per convenzione il trasposto della forma compagna che abbiamo introdotto per il VAR:

$$T_t = T = \begin{pmatrix} \varphi_1 & 1 & 0 & \ldots & 0 \\ \varphi_2 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi_{m-1} & 0 & 0 & \ldots & 1 \\ \varphi_m & 0 & 0 & \ldots & 0 \end{pmatrix}$$

Nel caso in cui $q > p$, si inseriscono comunque i termini φ_j di ordine superiore a p, ma non vengono usati nella transizione.

La matrice H_t è formata dal seguente vettore, con zeri alla fine se $q < m - 1$ (non si aggiungono termini come nella parte AR):

$$H_t = H = \begin{pmatrix} 1 \\ \vartheta_1 \\ \vdots \\ \vartheta_{m-1} \end{pmatrix}.$$

Le costanti c_t e d_t sono poste pari a zero, mentre nel caso in cui sia presente la media si può porre $d = c$.

Esempio (Processo AR(2))
Il processo AR(2) è definito come

$$y = \mu + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \xi_t,$$

dunque $m = \max\{p, q+1\} = 2$ e

$$y_t = (1 \quad 0) \begin{pmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{pmatrix} + 0 = \alpha_{1,t}.$$

La dinamica per la transizione è

$$\begin{pmatrix}\alpha_{1,t+1} & \alpha_{2,t+1}\end{pmatrix} = \begin{pmatrix}\varphi_1 & 1\\ \varphi_2 & 0\end{pmatrix}\begin{pmatrix}\alpha_{1,t}\\ \alpha_{2,t}\end{pmatrix} + \begin{pmatrix}1\\ 0\end{pmatrix}\xi_{t+1},$$

con matrice $H_t = \begin{pmatrix}1\\ \vartheta_1 = 0\end{pmatrix}$.

Osservazione

L'equazione di misura dice che il primo stato è esattamente pari a y, ed espandendo si ha

$$\alpha_{2,t} = \varphi_2\alpha_{1,t-1} \tag{27}$$

Mentre per il secondo stato, sostituendo (27), si ha

$$\begin{aligned}\alpha_{1,t+1} &= \varphi_1\alpha_{1,t} + \alpha_{2,t} + \xi_{t+1}\\ &= \varphi_1\alpha_{1,t} + \varphi_2\alpha_{1,t-1} + \xi_{t+1}\end{aligned}$$

e poiché $y = \alpha_{1,t+1}$, si ha

$$y_{t+1} = \varphi_1 y_t + \varphi_2 y_{t-1} + \xi_{t+1},$$

per cui si ha esattamente la rappresentazione AR(2).

Esempio (Local trend model)

Il local-trend model (LTM) generalizza il random walk con noise aggiungendo un drift:

$$\begin{aligned}y_t &= \mu + \varepsilon_t\\ \mu_{t+1} &= \mu_t + \beta_t + \eta_{1,t}\\ \beta_{t+1} &= \beta_t + \eta_{2,t}\end{aligned}$$

Se β_t fosse eliminato dalla seconda equazione, sarebbe un RW + noise.

Definendo $\alpha_t = (\mu_t \quad \beta_t)^\top$, per scrivere la forma state-space si può usare

$$\begin{aligned}y &= (1 \quad 0)\alpha_t + \varepsilon_t,\\ T_t &= \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}.\end{aligned}$$

Osservazione

Questa è una specificazione *gerarchica* dei fattori latenti: i modelli state-space permettono di costruire modelli gerarchici nello spazio e nel tempo,

Esempio (Processo VAR(p))
In un processo VAR(p)

$$y_t = \nu + \sum_{j=1}^{p} \Phi_j y_{t-j} + \varepsilon_t,$$

si può usare la forma compagna

$$\dots$$

ed utilizzare la matrice di selezione $J = (I_d \quad \mathbf{0}_d \quad \cdots \quad \mathbf{0}_d)$

Esempio (time-varying parameters VAR)
Andando oltre ai processi VAR, si possono costruire processi VAR con componenti dinamiche

$$y_t = \nu_t + \sum_{j=1}^{p} \Phi_{j,t} y_{t-j} + \varepsilon_t,$$

con $\Phi_{t,j}$ che segue una determinata dinamica. Definendo $\beta_t = \operatorname{vec} B_t$, si ha l'equazione di misura

$$y_t = \underbrace{(Z_{t-1}^\top \otimes I_d)}_{Z_t} \beta_t + \varepsilon_t,$$

a cui si può aggiungere una dinamica autoregressiva come nel modello di regressione:

$$\beta_{t+1} = \Theta \beta_t + \Omega^{1/2} \eta_t.$$

In particolare, nella dinamica autoregressiva si possono specificare forme molto varie.

Rispetto al modello lineare dinamico, dove y_t è scalare, nel VAR dinamico si ha $\beta_t \in \mathbb{R}^{(d^2 p + d) \times 1}$, per cui sia y_t sia il numero di stati potrebbero essere enormi e ci sono problematiche di elevata dimensionalità.

Esempio (Modello fattoriale dinamico)
Il modello fattoriale classico si può costruire come

$$y = \Lambda f + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2 I)$$

$$, f \sim \mathcal{N}(0, D),$$

da cui la varianza del modello

$$\mathbb{V}\left[y\right] = \Lambda D \Lambda^\top + \sigma^2 I.$$

Un analogo modello fattoriale dinamico si può ottenere attraverso la specificazione

dell'equazione di misura:

$$Y_t = \Lambda_0 f_t + \Lambda_1 f_{t-1} + \ldots + \Lambda_s f_{t-s} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, R),$$

dove Λ_j è la matrice dei factor loadings dinamici per f_{t-j} che seguono un processo VAR(p)

$$f_t = \Phi_1 f_{t-1} + \Phi_2 f_{t-2} + \ldots + \Phi_p f_{t-p} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q).$$

La versione più semplice utilizzerebbe la seguente specificazione

$$Y_t = \Lambda_0 f_t + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, R),$$

$$f_t = \Phi_1 f_{t-1} + \eta_t \qquad \eta_t \sim \mathcal{N}(0, Q).$$

che è esattamente un modello state-space che rappresenta l'estensione dinamica del modello fattoriale base.

Osservazioni

- Il modello con tutti i fattori migliora la previsione con una componente di persistenza, mentre la componente contemporanea è quella più rilevante per l'interpretazione.
- Aggiungere i fattori con lag porta ad avere un modello molto più grande, ma con la forma compagna si può scrivere in forma state-space e stimare.

Identificabilità dei modelli fattoriali

I modelli fattoriali non sono sempre identificati, ma si può dimostrare che le restrizioni che portano all'identificabilità possono essere le seguenti:

Teo. (Condizione sufficiente per l'identificabilità dei fattori)
Per avere identificabilità nel modello è sufficiente imporre che

$$\mathbb{V}\left[\eta_t\right] = I_q,$$

$$\Lambda_0 = \begin{pmatrix} \Lambda_{01} \\ \Lambda_{02} \end{pmatrix},$$

con Λ_{01} *triangolare inferiore e con elementi sulla diagonale strettamente positivi.*

Lezione 21: Filtraggio

2020-12-14

Modello *local-level* (o random walk + noise) è di tipo

$$y_t = \mu_t + \varepsilon_t \qquad \text{variabile osservata}$$

$$\mu_{t+1} = \mu_t + \eta_t \qquad \text{dinamica latente}$$

in particolare, il processo per μ_t è ignoto e si può fare inferenza solamente indirettamente tramite y_t.

Il filtro di Kalman è un algoritmo che processa tutte le $t = 1, \dots, T$ in modo da ottenere la stima per ogni tempo $t = 1, 2, \dots, T$ data da

$$\hat{\mu}_{t|t} = \mathbb{E}\left[\mu_t | Y_{1:t}\right].$$

In generale, se il modello è di tipo state-space gaussiano,

$$y_t = c_t + Z_t \alpha_t + G_t \varepsilon_t$$

$$\alpha_{t+1} = d_t + T_t \alpha_t + H_t \varepsilon_t,$$

$$\alpha_1 \sim \mathcal{N}_d(\alpha_{1|0}, P_{1|0})$$

allora si ha il seguente risultato:

Teo. (Filtro di Kalman)
Sotto le ipotesi, si ha che

$$\alpha_{t+1} | \mathcal{Y}_{1:t} \sim \mathcal{N}(\hat{\alpha}_{t+1|t}, P_{t+1|t})$$

$$\hat{\alpha}_{t+1|t} = \mathbb{E}\left[\alpha_{t+1} | y_{1:t}\right],$$

$$P_{t+1|t} = \mathbb{V}\left[\alpha_{t+1} | y_{1:t}\right]$$

e la previsione per y_{t+1} *è*

$$Y_{t+1} | \mathcal{Y}_{1:t} \sim \mathcal{N}(\hat{y}_{t+1|t}, F_{t+1|t})$$

$$\hat{y}_{t+1|t} = \mathbb{E}\left[Y_{t+1} | y_{1:t}\right],$$

$$F_{t+1|t} = \mathbb{V}\left[Y_{t+1} | y_{1:t}\right]$$

Teo. (Equazioni del filtro di Kalman)
La ricorsione del filtro di Kalman consiste delle seguenti equazioni:

$$\begin{aligned}
\nu_t &= y_t - c_t Z_t \hat{\alpha}_{t|t-1} && \textit{predittiva di } y_{t+1|t} \\
F_t &= Z_t P_{t|t-1} Z_t^\top + G_t G_t^\top && \textit{''} \\
K_t &= (T_t P_{t|t-1} Z_t^\top + H_t G_t^\top) F_t^{-1} && \textit{Kalman gain} \\
\hat{\alpha}_{t+1|t} &= d_t + T_t \hat{\alpha}_{t|t-1} + K_t \nu_t && \textit{predittiva di } \alpha_{t+1|t} \\
P_{t+1|t} &= T_t P_{t|t-1} T_t^\top + H_t H_t^\top - K_t F_t K_t^\top && \textit{''}
\end{aligned}$$

Equazione (1) e (2) danno le distribuzioni predittive per $y_{t+1|t}$, mentre (4) e (5) sono la predittiva degli stati α al tempo t. L'equazione (3) è quello che lega la previsione allo stato latente e prende il nome di *Kalman gain*.

Dim.
Il punto di partenza è la distribuzione predittiva al tempo t:

In generale,

$$\begin{aligned}
\mathbb{E}\left[y_t | \mathcal{Y}_{1:t-1}\right] &= \mathbb{E}\left[c_t + Z_t \alpha_t + G_t \varepsilon_t | Y_{1:t-1}\right] \\
&= c_t + Z_t \hat{\alpha}_{t|t-1},
\end{aligned}$$

dunque se non si conosce $\hat{\alpha}_{t|t-1}$ non si può ottenere la previsione per y_t.

Assumendo di aver osservato y_t, l'innovazione sarà

$$\begin{aligned}
\nu_t &= y_t - \mathbb{E}\left[y_t | \mathcal{Y}_{1:t-1}\right] \\
&= y_t - c_t - Z_t \hat{\alpha}_{t|t-1}. \\
&= Z_t \alpha_t + G_t \varepsilon_t - Z_t \hat{\alpha}_{t|t-1} \\
&= Z_t \underbrace{\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)}_{\text{stocastico}} + \underbrace{G_t \varepsilon_t}_{\text{stocastico}}
\end{aligned}$$

Il tutto diventa molto complesso, perché ci sono due componenti stocastiche che fanno parte dell'innovazione. Nel caso VAR si aveva un'innovazione data da una componente non stocastica.

$$\text{innov.} = y_t - \Phi y_{t-1}.$$

Le fonti di variazione sono

1. $\alpha_t - \hat{\alpha}_{t|t-1}$, tale che

$$\mathbb{E}\left[\alpha_t - \hat{\alpha}_{t|t-1} | \mathcal{Y}_{1:t-1}\right] = 0,$$

   con varianza

$$\mathbb{V}\left[\alpha_t - \hat{\alpha}_{t|t-1} | \mathcal{Y}_{1:t-1}\right] = P_{t|t-1}.$$

2. $\mathbb{E}\left[G_t\varepsilon_t|\mathcal{Y}_{1:t-1}\right] = 0.$

Dunque, si ha che

$$\begin{aligned}\mathbb{E}\left[\nu_t|\mathcal{Y}_{1:t-1}\right] &= 0\\ \mathbb{E}\left[\nu_t\right] &= 0\end{aligned}$$

Questo mostra che ν_t è effettivamente un'innovazione, in quanto

$$\mathrm{Cov}(\nu_t, \nu_{t-1}) = 0.$$

Per la varianza dell'innovazione, si ha

$$\begin{aligned}\mathbb{V}\left[\nu_t|\mathcal{Y}_{1:t}\right] &= \mathbb{E}\left[\nu_t\nu_t^\top|\mathcal{Y}_{1:t-1}\right]\\ &= \dots\\ &= Z_t\,\mathbb{E}\left[(\alpha_t - \hat{\alpha}_{t|t-1})(\alpha_t - \hat{\alpha}_{t|t-1})^\top|\mathcal{Y}_{1:t-1}\right]Z_t^\top + G_t\,\mathbb{E}\left[\varepsilon_t\varepsilon_t^\top|\mathcal{Y}_{1:t-1}\right]G_t^\top\\ &= Z_tP_{t|t-1}Z_t^\top + G_tG_t^\top\\ &= F_t.\end{aligned}$$

□

Riferimenti bibliografici

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*. New York: Springer Nature.

Diggle, P. et al. (2013). *Analysis of Longitudinal Data (Oxford Statistical Science): NCS P: 25*. Oxford: Oxford University Press, Usa.

Durbin, T. l. J. e Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. 2 edizione. Oxford: OUP Oxford.

Fahrmeir, L. e Tutz, G. (2010). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York; London: Springer Nature.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, N.J: Princeton Univ Pr.

Kroese, D. P. e Chan, J. C. C. (2013). *Statistical Modeling and Computation*. New York: Springer-Nature New York Inc.

Kwon, C. (2016). *Julia Programming for Operations Research: A Primer on Computing*. Charleston, SC: CreateSpace Independent Publishing Platform.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin Heidelberg: Springer-Verlag.

Robert, C. P. e Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer Nature.

Shumway, R. H. e Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. 4th ed. New York, NY: Springer.