Appunti di analisi delle serie storiche

Riccardo 'Jack' Lucchetti

25 marzo 2013

Indice

Prefazione **vii**

1 Introduzione **1**
1.1 Cos'è un processo stocastico e a che serve 1
1.2 Caratteristiche dei processi stocastici 2
1.3 Momenti 5
1.4 Qualche esempio 6

2 I processi ARMA **13**
2.1 L'operatore ritardo 13
2.2 Processi *white noise* 16
2.3 Processi MA 18
2.4 Processi AR 23
2.5 Processi ARMA 28
2.6 Uso dei modelli ARMA 32
2.6.1 Previsione 32
2.6.2 Analisi delle caratteristiche dinamiche 36
2.7 Stima dei modelli ARMA 39
2.7.1 Tecniche numeriche 40
2.7.2 Scelta degli ordini dei polinomi 41
2.7.3 Calcolo della verosimiglianza 44
2.8 In pratica 46

3 Processi integrati **59**
3.1 Caratteristiche delle serie macroeconomiche 59
3.2 Processi a radice unitaria 62
3.3 La scomposizione di Beveridge e Nelson 66
3.4 Test di radice unitaria 68
3.4.1 Distribuzione della statistica test 70
3.4.2 Persistenza di breve periodo 70
3.4.3 Nucleo deterministico 72
3.4.4 Test alternativi 73
3.4.5 Usare il cervello 74
3.4.6 Un esempio 75
3.5 Regressione spuria 76

4 Processi VAR **81**
4.1 Processi multivariati 81
4.2 I processi VAR 83
4.3 Stima dei VAR 88
4.4 VAR integrati 91
4.5 Uso dei VAR 93
4.5.1 Previsione 94
4.5.2 Analisi di causalità 95
4.5.3 Analisi dinamica 98

5 Cointegrazione **107**
5.1 Definizioni 107
5.2 Proprietà dei vettori di cointegrazione 108
5.3 Modelli a correzione d'errore 110
5.4 Il teorema di rappresentazione di Granger 116
5.4.1 Un po' di algebra matriciale 117
5.4.2 Il teorema vero e proprio 118
5.4.3 Nucleo deterministico 119
5.5 Tecniche di stima 122
5.5.1 La procedura di Johansen 123
5.5.2 Procedure alternative 128

6 Processi a volatilità persistente **133**
6.1 I fatti stilizzati 134
6.2 Processi ARCH e GARCH 137
6.2.1 Processi ARCH 137
6.2.2 Processi GARCH 140
6.2.3 Stima dei GARCH 141
6.3 Un esempio 142
6.4 Estensioni 145
6.4.1 Distribuzioni non-normali 145
6.4.2 Effetti asimmetrici 148
6.4.3 EGARCH 148
6.4.4 GARCH-in-mean 148
6.4.5 IGARCH 149
6.4.6 Modelli multivariati 149

7 Per approfondimenti **151**
7.1 In generale 151
7.2 Processi univariati 151
7.3 Processi VAR 152
7.4 Processi $I(1)$ e cointegrazione 152
7.5 Processi ad eteroschedasticità condizionale 153

Bibliografia **154**

Elenco delle figure

1.1 Variazioni mensili della produzione industriale USA 7
1.2 Variazioni produzione industriale USA – correlogramma 7
1.3 Inflazione USA 9
1.4 Inflazione USA – correlogramma 9
1.5 Indice Nasdaq – rendimenti giornalieri 10
1.6 Indice Nasdaq – Correlogramma 10
1.7 Indice Nasdaq – rendimenti giornalieri in valore assoluto . . . 11
1.8 Indice Nasdaq – Correlogramma dei valori assoluti 11

2.1 MA(1): $\theta = 0$ (*white noise*) 19
2.2 MA(1): $\theta = 0.5$ 20
2.3 MA(1): $\theta = 0.9$ 21
2.4 MA(1): Autocorrelazione di primo ordine in funzione di θ . . . 21
2.5 AR(1): $\varphi = 0$ (*white noise*) 26
2.6 AR(1): $\varphi = 0.5$ 27
2.7 AR(1): $\varphi = 0.9$ 27
2.8 AR(2): $\varphi_1 = 1.8$; $\varphi_2 = -0.9$ 29
2.9 Risposta di impulso per $y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$ . . . 38
2.10 Produzione industriale negli USA (dal 1921) 47
2.11 Logaritmo della produzione industriale negli USA (mensile) . . 48
2.12 Variazione percentuale della produzione industriale 49
2.13 Correlogrammi della produzione industriale 49
2.14 Risposte di impulso 53
2.15 Previsioni 53
2.16 Rappresentazione grafica di un numero complesso 55

3.1 log(PIL) 59
3.2 log(PIL) e trend deterministico 60
3.3 Residui 61
3.4 Δlog(PIL) 62
3.5 Random walk 64
3.6 Funzione di densità del test DF 71
3.7 Funzione di densità del test DF con intercetta 72

4.1 Autovalori della *companion matrix* 87
4.2 PIL e Consumi nell'UE 90

4.3 Risposte di impulso non strutturali 102
4.4 Risposte di impulso strutturali 103

5.1 VAR(1) stazionario: serie storiche simulate 112
5.2 VAR(1) stazionario: serie storiche simulate – diagramma XY . . 112
5.3 *Random walk*: serie storiche simulate 113
5.4 *Random walk*: serie storiche simulate – diagramma XY 114
5.5 Processo cointegrato: serie storiche simulate 114
5.6 Processo cointegrato: serie storiche simulate – diagramma XY . 115

6.1 Indice Nasdaq – logaritmi 134
6.2 Indice Nasdaq – rendimenti giornalieri 135
6.3 Rendimenti Nasdaq – valori assoluti 135
6.4 Rendimenti Nasdaq – distribuzione marginale 136
6.5 Rendimenti Nasdaq – residui e deviazione standard stimata . . 144
6.6 Rendimenti Nasdaq – serie standardizzata 145
6.7 Distribuzioni alternative alla normale 147

Prefazione

Questo scritto era nato come dispensa per il mio corso di Econometria. In quanto tale, non mi sono mai posto obiettivi particolarmente ambiziosi né per quanto riguarda il rigore, né per la completezza. L'obiettivo principale era, al contrario, quello di descrivere i concetti facendo perno principalmente sull'intuizione del lettore, cercando di motivare nel modo più esplicito possibile l'introduzione delle definizioni e dei risultati principali.

Le cose, poi, si sono evolute nel tempo e la dispensa è cresciuta: non la posso più usare come tale nel corso di Econometria di base, ma la uso per corsi più avanzati. La filosofia di base però è rimasta la stessa: un testo che si può "leggere", oltreché "studiare". Di conseguenza, a parte qualche eccezione, farò genericamente riferimento "alla letteratura" per spiegazioni, dimostrazioni e approfondimenti, senza citare fonti specifiche. Questo perché ho ritenuto più utile, dato lo scopo che mi propongo, raggruppare le indicazioni bibliografiche in un ultimo capitolo, che avesse anche la funzione di orientare il lettore nel *mare magnum* dell'econometria delle serie storiche.

Negli anni, ho avuto moltissimo *feedback* da parte di molte persone, che ringrazio per aver contribuito a migliorare il contenuto. Fra gli amici che fanno il mio stesso mestiere voglio ricordare (senza per questo chiamarli in correo) in particolare Gianni Amisano, Marco Avarucci, Emanuele Bacchiocchi, Nunzio Cappuccio, Francesca Di Iorio, Luca Fanelli, Massimo Franchi, Carlo Favero, Roberto Golinelli, Diego Lubian, Giulio Palomba, Matteo Pelagatti, Eduardo Rossi, Maurizio Serva, Stefano Siviero e Gennaro Zezza. Carlo Giannini merita una menzione a parte, perché senza di lui io probabilmente nella vita avrei fatto tutt'altro e questa dispensa non sarebbe mai esistita; sicuramente io sarei stato una persona peggiore.

Un pensiero riconoscente va poi a tutti coloro che si sono visti inflitta questa dispensa come libro di testo e mi hanno indotto ad essere più completo e chiaro (o meno incompleto ed oscuro, a seconda dei punti di vista) quando mi facevano notare, a parole o semplicemente con l'espressione del viso, che non ci si capiva niente. Non vorrei fare nomi perché sono troppi, ma devo fare un'eccezione per Gloria Maceratesi, che non posso non menzionare perché la sua efficienza di correttrice ha avuto del sovrumano. Grazie comunque a tutti quanti. Il fatto poi che questa dispensa sia liberamente disponibile su Internet ha anche indotto molti a scaricarla, e qualcuno mi ha anche scritto una mail con consigli e suggerimenti. Anche in questo caso, nutro grande riconoscenza, se non altro perché ha fatto bene al mio ego.

Un grazie grande come una casa va ad Allin Cottrell, che è la sbuffante locomotiva dietro il progetto *gretl*: per chi non lo sapesse, gretl è un pacchetto econometrico *free*[1] con cui sono stati realizzati tutti gli esempi contenuti in questa dispensa. Per saperne di più, e magari scaricarlo, andate su `http://gretl.sourceforge.net`.

Per quanto riguarda i prerequisiti, presuppongo che il lettore abbia già un certo grado di familiarità con i concetti probabilistici base (variabili casuali, valori attesi, condizionamento, vari modi di convergenza), con il modello OLS e con alcuni concetti base di teoria della stima, come identificazione e proprietà degli stimatori. Quindi, chi non se li è studiati già, può anche chiudere qui e andare a studiare. Gli altri, si mettano pure comodi, che andiamo a incominciare.

Alcuni passi sono scritti in un carattere più piccolo, su due colonne, come questo. Essi non sono indispensabili, e possono essere saltati senza pregiudizio della comprensione del resto. Certo però che, se li ho scritti, a qualcosa serviranno pure. Fate voi.

---

[1] Che vuol dire *anche* gratuito. L'espressione *free software*, però, di solito si traduce con "software libero", perché è disponibile il sorgente. È in ogni caso imperdonabile confondere il *free software* col *freeware*, che è semplicemente software che si può usare legalmente senza pagare.

Capitolo 1

Introduzione

1.1 Cos'è un processo stocastico e a che serve

I dati a cui vengono applicate le tecniche inferenziali che compongono il bagaglio dell'econometrico possono essere di due tipi: *cross-section*, nel caso in cui le osservazioni di cui disponiamo siano relative ad individui diversi, oppure *serie storiche*, quando ciò che abbiamo sono osservazioni, su una o più grandezze, protratte nel tempo[1].

Nel primo caso, pensare ad un insieme di N dati osservati come una delle possibili realizzazioni di N variabili casuali indipendenti ed identiche non è un'ipotesi troppo insostenibile: se rilevo peso e statura di N individui, non c'è ragione di pensare che

1. le caratteristiche fisiche dell'i-esimo individuo siano in qualche modo connesse a quelle degli altri individui (*indipendenza*);

2. la relazione fra peso e altezza che vale per l'i-esimo individuo sia diversa da quella che vale per tutti gli altri (*identicità*).

In questi casi, ci serviamo del concetto di realizzazione di una variabile casuale come metafora dell'i-esima osservazione, e l'apparato inferenziale appropriato non è diverso da quello standard, in cui l'indipendenza e l'identicità ci consentono di dire che

$$f(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f(x_i),$$

cioè che la funzione di densità del nostro campione è semplicemente la produttoria delle funzioni di densità delle singole osservazioni (le quali funzioni sono tutte uguali). Nel caso in cui lo strumento di analisi sia la regressione lineare, questo quadro di riferimento ci porta sostanzialmente alle cosiddette "ipotesi classiche", ampiamente analizzate al principio di qualunque corso di

[1] A dir la verità, un caso intermedio è dato dai cosiddetti dati *panel*, ma non ce ne occupiamo qui.

Econometria. Notate che questo tipo di ragionamento è perfettamente appropriato nella maggior parte dei casi in cui i dati da noi osservati provengano da un esperimento controllato, del tipo di quelli che usano i medici o i biologi.

Il caso delle serie storiche, tuttavia, presenta una differenza concettuale di base che richiede una estensione dei concetti probabilistici da utilizzare come metafora dei dati. Questa differenza consiste nel fatto che il tempo ha una direzione, e quindi esiste la storia.

In un contesto di serie storiche, infatti, la naturale tendenza di molti fenomeni ad evolversi in modo più o meno regolare porta a pensare che il dato rilevato in un dato istante t sia più simile a quello rilevato all'istante $t-1$ piuttosto che in epoche distanti; si può dire, in un certo senso, che la serie storica che analizziamo ha "memoria di sé". Questa caratteristica è generalmente indicata col nome di **persistenza**[2], e differenzia profondamente i campioni di serie storiche da quelli *cross-section*, perché nei primi l'ordine dei dati ha un'importanza fondamentale, mentre nei secondi esso è del tutto irrilevante.

Lo strumento che utilizziamo per far fronte all'esigenza di trovare una metafora probabilistica per le serie storiche osservate è il **processo stocastico**. Una definizione di processo stocastico non rigorosa, ma intuitiva e, per le nostre esigenze, sostanzialmente corretta può essere la seguente: *un processo stocastico è una sequenza infinitamente lunga di variabili casuali* o, se preferite, un vettore aleatorio di dimensione infinita. Un campione di T osservazioni consecutive nel tempo non viene quindi pensato tanto come una realizzazione di T variabili casuali distinte, quanto piuttosto come parte di *un'unica* realizzazione di un processo stocastico, la cui memoria è data dal grado di connessione fra le variabili casuali che lo compongono.

1.2 Caratteristiche dei processi stocastici

La definizione appena data (che nasconde astutamente una serie di complicazioni tecniche) rende ovvie una serie di proprietà dei processi stocastici piuttosto importanti per il seguito: dato un processo stocastico il cui t-esimo elemento[3] indichiamo con x_t,

- è possibile (concettualmente) definire una funzione di densità per il processo $f(\ldots, x_{t-1}, x_t, x_{t+1}, \ldots)$;
- è possibile marginalizzare tale funzione di densità per ogni sottoinsieme delle sue componenti; da questo consegue che sono definite le funzioni di densità marginali per ognuna delle x_t, ma anche per ogni coppia di elementi (x_t, x_{t+1}) e così via; il fatto poi che le osservazioni non siano indipendenti fra loro fa sì che la densità del campione non si può più rappresentare come una semplice produttoria delle marginali;

[2] In certi contesti, gli economisti amano anche dire **istèresi** (o isterèsi) per indicare più o meno la stessa cosa. Un caso tipico è quando si parla di disoccupazione.

[3] Ad essere pignoli, dovremmo utilizzare due notazioni diverse per il *processo stocastico* di cui stiamo parlando, e per un suo generico elemento. Se quest'ultimo viene indicato con x_t, il processo a cui appartiene dovrebbe essere scritto $\{x_t\}_{-\infty}^{+\infty}$. Considero superflua questa raffinatezza, e userò la stessa notazione sia per un processo che per il suo t-esimo elemento; non dovrebbero sorgere confusioni.

- se le funzioni di densità marginali hanno momenti, è possibile dire, ad esempio, che $E(x_t) = \mu_t$, $V(x_t) = \sigma_t^2$, $Cov(x_t, x_{t-k}) = \gamma_{k,t}$ e così via;
- allo stesso modo, è possibile definire funzioni di densità (coi relativi momenti) condizionali.

Le proprietà appena descritte fanno riferimento ai processi stocastici come strutture probabilistiche. Quando però vogliamo utilizzare queste strutture come base per procedure inferenziali, si aprono due problemi:

1. Se quella che osservo (peraltro non nella sua interezza) è una sola realizzazione delle molte possibili, la possibilità *logica* di fare inferenza sul processo non può essere data per scontata; infatti, non c'è modo di dire quali caratteristiche della serie osservata sono specifiche di *quella* realizzazione, e quali invece si ripresenterebbero anche osservandone altre.

2. Se anche fosse possibile usare una sola realizzazione per fare inferenza sulle caratteristiche del processo, è necessario che esso sia stabile nel tempo, cioè che i suoi connotati probabilistici permangano invariati, per lo meno all'interno del mio intervallo di osservazione.

Queste due questioni conducono alla definizione di due proprietà che i processi stocastici possono avere o non avere:

Stazionarietà Si parla di processo stocastico stazionario in due sensi: **stazionarietà forte** (anche detta **stretta**) e **stazionarietà debole**.

Per definire la stazionarietà forte, prendiamo in esame un sottoinsieme qualunque delle variabili casuali che compongono il processo; queste non devono necessariamente essere consecutive, ma per aiutare l'intuizione, facciamo finta che lo siano. Consideriamo perciò una 'finestra' aperta sul processo di ampiezza k, ossia un sottoinsieme del tipo $W_t^k = (x_t, \ldots, x_{t+k-1})$. Questa è naturalmente una variabile casuale a k dimensioni, con una sua funzione di densità che, in generale, può dipendere da t. Se però ciò non accade, allora la distribuzione di W_t^k è uguale a quella di W_{t+1}^k, W_{t+2}^k e così via. Siamo in presenza di stazionarietà forte quando questa invarianza vale per qualsiasi k. In altri termini, quando un processo è stazionario in senso forte le caratteristiche distribuzionali di tutte le marginali rimangono costanti al passare del tempo.

La stazionarietà debole, invece, riguarda solo finestre di ampiezza 2: si ha stazionarietà debole se tutte le variabili casuali doppie $W_t^2 = (x_t, x_{t+1})$, hanno momenti primi e secondi costanti nel tempo[4]; da questo discende che esistono anche tutti i momenti secondi incrociati $E(x_t \cdot x_{t+k})$, con k qualunque, e anch'essi non dipendono da t (anche se possono dipendere da k).

[4] È per questo motivo che la stazionarietà debole viene anche definita **stazionarietà in covarianza**.

A dispetto dei nomi, una definizione non implica l'altra; ad esempio, un processo può essere stazionario in senso forte ma non possedere momenti;[5] viceversa, la costanza nel tempo dei momenti non implica che le varie marginali abbiano la stessa distribuzione. In un caso, tuttavia, le due definizioni coincidono: questo caso — che è particolarmente importante per le applicazioni pratiche — è quello in cui il processo è **gaussiano**, ossia quando la distribuzione congiunta di un qualunque sottoinsieme di elementi del processo è una normale multivariata. Se un processo è gaussiano, stabilire che è stazionario in senso debole equivale a stabilire la stazionarietà stretta. Data la pervasività dei processi gaussiani nelle applicazioni ai dati, da un punto di vista operativo si adotta generalmente la definizione di stazionarietà debole, e quando si parla di stazionarietà senza aggettivi, è appunto a questa che ci si riferisce.

Ergodicità L'ergodicità è una condizione che limita la memoria del processo: un processo non ergodico è un processo che ha caratteristiche di persistenza così accentuate da far sì che un segmento del processo, per quanto lungo, sia insufficiente a dire alcunché sulle sue caratteristiche distributive. In un processo ergodico, al contrario, la memoria del processo è debole su lunghi orizzonti e all'aumentare dell'ampiezza del campione aumenta in modo significativo anche l'informazione in nostro possesso.

Le condizioni sotto le quali un processo stocastico stazionario è ergodico sono troppo complesse per essere descritte qui; per farmi capire, vi sottoporrò ad un'overdose di virgolette: euristicamente, si può dire che un processo è ergodico se eventi "molto" lontani fra loro possono essere considerati "virtualmente" indipendenti; osservando il processo per un lasso di tempo "abbastanza" lungo, è possibile osservare "quasi tutte" le sottosequenze che il processo è in grado di generare. In altri termini, si può dire che, in un sistema ergodico, se qualcosa *può* succedere allora prima o poi *deve* succedere. Il fatto che eventi lontani fra loro nel tempo possano essere considerati indipendenti da un punto di vista pratico è poi spesso sintetizzato nella seguente proprietà dei processi ergodici (che a volte viene usata come *definizione* di processo ergodico):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathrm{Cov}(x_t, x_{t-k}) = 0.$$

Di conseguenza, se un processo è ergodico, è possibile (almeno in linea di principio) usare le informazioni contenute nel suo svolgimento nel tempo per inferirne le caratteristiche. Esiste un teorema (detto appunto 'teorema ergodico') che dice che, se un processo è ergodico, l'osservazione di una sua realizzazione "abbastanza" lunga è equivalente, ai fini inferenziali, all'osservazione di un gran numero di realizzazioni.

Se, ad esempio, un processo ergodico x_t ha valore atteso μ, allora la sua media aritmetica *nel tempo* è uno stimatore consistente di μ (in formule,

[5] Esempio di processo stazionario in senso forte ma non debole: consideriamo una sequenza di variabili casuali $y_t = 1/x_t$, dove le x_t sono normali standard indipendenti. La sequenza delle y_t è una sequenza di variabili casuali identiche, indipendenti e senza momenti.

$T^{-1}\sum_{t=1}^{T} x_t \xrightarrow{\text{p}} \mu$), e quindi μ può essere stimato in modo consistente come se disponessimo di molte realizzazioni del processo anziché di una sola.

In linea generale, si può dire che l'inferenza è possibile solo se il processo stocastico che si sta studiando è stazionario ed ergodico. Va detto per altro che, se esistono dei metodi per sottoporre a test l'ipotesi di non stazionarietà (almeno in certi contesti, che esamineremo nel prosieguo), l'ipotesi di ergodicità non è testabile se si dispone di una sola realizzazione del processo, quand'anche fosse di ampiezza infinita.

1.3 Momenti

Nel caso di processi stocastici stazionari, avremo dunque che ogni elemento del processo x_t avrà un valore atteso finito e costante μ e una varianza finita e costante σ^2. Inoltre, risultano definite tutte le covarianze fra elementi diversi del processo, che saranno pari a

$$\gamma_k = E\left[(x_t - \mu)(x_{t-k} - \mu)\right] \tag{1.1}$$

e che sono note come **autocovarianze**. Si ricordi che la stazionarietà garantisce che queste quantità non sono funzioni di t; esse sono tuttavia funzioni di k, ed anzi si parla di funzione di autocovarianza, intendendo una funzione di k tale per cui $\gamma(k) = \gamma_k$. Va da sé che l'autocovarianza di ordine 0 non è che la varianza. Inoltre, la definizione è tale per cui $\gamma_k = \gamma_{-k}$, ossia la seguente espressione

$$E\left[(x_t - \mu)(x_{t-k} - \mu)\right] = E\left[(x_t - \mu)(x_{t+k} - \mu)\right]$$

è vera.

A questo punto, diventa banale definire le **autocorrelazioni**, che sono date da

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma^2} \tag{1.2}$$

Ovviamente, $\rho_0 = 1$. Queste quantità, se diverse da 0, descrivono la memoria del processo, e sono appunto l'elemento che rende i processi stocastici lo strumento teorico adatto a rappresentare serie storiche caratterizzate da persistenza. Infatti, se $\gamma_1 \neq 0$, allora si ha che

$$f(x_t|x_{t-1}) \neq f(x_t)$$

e di conseguenza

$$E(x_t|x_{t-1}) \neq E(x_t), \tag{1.3}$$

che si può tradurre: se x_{t-1} è noto, il valore atteso di x_t non è lo stesso che ci attenderemmo se x_{t-1} fosse incognito. Potremmo estendere l'insieme di variabili casuali su cui effettuiamo il condizionamento anche a x_{t-2}, x_{t-3} eccetera. Questo insieme di variabili casuali prende a volte il nome di **set informativo** al tempo $t-1$, e viene indicato con $\Im_{t-1}$.

A dire la verità, la definizione precisa di set informativo è un po' complessa: dovremmo parlare di σ-algebre ed essere più rigorosi su cosa si intende per condizionamento. L'argomento è in effetti appassionante, ma davvero non è questa la sede. Non ci si rimette molto, però, a considerare come set informativo un insieme di variabili casuali rispetto alle quali è possibile effettuare l'operazione di condizionamento. In un contesto di serie storiche, è abbastanza naturale supporre che il passato sia noto; di conseguenza, ha senso parlare di condizionamento di una variabile casuale al tempo t rispetto ai propri valori passati, perché se x_t è nota, allora lo sono anche x_{t-1}, x_{t-2} e così via. Naturalmente, nessuno esclude che nel set informativo al tempo t trovino posto anche variabili diverse da quella che stiamo condizionando. Anzi, in certi contesti (come ad esempio nella teoria delle aspettative razionali) l'idea di set informativo al tempo t viene usata come sinonimo di *tutta* l'informazione disponibile sul mondo al tempo t.

Se si osserva una realizzazione di ampiezza T di un processo stocastico x_t, si possono definire gli equivalenti campionari dei momenti teorici:

$$\begin{array}{rl} \text{media campionaria} & \hat{\mu} = T^{-1}\sum_{t=1}^{T} x_t \\ \text{varianza campionaria} & \hat{\sigma}^2 = T^{-1}\sum_{t=1}^{T}(x_t - \hat{\mu})^2 \\ \text{autocovarianza campionaria} & \hat{\gamma}_k = T^{-1}\sum_{t=k}^{T}(x_t - \hat{\mu})(x_{t-k} - \hat{\mu}) \end{array}$$

Se il processo è stazionario ed ergodico, si può dimostrare che queste quantità sono stimatori consistenti dei momenti del processo[6]. Tuttavia, queste statistiche hanno una loro utilità anche come statistiche descrittive. A parte la media campionaria, la cui interpretazione dò per scontata, l'analisi della funzione di autocorrelazione consente in molti casi di dare interessanti valutazioni qualitative sulle caratteristiche di persistenza della serie che stiamo considerando. Nella prossima sottosezione faremo qualche esempio.

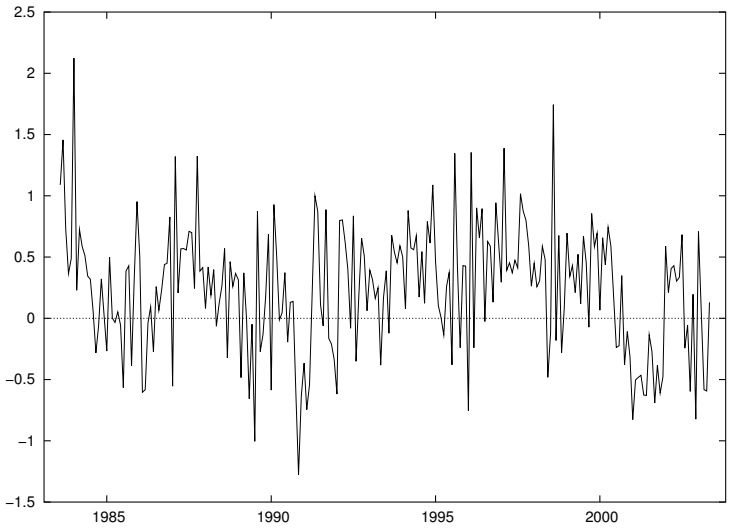
1.4 Qualche esempio

In che senso, allora, possiamo utilizzare i processi stocastici come idealizzazione del processo generatore dei dati? La situazione in cui ci si trova in pratica è press'a poco questa: abbiamo una serie storica; posto che la si possa considerare una realizzazione di un processo stocastico stazionario, ciò che vogliamo fare è trovare il processo che "meglio" rappresenta la serie. Più precisamente, ci chiederemo quale tipo di processo presenta realizzazioni che più somigliano alla serie osservata.

Consideriamo ad esempio la serie storica rappresentata in figura 1.1, che riporta i valori mensili, dall'agosto 1983 al luglio 2003, della variazione percentuale dell'indice della produzione industriale sul mese precedente per gli Stati Uniti. Come si vede, la serie oscilla in modo abbastanza regolare intorno ad un valore centrale, situato grosso modo fra 0 e 0.5%. In effetti, la media aritmetica delle osservazioni è pari a 0.253%. Se fossimo autorizzati a pensare che il processo che ha generato questi dati fosse stazionario ed ergodico, potremmo dire che tale valore è una stima del valore atteso non condizionale del processo.

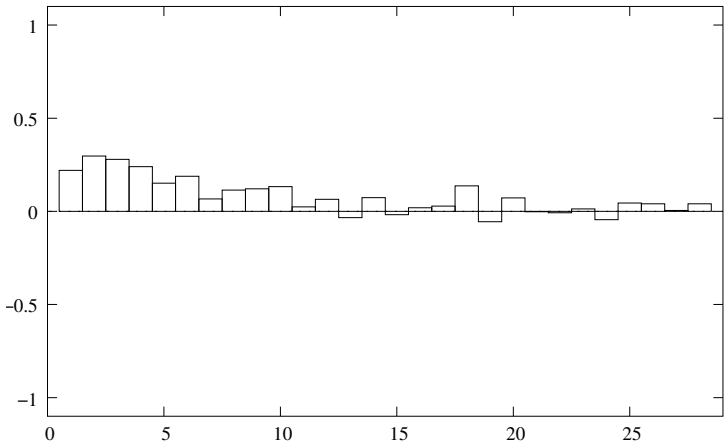
[6] Il lettore accorto noterà l'assenza della 'correzione per gradi di libertà': al denominatore della varianza campionaria, ed esempio, c'è T anziché $T-1$. Da un punto di vista asintotico le due

Figura 1.1: Variazioni mensili della produzione industriale USA



Ma questo processo (posto che esista) è stazionario? E se sì, è anche ergodico? E più in generale, quali sono le sue caratteristiche di persistenza? Guardando il grafico è difficile dare una risposta, perlomeno se non si ha una certa pratica. Un aiuto ci viene dall'analisi delle autocorrelazioni campionarie, riportate nella figura 1.2.

Figura 1.2: Variazioni produzione industriale USA – correlogramma



Una figura come la 1.2 si chiama **correlogramma**; il correlogramma è semplicemente un istogramma in cui ogni barretta riporta il valore dell'autocorrelazione ρ_k in funzione di k, che è in ascissa. In altre parole, il correlogramma si legge così: se indichiamo con y_t il dato al tempo t, la correlazione fra y_t e

formulazioni sono evidentemente equivalenti. Quel che succede in campioni finiti è di solito considerato irrilevante o troppo complicato per essere studiato.

y_{t-1} è il 22%, quella fra y_t e y_{t-2} è il 29.7% eccetera. Volendo fare un discorso propriamente statistico-inferenziale, dovremmo chiederci se queste statistiche sono stimatori di grandezze (le autocorrelazioni del processo) significativamente diverse da 0, ma per il momento possiamo accontentarci di considerarle statistiche descrittive, il cui significato è chiaro: osservazioni consecutive sono fortemente correlate, *ergo* difficilmente possiamo considerarle indipendenti, *ergo* c'è traccia di una certa persistenza. Allo stesso modo, questa persistenza sembra affievolirsi con l'andare del tempo: si direbbe che, man mano che la distanza fra le osservazioni aumenta, il valore assoluto della loro correlazione (che possiamo, a questo stadio, considerare un indicatore di persistenza) tende a diminuire: a 24 mesi di distanza la correlazione è decisamente più contenuta (-4.5%). Mettendo tutto insieme, si potrebbe dire che da un punto di vista qualitativo questo è quello che ci aspettiamo di vedere in una realizzazione di un processo stazionario ed ergodico: una persistenza che influenza sostanzialmente la serie nel breve periodo, ma che tutto sommato rimane un fenomeno "locale".

A questo punto, ci si potrebbe chiedere se la serie storica che stiamo osservando possa essere modellata statisticamente studiando la sua media condizionale così come si fa in un modello di regressione lineare. Se infatti in un modello lineare l'equazione $y_t = x_t'\beta + \epsilon_t$ scinde la variabile esplicativa in una media condizionale più un disturbo, nessuno ci vieta di rendere la media condizionale una funzione del set informativo $\Im_{t-1}$, e di stimare con gli OLS un modello come il seguente:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \epsilon_t. \tag{1.4}$$

Se lo facessimo, utilizzando ad esempio come set di condizionamento i valori fino a quattro mesi prima, otterremmo i risultati mostrati nella tabella 1.1.

Tabella 1.1: Stima OLS dell'equazione (1.4)

Coefficiente	Stima	Errore std.	Statistica t	p-value
β_0	0.086	0.038	2.2835	0.0233
β_1	0.069	0.066	1.0453	0.2970
β_2	0.207	0.065	3.1890	0.0016
β_3	0.192	0.064	2.9870	0.0031
β_4	0.118	0.065	1.8090	0.0718

Media della variable dipendente	0.224	Dev. std. della var. dipendente	0.511
Somma dei quadrati dei residui	51.296	Errore std dei residui ($\hat{\sigma}$)	0.473
R^2	0.156	$F(4,656)$	10.599

Se non vogliamo considerare questa stima come una semplice statistica descrittiva, allora le sue proprietà devono necessariamente essere studiate all'interno di un quadro di riferimento inferenziale appropriato. È proprio per questo che abbiamo bisogno di studiare i processi stocastici: per dare un significato probabilistico, se possibile, a statistiche come quelle che abbiamo appena visto. Nei capitoli successivi farò vedere come e perché la stima appena fatta ha effettivamente senso, e come vada interpretata.

Le cose, però, non sempre vanno così lisce: la figura 1.3 riporta la serie storica della variazione percentuale annua dell'indice dei prezzi al consumo, sempre per gli USA.

Figura 1.3: Inflazione USA

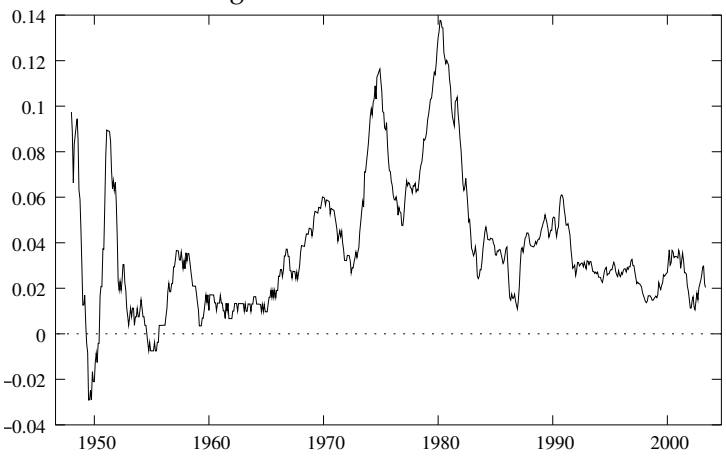
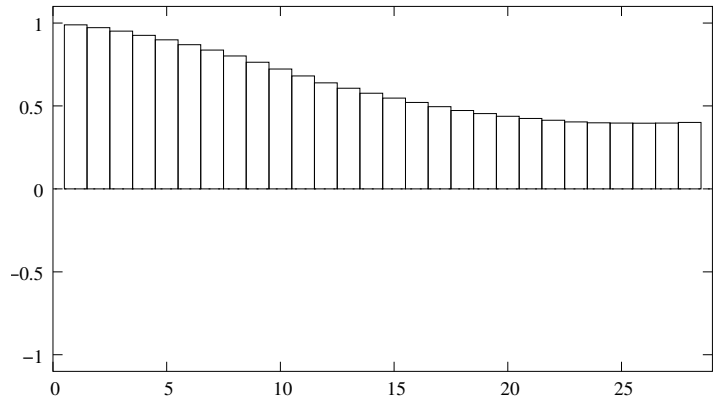


Figura 1.4: Inflazione USA – correlogramma

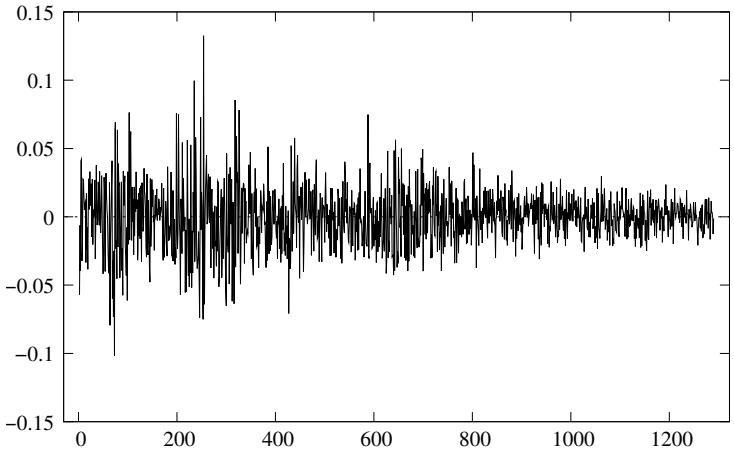


Siamo sicuri che una serie storica come questa possa essere generata da un processo stazionario? Come si vede, si alternano periodi (anche piuttosto lunghi) di inflazione alta e bassa. È lecito pensare che l'ipotetico processo che genera questa serie abbia una media costante, come richiesto per la stazionarietà? E per di più, diamo un'occhiata al correlogramma (figura 1.4): in questo caso, considerare la persistenza come un fenomeno di breve periodo è decisamente più temerario. L'autocorrelazione a 24 mesi è pari al 38.9%, e non dà mostra di scendere significativamente.

Serie storiche come questa, ad alta persistenza, sono estremamente comuni in economia ed in finanza; per essere analizzate, devono essere in qualche modo ricondotte a realizzazioni di processi stazionari. Questo, in molti casi, si può fare con strumenti appositi, che hanno dominato l'econometria delle

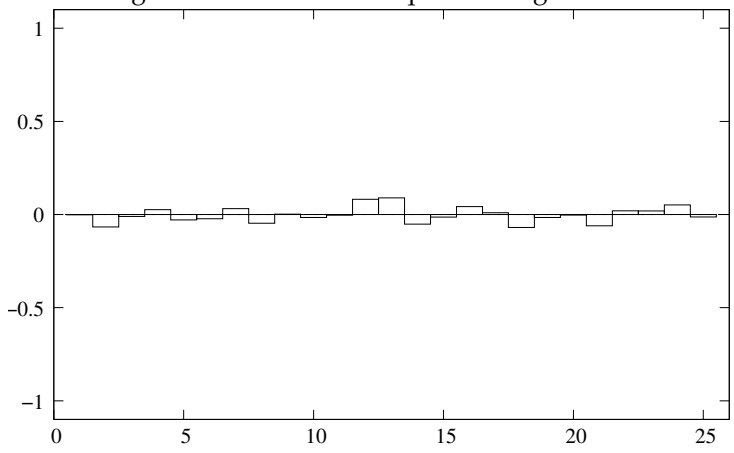
serie storiche negli ultimi due decenni del XX secolo. Siccome però sono un po' complessi, noi non li vedremo prima del capitolo 3. Portate pazienza.

Figura 1.5: Indice Nasdaq – rendimenti giornalieri



Chiudo questa carrellata di esempi con un caso opposto a quello precedente: la variazione percentuale (giornaliera) dell'indice Nasdaq dall'1/1/2000 al 28/2/2005, mostrato nella figura 1.5. L'aspetto della serie è — palesemente — molto diverso da quello delle serie mostrate prima: i dati fluttuano attorno ad un valore di poco superiore allo zero (la media aritmetica è -0.054 — in altri termini l'indice borsistico esaminato è sceso in media dello 0.054% al giorno negli ultimi 5 anni), senza che però siano visibili quelle onde lunghe che caratterizzavano le serie della produzione industriale o dell'inflazione. Questa impressione è confermata dal correlogramma (figura 1.6).

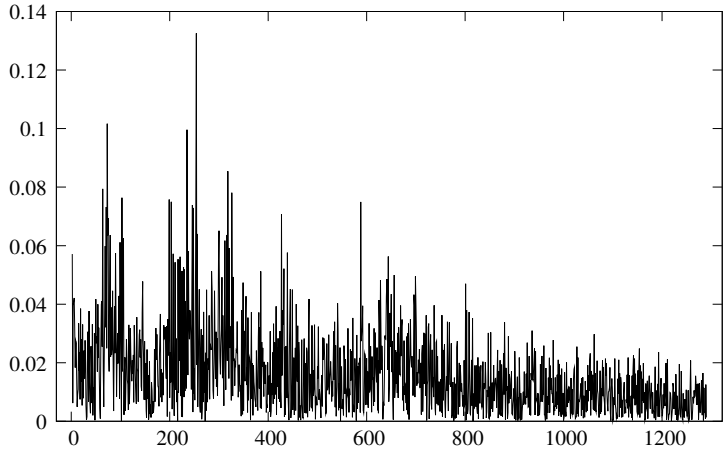
Figura 1.6: Indice Nasdaq – Correlogramma



Qui di persistenza se ne vede poca. E d'altronde è comprensibile: con buona pace dei fan dell'analisi tecnica, se ci fosse una regola "semplice" che

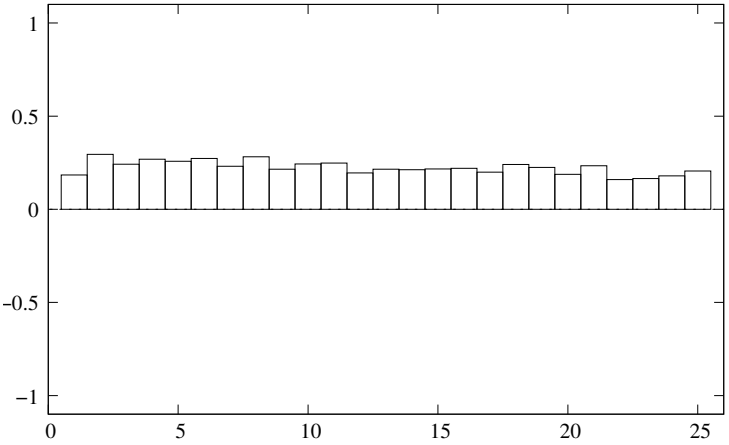
lega i rendimenti ai loro valori passati, qualunque cretino potrebbe mettersi a fare lo speculatore e trarne profitti illimitati[7].

Figura 1.7: Indice Nasdaq – rendimenti giornalieri in valore assoluto



Ma anche qui non mancano aspetti interessanti: infatti, l'andamento nel tempo della serie in figura è tipica di moltissime serie finanziarie. In particolare, è interessante considerare il fatto che l'andamento nel tempo dell'indice è caratterizzato da un'alternanza di periodi in cui la volatilità del mercato è più alta e da altri in cui le variazioni sono di entità più contenuta. La cosa si vede piuttosto bene considerando la serie storica dei valori assoluti dei rendimenti (figura 1.7).

Figura 1.8: Indice Nasdaq – Correlogramma dei valori assoluti



Qui, si vede bene, di persistenza ce n'è eccome. In questo caso, ciò che interessa è modellare statisticamente non tanto la persistenza della serie di per sé, ma piuttosto della sua volatilità.

[7]Chi è del mestiere noterà che sto volgarizzando la cosiddetta "legge dei mercati efficienti" con una disinvoltura che neanche Piero Angela si sognerebbe. Domando scusa.

Naturalmente, il concetto statistico nel quale si traduce la parola "volatilità" è la varianza (posto che i momenti secondi esistano). Come si vedrà in seguito, per analizzare serie di questo tipo si usano processi stocastici di natura particolare, in cui la persistenza eventualmente esistente nella serie si traduce nella dipendenza dal passato della varianza, anziché della media. In altre parole, le caratteristiche di persistenza di questi processi vengono sintetizzate nel fatto che

$$V(x_t|x_{t-1}) \neq V(x_t). \tag{1.5}$$

Si faccia il confronto con la (1.3): in questi processi, che si chiamano processi **condizionalmente eteroschedastici**, ciò che fa la differenza fra le distribuzioni marginali e quelle condizionali al set informativo $\Im_{t-1}$ è appunto la struttura dei momenti secondi, anziché dei momenti primi. Processi di questo tipo sono oramai di largo uso nella finanza empirica più avanzata.

Arrivati fin qui, il senso dell'operazione che ci accingiamo a compiere dovrebbe essere abbastanza chiaro. Nel capitolo seguente, faremo la conoscenza della classe di processi stocastici che fa da fondamento a tutta l'econometria delle serie storiche, e cioè i processi ARMA.

Capitolo 2

I processi ARMA

I processi ARMA costituiscono la famiglia di processi stocastici di gran lunga più utilizzati in econometria. Questa scelta ha ragioni teoriche e ragioni pratiche, che saranno illustrate nel seguito. Prima di analizzare le caratteristiche principali di tali processi, tuttavia, sono necessarie alcune definizioni di base, che formano l'oggetto dei prossimi paragrafi.

2.1 L'operatore ritardo

Tanto i processi stocastici che le serie storiche sono, in buona sostanza, sequenze di numeri. Capiterà molto spesso di dover manipolare tali sequenze, e lo faremo per mezzo di appositi operatori. L'**operatore ritardo** viene generalmente indicato con la lettera L nella letteratura econometrica (gli statistici preferiscono la B); è un operatore che si applica a sequenze di oggetti piuutosto generali, fra cui rientrano sia sequenze di variabili casuali (e cioè i processi stocastici) che sequenze di numeri (e cioè le loro traiettorie); tale operatore trasforma una sequenza x_t in un altra sequenza che ha la curiosa caratteristica di avere gli stessi valori di x_t, ma sfalsati di un periodo[1]. Se applicato ad una grandezza costante nel tempo, la lascia invariata. In formule,

$$Lx_t = x_{t-1}$$

L'applicazione ripetuta n volte di L viene indicata con la scrittura L^n, e quindi si ha $L^n x_t = x_{t-n}$. Per convenzione si pone $L^0 = 1$. L'operatore L è un operatore lineare, nel senso che, se a e b sono costanti, si ha $L(ax_t + b) = aLx_t + b = ax_{t-1} + b$.

La caratteristica più divertente dell'operatore L è che le sue proprietà appena enunciate permettono, in molte circostanze, di manipolarlo algebricamente come se fosse un numero. Questo avviene soprattutto quando si considerano *polinomi* nell'operatore L. Facciamo un paio di esempi semplici.

[1] In certi contesti, si utilizza anche il cosiddetto **operatore anticipo**, usualmente indicato con la lettera F e definito come l'inverso dell'operatore ritardo ($Fx_t = x_{t+1}$). Noi non lo useremo mai, ma è bello sapere che c'è.

Esempio 2.1.1 *Una squadra di calcio ha in classifica tanti punti quanti ne aveva alla giornata precedente, più quelli che ha guadagnato nell'ultimo turno. Chiamando rispettivamente queste sequenze c_t e u_t, si avrà*

$$c_t = c_{t-1} + u_t$$

La stessa cosa si sarebbe potuta scrivere adoperando l'operatore ritardo:

$$c_t = Lc_t + u_t \rightarrow c_t - Lc_t = (1-L)c_t = \Delta c_t = u_t$$

L'operatore Δ, che dovrebbe essere una vecchia conoscenza, è definito come $(1-L)$, ossia un polinomio di primo grado in L. L'espressione precedente non dice altro che la variazione dei punti in classifica è data dai punti guadagnati in ogni giornata.

Esempio 2.1.2 *Chiamiamo q_t il saldo demografico trimestrale per il comune di Rocca Cannuccia. È evidente che il saldo demografico annuale (cioè le nascite degli ultimi 12 mesi meno le morti nello stesso periodo) sono date da*

$$a_t = q_t + q_{t-1} + q_{t-2} + q_{t-3} = (1 + L + L^2 + L^3)q_t$$

Poiché $(1 + L + L^2 + L^3)(1 - L) = (1 - L^4)$ (moltiplicare per credere), "moltiplicando" l'espressione precedente[2] per $(1 - L)$ si ha

$$\Delta a_t = (1 - L^4)q_t = q_t - q_{t-4}$$

la variazione del saldo demografico annuale tra un trimestre ed il successivo non è che la differenza fra il saldo dell'ultimo trimestre e il corrispondente trimestre dell'anno precedente.

Le manipolazioni possono essere anche più complesse; in particolare ci sono due risultati di routine: il primo è che

$$\sum_{i=0}^{n} a^i = \frac{1 - a^{n+1}}{1 - a}$$

per $a \neq 1$. Se poi $|a| < 1$, si ha che $a^n \rightarrow 0$ e quindi $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$. Ponendo $a = \alpha L$, si può dire che, per $|\alpha| < 1$, i due operatori $(1 - \alpha L)$ e $(1 + \alpha L + \alpha^2 L^2 + \cdots)$ sono uno l'inverso dell'altro. In pratica, se $|a| < 1$, vale

$$(1 - \alpha L)(1 + \alpha L + \alpha^2 L^2 + \cdots) = 1,$$

da cui l'espressione (che incontreremo spesso)

$$(1 - \alpha L)^{-1} = \sum_{i=0}^{\infty} \alpha^i L^i,$$

che spesso si abbrevia anche in

$$\sum_{i=0}^{\infty} \alpha^i L^i = \frac{1}{1 - \alpha L}.$$

[2] Ad essere precisi, si dovrebbe dire: 'applicando all'espressione precedente l'operatore $(1 - L)$'.

Il secondo risultato riguarda i polinomi. Prendiamo un polinomio di n-esimo grado, e chiamiamolo $P(x)$. Per definizione, si ha

$$P(x) = \sum_{j=0}^{n} p_j x^j$$

Se $P(0) = p_0 = 1$, allora è possibile esprimere il polinomio di n-esimo grado come il prodotto di n polinomi di primo grado:

$$P(x) = \prod_{j=1}^{n} (1 - \lambda_j x) \tag{2.1}$$

i coefficienti λ_j non sono altro che i reciproci delle radici di $P(x)$, ossia quei valori per cui $P(\frac{1}{\lambda_j}) = 0$. Nessuno assicura che queste radici siano reali (per $n > 1$ possono anche essere numeri complessi), ma dal punto di vista teorico questo non ha alcuna rilevanza. Questo risultato è importante perché, unito al precedente, permette di invertire polinomi di qualunque grado.

Un altro trucchetto che a volte si usa è quello di valutare un polinomio $P(L)$ in $L = 1$. Evidentemente, l'espressione $P(1)$ è uguale a

$$P(1) = \sum_{j=0}^{n} p_j 1^j = \sum_{j=0}^{n} p_j$$

e quindi è semplicemente uguale ad un numero, dato dalla somma dei coefficienti del polinomio. Questo torna comodo quando si applica un polinomio ad una costante, visto che

$$P(L)\mu = \sum_{j=0}^{n} p_j \mu = \mu \sum_{j=0}^{n} p_j = P(1)\mu.$$

Vediamo un altro esempio:

Esempio 2.1.3 (Il moltiplicatore keynesiano) *Supponiamo che*

$$\begin{aligned} Y_t &= C_t + I_t \\ C_t &= \alpha Y_{t-1} \end{aligned}$$

Dove α è la propensione marginale al consumo, compresa fra 0 e 1. Combinando le due equazioni si ha

$$Y_t = \alpha Y_{t-1} + I_t \rightarrow (1 - \alpha L) Y_t = I_t;$$

in questo modello, quindi, applicando alla sequenza Y_t (la serie storica del reddito) il polinomio di primo grado $A(L) = (1 - \alpha L)$ si ottiene la serie storica degli investimenti, semplicemente perché $I_t = Y_t - C_t = Y_t - \alpha Y_{t-1}$.

Un risultato più interessante si ha invertendo l'operatore $A(L) = (1 - \alpha L)$:

$$Y_t = (1 + \alpha L + \alpha^2 L^2 + \cdots) I_t = \sum_{i=0}^{\infty} \alpha^i I_{t-i} :$$

la domanda aggregata al tempo t può essere vista come una somma ponderata dei valori presenti e passati dell'investimento. Se poi il flusso di investimenti è costante nel tempo, allora $I_t = \bar{I}$ può essere tirato fuori dalla sommatoria, e si ottiene il risultato standard da libro di macro elementare:

$$Y_t = \bar{I} \sum_{i=0}^{\infty} \alpha^i = \frac{\bar{I}}{1-\alpha}.$$

In questo ultimo caso si sarebbe anche potuto scrivere

$$A(1)Y_t = \bar{I} \Longrightarrow Y_t = \frac{\bar{I}}{1-\alpha}.$$

Il fatto che spesso si può maneggiare l'operatore L come se fosse un numero non vuol dire che lo si possa far sempre: bisogna sempre ricordare che Lx_t non è 'L per x_t', ma 'L applicato a x_t'. L'esempio seguente dovrebbe servire a mettere in guardia.

Esempio 2.1.4 *Date due sequenze x_t e y_t, definiamo una terza sequenza $z_t = x_t y_t$. È del tutto chiaro che $z_{t-1} = x_{t-1}y_{t-1}$. Tuttavia, potremmo essere tentati di fare il seguente ragionamento:*

$$z_{t-1} = x_{t-1}y_{t-1} = Lx_tLy_t = L^2x_ty_t = L^2z_t = z_{t-2}$$

che è evidentemente assurdo.

Manipolazioni come quelle viste in questo paragrafo possono essere effettuate su ogni tipo di sequenza, e quindi anche su processi stocastici. I paragrafi che seguono esaminano appunto che tipo di processi stocastici otteniamo con questi procedimenti.

2.2 Processi *white noise*

Il ***white noise*** è il processo stocastico più semplice che si può immaginare[3]: infatti, è un processo che possiede momenti (almeno) fino al secondo ordine; essi sono costanti nel tempo (quindi il processo è stazionario), ma non dànno al processo alcuna memoria di sé.

La stessa cosa si può dire in modo più formalizzato come segue: un processo *white noise*, il cui elemento t-esimo indicheremo con ϵ_t, presenta queste caratteristiche:

$$E(\epsilon_t) = 0 \tag{2.2}$$

$$E(\epsilon_t^2) = V(\epsilon_t) = \sigma^2 \tag{2.3}$$

$$\gamma_k = 0 \quad \text{per } |k| > 0 \tag{2.4}$$

Un *white noise* è quindi, in sostanza, un processo composto di un numero infinito di variabili casuali a media zero e varianza costante; queste variabili

[3]Il motivo per cui questo processo porta l'immaginifico nome di **rumore bianco** presenterebbe un certo interesse, ma gli strumenti analitici di cui si discute in questa dispensa non ci consentono di sviluppare questo punto. Pazienza.

casuali, inoltre, sono tutte incorrelate l'una all'altra. A rigore, questo non significa che esse siano indipendenti. Se però si parla di *white noise* gaussiano, ossia di un *white noise* in cui la distribuzione congiunta di tutte le coppie $(\epsilon_t, \epsilon_{t+k})$ sia una normale bivariata, allora sì. Ci sono due cose che vale la pena di far notare:

- Nel caso di normalità, una realizzazione di ampiezza N di un *white noise* può anche essere considerata del tutto legittimamente una realizzazione di N variabili casuali indipendenti ed identiche. In questo senso, un campione *cross-section* può essere visto come un caso particolare.

- Non c'è sostanziale differenza fra le condizioni che definiscono un *white noise* e le cosiddette "ipotesi classiche" sul termine di disturbo nel modello OLS, eccezion fatta per l'incorrelazione fra regressori e disturbi; non si sbaglierebbe riassumendo le ipotesi classiche nel modello OLS nella frase 'il termine di disturbo è un *white noise* incorrelato coi regressori'.

Un processo *white noise*, quindi, è un processo stocastico che non esibisce persistenza. In quanto tale, si potrebbe pensare che sia inadeguato a raggiungere lo scopo che ci eravamo prefissi nella premessa, cioè trovare una struttura probabilistica che possa servire da metafora per campioni di serie storiche che, invece, la persistenza ce l'hanno. Il passo in avanti decisivo, che vediamo nel prossimo paragrafo, sta nel considerare cosa succede applicando un polinomio nell'operatore ritardo ad un *white noise*.

Se volessi essere preciso, dovrei fare una distinzione fra diversi tipi di processi stocastici "senza memoria". A rigore, infatti, l'unico tipo di processo senza traccia di persistenza è quello composto da variabili casuali indipendenti. Spesso però si preferisce trattare processi che non siano così vincolanti per quanto riguarda le loro proprietà: ad esempio, la cosiddetta *differenza di martingala*, che è un concetto impiegato molto comunemente sia in statistica (soprattutto in teoria asintotica) che in economia (teoria delle aspettative razionali). In una differenza di martingala, la distribuzione è lasciata non specificata; ciò che caratterizza questo tipo di sequenza è la proprietà $E(x_t|\Im_{t-1}) = 0$. In questo contesto, l'unica cosa che interessa è il valor medio condizionale del processo, che non deve dipendere in alcun modo dal passato.

Un *white noise*, invece, è un concetto ancora diverso: la proprietà di incorrelazione fra elementi diversi assicura soltanto che la media condizionale non sia una funzione *lineare* del passato. Dimostrazione zippata:

$$\begin{aligned} E(x_t|\Im_{t-1}) = bx_{t-1} &\implies \\ E(x_t x_{t-1}|\Im_{t-1}) = bx_{t-1}^2 &\implies \\ E(x_t x_{t-1}) = E\left[E(x_t x_{t-1}|\Im_{t-1})\right] &= \\ = bE(x_{t-1}^2) \neq 0 & \end{aligned}$$

(ringraziamo la legge dei valori attesi iterati per la gentile collaborazione). Nulla esclude, però, che la media condizionale possa essere una funzione non lineare diversa da zero. In effetti, si possono costruire esempi di processi *white noise* che non sono differenze di martingala. Peraltro, non tutte le differenze di martingala sono dei *white noise*: la definizione di *white noise* comporta infatti condizioni ben precise sui momenti secondi, che in una differenza di martingala possono anche non esistere.
In pratica, però, questi concetti possono essere sovrapposti in modo abbastanza indolore: un *white noise* gaussiano, per esempio, è una sequenza di variabili casuali indipendenti a media 0, per cui è anche una differenza di martingala. Nel prosieguo, sarò molto elastico e considererò un *white noise* come processo senza memoria *tout court*.

2.3 Processi MA

Un **processo MA**, o processo **a media mobile** (MA sta appunto per Moving Average), è una sequenza di variabili casuali che può essere scritta nella forma

$$y_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i} = C(L)\epsilon_t$$

dove $C(L)$ è un polinomio di ordine q nell'operatore ritardo e ϵ_t è un *white noise*. Generalmente, e senza perdita di generalità, si pone $C(0) = \theta_0 = 1$. Se $C(L)$ è un polinomio di grado q, si dice anche che y_t è un processo MA(q), che si legge 'processo MA di ordine q'. Esaminiamo i suoi momenti: per quanto riguarda il momento primo, si ha

$$E(y_t) = E\left[\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right] = \sum_{i=0}^{q} \theta_i E(\epsilon_{t-i}) = 0$$

E quindi un processo MA ha media 0. A prima vista, si potrebbe pensare che questa caratteristica limiti fortemente l'applicabilità di processi MA a situazioni reali, visto che, in genere, non è detto che le serie storiche osservate oscillino intorno al valore 0. Tuttavia, la limitazione è più apparente che reale, visto che per ogni processo x_t per cui $E(x_t) = \mu_t$ si può sempre definire un nuovo processo $y_t = x_t - \mu_t$ a media nulla[4]. Se y_t è stazionario in covarianza, allora basta studiare y_t e poi ri-aggiungere la media per avere x_t.

Per quanto riguarda la varianza, il fatto che il momento primo sia nullo ci consente di scriverla come il momento secondo, ossia

$$V(y_t) = E(y_t^2) = E\left[\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)^2\right]$$

Sviluppando il quadrato[5], possiamo scomporre la somma in due parti distinte:

$$\left(\sum_{i=0}^{q} \theta_i \epsilon_{t-i}\right)^2 = \sum_{i=0}^{q} \theta_i^2 \epsilon_{t-i}^2 + \sum_{i=0}^{q} \sum_{j \neq i} \theta_i \theta_j \epsilon_{t-i} \epsilon_{t-j}$$

Dovrebbe essere ovvio, dalla proprietà del *white noise*, che il valore atteso della seconda sommatoria è nullo, cosicché

$$E(y_t^2) = E\left[\sum_{i=0}^{q} \theta_i^2 \epsilon_{t-i}^2\right] = \sum_{i=0}^{q} \theta_i^2 E(\epsilon_{t-i}^2) = \sum_{i=0}^{q} \theta_i^2 \sigma^2 = \sigma^2 \sum_{i=0}^{q} \theta_i^2 \tag{2.5}$$

che ha valore finito se $\sum_{i=0}^{q} \theta_i^2 < \infty$, cosa sempre vera se q è finito.

[4] Faccio notare *en passant* che in questo semplice esempio il processo x_t non è stazionario, secondo la definizione che ci siamo dati, ma il processo y_t sì.

[5] Attenzione: riprendo brevemente l'argomento di qualche pagina fa per far notare che $[C(L)\epsilon_t]^2$ è *diverso* da $C(L)^2\epsilon_t^2$. Pensate al semplice caso $C(L) = L$ e ve ne convincerete immediatamente.

Infine, con un ragionamento del tutto analogo perveniamo al calcolo delle autocovarianze: l'autocovarianza di ordine k è data da

$$E(y_t y_{t+k}) = E\left[\left(\sum_{i=0}^{q}\theta_i\epsilon_{t-i}\right)\left(\sum_{j=0}^{q}\theta_j\epsilon_{t-j+k}\right)\right] = \sum_{i=0}^{q}\theta_i\left(\sum_{j=0}^{q}\theta_j E(\epsilon_{t-i}\epsilon_{t-j+k})\right) \tag{2.6}$$

Sfruttando ancora le proprietà del *white noise*, si ha che $E(\epsilon_{t-i}\epsilon_{t-j+k}) = \sigma^2$ per $j = i + k$ e 0 in tutti gli altri casi, cosicché l'espressione precedente si riduce a:

$$\gamma_k = E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{q}\theta_i\theta_{i+k}$$

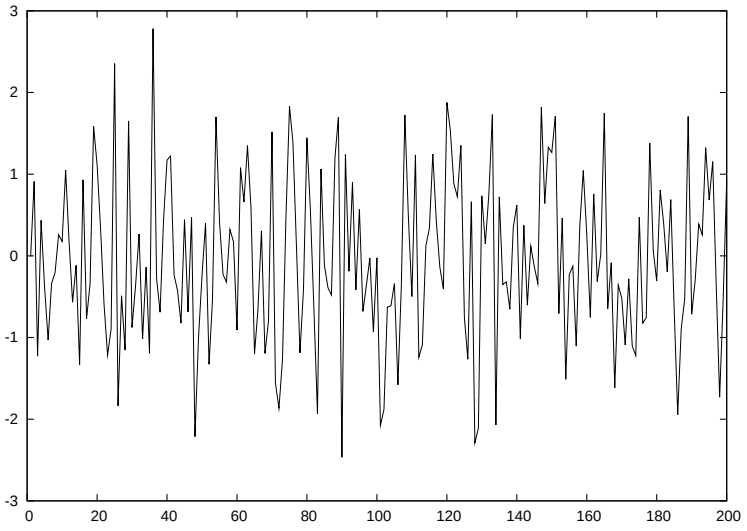
dove si intende che $\theta_i = 0$ per $i > q$.

Si noti che:

- L'espressione per la varianza è un caso particolare della formula precedente, ponendo $k = 0$;
- per $k > q$, le autocovarianze sono nulle.

Un processo MA(q), quindi, è un processo ottenuto come combinazione di diversi elementi di uno stesso *white noise* che presenta delle caratteristiche di persistenza tanto più pronunciate quanto più alto è il suo ordine. Quest'ultimo può anche essere infinito; in questo caso, tuttavia, l'esistenza dei momenti secondi (e quindi la stazionarietà) è garantita solo nel caso in cui $\sum_{i=0}^{q}\theta_i^2 < \infty$.

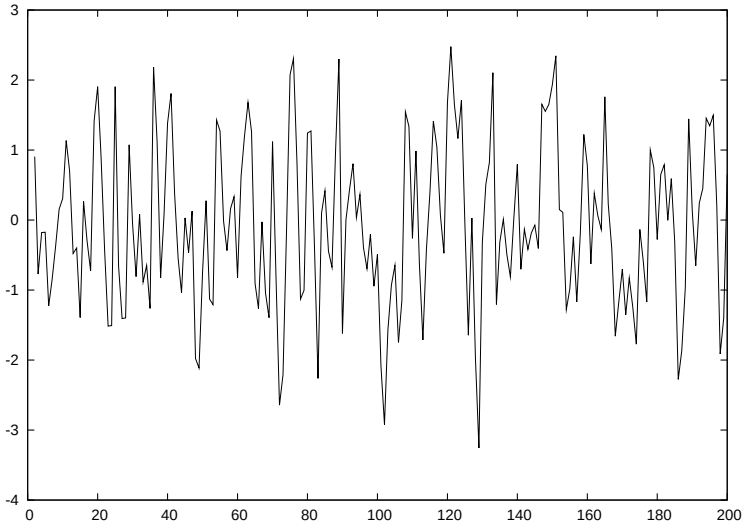
Figura 2.1: MA(1): $\theta = 0$ (*white noise*)



Esempio 2.3.1 *Consideriamo un processo MA(1) $x_t = \epsilon_t + \theta\epsilon_{t-1}$ e calcoliamo le sua autocovarianze: la sua varianza è data da*

$$E(x_t^2) = E\left(\epsilon_t + \theta\epsilon_{t-1}\right)^2 = E(\epsilon_t^2) + \theta^2 E(\epsilon_{t-1}^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) = (1+\theta^2)\sigma^2$$

Figura 2.2: MA(1): $\theta = 0.5$



Secondo la definizione, l'autocovarianza di ordine 1 è

$$E(x_t x_{t-1}) = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})\right]$$

Sviluppando il prodotto si ottiene

$$E(\epsilon_t\epsilon_{t-1}) + \theta E(\epsilon_t\epsilon_{t-2}) + \theta E(\epsilon_{t-1}^2) + \theta^2 E(\epsilon_{t-1}\epsilon_{t-2}) = \theta\sigma^2$$

L'autocorrelazione di ordine 1 è di conseguenza

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2}$$

In modo analogo si mostra che le autocovarianze di ordine superiore sono tutte nulle.

Tanto per avere un'idea più concreta, prendiamo un processo MA(1) di esempio e facciamone un grafico: se il processo è $y_t = \epsilon_t + \theta\epsilon_{t-1}$, l'andamento di y_t per diversi valori di θ è rappresentato nelle figure 2.1-2.3. Naturalmente, quando $\theta = 0$ (come nella figura 2.1) il processo è un *white noise*. Come si vede, al crescere di θ le caratteristiche di persistenza divengono più visibili (la serie si "smussa") e la sua varianza (misurata approssimativamente dall'ordine di grandezza delle ordinate) aumenta. Se avessimo simulato un processo MA di ordine superiore, la cosa sarebbe stata ancor più evidente.

Considerando più a fondo un processo MA(1), si possono fare alcune considerazioni interessanti. Come ho mostrato nell'esempio, l'autocorrelazione di ordine 1 di un processo MA(1) è data dalla formula $\rho_1 = \frac{\theta}{1+\theta^2}$. Questa relazione è rappresentata graficamente nella figura 2.4. Si può notare che il valore massimo che raggiunge ρ_1 è 0.5, in corrispondenza di $\theta = 1$; un discorso analogo, coi segni cambiati, vale per il punto di minimo. Inoltre, sappiamo dalle

Figura 2.3: MA(1): $\theta = 0.9$

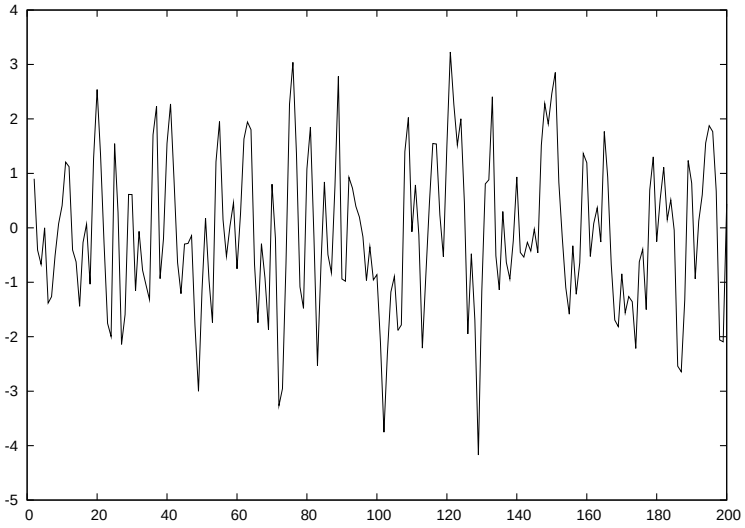
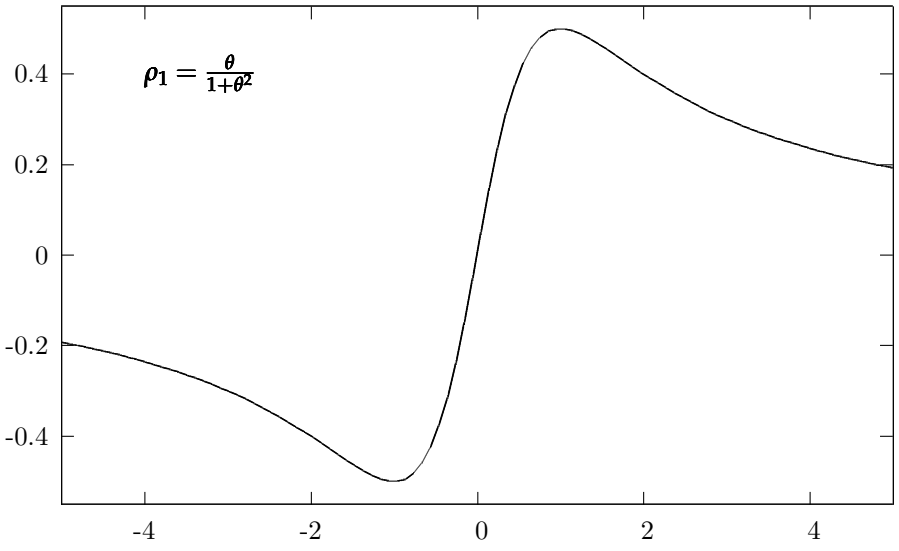


Figura 2.4: MA(1): Autocorrelazione di primo ordine in funzione di θ



considerazioni che ho fatto un paio di pagine fa (vedi equazione (2.6)), che tutte le autocorrelazioni di ordine maggiore di 1 sono nulle. Questo significa che il correlogramma di un processo MA(1) ha una sola barretta interessante (la prima), e anche quella è comunque vincolata a stare fra $-1/2$ e $1/2$.

Poniamoci ora un problema inferenziale: se volessimo rappresentare una certa serie storica come realizzazione di un processo MA(1), come potremmo

utilizzare le statistiche calcolabili sulla serie per ricavare delle stime dei parametri del processo (in questo caso, il parametro θ)? Naturalmente, questo procedimento sarebbe sostenibile solo nel caso in cui la nostra serie avesse un correlogramma empirico con valori moderati per l'autocorrelazione di primo ordine e trascurabili per le altre. Se così fosse, potremmo anche fare un ragionamento del tipo: se il processo che ha generato i dati è effettivamente un MA(1), allora è stazionario ed ergodico, per cui l'autocorrelazione campionaria converge in probabilità a quella teorica. In formule:

$$\hat{\rho}_1 \xrightarrow{\text{p}} \frac{\theta}{1+\theta^2};$$

poiché questa è una funzione continua di θ, posso invertirla e trovare uno stimatore consistente di θ col metodo dei momenti, ossia trovare quel valore $\hat{\theta}$ che soddisfa l'equazione

$$\hat{\rho}_1 = \frac{\hat{\theta}}{1+\hat{\theta}^2}; \tag{2.7}$$

Si vede facilmente che la soluzione della (2.7) è[6]

$$\hat{\theta} = \frac{1}{2\hat{\rho}_1}\left(1 - \sqrt{1-4\hat{\rho}_1^2}\right).$$

Si noti che, per l'esistenza dello stimatore, è necessario che $|\hat{\rho}_1| \leq 0.5$, ma in questo caso non c'è problema, perché stiamo appunto supponendo di avere a che fare con una serie in cui l'autocorrelazione di primo ordine non è troppo pronunciata.

In pratica, potremmo dire: visto che l'autocorrelazione campionaria è di — poniamo — 0.4, se sono convinto che il processo che ha generato i dati sia un MA(1), allora scelgo quel valore di θ tale per cui l'autocorrelazione teorica è anch'essa 0.4, ossia $\hat{\theta} = 0.5$. Naturalmente, questa strategia è perfettamente giustificata nella misura in cui la serie abbia effettivamente le caratteristiche di covarianza richieste, ossia una autocorrelazione di ordine 1 non troppo grande e autocorrelazioni successive trascurabili.

Ora, noi sappiamo che le cose non stanno sempre così: basta dare un'occhiata alle figure 1.2 a pagina 7 e 1.4 a pagina 9. È però vero che un processo MA di ordine superiore ha autocovarianze più articolate, e quindi si può congetturare che la stessa strategia potrebbe essere percorribile, almeno in teoria, a condizione di specificare un ordine del polinomio $C(L)$ abbastanza alto.

Facendo un passo più in là, ci si potrebbe chiedere se la congettura vale per qualunque struttura di autocovarianze. La risposta è nel mai abbastanza celebrato **teorema di rappresentazione di Wold**, di cui fornisco solo l'enunciato.

Teorema 1 (Teorema di rappresentazione di Wold) *Dato un qualunque processo stocastico y_t, stazionario in covarianza e a media 0, è sempre possibile trovare una*

[6] In effetti di valori ce ne sono due, perché la soluzione vera e propria sarebbe $\hat{\theta} = \frac{1 \pm \sqrt{1-4\hat{\rho}_1^2}}{2\hat{\rho}_1}$ (attenzione al simbolo $\pm$), ma per seguire l'argomento diamoci la regola di scegliere la soluzione interna all'intervallo $[-1,1]$, cioè quella riportata nel testo.

successione (non necessariamente finita) di coefficienti θ_i tali per cui

$$y_t = \sum_{i=0}^{\infty} \theta_i \epsilon_{t-i}$$

dove ϵ_t è un white noise.

In altri termini: per qualunque processo stocastico, purché stazionario, esiste un processo a media mobile che possiede la stessa struttura di autocovarianze. Questo risultato è di importanza enorme: esso ci dice, in sostanza, che qualunque sia la forma 'vera' di un processo stocastico stazionario, possiamo sempre rappresentarlo come un processo MA (al limite di ordine infinito). È per questo che, studiando i processi MA, stiamo di fatto studiando *tutti* i processi stazionari possibili, per lo meno per quanto riguarda le loro caratteristiche di media e di covarianza.

Il resoconto che ho appena dato del teorema di Wold non è proprio esatto: se guardate i libri seri, vi accorgete che il teorema in realtà si applica a una classe di processi più ampia. Meglio, che il teorema non dice esattamente quello che trovate scritto sopra. Per essere precisi, bisognerebbe dire che ogni processo stazionario di secondo ordine può essre scomposto in una parte "deterministica" (cioè perfettamente prevedibile dato il passato) più una parte a media mobile. Io ho furbescamente semplificato la definizione aggiungendo le parole "a media 0", escludendo così l'esistenza della parte deterministica, ma il messaggio rimane lo stesso.

2.4 Processi AR

Un'altra importante classe di processi è data dai processi **AR (AutoRegressivi)**. Questi processi forniscono, in un certo senso, una rappresentazione più intuitiva di una serie persistente di quella dei processi MA, poiché l'idea è che il livello della serie al tempo t sia una funzione lineare dei propri valori passati, più un *white noise*. Il nome deriva appunto dal fatto che un modello AR somiglia molto ad un modello di regressione in cui le variabili esplicative sono i valori passati della variabile dipendente.

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t \tag{2.8}$$

Non è ozioso notare che, in questo contesto, il *white noise* ϵ_t può essere interpretato in modo analogo al disturbo di un modello di regressione, cioè come la differenza fra y_t e la sua media condizionale; in questo caso, le variabili casuali che costituiscono l'insieme di condizionamento sono semplicemente il passato di y_t. I processi AR sono in un certo senso speculari ai processi MA perché, se un processo MA è un processo definito dall'applicazione di un polinomio nell'operatore L ad un *white noise*, un processo AR è definito come un processo l'applicazione al quale di un polinomio nell'operatore L produce un *white noise*. In simboli

$$A(L)y_t = \epsilon_t,$$

dove $A(L)$ è il solito polinomio in L (di grado p) con $A(0) = 1$ e $a_i = -\varphi_i$.

Per familiarizzarci con questo tipo di processi, iniziamo col considerare il caso più semplice: quello in cui $p = 1$ e il processo può essere scritto

$$y_t = \varphi y_{t-1} + \epsilon_t \longrightarrow (1 - \varphi L) y_t = \epsilon_t$$

Quali sono le caratteristiche di questo processo? Tanto per cominciare, vediamo come sono fatti i suoi momenti. I momenti di un processo AR(1) possono essere ricavati in diversi modi: uno piuttosto intuitivo è quello di supporre la stazionarietà del processo, e poi derivare le conseguenze di questa ipotesi. Supponiamo quindi che il processo abbia media costante μ. Quest'ipotesi implica

$$\mu = E(y_t) = \varphi E(y_{t-1}) + E(\epsilon_t) = \varphi \mu$$

L'espressione precedente può essere vera in due casi: o $\mu = 0$, nel qual caso è vera per qualsiasi valore di φ, oppure nel caso $\varphi = 1$, e allora l'espressione è vera per qualsiasi valore di μ, e la media del processo è indeterminata. In questo secondo caso si dice che il processo presenta una **radice unitaria**, perché il valore di z per cui $A(z) = 0$ è appunto 1; l'analisi di questa situazione, in cui accadono cose bizzarre, ha occupato pesantemente le menti dei migliori econometrici e le pagine delle riviste scientifiche negli ultimi vent'anni del XX secolo, e per molto tempo è stato considerato dagli economisti applicati un terreno impervio su cui è meglio non avventurarsi se non con una guida indigena. Noi ne parleremo nei capitoli 3 e 5. Per il momento, escludiamo dall'indagine i polinomi per cui $A(1) = 0$. Ne consegue che — nei casi che analizziamo qui — il processo ha media 0.

Un altro modo di derivare $E(y_t)$ è quello di rappresentare y_t come un processo a media mobile. Per farlo, utilizziamo i risultati riportati sopra sulla manipolazione dei polinomi. Se ci limitiamo ai casi in cui $|\varphi| < 1$ (condizione che chiaramente esclude la radice unitaria), si avrà che

$$A(L)^{-1} = (1 - \varphi L)^{-1} = 1 + \varphi L + \varphi^2 L^2 + \cdots = C(L)$$

e quindi la rappresentazione MA di y_t sarà

$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right) \epsilon_t = C(L)\epsilon_t$$

cioè un processo MA con $\theta_i = \varphi^i$, che ha media zero[7]; quindi, $E(y_t) = 0$.

Per quanto riguarda i momenti secondi, procediamo come sopra; supponiamo che il *white noise* ϵ_t abbia varianza pari a σ^2. Se indichiamo con V la varianza di y_t, e supponiamo che essa esista e sia costante nel tempo, avremo che

$$V = E(y_t^2) = E\left[(\varphi y_{t-1} + \epsilon_t)^2\right] = \varphi^2 V + \sigma^2 + 2\varphi E(y_{t-1}\epsilon_t)$$

L'ultimo elemento della somma è 0, poiché $y_{t-1} = C(L)\epsilon_{t-1}$, e quindi $E(y_{t-1}\epsilon_t)$ è una combinazione lineare di autocovarianze di un *white noise* (tutte nulle per

[7] La rappresentazione in media mobile di un processo AR(1) può anche essere ricavata col cosiddetto metodo delle "sostituzioni successive", che è più casareccio e meno elegante. Consideriamo che, se $y_t = \varphi y_{t-1} + \epsilon_t$, allora si avrà anche $y_{t-1} = \varphi y_{t-2} + \epsilon_{t-1}$; sostituiamo la seconda espressione nella prima e procediamo iterativamente.

definizione). Se ne deduce che

$$V = \varphi^2 V + \sigma^2 \Longrightarrow V = \frac{\sigma^2}{1-\varphi^2}$$

Lo stesso risultato poteva anche essere ottenuto dalla rappresentazione MA, notando che

$$V = \sigma^2 \sum_{i=0}^{\infty} \theta_i^2 = \sigma^2 \sum_{i=0}^{\infty} \varphi^{2i} = \sigma^2 \sum_{i=0}^{\infty} (\varphi^2)^i = \frac{\sigma^2}{1-\varphi^2}$$

L'espressione $V = \frac{\sigma^2}{1-\varphi^2}$ ci dice più di una cosa. In primo luogo, ci dice che solo se $|\varphi| < 1$ ha senso parlare di varianza stabile nel tempo (per $|\varphi| \geq 1$ non vale più l'ultima eguaglianza); questa condizione esclude dal novero dei processi AR(1) stazionari non solo quelli a radice unitaria, ma anche quelli a radice cosiddetta esplosiva ($|\varphi| > 1$).

La seconda considerazione nasce dal confronto di V, che è la varianza *non condizionale* di y_t, con σ^2, che è la varianza di $y_t|\Im_{t-1}$. V è sempre maggiore di σ^2, e la differenza è tanto maggiore quanto più φ è vicino a 1: tanto più persistente è il processo, tanto più la sua varianza condizionale al proprio passato sarà minore della sua varianza non condizionale. Vale a dire che la conoscenza del valore di y_{t-1} riduce l'incertezza sul valore di y_t quanto più persistente è la serie.

Rimangono da vedere le autocovarianze: l'autocovarianza di ordine 0 è V, che conosciamo già; l'autocovarianza di ordine 1 è data da

$$\gamma_1 = E(y_t y_{t-1}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-1}\right] = \varphi V$$

e più in generale

$$\gamma_k = E(y_t y_{t-k}) = E\left[(\varphi y_{t-1} + \epsilon_t) y_{t-k}\right] = \varphi \gamma_{k-1}$$

e si deduce che

$$\gamma_k = \varphi^k \frac{\sigma^2}{1-\varphi^2}$$

Oppure, partendo dalla rappresentazione MA, si ha che

$$E(y_t y_{t+k}) = \sigma^2 \sum_{i=0}^{q} \theta_i \theta_{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^i \varphi^{i+k} = \sigma^2 \sum_{i=0}^{q} \varphi^{2i+k}$$

che è uguale a

$$\gamma_k = \varphi^k \sigma^2 \sum_{i=0}^{q} \varphi^{2i} = \varphi^k \frac{\sigma^2}{1-\varphi^2}$$

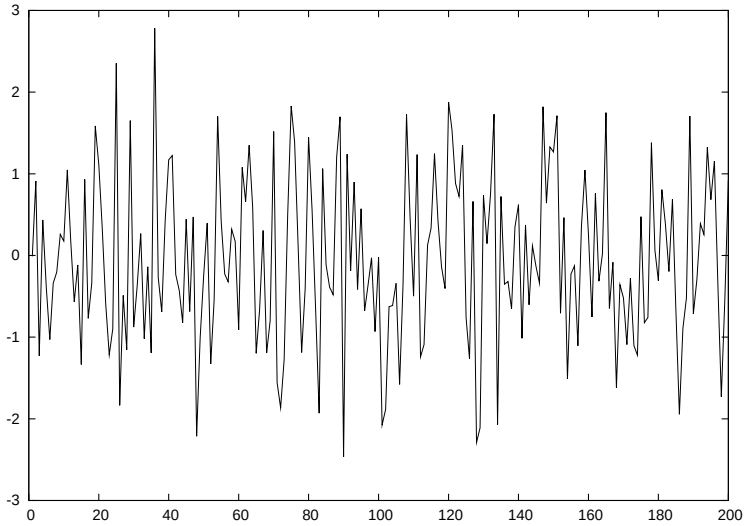
Le autocorrelazioni assumono in questo caso una forma molto semplice:

$$\rho_k = \varphi^k$$

Anche in questo caso è possibile dare un'interpretazione intuitiva del risultato: le autocorrelazioni, che sono un indice della memoria del processo,

sono tanto più grandi (in valore assoluto), tanto più grande (in valore assoluto) è φ, confermando l'interpretazione di φ come parametro di persistenza. In più, sebbene $\lim_{k\to\infty} \gamma_k = 0$, γ_k è sempre diverso da 0. In un certo senso, si può dire che la memoria del processo è infinita, anche se il passato molto remoto gioca un ruolo di fatto irrilevante.

Figura 2.5: AR(1): $\varphi = 0$ (*white noise*)



Vediamo anche qui un esempio. La figura 2.5 non rappresenta altro che il *white noise* già presentato in figura 2.1 come esempio sui processi MA(1). Applichiamo a questo *white noise* l'operatore $(1 - \varphi L)^{-1}$, con $\varphi = 0.5$ e $\varphi = 0.9$. Anche in questo caso, si nota un aumento delle caratteristiche di persistenza all'aumentare del parametro (φ in questo caso), anche se qui la cosa è molto più marcata.

Come nel caso dei processi MA, è facile generalizzare i processi AR al caso di media non nulla: supponiamo di aggiungere al modello AR(1) un'"intercetta":

$$y_t = \mu + \varphi y_{t-1} + \epsilon_t \to (1 - \varphi L) y_t = \mu + \epsilon_t$$

Invertendo il polinomio $A(L)$ si ha

$$y_t = \left(1 + \varphi L + \varphi^2 L^2 + \cdots\right)(\mu + \epsilon_t) = C(L)(\mu + \epsilon_t)$$

poiché l'applicazione di L a una costante la lascia inalterata, si ha

$$y_t = (1 + \varphi + \varphi^2 + \cdots)\mu + C(L)\epsilon_t = \frac{\mu}{1 - \varphi} + C(L)\epsilon_t$$

e quindi $E(y_t) = \frac{\mu}{1-\varphi}$.

La generalizzazione al caso AR(p) è piuttosto noiosa dal punto di vista dei maneggi algebrici che sono necessari: la difficoltà è legata fondamentalmente al fatto che la rappresentazione in media mobile del processo deriva

Figura 2.6: AR(1): $\varphi = 0.5$

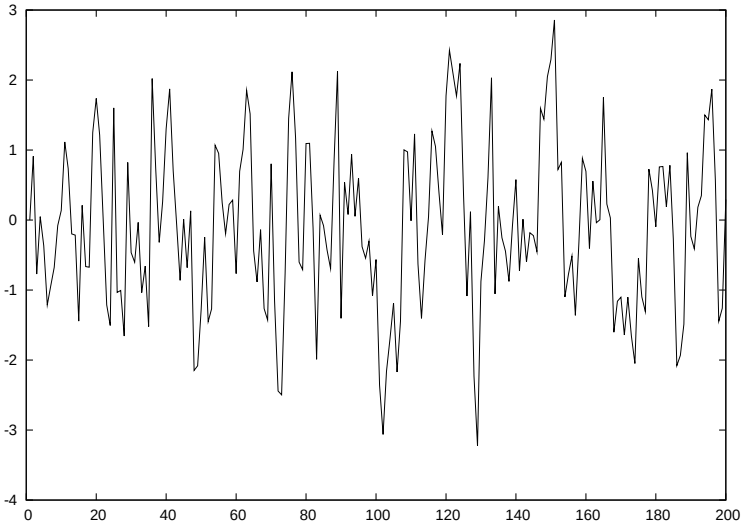
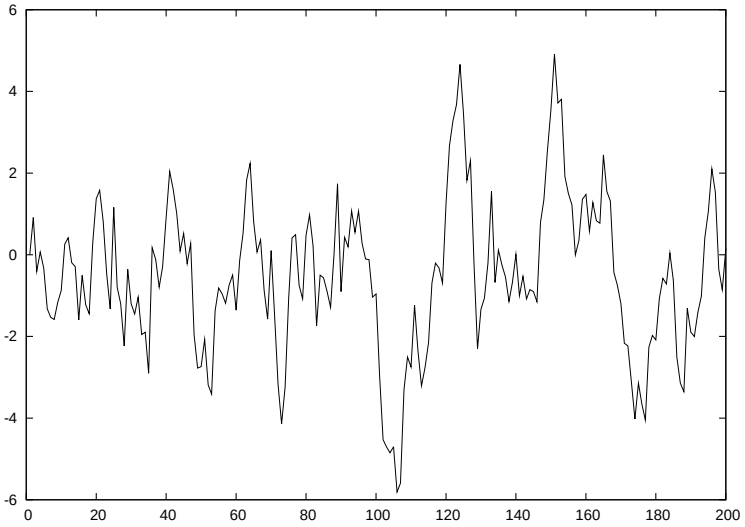


Figura 2.7: AR(1): $\varphi = 0.9$



dall'inversione di un polinomio di grado p-esimo. In pratica, si ha

$$C(L) = A(L)^{-1} = \prod_{j=1}^{p} (1 - \lambda_j L)^{-1}$$

dove le λ_j sono i reciproci delle radici di $A(L)$. D'altro canto, tale generalizzazione non porta grandi vantaggi alla comprensione intuitiva delle caratteristiche salienti di questi processi. Il punto fondamentale è che un processo

AR(p) è stazionario solo se $|\lambda_j| < 1$ per ogni j. Mi astengo dal dimostrarlo rigorosamente, ma il lettore curioso sappia che, tutto sommato, basta applicare la (2.1).

Nel caso in cui λ_j sia un numero complesso, ricordo qui che il suo valore assoluto è dato dalla formula $|a + bi| = \sqrt{a^2 + b^2}$. Se poi siete totalmente digiuni sull'argomento, magari potreste darvi una letta all'appendice a questo capitolo, che sta a pagina 54.

Altri fatti interessanti (non dimostro neanche questi) sono che un processo AR(p)

- ha memoria infinita, ma le autocorrelazioni decrescono al crescere di k in progressione geometrica;
- nel caso di "intercetta" diversa da 0, ha valore atteso $\frac{\mu}{A(1)}$, dove $A(1)$ è appunto il polinomio $A(z)$ valutato in $z = 1$ anziché in $z = L$ come al solito; in pratica, $A(1) = \sum_{i=0}^{p} a_i$.

L'unico aspetto che vale la pena di sottolineare del caso in cui l'ordine del processo autoregressivo p sia maggiore di 1 è che processi AR(p) possono avere andamenti ciclici: questo avviene se e solo se fra le radici del polinomio $A(z)$ c'è una coppia di numeri complessi coniugati. In questo caso, il processo assume un'andamento ciclico in cui l'ampiezza delle oscillazioni varia attorno ad un valore medio. Dovrebbe essere evidente che i processi di questo tipo sono i candidati naturali a modellare fenomeni economici caratterizzati da fasi cicliche.

Il motivo per cui esiste un legame fra numeri complessi ed andamenti ciclici sarebbe bellissimo da spiegare, ma purtroppo non posso farlo qui perché lo studente medio di una facoltà di Economia considera i numeri complessi e le funzioni trigonometriche una arcana stregoneria. Per gli stravaganti a cui piacciono queste cose, ho messo un'appendice a fondo capitolo cosicché i più non vengano disturbati.

Diamo un'occhiata ad un esempio: prendiamo il *white noise* di figura 2.5 ed utilizziamolo per costruire un processo AR(2) in cui il polinomio $A(z)$ non ha radici reali. Nella fattispecie,

$$y_t = 1.8y_{t-1} - 0.9y_{t-2} + \epsilon_t$$

e le radici sono

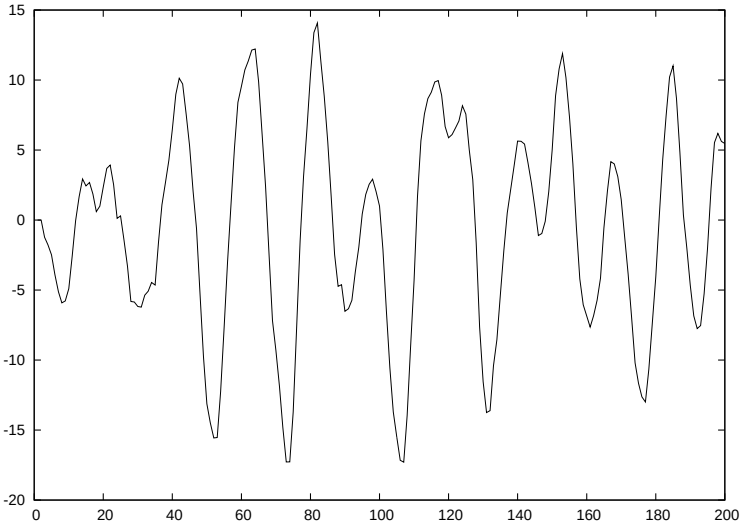
$$\lambda = \frac{1.8 \pm \sqrt{3.24 - 3.6}}{1.8} = 1 \pm \frac{i}{3},$$

ambedue maggiori di 1 in valore assoluto (sono "uno più qualcosa"). Come si nota nella figura 2.8, c'è un'alternanza più o meno regolare di 'picchi' e di 'pozzi'.

2.5 Processi ARMA

La classe dei processi ARMA comprende sia i processi AR che i processi MA come caso particolare. Un processo ARMA(p, q) è infatti definito da

Figura 2.8: AR(2): $\varphi_1 = 1.8$; $\varphi_2 = -0.9$



$$A(L)y_t = C(L)\epsilon_t \tag{2.9}$$

dove p è l'ordine del polinomio $A(L)$ e q è l'ordine del polinomio $C(L)$. Entrambi sono numeri finiti. I processi AR o MA sono quindi casi particolari ($q = 0$ e $p = 0$ rispettivamente).

Se il polinomio $A(L)$ ha tutte le sue radici maggiori di 1 in modulo, allora y_t può anche essere rappresentato in forma MA

$$y_t = A(L)^{-1}C(L)\epsilon_t = C^*(L)\epsilon_t$$

dove $C^*(L)$ è un polinomio di ordine infinito se $p > 0$. Tale condizione su $A(L)$ è necessaria e sufficiente affinché il processo sia stazionario.

Allo stesso modo, se il polinomio $C(L)$ è invertibile, allora y_t ammette una rappresentazione autoregressiva (di ordine infinito se $q > 0$)

$$C(L)^{-1}A(L)y_t = A^*(L)y_t = \epsilon_t$$

In questo caso, si dice anche che il *processo* è invertibile.

Le caratteristiche dei momenti di un processo ARMA(p,q) possono essere ricavate in modo concettualmente semplice (ma algebricamente esasperante) dalla sua rappresentazione in media mobile, e non le riporto qui. L'unica caratteristica che mi pare degna di menzione è che se aggiungiamo un'intercetta, si dimostra facilmente[8] che la media del processo è ancora $\frac{\mu}{A(1)}$. La cosa, poi, si può ulteriormente generalizzare (e, in parecchi casi, rendere più aderente

[8]Dimostrazione lampo: $A(L)y_t = \mu + C(L)\epsilon_t \Longrightarrow E[A(L)y_t] = \mu + E[C(L)\epsilon_t]$. Per la linearità degli operatori E e L, si ha che $A(L)E[y_t] = \mu + C(L)E[\epsilon_t] = \mu$. Ma se y_t è stazionario $E[y_t]$ esiste finito e costante, per cui $A(L)E[y_t] = A(1)E[y_t]$, da cui $E[y_t] = \frac{\mu}{A(1)}$.

alla realtà), prevedendo la possibilità di una media non nulla e variabile nel tempo, ovvero un processo del tipo

$$A(L)y_t = \mu(x_t, \beta) + C(L)\epsilon_t,$$

a somiglianza di un modello di regressione. Un modo alternativo di scrivere questa cosa è quello di pensare a un modello del tipo

$$y_t = \mu_t + u_t,$$

dove $\mu_t = \frac{1}{A(L)}\mu(x_t, \beta)$ e $u_t = \frac{C(L)}{A(L)}\epsilon_t$, ovvero come ad un modello di regressione dove gli errori sono dati da un processo ARMA(p, q). Come si vede, è facile passare da una rappresentazione all'altra.

Che senso ha studiare processi ARMA? In linea teorica, nessuna, visto che il teorema di rappresentazione di Wold ci dice che qualunque processo stazionario può essere rappresentato come un processo MA. Da un punto di vista pratico, tuttavia, c'è il problema che la rappresentazione di Wold è, in generale, infinita. Questo non è un problema a livello teorico, ma lo diventa nella pratica: la serie che osserviamo viene infatti pensata come realizzazione di un processo stocastico, i cui parametri sono i coefficienti dei polinomi nell'operatore L che ne determinano le caratteristiche di persistenza (più la varianza del *white noise*).

Se si considera una serie osservata come una realizzazione di un qualche processo stazionario, utilizzare un processo MA per riassumerne le caratteristiche di media e covarianza comporta quindi il problema inferenziale di stimare un numero potenzialmente infinito di parametri. Infatti, se pensiamo che y_t sia rappresentabile in forma MA come

$$y_t = B(L)\epsilon_t$$

niente ci assicura che il polinomio $B(L)$ non sia di ordine infinito. Si può però pensare di usare un'approssimazione di $B(L)$; in particolare, può darsi che si riescano a trovare due polinomi di ordine finito (e possibilmente basso) $A(L)$ e $C(L)$ tali per cui

$$B(z) \simeq \frac{C(z)}{A(z)}$$

Se l'uguaglianza fosse esatta, si potrebbe allora scrivere

$$A(L)y_t = C(L)\epsilon_t$$

Se l'uguaglianza vale solo in modo approssimato, allora si avrà

$$A(L)y_t = C(L)\epsilon_t^*$$

dove

$$\epsilon_t^* = \frac{A(L)}{C(L)}B(L)\epsilon_t$$

Il processo ϵ_t^* non è, a rigore, un *white noise*, ma se le sue autocovarianze non sono troppo grandi, può essere considerato tale a tutti i fini pratici. Si potrebbe

dire, da un'altra prospettiva, che considerare ϵ_t^* un *white noise* costituisce una metafora dei dati che non è molto più fuorviante di quella basata su ϵ_t, ossia sulla rappresentazione di Wold e che ha il vantaggio di basarsi su un numero finito di parametri.

In pratica, un modello ARMA viene costruito facendo un'ipotesi a priori sui gradi dei due polinomi $A(L)$ e $C(L)$ e poi, una volta stimati i coefficienti dei polinomi, esaminando le autocorrelazioni campionarie della serie corrispondente a ϵ_t^*. Se queste non sono troppo grandi, non ci sono problemi di sorta a considerare ϵ_t^* come un *white noise*[9].

L'esigenza di tener basso il numero dei parametri dei polinomi conduce, in certi casi, a lavorare con dei modelli noti come **ARMA moltiplicativi**, che si usano soprattutto per serie caratterizzate da persistenza stagionale, e che quindi sono anche conosciuti come **ARMA stagionali**, o **SARMA**.

Ad esempio: consideriamo la serie storica mensile delle presenza alberghiere nel comune di, che so, Riccione. È chiaro che c'è una forte stagionalità, nel senso che il dato di agosto somiglia probabilmente molto di più a quello di agosto dell'anno prima piuttosto che a quello di marzo dello stesso anno, che è più vicino nel tempo, ma "idealmente" più distante. Per semplicità, immaginiamo di voler utilizzare un modello autoregressivo puro, cioè senza parte MA. Un'applicazione bovina delle idee esposte fin qui condurrebbe, evidentemente, all'uso di un polinomio di ordine (almeno) 12, e quindi di una struttura con un discreto numero di parametri; molti di questi, però, sono probabilmente ridondanti, perché magari la media condizionale del mese di agosto dipende sì da agosto dell'anno prima, ma non si vede perché il dato di febbraio dovrebbe essere rilevante. Questa osservazione, di per sé, ci condurrebbe semplicemente ad utilizzare un polinomio $A(L)$ con dei "buchi", ossia dei coefficienti pari a 0. Un modo più elegante e più efficiente è quello di scrivere il polinomio dividendo gli effetti stagionali dagli altri. Consideriamo il polinomio dato da

$$A(L) = (1 - \varphi L)(1 - \Phi L^s) = 1 - \varphi L - \Phi L^s + (\varphi \cdot \Phi) L^{s+1},$$

dove s è il numero di sottoperiodi (cioè 12 per i mesi in un anno, e così via). Ovviamente, $A(L)$ è, in questo caso, un polinomio di ordine $s + 1$, i cui coefficienti sono tutti nulli, a parte tre: quelli di ordine 1, s e $s + 1$. Il numero di parametri che lo caratterizzano, però, è solo 2, perché il coefficiente di ordine $s + 1$ è il prodotto degli altri due, cosicché è possibile modellare un andamento stagionale anche piuttosto lungo tenendo sotto controllo il numero dei parametri necessari per farlo. In particolare, gli effetti stagionali sono sintetizzati nel solo parametro Φ, azzerando il quale gli effetti stagionali scompaiono.

Evidentemente, un giochino del genere può essere anche fatto sul polinomio $C(L)$, per cui il grado di flessibilità a cui si giunge può essere notevole senza che la dimensione dei parametri esploda in modo incontrollato. Generalizzando in modo ovvio l'espressione sopra, si ha un modello che può essere scritto come

$$A(L)B(L^s)y_t = C(L)D(L^s)\epsilon_t$$

[9] Sulle tecniche di stima, vedi il paragrafo 2.7

che contiene, appunto, le parti stagionali autoregressiva $B(L^s)$ e a media mobile $D(L^s)$. Se l'ordine dei polinomi $B(\cdot)$ e $D(\cdot)$ è zero, si ricade nel caso ARMA puro e semplice.

2.6 Uso dei modelli ARMA

Se i parametri di un processo ARMA sono noti, il modello può essere usato per due scopi: previsione dell'andamento futuro della serie e/o analisi delle sue caratteristiche dinamiche.

2.6.1 Previsione

Per quanto riguarda il primo punto, la miglior previsione per i valori futuri di y_t si può calcolare sulla base di questo ragionamento: definiamo come **previsore** di y_t una qualche funzione delle variabili contenute nel set informativo $\Im_{T-1}$. Un previsore, cioè, è una qualche regola che determina la previsione che facciamo su y_t dati i suoi valori precedenti, che supponiamo di conoscere. Chiamiamo questo valore $\hat{y}_t = f(y_{t-1}, y_{t-2}, \ldots)$. Naturalmente, questa regola ce la inventiamo noi, e si pone il problema di inventarcela in modo che funzioni "bene".

Se y_t è un processo ARMA (o rappresentabile come tale), una volta che abbiamo il modello nella forma $A(L)y_t = C(L)\epsilon_t$, un'ipotesi sulla distribuzione di ϵ_t ci mette in condizione, almeno in linea di principio, di determinare la distribuzione della variabile casuale $y_t|\Im_{T-1}$. È evidente che questo ci mette in grado anche di determinare la distribuzione condizionale dell'*errore* di previsione, cioè della variabile

$$e_t = y_t - \hat{y}_t.$$

La distribuzione di $e_t|\Im_{T-1}$ diventa rilevante se dobbiamo scegliere quale funzione usare come previsore. A rigore, una scelta ottimale dovrebbe essere fatta secondo questo criterio:

1. in primo luogo, scegliamo una funzione $c(e_t)$ (cosiddetta di *perdita*), che associa un costo all'errore di previsione. In generale, si ha che $c(0) = 0$ (il costo di una previsione perfetta è 0) e $c(e_t) \geq 0$ per $e_t \neq 0$.

2. Definiamo a questo punto la perdita attesa come

$$c^* = E\left[c(e_t)|\Im_{T-1}\right] = E\left[c(y_t - \hat{y}_t)|\Im_{T-1}\right];$$

   la grandezza c^* è il costo che in media ci tocca sostenere a causa delle previsioni sbagliate. Naturalmente vogliamo che essa sia più piccola possibile.

3. Siccome c^* è una funzione di $\hat{y}_t$, scegliamo $\hat{y}_t$ in modo tale da minimizzare c^*, ossia *definiamo* $\hat{y}_t$ come quella funzione che minimizza il costo atteso dell'errore di previsione.

Dovrebbe essere chiaro a questo punto che quale sia il miglior previsore dipende dalle caratteristiche della funzione di perdita e per ogni problema pratico il previsore ottimo può essere diverso. L'esempio che faccio sempre è la prenotazione di un ristorante: poiché in questo caso la funzione di perdita è asimmetrica (meglio avere sedie vuote che gente in piedi), conviene sempre prenotare per un numero di persone leggermente superiore di quello che realmente si pensa.

Per fortuna, però, la faccenda diventa molto meno intricata se la funzione di perdita è quadratica, cioè se $C(e_t) = \kappa e_t^2$ per κ positivo qualunque. In questo caso (che spesso può essere preso come approssimazione soddisfacente della funzione di costo più appropriata) si può dimostrare che $\hat{y}_t$ coincide con il valore atteso condizionale:

$$C(e_t) = \kappa e_t^2 \Longrightarrow \hat{y}_{T+1} = E(y_{T+1}|\Im_T).$$

Questa proprietà è così comoda che nella stragrande maggioranza dei casi si prende la media condizionale come previsore senza neanche giustificare la scelta.

Dato un insieme di osservazioni che vanno da 1 a T, ammettiamo perciò che il miglior previsore di y_{T+1} sia la sua media condizionale al set informativo di cui disponiamo, ossia

$$\hat{y}_{T+1} = E(y_{T+1}|\Im_T). \tag{2.10}$$

Nel caso di un modello AR puro, la soluzione è banale, poiché tutti i valori di y fino al tempo T sono noti, e quindi $E(y_{t-k}|\Im_T) = y_{t-k}$ per qualunque $k \geq 0$:

$$E(y_{T+1}|\Im_T) = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1} + E(\epsilon_{T+1}|\Im_T)$$

ma il valore di $E(\epsilon_{T+1}|\Im_T)$ è evidentemente 0, poiché l'assenza di memoria del *white noise* garantisce[10] che non ci sia informazione disponibile al presente sul futuro di ϵ; di conseguenza, $E(\epsilon_{T+1}|\Im_T) = E(\epsilon_{T+1}) = 0$. La previsione di y_{T+1} è quindi

$$\hat{y}_{T+1} = \varphi_1 y_T + \cdots + \varphi_p y_{T-p+1} \tag{2.11}$$

Visto che ancora stiamo sul teorico, qui stiamo assumendo che il set informativo a nostra disposizione si estenda infinitamente all'indietro nel passato, cosa che ci semplifica molto le cose, perché significa che $\hat{y}_{T+1}$ è facilmente calcolabile tramite la (2.11). Se il nostro set informativo (come accade nella realtà) si interrompe ad una qualche data iniziale, il meccanismo vale ancora per processi stazionari, anche se in modo approssimato.

Per la previsione a due periodi in avanti, ripetiamo il ragionamento precedente partendo dall'espressione:

$$\hat{y}_{T+2} = E(y_{T+2}|\Im_T) = \varphi_1 E(y_{T+1}|\Im_T) + \cdots + \varphi_p y_{T-p+2} + E(\epsilon_{T+2}|\Im_T)$$

che si dimostra facilmente essere pari a

$$\hat{y}_{T+2} = \varphi_1 \hat{y}_{T+1} + \cdots + \varphi_p y_{T-p+2}$$

[10] Il lettore pignolo farà rimarcare che qui sto implicitamente assumendo che ϵ_t sia una differenza di martingala, che non necessariamente coincide con un *white noise*. Ebbene sì, lo sto assumendo.

e più in generale

$$\hat{y}_{T+k} = \varphi_1 \hat{y}_{T+k-1} + \cdots + \varphi_p \hat{y}_{T+k-p},$$

dove naturalmente $\hat{y}_{T+k} = y_{T+k}$ per $k \leq 0$. Si noti l'intrigante parallelismo fra $A(L)y_t = \epsilon_t$ e $A(L)\hat{y}_t = 0$, a cui si arriva facilmente considerando il valore atteso (condizionale a $\Im_{t-1}$) della prima delle due espressioni.

Esempio 2.6.1 *Dato un processo AR(2) così parametrizzato*

$$y_t = 0.9y_{t-1} - 0.5y_{t-2} + \epsilon_t,$$

supponiamo di osservarne una realizzazione, e che le ultime due osservazioni siano pari a: $y_{T-1} = 2$ e $y_T = 1$. La miglior previsione per y_{T+1} è quindi

$$\hat{y}_{T+1} = 0.9 \times 1 - 0.5 \times 2 = -0.1$$

per la previsione di y_{T+2} risulta

$$\hat{y}_{T+2} = 0.9 \times (-0.1) - 0.5 \times 1 = -0.59$$

e si può continuare; per la cronaca, i cinque valori seguenti sono -0.481, -0.1379, 0.11639, 0.173701, 0.098136

Naturalmente, la valutazione della media condizionale dà un valore puntuale, ma non dice nulla sull'attendibilità della previsione, cioè sulla dispersione dell'errore che ci attendiamo di commettere.

In termini più statistici, è necessario valutare anche la varianza dell'errore di previsione. Questo non è un argomento su cui vorrei intrattenermi più di tanto. Al lettore interessato mi limito a suggerire, oltre ai soliti riferimenti bibliografici che trova in fondo, che un utile esercizio può essere quello di provare che, nel caso di un AR(1),

$$V(\hat{y}_{T+k}) = \sigma^2 \frac{1 - \varphi^{2k}}{1 - \varphi^2}$$

Può essere interessante notare che la varianza dell'errore di previsione è sempre minore della varianza non condizionale di y_t: questo significa che sfruttare le caratteristiche di persistenza della serie storica permette di rendere meno incerto il suo comportamento futuro. Peraltro, per $k \to \infty$, le due varianze tendono a coincidere, e questo avviene perché nei processi AR(1) stazionari la persistenza ha sempre un carattere di breve periodo. La conoscenza dello stato del sistema oggi non è informativa sul futuro remoto del sistema stesso più di quanto non lo sia la sua distribuzione non condizionale: per k abbastanza grande, y_t e y_{t+k} sono virtualmente incorrelate (e quindi, se gaussiane, virtualmente indipendenti).

In pratica, poi, le cose sono un tantino più complicate. Intanto perché qui stiamo ipotizzando di conoscere i veri parametri del processo, quando in realtà di solito lavoriamo con delle stime, e quindi la varianza dell'errore di previsione dipende non solo dalla variabilità intrin-

seca del processo, ma anche dal fatto che esiste incertezza sui parametri del processo stesso. Ad esempio, nel caso di un processo AR(1) per cui avessimo una stima $\hat{\varphi}$ del parametro, il ragionamento fin qui seguito ci condurrebbe ad analizzare

$$\hat{y}_{T+k} = E(y_{T+k}|\Im_T) = E(\hat{\varphi} \cdot y_{T+k-1}|\Im_T),$$

dove $\hat{\varphi}$ non può essere "tirato fuori" dell'operatore valore atteso perché è uno stimatore e non una costante. Va detto, peraltro, che questa distinzione è importante per questioni teoriche, ma in pratica la distinzione scompare e normalmente si fa uso dei parametri stimati come se fossero quelli veri.

Nel caso più generale di processi ARMA, le previsioni si possono fare applicando ancora lo stesso concetto. In particolare, si noti che, se $\Im_{t-1}$ non ha limite temporale inferiore, allora esso comprende non solo tutti i valori passati di y_t, ma anche quelli di ϵ_t: se infatti il processo è invertibile, si può scrivere

$$C(L)^{-1}A(L)y_t = G(L)y_t = \epsilon_t$$

da cui

$$\epsilon_{t-k} = y_{t-k} + g_1 y_{t-k-1} + g_2 y_{t-k-2} + \cdots$$

e quindi sono noti (nel senso "ricavabili da $\Im_{t-1}$") anche tutti i valori del *white noise* fino al tempo $t-1$. A questo punto, si può applicare ad ogni ingrediente di un modello ARMA l'operatore valore atteso condizionale. Il fatto che poi il set informativo a nostra disposizione non sia infinito rappresenta solo un problema di minore entità. Se infatti noi abbiamo solo osservazioni nell'arco di tempo $\{0 \cdots T\}$, una soluzione molto comoda è quella di estendere il nostro set informativo all'indietro usando i valori medi non condizionali di $y_{-1}, y_{-2}, \cdots$ eccetera. Se il processo è stazionario ed ergodico, al crescere del campione non c'è differenza[11].

Esemplifico nel caso di un ARMA(1,1), perché una volta capito il concetto la generalizzazione è banale. Supponiamo quindi di sapere che il processo ha la forma

$$y_t = \varphi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}.$$

Mettiamoci all'istante 0, in cui non abbiamo alcuna osservazione. Qual è la migliore previsione che possiamo fare su y_1? Visto che non abbiamo dati, la media condizionale coincide con la media marginale, e quindi $\hat{y}_1 = E(y_1) = 0$. Passa un periodo, e osserviamo il dato effettivo y_1. A questo punto, possiamo calcolare l'errore di previsione per il periodo 1, ossia $e_1 = y_1 - \hat{y}_1$; poiché $\hat{y}_1$ è 0, per i motivi che abbiamo appena detto, ne consegue che $e_1 = y_1$. A questo punto, possiamo calcolare $\hat{y}_2$, con la seguente formula:

$$\hat{y}_2 = E(y_2|\Im_1) = E(\varphi y_1 + \epsilon_2 + \theta\epsilon_1|\Im_1) = \varphi E(y_1|\Im_1) + E(\epsilon_2|\Im_1) + \theta E(\epsilon_1|\Im_1).$$

Ragioniamo un addendo per volta, tenendo a mente che $\Im_1 = y_1$: evidentemente, i primi due termini non pongono problemi, perché $E(y_1|\Im_1) = y_1$ (è ovvio) e $E(\epsilon_2|\Im_1) = 0$ (per ipotesi). Ma che dire di $E(\epsilon_1|\Im_1)$? Poiché ϵ_1 è anche interpretabile come l'errore di previsione che si commetterebbe al tempo 0 se il set informativo fosse infinito, allora *la miglior previsione possibile che*

[11] Il calcolo *esatto*, volendo, si può fare. Ci sono molti modi, ma il più comune — anche perché facilmente automatizzabile – è quello di usare un attrezzo che si chiama **filtro di Kalman**. Per chi vuole saperne di più, c'è la letteratura.

possiamo fare sull'errore di previsione al tempo 1 è esattamente l'errore di previsione che abbiamo effettivamente commesso. In base a questo ragionamento, possiamo formulare la nostra previsione su y_2 come

$$\hat{y}_2 = \varphi y_1 + \theta e_1.$$

Facciamo passare un altro periodo, e osserviamo y_2; da qui calcoliamo e_2, e il giochino prosegue, nel senso che a questo punto abbiamo tutto quel che ci serve per calcolare $\hat{y}_3 = \varphi y_2 + \theta e_2$, eccetera eccetera. In pratica, le previsioni un passo in avanti su processi del tipo

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

si fanno così:

$$\hat{y}_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}, \tag{2.12}$$

ovvero utilizzando i valori effettivamente osservati delle y_t e i valori degli errori di previsione passati al posto delle ϵ_{t-i}.

Piccola digressione. Ci si potrebbe legittimamente chiedere quale sia il valore pratico delle previsioni fatte in questo modo; in effetti, noi abbiamo sempre a che fare con serie storiche a cui associamo processi ARMA solo come rappresentazione stilizzata e approssimata. In altri termini, le caratteristiche *storiche* di persistenza della serie vengono sintetizzate giocando a far finta che la serie che osserviamo sia una realizzazione di un processo ARMA che, guarda caso, ha proprio quelle caratteristiche.
Non c'è nessun motivo *logico*, però, per cui una approssimazione che andava bene per il passato continui ad andare bene per il futuro. Per considerare attendibile una previsione di una serie storica fatta in questo modo, è necessario assumere, più o meno implicitamente, che l'insieme di circostanze che hanno fino ad oggi congiurato a far sì che quel certo processo fosse una buona approssimazione dell'andamento di quella certa serie storica continuino a valere per l'orizzonte temporale che ci interessa prevedere.
Questa condizione è spesso verosimile quando la serie è una descrizione di un fenomeno fisico (ad esempio, la temperatura rilevata giornalmente all'aeroporto di Falconara alle ore 8 del mattino) ragionevolmente stabile. Tuttavia, nel caso di fenomeni economici questa può essere un'ipotesi piuttosto coraggiosa, in quanto la catena causale di eventi che concorrono a determinare il valore della serie in un dato momento è verosimilmente più instabile: riterrei poco serio fare una previsione del prezzo del petrolio greggio che si basi esclusivamente su un processo ARMA e che, ad esempio, non tenga conto della situazione politica interna del Venezuela. Per meglio dire, la previsione di un modello ARMA va presa per buona come previsione condizionale ad uno scenario: *se e solo se* la situazione politica in Venezuela (e in Iran, e negli Stati Uniti, eccetera eccetera) rimane più o meno quella di oggi, *allora* si può dire che eccetera eccetera.

2.6.2 Analisi delle caratteristiche dinamiche

Questo aspetto è generalmente indagato facendo uso della cosiddetta **funzione di risposta di impulso**. Cos'è la funzione di risposta di impulso? La risposta a questa domanda passa attraverso una considerazione che possiamo fare alla luce di quanto detto nel sottoparagrafo precedente: consideriamo l'equazione

$$y_t = E[y_t | \Im_{t-1}] + \epsilon_t = \hat{y}_t + \epsilon_t,$$

che segue dall'equazione (2.10).

Il valore di y_t può quindi essere interpretato come la somma di due componenti: una ($\hat{y}_t$) che, almeno in linea di principio, è perfettamente prevedibile dato il passato; l'altra (ϵ_t) assolutamente imprevedibile. In altri termini, si può pensare che il valore di y_t dipenda da una componente di persistenza a cui si somma un disturbo, o, come si usa dire, **shock** casuale che riassume tutto ciò che è successo al tempo t che non poteva essere previsto. L'effetto di questa componente, tuttavia, si riverbera anche nel futuro della serie y_t attraverso l'effetto persistenza. È per questo che, sovente, il *white noise* ϵ_t viene chiamato, in forma più neutra, *errore di previsione ad un passo* o *innovazione*.

L'idea, a questo punto, è la seguente: se scriviamo il processo in forma MA

$$y_t = A(L)^{-1}C(L)\epsilon_t = B(L)\epsilon_t$$

si può pensare all'i-esimo coefficiente del polinomio $B(L)$ come all'effetto che lo shock avvenuto i periodi addietro ha sul valore attuale di y, o, equivalentemente, all'impatto che gli avvenimenti di oggi avranno sulla serie studiata fra i periodi.

$$b_i = \frac{\partial y_t}{\partial \epsilon_{t-i}} = \frac{\partial y_{t+i}}{\partial \epsilon_t}$$

La funzione di risposta di impulso, insomma, è data semplicemente dai coefficienti della rappresentazione MA del processo, e viene generalmente esaminata con un grafico che ha in ascissa i valori di i ed in ordinata i valori di b_i.

Per calcolarsi la rappresentazione di Wold di un processo ARMA di cui siano noti i parametri, quindi, bisogna calcolarsi il polinomio inverso di $A(L)$. Questo può essere piuttosto noioso, specie se l'ordine della parte autoregressiva è alto. Un algoritmo di calcolo decisamente più semplice, che può essere implementato anche su un comune foglio elettronico, è il seguente:

1. Definite una serie e_t che contiene tutti zeri fuorché per un periodo, in cui vale 1. Detto in un altro modo, definite una e_t per cui $e_0 = 1$, e $e_t = 0$ per $t \neq 0$.

2. Definite una serie i_t, che imponete uguale a 0 per $t < 0$; per $t \geq 0$, invece, valga $A(L)i_t = C(L)e_t$.

I valori che otterrete per la serie i_t sono esattamente i valori della funzione di risposta di impulso.

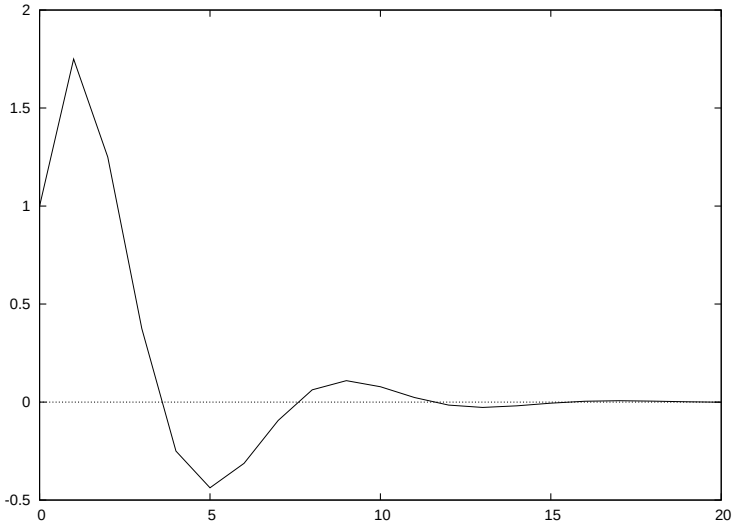
Esempio 2.6.2 *Prendiamo ad esempio un processo ARMA(2,1) così fatto:*

$$y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$$

e diciamo che, al tempo t, si è verificato un "evento imprevedibile" pari a 1 (ossia $\epsilon_t = 1$). Che effetto ha questo sui valori di y dal tempo t in poi? Ragioniamo con calma. Al tempo t, evidentemente, l'effetto è 1, poiché ϵ_t agisce direttamente su y_t e non influenza le sue altre componenti. Al tempo $t + 1$, avremo che

$$y_{t+1} = y_t - 0.5y_{t-1} + \epsilon_{t+1} + 0.75\epsilon_t,$$

Figura 2.9: Risposta di impulso per $y_t = y_{t-1} - 0.5y_{t-2} + \epsilon_t + 0.75\epsilon_{t-1}$



e l'effetto di ϵ_t su y_{t+1} sarà duplice: da una parte, esso compare direttamente, associato ad un coefficiente di 0.75; dall'altra, bisogna tenere conto del fatto che l'effetto di ϵ_t è anche contenuto in y_t, a cui è associato un coefficiente pari a 1: l'effetto totale sarà perciò di 1.75. Andando avanti ancora di un periodo, l'effetto diretto scompare e rimane soltanto quello generato dai valori ritardati della y. Facendo un po' di conti, si ha che l'effetto di ϵ_t su y_{t+2} è 1.25. La seguente tabellina forse aiuta:

t	e_t	i_t
-2	*0*	0
-1	*0*	0
0	*1*	$i_{-1} - 0.5i_{-2} + e_0 + 0.75e_{-1} = 1$
1	*0*	$i_0 - 0.5i_{-1} + e_1 + 0.75e_0 = 1.75$
2	*0*	$i_1 - 0.5i_0 + e_2 + 0.75e_1 = 1.25$
3	*0*	$i_2 - 0.5i_1 + e_3 + 0.75e_2 = 0.375$
$\vdots$	$\vdots$	$\vdots$

Chi ha la pazienza di andare avanti fino a 20 periodi potrà costruirsi un grafichetto come quello mostrato in figura 2.9, da cui si vede abbastanza chiaramente che la funzione, dopo 8 periodi, riproduce (in modo ovviamente attenuato) più o meno la stessa dinamica. Di conseguenza, sarà lecito aspettarsi che una realizzazione di questo processo evidenzierà degli andamenti ciclici di ampiezza 8 periodi (circa).

Da quanto abbiamo detto fin qui, uno potrebbe essere indotto a pensare che la rappresentazione di Wold e la funzione di risposta di impulso siano la stessa cosa. Non è proprio vero: la funzione di risposta di impulso si può calcolare sempre, anche se il processo non fosse stazionario in covarianza. Se però lo è, allora la funzione di risposta di impulso coincide coi coefficienti della rappresentazione di Wold.

2.7 Stima dei modelli ARMA

Fino ad ora abbiamo fatto finta che il processo stocastico che sovrapponiamo ai dati per interpretarli fosse governato da parametri noti. Se questi ultimi noti non sono (e non lo sono mai), si possono utilizzare delle loro stime. La tecnica di base per la stima dei parametri di un processo ARMA è la massima verosimiglianza. Di solito si assume che il processo sia normale, cosicché la forma della funzione di densità delle osservazioni è nota e trattabile.

Può essere utile richiamare brevemente cosa si intende per funzione di verosimiglianza. La verosimiglianza è la funzione di densità del campione, calcolata nel punto corrispondente al campione osservato. Essa dipenderà da un vettore ψ di parametri incogniti, che ne determinano la forma. Per questo la scriviamo $L(\psi)$. Massimizzando questa funzione rispetto a ψ si ottiene la stima di massima verosimiglianza.

Esempio 2.7.1 *Se lanciamo una moneta, e otteniamo "testa", abbiamo una realizzazione di una variabile casuale che assume valore 1 (testa) con probabilità p e 0 con probabilità $1-p$; in questo caso, la verosimiglianza è la probabilità di osservare il campione che si è effettivamente osservato, dato il parametro $p \in [0,1]$, vale a dire $L(p) = p$; la stima di massima verosimiglianza in questo esempio è 1. Se avessimo ottenuto "croce", la verosimiglianza avrebbe assunto la forma $L(p) = 1-p$, e la stima di massima verosimiglianza sarebbe stata 0.*

Se lanciamo 2 monete, avremmo i seguenti possibili esiti:

Campione	*L(p)*	*Punto di massimo*
TT	p^2	*1*
TC	$p(1-p)$	*0.5*
CT	$(1-p)p$	*0.5*
CC	$(1-p)^2$	*0*

eccetera.

Quando osserviamo una realizzazione di un processo stocastico (o, per meglio dire, una serie storica che possiamo pensare come tale) $x_1,\ldots,x_T$, la funzione di verosimiglianza non è altro che la funzione di densità congiunta della parte di processo osservata, ossia la funzione di densità marginale del vettore aleatorio $(x_1,\ldots,x_T)$, calcolata nei valori osservati; nel caso di un processo ARMA del tipo

$$A(L)x_t = \mu + C(L)\epsilon_t$$

essa dipenderà dal vettore di parametri $\psi = \{\mu;\varphi_1\ldots\varphi_p;\theta_1\ldots\theta_q;\sigma^2\}$.

Se supponiamo (come generalmente si fa) che il processo sia gaussiano, la funzione di verosimiglianza non è che la funzione di densità di una normale multivariata:

$$L(\psi) = f(x;\psi) = (2\pi)^{-\frac{T}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-k)'\Sigma^{-1}(x-k)\right\}$$

dove x è il vettore $(x_1,\ldots,x_T)$ delle T osservazioni; k e Σ sono i suoi momenti primi e secondi, che dipendono da ψ. Ad esempio, l'elemento ij della matrice

Σ non è che l'autocovarianza di ordine $|i-j|$ la quale, come sappiamo, è una funzione dei parametri del processo ARMA.

È possibile dimostrare che gli stimatori di massima verosimiglianza di processi ARMA gaussiani sono consistenti, asintoticamente normali ed asintoticamente efficienti. Inoltre, sotto condizioni piuttosto blande, le proprietà di consistenza e normalità asintotica vengono conservate anche quando la vera distribuzione del processo non sia normale (si parla in questo caso di stime di quasi-massima verosimiglianza).

Da un punto di vista teorico, è detto tutto. Da un punto di vista pratico, i problemi sono appena all'inizio. Innanzitutto, va detto che come al solito non si lavora sulla funzione $L(\psi)$, ma sul suo logaritmo

$$\log L(\psi) = l(\psi) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\left[\log|\Sigma| + (x-k)'\Sigma^{-1}(x-k)\right]$$

ma questo è irrilevante. I problemi principali sono tre:

1. in primo luogo, il sistema di equazioni che risulta uguagliando a 0 il vettore dello *score* $s(\psi) = \frac{\partial l(\psi)}{\partial \psi}$ è non lineare, ed in generale non si riesce a risolvere analiticamente, per cui non sono disponibili espressioni che permettano di calcolare gli elementi di $\hat{\psi}$ come semplici funzioni dai dati;

2. non è noto l'ordine dei polinomi $A(L)$ e $C(L)$ adatti a rappresentare in modo adeguato il processo di cui pensiamo x_t sia una realizzazione;

3. per calcolare $L(\psi)$ bisogna conoscere la matrice Σ; ora, la matrice Σ è una matrice $T \times T$, e quando T è grande (già nell'ordine delle decine, ma i campioni di serie storiche possono assumere dimensioni dell'ordine di *decine di migliaia* di osservazioni) anche il semplice calcolo dei suoi elementi in funzione di ψ è un problema non da poco, per non parlare del suo determinante o della sua inversa.

2.7.1 Tecniche numeriche

Il primo problema meriterebbe una disamina approfondita, ma veramente questa non è la sede. Dico solo che il problema cessa di essere tale se si dispone delle risorse di calcolo adeguate. Esistono algoritmi numerici che permettono di trovare il massimo delle funzioni di verosimiglianza senza grandi difficoltà, e tutti i pacchetti econometrici ne sono forniti.

Banalizzando al massimo, si può dire che questi algoritmi servono a trovare il massimo di una funzione una volta che si sia in grado di calcolare questa funzione, e la sua derivata prima, per un punto qualunque dello spazio parametrico. Se la funzione è liscia, la sua derivata (il *gradiente*) è 0 sul massimo.

Più o meno, si procede così:

1. Si parte da un punto preso "a caso" ψ_0 ;

2. si calcola lo *score* $s(\psi_0)$;

3. se $s(\psi_0)$ è "piccolo", stop. Se no, si calcola una direzione $d(s(\psi_0))$ nella quale spostarsi;

4. si calcola $\psi_1 = \psi_0 + d(s(\psi_0))$;

5. si rimpiazza ψ_0 con ψ_1 e si riparte dal punto 2.

Ci sono molti algoritmi di questo tipo: sostanzialmente, ognuno calcola a modo suo il vettore di direzione $d(s(\theta_0))$, cercando di far sì che $\ell(\psi_1) > \ell(\psi_0)$ ad ogni iterazione, cosicché prima o poi si arriva sul massimo.

Qualcuno avrà notato che l'algoritmo di cui prima dipende in modo cruciale da quel che succede al punto 3. Finché $s(\psi_0)$ non è "piccolo", l'algoritmo va avanti. Indi, dobbiamo decidere cosa vuol dire che un vettore è "piccolo", per non correre il rischio di trasformare la nostra CPU in una palla di fuoco.
Questo è uno di quesi casi in cui una domanda è semplice solo in apparenza: ci sono vari modi di rispondere e come al solito, nessuno è "giusto". Uno dei criteri più comuni è decidere che $s(\psi_0)$ è "zero" quando nessuno dei suoi elementi eccede in valore assoluto una soglia prefissata, tipo 1.0E-07 o giù di lì. Ma non è l'unico e non è necessariamente il migliore.

Di solito, le funzioni di verosimiglianza che si hanno in questi casi sono piuttosto lisce e la possibilità di avere massimi multipli è trascurabile. Di conseguenza, basta essere in grado di calcolare le derivate prime della verosimiglianza per qualunque vettore ψ per essere in grado arrivare — prima o poi — sul massimo. Inoltre, la letteratura si è scatenata per almeno vent'anni sul problema particolare delle funzioni di verosimiglianza di modelli ARMA gaussiani, cosicché gli algoritmi che oggi si trovano precotti nei pacchetti econometrici sono particolarmente stabili ed efficienti.

2.7.2 Scelta degli ordini dei polinomi

Per quanto riguarda il secondo problema, la scelta dell'ordine dei polinomi $A(L)$ e $C(L)$ è un'operazione di alto artigianato, che richiede conoscenze teoriche, esperienza ed occhio.

Il modo in cui si procede di solito[12] è basato sul fatto che esistono dei criteri (ossia, delle procedure di test) che ci permettono di stabilire se un processo è un *white noise* o possiede della persistenza; il più diffuso è il cosiddetto test di **Ljung-Box**, che è basato sul fatto che in grandi campioni le autocovarianze campionarie tendono a 0 nel caso di un *white noise*: la statistica test vera e propria è

$$LB(p) = T(T+2)\sum_{i=1}^{p}\frac{\hat{\rho}_i^2}{T-i};$$

si noterà che essa è sostanziamente una somma ponderata dei quadrati delle autocorrelazioni campionarie fino all'ordine p. Più queste sono piccole, più il test viene piccolo; l'ipotesi nulla è che il vero valore di tutte le autocorrelazioni fino all'ordine p sia 0, e i valori critici sono quelli della χ_p^2.

[12] Questa è una descrizione semplificata in modo quasi insolente di quella che di solito viene descritta come metodologia di **Box-Jenkins**, per cui rimando alla letteratura.

L'idea che, sotto le ipotesi di ergodicità e stazionarietà, le autocorrelazioni campionarie siano stimatori consistenti di quelle teoriche può essere sfruttata anche in modo più generale. Come abbiamo già visto, infatti, ci sono delle relazioni ben precise fra ordine dei polinomi e autocorrelazioni. Dall'esame delle autocorrelazioni campionarie si può fare un'ipotesi di partenza sugli ordini dei polinomi. Se, ad esempio, si nota che le autocorrelazioni campionarie si interrompono bruscamente al di là di un certo ordine q, si può pensare di usare un modello MA(q), le cui autocorrelazioni teoriche hanno la stessa caratteristica. Se invece le autocorrelazioni digradano dolcemente, forse è meglio un processo AR. Questa fase è nota nella letteratura statistica come fase di **identificazione**. Questo termine genera a volte un po' di confusione, perché normalmente in econometria la parola "identificazione" vuol dire un'altra cosa[13].

In questa fase, si adoperano a volte anche statistiche note come **autocorrelazioni parziali** (le quali, in pratica, non si usano che a questo scopo). Definire le autocorrelazioni parziali rigorosamente è un po' macchinoso. Si fa prima a dire come si calcolano: l'autocorrelazione parziale di ordine p si calcola facendo una regressione di y_t su una costante e $y_{t-1} \dots y_{t-p}$. Il coefficiente associato a y_{t-p} che risulta è l'autocorrelazione parziale di ordine p. Queste grandezze si interrompono bruscamente nel caso di modelli AR puri, e scendono gradualmente nel caso di modelli MA puri.

Mi permetto una piccola tirata polemica: ancora oggi, chi insegna queste cose è portato ad ammorbare i propri studenti con queste tecniche un po' da rimedio della nonna per scegliere p e q facendo elaborate considerazioni sulla forma delle funzione di autocorrelazione e tacendo pudicamente il fatto che nella maggioranza dei casi che si incontrano in pratica o uno ha un occhio molto allenato oppure non ci si capisce niente. Il fatto è che queste tecniche sono state inventate in un epoca in cui un computer era una rarità da scienziati, e fare una stima di un ARMA era difficile e costoso, per cui tante prove non si potevano fare ed era essenziale avere un'idea il più possibile precisa di possibili valori di p e q prima di tentare la stima. Oggi stimare un modello ARMA è ridicolmente facile, e l'arte dell'interpretazione dei correlogrammi la lasciamo agli appassionati del genere *vintage*.

Una volta fatta la stima, si controlla se i 'residui' sono *white noise*, quasi sempre col test di Ljung-Box o con statistiche equivalenti. Un'altra classe di statistiche che si usano in questo contesto sono i cosiddetti *criteri di informazione*, come ad esempio quello di **Akaike** (spesso abbreviato in AIC) o quello di **Schwartz** (spesso abbreviato in BIC); l'uso di queste statistiche è motivato con concetti presi dalla teoria dell'informazione, ma mi contento di rinviare alla letteratura per i dettagli. Qui mi basta dire che fra due modelli, quello "migliore" dovrebbe avere un indice AIC o BIC più basso, in quanto tutti questi

[13] Ricordo brevemente cosa si intende per identificazione di un modello nell'accezione comune in econometria: un modello econometrico si dice sotto-identificato se esiste più di una rappresentazione dei dati coerente con ciò che si osserva. In pratica, non è possibile decidere sulla base dei dati se sia più giusta la rappresentazione A o la rappresentazione B; in questi casi, si usa l'espressione "equivalenza osservazionale". Se il modello è parametrico (come nella maggior parte dei casi), esso è identificato se la funzione di verosimiglianza ha un solo massimo assoluto; di conseguenza, una condizione necessaria per l'identificazione è la non singolarità dell'Hessiano nel punto di massimo. L'identificazione è, chiaramente, a sua volta condizione necessaria per l'esistenza di uno stimatore consistente.

criteri possono essere scritti nella forma

$$C = -2L(\theta) + c(k,T)$$

dove k è il numero di parametri stimati e T è l'ampiezza campionaria; la funzione $c(k,T)$ è crescente in k, per cui a parità di verosimiglianza viene scelto il modello più parsimonioso. Ad esempio, per il criterio di Schwartz, $c(k,T) = k\log(T)$.

In questa fase, è importante non scegliere degli ordini dei polinomi troppo alti, per il cosiddetto problema dei **fattori comuni**: dato un processo ARMA(p,q) della forma

$$A(L)x_t = C(L)\epsilon_t$$

è chiaro che, applicando l'operatore $(1-\beta L)$ ad entrambi i lati dell'uguaglianza, la relazione continua ad essere vera. Chiamiamo

$$A_\beta(L) = (1-\beta L)A(L)$$

e

$$C_\beta(L) = (1-\beta L)C(L)$$

e quindi

$$A_\beta(L)x_t = C_\beta(L)\epsilon_t. \tag{2.13}$$

Si noti che la rappresentazione di Wold basata sul modello ARMA($p+1$,$q+1$) è assolutamente la stessa di quella basata sul modello ARMA(p,q), perché i fattori $(1-\beta L)$ si semplificano. Il processo x_t, quindi, ha una rappresentazione ARMA($p+1$,$q+1$) del tutto equivalente. Poiché questo è vero per qualunque valore di β, è ovvio che il modello non è identificato (nel senso econometrico; vedi nota 13), perché ogni valore di β è equivalente dal punto di vista osservazionale e quindi non è stimabile (il valore della funzione di verosimiglianza è lo stesso per qualunque β, e quindi non c'è un massimo unico: di massimi ce ne sono infiniti, uno per ogni valore di β.).

Detta in un altro modo, esistono infiniti polinomi $A_\beta(L)$ e $C_\beta(L)$ che conducono alla stessa rappresentazione di Wold, e quindi alla stessa funzione di autocovarianza. L'equivalenza osservazionale nasce esattamente dal fatto che le autocovarianze campionarie non ci pemettono di discriminare fra valori diversi di β.

Faccio un esempio che forse è meglio: che tipo di processo è

$$y_t = 0.5y_{t-1} + \epsilon_t - 0.5\epsilon_{t-1}?$$

Facile, direte: è un ARMA(1,1). Giusto. Però è anche un *white noise*; infatti

$$y_t = \frac{1-0.5L}{1-0.5L}\epsilon_t = \epsilon_t.$$

In pratica abbiamo scritto un *white noise* come un ARMA(1,1). Quest'ultima rappresentazione è ridondante, ma non *sbagliata*. La cosa importante da notare è che il numero 0.5 è del tutto irrilevante: avrei potuto usare 0.7, 0.1 o che so io. Di rappresentazioni "non sbagliate" ce ne sono infinite.

Da un punto di vista pratico, modellare un ARMA(p,q) con un ARMA($p+1$,$q+1$) porta ogni sorta di problemi. Intanto, perché l'algoritmo numerico fa fatica a convergere (e non sorprende, visto che non c'è un massimo unico). In secondo luogo, perché (anche ammesso che la convergenza alla fine avvenga), il punto di massimo che troviamo è solo una delle infinite rappresentazioni possibili del modello[14].

Di solito, ci si accorge di questa situazione dal fatto che gli errori standard stimati dei coefficienti esplodono; questo succede perché, tentando di stimare un modello non identificato, la matrice di informazione che viene stimata tende ad una matrice singolare. Invertendola, vengono fuori numeri giganteschi per la matrice varianze-covarianze dei coefficienti.

2.7.3 Calcolo della verosimiglianza

Il terzo problema è più intrigante: bisogna, in sostanza, scrivere la funzione di verosimiglianza con un'espressione alternativa che non richieda il calcolo di matrici di dimensione sproporzionata. Questo argomento è stato studiato a fondo, ed è bene rinviare alla letteratura per una discussione esauriente, ma in questa sede voglio illustrare una tecnica piuttosto interessante, che va sotto il nome di **fattorizzazione sequenziale**.

Per illustrare questa tecnica, sarà utile partire dalla definizione di probabilità condizionata, che è

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

da cui

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Se applichiamo questa regola alla funzione di densità di una variabile casuale doppia, otteniamo

$$f(x,y) = f(y|x)f(x) = f(x|y)f(y) \tag{2.14}$$

Il giochino può essere ripetuto anche con una variabile casuale tripla, ottenendo

$$f(x,y,z) = f(x|y,z)f(y,z) = f(y|x,z)f(x,z) = f(z|x,y)f(x,y) \tag{2.15}$$

Mettendo assieme le due relazioni (2.14) e (2.15), è chiaro che si può scrivere, ad esempio,

$$f(x,y,z) = f(z|x,y)f(x,y) = f(z|x,y)f(y|x)f(x)$$

e quindi una funzione di densità congiunta di n variabili casuali può essere scritta

$$f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i|x_1,\ldots,x_{i-1})$$

[14] È vero che tutte queste rappresentazioni hanno la stessa rappresentazione di Wold, per cui le previsioni a cui portano e le risposte di impulso che generano sono identiche, ma abbiamo il problema che qualunque tipo di test in questo caso ci è precluso. Infatti, il punto di massimo che troviamo è soltanto uno degli infiniti possibili, e quindi l'Hessiano della funzione di verosimiglianza è singolare. Poiché tutte le statistiche test sono basate in qualche modo sulla curvatura della funzione di verosimiglianza, è chiaro che i test non si possono fare

così da trasformare una funzione di molte variabili in una produttoria di funzioni di una variabile. Si noti che, quando le variabili x_i sono indipendenti, $f(x_i|x_1,\ldots,x_{i-1}) = f(x_i)$ e (come è noto), la funzione di densità congiunta è il prodotto delle marginali.

Poiché la funzione di verosimiglianza non è che la funzione di densità del campione, questa stessa scomposizione può essere applicata alla funzione di verosimiglianza, e sarà tanto più utile quanto più semplici sono le espressioni delle densità condizionali. Inoltre, la caratteristica dei modelli ARMA per cui

$$y_t = E(y_t|\Im_{t-1}) + \epsilon_t$$

(vedi equazione (2.10)) fa sì che, condizionatamente a $\Im_{t-1}$, la distribuzione di y_t sia la stessa di ϵ_t, e quindi la verosimiglianza può essere scritta in modo equivalente anche in termini degli errori di previsione ad un passo in avanti anziché delle y_t.

Fare un esempio risulta particolarmente facile nel caso di processi AR puri. In questo caso, infatti, se il processo è

$$y_t = \mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \epsilon_t$$

e ϵ_t è normale, la funzione di densità $f(y_t|\Im_{t-1})$ è semplicemente una normale:

$$y_t|\Im_{t-1} \sim N\left(\mu + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p}, \sigma^2\right)$$

Nel caso particolare di un processo AR(1) si avrà che

$$y_t|\Im_{t-1} \sim N\left(\mu + \varphi y_{t-1}, \sigma^2\right)$$

e quindi $f(y_t|\Im_{t-1}) = f(y_t|y_{t-1})$: l'unico elemento di $\Im_{t-1}$ che conta è l'ultimo. Di conseguenza, la verosimiglianza potrà essere scritta come

$$L(\mu,\varphi,\sigma^2) = f(y_1) \times \prod_{t=2}^{T} f(y_t|y_{t-1})$$

Passando ai logaritmi si ottiene

$$\begin{aligned} l(\mu,\varphi,\sigma^2) &= \log f(y_1) + \tilde{l}(\mu,\varphi,\sigma^2) = \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{(y_t - \mu - \varphi y_{t-1})^2}{\sigma^2}\right) \\ &= \log f(y_1) - \frac{1}{2}\sum_{i=2}^{T}\left(\log 2\pi + \log\sigma^2 + \frac{e_t^2}{\sigma^2}\right), \end{aligned}$$

dove ho usato la notazione $e_t = y_t - E(y_t|\Im_{t-1})$.

Se il primo addendo fosse zero, il resto sarebbe uguale ad una normalissima funzione di log-verosimiglianza per un modello lineare in cui la variabile dipendente è y_t, la variabile esplicativa è y_{t-1} (più l'intercetta) e il termine di disturbo è normale con media 0 e varianza σ^2.

Sappiamo già che per tale modello gli stimatori di massima verosimiglianza coincidono con quelli OLS, cosicché se non fosse per il primo termine potremmo usare semplicemente la tecnica OLS. Tuttavia, per campioni molto

grandi, il peso del primo addendo nel determinare la verosimiglianza totale diventa irrilevante: a differenza del secondo, infatti, il primo addendo non cresce all'aumentare di T. Le stime OLS (che massimizzano $\tilde{l}(\mu, \varphi, \sigma^2)$) tendono quindi a quelle di massima verosimiglianza, e asintoticamente ne condividono le proprietà[15].

Questo ragionamento fila anche per modelli AR(p): in questo caso, il primo elemento della verosimiglianza diventa $\log f(y_1, \ldots, y_p)$, ma l'argomento rimane invariato. È peraltro vero che, sotto condizioni abbastanza generali, lo stimatore OLS dei parametri di un processo autoregressivo stazionario è uno stimatore consistente e asintoticamente normale anche se il processo ϵ_t non è gaussiano; quindi, anche se non è corretto vedere lo stimatore OLS come asintoticamente equivalente a quello di massima verosimiglianza, comunque non è improprio adoperarlo. In questo caso, che va sotto il nome di *regressione dinamica*, le proprietà asintotiche degli stimatori OLS possono essere provate facendo ricorso a teoremi limite che solitamente in un corso di Econometria si affrontano da qualche altra parte, e quindi non riporto qui.

Nel caso di modelli in cui sia presente una parte a media mobile, il discorso si complica solo di poco, se manteniamo l'ipotesi di normalità. Infatti, come ho già fatto rilevare al paragrafo 2.6.1, in un modello ARMA il *white noise* che guida il processo può essere interpretato come la differenza fra y_t e il suo valore atteso condizionale a $\Im_{t-1}$ (vedi eq. (2.10)). Di conseguenza, se assumiamo che la distribuzione di y_t condizionale a $\Im_{t-1}$ sia una normale, se ne deduce che gli errori di previsione ad un passo sono una sequenza di normali incorrelate (e quindi indipendenti) a media 0 e varianza costante, per cui la verosimiglianza può essere calcolata molto semplicemente utilizzando gli errori di previsione.

2.8 In pratica

Per far capire come funziona il tutto, facciamo il caso di voler impostare un modello per una serie storica di cui abbiamo già parlato nell'introduzione, e cioè la produzione industriale USA[16]. Diciamo che la serie a nostra disposizione è quella disegnata in figura 2.10.

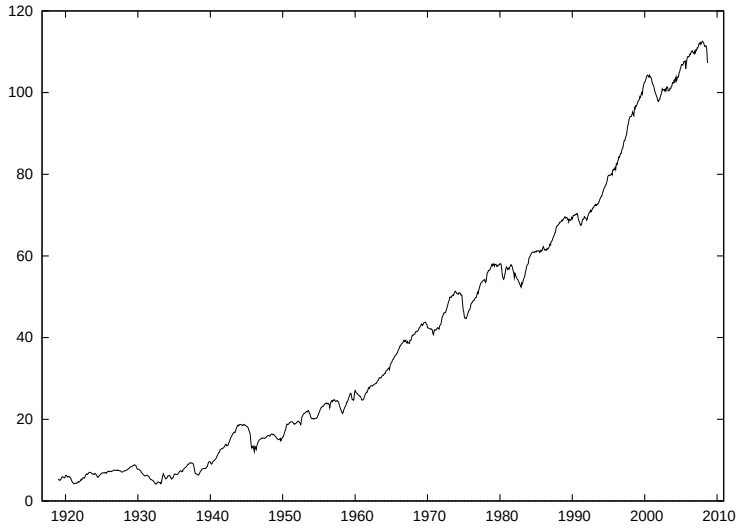
Tanto per cominciare, noterete che la serie è molto lunga: va dal gennaio 1921 al settembre 2008. Bene, dirà qualcuno, obbedendo al condizionamento pavloviano del campione che più è grande, meglio è. Male, dico io. Ricordiamoci che quel che stiamo cercando è un processo stocastico di cui è pensabile stiamo osservando una realizzazione.

Ora, il processo stocastico "vero" che ha generato questa serie (ammesso che esista) è senz'altro qualcosa che ha cambiato profondamente i suoi con-

[15] Si parla, in casi come questo, di verosimiglianza **condizionale**, per indicare che la funzione di verosimiglianza che stiamo usando considera le prime p osservazioni come fisse, e quindi fa riferimento alla distribuzione di $y_{p+1} \ldots y_T$ *condizionale a* $y_1 \ldots y_p$. Esistono anche tecniche per massimizzare la cosiddetta verosimiglianza **esatta**, cioè quella che tiene conto anche della distribuzione delle prime p osservazioni, ma asintoticamente non fa differenza.

[16] Per chi è pratico di queste cose, preciso fin da subito che ho fatto il furbo e ho usato la serie destagionalizzata. Ma chi è pratico di queste cose immagina facilmente il perché.

Figura 2.10: Indice destagionalizzato della produzione industriale negli USA (mensile dal 1921)



notati durante la nostra finestra di osservazione. Nell'arco di tempo descritto dai nostri dati ci sono il delitto Matteotti, l'invenzione della penna biro, i film di Totò, *Voodoo Chile* e Google. Risulta un po' ardito postulare l'esistenza di una rappresentazione dei dati che attraversa indenne tutto questo ed è buona tanto allora quanto oggi. Nei termini del primo capitolo, potremmo dire con una certa tranquillità che il "vero" processo stocastico che ha generato i dati non è stazionario. Se proprio vogliamo ingabbiare la serie in un processo stazionario, conviene accorciare il campione. In questo caso, gli economisti amano dire che escludiamo i cosiddetti *break strutturali*[17]; si noti che questo ragionamento si può fare *senza neanche guardare i dati*.

Con procedura del tutto arbitraria (tanto è un esempio), decido che il mondo in cui viviamo oggi è cominciato nel gennaio 1984. Già che ci siamo, decidiamo di lavorare non sul numero indice vero e proprio, ma sul suo logaritmo. Questa è una procedura molto diffusa, e serve a far sì che si possa dare un'interpretazione più naturale ai numeri di cui è composta la serie, visto che le sue differenze prime sono più o meno variazioni percentuali[18]. Un'altra cosa che vale la pena di fare è escludere dalle nostre considerazioni per il momento le ultime tre osservazioni; vedremo poi il perché.

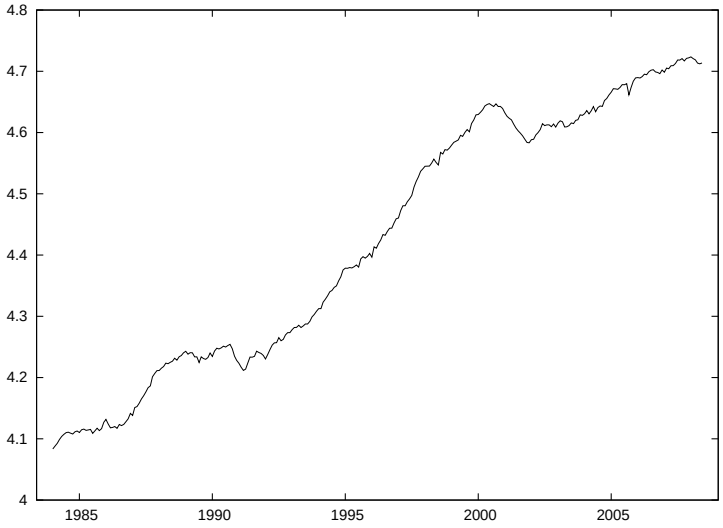
Il risultato è la serie in figura 2.11, che è abbastanza lunga da consentirci di dire qualcosa di interessante (294 osservazioni), ma al contempo ci racconta una storia ragionevolmente omogenea.

Possiamo, a questo punto, sostenere di osservare una realizzazione di un processo stazionario? Di nuovo, la risposta è "probabilmente no". In questo caso, però, il problema non nasce dalla disomogeneità del campione, ma dal

[17] È peraltro vero che esistono metodi di lavorare con serie storiche con *break* strutturali al loro interno, ma questi metodi sono ancora troppo esoterici per parlarne in questa dispensa.

[18] Ricordo che $\log(y_t) - \log(y_{t-1}) = \log(1 + \frac{\Delta y_t}{y_{t-1}}) \simeq \frac{\Delta y_t}{y_{t-1}}$

Figura 2.11: Logaritmo della produzione industriale negli USA (mensile)



fatto che la serie in figura 2.11 presenta un chiaro trend crescente, che evidentemente ci preclude di pensare che il processo sia stazionario. Si può pensare a un processo stazionario intorno a un trend deterministico, ossia ad una cosa del tipo $Y_t = (a + b \cdot t) + u_t$, dove u_t è un qualche processo ARMA. Oltretutto, questa non sarebbe nemmeno un'idea irragionevole, perché il parametro b potrebbe essere interpretato come il tasso esogeno di progresso tecnico di lungo periodo. Tuttavia, questa idea non regge, per motivi che spiegherò nel capitolo 3 (scusate). Anticipo soltanto che il problema fondamentale sta nel fatto che, anche togliendo via un trend deterministico, questa serie è troppo persistente per poter dire che il processo che l'ha generata è stazionario.

Una possibilità alternativa è quella di trasformare la serie in modo tale da poterla ancora interpretare, ma nel frattempo eliminare il problema. In questo caso, ci caviamo d'impaccio con una differenziazione e cosideriamo $y_t = 100 \cdot \Delta Y_t$, che potete ammirare in figura 2.12 ed è, come ho accennato prima, più o meno il tasso di variazione percentuale della produzione industriale rispetto al mese precedente.

La figura 2.13, invece, mostra i correlogrammi totale e parziale. Le due lineette tratteggiate orizzontali che circondano il correlogramma vero e proprio rappresentano la costante $\pm 1.96/\sqrt{T}$, dove T è l'ampiezza campionaria: visto che abbiamo 294 osservazioni, la costante è circa 0.11. Queste lineette vengono spesso inserite nei correlogrammi per rendere immediato il seguente ragionamento: le autocorrelazioni campionarie $\hat{\rho}_k$ sono stimatori consistenti delle vere autocorrelazioni ρ_k. Se per $\rho_k = 0$, allora si può dimostrare che $\sqrt{T}\hat{\rho}_k \xrightarrow{\text{d}} N(0,1)$. Di conseguenza, l'intervallo $\pm 1.96/\sqrt{T}$ è l'intervallo di accettazione al 95% del test per l'ipotesi $\rho_k = 0$; in pratica, le autocorrelazioni fuori banda sono "statisticamente significative". Non possiamo fare a meno di osservare che di autocorrelazioni significative ce ne sono almeno cinque o sei, per cui possiamo ragionevolmente escludere l'ipotesi che y_t non abbia

Figura 2.12: Variazione percentuale della produzione industriale

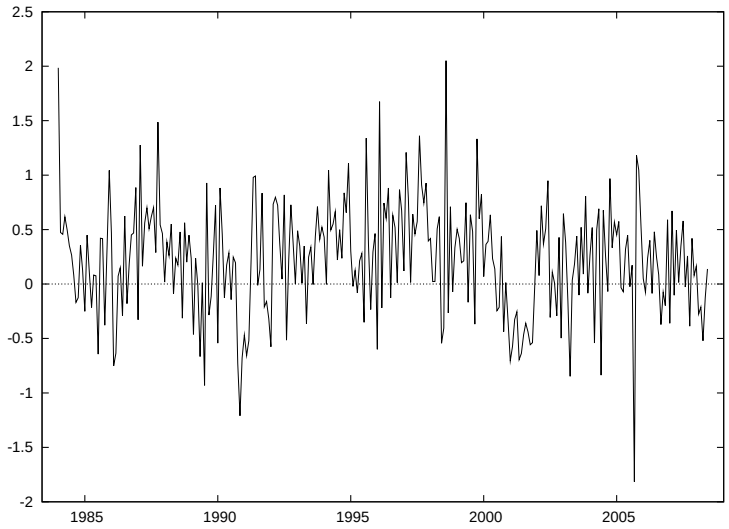
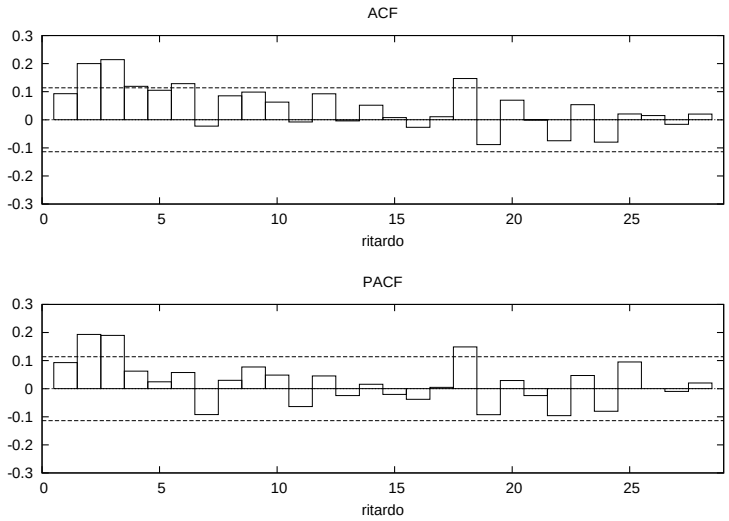


Figura 2.13: Variazione percentuale della produzione industriale – correlogramma parziale e totale



persistenza (cosa che d'altronde un occhio allenato vede anche dalla figura 2.12). In più, calcolando il test di Ljung-Box fino a 12 ritardi troviamo un valore della statistica test di 49.9704; considerando che sotto la nulla questa è una realizzazione di una χ^2 con 12 gradi di libertà, direi che la nulla si può rifiutare baldanzosamente.

Di persistenza, pertanto, ce n'è. Se, quindi, un modello ARMA può essere una buona idea, il correlogramma non ci dice con chiarezza quali siano gli ordini dei polinomi $A(L)$ e $C(L)$ (quelli che abbiamo fin qui chiamato p e q) da utilizzare.

Visto che il correlogramma parziale si ferma a 3, una possibilità è quella di un AR(3), ma siamo piuttosto al limite, per cui può essere il caso di provare più di una soluzione. In questo caso, possiamo anche optare per una strategia "a tappeto". Stabiliamo un ordine massimo per p e q (diciamo 5) e proviamo tutte le combinazioni possibili.

Tabella 2.1: Modelli ARMA per la produzione industriale USA: criteri di informazione

Akaike (AIC)

AR \ MA	0	1	2	3	4	5
0	446.06	446.15	439.29	432.49	430.84	430.48
1	445.42	429.84	425.46	427.10	428.69	430.64
2	435.78	426.33	427.27	428.89	430.52	430.42
3	426.36	426.69	428.67	428.62	**421.66**	423.66
4	427.06	428.66	428.71	421.90	423.65	425.60
5	428.84	430.66	426.39	423.69	425.64	426.33

Schwartz (BIC)

AR \ MA	0	1	2	3	4	5
0	453.43	457.20	454.03	450.91	452.94	456.26
1	456.47	444.57	**443.88**	449.20	454.47	460.10
2	450.51	444.75	449.37	454.68	459.99	463.57
3	444.78	448.79	454.45	458.09	454.81	460.49
4	449.16	454.45	458.18	455.06	460.49	466.12
5	454.63	460.13	459.54	460.52	466.15	471.82

Hannan-Quinn (HQC)

AR \ MA	0	1	2	3	4	5
0	449.01	450.57	445.19	439.87	439.69	440.80
1	449.84	435.74	**432.83**	435.95	439.01	442.44
2	441.68	433.71	436.12	439.22	442.32	443.70
3	433.73	435.54	439.00	440.42	434.94	438.41
4	435.91	438.99	440.51	435.18	438.40	441.83
5	439.17	442.46	439.67	438.44	441.86	445.32

È naturale, in questo contesto, usare i criteri di informazione per paragonare fra loro i vari modelli stimati. Qui, ne useremo tre, e cioè il criterio di Akaike (AIC), quello di Schwartz (BIC) e quello di Hannan e Quinn (HQC). Come è noto, la scelta si fa prendendo il modello per cui il criterio è minore.

Ora, date un'occhiata alla tabella 2.1. Per ognuno dei tre criteri, il numero per riga indica l'ordine AR e quello per colonna l'ordine MA; ho evidenziato col grassetto il minimo per ogni tabella. Come si vede, la scelta non è univoca: se il BIC e lo HQC indicano un ARMA(1,2), l'AIC si concede un faraonico ARMA(3,4). Questo non è un caso: per come sono costruiti i criteri, si ha di solito che l'AIC tende a essere piuttosto permissivo in termini di numero totale di parametri, mentre il BIC ha la tendenza opposta. Lo HQC generalmente opta per una via di mezzo. È peraltro vero che il criterio di Akaike, che ha una sua importanza storica perché è stato il primo di tutta la stirpe, ha il difetto, rispetto agli altri, di non essere *consistente*: si può dimostrare che la probabilità che esso selezioni il modello "giusto" non va ad 1 asintoticamente, come accade per gli altri due; di conseguenza, oggi è un po' in discredito.

In ogni caso, non c'è un vincitore chiaro. Quindi, guardiamo questi due modelli un po' più da vicino. Entrambi esibiscono un correlogramma degli errori di previsione a un passo (quelli che in un contesto di regressione chiameremmo i residui) assolutamente piatto e il relativo test di Ljung-Box accetta la nulla senza problemi. Non ve li faccio neanche vedere, il succo è che ognuno di questi modelli cattura adeguatamente la persistenza che c'è. Bisogna vedere quale dei due lo fa nel modo più efficace.

Tabella 2.2: Modello ARMA(1,2)

	Coefficient	Std. Error	z-stat	p-value
const	0.2253	0.0564	3.9943	0.0001
ϕ_1	0.8034	0.0942	8.5251	0.0000
θ_1	−0.7865	0.1033	−7.6154	0.0000
θ_2	0.1779	0.0680	2.6173	0.0089

Mean dependent var	0.221113	S.D. dependent var	0.514050
Mean of innovations	−0.004798	S.D. of innovations	0.490280
Log-likelihood	−207.7288	Akaike criterion	425.4577
Schwarz criterion	443.8756	Hannan–Quinn	432.8335

		Real	Imaginary	Modulus	Frequency
AR					
	Root 1	1.2447	0.0000	1.2447	0.0000
MA					
	Root 1	2.2106	−0.8573	2.3710	−0.0589
	Root 2	2.2106	0.8573	2.3710	0.0589

Il modello ARMA(1,2) (mostrato in tabella 2.2) è il più parsimonioso dei due; certo, non è un gran che come capacità di fittare i dati: la stima di σ (lo scarto quadratico medio degli errori di previsione a un passo) è circa 0.49, che è non è molto più piccola dello scarto quadratico medio della serie osservata (0.514). In altri termini: la dispersione non condizionale di y_t è appena inferiore alla dispersione della distribuzione condizionata a $\Im_{t-1}$. La persistenza qui, c'è, ma prevede un po' pochino.

Il quadro che ci si presenta considerando il modello ARMA(3,4) (tabella 2.3) è, a prima vista, molto diverso. In questo caso, se uno si limitasse a considerare i singoli parametri (cosa che la lettura dei tabulati di regressione

Tabella 2.3: Modello ARMA(3,4)

	Coefficient	Std. Error	z-stat	p-value
const	0.2258	0.0564	4.0051	0.0001
ϕ_1	−0.3520	0.0954	−3.6895	0.0002
ϕ_2	−0.0763	0.1097	−0.6953	0.4869
ϕ_3	0.7950	0.0954	8.3306	0.0000
θ_1	0.4101	0.1075	3.8163	0.0001
θ_2	0.2785	0.1438	1.9369	0.0528
θ_3	−0.5627	0.1358	−4.1448	0.0000
θ_4	0.1691	0.0676	2.5015	0.0124

Mean dependent var	0.221113	S.D. dependent var	0.514050
Mean of innovations	−0.005004	S.D. of innovations	0.477041
Log-likelihood	−201.8311	Akaike criterion	421.6622
Schwarz criterion	454.8144	Hannan–Quinn	434.9386

			Real	Imaginary	Modulus	Frequency
AR						
	Root	1	−0.5780	0.8189	1.0023	0.3478
	Root	2	−0.5780	−0.8189	1.0023	−0.3478
	Root	3	1.2520	0.0000	1.2520	0.0000
MA						
	Root	1	−0.5854	−0.8107	1.0000	−0.3495
	Root	2	−0.5854	0.8107	1.0000	0.3495
	Root	3	2.2492	−0.9246	2.4319	−0.0621
	Root	4	2.2492	0.9246	2.4319	0.0621

ci abitua, purtroppo, a fare) vedrebbe che in maggioranza sembrano essere statisticamente diversi da 0; peraltro, la dispersione degli errori di previsione si riduce un tantino (0.477), anche se non c'è da fare salti di gioia. Tuttavia, ci si può rendere conto che questa stima è un po' farlocca guardando le radici dei polinomi $A(L)$ e $C(L)$ stimati. Fra le radici di $A(L)$ c'è la coppia di numeri complessi $-0.5780 \pm 0.8189i$ (che tra l'altro è sospettosamente vicina a 1 in modulo) cui fa da contraltare la coppia di numeri complessi $-0.5854 \pm 0.8107i$ fra le radici di $C(L)$.

È chiaro che qui siamo in presenza di fattori comuni (cosa siano i fattori comuni lo trovate a pag. 43). Cioè, la stima nella tabella 2.3 è una stima gonfiata con parametri che non servono a nulla. Tant'è che, se andate a vedere le altre radici dei due polinomi, vi accorgete che sono praticamente le stesse del modello ARMA(1,2); in altre parole, quello che vedete nella tabella 2.3 è lo stesso modello della tabella 2.2 sotto mentite spoglie.

Queste considerazioni vengono corroborate in modo abbastanza palese osservando la figura 2.14, che contiene, per i tre modelli esaminati qui, la funzione di risposta di impulso fino a 24 passi, ossia una rappresentazione grafica dei primi 24 coefficienti del polinomi $\frac{C(L)}{A(L)}$. Notate come le risposte di impulso per i modelli ARMA(1,2) e ARMA(3,4) siano praticamente indistinguibili. Ciò implica che, in pratica, questi due modelli sono due possibilità alternative per sintetizzare un modo di persistenza della serie che è sempre lo stesso e le ragioni per cui la rappresentazione ARMA(3,4) sia da ritenersi ridondante sono ovvie.

Figura 2.14: Risposte di impulso

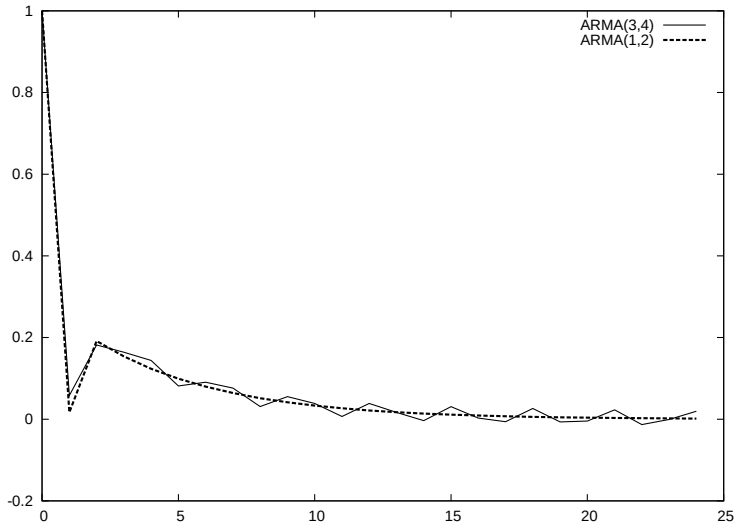
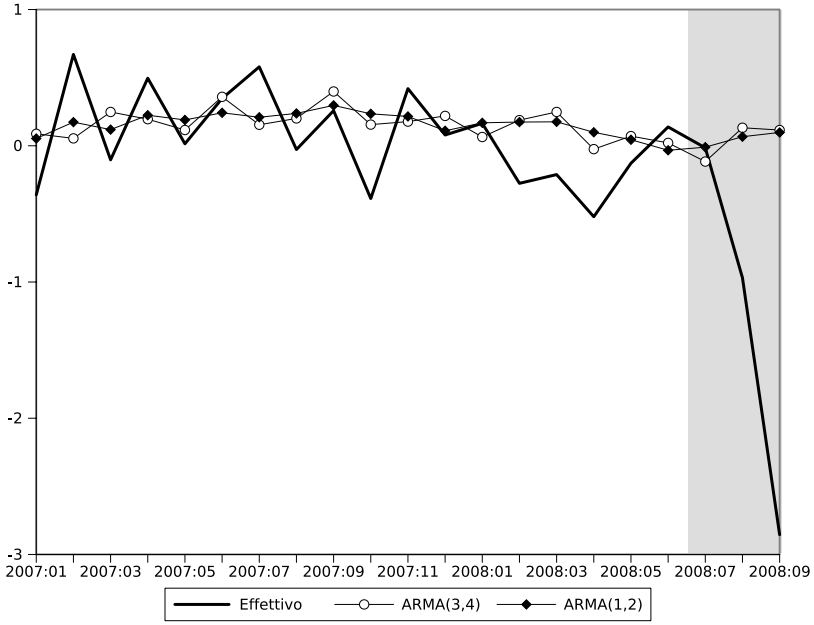


Figura 2.15: Previsioni



Un altro modo di vedere la sostanziale equivalenza dei due modelli è quello di considerare cosa succede utilizzandoli come modelli previsivi: la figura 2.15 mostra le previsioni fino a settembre fatta con i nostri due modelli.

Vi ricordo che ci eravamo tenuti nel cassetto le ultime tre osservazioni: i dati da luglio a settembre 2008 non erano stati usati per la stima. Per cui, quelle che vedete nell'area non ombreggiata sono le previsioni di y_t basate su $\Im_{t-1}$. Quelle nell'area ombreggiata, invece, sono basate sul set informativo che si ferma a giugno.

Qui ci sono alcune osservazioni interessanti da fare: in primo luogo, finché stiamo dentro il campione le previsioni non sono grossolanamente sbagliate, ma non sono nemmeno un gran che. Il fatto che la persistenza della serie sia, pur non trascurabile, poco informativa emerge dal fatto che i modelli forniscono previsioni che non si discostano mai molto dalla media aritmetica della serie (la quale a sua volta potrebbe essere vista come un previsore basato sulla distribuzione non condizionata). Fra l'altro, i due modelli danno previsioni quasi identiche. Questo non deve sorprenderci, avendo appurato che le due rappresentazioni di Wold sono, appunto, quasi identiche.

Dove i due modelli fanno una pessima figura è piuttosto nelle previsioni fuori campione. Come si vede, la produzione industriale subisce un vero e proprio tracollo, assolutamente non previsto dai nostri due modellini. Perché?

Il motivo è semplice. A parte il fatto che in generale i modelli di questo tipo prevedono decentemente solo a pochi passi in avanti, c'è una ragione ben più cogente: la crisi economica mondiale. E la crisi economica mondiale è una cosa che entro certi limiti si poteva prevedere, ma sicuramente non solo col set informativo fornito dalla storia della produzione industriale americana. Ci sarebbe stato bisogno di sapere cose sui mutui *subprime*, sul prezzo del petrolio, sulla bilancia commerciale americana, sulla struttura finanziaria islandese e così via. E magari di aver letto Minsky. Tutto questo, nel nostro set informativo, non c'è (o ce ne sono appena dei pallidi riflessi). Morale: in economia, un modello ARMA può servire, tutt'al più, a fare previsioni a breve in periodi tranquilli. Ma se davvero volete fare gli aruspici, usate gli ARMA, ma leggete anche i giornali.

Appendice: L'ABC sui numeri complessi

Me l'hanno chiesto in tanti, non potevo esimermi. In questa appendice ci sono alcune brevissime e lacunosissime nozioni su come funzionano i numeri complessi e, in particolare, il motivo per cui radici complesse in un polinomio autoregressivo danno luogo a fenomeni ciclici.[19]

Iniziamo da i, che è definito dalla proprietà $i^2 = -1$. Il numero i viene chiamato *unità immaginaria*, largamente per ragioni storiche. Un suo qualunque multiplo reale si chiama *numero immaginario*. Per esempio, $3i$ è un numero il cui quadrato è -9.

[19] Chi volesse approfondire, trova facilmente tutto il necessario in rete; per esempio, su Wikipedia.

Un numero complesso è la somma di un numero reale più uno immaginario, cioè una cosa del tipo $z = a + bi$, dove a e b sono numeri reali, che si chiamano rispettivamente *parte reale* (spesso scritto $\Re(z)$) e *parte immaginaria* (spesso scritto $\Im(z)$).

Ogni numero complesso ha il suo *coniugato*, che è semplicemente lo stesso numero col segno cambiato nella parte immaginaria. L'operazione di coniugazione si scrive con una barretta. Quindi, se $z = a + bi$, $\bar{z} = a - bi$. Si noterà che $z \cdot \bar{z}$ è per forza un numero reale positivo:

$$z \cdot \bar{z} = (a + bi)(a - bi) = a^2 - (b^2 i^2) = a^2 + b^2,$$

tant'è che il valore assoluto, o modulo, di un numero complesso è definito come

$$|z| = \sqrt{z \cdot \bar{z}} = \sqrt{a^2 + b^2} = \rho;$$

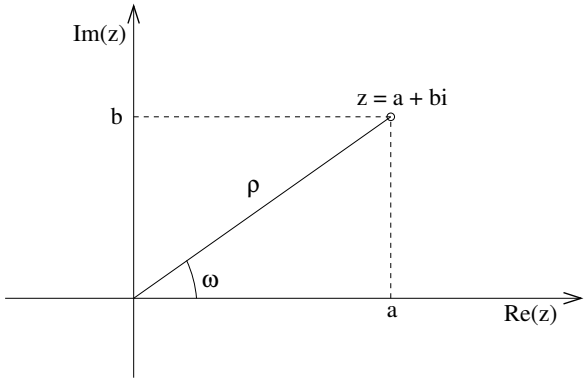
si noti che la definizione comprende come caso particolare quella per i numeri reali, che si ha quando $b = 0$.

In pratica, un numero complesso è un oggetto identificato da due numeri reali, cosicché tanto vale pensarlo come un punto su un piano. Per convenzione, questo piano viene pensato con la parte reale in ascissa e quella immaginaria in ordinata. Questo fa sì che, alternativamente, si possa rappresentare lo stesso punto in termini della sua distanza dall'origine ρ e dell'angolo che forma con l'asse delle ascisse ω, e cioè come $\rho(\cos\omega + i\sin\omega)$. Evidentemente, questo si traduce nelle due relazioni

$$\cos\omega = \frac{\Re(z)}{\rho} \qquad \sin\omega = \frac{\Im(z)}{\rho}.$$

La figura 2.16 dovrebbe far visualizzare la cosa con un certo agio. Aggiungo che, se $\rho = 1$, il punto corrispondente sta su un cerchio di raggio 1 e centro nell'origine, da cui l'espressione "cerchio unitario". Ovviamente, se $\rho > 1$, il punto sta fuori da detto cerchio e se $\rho < 1$ sta dentro.

Figura 2.16: Rappresentazione grafica di un numero complesso



Per ragioni che non sto a dire, vale anche la seguente uguaglianza:

$$z = a + bi = \rho \exp(i\omega),$$

dove l'operazione di coniugazione si traduce nel cambio di segno nell'argomento della funzione esponenziale:

$$\bar{z} = a - bi = \rho \exp(-i\omega),$$

Avere due rappresentazioni è comodo perché la prima funziona bene per la somma:

$$z_1 + z_2 = (a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + a_2) + (b_1 + b_2)i$$

e la seconda per il prodotto e l'elevamento a potenza:

$$z_1 \cdot z_2 = [\rho_1 \exp(i\omega_1)][\rho_2 \exp(i\omega_2)] = \rho_1 \rho_2 e^{i(\omega_1 + \omega_2)}$$

$$z^k = \rho^k e^{i\omega k}.$$

Infine, poiché $\Re(z) = \frac{z+\bar{z}}{2} = \rho \frac{e^{i\omega} + e^{-i\omega}}{2}$, si ha anche

$$\cos\omega = \frac{e^{i\omega} + e^{-i\omega}}{2}$$

e c'è anche una formula simile per la funzione seno:

$$\sin\omega = \frac{e^{i\omega} - e^{-i\omega}}{2i}$$

Più in generale, si può dimostrare che

$$\frac{e^{i\omega k} + e^{-i\omega k}}{2} = \cos(\omega k) \quad \Longrightarrow \quad \frac{z^k + \bar{z}^k}{2} = \rho^k \cos(\omega k) \tag{2.16}$$

$$\frac{e^{i\omega k} - e^{-i\omega k}}{2i} = \sin(\omega k) \quad \Longrightarrow \quad \frac{z^k - \bar{z}^k}{2i} = \rho^k \sin(\omega k) \tag{2.17}$$

A questo punto, abbiamo tutto il necessario per capire che legame c'è fra polinomi autoregressivi con radici complesse e fenomeni ciclici: supponete di avere un polinomio in L di ordine 2 con radici complesse coniugate:

$$A(L) = 1 - \varphi_1 L - \varphi_2 L^2 = (1 - zL)(1 - \bar{z}L).$$

Posto che $|z|$ sia minore di 1 per assicurare l'invertibilità, calcoliamo come è fatto l'inverso di $A(L)$, ciò che ci conduce alla rappresentazione di Wold. Si ha

$$A(L)^{-1} = (1 - zL)^{-1}(1 - \bar{z}L)^{-1} = (1 + zL + z^2L^2 + \cdots)(1 + \bar{z}L + \bar{z}^2L^2 + \cdots)$$

Come vedremo ora, non solo questo polinomio infinito ha tutti coefficienti reali, ma questi coefficienti oscillano intorno allo zero: infatti,

$$\begin{aligned} A(L)^{-1} \quad = \quad & 1 + \\ & (z + \bar{z})L + \\ & (z^2 + z\bar{z} + \bar{z}^2)L^2 + \\ & (z^3 + z^2\bar{z} + z\bar{z}^2 + \bar{z}^3)L^3 + \cdots \end{aligned}$$

e più in generale, se scriviamo

$$A(L)^{-1} = 1 + \theta_1 L + \theta_2 L^2 + \cdots$$

si ha che

$$\theta_k = \sum_{j=0}^{k} z^j \bar{z}^{k-j}$$

Il risultato che ci interessa si può trovare notando che

$$\theta_{k+1} = z^{k+1} + \bar{z}\theta_k = \bar{z}^{k+1} + z\theta_k,$$

da cui

$$z^{k+1} - \bar{z}^{k+1} = (z - \bar{z})\theta_k;$$

usando la (2.17) si ottiene

$$\theta_k = \rho^k \frac{\sin[\omega(k+1)]}{\sin\omega}$$

In pratica, la funzione di risposta di impulso è il prodotto fra due funzioni: $\frac{\sin[\omega(k+1)]}{\sin\omega}$, che è una funzione periodica in k di periodo $2\pi/\omega$ (e quindi, tanto più piccolo è ω, tanto più è lungo il ciclo). L'altra, ρ^k, che smorza le fluttuazioni al crescere di k visto che, per ipotesi, $|\rho| < 1$, e quindi ovviamente $\rho^k \to 0$.

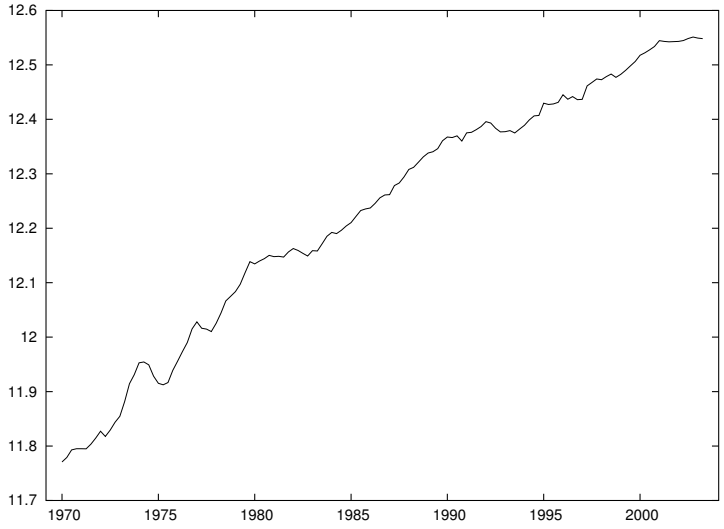
Capitolo 3

Processi integrati

3.1 Caratteristiche delle serie macroeconomiche

Tutto l'apparato descritto nei capitoli precedenti presuppone che le serie storiche economiche possano essere modellate come realizzazioni di processi stocastici stazionari. Questa, disgraziatamente, non è la situazione standard quando ci si occupa di serie storiche macroeconomiche. Osserviamo, tanto per fare un esempio, l'andamento nel tempo del logaritmo del Pil italiano trimestrale a prezzi 1990 (destagionalizzato), che è mostrato in figura 3.1.

Figura 3.1: log(PIL)



Come si vede, la serie esibisce un chiaro andamento crescente nel tempo, cosa che di per sé preclude la possibilità di modellarla con un processo stazionario, in quanto sarebbe opportuno usare un processo la cui media cambi nel tempo. Si potrebbe pensare, però, di modellare la serie in questo modo:

supponiamo che la serie segua un trend di crescita stabile nel tempo (dato sostanzialmente dal progresso tecnico e dall'accumulazione di capitale), che possiamo in prima approssimazione supporre una funzione lineare del tempo. A questo sentiero si sovrappone l'effetto "ciclo economico", o "congiuntura", che si può pensare come rappresentabile con un processo stazionario, perché il ciclo è un fenomeno di breve periodo a media 0 per definizione. Avremo dunque una relazione del tipo:

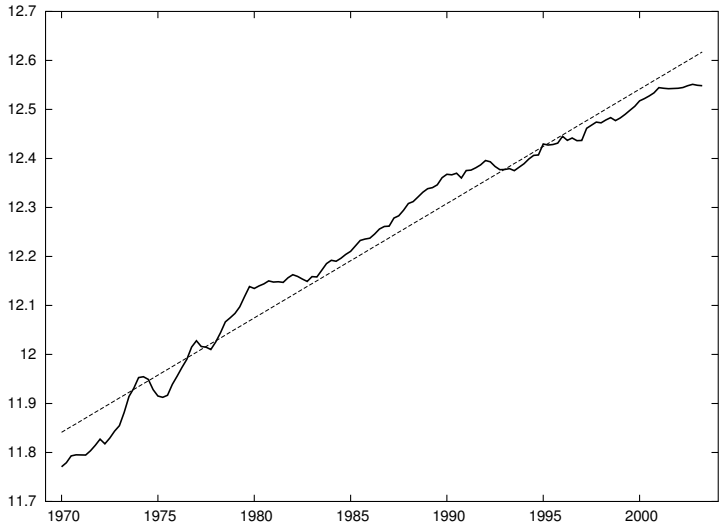
$$y_t = \alpha + \beta t + u_t$$

dove u_t è un qualche processo stocastico stazionario a media 0.

Il processo y_t testè descritto non è, a rigore, un processo stazionario, poiché $E(y_t) = \alpha + \beta t$, e quindi la media di y_t non è costante (per $\beta \neq 0$). Tuttavia, la non stazionarietà del processo è limitata a questo aspetto. Se si considerano le variazioni dal trend, quel che rimane è un processo stazionario, che può essere analizzato con le tecniche viste qualche pagina fa. È per questo motivo che i processi di questo tipo vengono denominati processi stazionari intorno ad un trend, o **processi TS** (dall'inglese *Trend-Stationary*).

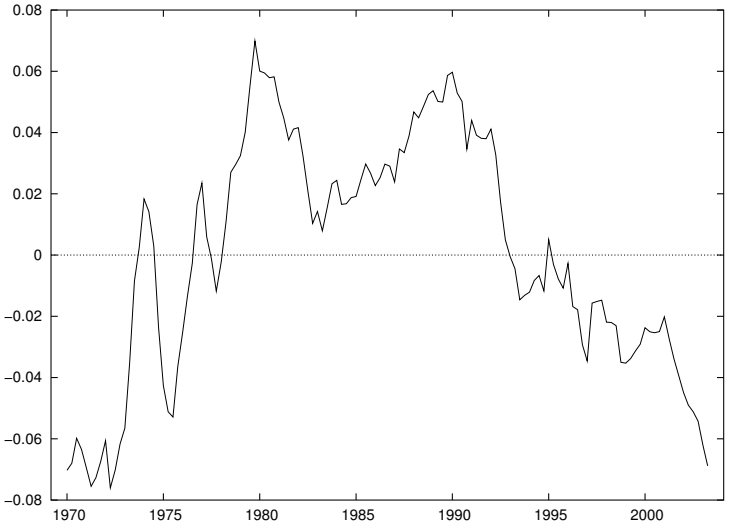
Se il processo y_t è TS, allora i parametri α e β si possono stimare in modo consistente con gli OLS, e l'inferenza funziona nel modo standard[1]. Una volta ottenuti $\hat{\alpha}$ e $\hat{\beta}$, sarà facile ottenere una scomposizione trend-ciclo della serie: il trend (che sarà una funzione lineare del tempo) sarà dato dalla serie $\hat{y}_t$, mentre il ciclo sarà dato dalla serie $\hat{u}_t$; nel nostro caso, le due serie sono mostrate nelle figg. 3.2 e 3.3.

Figura 3.2: log(PIL) e trend deterministico



[1]Questo caso è uno dei cavalli di battaglia di qualunque testo di teoria asintotica. Chi è interessato guardi lì.

Figura 3.3: Residui



Già dal grafico si vede 'a occhio' che i residui non sono *white noise*. Tuttavia, è pensabile che si riesca a trovare un processo ARMA di un qualche ordine che renda conto delle caratteristiche di persistenza di $\hat{u}_t$.

Una possibilità alternativa di rappresentazione della serie può essere quella di considerare la serie Δy_t. Visto che stiamo lavorando coi logaritmi, questa serie (l'andamento della quale è mostrato in figura 3.4) può essere interpretata come la serie storica dei tassi di crescita trimestrali.

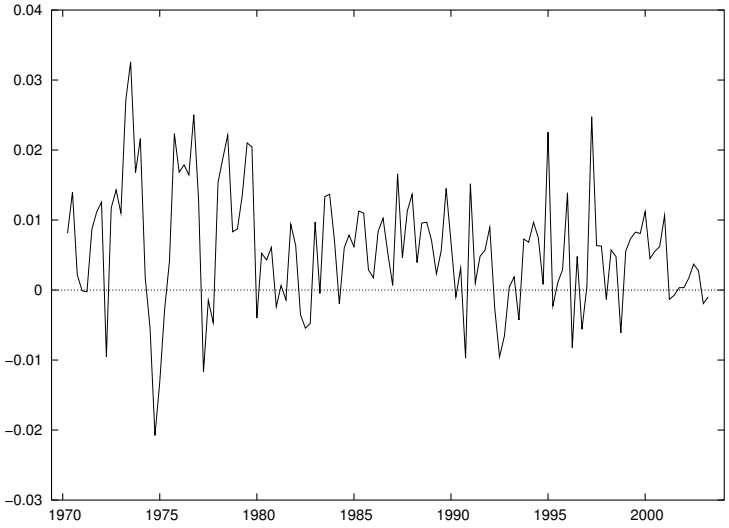
Poiché il saggio di crescita dovrebbe (ragionevolmente) fluttuare all'interno di una banda, si può immaginare di rappresentare la serie Δy_t per mezzo di un processo stazionario a media magari non nulla. Si noti che, in questo contesto, y_t è un processo a radice unitaria. Infatti, se Δy_t è stazionario ammette una rappresentazione di Wold del tipo

$$\Delta y_t = \mu + C(L)\epsilon_t,$$

dove μ è il tasso medio di crescita. Questa espressione può anche essere letta come la descrizione di y_t come un processo ARMA, in cui $A(L) = 1 - L$; di conseguenza, il valore di z per cui $A(z) = 0$ è 1. Come sappiamo, i processi a radice unitaria non sono stazionari, e quindi y_t non è stazionario, ma la sua differenza prima sì. In questo caso, si dice che y_t è stazionario in differenza, o DS (*Difference-Stationary*). Un'altra espressione che si usa frequentemente è che y_t è un processo $I(1)$ (che si legge **integrato di ordine uno**), a dire che y_t va differenziato una volta perché il risultato sia stazionario.

Una di quelle questioni da cornicetta: cosa succede differenziando un processo $I(0)$? Ovviamente il processo che ne risulta è stazionario, ma sappiamo che, per costruzione, ha una radice unitaria nella sua rappresentazione MA, per cui non è invertibile. Qualcuno sostiene che ha senso estendere la notazione appena illustrata ad interi negativi. Ad esempio, se u_t è $I(0)$,

Figura 3.4: Δlog(PIL)



aderendo a questa convenzione si può dire che Δu_t è $I(-1)$.
Come tutte le convenzioni, è buona nella misura in cui è utile. Io personalmente non ho ancora deciso al riguardo, ma è giusto che ve lo dica.

Per apprezzare adeguatamente le differenti conseguenze che scaturiscono dalla scelta di modellare una serie come un processo TS o come un processo DS è necessario analizzare nel dettaglio la caratteristiche dei processi a radice unitaria, ciò che rappresenta l'oggetto dei prossimi paragrafi.

3.2 Processi a radice unitaria

Come ho appena detto, un processo $I(1)$ è un processo che non è stazionario, ma è stazionaria la sua differenza prima. Più in generale, si definisce come processo $I(d)$ un processo la cui differenza d-esima è stazionaria. Per quel che ci riguarda, noi ci occuperemo solo dei casi in cui d è 0 oppure 1, anche se non manca una copiosa mole di letteratura dedicata a casi più esotici.

Il primo processo $I(1)$ di cui analizziamo le proprietà è il cosiddetto *random walk*. La definizione è semplice: y_t è un *random walk* se Δy_t è un *white noise*. Una cosa che va notata immediatamente è che per questo processo vale la relazione $y_t = y_{t-1} + \epsilon_t$; di conseguenza, sostituendo ripetutamente i valori passati di y_{t-1} si ha

$$y_t = y_{t-n} + \sum_{i=0}^{n-1} \epsilon_{t-i}$$

Quali sono le caratteristiche di un processo come questo? Tanto per cominciare, rendiamo le cose più semplici: supponiamo che il processo abbia avuto

inizio ad un tempo remoto, che chiamiamo tempo 0, e che a quella data il valore di y_t fosse 0. In questo caso, l'espressione precedente si riduce a

$$y_t = \sum_{i=1}^{t} \epsilon_i \tag{3.1}$$

Si noti che questa espressione può essere considerata una specie di rappresentazione a media mobile di y_t, in cui tutti i coefficienti sono pari a 1. È chiaro che, ad ogni istante t, la media del processo è 0. Se fosse solo per la media, quindi, il processo sarebbe stazionario. La varianza, però, non è costante, in quanto y_t è la somma di t v.c. indipendenti ed identiche con varianza (diciamo) σ^2; ne consegue che la varianza di y_t è $t\sigma^2$, e quindi cresce nel tempo. Da qui, e dal fatto che $Cov(y_t, y_s) = \sigma^2 \min(t, s)$ (dimostrarlo è un utile esercizio), consegue che y_t non è stazionario.

Per molti aspetti, conviene considerare un *random walk* come un caso limite di un AR(1) in cui le caratteristiche di persistenza sono così esasperate da modificare il processo nelle sue caratteristiche qualitative. In particolare, un *random walk* è, come abbiamo già detto, non stazionario. In più, si può considerare la funzione di risposta d'impulso ad esso associata: sebbene la rappresentazione di Wold non esista, la funzione di risposta di impulso è perfettamente definita come $IRF_k = \frac{\partial y_{t+k}}{\partial \epsilon_t}$. In questo caso, essa vale 1 per ogni valore di k, per cui è piatta e non decade esponenzialmente come nel caso stazionario: ciò significa che l'effetto di uno shock al tempo t permane indefinitamente nel futuro.

Quest'ultima caratteristica fa anche sì che i *random walk* non condividano con i processi stazionari la caratteristica di essere ***mean-reverting***. Se un processo è *mean-reverting*, esso presenta la tendenza a muoversi preferenzialmente verso il suo valore atteso; per un processo a media 0, significa che il grafico del processo interseca 'frequentemente' l'asse delle ascisse. Più formalmente, la locuzione è di solito impiegata per descrivere un processo la cui funzione di risposta di impulso tende asintoticamente a 0.

Tabella 3.1: AR(1) stazionario *versus random walk*

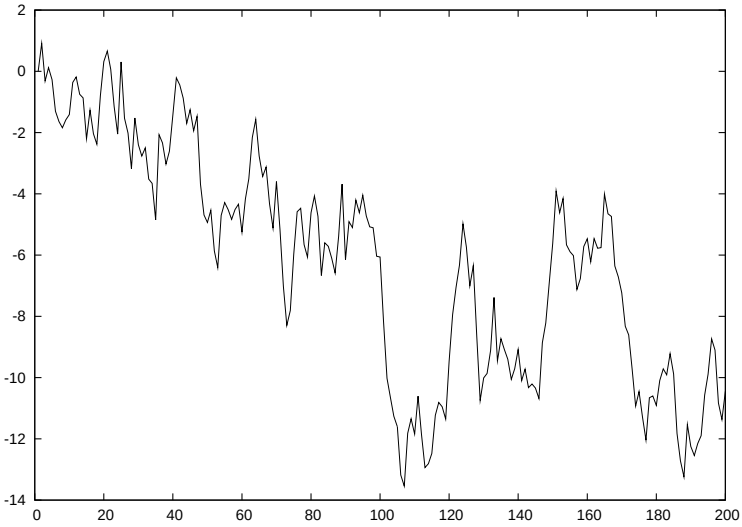
$y_t = \varphi y_{t-1} + \epsilon_t$

	$\lvert\varphi\rvert < 1$	$\varphi = 1$
Varianza	Finita	Illimitata
Autocorrelazioni	$\rho_i = \varphi^i$	$\rho_i = \sqrt{1 - \frac{i}{t}}$
mean-reverting	Sì	No
Memoria	Temporanea	Permanente

La tabella 3.1 (rubata a Banerjee *et al.* (1993)) evidenzia le differenze fra un AR(1) stazionario e un *random walk*.

Appare qui evidente che la scelta fra un processo TS e un processo DS per la modellazione di una variabile come, che so, il PIL comporta delle conseguenze notevoli per l'analisi dell'andamento di tale variabile nel lungo periodo. Se il PIL fosse rappresentabile come realizzazione di un processo TS, nel lungo periodo ciò che conta per la crescita economica è l'andamento del trend esogeno (tecnologia o che altro); una crisi congiunturale può avere un effetto depressivo, ma questo è solo **temporaneo**: il sistema ha una sua tendenza intrinseca a ritornare sul trend di lungo periodo.
Viceversa, se la metafora più appropriata per la serie storica del PIL fosse un processo DS, dovremmo concludere che esistono shock **permanenti** che non verranno mai riassorbiti: le colpe (o i meriti) dei padri ricadranno sui figli dei figli dei figli, e anche più in là.
Questa sarà anche una visione inquietante, ma in certi casi può essere del tutto appropriata: chi ha detto che la tecnologia marci ad un tasso esogeno e fisso? L'Alto Medioevo è durato parecchio, ma il progresso tecnologico ha lasciato un po' a desiderare. E poi, una volta che una cosa è stata inventata, non si può dis-inventarla (a meno di invasioni barbariche o guerre nucleari): una volta che il progresso tecnologico c'è stato, c'è anche per tutti quelli che vengono dopo. Il dibattito sull'argomento è ricco e fiorente, ma io mi contento di aver dato un'idea.

Figura 3.5: Random walk



Che aspetto ha 'a occhio' un *random walk*? Riportiamo alla mente la figura 2.7 a pag. 27 e confrontiamola con la figura 3.5. Come quelle volpi dei miei lettori avranno già intuito, in figura 3.5 è rappresentato un *random walk* i cui incrementi (le ϵ_t) non sono altro che il *white noise* usato per generare la serie mostrata in figura 2.7. In pratica, l'unica differenza fra la figura 2.7 e la figura 3.5 è il coefficiente dell'autoregressivo, che è pari a 0.9 nel primo caso e pari a 1 nel secondo. Si noti l'aumento di persistenza della serie indotto dalla radice unitaria.

Un aspetto caratteristico dei *random walk* è quello per cui l'assenza di *mean reversion* provoca periodi — anche molto lunghi — in cui la serie presenta un andamento crescente o decrescente piuttosto marcato. Ad esempio, chi non sapesse che la serie disegnata in figura 3.5 è frutto del puro caso, potrebbe

anche lanciarsi a commentare il 'chiaro' andamento decrescente del primo tratto con la 'crisi' attorno all'osservazione 100 e la conseguente 'ripresa' (e potrei continuare). È per questa caratteristica che spesso, quando si parla di *random walk* o più in generale di processi $I(1)$, si parla di trend **stocastici**, opposto a trend **deterministici**, che sono semplici funzioni del tempo.

Naturalmente, nessuno esclude che ci possano essere effettivamente dei trend deterministici sovrapposti a quelli puramente stocastici. Questo accade, ad esempio, nei cosiddetti *random walk* con *drift*. Questi processi sono semplicemente processi per cui si ha

$$\Delta y_t = \mu + \epsilon_t$$

e quindi y_t è un *random walk* a cui si sovrappone una funzione lineare del tempo. Se il *drift*, cioè la costante μ, è positivo, si avrà un processo che tende a salire, ma con fluttuazioni intorno a questo trend via via più marcate al passare del tempo. Perché questo avvenga lo si vede bene considerando la (3.1), che in questo caso si modifica in

$$y_t = \sum_{i=0}^{t} \epsilon_i + \mu \cdot t \tag{3.2}$$

dove il secondo termine della somma non è che un trend lineare con pendenza μ; in casi più generali si hanno cose del tipo

$$\Delta y_t = d_t + \epsilon_t$$

dove d_t è una qualche funzione deterministica del tempo: trend polinomiali di varia natura (cioè cose del tipo $d_t = \delta_0 + \delta_1 t + \cdots$), dummy stagionali e così via. È interessante, in questo caso, notare che vale un risultato che comprende il caso del semplice *drift* come caso particolare: se d_t è un polinomio in t di grado k, allora dentro y_t sarà presente un polinomio in t di grado $k+1$. Nel caso del *drift*, infatti, $d_t = \mu$, cioè un polinomio di ordine 0; analogamente, si mostra che un trend lineare incluso in Δy_t produce un trend quadratico in y_t, eccetera. Questo accade semplicemente perché, se μ_t è un polinomio in t di ordine p, allora $\Delta\mu_t$ è un polinomio di ordine $p-1$ (provare per credere).

Il caso del *random walk* si estende in modo piuttosto indolore al caso in cui gli incrementi del processo non sono un *white noise*, ma più in generale un qualunque processo stocastico stazionario: rappresentando quest'ultimo come un processo ARMA, avremo una rappresentazione del tipo

$$A(L)\Delta y_t = C(L)\epsilon_t,$$

dove ometto una eventuale parte deterministica per tenere semplice la notazione.

In questi casi parliamo genericamente di processi $I(1)$; come vedremo fra breve, processi di questo tipo condividono col caso particolare del *random walk* molte delle sue caratteristiche salienti, come quella di non possedere momento secondo costante, di non essere *mean-reverting* e di possedere memoria infinita, anche se le cose si fanno più articolate, perché la presenza dei polinomi $A(L)$

e $C(L)$ conferisce al processo una memoria di breve periodo, oltre a quella di lungo periodo.

Sebbene infatti la distinzione fra processi integrati e processi stazionari sia perfettamente definita ed assolutamente univoca, quando si osservano realizzazioni finite di processi $I(1)$ si possono dare delle situazioni in cui le differenze si fanno più sfumate, e questo perché la memoria di breve periodo si sovrappone a quella di lungo creando effetti curiosi: consideriamo ad esempio due processi definiti da

$$\begin{aligned} x_t &= 0.99999 x_{t-1} + \epsilon_t \\ y_t &= y_{t-1} + \epsilon_t - 0.99999 \epsilon_{t-1} \end{aligned}$$

A rigore, x_t è $I(0)$, mentre y_t è $I(1)$. Tuttavia, nel caso di x_t la radice del polinomio $A(L)$ non è 1, ma poco ci manca; a tutti i fini pratici, in campioni finiti una realizzazione di x_t è del tutto indistinguibile da quella di un *random walk*. Viceversa, y_t può essere scritto nella forma

$$(1 - L)y_t = (1 - 0.99999L)\epsilon_t = A(L)y_t = C(L)\epsilon_t$$

così da rendere evidente che i polinomi $A(z)$ e $C(z)$ sono molto vicini a poter essere 'semplificati'; da un punto di vista pratico, qualunque realizzazione di y_t non presenta apprezzabili differenze da un *white noise*.

3.3 La scomposizione di Beveridge e Nelson

Per quel che abbiamo visto nel paragrafo precedente, un *random walk* è un caso particolare di processo I(1). Un processo del tipo $y_t = x_t + u_t$ (dove x_t è un *random walk* e u_t è un processo I(0) qualunque) è un processo integrato, perché non è stazionario se non dopo una differenziazione, ma non è un *random walk*, perché Δy_t non è un *white noise*.

Se fosse possibile scrivere in questa forma un processo I(1) ci sarebbe un vantaggio immediato: si potrebbero attribuire a due componenti distinte le caratteristiche di non stazionarietà da un lato, e di persistenza di breve periodo dall'altro. In questo senso, si può pensare che y_t venga scisso in due componenti: una permanente, o di lungo periodo, che è data da x_t, ed una transitoria, o di breve periodo, data da u_t. Poiché u_t è per definizione un processo a media 0, y_t può essere pensato come un processo che fluttua intorno a x_t, senza che queste fluttuazioni siano mai troppo pronunciate.

La cosa interessante è che *qualsiasi* processo I(1) può essere pensato come la somma di un *random walk* e di un processo I(0). Questa scomposizione è nota come **scomposizione di Beveridge e Nelson** a volte anche detta **scomposizione BN**. La scomposizione BN può essere illustrata partendo da una proprietà dei polinomi quasi banale: dato un polinomio $C(z)$ di ordine q, è sempre possibile trovare un polinomio $C^*(z)$, di ordine $q - 1$, tale per cui

$$C(z) = C(1) + C^*(z)(1 - z).$$

La dimostrazione non è difficile: naturalmente, $D(z) = C(z) - C(1)$ è ancora un polinomio di ordine q, poiché $C(1)$ è una costante (la somma dei coefficienti di $C(z)$).

Tuttavia, segue dalla definizione che $D(1) = 0$, e quindi 1 è una radice del polinomio $D(z)$. Esso, allora, può anche essere scritto $D(z) = C^*(z)(1-z)$, dove $C^*(z)$ è un polinomio di grado $q-1$. In altri termini, il polinomio $C^*(z)$ risulta definito da

$$C^*(z) = \frac{C(z) - C(1)}{1 - z},$$

da cui l'espressione nel testo. Non ho voglia di spiegare il perché, ma dimostrare che

$$c_i^* = -\sum_{j=i+1}^{q} c_j$$

può essere un simpatico esercizio.

Prendiamo ora un processo I(1) arbitrario, e chiamiamolo y_t. Il processo Δy_t è di conseguenza un I(0), e quindi deve avere una rappresentazione di Wold che possiamo scrivere in questo modo:

$$\Delta y_t = C(L)\epsilon_t$$

Applicando a $C(L)$ la scomposizione polinomiale appena illustrata, possiamo anche scrivere

$$\Delta y_t = [C(1) + C^*(L)(1-L)]\,\epsilon_t = C(1)\epsilon_t + C^*(L)\Delta\epsilon_t \tag{3.3}$$

Se definiamo un processo μ_t tale per cui valga $\Delta\mu_t = \epsilon_t$ (ossia un *random walk* i cui incrementi siano dati da ϵ_t), si arriva a

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t = P_t + T_t \tag{3.4}$$

dove $P_t = C(1)\mu_t$ è un *random walk* che chiamiamo componente permanente e $T_t = C^*(L)\epsilon_t$ è un processo I(0) che chiamiamo componente transitoria.

Esempio 3.3.1 (Semplice) *Prendiamo un processo integrato di ordine 1 y_t per cui valga*

$$\Delta y_t = \epsilon_t + 0.5\epsilon_{t-1} = (1 + 0.5L)\epsilon_t = C(L)\epsilon_t$$

dove ϵ_t è un white noise*. Poiché*

$$C(1) = 1.5 \qquad C^*(L) = -0.5$$

si ha

$$y_t = 1.5\mu_t - 0.5\epsilon_t$$

Esempio 3.3.2 (Più complicato) *Supponiamo che Δy_t sia rappresentabile come un ARMA(1,1)*

$$(1 - \varphi L)\Delta y_t = (1 + \theta L)\epsilon_t$$

e quindi $C(L) = \frac{1+\theta L}{1-\varphi L}$.

$C(1)$ è facile da calcolare, ed è uguale a $\frac{1+\theta}{1-\varphi}$. Il calcolo di $C^*(L)$ è un po' più lungo ma non più difficile; si arriva a dimostrare che

$$C^*(L) = -\frac{\varphi + \theta}{1 - \varphi}(1 - \varphi L)^{-1}$$

Il risultato finale è

$$\begin{aligned} y_t &= P_t + T_t \\ P_t &= \frac{1+\theta}{1-\varphi}\mu_t \\ T_t &= -\frac{\varphi+\theta}{1-\varphi}(1-\varphi L)^{-1}\epsilon_t \end{aligned}$$

Si noti che T_t è un processo autoregressivo di ordine 1, tanto più persistente quanto maggiore è $|\varphi|$. Di conseguenza, y_t può essere rappresentato come un random walk *più un processo AR(1) stazionario che gli fluttua attorno.*

Un'interpretazione interessante della grandezza $C(1)$ è quella di misura della persistenza di un dato processo, poiché misura la frazione dello shock che permane nel processo dopo un tempo 'infinito'. È possibile controllare che, applicando la scomposizione qui descritta ad un processo stazionario, $C(1) = 0$, mentre $C(1) \neq 0$ nel caso di processi $I(1)$. Intuitivamente, questa interpretazione può anche essere motivata osservando che $C(1)$ è un coefficiente che determina il peso del *random walk* sul processo. Nel caso del processo $I(1)$ esaminato alla fine della sezione precedente, che somigliava tanto ad un *white noise*, il coefficiente $C(1)$ risulta essere appena 0.00001.

L'utilità della scomposizione BN è duplice: da un punto di vista pratico, è uno strumento che viene spesso utilizzato in macroeconometria quando si tratta di separare trend e ciclo in una serie storica. In poche parole, data una serie storica che ci interessa scomporre in trend e ciclo, si stima un modello ARMA sulle differenze prime, dopodiché si applica la scomposizione BN a partire dai parametri stimati. La scomposizione BN non è l'unico strumento per raggiungere lo scopo, e non è immune da critiche[2], ma su questo, come al solito, rinvio alla letteratura specializzata.

L'altro uso che si fa della scomposizione BN è teorico. Con un nome diverso (*scomposizione in martingala*), gioca un ruolo fondamentale nella letteratura probabilistica sui processi stocastici quando si devono analizzare certe proprietà asintotiche. Questo a noi non interessa, ma della scomposizione BN faremo sistematico uso nell'analisi dei sistemi cointegrati, di cui parleremo più avanti.

3.4 Test di radice unitaria

I processi integrati, così come visti finora, hanno delle caratteristiche che li rendono molto interessanti, sia da un punto di vista formale (perché rappresentano un esempio di processi non stazionari), che da un punto di vista pratico (perché le loro realizzazioni somigliano in modo spiccato alle serie storiche che siamo abituati ad incontrare in macroeconomia).

Non abbiamo, però, ancora esaminato le conseguenze della non stazionarietà dei processi di questo tipo per la possibilità di fare inferenza sulle loro

[2]Una, ad esempio è: dove sta scritto che la componente di lungo periodo debba essere per forza un *random walk*, anziché un qualche altro tipo di processo I(1)?

realizzazioni. Ricordo che, fino ad ora, abbiamo sempre supposto la stazionarietà dei processi per cui ci interessava fare inferenza. Nel caso dei processi $I(1)$, le cose si fanno più complesse.

Cominciamo con una banalità: se y_t è $I(1)$, allora Δy_t è $I(0)$ per definizione, e quindi tutto il bagaglio di conoscenze fin qui accumulato sulla stima dei parametri che caratterizzano i processi stazionari può essere riciclato senza problemi stimando un modello del tipo

$$A(L)\Delta y_t = C(L)\epsilon_t$$

e quindi modelleremo un tasso di crescita anziché il (logaritmo del) PIL, il tasso d'inflazione anziché (il logaritmo del) l'indice dei prezzi, e così via. È comune riferirsi a questo tipo di modelli come a modelli ARIMA, cioè ARMA integrati, nella letteratura statistica (che al proposito è sconfinata).

Una strategia di questo tipo, però, presuppone che si sappia esattamente se una serie è integrata o stazionaria[3]. A meno di rivelazioni soprannaturali, di solito questa è una cosa che non si sa; o per meglio dire, non è quasi mai possibile stabilire *a priori* se una certa serie può essere rappresentata meglio con un processo $I(0)$ oppure $I(1)$.

Questa decisione, però, può essere presa sulla base dei dati stessi. Una prima idea potrebbe essere semplicemente quella di osservare il grafico dell'andamento nella serie nel tempo. Se un processo è stazionario, non può presentare un andamento regolare crescente o decrescente, e quindi si potrebbe pensare di considerare stazionario un processo che oscilla attorno ad un valore costante, e non stazionario altrimenti.

Tale regola, che con un po' di occhio e di esperienza non è del tutto da buttar via, risulta però troppo semplicistica, e questo per almeno tre motivi: in primo luogo, perché un giudizio del genere è piuttosto soggettivo e scarsamente formalizzabile; in secondo luogo, perché può benissimo darsi che un processo sia stazionario attorno ad un trend deterministico (come si è visto qualche pagina fa); infine, perché esiste anche la possibilità che un processo effettivamente $I(1)$ dia luogo a realizzazioni che non presentano una tendenza particolarmente marcata a salire o a scendere. Per tutte queste ragioni, è necessaria una regola di decisione meno arbitraria e più affidabile. Regole di decisione di questo tipo sono note come **test di radice unitaria**.

Di test di radice unitaria ce n'è più d'uno[4]. Quelli più usati discendono però da un'impostazione comune, che illustrerò per sommi capi. Partiamo da

[3]Questa è una semplificazione piuttosto grossolana: a parte il fatto che, usando concetti appena più complessi di quelli di cui parlo qui, si possono dare esempi di processi che non sono né $I(0)$ né $I(1)$, ricordo che l'integrazione non è una caratteristica della serie storica, ma del processo stocastico che adottiamo per darne una rappresentazione statistica.

A voler essere rigorosi, dovremmo dire "... che si sappia esattamente se la serie storica osservata è rappresentata meglio da un processo stocastico stazionario o integrato di ordine 1", e la questione, a questo punto, potrebbe spostarsi sul significato di "meglio". Sottigliezze di questo tipo sono peraltro completamente ignorate dalla quasi totalità della macroeconomia contemporanea, e quindi non vale la pena di perderci il sonno.

[4]Pallido eufemismo. Ce n'è una marea. Anzi, c'è chi ha detto che di test di radice unitaria ce ne sono addirittura troppi. Chi fosse particolarmente interessato a questo argomento non può sottrarsi ad un esame della letteratura rilevante, che è vasta e complessa.

un processo autoregressivo del primo ordine che chiamiamo y_t:

$$y_t = \varphi y_{t-1} + u_t. \tag{3.5}$$

Per definizione, deve valere la relazione

$$\Delta y_t = \rho y_{t-1} + u_t \tag{3.6}$$

dove u_t è un *white noise* e $\rho = \varphi - 1$, cosicché il processo è stazionario solo se $\rho < 0$. Viceversa, se $\rho = 0$ siamo in presenza di un processo $I(1)$.

Visto che l'equazione appena scritta assomiglia sospettosamente ad un modello di regressione, si potrebbe congetturare che un test di radice unitaria non sia altro che un test t di azzeramento del parametro ρ, ossia un test basato sulla statistica

$$t_\rho = \frac{\hat{\rho}}{\sqrt{\widehat{\mathrm{Var}(\hat{\rho})}}}, \tag{3.7}$$

dove i 'cappelli' indicano come di consueto le stime OLS. Il test, in questo caso, avrebbe come ipotesi nulla la non stazionarietà del processo (se $\rho = 0$, allora $\varphi = 1$), e la stazionarietà come ipotesi alternativa ($\rho < 0$, e quindi $\varphi < 1$). La congettura è in effetti corretta, anche se ci sono almeno tre osservazioni da fare.

3.4.1 Distribuzione della statistica test

La prima osservazione riguarda il fatto che, sotto l'ipotesi nulla, la distribuzione del test t per l'azzeramento di ρ non è né una t di Student in campioni finiti, come accade nel modello lineare classico, né asintoticamente Gaussiana, come invece accade nel caso di processi stazionari. La sua distribuzione asintotica è invece una distribuzione un po' bislacca, per cui non esiste una espressione compatta né per la funzione di densità né per la funzione di ripartizione[5].

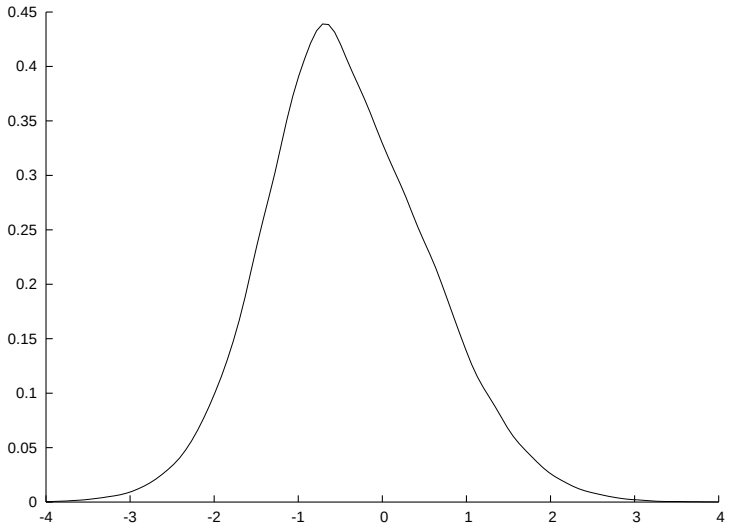
Una stima numerica della funzione di densità di questa statistica è mostrata in figura 3.6. I quantili di questa distribuzione vanno calcolati attraverso simulazioni numeriche, ed i primi che l'hanno fatto sono stati Dickey e Fuller nel 1976, ragion per cui talvolta questa distribuzione viene chiamata distribuzione DF, o Dickey-Fuller. Per lo stesso motivo, anche il test è noto come **test DF**. La conseguenza immediata di ciò è che per accettare o rifiutare l'ipotesi nulla bisogna consultare apposite tavole che però non mancano mai nei libri di econometria moderni né, men che meno, nei pacchetti econometrici.

3.4.2 Persistenza di breve periodo

Una seconda osservazione riguarda il fatto che, in generale, non è detto che y_t sia un *random walk* (ossia che u_t sia un *white noise*). È possibile, cioè, che Δy_t presenti esso stesso caratteristiche di persistenza, anche se di breve periodo.

[5]Questa distribuzione è definita in termini di integrali di moti browniani. Un moto browniano, o processo di Wiener, è un processo stocastico in tempo continuo di cui non dò la definizione, ma che sostanzialmente può essere pensato come un *random walk* in cui l'intervallo fra le osservazioni è infinitesimo.

Figura 3.6: Funzione di densità del test DF



In questi casi, che poi nella pratica sono quelli che più comunemente si incontrano, è necessario fare in modo che la distribuzione del test non risenta della memoria di breve periodo contenuta in u_t. Uno dei modi più diffusi è quello di supporre che le caratteristiche di memoria di Δy_t possano essere approssimate in modo soddisfacente da un processo AR(p) scegliendo un valore di p abbastanza grande.

Facciamo un esempio. Supponiamo di partire da una formulazione in livelli analoga alla (3.5):

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \epsilon_t. \tag{3.8}$$

In questo caso, supponiamo che ϵ_t sia un *white noise*, ossia che l'eventuale persistenza di breve periodo sia completamente catturata dalla parte autoregressiva.

Sfruttando il fatto che, per definizione, $y_{t-k} = y_{t-1} - \sum_{i=1}^{k-1} \Delta y_{t-i}$ (è facile convincersene controllando il caso $k = 2$), la (3.8) può essere riparametrizzata come segue[6]:

$$\Delta y_t = \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_p \Delta y_{t-p+1} + u_t \tag{3.9}$$

dove $\rho = (\varphi_1 + \cdots + \varphi_p) - 1$ e $\gamma_i = -(\varphi_{i+1} + \cdots + \varphi_p)$. In questo caso, il test prende il nome di **test ADF** (Augmented Dickey-Fuller), ed è il test t di azzeramento del parametro ρ nella regressione (3.9).

[6]Lettori particolarmente volonterosi possono controllare che qui non facciamo altro che applicare la scomposizione BN: infatti, scrivendo

$$y_t = B(L) y_{t-1} + u_t$$

si può andare avanti notando che

$$\Delta y_t = [B(L) - 1] y_{t-1} + u_t.$$

Il risultato del testo segue applicando la scomposizione BN al polinomio $H(L) = B(L) - 1$.

Se il valore scelto di p è abbastanza alto, e quindi la correzione è efficace, la distribuzione del test ADF è la stessa del test DF. Cosa vuol dire "abbastanza alto"? Vuol dire semplicemente che u_t deve essere, per lo meno ai fini pratici, un *white noise*. In pratica, spesso si usano gli stessi criteri di selezione di p che si usano per il problema analogo in ambito stazionario, di cui ho parlato nella sezione 2.7, e cioè si sceglie p in modo da minimizzare criteri del tipo Akaike o Schwartz. Un modo affine di risolvere questo problema è stato proposto da Phillips e Perron, e il cosiddetto test di Phillips e Perron (chiamato familiarmente **test PP**) si affianca oramai al test ADF in parecchi pacchetti.

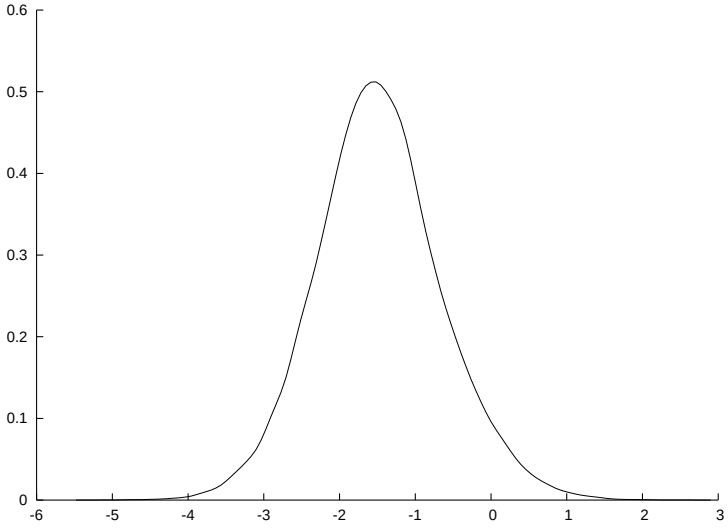
3.4.3 Nucleo deterministico

Infine, va menzionato il fatto che la distribuzione del test (sia del tipo ADF che del tipo PP) non è invariante al nucleo deterministico che si inserisce nella regressione. Finora abbiamo esaminato il caso di un *random walk* senza drift. Nel caso in cui un drift sia effettivamente presente nel processo che ha generato la serie in esame, esso va incluso anche nella regressione usata per calcolare il test. Ma come si fa a sapere se il drift c'è oppure no? Il problema è che non si sa. Di conseguenza, la cosa migliore è quella di mettercelo, e quindi stimare una regressione del tipo

$$\Delta y_t = \mu + \rho y_{t-1} + \varphi_1 \Delta y_{t-1} + \cdots + \varphi_p \Delta y_{t-p} + u_t \tag{3.10}$$

in cui, tutt'al più, μ varrà 0 nel caso in cui il drift non ci sia.

Figura 3.7: Funzione di densità del test DF con intercetta



Disgraziatamente, è possibile dimostrare che in questo caso la distribuzione asintotica del test di azzeramento è diversa da quella vista in precedenza. Come se non bastasse, in realtà le distribuzioni rilevanti sono due: una — nonstandard, anch'essa tabulata, e mostrata nella figura 3.7 — nel caso in cui

il vero valore di μ sia 0; nel caso in cui $\mu \neq 0$, invece, viene fuori che la distribuzione asintotica del test è (forse sorprendentemente) normale, anche se la dimensione campionaria deve essere molto grande perché l'approssimazione sia soddisfacente.

Come si vede, la cosa diventa un tantino ingarbugliata già a questo stadio; se poi si analizza il caso in cui nella (3.10) si aggiunge anche un trend deterministico lineare, si ha un'altra distribuzione ancora. Questa molteplicità di situazioni è forse uno degli aspetti che lascia più perplessi quando ci si accosta ai test di radice unitaria. In realtà se ne viene fuori, ma con molta pazienza e facendo una serie di distinguo per i quali, però, rinvio alla letteratura specializzata, ritenendo esaurito il compito di introduzione divulgativa che mi propongo qui; un problema molto simile tornerà nel capitolo 5, ma ne parleremo a tempo debito.

3.4.4 Test alternativi

Il test ADF assume come ipotesi nulla l'esistenza della radice unitaria, e così le sue varianti tipo il test Phillips-Perron; ci sono invece test che partono dalla nulla di stazionarietà. Il più noto di questi ultimi è il cosiddetto **test KPSS**, di cui spiego l'intuizione base. Se y_t fosse stazionario attorno ad un trend deterministico, allora una regressione del tipo

$$y_t = \beta_0 + \beta_1 \cdot t + u_t$$

dovrebbe produrre dei residui I(0). Fatta la regressione, si prendono i residui OLS e si cumulano, producendo una nuova serie $S_t = \frac{1}{T}\sum_{s=1}^{t} \hat{u}_s$; sotto la nulla, questa serie è pensabile (per campioni molto grandi) come una realizzazione di un processo un po' strano[7], perché per costruzione si ha non solo che $S_0 = 0$, ma anche che $S_T = 0$. In questo caso, si può dimostrare che la somma dei quadrati di S_t (opportunamente normalizzata) converge in distribuzione ad una variabile casuale che è sempre la stessa per qualunque processo stazionario. Se invece y_t non è stazionario, la statistica diverge. Di conseguenza, l'intervallo di accettazione va da 0 ad un certo valore critico che, anche in questo caso, è stato tabulato.

L'espressione "opportunamente normalizzata" che ho usato al capoverso precedente è volutamente un po' vaga: infatti, si può mostrare che l'ingrediente essenziale di questa normalizzazione è la varianza di lungo periodo di $\hat{u}_t$: quest'ultima è definita come la somma di tutte le sue autocovarianze (da meno a più infinito). Spesso, questa quantità viene stimata in modo non parametrico tramite la statistica $\hat{\omega}^2$, che è definita come

$$\hat{\omega}^2(m) = T^{-1} \sum_{t=m}^{T-m} \left[\sum_{i=-m}^{m} w_i \hat{u}_t \hat{u}_{t-i} \right],$$

dove m è noto come *window size* e i termini w_i sono i cosiddetti *pesi di Bartlett*, definiti da $w_i = 1 - \frac{|i|}{m+1}$. Si dimostra che, per m abbastanza grande, $\hat{\omega}^2(m)$ fornisce una stima consistente della varianza di lungo periodo. Il problema principale è la scelta di m, e qui regole precise non esistono: la teoria asintotica dice solo che m deve essere proporzionale a $T^{1/3}$, il che in pratica equivale a una licenza di fare come a uno gli pare. Il consiglio che dò io è di provare vari valori di m e vedere quando la statistica si stabilizza.

---

[7] Si chiama un *ponte browniano*, curiosoni.

Il test si può fare anche senza un trend, cosicché le $\hat{u}_s$ sono semplicemente gli scarti di y_t dalla sua media. Evidentemente, in questo caso l'ipotesi nulla è che il processo sia stazionario *tout court*. I valori critici cambiano, ma anche questi sono stati tabulati.

Secondo me, è sempre buona norma provare a testare l'ordine di integrazione di una serie in tutti e due i modi. Di solito, le indicazioni coincidono, nel senso che se il KPSS accetta l'ADF rifiuta e viceversa. Tuttavia, non è raro che questi test non diano indicazioni coerenti; capita sovente, cioè, che rifiutino (o accettino) la rispettiva nulla sia il test ADF che il test KPSS.

Infine, menziono il fatto che alcuni ritengono ormai superata l'idea stessa di fare test di ipotesi sull'ordine di integrazione in un contesto multivariato. Se abbiamo a che fare con più di una serie, si può procedere ad una batteria di test ADF o simili su ognuna di esse, naturalmente. Però forse è più intelligente partire direttamente da una rappresentazione multivariata (di cui parlerò nel capitolo 4), ciò che conduce al cosiddetto test di Johansen (di cui parlerò nel capitolo 5).

3.4.5 Usare il cervello

Una parola di commento sui test di radice unitaria: accade molto spesso che applicando un test di radice unitaria ad una serie storica la quale, ragionevolmente, dovrebbe fluttuare all'interno di una banda più o meno ampia, non sia possibile rifiutare l'ipotesi di radice unitaria. Questo avviene, ad esempio, quasi sempre con tassi di disoccupazione, tassi di inflazione, o tassi di interesse (reali o nominali). È comune, a questo punto, che qualcuno alzi la mano e dica: "Come è possibile che il tasso sui BOT sia $I(1)$? Era già al 12% al tempo dei babilonesi!"

Si possono dare, a questa obiezione, due risposte. Una è quella di dimostrare la propria adesione dogmatica al culto del p-value dicendo: "Il test viene così! Che ci posso fare?"; un'altra, che secondo me è più intelligente, è di far notare che *nel campione a nostra disposizione* il tasso sui BOT ha evidentemente un grado di persistenza tale per cui è meglio, da un punto di vista di aderenza ai dati, pensarlo come una realizzazione di un processo $I(1)$ che $I(0)$.

Non diciamo che la serie sia $I(1)$: in realtà, ammesso e concesso che abbia senso pensare la nostra serie storica dei tassi di interesse come realizzazione di un qualche processo stocastico, lo sa il diavolo che processo è; noi stiamo solo scegliendo all'interno di una classe limitata di processi (gli ARIMA) la parametrizzazione più appropriata per descrivere i dati. Se poi avessimo osservazioni su migliaia di anni, sospetto che il processo più adeguato a rappresentare l'andamento nel tempo dei tassi di interesse da Hammurabi in avanti sarebbe un $I(0)$, ma non credo che saremo mai nelle condizioni di stabilirlo.

È un problema di *rappresentazione dei dati*: con un test di radice unitaria non stiamo veramente decidendo se il processo è $I(1)$ oppure $I(0)$. Stiamo soltanto decidendo se è più conveniente rappresentare i dati che abbiamo con un processo stazionario o integrato.

Una metafora che io trovo calzante è quella della curvatura della Terra. Per molti secoli si è creduto che la Terra fosse piatta semplicemente perché non c'era ragione di pensarla rotonda: la curvatura della Terra diventa un problema solo quando si ha a che fare con grandi distanze tant'è che, ad esempio, bisogna tenerne conto per calcolare le rotte delle navi transoceaniche. Se però bisogna costruire una casa, o anche uno stadio o un parcheggio o un centro commerciale, la scala del problema è tale per cui la curvatura del globo diventa trascurabile (e infatti gli ingegneri la trascurano senza che le case crollino per questo).

Allo stesso modo, un processo stazionario può avere un grado di persistenza tale per cui le sue realizzazioni diventano "evidentemente" stazionarie solo dopo moltissimo tempo. Un test di radice unitaria condotto su un campione non così ampio accetta, probabilmente, la nulla. In questo caso, il test ci dice semplicemente che forse è meglio, nel nostro campione, differenziare la serie.

3.4.6 Un esempio

Sottoponiamo a test di radice unitaria la serie storica vista in apertura di capitolo, e cioè il logaritmo del PIL italiano a prezzi costanti (vedi fig. 3.1).

Per applicare il test ADF, bisogna scegliere l'ordine dei ritardi più adeguato a rappresentare la serie come processo autoregressivo. Ci sono molti modi per farlo. Uno dei più semplici è quello di fare delle regressioni del tipo

$$y_t = b_t + \sum_{i=1}^{p} \varphi_i y_{t-i} + \epsilon_t$$

per diversi valori di p e scegliere il p che minimizza il criterio di Schwarz; b_t è il nucleo deterministico che riteniamo più appropriato. Nel nostro caso, proviamo sia una costante ($b_t = a$) che una costante più trend ($b_t = a + b \cdot t$). I risultati sono sintetizzati nella tabella seguente:

p	C	C+T
1	-881.341	-877.246
2	-886.637	-884.346
3	-874.406	-872.165
4	-866.458	-863.827
5	-856.408	-853.772
6	-846.929	-844.136

Come si vede, il criterio BIC risulta minimizzato, in ambo i casi, dalla scelta $p = 2$. Procediamo, pertanto alla regressione sul modello riparametrizzato; nel caso del modello con sola costante si ha:

$$y_t = a + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \epsilon_t \Longrightarrow \Delta y_t = a + \varphi y_{t-1} + \gamma_1 \Delta y_{t-1} + \epsilon_t$$

dove $\varphi = \varphi_1 + \varphi_2 - 1 = -A(1)$ e $\gamma_1 = -\varphi_2$. Il test ADF consiste appunto nell'azzeramento di φ. Riporto qui la regressione per esteso:

Coefficiente	Stima	Errore Std.	t-stat.
a	0.07992	0.03877	2.0614
ϕ	-0.00623	0.00316	-1.9697
γ_1	0.35290	0.08172	4.3183

La statistica test è la statistica t relativa al parametro ϕ, e cioè -1.9697. Confrontandola con le apposite tavole (che *non* sono quelle della normale o della t) si scopre che il valore critico al 95% in questo caso è circa -2.89, per cui siamo ancora nell'area di accettazione della nulla di non stazionarietà. Alcuni pacchetti fanno anche di meglio, e cioè calcolano direttamente il p-value del test, che in questo caso risulta pari al 30.05%. Ripetendo il test per il modello con trend le cose non cambiano di molto:

Coefficiente	Stima	Errore Std.	t-stat.
a	0.42462	0.21979	1.9319
b	0.00018	0.00011	1.5930
ϕ	-0.03541	0.01858	-1.9052
γ_1	0.37087	0.08202	4.5217

Anche in questo caso, la statistica test (-1.9052) risulta più alta del valore critico (circa -3.45) per cui si accetta la nulla; la stessa conclusione, naturalmente, l'avremmo ottenuta se avessimo avuto un pacchetto che ci calcola il p-value, visto che in questo caso risulta pari al 65.17%.

La stessa procedura, applicata a Δy_t anziche a y_t, e quindi ai tassi di variazione del PIL (vedi fig 3.4) produce invece un netto rifiuto della nulla. Non riporto qui i risultati per brevità, fidatevi.

Conclusione? Qui le cose sembrano piuttosto chiare: y_t ha una radice unitaria, Δy_t no, per cui concludiamo che il processo più adeguato a rappresentare la serie è $I(1)$.

3.5 Regressione spuria

Nella breve ed incompleta presentazione dei test di radice unitaria fatta al paragrafo precedente sarà saltato all'occhio del lettore che, quando si fa inferenza con processi integrati, molte delle confortevoli certezze che ci accompagnano nel mondo della stazionarietà cedono il posto a risultati inconsueti. Questo stato di cose è ancora più eclatante quando si analizza il fenomeno della **regressione spuria**.

Prendiamo due processi y_t e x_t così definiti:

$$\begin{cases} y_t = y_{t-1} + \eta_t \\ x_t = x_{t-1} + \epsilon_t \end{cases} \tag{3.11}$$

dove η_t e ϵ_t sono due *white noise* indipendenti fra loro. È evidente che y_t e x_t sono due *random walk* che hanno ben poco a che spartire l'uno con l'altro. Ci si attenderebbe di non trovare traccia di relazioni statisticamente significative fra y_t e x_t. Così è, ma le cose non sono così semplici.

Se si tentasse di analizzare l'eventuale presenza di relazioni fra x_t e y_t impostando un modello di regressione lineare, si finirebbe con lo stimare un'equazione del tipo

$$y_t = \alpha + \beta x_t + u_t. \tag{3.12}$$

A prima vista, si potrebbe pensare che l'assenza di relazioni fra y_t e x_t comporti

1. che l'indice R^2 sia "basso";
2. che lo stimatore OLS di β converga in probabilità a 0;
3. che un test t di azzeramento di β, perlomeno in grandi campioni, rientri nella banda di accettazione dell'ipotesi nulla data dalle tavole della normale standardizzata; detto in parole povere, che la statistica t relativa al coefficiente β sia compresa fra -2 e 2 in 19 casi su 20.

Ebbene, nessuna di queste tre cose avviene nel caso in esame; al contrario:

1. l'indice R^2 converge in distribuzione ad una variabile casuale non degenere;
2. lo stimatore OLS di β converge in distribuzione ad una variabile casuale;
3. un test t di azzeramento di β porta, usando i valori critici della normale standardizzata, al rifiuto dell'ipotesi nulla, tanto più frequentemente quanto più grande è il campione (!).

È evidente che, sulla base di una regressione così, un ricercatore incauto, il quale non si ponga il problema dell'ordine di integrazione delle variabili, potrebbe "scoprire" relazioni fra variabili assolutamente inesistenti nella realtà: da qui l'espressione 'regressione spuria'[8].

Tabella 3.2: regressione spuria: Esperimento di Monte Carlo

Ampiezza campionaria	Percentuale di rifiuti
20	47.7%
50	66.4%
100	75.9%
200	83.5%
1000	92.5%

40000 simulazioni per ogni ampiezza campionaria

Per capire meglio la cosa, date un'occhiata alla tabella 3.2, in cui è evidenziato il risultato di un piccolo esperimento di Monte Carlo: ho simulato un sistema uguale a quello presentato dalla (3.11), con $E(\epsilon_t^2) = E(\eta_t^2) = 1$ per

[8] Il fenomeno era già stato osservato negli anni Venti. È solo con gli anni Settanta e Ottanta, però, che viene portato all'attenzione generale (per merito di Granger e Newbold) ed analizzato in profondità (per merito di P. C. B. Phillips).

diverse ampiezze campionarie. Fatta una regressione di y_t su una costante e su x_t (come quella presentata nella (3.12)), ho calcolato il valore del test t di azzeramento di β confrontandolo poi con il valore critico della t di Student al 95%. Seguendo questa procedura, si arriva ad una percentuale di rifiuti che, come si vede, non solo è abbastanza alta da essere imbarazzante, ma cresce al crescere dell'ampiezza campionaria.

Questi risultati, ad un esame più attento, non dovrebbero però sorprendere più di tanto. Per $\beta = 0$, infatti, l'espressione (3.12) si riduce a $y_t = \alpha + u_t$; se y_t è $I(1)$, delle due l'una: o u_t è $I(0)$, ma in questo caso l'equazione è contraddittoria, o u_t è anch'esso $I(1)$, e allora tutti i teoremi limite vanno a farsi benedire. In altri termini, non c'è un valore di β che renda la (3.12) una descrizione corretta dei dati; il valore $\beta = 0$ non è più giusto né più sbagliato che $\beta = 1$ o $\beta = -1$; il β "vero" *non esiste*. Un esame dell'equazione, infatti, rivela che che non esiste alcun meccanismo che renda conto della persistenza di y_t; di quest'ultima deve — necessariamente — farsi carico il termine di disturbo.

In pratica, questo stato di cose diventa evidente osservando che la stima della (3.12) con il metodo OLS scarica tutta la persistenza di y_t sui residui $\hat{u}_t$, che risultano fortemente autocorrelati. Anzi, è possibile dimostrare che, in presenza di questo fenomeno, la statistica Durbin-Watson[9] converge in probabilità a 0. Dirò di più: una regola rozza ma efficace per segnalare se una regressione è spuria o no è quella di confrontare l'indice R^2 col valore della statistica DW. Se il primo è maggiore della seconda, c'è di che insospettirsi (anche se, va detto, questo non va preso come un test a tutti gli effetti; è semplicemente un suggerimento euristico contenuto nell'articolo originale di Granger e Newbold).

Ora, se la regressione è uno strumento che può dare risultati fuorvianti se usato con realizzazioni di processi $I(1)$, a cui tipicamente le serie storiche macroeconomiche somigliano molto, vuol dire che non si può usare la regressione sulle serie storiche macro? Non è detto.

Innanzitutto, va detto che una gran parte degli aspetti apparentemente paradossali appena tratteggiati va imputata al fatto che non c'è nessun valore di β compatibile con una corretta descrizione dei dati, come ho detto poco fa. Se avessimo stimato una cosa del tipo

$$y_t = \alpha + \varphi y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

avremmo avuto che

$$\begin{pmatrix} \hat{\varphi} \\ \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \xrightarrow{\text{p}} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

cioè le stime OLS convergono ai valori 'veri' dei parametri. Una corretta specificazione dinamica (una specificazione, cioè, che permetta ai disturbi di somigliare ad un *white noise*) è un bel passo avanti. Parleremo meglio di questo

[9] Ricordo che la statistica Durbin-Watson è una venerabile statistica escogitata nella preistoria dell'econometria per controllare se ci fosse autocorrelazione di ordine 1 nei residui di un OLS. Tale statistica veniva anche, spesso, usata temerariamente come statistica test. Oggi siamo nel ventunesimo secolo e abbiamo strumenti più sofisticati per fare meglio la stessa cosa, ma i pacchetti ancora la riportano.

risultato, e di altri, nella sezione 4.4. La cosa più importante da dire, però, è che una regressione fra variabili integrate *può* avere un senso, ed anzi in determinate circostanze rappresenta un modo sbrigativo ma efficace di misurare relazioni statistiche a cui è possibile attribuire un significato ben preciso dal punto di vista della teoria economica che sovraintende al modello stimato. Questo accade quando le variabili a destra e a sinistra del segno di uguale sono **cointegrate**. Cosa sia la cointegrazione, lo vedremo nel capitolo 5.

Capitolo 4

Processi VAR

4.1 Processi multivariati

È piuttosto raro che un fenomeno complesso come quelli che di solito si studiano in economia possa essere descritto da una sola variabile. Ci si trova molto più comunemente nella situazione in cui i fatti a cui ci si interessa non possano essere riassunti in modo soddisfacente se non usando più di una grandezza.

In un contesto statico, questo conduce naturalmente all'uso di variabili casuali multiple, o, dir che si voglia, vettori aleatori. Le tecniche per lo studio delle realizzazioni dei vettori aleatori sono moltissime, e fra queste c'è l'analisi di regressione, croce e delizia degli studenti di Econometria.

In un contesto dinamico è necessario operare una generalizzazione definendo il concetto di **processo stocastico multivariato**. Anche in questo caso, non mi sforzerò di essere rigoroso: basterà dire che un processo stocastico multivariato è un processo stocastico i cui elementi non sono variabili casuali semplici, ma multiple; in alternativa, si può pensare ad un processo stocastico multivariato come ad un vettore i cui elementi sono processi stocastici univariati. Se, ad esempio, pensiamo alla rilevazione giornaliera del tasso di cambio euro/dollaro come alla realizzazione di un processo univariato, possiamo pensare alla rilevazione giornaliera dei tassi di cambio euro/dollaro, euro/yen, euro/sterlina eccetera come alla realizzazione di un processo multivariato.

Tale definizione rende pressoché ovvia l'estensione al caso multivariato di molti dei concetti visti in precedenza a proposito dei processi stocastici univariati: ad esempio, le definizioni di stazionarietà ed ergodicità rimangono immutate.

Per quanto riguarda i momenti, per processi debolmente stazionari sarà possibile definire i momenti primi e secondi come segue:

$$\begin{cases} E(y_t) = \mu \\ E\left[(y_t - \mu)(y_{t-k} - \mu)'\right] = \Gamma_k \end{cases}$$

dove, se il processo y_t ha n elementi, μ è un vettore $n \times 1$ e Γ_k è una matrice $n \times n$. Per $k = 0$, essa non è che la matrice varianze-covarianze del vettore y_t;

per $k \neq 0$, l'elemento ij di Γ_k rappresenta la covarianza fra l'i-esimo elemento di y_t ed il j-esimo elemento di y_{t-k}. La matrice di autocovarianze è definita in modo tale che $\Gamma_k = \Gamma'_{-k}$ e quindi, in generale, $\Gamma_k \neq \Gamma_{-k}$. Si noti che, per $n = 1$, queste definizioni coincidono con quelle date in precedenza per processi univariati.

Del pari, la definizione di un *white noise* multivariato è piuttosto semplice: chiamiamo in questo modo un processo ϵ_t tale per cui

$$\begin{aligned} E(\epsilon_t) &= 0 \\ \Gamma_k = E(\epsilon_t \epsilon_{t-k}) &= \begin{cases} \Sigma & \text{per } k = 0 \\ 0 & \text{per } k \neq 0 \end{cases} \end{aligned}$$

La definizione di *white noise* multivariato è quindi molto simile a quella di *white noise* univariato (del resto, la seconda è un caso particolare della prima). Va notato, peraltro, che Σ è una matrice di varianze e covarianze generica, e pertanto simmetrica e semidefinita positiva, ma non necessariamente diagonale. Di conseguenza, il fatto che un processo multivariato sia un *white noise* esclude la correlazione fra ogni elemento del processo e la storia passata di *tutto* il processo, ma non esclude che possa esserci correlazione fra elementi contemporanei.

Anche l'operatore L può essere applicato in modo del tutto analogo: $Lx_t = x_{t-1}$ anche nel caso in cui x_t sia un vettore. Le cose si fanno più articolate se consideriamo espressioni del tipo

$$x_t + Ax_{t-1} = (I + AL)x_t$$

dove A è una matrice quadrata. In questo caso l'espressione $(I + AL)$ è un operatore — funzione dell'operatore L — matriciale. Esso può essere visto in due modi equivalenti:

Polinomio matriciale L'operatore $(I + AL)$ è la somma di due matrici, ognuna delle quali "moltiplica" l'operatore L per una potenza diversa. Si può pensare a $(I + AL)$ come ad un polinomio di ordine 1 nell'operatore L in cui il primo coefficiente è la matrice identità ed il secondo è la matrice A.

Matrice di polinomi L'operatore $(I + AL)$ è una matrice i cui elementi sono polinomi di ordine 1; ad esempio, l'elemento ij di $(I + AL)$ è $\delta_{ij} + a_{ij}L$, dove δ_{ij} è il cosiddetto 'delta di Kronecker', che è uguale a 1 per $i = j$ e 0 altrimenti.

La generalizzazione al caso di polinomi di ordine p dovrebbe essere immediata, così che un'espressione del tipo

$$C(L)x_t = C_0 x_t + C_1 x_{t-1} + \cdots + C_p x_{t-p}$$

non dovrebbe destare alcuno stupore.

Il fatto di poter interpretare un operatore tipo $C(L)$ come una matrice di polinomi comporta anche che l'inversione di tali operatori segue le normali regole di inversioni di matrici, cosa che può tornare comoda in più di un caso.

4.2 I processi VAR

I processi VAR costituiscono la generalizzazione multivariata del processi AR. Un processo VAR di ordine p, infatti, può essere scritto in questo modo:

$$A(L)y_t = \epsilon_t \rightarrow y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + \epsilon_t$$

dove $A(L)$ è un polinomio matriciale di ordine p e ϵ_t è un *white noise* vettoriale. Questi processi fanno parte della più ampia famiglia dei processi VARMA, che hanno una struttura ARMA vettoriale ($A(L)y_t = C(L)\epsilon_t$); questi ultimi, però, sono molto complicati da stimare quando il polinomio $C(L)$ ha un ordine maggiore di 0, e quindi la stragrande maggioranza delle applicazioni empiriche fa uso dei modelli VAR, che come vedremo possono essere stimati in modo semplice e consistente con gli OLS, piuttosto che dei VARMA.

Molte delle considerazioni che abbiamo fatto in precedenza a proposito dei modelli AR si estendono in modo piuttosto banale ai modelli VAR. Il fatto che però in un modello multivariato abbiamo a che fare con dei polinomi matriciali anziché scalari impone una serie di considerazioni aggiuntive. Tanto per cominciare, possiamo chiederci se è possibile, come nel caso univariato, esprimere un processo VAR in forma di processo a media mobile multivariato (VMA). La risposta è evidentemente legata all'invertibilità dell'operatore $A(L)$, il che ci porta a valutare sotto quali condizioni $A(L)$ possieda un'inversa.

Nel caso di processi univariati, avevamo visto a suo tempo che bisognava vedere se i valori assoluti delle radici di $A(L)$ erano tutte maggiori di 1. Consideriamo allora un VAR di ordine 1:

$$y_t = Ay_{t-1} + \epsilon_t \tag{4.1}$$

In questo caso $A(L) = I - AL$ è un polinomio matriciale di primo grado in L. Poiché $y_{t-1} = Ay_{t-2} + \epsilon_{t-1}$, posso sostituire questa espressione nella (4.1), ottenendo

$$y_t = A^2 y_{t-2} + \epsilon_t + A\epsilon_{t-1}$$

dove $A^2 = A \cdot A$; ripetendo questo procedimento n volte si ha

$$y_t = A^{n+1} y_{t-n-1} + \epsilon_t + A\epsilon_{t-1} + \cdots + A^n \epsilon_{t-n}$$

Al crescere di n, il primo addendo 'scompare' se $\lim_{n\to\infty} A^n = 0$; questo accade se tutti gli autovalori di A (ossia i valori di λ che rendono vera l'espressione $|A - \lambda I| = 0$) sono minori di 1 in valore assoluto[1]. Si può dimostrare che questa condizione sugli autovalori di A è necessaria e sufficiente perché il processo sia stazionario in covarianza. Essa può anche essere espressa in modo equivalente dicendo che il processo è stazionario se $|A(z)| = 0$ non ha soluzioni per $|z| \leq 1$ (provarlo può essere un buon esercizio)[2]. In questo caso,

[1] Si noti il parallelismo con un processo AR(1), in cui la rappresentazione in media mobile è ben definita se $|\alpha| < 1$.

[2] Particolare curioso: a differenza del caso univariato, non è detto che invertendo un polinomio matriciale di ordine finito se ne ottenga uno di ordine infinito. Chi vuole fare la prova, consideri

è possibile definire la rappresentazione VMA di y_t come:

$$y_t = \epsilon_t + A\epsilon_{t-1} + \cdots = \sum_{i=0}^{\infty} A^i \epsilon_{t-i}.$$

Sebbene le condizioni di stazionarietà possano essere derivate in modo abbastanza semplice anche per un VAR di ordine p, in generale lavorare con modelli VAR(p) è molto più noioso dal punto di vista algebrico che farlo con dei VAR(1). Fortunatamente, esiste un modo per scrivere un VAR di ordine p qualunque come un VAR(1), che va sotto il nome di rappresentazione in **companion form**.[3] Consideriamo ad esempio un processo VAR di ordine 3

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + A_3 y_{t-3} + \epsilon_t$$

Aggiungendo a questa equazione le due identità

$$\begin{aligned} y_{t-1} &= y_{t-1} \\ y_{t-2} &= y_{t-2} \end{aligned}$$

otteniamo un sistema di tre equazioni che è possibile scrivere in forma matriciale come segue:

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{bmatrix} = \begin{bmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \\ 0 \end{bmatrix} \tag{4.2}$$

o, in forma più abbreviata,

$$x_t = \tilde{A} x_{t-1} + \eta_t.$$

L'espressione precedente definisce un VAR(1) in cui il processo autoregressivo vettoriale non è più y_t, bensì x_t, che risulta dall'accostamento verticale di y_t, y_{t-1} e y_{t-2}. Se A è quadrata di ordine n, la matrice $\tilde{A}$ è quadrata di ordine $3n$, e η_t è un *white noise* multivariato la cui matrice di varianze-covarianze è sì singolare, ma continua ad essere simmetrica e semidefinita positiva. La condizione di stazionarietà, a questo punto, è una condizione imposta sui $3n$ autovalori di A. La generalizzazione al caso di un VAR di ordine p dovrebbe essere banale: in questo caso la matrice *companion* è fatta così:

$$\tilde{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I & 0 & \cdots & 0 \\ 0 & I & \cdots & 0 \\ \vdots & \vdots & & \ddots \end{bmatrix},$$

il seguente sistema:

$$\begin{aligned} y_t &= \theta x_{t-1} + \epsilon_{1,t} \\ x_t &= \epsilon_{2,t}. \end{aligned}$$

Bizzarro, eh?

[3] Quando posso, io parlo in italiano, ma qui non ce la faccio. La locuzione 'forma compagna', che qualcuno usa, a me evoca barbuti in eskimo che parlano attraverso un megafono. Ma che ne sapete voi, sbarbatelli?

che ha $n \cdot p$ autovalori: tutti devono essere minori di 1 in modulo perché il VAR sia stazionario.

Esempio 4.2.1 (Un AR(2) riscritto come un VAR(1)) *Naturalmente, si può pensare ad un processo AR come un caso particolare di VAR in cui la dimensione del processo stocastico è 1. Prendiamo un AR(2) e riscriviamolo in* companion form, *controllando l'equivalenza delle condizioni di stazionarietà: se*

$$y_t = 1.3y_{t-1} - 0.4y_{t-2} + \epsilon_t,$$

allora il polinomio di cui dobbiamo trovare le radici è $A(z) = 1 - 1.3z + 0.4z^2$*; poiché il polinomio è di secondo grado,* $A(z) = 0$ *si risolve facilmente con la formula*

$$z = \frac{1.3 \pm \sqrt{1.69 - 1.6}}{0.8} \Longrightarrow \begin{cases} z_1 = 2 \\ z_2 = 1.25 \end{cases}$$

Le radici del polinomio sono pertanto maggiori di 1 in valore assoluto. Il processo è stazionario.

Proviamo adesso a scriverlo in companion form*: avremo*

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1.3 & -0.4 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}$$

e cioè un VAR(1) in x_t*, dove*

$$x_t = \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix}.$$

In base a quanto abbiamo detto sopra, il processo è stazionario se gli autovalori della matrice A sono minori di uno in valore assoluto. Gli autovalori di A sono quei numeri λ *che soddisfano l'equazione*

$$\begin{vmatrix} 1.3 - \lambda & -0.4 \\ 1 & -\lambda \end{vmatrix} = 0$$

Calcolando il determinante si arriva ad un'equazione di secondo grado:

$$\lambda^2 - 1.3\lambda + 0.4 = 0,$$

le cui soluzioni sono date da

$$\lambda = \frac{1.3 \pm \sqrt{1.69 - 1.6}}{2} \Longrightarrow \begin{cases} \lambda_1 = 0.8 \\ \lambda_2 = 0.5 \end{cases}$$

Si noti che $\lambda_1 = z_2^{-1}$ *e* $\lambda_2 = z_1^{-1}$ *(non è un caso). Comunque, poiché ambedue gli autovalori sono minori di 1 in valore assoluto, concludiamo che il processo è stazionario.*

Questo esempio mi dà anche un buon pretesto per illustrare una cosa di cui parlavo poco fa, e cioé che una matrice di polinomi si può manipolare algebricamente come una matrice "normale". Se infatti partissimo dalla companion form, *si mostra che* y_t *è un AR(2) con una banale inversione di matrice. Infatti*

$$\left(I - \begin{bmatrix} 1.3 & -0.4 \\ 1 & 0 \end{bmatrix} L \right) x_t = \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix},$$

da cui

$$x_t = \begin{bmatrix} 1-1.3L & 0.4L \\ -L & 1 \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix}.$$

Se ora applichiamo le regole per invertire una matrice trattando L come se fosse un numero, scopriamo che

$$\begin{bmatrix} 1-1.3L & 0.4L \\ -L & 1 \end{bmatrix}^{-1} = \frac{1}{1-1.3L+0.4L^2} \begin{bmatrix} 1 & -0.4L \\ L & 1-1.3L \end{bmatrix},$$

da cui

$$(1-1.3L+0.4L^2) \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & -0.4L \\ L & 1-1.3L \end{bmatrix} \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix},$$

e quindi, appunto, $y_t = 1.3y_{t-1} - 0.4y_{t-2} + \epsilon_t$.

Scrivere un VAR in *companion form* è utile per tante cose: una di queste è il calcolo delle matrici della rappresentazione VMA. Esemplifico con un VAR di ordine 2 perché una volta capito il principio la generalizzazione è banale. Supponiamo di partire dal processo

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \epsilon_t$$

in cui le matrici A_1 e A_2 sono note, e di voler calcolare la sequenza di matrici C_i relative alla sua rappresentazione MA, ossia

$$y_t = \epsilon_t + C_1\epsilon_{t-1} + C_2\epsilon_{t-2} + \cdots$$

Come abbiamo visto, nel caso di un VAR(1) si ha $C_i = A^i$. Di conseguenza, riscrivendo il VAR in *companion form* come nella 4.2 si ha

$$x_t = \tilde{A}x_{t-1} + \eta_t = \eta_t + \tilde{A}\eta_{t-1} + \tilde{A}^2\eta_{t-2} + \cdots$$

ossia

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \epsilon_t \\ 0 \end{bmatrix} + \tilde{A} \begin{bmatrix} \epsilon_{t-1} \\ 0 \end{bmatrix} + \tilde{A}^2 \begin{bmatrix} \epsilon_{t-2} \\ 0 \end{bmatrix} + \cdots;$$

dovrebbe essere evidente dall'espressione di cui sopra che C_i è semplicemente il blocco $n \times n$ che occupa l'angolo a nord-ovest di $\tilde{A}^i$; in pratica, per calcolare i primi k elementi della sequenza delle matrici C_i basta costruire la matrice *companion*, moltiplicarla per se stessa k volte e prendere ogni volta l'angolo in alto a sinistra.

Esempio 4.2.2 *Supponiamo che*

$$A_1 = \begin{bmatrix} 0.8 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.1 & 0.3 \\ -0.2 & -0.2 \end{bmatrix}.$$

La matrice companion *è, naturalmente,*

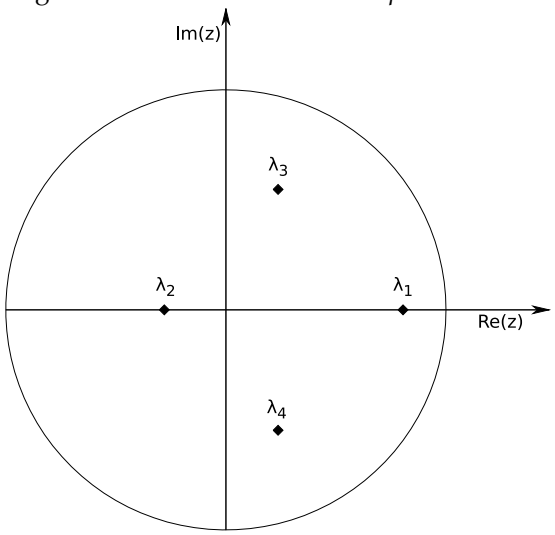
$$\tilde{A} = \begin{bmatrix} 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Le quantità di nostro interesse sono innanzitutto i suoi autovalori: calcolarli a mano è duretta assai, ma con l'aiuto dei nostri amici elaboratori ci vuole relativamente poco ad appurare che sono pari a

$$\lambda = [0.80481, -0.27953, 0.23736 \pm 0.54705i]$$

Questi possono essere rappresentati graficamente come punti sul piano complesso (vedi fig. 4.1). Si noti che tutti sono compresi all'interno del cerchio unitario, per cui il VAR è stazionario.

Figura 4.1: Autovalori della *companion matrix*



Calcoliamo ora le matrici della rappresentazione in media mobile. Le prime potenze della matrice companion *sono:*

$$\tilde{A}^0 = I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\tilde{A}^1 = \tilde{A} = \begin{bmatrix} 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\tilde{A}^2 = \tilde{A} \cdot \tilde{A} = \begin{bmatrix} 0.62 & -0.3 & -0.04 & 0.36 \\ 0.4 & -0.28 & 0.06 & 0.02 \\ 0.8 & -0.6 & 0.1 & 0.3 \\ 0.2 & 0.2 & 0.2 & -0.2 \end{bmatrix}$$

$$\tilde{A}^3 = \tilde{A} \cdot \tilde{A} \cdot \tilde{A} = \begin{bmatrix} 0.396 & -0.072 & 0.002 & 0.24600 \\ 0.324 & -0.276 & -0.016 & 0.176 \\ 0.62 & -0.3 & -0.04 & 0.36 \\ 0.4 & -0.28 & 0.06 & 0.02 \end{bmatrix}$$

Vedete? Poiché in basso a sinistra la matrice companion *ha una matrice identità, le righe scorrono verso il basso man mano che si itera.*

Prendendo i rispettivi blocchi 2×2 *in alto a sinistra, otteniamo le matrici della rappresentazione MA:*

$$C_0 = I \quad C_1 = \begin{bmatrix} 0.8 & -0.6 \\ 0.2 & 0.2 \end{bmatrix} \quad C_2 = \begin{bmatrix} 0.62 & -0.3 \\ 0.4 & -0.28 \end{bmatrix} \quad C_3 = \begin{bmatrix} 0.396 & -0.072 \\ 0.324 & -0.276 \end{bmatrix}$$

Dovrebbe essere evidente che questa procedura è il classico caso in cui i conti sono brutti e noiosi da fare a mano, ma molto facili da far fare a un elaboratore.

4.3 Stima dei VAR

Comincio dalla fine: i parametri di un VAR si possono stimare in modo consistente con una serie di regressioni OLS. Vediamo perché. Un VAR n-variato di ordine p può essere considerato un sistema di n equazioni dalla forma

$$y_{it} = \sum_{j=1}^{p} \left(a_{i1j} y_{1t-j} + \cdots + a_{inj} y_{nt-j} \right) + \epsilon_{it} \tag{4.3}$$

Per $n = 2$ e $p = 1$, si avrebbe ad esempio

$$\begin{aligned} y_{1t} &= a_{11} y_{1t-1} + a_{12} y_{2t-1} + \epsilon_{1t} \\ y_{2t} &= a_{21} y_{1t-1} + a_{22} y_{2t-1} + \epsilon_{2t} \end{aligned}$$

Il fatto che normalmente p non è noto può essere affrontato con metodi sostanzialmente non differenti da quelli di cui ho parlato nel paragrafo 2.7: in due parole, si fanno dei test preliminari che ci consentono di porre un limite al numero di ritardi necessario perché un VAR riesca a dar conto della persistenza presente nei dati. D'ora in poi, facciamo finta che l'ordine del VAR sia noto.

A questo punto, ognuna delle n equazioni che compongono la (4.3) potrebbe essere vista come un modello di regressione dinamica (vedi la discussione alla fine del sottoparagrafo 2.7.3); in questo caso, si può dimostrare che l'applicazione degli OLS produce stime consistenti e asintoticamente normali di tutti i parametri a_{ij}. Da un punto di vista econometrico, la stima di un VAR è un'operazione che può essere interpretata come la stima della forma ridotta di un modello ad equazioni simultanee. A differenza di quest'ultimo, però, un VAR non contiene restrizioni di identificazione, in quanto lo scopo di chi stima un VAR (come più in generale di chi usa modelli di analisi delle serie storiche) non è quello di spiegare il perché e il percome delle cose di questo mondo, ma solo di trovare una descrizione statisticamente accurata delle caratteristiche di persistenza di un insieme di serie. È per questo motivo che, al tempo della loro comparsa sulla scena, i modelli VAR vennero etichettati come modelli "a-teorici".

Purtroppo, i sistemi di equazioni simultanee sono considerati irrimediabilmente *démodé*, per cui non tutti li hanno studiati; vi dò un rapido promemoria.
Un sistema di equazioni simultanee si può rappresentare in due modi: nella **forma strutturale** il sistema può essere rappresentato come

$$\Gamma y_t = Bx_t + u_t,$$

in cui il vettore y_t contiene n variabili endogene, il vettore x_t contiene k esogene e u_t è un vettore di disturbi. Le matrici Γ e B contengono parametri comportamentali, a cui ci piace dare un'interpretazione economica, e sono quelli che vorremmo poter stimare. Il problema è che i parametri contenuti in queste matrici, però, non si possono stimare consistentemente con gli OLS, ciò che conduce a definire la **forma ridotta** del sistema:

$$y_t = \Pi x_t + w_t.$$

Nella forma ridotta, si ha semplicemente $\Pi = \Gamma^{-1}B$ e $w_t = \Gamma^{-1}u_t$. A differenza della forma strutturale, la forma ridotta si può stimare usando il metodo OLS per ognuna delle equazioni del sistema, ma i coefficienti contenuti nella matrice Π non hanno un'interpretazione economica. Una volta però ottenuta una stima consistente di Π (chiamiamola $\widehat{\Pi}$) potremmo definire in modo implicito degli stimatori consistenti di Γ e B (chiamiamoli $\widehat{\Gamma}$ e $\widehat{B}$) come quelle statistiche che soddisfano la relazione $\widehat{\Gamma}\widehat{\Pi} = \widehat{B}$. Tuttavia, il numero di elementi della matrice Π è minore del numero di parametri contenuti in Γ e B, e quindi è impossibile definire queste statistiche in modo univoco, a meno che non si pongano dei vincoli sulle matrici Γ e B. Le cosiddette "condizioni di identificazione" non sono altro che l'imposizione di un certo numero di vincoli sugli elementi di Γ e B. Ma forse è meglio che vi prendiate un testo (serio) di econometria.

L'eventuale presenza di regressori aggiuntivi di norma non costituisce un problema, nella misura in cui può essere fatta ricadere nelle fattispecie coperte dai teoremi asintotici che riguardano le regressioni dinamiche. È pertanto possibile (e si fa pressoché sempre) aggiungere parti deterministiche che tengano conto di alcune caratteristiche dei dati, come ad esempio una costante se le y_{it} hanno media non nulla, o un trend, o variabili *dummy* per segnalare eventi eccezionali o effetti stagionali.

Per quanto riguarda la stima della matrice Σ, anche in questo caso le cose sono piuttosto agevoli. Infatti, l'applicazione del metodo OLS a tutte le equazioni produce n serie di residui $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$; si può mostrare che

$$\frac{1}{T}\hat{\epsilon}_i'\hat{\epsilon}_j \xrightarrow{\text{p}} \Sigma_{ij}$$

e quindi la covarianza campionaria fra i residui è uno stimatore consistente dell'elemento ij della matrice varianze-covarianze di ϵ_t.

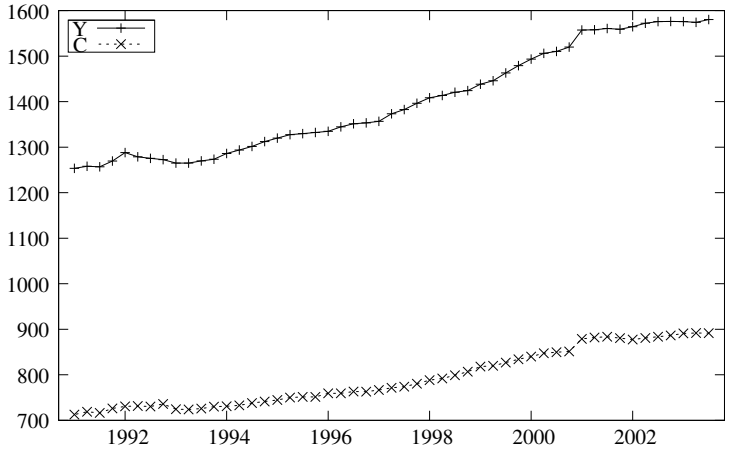
Esempio 4.3.1 *Prendiamo i dati su reddito e consumo per l'Unione Europea dal primo trimestre 1991 al terzo trimestre 2003 (la fonte è la BCE, i dati sono a prezzi costanti e destagionalizzati). Le serie sono mostrate nella figura 4.2.*

Passiamo i dati in logaritmo, e decidiamo tanto per fare un esempio che una rappresentazione statisticamente appropriata dei dati sia un VAR di ordine 1, il cui nucleo deterministico contiene una costante ed un trend[4]*. In pratica, supporremo che i nostri dati siano una realizzazione del seguente processo stocastico:*

$$\begin{bmatrix} c_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu_{0c} + \mu_{1c} \cdot t \\ \mu_{0y} + \mu_{1y} \cdot t \end{bmatrix} + A \begin{bmatrix} c_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix},$$

[4]Chi sa già queste cose adesso mi salterà alla gola, perché non ho tenuto conto di questo e di quello; in particolare, tutti i problemi di cui ho parlato nel capitolo 3 e di cui mi accingo a parlare nel capitolo 5 sono bellamente ignorati. Risposta: lo so. Mi serve un esempio maneggevole.

Figura 4.2: PIL e Consumi nell'UE



dove ovviamente A è una matrice 2×2 di parameteri. Questi, più i vari μ, possono essere stimati con i minimi quadrati ordinari. Si perviene ai risultati mostrati in tavola 4.1. Il valore stimato della matrice A è, pertanto,

$$\widehat{A} = \begin{bmatrix} 0.4453 & 0.5567 \\ -0.2010 & 1.1199 \end{bmatrix}.$$

Tabella 4.1: Risultati delle stime OLS

Equazione per c_t				
	Coeff.	S.E.	t-stat.	p-value
μ_{0c}	-0.3218	0.4134	-0.7780	0.4403
μ_{1c}	-0.0002	0.0003	-0.4940	0.6235
c_{t-1}	0.4453	0.1454	3.0620	0.0037
y_{t-1}	0.5567	0.1645	3.3830	0.0015
	$R^2 = 0.9947$		$\hat{\sigma} = 0.0057111$	

Equazione per y_t				
	Coeff.	S.E.	t-stat.	p-value
μ_{0y}	0.4663	0.3777	1.2350	0.2232
μ_{1y}	0.0004	0.0003	1.4260	0.1606
c_{t-1}	-0.2010	0.1329	-1.5130	0.1371
y_{t-1}	1.1199	0.1503	7.4510	0.0000
	$R^2 = 0.9961$		$\hat{\sigma} = 0.00521754$	

Dalle regressioni appena viste estraiamo i residui. Calcolando la loro matrice varianze-covarianze campionaria, si ottiene la stima della matrice Σ:

$$\hat{\Sigma} = \begin{bmatrix} 3.2617 \cdot 10^{-5} & 2.1389 \cdot 10^{-5} \\ 2.1389 \cdot 10^{-5} & 2.7223 \cdot 10^{-5} \end{bmatrix},$$

4.4 VAR integrati

Nel capitolo 3 ci siamo dilungati sulle caratteristiche dei processi integrati e sui motivi per cui questo tipo di processi è così importante da studiare. A questo punto, chiediamoci se e cosa si possa generalizzare ai VAR.

Come argomentato nella sezione 4.2, dato un VAR a n variabili

$$A(L)y_t = \epsilon_t,$$

esso è stazionario se le soluzioni dell'equazione $|A(z)| = 0$ sono tutte maggiori di uno in modulo; una condizione equivalente la si ha sugli autovalori della *companion matrix*, che devono essere tutti minori di uno. Che succede se $|A(z)| = 0$ ha qualche soluzione per $|z| = 1$?

In ambito univariato, non c'è possibilità intermedia: o $A(1)$ è zero, o non lo è. In ambito multivariato, diventa importante chiedersi *quante* soluzioni esistono o, equivalentemente, quanti autovalori unitari ha la matrice *companion*.

Come abbiamo detto, se tutti gli autovalori sono minori di 1, non c'è problema perché il processo è stazionario. Se invece ce ne sono uno o più, ci sono vari casi di interesse. Il primo caso sorge quando $A(1)$ è una matrice di zeri. In questo caso, ci fa comodo usare la scomposizione BN, che è valida anche nel caso di processi multivariati:

$$A(L) = A(1) + A^*(L)\Delta$$

dove ovviamente $A(L)$ è un polinomio matriciale, e di conseguenza, sono matrici anche $A(1)$ e $A^*(L)$. Se applichiamo questa scomposizione al nostro VAR, otteniamo

$$A^*(L)\Delta y_t = \epsilon_t,$$

Se Δy_t è stazionario, concludiamo che l'intero processo doveva essere differenziato una volta, e quindi è $I(1)$. In queste circostanze, non c'è altro da fare che effettuare la nostra analisi empirica sulle serie in differenze anziché in livelli. In questo caso, si può mostrare che il numero di autovalori unitari della matrice *companion* è esattamente n.[5]

In questi casi, naturalmente, non esiste una rappresentazione di Wold, ma nessuno impedisce di calcolare una rappresentazione in media mobile non-Wold tramite la funzione di risposta di impulso come una successione di matrici C definite come

$$(C_n)_{ij} = \frac{\partial y_{it}}{\partial \epsilon_{jt-n}};$$

Bene: in questi casi si dimostra che la sequenza C_n non converge a 0 come nel caso stazionario, ma ad una matrice limite non nulla. Si noti il parallelo con un processo univariato $I(1)$, dove la funzione di risposta di impulso rimane indefinitamente a un livello diverso da 0.

[5] Attenzione: il converso non è vero. Si possono costruire esempi di matrici *companion* con n autovalori unitari senza che il processo risultante sia $I(1)$.

Esempio 4.4.1 *Considerate il seguente VAR(2):*

$$\begin{aligned} y_t &= 1.1y_{t-1} - 0.1x_{t-1} - 0.1y_{t-2} + 0.1x_{t-2} + \epsilon_{1t} \\ x_t &= 1.3x_{t-1} - 0.3x_{t-2} + \epsilon_{2t}. \end{aligned}$$

Che x_t sia $I(1)$ si vede a occhio, visto che con due passaggi semplici semplici si arriva a $\Delta x_t = 0.3\Delta x_{t-1} + \epsilon_{2t}$, che è un AR(1) stazionario. Per quanto riguarda y_t, la cosa è un po' più complicata, ma anche qui si può arrivare a scrivere

$$\Delta y_t = 0.1\Delta y_{t-1} - 0.1\Delta x_{t-1} + \epsilon_{1t};$$

visto che $\Delta x_t \sim I(0)$, è facile concludere che Δy_t è anch'esso stazionario, per cui $y_t \sim I(1)$.

Il polinomio $A(L)$ si può scrivere come

$$A(L) = \begin{bmatrix} 1 - 1.1L + 0.1L^2 & 0.1L - 0.1L^2 \\ 0 & 1 - 1.3L + 0.3L^2 \end{bmatrix}$$

da cui è facile dedurre che

$$A(1) = \begin{bmatrix} 1 - 1.1 + 0.1 & 0.1 - 0.1 \\ 0 & 1 - 1.3 + 0.3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Applicando la scomposizione BN, possiamo scrivere

$$A(L) = A(1) + A^*(L)\Delta = 0 + \begin{bmatrix} 1 - 0.1L & 0.1L \\ 0 & 1 - 0.3L \end{bmatrix}\Delta,$$

cosicché

$$\begin{aligned} \Delta y_t &= 0.1\Delta y_{t-1} - 0.1\Delta x_{t-1} + \epsilon_{1t} \\ \Delta x_t &= 0.3\Delta x_{t-1} + \epsilon_{2t}. \end{aligned}$$

che è un bel VAR(1) stazionario.

Per finire, consideriamo la matrice companion*:*

$$\tilde{A} = \begin{bmatrix} 1.1 & -0.1 & -0.1 & 0.1 \\ 0 & 1.3 & 0 & -0.3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Come si può controllare, i suoi autovalori sono 1, 1, 0.3 e 0.1.

Infine, volendo rappresentare il processo in forma VMA, facendo il giochino delle potenze di $\tilde{A}$ come nell'esempio 4.2.2, si ottiene che

$$C_0 = I, \quad C_1 = \begin{bmatrix} 1.1 & -0.1 \\ 0 & 1.3 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 1.11 & -0.14 \\ 0 & 1.39 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 1.111 & -0.153 \\ 0 & 1.417 \end{bmatrix}$$

e così via; chi volesse continuare, troverà che C_n non converge a 0 per $n \to \infty$.

La cosa si fa più interessante se il numero di autovalori unitari fosse compreso fra 0 e n; Questo caso bizzarro, in cui $A(1)$ non è zero ma non è neanche invertibile, conduce alla cosiddetta **cointegrazione**, in cui il processo è $I(1)$, ma ha delle caratteristiche peculiari che esamineremo nel capitolo 5. Un capitolo a parte è necessario perché il fatto che $A(1)$ abbia rango ridotto ha importanti conseguenze sia sotto il profilo inferenziale che, soprattutto, sotto quello interpretativo. Non per niente la cointegrazione è stata il grande *hot topic* dell'econometria delle serie storiche di fine '900.

Dal punto di vista della stima, per fortuna l'eventuale presenza di integrazione produce meno disastri di quanto non si possa temere. In particolare:

- L'uso dei criteri di informazione per la scelta dell'ordine del VAR funziona sempre, indipendentemente dal fatto che il VAR sia stazionario o meno;
- il metodo OLS produce in ogni caso stime consistenti dei parametri, anche se la distribuzione limite delle statistiche OLS può diventare non-standard in presenza di radici unitarie. Se ci interessa solo una stima puntuale (ad esempio, per usare il VAR stimato solo come macchina previsiva), il problema non si pone. Se però ci serve fare dei test (come per esempio il cosiddetto test di Granger-causalità, di cui parlerò nella sezione 4.5.2), allora bisogna fare attenzione.

Questi risultati (più tanti altri) sono tutti contenuti in un articolo (Sims *et al.* (1990)), che a mio modesto parere è uno dei più begli articoli di econometria delle serie storiche di sempre. In estrema sintesi, si fa vedere nell'articolo che la consistenza degli stimatori OLS per i parametri di un VAR non è messa a repentaglio dall'esistenza di eventuali radici unitarie, purché la dinamica del VAR stimato sia abbastanza ampia da permettere ai disturbi di essere — più o meno — dei *white noise*. Nello stesso articolo poi c'è una discussione molto interessante delle conseguenze che la presenza di radici unitarie ha sulla distribuzione di tali stimatori, e le proprietà distribuzionali dei test DF e ADF emergono come caso particolare in modo molto elegante.

4.5 Uso dei VAR

I VAR — come i loro fratelli minori univariati — vengono ampiamente usati per la previsione e per l'analisi delle caratteristiche dinamiche delle serie che li compongono. Gli usi che si possono fare delle stime di un VAR sono molti, ma qui voglio parlare di tre applicazioni, che sono quelle più comuni in macroeconometria:

1. Previsione
2. Analisi di causalità
3. Analisi dinamica

Come si vedrà, molti dei concetti che adopereremo sono delle naturali estensioni di quelli già analizzati nell'ambito dei processi univariati. La natura

multivariata dei VAR, tuttavia, apre delle prospettive interessanti nonché dei problemi che nel caso univariato non si pongono.

4.5.1 Previsione

Il primo punto non richiede spiegazioni che non richiamino le considerazioni già fatte sull'uso dei processi AR in sede previsiva (vedi 2.6.1). Utilizzando come previsore il valore atteso condizionale, l'unica avvertenza da fare qui è che naturalmente il set informativo sul quale effettuare il condizionamento comprende il passato di più serie (tutte quelle che compongono il VAR) anziché di una sola.

Una volta stimate le matrici $\hat{A}_1, \ldots, \hat{A}_p$ coi metodi di cui dicevamo poc'anzi, la previsione di y_{T+k} sarà data da una semplice generalizzazione del caso univariato

$$\hat{y}_{T+k} = \hat{A}_1 \hat{y}_{T+k-1} + \cdots + \hat{A}_p \hat{y}_{T+k-p}$$

dove anche in questo caso $\hat{y}_{T+k} = y_{T+k}$ per $k \leq 0$. Esistono poi delle espressioni — che qui non riporto — per calcolare anche la deviazione standard dei valori previsti, così da poter impostare l'attività di previsione in termini propriamente statistici.

Come si vede, la questione è molto semplice. Il fatto poi che i parametri contenuti nelle matrici A_i possano essere stimati con tecniche molto semplici (gli OLS) ha fatto sì che l'analisi VAR sia da almeno vent'anni la tecnica standard per la previsione con modelli macroeconomici di piccole dimensioni. Naturalmente, questa non è tutta la storia. Esistono modi di lavorare coi VAR più raffinati, quando si tratta di fare previsioni, ma esulano dal nostro ambito.

Altra piccola digressione (parente stretta di quella fatta nel sottoparagrafo 2.6.1). L'uso dei VAR come strumento per la previsione è l'oggetto di un certo grado di ironia nella professione. Effettivamente, un modello VAR "base" come quelli qui presentati costituisce una stilizzazione dei fatti empirici molto drastica, tutto sommato inadeguata a catturare fenomeni complessi; basti pensare, ad esempio, al fatto che in un VAR non c'è modo semplice per inserire le informazioni a priori che possiamo avere sul futuro, come ad esempio cambiamenti di regime nella politica economica e così via.

C'è una pagina web[6] di barzellette sugli economisti in cui ce n'è una che calza a pennello:

> Forecasting is like trying to drive a car blindfolded and following directions given by a person who is looking out of the back window[7].

In realtà, questa critica è ingenerosa: nessuno si sognerebbe di tacciare per meccanica o ridicola un'affermazione del tipo "un avanzo nella bilancia commerciale oggi contribuirà a ridurre la disoccupazione fra un anno, nella misura in cui gli avanzi (disavanzi) passati hanno influenzato la disoccupazione a un anno di distanza". Certamente, una affermazione così è parziale e schematica, e non tiene conto di tante cose, ma può rappresentare un'ottima base per ragionare più di fino.

Un VAR è uno strumento per rendere "semiautomatici" ragionamenti di questo genere. Una previsione ottenuta con un modello necessita sempre di essere vagliata alla luce delle caratteristiche qualitative del fenomeno. In questo senso, si sarebbe quasi tentati di sostenere il punto apparentemente paradossale secondo cui servono di più le previsioni sbagliate che

[6] `http://netec.mcc.ac.uk/JokEc.html`

[7] Fare previsioni è come tentare di guidare bendato seguendo le istruzioni di uno che guarda dal lunotto.

quelle giuste, perché è attraverso le prime che abbiamo segnali se il mondo sta cambiando (o se, più banalmente, il nostro modello previsivo ha qualche problema).

4.5.2 Analisi di causalità

Un'altra applicazione per la quale i VAR sono molto usati è l'analisi della causalità. In generale, le relazioni di causa-effetto sono molto complesse da stabilire in un'analisi empirica di dati economici. Se osserviamo un'alta correlazione fra due variabili X e Y, possiamo dire tutt'al più che quelle due variabili presentano una spiccata tendenza a muoversi insieme, ma in assenza di altre informazioni non possiamo dire nulla sui nessi causali che le collegano. Potrebbe darsi che X sia la causa di Y, che Y sia la causa di X o addirittura che ci sia una terza variabile Z (non osservata o non considerata) che sia la causa comune di entrambe. Tutte e tre queste situazioni darebbero luogo allo stesso fenomeno osservabile, cioè un alto grado di correlazione fra X e Y.

A volte è la teoria economica a venirci in aiuto: se, ad esempio, osservassimo che il prezzo di un bene cresce sistematicamente al crescere della quantità scambiata, potremmo suggerire un'interpretazione basata su uno spostamento verso destra della curva di domanda, cui corrisponde una curva di offerta stabile. In questo caso, avremmo buon gioco a sostenere che è stato l'incremento di domanda a far aumentare la quantità, e, *di conseguenza*, il prezzo del bene.

In molte circostanze, tuttavia, la teoria non offre indicazioni univoche: in tali casi, esiste una definizione di causalità che offre la possibilità di determinare il senso del nesso causa-effetto su basi puramente statistiche, ed è basata sul seguente principio: **la causa precede sempre l'effetto**. Si suppone, in altri termini, che se X causa Y, il nesso causale richieda per prodursi un tempo minimo, durante il quale osserviamo lo spostamento di X, e solo dopo il suo effetto, cioè lo spostamento di Y. Viceversa, se X non causasse Y, variazioni in X non dovrebbero produrre variazioni sistematiche nei valori *futuri* di Y.

Volendo essere più precisi, si può definire la **causalità secondo Granger**, o **Granger-causalità** in questo modo[8]:

$$X \; GC \; Y \iff E(y_t|y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots) \neq E(y_t|y_{t-1}, y_{t-2}, \ldots)$$

ossia se le osservazioni sul passato di X sono di una qualche utilità nel predire Y[9]; GC si legge *Granger-causa*, voce del raccapricciante verbo *Granger-causare*; il suo contrario è NGC(non Granger-causa).

[8]Questa definizione è stata introdotta negli anni '70 da C.W.J. Granger, e quindi prende il nome da lui. Granger, comunque, ha precisato più di una volta che ritiene questo concetto di causazione valido solo in prima approssimazione, e che ne stigmatizza l'uso indiscriminato. Il fatto inoltre che più di una volta egli stesso abbia affermato di aver semplicemente introdotto in econometria una definizione del matematico Norbert Wiener fa pensare più ad una presa di distanza che non ad un impeto di modestia.

[9]Si noti che dalla definizione di Granger-causalità consegue se $X \; GC \; Y$, non è detto che $Y \; NGC \; X$. Del pari, se $X \; NGC \; Y$, non è detto che $Y \; GC \; X$.

In un VAR bivariato, tale definizione si traduce immediatamente: infatti se il vettore $z_t = (y_t, x_t)$ può essere rappresentato come un VAR

$$\begin{bmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{bmatrix} \begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

si ha che $x \; NGC \; y \iff A_{12}(L) = 0$; si noti che, se il VAR è stazionario, un test dell'ipotesi $x \; NGC \; y$ può essere condotto tramite un semplice test F: infatti, se scriviamo per esteso la prima equazione del VAR,

$$y_t = \alpha_1 y_{t-1} + \beta_1 x_{t-1} + \alpha_2 y_{t-2} + \beta_2 x_{t-2} + \cdots + \alpha_p y_{t-p} + \beta_p x_{t-p} + \epsilon_{1t}$$

l'ipotesi di assenza di Granger-causalità da x a y è equivalente all'ipotesi

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

che è una restrizione lineare sui parametri dell'equazione. Poiché (se il VAR è stazionario) l'inferenza su questi ultimi può essere condotta in modo asintoticamente valido con i consueti strumenti OLS, il test di Granger-causalità viene ricondotto alla più generale teoria del test di ipotesi nel modello OLS.

Esempio 4.5.1 *Supponiamo di considerare il seguente VAR bivariato:*

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} 0.8 & -0.4 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}$$

con $\Sigma = E\left(\begin{bmatrix} u_t \\ v_t \end{bmatrix} \begin{bmatrix} u_t & v_t \end{bmatrix} \right) = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$.

In questo caso, si ha che X GC Y, ma non viceversa, perché i valori passati di Y non aiutano a prevedere X. Si noti, peraltro, che

$$E(x_t | y_t, \Im_{t-1}) = 0.96 x_{t-1} + 0.4 y_t - 0.32 y_{t-1}$$

e quindi i valori passati di Y hanno una utilità nel prevedere X, ma solo quando nell'insieme di condizionamento entri anche il valore contemporaneo di Y.

La semplicità della sua applicazione ha fatto sì che la Granger-causalità venisse impiegata in un numero incalcolabile di articoli applicati, spesso nel tentativo di confermare o confutare questa o quella ipotesi teorica. In realtà, il concetto di Granger-causalità è troppo meccanico per essere considerato un sinonimo dell'idea di causalità che come economisti abbiamo in mente. In particolare, due sono le critiche che vengono fatte in letteratura all'applicazione acritica di questo strumento.

La prima è una critica, per così dire, statistica, ed è imperniata sul fatto che una variabile X può essere trovata ad essere Granger-causale per un'altra variabile Y o meno a seconda di quali altre variabili siano presenti nel sistema. Per chiarire questo concetto, facciamo un esempio (che *non* c'entra con la Granger causalità): supponiamo di avere una normale trivariata (Y, X, Z) a media 0 e con matrice varianze-covarianze $\begin{bmatrix} 1 & 0 & \alpha \\ 0 & 1 & \beta \\ \alpha & \beta & 1 \end{bmatrix}$. Si vede subito che

$$E(Y|X) = 0 \qquad E(Y|X,Z) = \frac{-\alpha\beta}{1-\beta^2} X + \frac{\alpha}{1-\beta^2} Z$$

e quindi concludere che, se α e β sono diversi da 0, X 'non spiega' Y è corretto solo se non si considera Z. In termini di Granger-causalità, un test che accetti (o rifiuti) la nulla di assenza di Granger-causalità in un VAR bivariato potrebbe rifiutarla (o accettarla) in un VAR trivariato, e quindi i test devono sempre essere considerati validi *all'interno del set di condizionamento che si è scelto*. Modificando il contenuto del vettore di variabili che costituisce il VAR i risultati del test di Granger-causalità possono essere diversi[10]. Pare[11], ad esempio, che la Banca Centrale Europea utilizzi sistematicamente la produzione industriale belga come previsore della congiuntura tedesca. Ovviamente, non è che il Belgio traini la Germania; molto più semplicemente, lavorano su commessa.

La seconda critica è più, come si diceva una volta, a monte; essa fa riferimento al fatto che il concetto *logico* di causa-effetto prescinde da ciò che accade nel tempo fisico. In particolare, è possibile che la causa si manifesti solo dopo l'effetto, quando questo è influenzato dalle aspettative. In questo senso, aveva probabilmente ragione Maddala a dire che ci sarebbero stati meno problemi se Granger avesse usato la parola *precedenza* anziché la parola causalità. L'esempio che si fa di solito è un po' abusato, ma rende l'idea: il fatto che gli acquisti natalizi vengano fatti prima di Natale non ci autorizza a dire che la celebrazione del Natale il 25 dicembre sia causata dall'aumento di vendite di trenini e cravatte[12].

Un esempio più interessante per noi economisti è dato dalla storia economica recente: è abbastanza naturale pensare che la causa della discesa dei tassi di interesse in Italia fra il 1995 e il 1997 sia dovuta al fatto che l'ingresso dell'Italia nell'Unione Monetaria Europea è stata vista dagli operatori come sempre meno improbabile. Tuttavia, se applicassimo rigidamente il criterio *post hoc, ergo propter hoc*[13], dovremmo concludere che l'ingresso dell'Italia nell'UME è stato *provocato* dalla discesa dei tassi. In parte questo è vero (la riduzione dei tassi ha provocato un alleggerimento del servizio del debito, e di conseguenza un miglioramento del bilancio pubblico), ma sicuramente questa non è tutta la storia: c'è una parte importante di spiegazione che non è compresa nell'influsso del calo dei tassi sul servizio del debito, che ha a che fare con le aspettative, con la credibilità del governo, con la congiuntura internazionale; ma questa storia la trovate — detta meglio — in tanti altri posti, e quindi non la trovate qui.

In conclusione, mi piace citare uno dei miei econometrici preferiti, Adrian Pagan, che in uno dei suoi splendidi articoli di rassegna ebbe a scrivere (la traduzione è mia):

> C'è stata molta analisi di alto livello su questo argomento, ma l'impressione che ho avuto dopo la lettura è che [la Granger causalità] sia stata una delle più spiacevoli vicende accadute all'econometria in vent'anni, e che ha probabilmente prodotto più risultati assurdi di qualunque altra cosa in questo periodo.

[10] Se si vuole, l'intero argomento può essere letto come un caso particolare del cosiddetto problema delle variabili omesse nel modello lineare.

[11] Non sono stato in grado di reperire questa informazione da fonti ufficiali, ma poiché mi viene da Beppe Parigi, che è un grande, ci credo.

[12] Un altro carino è: le previsioni del tempo causano la pioggia?

[13] Dopo la tal cosa, quindi a causa della tal cosa.

4.5.3 Analisi dinamica

Lo strumento principe per l'analisi dinamica di un processo VAR è, come nel caso univariato, la funzione di risposta di impulso[14], già definita ed analizzata nel sottoparagrafo 2.6.2. Il lettore ricorderà che, in quella sede, abbiamo motivato l'utilizzo della funzione di risposta di impulso interpretando il *white noise* che compare nella rappresentazione ARMA del processo come l'informazione aggiuntiva che, per così dire, entra nella memoria della serie ad ogni istante di rilevazione. Qui, tuttavia, le cose sono più complicate, perché abbiamo n variabili e n shock, cosicché per ogni dato periodo la risposta d'impulso è una matrice $n \times n$.

La letteratura che affronta questo tipo di tematiche è nota come letteratura sui cosiddetti **VAR strutturali**, ed ha raggiunto livelli di notevole articolazione e complessità. Qui mi limiterò a una disamina molto introduttiva.

Prendiamo un VAR stazionario

$$A(L)y_t = \epsilon_t \tag{4.4}$$

e partiamo dalla sua rappresentazione in media mobile:

$$y_t = \sum_{i=0}^{\infty} C_i \epsilon_{t-i}.$$

Tanto per stabilire la notazione, definiamo la funzione di risposta di impulso così:

$$h(i,j,n) = (C_n)_{ij} = \frac{\partial y_{it}}{\partial \epsilon_{jt-n}};$$

come interpretare questa quantità? Si può tentare un parallelo col caso univariato: $h(i,j,n)$ è la risposta dell'i-esima variabile al j-esimo shock dopo n periodi. Poiché il vettore ϵ_t rappresenta lo scarto fra y_t ed il suo valore atteso condizionale al set informativo $\Im_{t-1}$, spesso si fa riferimento a ϵ_t semplicemente chiamandolo "errore di previsione ad un passo" (vedi tutta la discussione al sottoparagrafo 2.6.2);

Tuttavia, rispetto al caso univariato, qui si può dire qualcosa in più. Infatti, con una sola serie i motivi per cui la nostra previsione è sbagliata rimangono insondabili: semplicemente, le cause vanno ricercate al di fuori del set informativo di condizionamento. È per questo che ci limitiamo a leggere lo shock come un qualche evento imprevedibile (e infatti, imprevisto).

Nel caso multivariato, possiamo aggiungere una considerazione: per ogni periodo, noi abbiamo n errori di previsione, uno per ogni serie. Di conseguenza, possiamo chiederci quali siano i motivi a monte degli errori di previsione ragionando — indirettamente — sulle loro covarianze.

Esempio 4.5.2 *Immaginiamo di osservare nel tempo il prezzo di un bene p_t e la quantità scambiata q_t e che valga la rappresentazione*

$$A(L) \begin{bmatrix} p_t \\ q_t \end{bmatrix} = \begin{bmatrix} \mu_{p,t} \\ \mu_{q,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{p,t} \\ \epsilon_{q,t} \end{bmatrix}.$$

[14] L'altro strumento che si usa di solito è la cosiddetta FEVD (*Forecast Error Variance Decomposition*), che però è eccessivamente pesante da trattare qui, e quindi rinvio come al solito ai testi sacri.

Da economisti, potremmo essere tentati di fare un ragionamento del tipo: se i segni dei due errori di previsione coincidono, il motivo principale del mio errore di previsione è uno spostamento della curva di domanda. Se invece sono diversi, evidentemente lo spostamento ha riguardato la curva di offerta.

Più precisamente: immaginiamo che le curve di domanda e di offerta siano soggette a shock casuali incorrelati fra loro. Se la curva di offerta stesse ferma, uno spostamento imprevedibile della curva di domanda produrrebbe un errore nella previsione sia del prezzo che della quantità dello stesso segno. *Chiaramente, rovesciando il discorso si vede che uno spostamento non prevedibile della curva di offerta implicherebbe errori di previsione* di segno opposto.

In pratica, entrambi gli shock strutturali sulla curve di domanda e di offerta si propagano agli errori di previsione di ambedue le equazioni.

Più formalmente, potremmo immaginare che il nostro vettore degli errori di previsione sia una funzione (che immaginiamo lineare per semplicità) dei movimenti nelle relazioni comportamentali, che chiamiamo **shock strutturali**, per cui potremmo scrivere

$$\epsilon_t = Bu_t \tag{4.5}$$

dove assumiamo che B sia quadrata e invertibile. Si noti che il vettore ϵ_t è osservabile (o per lo meno stimabile), mentre il vettore u_t e la matrice B no.

Se B fosse nota, si aprirebbero delle possibilità molto interessanti. Innanzitutto, potremmo ricostruire la storia degli shock strutturali (via $u_t = B^{-1}\epsilon_t$), ma sopratutto potremmo calcolarci le risposte di impulso strutturali: mettendo insieme le equazioni (4.4) e (4.5) si ha

$$A(L)y_t = Bu_t \tag{4.6}$$

e quindi

$$y_t = [A(L)]^{-1} Bu_t = Bu_t + C_1 \cdot Bu_{t-1} + C_2 \cdot Bu_{t-2} + \cdots$$

per cui

$$IRF(i,j,n) = \frac{\partial y_{it}}{\partial u_{jt-n}} = (C_n \cdot B)_{ij}.$$

Dovrebbe essere palese che la risposta di impulso rispetto all'errore di previsione dice poco o nulla dal punto di vista intrepretativo, mentre la risposta di impulso allo shock strutturale ci permetterebbe di valutare come rispondono nel tempo le quantità osservabili (nell'esempio precedente, prezzo e quantità) rispetto ad uno shock che impatta su una relazione comportamentale (nell'esempio precedente, la funzione di domanda o quella di offerta), e per questo vien detto "strutturale".

Come dicevamo, la matrice B va stimata. Ma questo non è affar semplice, perché l'unica statistica osservabile che può servire da base per la stima è la matrice di varianze-covarianze di ϵ_t, cioè Σ. Se, senza perdita di generalità, normalizziamo gli shock strutturali ad avere varianza 1, si deduce dalla (4.5) che

$$\Sigma = BB'; \tag{4.7}$$

visto che Σ è stimabile molto facilmente con la matrice varianze-covarianze dei residui del VAR, una matrice $\hat{B}$ per cui valga $\widehat{\Sigma} = \hat{B}\hat{B}'$ è una stima valida di B. Questo però ci aiuta solo fino ad un certo punto, poiché — come sappiamo dall'algebra delle matrici — per ogni matrice simmetrica e positiva definita Σ esistono infinite matrici B che soddisfano la (4.7). In pratica, abbiamo un problema di sottoidentificazione: esiste un'infinità non numerabile di matrici $\hat{B}$ che sono equivalenti dal punto di vista osservazionale. Se vogliamo avere *una* stima di B, dobbiamo imporre dei vincoli tali per cui l'equazione (4.7) abbia una e una sola soluzione.

In questa sede, darò soltanto un'esposizione introduttiva della prima soluzione storicamente data a questo problema: essa è anche quella più diffusa ancora oggi[15], ed è quella di **triangolarizzare** il sistema di equazioni che compone il VAR. Si può dimostrare che l'equazione (4.7) ha una sola soluzione se si impone che la matrice B sia triangolare bassa, ovvero tutti i suoi elementi b_{ij} siano nulli per $j > i$. In questo caso, la scomposizione di Σ nel prodotto di B per B trasposto prende il nome di **scomposizione di Cholesky**: essa stabilisce che qualunque matrice simmetrica e definita positiva V può sempre essere scritta come il prodotto di una matrice triangolare bassa L per la sua trasposta, ossia $V = LL'$, e che L è unica[16].

Nel nostro caso, data una stima consistente di Σ, è possibile ricavare $\hat{B}$ da $\widehat{\Sigma}$, perché B è una funzione continua di Σ. Una volta ottenute queste stime, si può passare all'analisi delle risposte di impulso, che possono essere calcolate a partire dalle matrici $\hat{C}_n\hat{B}$.

Ma qual è il significato della scomposizione triangolare? Se B è triangolare bassa, allora la (4.5) si può scrivere più per esteso a questa maniera:

$$\begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \vdots \\ \epsilon_{nt} \end{bmatrix} = \begin{bmatrix} b_{11} & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \\ \vdots \\ u_{nt} \end{bmatrix} = \begin{bmatrix} b_{11}u_{1t} \\ b_{21}u_{1t} + b_{22}u_{2t} \\ \vdots \\ b_{n1}u_{1t} + b_{n2}u_{2t} + \cdots + b_{nn}u_{nt} \end{bmatrix}.$$

Come si vede bene, la triangolarità di B implica che il primo errore di previsione sia funzione solo del primo shock strutturale, il secondo errore di previsione sia funzione dei primi due, e in generale $\epsilon_{it} = \sum_{j=1}^{i} b_{ij}u_{jt}$ (notare che l'estremo superiore della sommatoria è i). In altri termini, l'ordinamento degli elementi all'interno del vettore y_t non è arbitrario, ma segue una precisa logica: l'errore di previsione a un passo sulla prima variabile del VAR è funzione solo del primo shock strutturale, cosicché *per costruzione* il primo shock strutturale coincide (a meno di un fattore di scala) con l'errore di previsione della prima variabile. Il secondo shock strutturale, invece, risulta identificato per differenza: visto che sulla seconda variabile impattano solo i primi due shock strutturali, il secondo shock u_{2t} strutturale risulta definito come il secondo errore di previsione ϵ_{2t} "al netto di" u_{1t} (o di ϵ_{1t}, che è lo stesso).

[15] Questo avviene sia per motivi di ordine storico, sia perché la sua semplicità computazionale fa sì che molti pacchetti econometrici la implementino come opzione standard.

[16] La scomposizione di Cholesky si può calcolare a mano senza grande fatica solo nei casi in cui la matrice V è piccola. Comunque questo è lavoro per gli elaboratori.

Spero che da questo ragionamento risulti ovvio che la scelta dell'ordinamento delle variabili è assolutamente cruciale nell'interpretazione dei risultati di un VAR triangolarizzato. Attenzione, però. L'ipotesi di cui sopra è meno forte di quanto non appaia a prima vista. Infatti, se B è triangolare bassa il vincolo riguarda solo le risposte di impulso al tempo 0; dopo un periodo, ognuno degli shock strutturali può avere impatto non nullo su ognuna delle variabili.

Esempio 4.5.3 *Proseguiamo nell'esempio 4.5.2. Potrei immaginare, ad esempio, che uno shock sulla domanda del bene non impatti istantaneamente sul prezzo, magari perché i produttori hanno bisogno di un po' di tempo per rendersi conto dell'incremento delle vendite, adeguare i listini e far sì che il nuovo prezzo si propaghi lungo la catena distributiva. In questo caso, uno shock di domanda influenza — nell'immediato — solo la quantità scambiata (immaginate che il bene di cui stiamo parlando sia lo spumante e lo shock sulla domanda provenga dall'Italia che vince i mondiali).*

L'errore di previsione sul prezzo, pertanto, non conterrà lo shock di domanda, ma solo quello di offerta, mentre l'errore di previsione sulla quantità sarà funzione di tutti e due. In formule:

$$A(L)\begin{bmatrix} p_t \\ q_t \end{bmatrix} = \begin{bmatrix} \mu_{p,t} \\ \mu_{q,t} \end{bmatrix} + \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix}\begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

Notate che quest'ipotesi non implica che lo shock di domanda non impatti mai sul prezzo. Implica soltanto la neutralità istantanea *del prezzo alla domanda, vale a dire che uno shock alla domanda si fa sentire sul prezzo solo dopo almeno un periodo.*

Questo tipo di ragionamento ottiene l'identificazione imponendo una gerarchia fra le variabili che compongono il VAR in base alla loro reattività rispetto agli shock strutturali, e le variabili vengono ordinate per reattività crescente: nel gergo da sala macchine degli economisti applicati il precetto è "le variabili più esogene vanno prima". Naturalmente, c'è un certo grado di arbitrarietà in questo. Spesso, per giustificare un certo ordinamento si fa riferimento a fattori istituzionali o ad asimmetrie informative ed è chiaro che in generale l'idea che una variabile ci metta almeno un periodo a muoversi in risposta ad uno shock è tanto più difendibile quanto più il periodo è breve: nell'esempio prezzo/quantità appena visto la restrizione di identificazione (il prezzo non risponde nell'immediato a uno shock di domanda) potrebbe essere perfettamente giustificata su dati settimanali, un po' meno su dati mensili e diventa decisamente spericolata su dati trimestrali o annuali.

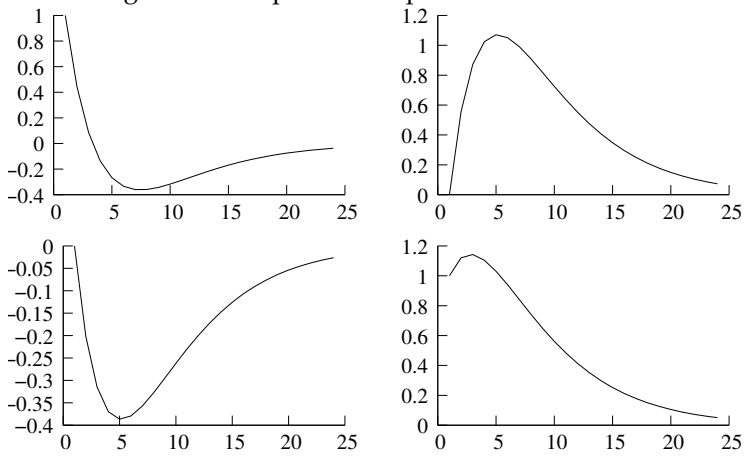
Esempio 4.5.4 *Riprendiamo l'esempio 4.3.1. Come senz'altro ricorderete, avevamo a che fare con un VAR di ordine 1, in cui la stima della matrice A era*

$$\widehat{A} = \begin{bmatrix} 0.4453 & 0.5567 \\ -0.2010 & 1.1199 \end{bmatrix}.$$

Calcolando la rappresentazione di Wold

$$(I - AL)^{-1} = I + AL + A^2L^2 + \cdots$$

Figura 4.3: Risposte di impulso non strutturali



otteniamo le risposte di impulso non strutturali; con un po' di conti si ottiene

$$A^2 = \begin{bmatrix} 0.086421 & 0.87135 \\ -0.31464 & 1.1423 \end{bmatrix} \qquad A^3 = \begin{bmatrix} -0.13667 & 1.0239 \\ -0.36974 & 1.1041 \end{bmatrix}$$

e così via. Graficamente, le risposte di impulso hanno l'aspetto presentato in figura 4.3. Il grafico in alto a sinistra mostra la risposta di c_t rispetto a ϵ_{1t}, quello in alto a destra mostra la risposta di c_t rispetto a ϵ_{2t}. La seconda riga mostra invece le risposte di impulso di y_t.

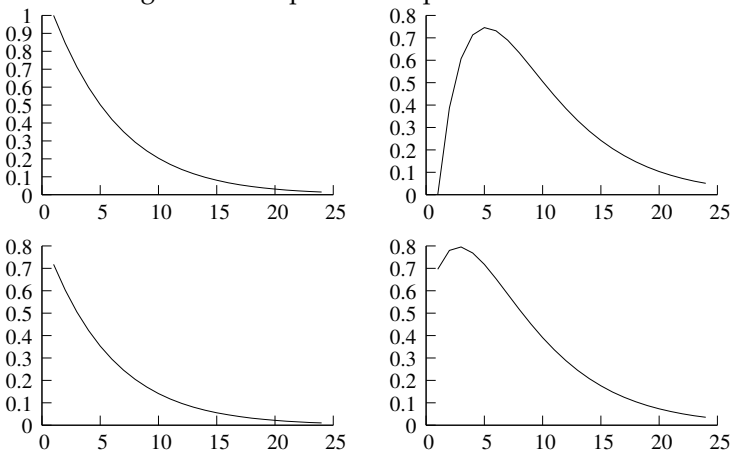
Quale può essere l'interpretazione economica della figura 4.3? Poco o niente, direi. Infatti, quelle rappresentate sono le risposte delle variabili osservate agli errori di previsione ad un passo.

Consideriamo invece la versione triangolarizzata. Il fatto che la prima variabile sia il consumo può suggerirci di interpretare il secondo shock strutturale (quello che non entra nell'equazione per c_t) come uno shock alle componenti della domanda autonoma. Ad esempio, un miglioramento inatteso della bilancia commerciale. È un'interpretazione Keynesiana molto passé*, ma per l'esempio va bene. In questo mondo che sto immaginando, le famiglie non si accorgono immediatamente dell'aumento del reddito, e iniziano a spendere solo dopo un periodo. Gli errori di previsione sul consumo sono da imputare esclusivamente al primo shock strutturale, che chiameremo di conseguenza il primo shock strutturale "shock sui consumi" e il secondo "shock autonomo".*

La scomposizione di Cholesky della matrice di correlazione di ϵ_t è uguale a

$$C = \begin{bmatrix} 1 & 0 \\ 0.71780 & 0.69625 \end{bmatrix}$$

Figura 4.4: Risposte di impulso strutturali



e le risposte di impulso strutturali sono calcolate come segue:

$$
\begin{aligned}
C_0 = A^0 \cdot C &= \begin{bmatrix} 1 & 0 \\ 0.71780 & 0.69625 \end{bmatrix} \\
C_1 = A^1 \cdot C &= \begin{bmatrix} 0.84492 & 0.38759 \\ 0.60286 & 0.77974 \end{bmatrix} \\
C_2 = A^2 \cdot C &= \begin{bmatrix} 0.71188 & 0.60668 \\ 0.50530 & 0.79532 \end{bmatrix}
\end{aligned}
$$

eccetera. Il risultato è mostrato in figura 4.4. I quattro grafici vanno interpretati tenendo conto che per riga abbiamo le variabili (consumo e PIL), per colonna abbiamo gli shock strutturali, cosicché il grafico in alto a destra mostra l'effetto sul consumo delle shock autonomo. Notate che (come abbiamo posto per ipotesi) questo grafico parte da 0.

La sua interpretazione sarebbe la seguente: dato uno shock alla domanda autonoma, le famiglie non se ne accorgono subito, e infatti la curva parte da zero. Dopo un trimestre, però, si mette in moto una specie di moltiplicatore keynesiano che porta le famiglie ad aumentare i consumi per 5-6 periodi, dopodiché l'effetto comincia a declinare.

Un metodo alternativo di ottenere l'identificazione è quello di imporre vincoli non su ciò che accade istantaneamente, ma piuttosto su ciò che accade nel lungo periodo: quest'idea è stata introdotta da Blanchard e Quah in un celebre articolo uscito nel 1989, e da allora è stata sfruttata molte volte e sviluppata in molte direzioni.

Queste restrizioni di lungo periodo, anziché essere motivate con considerazioni di tipo istituzionale, provengono da un atteggiamento più incline a prendere sul serio la teoria economica, che spesso non è in grado di fare previsioni su quel che accade nel breve periodo, ma solo sul lungo.

Per illustrare come funziona il tutto, riprendo l'idea originale di Blanchard e Quah: per descrivere lo stato dell'economia reale usiamo due variabili (sta-

zionarie per ipotesi), e cioè il tasso crescita del PIL (ΔY_t) e il tasso di disoccupazione (U_t). I movimenti nel tempo di queste due variabili originano da shock alla domanda aggregata, come ad esempio la politica fiscale o la politica monetaria, e da shock all'offerta aggregata, come ad esempio innovazioni tecnologiche.

L'idea "teorica" è che nel lungo periodo il *livello* del PIL è interamente determinato dalla tecnologia, e cioè dall'offerta aggregata. Fluttuazioni nella domanda aggregata, come quelle indotte dalla politica economica, impattano il PIL in modo soltanto transitorio.

In che modo quest'idea ci fornisce il vincolo necessario a identificare B? Per vedere come imporre un siffatto vincolo, partiamo dalla rappresentazione in media mobile:

$$y_t = \Theta(L)\epsilon_t = \epsilon_t + \Theta_1\epsilon_{t-1} + \cdots$$

che nel caso in esame è

$$\begin{bmatrix} \Delta Y_t \\ U_t \end{bmatrix} = B \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix} + \Theta_1 B \begin{bmatrix} u_{t-1}^s \\ u_{t-1}^d \end{bmatrix} + \dots$$

Se il VAR è stazionario, ovviamente $\Theta_n \to 0$. Tuttavia, a noi non interessa l'impatto degli shock su ΔY_t, bensì sul livello Y_t. Evidentemente,

$$Y_{t+k} = Y_t + \Delta Y_{t+1} + \Delta Y_{t+2} + \cdots + \Delta Y_{t+k}$$

e quindi l'impatto di lungo periodo sarà dato da

$$\lim_{k\to\infty} \frac{\partial Y_{t+k}}{\partial u_t^d} = \frac{\partial \sum_{i=0}^{\infty} \Delta Y_{t+i}}{\partial u_t^d} = \sum_{i=0}^{\infty} \frac{\partial \Delta Y_{t+i}}{\partial u_t^d} = \sum_{i=0}^{\infty} [\Theta_i B]_{1,2} = [\Theta(1)B]_{1,2}$$

Di conseguenza, per analizzare quale sia l'impatto degli shock sul livello del PIL, bisogna considerare la *cumulata* degli shock su ΔY_t. L'ipotesi teorica è che questa quantità vada a 0, cioè

$$\lim_{k\to\infty} \frac{\partial Y_{t+k}}{\partial u_t^d} = 0. \tag{4.8}$$

Il vincolo che ci serve, pertanto, si ottiene imponendo che la matrice $\Theta(1)B$ sia triangolare bassa. La matrice $\Theta(1)$ è facile da stimare una volta che siano stati stimati i parametri del VAR: infatti, poiché

$$\Theta(L) = A(L)^{-1}$$

una stima di $\Theta(1)$ è semplicemente

$$\widehat{\Theta(1)} = \hat{A}(1)^{-1}.$$

Naturalmente, perché questo funzioni è necessario che $A(1)$ sia invertibile. Questo esclude processi con radici unitarie.[17]

[17] Non è vero: esiste una interessantissima generalizzazione al caso cointegrato che, però, non affronto. Chi è interessato cerchi su Google "King Plosser Stock Watson". Non dico di più.

Per illustrare il senso del vincolo, viene utile dire che la matrice Σ è per definizione la matrice varianze-covarianze di ϵ_t, ma può anche essere vista come $V(y_t|\Im_{t-1})$, cioè la matrice varianze-covarianze della distribuzione condizionale di y_t. Come è fatta, invece, la matrice varianze-covarianze della distribuzione marginale di y_t, ovvero $V(y_t)$? Consideriamo, di nuovo, la rappresentazione in media mobile:

$$y_t = \Theta(L)\epsilon_t = \epsilon_t + \Theta_1\epsilon_{t-1} + \cdots$$

Visto che le ϵ_t sono incorrelate fra loro nel tempo (è un *white noise* vettoriale) si ha che

$$V(y_t) = V(\epsilon_t) + \Theta_1 V(\epsilon_{t-1})\Theta_1' + \cdots = \Sigma + \Theta_1\Sigma\Theta_1' + \cdots = \Theta(1)\Sigma\Theta(1)'$$

Visto che $\Sigma = BB'$, possiamo scrivere

$$\Theta(1)\Sigma\Theta(1)' = [\Theta(1)B][\Theta(1)B]'$$

Il vincolo corrispondente alla restrizione di lungo periodo è che la matrice $[\Theta(1)B]$ sia triangolare bassa. In pratica, una restrizione di lungo periodo può essere pensata come una restrizione che imponiamo sulla matrice varianze-covarianze marginale anziché su quella condizionale.

A questo punto, almeno in linea di principio, possono essere ritrovati gli elementi di B. Nel caso Blanchard-Quah si avrebbe che

$$\Theta(1)_{11}B_{12} + \Theta(1)_{12}B_{22} = 0 \Longrightarrow B_{12} = -\frac{\Theta(1)_{12}}{\Theta(1)_{11}}B_{22},$$

per cui, se chiamiamo $\theta = -\frac{\Theta(1)_{12}}{\Theta(1)_{11}}$, si ha

$$BB' = \begin{bmatrix} B_{11}^2 + \theta^2 B_{22}^2 & B_{11}B_{21} - \theta B_{22} \\ B_{11}B_{21} - \theta B_{22} & B_{21}^2 + B_{22}^2 \end{bmatrix} = \Sigma = \begin{bmatrix} \sigma_{\Delta_Y}^2 & \sigma_{\Delta_Y,U} \\ \sigma_{\Delta_Y,U} & \sigma_U^2 \end{bmatrix}$$

e per trovare la matrice B bisogna risolvere il seguente sistema di equazioni:

$$\begin{aligned} \sigma_{\Delta_Y}^2 &= B_{11}^2 + \theta^2 B_{22}^2 \\ \sigma_{\Delta_Y,U} &= B_{11}B_{21} - \theta B_{22} \\ \sigma_U^2 &= B_{21}^2 + B_{22}^2 \end{aligned}$$

Dalle considerazioni svolte in queste ultime pagine risulta evidente che ogni analisi dinamica produce risultati che dipendono in modo cruciale dalle condizioni di identificazione che sono state imposte. Non è impossibile (anzi, è tutt'altro che raro), che due studiosi possano pervenire ad interpretazioni completamente diverse delle proprietà dinamiche di uno stesso sistema semplicemente perché hanno scelto diverse ipotesi per l'identificazione (ad esempio, un diverso ordinamento delle variabili). Questa caratteristica è sovente considerata un punto di debolezza della metodologia dei VAR strutturali.

Io personalmente non condivido. Vorrei sottolineare che nell'analisi di un problema complesso non c'è un modo "giusto" e un modo "sbagliato" di procedere[18]. Di conseguenza, nell'analisi strutturale di un VAR possono essere

[18] "...there is always a well-known solution to every human problem — neat, plausible, and wrong". H. L. Mencken.

ragionevoli approcci diversi e, com'è ovvio, i risultati nei vari casi cambiano perché cambiano le definizioni degli shock le risposte ai quali vengono misurate. Se poi sia più utile, ragionevole, informativo, illuminante considerare $E(y_t|x_t)$ o $E(x_t|y_t)$ dipende in buona parte dal problema in esame e dai gusti personali.

Capitolo 5

Cointegrazione

5.1 Definizioni

In questo capitolo considereremo spesso *combinazioni lineari* di processi univariati. Detto in un altro modo, parleremo delle proprietà di processi che possono essere definiti come

$$z_t = \beta' y_t,$$

dove y_t è un processo stocastico multivariato (stazionario o meno) e β è una matrice di costanti.

Non l'abbiamo detto mai esplicitamente, ma dovrebbe essere intuitivo che la combinazione lineare di due o più processi stazionari è ancora un processo stazionario. Si può dimostrare, ma mi contento dell'intuizione. Inoltre, dovrebbe essere piuttosto chiaro (e qui faccio ancora appello all'intuizione) che una combinazione lineare, ad esempio una somma, fra un processo $I(1)$ e un processo $I(0)$ è un processo $I(1)$. Continuando di questo passo, si potrebbe pensare che una combinazione lineare di due processi $I(1)$ sia un processo $I(1)$. In realtà, questo non è sempre vero. Prendiamo per esempio questo caso ultrasemplificato:

$$\begin{cases} x_{1t} = x_{1t-1} + \epsilon_t \\ x_{2t} = x_{1t} + u_t \end{cases}$$

dove ϵ_t e u_t sono due processi $I(0)$ generici.

È del tutto evidente che x_{1t} è $I(1)$; è altrettanto evidente che anche x_{2t} è $I(1)$, poiché risulta dalla somma di un processo $I(1)$ e di un processo $I(0)$. Consideriamo però una particolare combinazione lineare di x_{1t} e x_{2t}, cioè la loro differenza: $z_t = x_{2t} - x_{1t}$. Dalle definizioni sopra date, è ovvio che $z_t = u_t$, che è $I(0)$ per ipotesi.

A questo punto, possiamo enunciare un principio più generale sulle combinazioni lineari fra processi: se x_t è $I(d)$ e y_t è $I(b)$, allora $z_t = x_t + \lambda y_t$ è $I(c)$, dove

$$\begin{cases} c = \max(d,b) & \text{per } d \neq b \\ c \leq \max(d,b) & \text{per } d = b \end{cases}$$

Quando $c < \max(d,b)$ (cioè quando la diseguaglianza vale in senso stretto) si ha **cointegrazione**. Noi (come peraltro il 99% della letteratura in merito) ci

interesseremo al caso in cui una combinazione lineare di processi $I(1)$ produce un processo $I(0)$. Supponiamo quindi di avere un processo $I(1)$ multivariato di dimensione n, che chiamiamo y_t; questo equivale a dire che ognuno degli n processi che compongono il vettore y_t è $I(1)$.

Noi diremo che c'è cointegrazione se c'è almeno un vettore β tale per cui la combinazione $z_t = \beta' y_t$ è $I(0)$; se β ha questa proprietà, allora prende il nome di **vettore di cointegrazione**. Come vedremo, di questi vettori possono essercene più di uno: se questi vettori vengono raccolti in una matrice, quest'ultima la chiamiamo **matrice di cointegrazione**; il numero di vettori di cointegrazione linearmente indipendenti (ossia il rango della matrice di cointegrazione) prende il nome di **rango di cointegrazione**. Nell'esempio precedente, $y_t' = [x_{1,t}, x_{2,t}]$ e $\beta' = [-1, 1]$; il rango di cointegrazione è 1, e quindi β è anche la matrice di cointegrazione. Un processo $I(1)$ multivariato per cui esista almeno un vettore di cointegrazione viene anche detto **sistema cointegrato**.

5.2 Proprietà dei vettori di cointegrazione

Già a questo stadio, sui vettori di cointegrazione si possono dire un paio di cose interessanti: in primo luogo, che per un vettore $I(1)$ y_t di dimensione $n \times 1$ il rango di cointegrazione può essere al massimo uguale a $n - 1$; se infatti esistessero n vettori di cointegrazione linearmente indipendenti, essi si potrebbero raccogliere in una matrice di cointegrazione non singolare B: in questo caso, avremmo che $B' y_t = z_t$; ma se B è invertibile, dovrebbe anche valere $y_t = (B')^{-1} z_t$, che è evidentemente assurda, visto che a sinistra del segno di uguale c'è una cosa che è $I(1)$ (per definizione), e a destra una cosa che è $I(0)$, perché è una combinazione lineare di processi $I(0)$.

La seconda cosa che si può dire è che, dati uno o più vettori di cointegrazione, ogni loro combinazione lineare è ancora un vettore di cointegrazione. Infatti, se β è una matrice di cointegrazione con r righe e n colonne,

$$\beta' y_t = z_t$$

e z_t è $I(0)$; ovviamente, possiamo scrivere qualunque combinazione lineare delle righe di β' come $b' = K\beta'$, dove K è una qualsiasi matrice con r colonne. Definiamo adesso un processo w_t come

$$w_t = b' y_t = K z_t.$$

poiché w_t è una combinazione lineare di processi $I(0)$, è anch'esso un processo $I(0)$, di modo che anche b' è un vettore (o una matrice) di cointegrazione. A volerla dire più difficile, la matrice di cointegrazione è definita solo a meno di una trasformazione lineare invertibile. Come si vedrà, la cosa ha una sua importanza quando si fanno le stime. Ma di questo parleremo più avanti.

Tutto questo è divertente e ci sarà anche utile nel prosieguo, ma ci parla delle proprietà dei vettori di cointegrazione senza dir nulla sulla loro interpretazione, che è di gran lunga più interessante. Consideriamo cosa succede a due variabili $I(1)$ cointegrate: essendo processi DS, avranno tutte le caratteristiche con cui vi ho intrattenuto nel paragrafo 3.2, come ad esempio l'assenza

di *mean-reversion* e così via. Esiste, però, una (e, in questo esempio, una sola) loro combinazione lineare che è stazionaria, cioè fluttua attorno ad un valor medio ben definito, che ha solo una memoria di breve periodo e non di lungo e non mostra nessuna tendenza a scapparsene via verso qualche asintoto.

Poiché come economisti siamo abituati a pensare in termini di equilibrio, forte è la tentazione di assegnare a questa particolare combinazione lineare uno status interpretativo privilegiato, cioè quello di **relazione di equilibrio**; le variabili possono andare dove vogliono, ma c'è fra loro una relazione che, magari approssimativamente, vale sempre. Se vogliamo pensare alle variabili $I(1)$ come a dei viandanti perenni, i più romantici possono pensare a due variabili cointegrate come a una **coppia** di viandanti perenni; la cointegrazione unisce per sempre i loro destini: vagheranno in eterno, ma sempre legati[1].

Esempio 5.2.1 (La teoria quantitativa della moneta) *Dovendo esprimere in forma semplificata la teoria quantitativa della moneta, si può dire che secondo tale teoria esiste una proporzione stabile fra la quantità di moneta presente in un sistema e il valore delle transazioni effettuate: questo rapporto si chiama velocità di circolazione. In formule:*

$$MV = PY$$

Oppure, si può dire che la velocità di circolazione è data dal rapporto fra PIL e quantità reale di moneta

$$V = \frac{Y}{M/P}$$

In logaritmi, si ha che

$$v = y - m$$

dove y è il logaritmo del PIL e m è il logaritmo della moneta reale. Un estremista della teoria quantitativa della moneta direbbe che v è una quantità fissa. Uno meno estremista potrebbe dire che v è una quantità che, osservata nel tempo, presenta fluttuazioni più o meno persistenti attorno ad un valore centrale: la serie storica di v_t somiglia pertanto alla realizzazione di un processo $I(0)$. Se y_t e m_t sono rappresentabili come processi $I(1)$, sostenere la teoria quantitativa della moneta equivale, più o meno, ad affermare che le serie y_t e m_t cointegrano, e che il vettore di cointegrazione che le lega è $(1, -1)$.

Esempio 5.2.2 (La parità dei poteri d'acquisto) *Consideriamo le economie di due paesi, il paese A e il paese B. Chiamiamo P_t^A e P_t^B il livello dei prezzi nei due paesi al tempo t. Se P_t^A sale, mentre P_t^B rimane fermo, evidentemente le merci prodotte nel paese B diventano più competitive di quelle prodotte nel paese A. Questo (la faccio breve, se volete tutta la storia la studiate in Economia Internazionale) produce una catena di conseguenze che portano a fare sì che la moneta del paese A si svaluti progressivamente rispetto a quella del paese B, e quindi che il tasso di cambio E_t tenda anch'esso a salire nel tempo.*

Se vige fra queste due economie la parità dei poteri d'acquisto (spesso detta più brevemente PPP, dall'inglese Purchasing Power Parity*), il tasso di cambio fra le*

[1] Un articolo, peraltro meritorio, apparso sull'*American Statistician* suggerisce la similitudine meno sentimentale e più bukowskiana di un'ubriaca e il suo cane.

monete dei paesi A e B si muoverà nel tempo in modo tale da compensare le differenze nei tassi d'inflazione nei due paesi. Potremmo quindi ipotizzare che la quantità

$$\frac{P_t^B E_t}{P_t^A}$$

tenda a rimanere più o meno costante nel tempo. Come nell'esempio precedente, sostenere questa ipotesi è equivalente a sostenere che, se le serie sono $I(1)$*, la relazione*

$$p_t^B + e_t - p_t^A$$

(dove di nuovo le lettere minuscole indicano i logaritmi) è una relazione di cointegrazione.

Da questi esempi dovrebbe essere chiaro che la cointegrazione è un concetto statistico che ci permette di formulare in modo empirico l'idea di una relazione di equilibrio di lungo periodo fra due o più variabili: se esiste una relazione di cointegrazione, siamo autorizzati a pensare che tale relazione non varrà mai esattamente, ma le deviazioni da essa saranno comunque temporanee e limitate.

I parametri che definiscono tale relazione possono essere definiti *a priori* dalla teoria economica (come negli esempi appena fatti), ma possono darsi dei casi in cui la teoria non indica valori numerici precisi per i parametri che compaiono nella relazione d'equilibrio: in questi casi, tutt'al più la teoria indica il segno. Da qui la grande attenzione che ha ricevuto, negli ultimi vent'anni del secolo scorso, il problema di stimare le relazioni di cointegrazione anziché postularle. C'è da aspettarsi, dopo quel che abbiamo visto a proposito dei test di radice unitaria e della regressione spuria, che la stima dei vettori di cointegrazione sia una faccenda tutt'altro che banale. Per capire come muoversi, ci sarà utile affrontare, prima dei problemi inferenziali veri e propri, il tema della rappresentazione dei sistemi cointegrati.

5.3 Modelli a correzione d'errore

Partiamo dal caso più semplice: quello in cui il sistema ammette una rappresentazione VAR di ordine 1. Consideriamo quindi un processo y_t di dimensione n, del tipo

$$y_t = Ay_{t-1} + \epsilon_t \tag{5.1}$$

sottraendo y_{t-1} da ambo i lati si arriva a

$$\Delta y_t = \Pi y_{t-1} + \epsilon_t \tag{5.2}$$

dove $\Pi = A - I$. Si noti la somiglianza con l'equazione (3.6), che ci era servita ad illustrare il test DF. In ambito univariato, ovviamente Π può essere solo 0 o diverso da 0. In ambito multivariato, si apre una casistica intermedia, che è precisamente quella che ci interessa, legata al rango della matrice Π. Se chiamiamo questo rango r, abbiamo tre possibili casi:

$r = 0$: in questo caso, $\Pi = 0$, per cui y_t è un *random walk* multivariato, e non esiste cointegrazione;
$r = n$: y_t non è $I(1)$, ma $I(0)$, poiché Π è invertibile (vedi par. 5.2);
$0 < r < n$: y_t è un sistema cointegrato.

L'analisi di questo ultimo caso occuperà il resto del capitolo: come vedremo, r è il rango di cointegrazione; inoltre, Π può essere scritta come $\alpha\beta'$, dove α e β sono due matrici $n \times r$; in particolare, β è la matrice di cointegrazione. Messe tutte assieme, queste proprietà implicano che la (5.2) può essere riscritta come

$$\Delta y_t = \alpha z_{t-1} + \epsilon_t \tag{5.3}$$

Dove $z_t = \beta' y_t$ è un vettore $r \times 1$. Poiché β è la matrice di cointegrazione, z_t è $I(0)$.

Qual è il significato delle variabili z_t? Esse rappresentano la serie storica delle deviazioni dalle relazioni di cointegrazione. In questa ottica, la (5.3) dice una cosa quasi banale: il movimento di un sistema cointegrato è determinato da due fattori. Uno (ϵ_t) è casuale, e nel caso specifico del nostro esempio è un *white noise*; l'altro (αz_{t-1}) è determinato dall'ampiezza — al periodo precedente — della deviazione dalla relazione di cointegrazione, ossia da quella che dal punto di vista interpretativo potremmo chiamare l'entità del disequilibrio al tempo $t-1$. La matrice α si chiama **matrice dei pesi** (viene talvolta chiamata matrice dei *loadings* da quelli che amano sfoggiare il loro inglese), perché il suo elemento *ij* ci dice qual è l'effetto sulla *i*-esima variabile del *j*-esimo elemento di z_{t-1}.

Esempio 5.3.1 *Facciamo di nuovo l'esempio della teoria quantitativa della moneta. Supponiamo che, in un dato istante di tempo t, ci sia più moneta nel sistema di quanto previsto dalla relazione di equilibrio. In altri termini, la velocità di circolazione è al di sotto del suo valore medio. Ebbene, la variazione fra t e $t+1$ del PIL e della moneta reale (in logaritmi) sarà data dalla somma di due componenti, che possiamo riassumere così:*

$$\begin{cases} \Delta y_{t+1} = \alpha_1(y_t - m_t) + \epsilon_{1t} \\ \Delta m_{t+1} = \alpha_2(y_t - m_t) + \epsilon_{2t} \end{cases}$$

Ricordo che la grandezza $(m_t - y_t)$ è interpretabile come il logaritmo della velocità di circolazione; un valore più basso del normale della velocità di circolazione provoca quindi un aggiustamento, al periodo successivo, tanto del reddito reale che della quantità reale di moneta. Se, tanto per dire, α_2 fosse positivo, vorrebbe dire che in presenza di troppa moneta, Δm_t dev'essere negativo, ossia m_t deve scendere. Questo può accadere, ad esempio, per l'aumento dei prezzi. Secondo i monetaristi, questo è il meccanismo che spiega l'inflazione.

Questo meccanismo ha un nome ben preciso, ed è **meccanismo a correzione d'errore**, o più in breve **ECM** (dall'inglese *Error Correction Mechanism*); un VAR riscritto in forma ECM spesso viene anche chiamato VECM (ECM vettoriale). I modelli ECM occupano giustamente un posto di primo piano nell'econometria moderna, proprio perché rappresentano lo snodo che collega analisi delle serie storiche e teoria economica, breve periodo e lungo periodo.

Se un sistema cointegrato ha rango di cointegrazione r, si può dire che esistono r relazioni di equilibrio di lungo periodo, e quindi r processi stazionari che descrivono l'andamento nel tempo degli squilibri da tali relazioni. Ove questi squilibri (che abbiamo chiamato z_t) siano diversi da 0, si avrà un movimento nel vettore y_{t+1} tale per cui lo squilibrio tende a venire riassorbito.

Figura 5.1: VAR(1) stazionario: serie storiche simulate

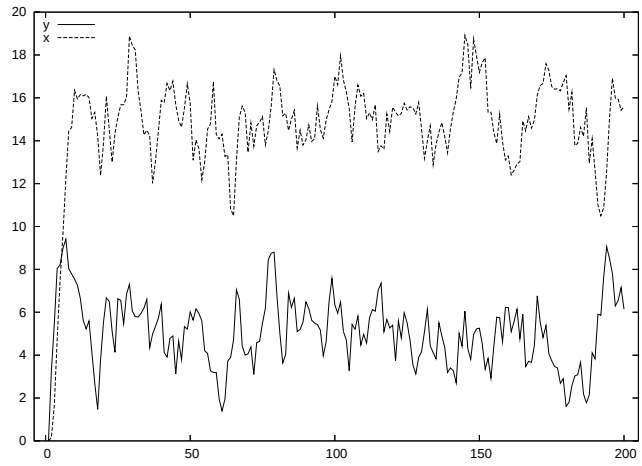
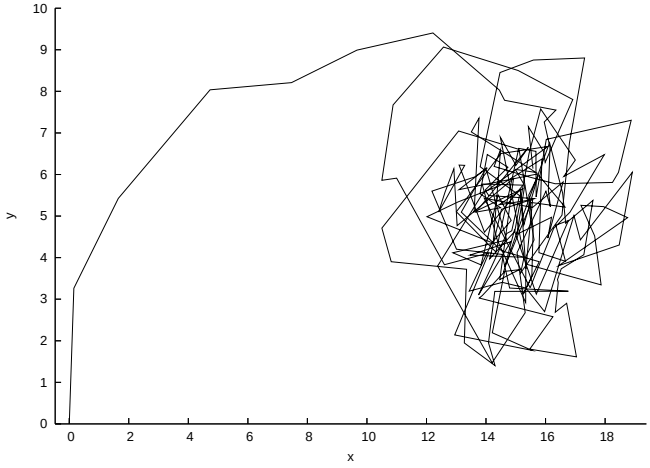


Figura 5.2: VAR(1) stazionario: serie storiche simulate – diagramma XY



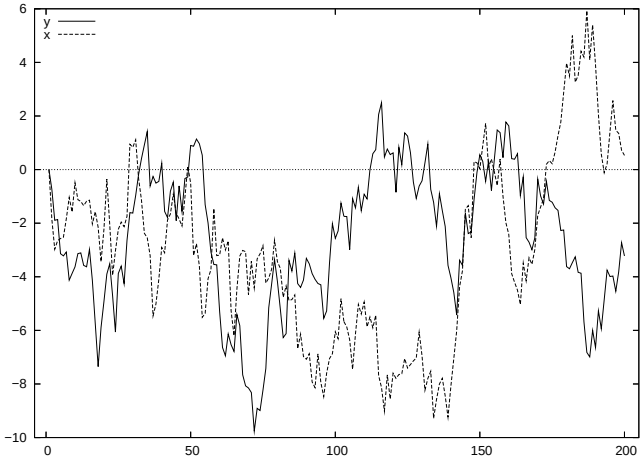
Un concetto che spesso si rivela utile per la comprensione di questa idea è quello di **attrattore**. Questo concetto è molto generale, ma si può rendere l'idea dicendo che un attrattore è un insieme di punti attorno ai quali un sistema dinamico tende a trovarsi. Facciamo un esempio che è meglio: tanto

per cominciare, consideriamo un VAR(1) stazionario così formulato:

$$\begin{aligned} y_t &= 4 + 0.8y_{t-1} - 0.2x_{t-1} + \epsilon_{1t} \\ x_t &= 2 + 0.2y_{t-1} + 0.8x_{t-1} + \epsilon_{2t} \end{aligned}$$

Facendo un minimo di conti, si vedrà che il valore atteso non condizionale del vettore $[y_t, x_t]$ è $[5, 15]$. Simuliamo il processo facendolo partire dal punto $[0, 0]$ e otteniamo un grafico come quello mostrato in figura 5.1; come si vede, le serie tendono ad oscillare intorno alla media, che funziona da attrattore, tant'è che partendo da un punto "lontano" la prima cosa che fanno è riportarsi nei paraggi dell'attrattore, che è appunto la coppia di valori $[5, 15]$. La cosa si vede anche meglio nella figura 5.2, in cui i valori delle serie sono rappresentati su un piano, cioè come un punto per ogni osservazione. Qui si vede benissimo il cammino che il sistema compie nel tempo: parte dal punto $[0, 0]$, si porta rapidamente nei pressi del punto attrattore, dopodiché ronza là attorno.

Figura 5.3: *Random walk*: serie storiche simulate



Se invece facciamo la stessa cosa con un *random walk* bivariato, succedono cose molto diverse: in questo caso, non esiste un punto a cui le serie tendono: l'attrattore è l'intero piano. Infatti, facendo partire due ubriachi dal bar $[0, 0]$, questi vagano per la città senza puntare da nessuna parte (vedi figure 5.3 e 5.4). Faccio notare che la figura 5.4 permette anche di farsi un'idea intuitiva del perché la regressione fra y_t e x_t è, in questo caso, spuria: la retta che passa il più possibile in mezzo ai punti avrà, con ogni probabilità, una pendenza diversa da 0 senza che questo autorizzi a dedurne alcun legame sistematico fra le due variabili.

Esaminiamo ora un processo cointegrato. In questo caso, il processo generatore dei dati è

$$\begin{aligned} y_t &= 0.5y_{t-1} + 0.5x_{t-1} + \epsilon_{1t} \\ x_t &= -0.2y_{t-1} + 1.2x_{t-1} + \epsilon_{2t}, \end{aligned}$$

Figura 5.4: *Random walk*: serie storiche simulate – diagramma XY

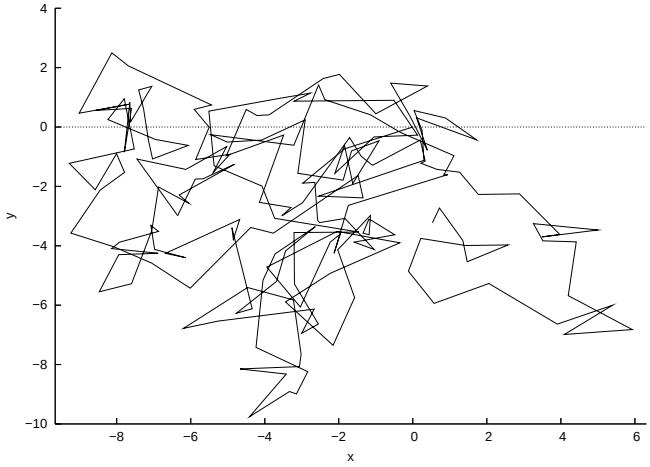


Figura 5.5: Processo cointegrato: serie storiche simulate

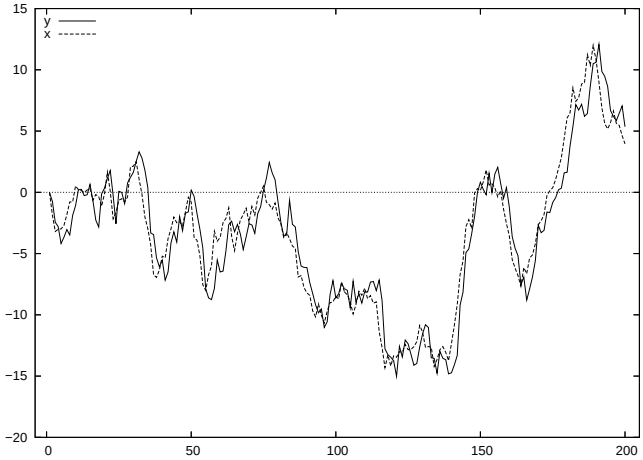
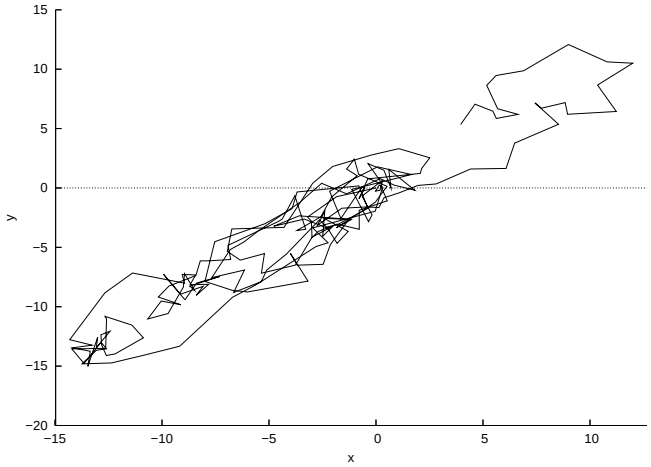


Figura 5.6: Processo cointegrato: serie storiche simulate – diagramma XY



che si può scrivere in forma VECM come

$$\begin{aligned} \Delta y_t &= -0.5 z_{t-1} + \epsilon_{1t} \\ \Delta x_t &= -0.2 z_{t-1} + \epsilon_{2t}, \end{aligned}$$

dove $\beta' = [1, -1]$, e quindi $z_t = y_t - x_t$.

Il grafico 5.5 mostra il tipico andamento di due serie cointegrate con vettore di cointegrazione $[1, -1]$ (se vi piace, padrone-cane). Ancora più interessante, però è quello che si vede nel diagramma XY: in questo caso, infatti, l'attrattore è costituito da tutti i punti per cui $z_t = 0$, e cioè la retta a 45°. Notate che le due serie hanno il classico comportamento da processo $I(1)$ se considerate separatamente; tuttavia, il sistema tende a non allontanarsi mai dalla retta. In altre parole, quando si dovesse verificare uno squilibrio ($z_t \neq 0$), il sistema torna per suo conto verso l'attrattore. A differenza del caso stazionario, però, l'attrattore non è un singolo punto, ma un insieme infinito di punti, per cui nel lungo periodo non è dato sapere esattamente dove il sistema si troverà: tutto quel che sappiamo è che si troverà vicino a qualche punto dell'attrattore.

Più in generale, si può mostrare che in un sistema cointegrato con n variabili e rango di cointegrazione r, l'attrattore è un iperpiano a $n - r$ dimensioni.

Una generalizzazione a costo zero che possiamo fare sin d'ora è quella per cui possiamo considerare un ECM anche un'equazione della forma (5.3) dove il termine ϵ_t sia sostituito da un generico processo $I(0)$. L'interpretazione rimane più o meno la stessa. Va comunque ricordato che, nella letteratura applicata, gli ECM che si incontrano più spesso sono quelli in cui la persistenza nel termine di disturbo viene eliminata attraverso un autoregressivo, pervenendo ad un modello del tipo

$$\Delta y_t = \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t \tag{5.4}$$

che, se z_t fosse una serie osservabile, potrebbe essere tranquillamente stimato con gli OLS, poiché non ci sono serie non stazionarie nell'equazione (5.4). Il problema spesso è che z_t non è osservabile perché non conosciamo β. Ma non anticipiamo. Una cosa che va detta ora è però che un'equazione come la (5.4) può essere fatta derivare da una rappresentazione VAR di y_t usando la scomposizione BN. Infatti, supponiamo che $A(L)y_t = \epsilon_t$. Possiamo anche scrivere

$$y_t = B(L)y_{t-1} + \epsilon_t$$

dove $B(z) = \frac{I-A(z)}{z}$. Applicando la scomposizione BN a $B(L)$ possiamo scrivere

$$y_t = [B(1) + \Gamma(L)\Delta]\, y_{t-1} + \epsilon_t$$

Sottraendo da ambo i lati y_{t-1}, si arriva a

$$\Delta y_t = [B(1) - I]\, y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t.$$

Dalla definizione è immediato controllare che $B(1) = I - A(1)$, cosicché

$$\Delta y_t = -A(1)y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t \tag{5.5}$$

e si noti, di nuovo, la somiglianza di questa equazione con la (3.9). La presenza di cointegrazione è legata al rango di $A(1)$, che svolge lo stesso ruolo che svolgeva Π nell'equazione (5.2).

5.4 Il teorema di rappresentazione di Granger

Come abbiamo visto, la presenza di cointegrazione in un processo stocastico multivariato apre la possibilità di leggere alcuni dei parametri del processo (la matrice di cointegrazione) in modo estremamente fruttuoso per quanto riguarda l'interpretazione economica. In particolare, questa interpretazione diviene immediata quando si possa scrivere il processo sotto forma VECM. Nel paragrafo precedente, abbiamo supposto che si potesse. In questo paragrafo, preciseremo meglio quale sia la natura del legame fra processi cointegrati e modelli ECM: questo legame rappresenta l'oggetto del **teorema di rappresentazione di Granger**[2]. Come vedremo, l'analisi di questo teorema ha come sottoprodotto una notevole quantità di cose che possiamo dire sui sistemi cointegrati.

Per esplorare le proprietà di un sistema cointegrato y_t vengono alla mente almeno due possibilità: una legata all'ipotesi che il sistema possa essere scritto come un VAR di un qualche ordine (anche infinito)

$$A(L)y_t = \epsilon_t \tag{5.6}$$

[2]In una versione ormai antica di questa dispensa, questa nota diceva: "Sì, è lo stesso Granger della causalità e della regressione spuria. Perché a quest'uomo non abbiano ancora dato il premio Nobel per me è un mistero.". Bè, alla fine gliel'hanno dato, nel 2003, ma sospetto di avere influenzato gli accademici di Svezia in modo molto marginale.

con l'associata rappresentazione ECM, e una legata al fatto che, se y_t è $I(1)$, allora Δy_t è $I(0)$, e quindi deve avere una rappresentazione di Wold del tipo

$$\Delta y_t = C(L)\epsilon_t. \tag{5.7}$$

Che relazioni intercorrono fra queste due rappresentazioni? La risposta è appunto contenuta nel teorema di rappresentazione di Granger. A dir la verità, questo benedetto teorema è una specie di idra a nove teste, perché non ha un enunciato ben preciso, ma si riferisce in generale a tutto quel che si può dire sulla rappresentazione di un sistema cointegrato. Questo rende la sua digestione un po' ardua, ma ripaga ampiamente perché mostra come vari aspetti di un sistema cointegrato possano essere considerati da più di un punto di vista. Al proposito, mi piace citare il Poeta:

> Notate che ogni proposizione, ogni teorema, ogni oggetto di speculazione, ogni cosa ha non solo due ma infinite facce, sotto ciascuna delle quali si può considerare, contemplare, dimostrare e credere con ragione e verità.
> G. Leopardi, "Zibaldone", 2527-8

Siccome non ho l'ambizione di far meglio quel che è stato già fatto egregiamente da teste migliori della mia, non offrirò una vera e propria dimostrazione di questo teorema, ma mi limiterò soltanto ad esplorarne i punti salienti. Ma prima, una piccola premessa.

5.4.1 Un po' di algebra matriciale

Nel seguito, userò un operatore matriciale che non tutti conoscono, per cui si usa il segno "$\perp$" e che si legge "ortogonale". Quindi, $X_\perp$ si legge "X ortogonale".

Se X è una qualunque matrice $l \times m$, con $l > m$ e rango m, $X_\perp$ è una matrice $l \times (l - m)$ le cui colonne sono linearmente indipendenti e ortogonali a quelle di X.[3] In altri termini, $X'_\perp X = 0$ per definizione[4] e non ci sono vettori ortogonali a X che non siano combinazioni lineari delle colonne di $X_\perp$. L'operatore "$\perp$" a noi servirà soprattutto per uno scopo: se la matrice A ha la proprietà

$$x'A = 0,$$

allora ne consegue che ogni colonna di A deve essere ortogonale a x. Ma siccome ogni vettore ortogonale a x è una combinazione lineare di $x_\perp$, allora A deve poter essere scritta come $A = x_\perp B$, dove B è "qualcos'altro" che varierà a seconda del contesto.

Esempio 5.4.1 *Tanto per farsi un'idea, il lettore è invitato a controllare le proprietà appena enunciate dell'operatore $\perp$ sulle matrici*

$$w = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

[3] Per i patiti dell'algebra: le colonne di $X_\perp$ definiscono uno spazio vettoriale noto come **spazio nullo** di X. Di conseguenza, $X_\perp$ non è unica. Ma non importa.

[4] Attenzione: $X'_\perp X = 0$ **non** implica $XX'_\perp = 0$.

e

$$w'_\perp = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}.$$

Considerate ora la matrice

$$A = \begin{bmatrix} 2 & -1 \\ -2 & 1 \\ -2 & 1 \end{bmatrix};$$

si può controllare facilmente che $w'A = 0$ (basta moltiplicare). Tanto ci basta per dire che deve esistere una matrice B tale per cui $A = w_\perp B$. Per la cronaca, tale matrice è in questo caso $B = [2 \quad -1]$.

5.4.2 Il teorema vero e proprio

Teorema 2 (Teorema di rappresentazione di Granger) *Per ogni sistema cointegrato esiste una rappresentazione ECM; se esiste un rappresentazione ECM e le serie sono integrate, allora sono cointegrate.*

La dimostrazione è piuttosto complessa, ma nell'appendice a questo capitolo ne dò una traccia per capire i punti fondamentali. Il succo del teorema è questo: un sistema cointegrato può essere espresso in due modi equivalenti, corrispondenti alle rappresentazioni autoregressiva e a media mobile. La rappresentazione derivante da quella autoregressiva è quella ECM, già vista nella (5.4), che riporto qui per completezza:

$$\Delta y_t = \alpha\beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \epsilon_t$$

La rappresentazione derivante da quella a media mobile è, se vogliamo, un adattamento al caso cointegrato della scomposizione di Beveridge e Nelson, che possiamo scrivere così

$$y_t = \left[\beta_\perp H \alpha'_\perp\right] \mu_t + C^*(L)\epsilon_t, \tag{5.8}$$

dove μ_t è definito dalla proprietà $\Delta\mu_t = \epsilon_t$ e H è una matrice invertibile i cui dettagli sono, per il momento, inessenziali[5]. Naturalmente, visto che ϵ_t è un *white noise* vettoriale, μ_t è un *random walk* vettoriale per definizione. Notate che la matrice fra parentesi quadre nella (5.8) è semplicemente $C(1)$; essa ha, in questo caso, la caratteristica di annullarsi se premoltiplicata per β', per le proprietà dell'operatore $\perp$. In altri termini, $C(1)$ è singolare, con rango $(n-r)$.

La rappresentazione (5.8) viene di solito detta rappresentazione di Stock e Watson, o rappresentazione **a trend comuni**, ed ora vediamo il perché. Consideriamo la serie μ_t: come ho già detto, questa serie è un *random walk* a n dimensioni, e quindi un processo $I(1)$ non cointegrato. Costruiamo ora una serie η_t nel seguente modo:

$$\eta_t = \alpha'_\perp \mu_t$$

[5]Chi fosse interessato, guardi l'appendice a questo capitolo.

Evidentemente, η_t è un *random walk* a $n-r$ dimensioni, perché la matrice $\alpha_\perp$ è una matrice $n \times (n-r)$. È quindi possibile riscrivere la (5.8) come segue:

$$y_t = F\eta_t + u_t$$

dove $F = \beta_\perp H$ è una matrice $n \times (n-r)$ e $u_t = C^*(L)\epsilon_t$ è per ipotesi un processo stazionario. Scrivendo y_t in questo modo si vede chiaramente che è possibile pensare un sistema cointegrato come un sistema in cui esistono un certo numero $(n-r)$ di trend stocastici inosservabili, i quali si palesano sulle serie osservabili attraverso la matrice F; le serie che noi osserviamo contengono una parte $I(1)$ (data da combinazioni lineari di questi trend stocastici) e una parte $I(0)$ (data da u_t) . La cointegrazione esiste appunto perché $n-r < n$, e quindi la combinazione lineare di $\beta' y_t$ è un processo stazionario semplicemente in forza della relazione $\beta' F = 0$. In pratica, moltiplicando β' per y_t neutralizziamo l'effetto dei trend comuni.

Esempio 5.4.2 *Mettiamo il caso di avere due serie cointegrate x_t e y_t, con $\beta = (1,-1)'$; di conseguenza, si avrà che $z_t = x_t - y_t$ è $I(0)$. In questo caso H è uno scalare, e $F = \beta_\perp H$ è proporzionale al vettore $(1,1)'$.*

Le due serie, quindi, possono essere scritte come la somma di un processo $I(1)$ che è lo stesso per tutte e due, più una parte stazionaria. In pratica, le due serie fluttuano attorno allo stesso trend stocastico, e non si allontanano mai proprio perché ambedue tendono a stare nei paraggi del trend comune. È chiaro che la cointegrazione nasce proprio dal fatto che prendendo la differenza fra le due serie il trend stocastico scompare, e rimangono solo le oscillazioni attorno ad esso.

Si noti che questo caso ultrasemplificato si adatta benissimo ad essere contestualizzato se consideriamo l'esempio, fatto qualche pagina fa, della teoria quantitativa della moneta. In questo caso, sia lo stock di moneta che il reddito viaggiano attorno ad un trend comune, che con un po' di fantasia si potrebbe anche pensare come il sentiero di crescita di lungo periodo dell'intera economia.

5.4.3 Nucleo deterministico

La situazione diviene più articolata se si considera il caso in cui il processo Δy_t possa avere media non nulla. Tipicamente si considerano processi del tipo

$$\Delta y_t = d_t + \alpha\beta' y_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t$$

cioè un processo ECM vettoriale a cui viene aggiunta una parte deterministica, che di solito contiene un'intercetta ed un trend lineare, ossia $d_t = \delta_0 + \delta_1 t$.

In analogia con quanto visto nel paragrafo 3.2, si potrebbe pensare che, se d_t è un polinomio di ordine p nel tempo (cioè una cosa del tipo $\delta_0 + \delta_1 t + \cdots + \delta_p t^p$), allora nella serie in livelli y_t sarà presente, in generale, un polinomio di ordine $p+1$. Nel caso della cointegrazione, tuttavia, ciò non è necessariamente vero. Vediamo perché considerando nuovamente, per semplicità, il caso in cui la serie in livelli sia rappresentabile come un VAR(1). Se

$$y_t = d_t + Ay_{t-1} + \epsilon_t$$

possiamo riscrivere il modello in forma ECM accorpando la parte deterministica al termine di disturbo come segue:

$$\Delta y_t = \alpha\beta' y_{t-1} + (\epsilon_t + d_t)$$

Se definiamo $u_t = \epsilon_t + d_t$, possiamo rifare tutto il ragionamento di prima usando u_t in luogo di ϵ_t. In questo caso avremmo che

$$y_t = \left[\beta_\perp H\alpha'_\perp\right]\tilde{\mu}_t + C^*(L)u_t$$

dove $\Delta\tilde{\mu}_t = u_t = \epsilon_t + d_t$. Il processo $\tilde{\mu}_t$ risulta cioè dalla somma di un *random walk* multivariato (e fin qui nulla di nuovo) più una parte deterministica la cui differenza prima è un polinomio nel tempo di ordine p, ossia un polinomio di ordine $p+1$.

Consideriamo il caso particolare in cui $d_t = \delta_0$; il polinomio in t presente in $\tilde{\mu}_t$ sarà una cosa del tipo $\delta_0 t + k$; in pratica, un *random walk* con drift multivariato. Poiché in un sistema cointegrato la matrice $\alpha'_\perp$ ha rango $(n-r)$, può benissimo darsi che $\alpha'_\perp\delta_0 = 0$. In altri termini, se δ_0 può essere espresso come combinazione lineare delle colonne della matrice α, la componente di ordine 1 del polinomio contenuto in $\tilde{\mu}_t$ si annulla, cosicché il trend lineare presente in $\tilde{\mu}_t$ non compare in y_t. Peraltro, se $\delta_0 = \alpha d$, allora la rappresentazione VECM può essere scritta come segue:

$$\Delta y_t = \alpha(d + \beta' y_{t-1}) + \epsilon_t$$

e quindi la relazione di equilibrio di lungo periodo viene a contenere un'intercetta, data dal vettore d.

Più in generale, se ho un polinomio del tipo

$$d_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \ldots + \delta_p t^p$$

in cui i vari δ_i sono vettori a n elementi, non è detto che il polinomio $\alpha'_\perp d_t$, abbia esso stesso grado p: bisogna vedere se $\alpha'_\perp\delta_p$ è zero oppure no. Come ormai sappiamo, se $\alpha'_\perp\delta_p = 0$, vuol dire che δ_p è una qualche combinazione lineare delle colonne di α.

In pratica, si hanno cinque casi, che di solito vengono ordinati dal più al meno restrittivo:

$d_t = 0$	In questo caso, la parte deterministica non c'è affatto. I trend comuni sono *random walk* senza drift e le z_t hanno media 0. I dati non presentano traccia di trend deterministici e fluttuano attorno allo 0.
$d_t = \delta_0; \alpha'_\perp\delta_0 = 0$	Qui accade una cosa più complessa: la rappresentazione ECM vettoriale ha un'intercetta, che però non dà origine ad un trend lineare nella rappresentazione a trend comuni, perché questi ultimi non hanno drift. I dati non hanno trend deterministici, ma fluttuano attorno ad un valore diverso da 0. Gli squilibri z_t hanno media diversa da 0, così che si può parlare di un'intercetta nella relazione di cointegrazione.

$d_t = \delta_0; \alpha'_\perp \delta_0 \neq 0$	In questa situazione, l'intercetta del VECM non rispetta alcun vincolo particolare, e quindi in linea teorica abbiamo un'intercetta sia nella relazione di cointegrazione (che non genera un trend nelle serie osservate) sia fuori (da cui un drift nella rappresentazione a trend comuni, e quindi un trend nelle serie osservate). In pratica, però, queste due intercette vengono di solito sommate, e la relazione di cointegrazione viene presentata senza intercetta.
$d_t = \delta_0 + \delta_1 t; \alpha'_\perp \delta_1 = 0$	Il caso parallelo al caso 2: la relazione di cointegrazione ha un trend lineare, che però non si traduce in un trend quadratico nei livelli.
$d_t = \delta_0 + \delta_1 t; \alpha'_\perp \delta_1 \neq 0$	Qui non ci sono restrizioni di sorta, e le serie esibiscono un trend quadratico.

Il caso 2 è quello che, di solito, rimane più misterioso: cercherò di far capire come funziona con un esempio.

Esempio 5.4.3 *Supponiamo di avere un processo x_t, che è un* random walk *con drift:*

$$x_t = m + x_{t-1} + \varepsilon_t; \tag{5.9}$$

come sappiamo, x_t è un processo $I(1)$ che fluttua attorno ad un trend deterministico; quest'ultimo ha pendenza m. Se $m = 0$, il trend deterministico scompare e abbiamo un random walk *puro.*

Considerate ora un secondo processo:

$$y_t = k + x_t + u_t, \tag{5.10}$$

dove u_t è un white noise. *Visto che u_t è stazionario per definizione, x_t e y_t cointegrano, perché la loro differenza $z_t = y_t - x_t$ è stazionaria, e cioè un white noise con media k.*

Sostituendo la (5.9) nella (5.10), rappresentiamo il sistema come un VAR(1)

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix}, \tag{5.11}$$

che si può trasformare in forma VECM sottraendo da tutti e due i lati il vettore $[y_{t-1}, x_{t-1}]'$:

$$\begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} = \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix}, \tag{5.12}$$

dove la matrice $\begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$ è quella che avevamo chiamato Π nell'equazione (5.2). Notate che la matrice Π è singolare (ha rango 1), ciò che consente di scriverla come $\alpha\beta'$:

$$\begin{aligned} \begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} &= \begin{bmatrix} k+m \\ m \end{bmatrix} + \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix} = \\ &= \mu_0 + \alpha\beta' \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \eta_t = \mu_0 + \alpha z_{t-1} + \eta_t. \end{aligned}$$

Considerate ora le tre possibilità:

1. *$m \neq 0$: Come abbiamo detto, x_t ha un trend. Ne consegue (dalla (5.10)) che ce l'ha pure y_t, perché in media si mantiene a una distanza k da x_t. Il vettore μ_0 non è soggetto a restrizioni e siamo nel caso 3.*

2. *$m = 0$ e $k \neq 0$: Qui, x_t non ha trend deterministico, e per conseguenza neanche y_t. Tuttavia, la distanza media fra y_t e x_t è diversa da 0. Il vettore μ_0 è*

$$\mu_0 = \begin{bmatrix} k \\ 0 \end{bmatrix},$$

che, si noti, non è nullo e quindi il VECM in (5.12) ha un'intercetta. Questa è soggetta al vincolo che il suo secondo elemento sia zero. Più in generale, μ_0 è un multiplo del vettore α. Possiamo pertanto riscrivere il VECM come segue:

$$\begin{bmatrix} \Delta y_t \\ \Delta x_t \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & -k \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \\ 1 \end{bmatrix} + \begin{bmatrix} u_t + \varepsilon_t \\ \varepsilon_t \end{bmatrix},$$

e cioè incorporando l'intercetta nel vettore di cointegrazione. Questo è il caso 2 (quello con "costante vincolata").

3. *$m = 0$ e $k = 0$: Questo è il caso 1, il più restrittivo. x_t e y_t non hanno trend, e la distanza fra loro è in media 0. Il VECM non ha intercetta.*

Spesso, la scelta fra queste diverse possibilità si basa sia sull'osservazione delle serie che su un qualche tipo di ragionamento a priori. Se le due serie hanno un chiaro trend lineare, allora imporre restrizioni sull'intercetta è inappropriato. In caso contrario, ci si può chiedere se abbia senso includere un intercetta nella relazione di cointegrazione. Il caso di scuola nasce quando esaminiamo due tassi di interesse: normalmente un trend questi non ce l'hanno[6]*, ma il VAR può benissimo avere un'intercetta, perché la loro differenza (lo "spread") può avere media non nulla, (per esempio, per via di un premio al rischio o di liquidità).*

Questo può sembrare un arzigogolo algebrico privo di alcuna rilevanza pratica; in realtà, questa caratteristica dei sistemi cointegrati diventa a volte cruciale in fase di stima, come vedremo nel prossimo paragrafo.

5.5 Tecniche di stima

Quando decidiamo di stimare i parametri di un sistema cointegrato, possiamo trovarci — grosso modo — in tre situazioni:

[6]Giustamente (ricordate il discorso sui babilonesi che ho fatto a proposito dei test di radice unitaria al sottoparagrafo 3.4.5?). Però attenzione, in certi contesti ha perfettamente senso descrivere il movimento nei tassi come qualcosa che si sovrappone a un trend lineare. I tassi di interesse italiani sono scesi regolarmente negli ultimi 15 anni del XX secolo, per lo più per motivi macro (discesa dell'inflazione e così via). Se non vogliamo modellare il quadro di riferimento macroeconomico, ci mettiamo un bel trend e via. Ma con campioni più lunghi il problema normalmente non si pone.

1. La matrice di cointegrazione è nota, e di conseguenza anche il suo rango: questo può essere, ad esempio, il caso in cui si parta da un'ipotesi di lavoro che assume *a priori* come valida la teoria delle parità dei poteri di acquisto.

2. Il rango di cointegrazione è noto, ma la matrice di cointegrazione no. Può darsi questo caso, ad esempio, se riteniamo che inflazione e disoccupazione formino un sistema cointegrato, in una specie di curva di Phillips; noi supponiamo che la curva di Phillips esista, ma non ne conosciamo la pendenza.

3. Non sono noti a priori né rango né matrice di cointegrazione.

5.5.1 La procedura di Johansen

Nel terzo caso, che è anche il più generale (se anche si fanno delle ipotesi può essere scientificamente interessante andare a vedere se tengono empiricamente), normalmente si fa ricorso ad un metodo di stima escogitato da Søren Johansen, noto appunto come **metodo di Johansen**, che è piuttosto complesso[7], per cui ne darò una descrizione sintetica ai limiti della reticenza; mai come ora, rinvio alla letteratura.

Mi limito a dire che si tratta di una procedura in cui si suppone che il sistema cointegrato possa essere rappresentato come un VAR di ordine finito con errori gaussiani. Il punto di partenza è quello di riparametrizzare il sistema sotto forma ECM

$$\Delta y_t = d_t + \Pi y_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t; \tag{5.13}$$

implicitamente, supponiamo che l'ordine del VAR p sia noto. Ma trovare l'ordine del VAR più adeguato a rappresentare i dati, per fortuna, non pone problemi particolari, perché si possono utilizzare metodi standard (e cioè i soliti test di ipotesi e/o i criteri di informazione).

A questo punto, si può impostare il problema della stima in un contesto di massima verosimiglianza. Normalmente, lo stimatore di massima verosimiglianza dei parametri di un modello di regressione lineare con errori gaussiani non è altro che l'OLS. In questo caso, però, bisogna anche tenere conto dei vincoli legati alla presenza di cointegrazione: infatti, il rango della matrice Π è uguale al rango di cointegrazione r, per cui vogliamo che il metodo di stima garantisca una stima di Π con rango ridotto (cosa che l'OLS non fa).

Quantificazione del rango di cointegrazione

Il primo problema, pertanto, è quantificare il rango di cointegrazione r. Per fare questo, la procedura di Johansen prevede un test sul rango della matrice Π nell'equazione 5.13. Anzi, due. Ambedue sono legati al fatto che in una

[7] Dal punto di vista pratico, stimare un sistema cointegrato con il metodo di Johansen è invece piuttosto semplice, visto che il metodo di Johansen è disponibile — precotto — in quasi tutti i pacchetti econometrici più diffusi, che anzi spesso lo propongono come unica scelta.

matrice semidefinita positiva il numero di autovalori positivi è uguale al suo rango, e gli altri sono zero.

I test, pertanto, funzionano così: per prima cosa, viene definita una matrice M (i cui dettagli ometto) che per costruzione è semidefinita positiva ed ha lo stesso rango di Π. Il vantaggio di lavorare con M anziché Π è che la semidefinitezza[8] di M assicura che tutti i suoi autovalori siano reali non negativi.

Di questa matrice è disponibile una stima consistente $\hat{M}$, con la conseguenza che gli autovalori di $\hat{M}$ (chiamiamoli $\hat{\lambda}$) sono a loro volta stimatori consistenti degli n autovalori di M. A questo punto, li si ordina dal più grande $\hat{\lambda}_1$ al più piccolo $\hat{\lambda}_n$ e si imposta un test di azzeramento *del più piccolo*.

Se la nulla viene rifiutata, stop: se λ_n è positivo, allora sono positivi tutti, per cui Π ha rango pieno e il sistema è stazionario. Altrimenti, si passa a considerare λ_{n-1}. Qui sono possibili due strade:

1. si può fare un test la cui nulla è $\lambda_{n-1} = 0$, dando per scontato che $r < n$ e testando così l'ipotesi $r < n - 1$, e questo è il cosiddetto test λ-max;

2. in alternativa, si può usare un test la cui nulla è $\lambda_n = \lambda_{n-1} = 0$, ossia un test congiunto di azzeramento degli ultimi due autovalori che non dà per scontato che $r < n$; questo è il cosiddetto test traccia.

Se l'ipotesi viene accettata, si prosegue con $\hat{\lambda}_{n-2}$, e così via. Al primo rifiuto, ci si ferma, ottenendo così una stima di r. Se poi non si rifiuta mai la nulla, vuol dire che $r = 0$ e non c'è cointegrazione.

Confrontando l'equazione (5.13) con quella che forma la base del test Dickey-Fuller,

$$\Delta y_t = d_t + \rho y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \epsilon_t;$$

si vede subito che il test di Johansen è, di fatto, l'equivalente multivariato del test ADF. L'unica differenza è che il coefficiente ρ è uno scalare e quindi o è 0 o è invertibile; nel caso di sistemi cointegrati, invece, la casistica intermedia, in cui Π è singolare senza essere nulla, è quella che più ci interessa. È peraltro vero che, se il test dovesse accettare l'ipotesi secondo cui la matrice Π ha rango pieno, allora dovremmo concludere non solo che non c'è cointegrazione, ma anche che il sistema non è nemmeno $I(1)$, bensì stazionario. In questo senso va letta l'affermazione, che ho fatto qualche pagina fa (e cioè nel sottoparagrafo 3.4.4) secondo cui il test di Johansen può sostituire i test univariati tipo Dickey-Fuller.

Non sorprenderà, perciò, sapere che i test di Johansen si distribuiscono asintoticamente in modo non standard. Oltretutto, così come accade nel test ADF, queste distribuzioni limite non sono invarianti alla parte deterministica del VAR, la quale, a sua volta, rischia di essere piuttosto ingarbugliata, alla luce delle considerazioni fatte al paragrafo precedente. Bisogna decidere, infatti, quale sia il grado più appropriato del polinomio in t da inserire nella parte deterministica: di solito, la scelta è fra una costante o costante più

[8] Ma si dice in italiano "semidefinitezza"? Mah, facciamo di sì.

trend. Tuttavia, come abbiamo visto al paragrafo precedente, non è detto che alcuni dei vettori di parametri della parte deterministica non scompaiano una volta moltiplicati per $\alpha_\perp$: questo conduce alle cinque possibilità diverse di cui ho parlato poc'anzi, per ognuna delle quali c'è un set di valori critici da consultare.

Stima dei vettori di cointegrazione

Una volta stimato il rango di cointegrazione (o, per meglio dire, trovato il rango di cointegrazione che offre una descrizione statisticamente sostenibile dei dati), si può stimare β.

Un problema preliminare che si pone a questo punto è però che — come facevo notare nel paragrafo 5.2 — la matrice β non è identificata. Infatti, se β è una matrice di cointegrazione, lo è anche $b = \beta K$, dove K è una qualunque matrice $(r \times r)$ non singolare. Di conseguenza, esiste un numero infinito di matrici $n \times r$ che sono equivalenti dal punto di vista osservazionale. Questo argomento può anche essere visto in un altro modo: supponiamo di conoscere la matrice Π. Ora, questa matrice può essere rappresentata sia come

$$\Pi = \alpha\beta'$$

che come

$$\Pi = \alpha K^{-1} K\beta' = ab'$$

e non è che la rappresentazione di Π basata su α e β sia più 'vera' di quanto non lo sia quella basata su a e b: semplicemente, sono equivalenti. Siamo nel classico problema di sottoidentificazione.

Se non siete completamente digiuni di algebra lineare, sappiate che il problema della sottoidentificazione può essere anche visto dal punto di vista geometrico: infatti, si può mostrare che i vettori di cointegrazione formano una base per l'iperspazio di dimensione r, ortogonale a quello dell'attrattore. Come è noto, la scelta di una base è arbitraria.

Come si esce dalla sottoidentificazione? Imponendo vincoli. Si può dimostrare che il numero minimo di vincoli che rende i parametri identificati è pari a r^2. Nella procedura originariamente proposta da Johansen l'identificazione è ottenuta imponendo tali vincoli su una particolare forma quadratica in β, su cui non mi dilungo. Un'approccio alternativo all'indentificazione, che dà luogo alla cosiddetta **rappresentazione triangolare**, è stato proposto da P. C. B. Phillips, che consiste nell'assumere che le prime r righe di β siano una matrice identità. Quindi, si può scrivere

$$\tilde{\beta} = \begin{bmatrix} I \\ -\tilde{\beta}_2 \end{bmatrix} \tag{5.14}$$

dove $\tilde{\beta}_2$ è, invece, libero da vincoli[9].

[9] Nessuno impedisce di imporre ulteriori vincoli su $\tilde{\beta}_2$, naturalmente. In questo caso, però, la meccanica della stima è un tantino più complessa.

Per fortuna, non servono algoritmi iterativi, ma basta calcolarsi gli autovettori della matrice $\hat{M}$ di cui ho parlato a proposito dei test traccia e λ-max. Dal punto di vista delle proprietà asintotiche, lo stimatore di β ha varie caratteristiche inusuali: la più notevole è quella chiamata **superconsistenza**.

Piccolo ripasso di teoria asintotica: di solito, quando uno stimatore è consistente, si ha che la sua distribuzione collassa a una degenere per $T \to \infty$, ciò che è detto in modo succinto dall'espressione $\hat{\theta} \xrightarrow{\text{p}} \theta_0$. Quindi, $\hat{\theta} - \theta_0 \xrightarrow{\text{p}} 0$. Se però moltiplichiamo questa quantità per $\sqrt{T}$, otteniamo una cosa che non collassa e non diverge, ma si stabilizza su una distribuzione limite. Nei casi che si incontrano di solito nel mondo della stazionarietà, $\sqrt{T}(\hat{\theta} - \theta_0)$ diventa, al crescere di T, un variabile casuale sempre più simile a una normale a media 0. Nel caso in esame qui, invece, per ottenere una distribuzione limite, la differenza $(\hat{\beta} - \beta)$ deve essere moltiplicata per T anziché $\sqrt{T}$. In pratica, la velocità con cui $\hat{\beta}$ collassa verso β è molto più grande; detto altrimenti, la dispersione dello stimatore è proporzionale a T^{-1} anziché a $T^{-1/2}$. Questo non vuol dire che in campioni finiti questo stimatore sia particolarmente preciso: è stato notato che la sua distorsione può essere piuttosto seria, e con le ampiezze campionarie che si hanno di solito, il fatto che converga più velocemente è una magra consolazione.

Inserita la stima di β nell'ECM, tutto il resto (cioè i parametri che controllano la dinamica di breve periodo) si stima con gli OLS. Tutto il coacervo di stimatori che viene fuori è consistente (al tasso $\sqrt{T}$), asintoticamente efficiente, e permette in molti casi di fare test di ipotesi che si distribuiscono come delle tranquillizzanti χ^2.

Tenete presente che di solito i pacchetti fanno più o meno tutto da soli, nel senso che una volta decisi

1. l'ordine del VAR
2. il nucleo deterministico
3. il rango di cointegrazione

le stime vengono fuori tutte insieme. Per chi ci vuole provare, però, farlo "a mano" è piuttosto istruttivo.

Vincoli sullo spazio di cointegrazione

Una possibilità intrigante è quella di stimare la matrice di cointegrazione imponendo un numero di vincoli superiore a quello strettamente necessario per ottenere l'identificazione. Naturalmente, questo conduce ad avere una stima vincolata ed una non vincolata, con l'ovvia conseguenza che le due possono essere confrontate sulla base di un'apposita statistica test, la quale a sua volta ha spesso interpretazioni interessanti.

Facciamo un esempio: supponiamo di analizzare per mezzo di un VECM un sistema di tre tassi di interesse. In questo caso, è ragionevole pensare che le relazioni di lungo periodo abbiano la forma seguente:

$$r_t^1 = \pi_1 + r_t^3 + z_t^1 \tag{5.15}$$
$$r_t^2 = \pi_2 + r_t^3 + z_t^2. \tag{5.16}$$

Che va letta: fra il tasso 1 e il tasso 3 c'è una differenza sistematica π_1, che dipende dalle diverse caratteristiche di liquidità e/o rischosità dei due titoli. A questa relazione si sovrappone un disturbo stazionario z_t^1; lo stesso discorso vale per il tasso 2. In notazione matriciale, lo stesso si può dire in modo equivalente come

$$z_t = \beta' y_t = \begin{bmatrix} 1 & 0 & -\beta_1 & -\pi_1 \\ 0 & 1 & -\beta_2 & -\pi_2 \end{bmatrix} \begin{bmatrix} r_t^1 \\ r_t^2 \\ r_t^3 \\ 1 \end{bmatrix} \sim I(0) \tag{5.17}$$

dove la nostra ipotesi teorica è che $\beta_1 = \beta_2 = 1$.

Questo implica, evidentemente:

1. che il rango di cointegrazione r sia pari a 2;
2. che il VECM abbia la forma

   $$\Delta y_t = \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t;$$

   per le ragioni esposte al sottoparagrafo precedente: ossia, il VECM ha un'intercetta, ma che questa è del tipo $\mu = \alpha\pi$, cosicché l'intercetta compare nella relazione di lungo periodo e basta.
3. che la matrice di cointegrazione β possa essere scritta come nella (5.17). Si noti che gli unici due elementi incogniti di β sono π_1 e π_2.

Immaginate che il calcolo degli autovettori ritorni la seguente stima:

$$\beta^* = \begin{bmatrix} -0.26643 & -0.49177 \\ 0.62064 & 0.29382 \\ -0.33026 & 0.21660 \\ -0.97485 & -0.095877 \end{bmatrix}$$

Come ho cercato di argomentare prima, questa è una stima consistente di due vettori che formano lo spazio di cointegrazione, ma una qualunque combinazione lineare delle colonne di β^* va bene lo stesso. Se passiamo alla rappresentazione triangolare otteniamo

$$\hat{\beta} = \beta^* (\beta^*_{[1:2,1:2]})^{-1} = \begin{bmatrix} -0.26643 & -0.49177 \\ 0.62064 & 0.29382 \\ -0.33026 & 0.21660 \\ -0.97485 & -0.095877 \end{bmatrix} \begin{bmatrix} -0.26643 & -0.49177 \\ 0.62064 & 0.29382 \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} -0.26643 & -0.49177 \\ 0.62064 & 0.29382 \\ -0.33026 & 0.21660 \\ -0.97485 & -0.095877 \end{bmatrix} \begin{bmatrix} 1.0453 & 2.8907 \\ -0.61682 & 0.27548 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1.02 & -0.97 \\ -1 & -2 \end{bmatrix}$$

Le stime dei parametri nelle nostre relazioni di equilibrio (5.17) sono quindi

$$\hat{\beta}_1 = 1.02 \qquad \hat{\pi}_1 = 1$$
$$\hat{\beta}_2 = 0.97 \qquad \hat{\pi}_2 = 2$$

Ci contentiamo? Forse no: mentre sui due parametri π_1 e π_2 non c'era nessuna ipotesi teorica, ci sarebbe piaciuto che $\hat{\beta}_1$ e $\hat{\beta}_2$ venissero uguali a 1. Ma, come sa chi fa le stime, in questo caso la cosa appropriata da fare è domandarsi: "è sostenibile l'ipotesi secondo cui $\beta_1 = \beta_2 = 1$?". In altri termini, si potrebbe, o dovrebbe, fare un test.

In linea di principio, il test si può fare in vari modi. Nella pratica, però, usano quasi tutti un test di tipo LR, che comporta la stima del modello sotto vincolo e il confronto delle due log-verosimiglianze. A questo punto, ho una buona notizia e una cattiva. Quella buona è che il test si distribuisce, sotto la nulla, coma una χ^2 e una volta tanto non c'è bisogno di tavole esotiche. La cattiva notizia è che la stima del modello vincolato è spesso notevolmente più difficile dal punto di vista numerico: c'è bisogno di un algoritmo iterativo la cui convergenza a volte richiede pianto e stridor di denti.

Anche con la descrizione da Bignami che ne ho fatto qui, il lettore si renderà conto che la procedura di Johansen si presta molto bene ad essere automatizzata. Per questo, e per altri motivi, è diventata ormai lo standard di riferimento.

Come putroppo succede in questi casi, è comparso nella professione un certo grado di adesione acritica, per cui chi fa le cose così "fa bene", chi no "sbaglia" e si fanno le gare a chi è più ortodosso. Questo sicuramente avviene in Europa, negli Stati Uniti forse un po' meno.

Non so a voi, ma a me il pensiero unico mette tristezza.

5.5.2 Procedure alternative

Se la matrice β fosse nota, problemi inferenziali non ce ne sarebbero. Poiché tutte le serie che compaiono in un modello del tipo

$$\Delta y_t = d_t + \alpha z_{t-1} + \sum_{i=1}^{p} \Gamma_i \Delta y_{t-i} + \epsilon_t \tag{5.18}$$

sono stazionarie, si possono usare gli OLS senza alcun problema ed ottenere delle stime che godono di tutte le belle proprietà che sappiamo.

Il problema è che spesso non si conosce β, e quindi la serie z_t non è osservabile. Possiamo però pensare di ottenere una stima della matrice di cointegrazione, e di usare quella come se fosse il vero valore di β. Questo è esattamente ciò che facciamo nella procedura di Johansen, in cui β viene stimato col metodo della massima verosimiglianza.

Esistono, però, modi alternativi di stimare β, che possono venire comodo di tanto in tanto. Dato che ci stiamo muovendo nel campo della non stazionarietà, si potrebbe congetturare che questo sia un affare complicato. Una volta tanto, però, le bizzarrie inferenziali dei processi $I(1)$ ci vengono in soccorso. Infatti, è possibile dimostrare che, la matrice $\tilde{\beta}_2$ può essere stimata usando quella stessa regressione statica di cui tanto abbiamo sparlato quando trattavamo la regressione spuria.

Consideriamo la (5.14). Naturalmente $\tilde{\beta}$ è una matrice di cointegrazione, e quindi $\tilde{\beta}' y_t$ è stazionario. Questo implica che possiamo scrivere

$$\tilde{\beta}' y_t = y_{1t} - \tilde{\beta}_2' y_{2t} = z_t$$

e quindi

$$y_{1t} = \tilde{\beta}_2' y_{2t} + z_t$$

dove y_{1t} sono i primi r elementi di y_t, y_{2t} sono i restanti $n - r$ elementi e z_t è un qualche processo $I(0)$. Si può dimostrare che la regressione di y_{1t} su y_{2t} produce uno stimatore di $\tilde{\beta}_2$ consistente, anzi superconsistente (vedi sopra).

Fra l'altro, la sua consistenza non è messa a repentaglio — come accade nel caso stazionario — da un eventuale effetto 'da equazioni simultanee'. C'è infine da dire che la distribuzione di questo stimatore non è asintoticamente normale, cosicché non è possibile fare test di ipotesi sugli elementi della matrice di cointegrazione stimata in questo modo; questo, tuttavia, non è un problema nella misura in cui quella che interessa è una stima puntuale.

Il lettore sarà forse colto, a questo punto, da un certo qual turbamento: non avevamo forse detto, a proposito della regressione spuria, che l'OLS su variabili integrate produce spazzatura? Perché adesso funziona?
La risposta precisa richiede un notevole dispiego di teoria asintotica. Lo stile informale di questa dispensa, però, mi consente di produrre un argomento euristico che, credo, convincerà i più: considerate il caso in cui y_t e x_t siano cointegrate, e quindi $z_t = y_t - \beta x_t$ è $I(0)$. Considerate ora la funzione

$$S(\theta) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \theta x_t)^2$$

Valutando questa funzione in $\theta = \beta$ ottenete una statistica che converge in probabilità ad un numero finito (la varianza di z_t). Viceversa, scegliendo un qualunque altro $\theta \neq \beta$, la differenza $(y_t - \theta x_t)$ diventa $I(1)$, con la conseguenza che il momento secondo non esiste e la statistica diverge. Poiché il mestiere dell'OLS è quello di scegliere θ così che $S(\theta)$ sia la più piccola possibile, è chiaro che l'OLS tenderà a scegliere un valore di θ simile a β.
Questo ragionamento può anche essere esemplificato graficamente: confrontate le figure 5.4 (pag. 114) e 5.6 (pag. 115). Visto che la statistica OLS ha per sua natura la tendenza a "passare in mezzo ai punti", si capisce anche ad occhio che nel primo caso succedono cose poco chiare, mentre nel secondo caso è perfettamente ragionevole che l'OLS tenda a riprodurre quello che ora sappiamo essere l'attrattore.

Queste considerazioni suggeriscono che intanto si può stimare β con una regressione in livelli, e poi modellare il breve periodo sostituendo alle z_{t-1} nella (5.18) i residui (ritardati di un periodo) della regressione in livelli. Si può dimostrare che l'uso di $\hat{\beta}$ anziché del vero vettore di cointegrazione β non fa differenza (asintoticamente parlando). Se poi per sicurezza si vuol controllare che le $\hat{z}_t$ siano effettivamente $I(0)$, è possibile utilizzare un test ADF. Attenzione, però, che in questo caso le tavole dei valori critici sono lievemente diverse da quelle per il test ADF vero e proprio; per fortuna, sono state anche queste calcolate per mezzo di simulazioni.

Questo approccio, noto come approccio di **Engle-Granger**, è estremamente semplice, e non richiede apparati di calcolo che non siano quelli della regressione OLS. I problemi che sorgono sono sostanzialmente due:

1. gli stimatori dei parametri β non sono efficienti;

2. non è sempre possibile sottoporre i coefficienti così stimati a test di ipotesi standard, in quanto possono venir fuori distribuzioni diverse dalla χ^2; in particolare, segnalo che effettuare test di Granger-causalità in sistemi cointegrati è una faccenda non banale, su cui la letteratura si è accapigliata per qualche anno prima di arrivare a delle conclusioni definitive.

Problemi di questo tipo possono venir superati facendo ricorso alla procedura di Johansen o a procedure alternative; queste, in sostanza, introducono dei correttivi al primo stadio della procedura di Engle-Granger. In pratica, si usano delle regressioni modificate per stimare il vettore (o i vettori) di cointegrazione in modo che lo stimatore sia efficiente e si possano fare test di ipotesi sugli elementi di β usando distribuzioni standard.

Le trovate più ingegnose che si usano comunemente sono il cosiddetto stimatore **Fully Modified OLS** (o FM-OLS per brevità) di Phillips e Hansen e il **Dynamic OLS** (o DOLS), inventato indipendentemente da Saikkonen e dalla premiata ditta Stock & Watson. Per i soliti motivi di concisione, però, non li illustro qui e me la cavo col solito richiamo alla letteratura.

Appendice: Traccia di dimostrazione del teorema di rappresentazione di Granger

Consideriamo un processo stocastico $I(1)$ a n dimensioni y_t. Poiché Δy_t è stazionario, deve avere una rappresentazione di Wold

$$\Delta y_t = C(L)\epsilon_t \tag{5.19}$$

Applichiamo la scomposizione di Beveridge-Nelson a y_t e otteniamo

$$y_t = C(1)\mu_t + C^*(L)\epsilon_t \tag{5.20}$$

Questa scomposizione esiste per qualunque processo $I(1)$ multivariato; se però il sistema è cointegrato, con rango di cointegrazione r, deve esistere una matrice β di dimensione $n \times r$ tale per cui $\beta' y_t$ è stazionario. Premoltiplicando la (5.20) per β' si vede immediatamente che l'unica condizione per cui questo è possibile è che

$$\beta' C(1) = 0.$$

Poiché $C(1)$ è una matrice $n \times n$, questo può succedere solo se $C(1)$ è singolare; in particolare, il suo rango non deve essere superiore a $n - r$. Se è così, $C(1)$ può sempre essere scritta come

$$C(1) = \beta_\perp \lambda' \tag{5.21}$$

Nel nostro caso, la condizione $\beta' C(1) = 0$ è assicurata solo se $C(1)$ "comincia con" $\beta_\perp$. La matrice λ sarà un'altra cosa sulla cui natura ancora non sappiamo nulla, se non che è una matrice $n \times (n - r)$, ma ci arriveremo presto.

Supponiamo ora che esista una rappresentazione VAR di y_t, del tipo $A(L)y_t = \epsilon_t$; il nostro obiettivo è quello ora di provare che $A(1)$ può essere scritta come $\alpha\beta'$, e quindi vale la rappresentazione ECM. Per mostrare questo, si considerino le equazioni (5.6) e (5.7); è chiaro che, se ambedue le rappresentazioni VAR e VMA sono valide, deve valere

$$A(z)C(z) = C(z)A(z) = I \cdot (1 - z) \tag{5.22}$$

Ponendo $z = 1$ nella (5.22), è facile dedurre che

$$A(1)C(1) = C(1)A(1) = 0 \tag{5.23}$$

Mettendo insieme la (5.21) e la (5.23), la prima cosa che possiamo dire su $A(1)$ è che, se $C(1)$ "comincia con" $\beta_\perp$, allora $A(1)$ deve "finire con" β'; per mezzo dello stesso ragionamento, poi, dovrebbe essere altrettanto chiaro che $A(1)$ deve "cominciare con" $\lambda_\perp$. Riassumendo, possiamo dire che $A(1)$ deve essere della forma

$$A(1) = \lambda_\perp K \beta'$$

dove K è una qualche matrice $r \times r$ invertibile. Il gioco è fatto. Battezziamo $\alpha = \lambda_\perp K$ ed ecco che $A(1)$ può essere scritta come $\alpha\beta'$, come volevasi dimostrare.

L'ultima cosa che è rimasta in sospeso è questa matrice λ di cui non sappiamo nulla. Possiamo provare a vedere in che relazione sta con α. Sappiamo che $\alpha = \lambda_\perp K$, e quindi $\lambda'\alpha = 0$; ne consegue che $C(1)$ deve avere la forma

$$C(1) = \beta_\perp H \alpha'_\perp \tag{5.24}$$

dove H è una matrice $(n-r) \times (n-r)$, i cui elementi sono funzioni dei parametri della rappresentazione autoregressiva.

È possibile dimostrare che $H = \left[\alpha'_\perp A^*(1)\beta_\perp\right]^{-1}$, cosicché

$$C(1) = \beta_\perp \left[\alpha'_\perp A^*(1)\beta_\perp\right]^{-1} \alpha'_\perp$$

Fatto. Anche se non è evidente a prima vista, questi passaggi hanno prodotto un risultato molto importante: il fatto di poter ricondurre la matrice $C(1)$ ai parametri dell'ECM, basandosi sul solo fatto che $C(1)$ deve essere singolare ci assicura che la rappresentazione ECM di un sistema cointegrato esiste *sempre*.

La matrice H che compare nella (5.24) non ha un'interpretazione immediatissima, ma, per gli amanti del genere, vediamo come è fatta: applichiamo la scomposizione BN ai polinomi che compaiono nella (5.22)

$$\left[A(1) + A^*(z)(1-z)\right]\left[C(1) + C^*(z)(1-z)\right] = I \cdot (1-z)$$

Poiché in un sistema cointegrato si ha $A(1)C(1) = 0$, espandendo l'espressione precedente si ha

$$\left[A(1)C^*(z) + A^*(z)C(1)\right](1-z) + A^*(z)C^*(z)(1-z)^2 = I(1-z)$$

'Semplificando' $(1-z)$ si ottiene

$$A(1)C^*(z) + A^*(z)C(1) + A^*(z)C^*(z)(1-z) = I$$

che, valutata in $z = 1$, dà

$$\alpha\beta' C^*(1) + A^*(1)\beta_\perp H \alpha'_\perp = I$$

La matrice all'espressione precedente può essere premoltiplicata per $\alpha'_\perp$ e postmoltiplicata per $A^*(1)\beta_\perp$, ottenendo

$$\alpha'_\perp A^*(1)\beta_\perp H \alpha'_\perp A^*(1)\beta_\perp = \alpha'_\perp A^*(1)\beta_\perp$$

Se la matrice a destra del segno di uguale è invertibile, si ha evidentemente $H = \left[\alpha'_\perp A^*(1)\beta_\perp\right]^{-1}$. Accenno solo al fatto che questa inversa esiste se il sistema è effettivamente $I(1)$. In processi con ordine di integrazione superiore, la matrice non è invertibile e il teorema si può generalizzare, ma tutto diventa un po' esoterico.

Capitolo 6

Processi a volatilità persistente

Fino ad ora, ci siamo occupati della persistenza nelle serie storiche da un punto di vista un po' riduttivo. Infatti, abbiamo definito un processo come persistente se

$$f(x_t) \neq f(x_t|\Im_{t-1}), \tag{6.1}$$

ciò che implica la non indipendenza del singolo elemento del processo x_t dalla sua storia. Nei capitoli precedenti, però, non abbiamo davvero studiato la persistenza dei processi secondo la definizione appena data, ma piuttosto un caso particolare, e cioè quello in cui

$$E(x_t) \neq E(x_t|\Im_{t-1}), \tag{6.2}$$

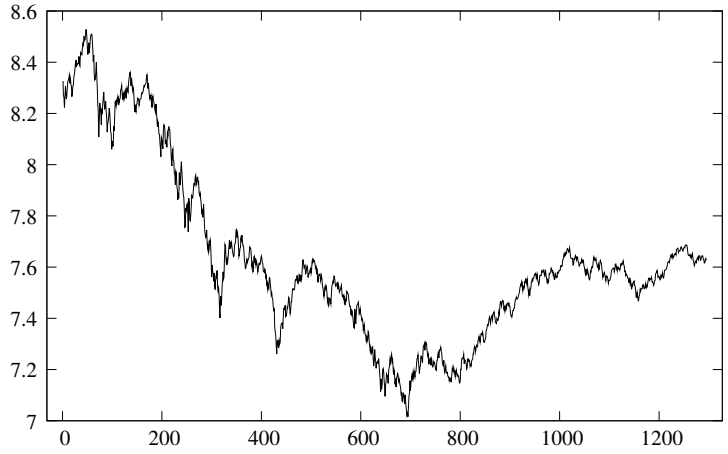
dando per implicito che, una volta modellata adeguatamente la persistenza nei momenti primi, tutti gli aspetti di persistenza interessanti fossero trattati in modo soddisfacente. In un modello ARMA, la media condizionale di y_t dipende da $\Im_{t-1}$ nei modi che abbiamo visto nei capitoli precedenti, ma tutte le altre caratteristiche distributive rimangono indipendenti dalla storia del processo: ad esempio, se confrontiamo la varianza marginale e quella condizionale di un processo AR(1) abbiamo che sono, sì, diverse, ($\frac{\sigma^2}{1-\alpha^2} \neq \sigma^2$) ma la loro diversità non dipende in alcun modo da $\Im_{t-1}$.

Così come è chiaro che la (6.2) implica la (6.1), è però altrettanto vero che la (6.1) non implica la (6.2). È possibile, quindi, che esistano degli aspetti di persistenza meritevoli di attenzione che non riguardano i momenti primi del processo, ma piuttosto altre sue caratteristiche distribuzionali.

In questo capitolo parleremo di processi chiamati **GARCH**, in cui la persistenza si avverte attraverso i momenti secondi, e abbiamo ottimi motivi per farlo: questi processi, infatti, hanno delle proprietà che replicano piuttosto bene alcune caratteristiche comunemente riscontrate in serie storiche finanziarie. Nel prossimo paragrafo, diremo quali sono.

6.1 I fatti stilizzati

Figura 6.1: Indice Nasdaq – logaritmi



Nella figura 6.1 mostro il logaritmo naturale dell'indice Nasdaq rilevato giornalmente dal 3 gennaio 2000 al 28 febbraio 2005. Come è evidente già dal grafico, e confermato da qualunque test di radice unitaria, non c'è verso di poter considerare questa serie una realizzazione di un processo $I(0)$. Questo, d'altronde, era un risultato almeno in parte prevedibile, visto che in un mercato efficiente il rendimento di una attività finanziaria deve essere una differenza di martingala, o per lo meno ci deve somigliare[1].

Differenziamo la serie, e otteniamo i rendimenti, mostrati in figura 6.2 (che è una copia conforme di quella che sta a pag. 10). I rendimenti, come c'era da aspettarsi, non evidenziano correlazioni significative (vedi pag. 10). Notate, però, che l'ampiezza delle oscillazioni varia nel tempo, e che periodi, anche lunghi, di bassa volatilità si alternano a periodi di alta volatilità. Questi "grappoli" di volatilità si chiamano appunto *volatility clusters*. Questo semplice fatto suggerisce che, anche se non c'è persistenza nella media, potrebbe essercene nella varianza, o più in generale nella volatilità.

Per chi si occupa di finanza, questa è una caratteristica molto importante: infatti, la volatilità di un mercato è strettamente connessa al suo livello di rischio, cosicché la possibilità di prevedere la volatilità di un mercato è un dato essenziale in qualunque attività di *asset allocation*.

[1] In effetti, ci sono ottimi motivi teorici per cui il prezzo di una attività finanziaria non debba essere per forza una differenza di martingala. A parte l'ovvia constatazione che non necessariamente i mercati sono davvero efficienti, c'è anche da dire che il risultato deriva da una serie di assunzioni opinabili, fra cui l'assenza di asimmetrie informative, il fatto che esista un "agente rappresentativo" che è un po' la media di tutti gli altri, che questo agente rappresentativo sia neutrale al rischio, anziché avverso e così via. Non lo dico io che sono un sovversivo, lo dice il principe degli economisti conservatori, cioè il premio Nobel Robert Lucas. Ma qui stiamo cercando il pelo nell'uovo, perché poi, di fatto, una regola meccanica che preveda i rendimenti non l'ha trovata ancora nessuno (per quanto, se qualcuno l'avesse trovata non lo verrebbe a dire a me e a voi).

Figura 6.2: Indice Nasdaq – rendimenti giornalieri

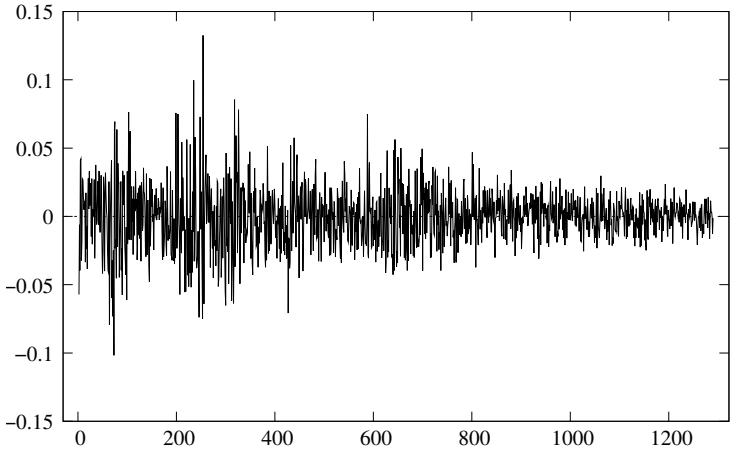
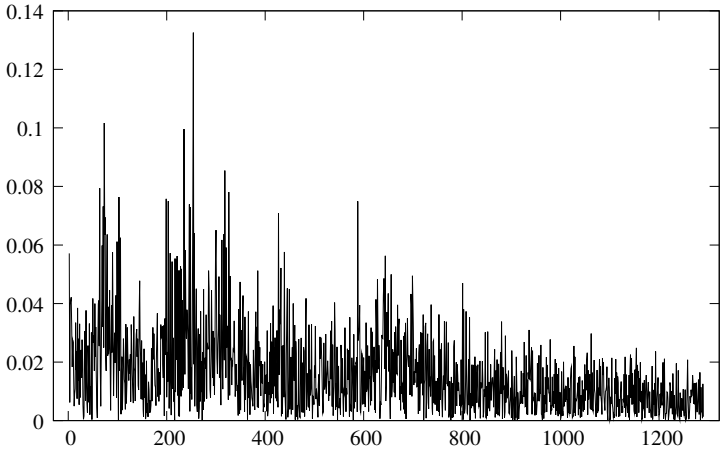


Figura 6.3: Rendimenti Nasdaq – valori assoluti

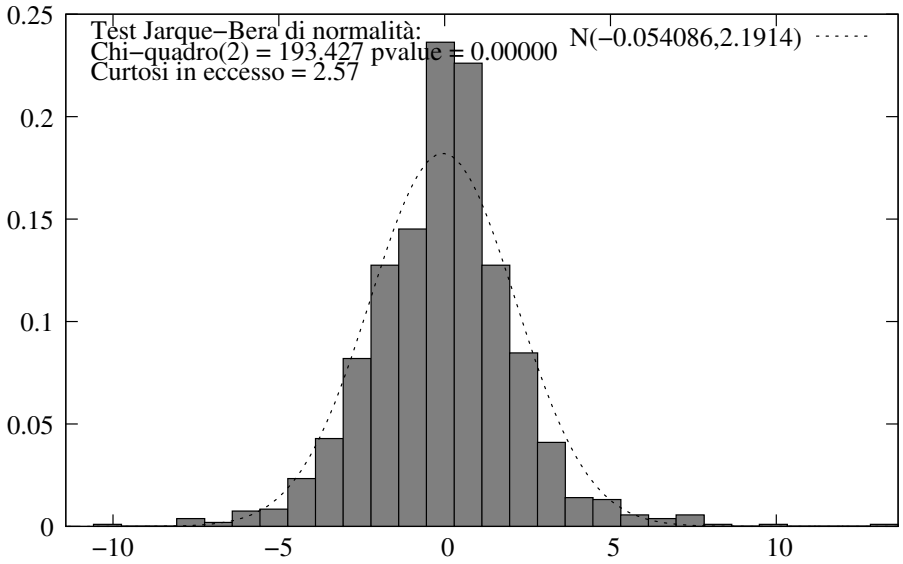


La persistenza nella volatilità si vede molto bene osservando la figura 6.3 (anche questa già vista nell'introduzione), che riporta i valori assoluti dei rendimenti, un indice come un altro della volatilità. Si vede bene che questa serie è soggetta ad una certa persistenza: le sue autocorrelazioni sono ripotrate nella tabella 6.1, e sono tutte statisticamente significative.

Tabella 6.1: Autocorrelazioni dei valori assoluti

Ritardo	1	2	3	4	5	6	7	8	9	10
$\hat{\rho}$	0.184	0.295	0.269	0.258	0.273	0.273	0.231	0.282	0.215	0.243

Figura 6.4: Rendimenti Nasdaq – distribuzione marginale



Un altro fatto molto comune quando si osservano serie finanziarie ad alta frequenza riguarda la forma della loro distribuzione marginale: in un processo ARMA del tipo $A(L)y_t = C(L)\epsilon_t$, si dimostra facilmente che, se la distribuzione congiunta delle ϵ_t è una normale multivariata, allora tale è anche la distribuzione delle y_t.

La distribuzione marginale della serie dei rendimenti dell'indice Nasdaq mette in luce, invece, evidenti tracce di non-normalità (vedi figura 6.4), date soprattutto da un eccesso di curtosi[2]. Questo suggerisce, per lo meno, che se volessimo adattare alla serie osservata un modello ARMA, questo non potrebbe avere disturbi gaussiani, ma dovremmo scegliere un qualche altro tipo di distribuzione, possibilmente dalla code più spesse. In realtà, come vedremo, questo fatto fa il paio con l'altro, in quanto è normale trovare un valore della curtosi piuttosto elevato quando si è in presenza di eteroschedasticità.

[2]Ricordo che l'indice di curtosi per una normale è pari a 3. Poiché la normale fa da pietra di paragone, la curtosi in eccesso è semplicemente la curtosi meno 3. Una distribuzione che abbia curtosi maggiore di 3 si dice *leptocurtica*.

La classe di processi stocastici GARCH, che esamineremo fra poco, ha conosciuto una popolarità enorme appunto perché riesce, a partire da concetti relativamente semplici, a replicare le caratteristiche che abbiamo testè tratteggiato.

6.2 Processi ARCH e GARCH

6.2.1 Processi ARCH

In generale, un processo eteroschedastico (che per il momento supponiamo osservabile) a media 0 può sempre essere scritto nella forma

$$\epsilon_t = u_t \sqrt{h_t}, \tag{6.3}$$

dove u_t è un processo a media 0 e varianza 1 (opzionalmente con persistenza, ma per il momento facciamo conto di no per non complicare inutilmente le cose) e h_t è una sequenza, che può essere deterministica o, caso più interessante, può essere a sua volta un processo stocastico, che supporremo sempre indipendente da u_t. Naturalmente, dalla definizione di varianza si ha che $V(\epsilon_t) = E(h_t u_t^2)$; poiché u_t e h_t sono indipendenti, il valore atteso del prodotto è il prodotto dei valori attesi e quindi $V(\epsilon_t) = E(h_t)E(u_t^2) = E(h_t)$. Si noti che la varianza è costante (e quindi il processo è omoschedastico) solo nel caso in cui $E(h_t)$ è una sequenza costante. In tutti gli altri casi, si ha eteroschedasticità.

È interessante notare, già a questo stadio, che processi eteroschedastici di questo tipo hanno una curtosi maggiore di quella di u_t. La dimostrazione è in sostanza una semplice applicazione della mai abbastanza lodata disuguaglianza di Jensen[3]: il coefficiente di curtosi di ϵ_t può essere scritto come

$$\kappa_\epsilon = \frac{E(\epsilon_t^4)}{E(\epsilon_t^2)^2} = \frac{E(h_t^2 u_t^4)}{E(h_t u_t^2)^2}$$

ma siccome u_t e h_t sono indipendenti, si ha

$$\kappa_\epsilon = \frac{E(h_t^2)E(u_t^4)}{E(h_t)^2 E(u_t^2)^2} = \frac{E(h_t^2)}{E(h_t)^2}\kappa_u;$$

poiché (lemma di Jensen) $E(h_t^2) > E(h_t)^2$, se ne deduce che $\kappa_\epsilon > \kappa_u$.

In un processo eteroschedastico di questo tipo, quindi, se u_t è normale (e quindi ha curtosi pari a 3), ϵ_t sarà sicuramente leptocurtica. Si noti che già stiamo mettendo insieme due dei principali fatti stilizzati di cui si parlava poc'anzi.

Chiaramente, la modellazione statistica di processi di questo tipo passa attraverso la specificazione di una forma funzionale per h_t; altrettanto chiaramente, più questa è semplice, tanto meglio. I processi di tipo ARCH vengono

[3]Per chi se lo fosse dimenticato: la disuguaglianza di Jensen dice che, se $g(\cdot)$ è una funzione convessa, allora $E[g(X)] > g[E(X)]$. Esempio: $E(X^2)$ è sempre maggiore di $E(X)^2$ (e infatti la varianza è sempre positiva). È immediato dimostrare che se la funzione è concava anziché convessa, la diseguaglianza cambia di verso.

fuori quando h_t è una funzione lineare dei valori passati di ϵ_t al quadrato:

$$h_t = c + \sum_{i=1}^{p} a_i \epsilon_{t-i}^2,$$

per cui h_t è una funzione deterministica di variabili che stanno in $\Im_{t-1}$ (si noti che la sommatoria parte da 1 e non da 0). Di conseguenza,

$$V(\epsilon_t|\Im_{t-1}) = E(h_t|\Im_{t-1}) = h_t.$$

È per questo che si usa la sigla ARCH (*AutoRegressive Conditional Heteroskedaticity*): perché siamo in presenza di processi condizionalmente eteroschedastici in cui l'eteroschedasticità deriva da un meccanismo autoregressivo. Ma per capire meglio questo punto, consideriamo il caso più elementare.

Un modello ARCH(p) identifica un processo ϵ_t (che, per il momento, consideriamo per semplicità incorrelato e a media 0) in cui la varianza condizionale è data da

$$h_t = V(\epsilon_t|\Im_{t-1}) = c + A(L)\epsilon_{t-1}^2, \tag{6.4}$$

dove $A(L)$ è un polinomio di ordine $p-1$ (attenzione al -1).

Un aspetto interessante da notare, che oltretutto è particolarmente utile nell'analizzare le proprietà del processo, è che la (6.4) implica che i *quadrati* della serie osservata sono un processo AR(p). Consideriamo infatti la differenza fra ϵ_t^2 e la sua media condizionale:

$$\eta_t = \epsilon_t^2 - h_t.$$

Chiaramente, questa definizione comporta che η_t sia una differenza di martingala:

$$E[\eta_t|\Im_{t-1}] = E\left[\epsilon_{t-1}^2|\Im_{t-1}\right] - E[h_t|\Im_{t-1}] = h_t - h_t = 0.$$

Se postulassimo che esistono anche i suoi momenti secondi, tutte le autocorrelazioni di η_t sarebbero 0, per cui potremmo addirittura dire che η_t è un *white noise*, ma poco importa per il momento[4].

Dalla definzione di η_t discende, ovviamente, anche

$$\epsilon_t^2 = c + A(L)\epsilon_{t-1}^2 + \eta_t \Longrightarrow [1 - A(L)L]\epsilon_t^2 = c + \eta_t \tag{6.5}$$

e siccome il polinomio $[1 - A(L)L]$ è di grado p, allora ϵ_t^2 è un AR(p). Se il polinomio $[1 - A(L)L]$ non ha radici unitarie o esplosive, allora ϵ_t^2 ha un valore atteso non condizionale finito pari a

$$E(\epsilon_t^2) = \frac{c}{1 - A(1)} = E(h_t).$$

Poiché supponiamo che ϵ_t abbia media 0, il valore atteso del suo quadrato è anche la sua varianza.

[4]Certo, sarebbe un *white noise* un po' strano: ad esempio, è vero che la sua media condizionale è zero, ma non vuol dire che η_t è *indipendente* da $\Im_{t-1}$: ad esempio, è ovvio che il suo supporto (ossia l'insieme dei valori su cui è definita) è limitato verso il basso, visto che per $\eta_t < h_t$ si avrebbe un quadrato negativo (?) nella (6.5).

Esempio 6.2.1 (ARCH(1)) *Un esempio di processo ARCH(1) è dato da:*

$$h_t = 0.1 + 0.8\epsilon_{t-1}^2.$$

La varianza condizionale del processo, quindi, è variabile. Tuttavia, in questo caso si può scrivere

$$[1 - 0.8L]\epsilon_t^2 = 0.1 + \eta_t$$

e si vede facilmente che la sua varianza non condizionale è costante e pari a

$$E\left[\epsilon_{t-1}^2\right] = E(h_t) = V(\epsilon_t) = 0.5.$$

Conclusione: ϵ_t è un processo che ha varianza finita e costante, se consideriamo la sua distribuzione marginale, ma la varianza *condizionale al set informativo* $\Im_{t-1}$ non è che h_t, che varia al variare di t. È per questo che parliamo di eteroschedasticità condizionale.

Si noti che, in questo contesto, la semplice struttura della legge di moto per la varianza condizionale ci permette anche di costruire in modo piuttosto semplice un test, pertendo dal test di eteroschedasticità di Breusch e Pagan, ampiamente usato in un constesto di regressione lineare: in questo test, si assume che h_t sia rappresentabile come una qualche funzione del tipo

$$h_t = H(z_t'\gamma),$$

dove $H(\cdot)$ è una qualche funzione derivabile, non importa quale. Si arriva a dimostrare, con procedimento piuttosto ingegnoso, che con una regressione ausiliaria del tipo

$$e_t^2 = \phi_0 + z_t'\phi_1 + \text{residui}$$

si può costruire una statistica test, basata sui moltiplicatori di Lagrange, moltiplicando l'R^2 per il numero di osservazioni. Questa statistica ha, sotto appropriate ipotesi, una distribuzione asintotica χ^2 con tanti gradi di libertà quanti sono gli elementi di ϕ_1. Nel nostro caso, la funzione $H(\cdot)$ è l'equazione (6.4); di conseguenza, per fare un test sulla presenza di effetti ARCH basta fare una regressione OLS dei quadrati dei residui di un modello per la media condizionale sui loro ritardi. Questo test è appunto noto come **test LM-ARCH**.

Un processo ARCH, quindi, presenta varie caratteristiche che lo rendono molto appetibile per modellare serie come quelle che abbiamo visto al paragrafo precedente. Il problema che spesso si pone è però che la struttura di persistenza della volatilità è tale per cui sarebbero necessari polinomi di ordine piuttosto alto per replicarla: considerate la tabella 6.1. È evidente che le autocorrelazioni sono tutte "piccole", e però un autoregressivo di ordine basso non è molto adatto, in quanto la struttura di autocorrelazione è molto persistente.

Questo rischia di essere un problema per via dei vincoli sui parametri che è sensato imporre in un ARCH. Infatti, è necessario che tutti i coefficienti nel polinomio $A(L)$ nella (6.4) siano positivi: se così non fosse, sarebbe possibile il verificarsi di un evento che porta la varianza condizionale ad essere negativa (ammetterete che è imbarazzante). D'altro canto, non basta che i coefficienti

siano positivi: se vogliamo che il processo possieda anche una varianza marginale, dobbiamo anche escludere le radici unitarie. Tutti questi vincoli spesso fanno sorgere seri problemi computazionali: ne parlerò fra poco.

Morale: c'è la possibilità che l'ordine del polinomio $A(\cdot)$ sia troppo alto per essere gestibile. Utilizzando la stessa logica che ci ha condotti alla specificazione dei modelli ARMA, si potrebbe alleviare il problema (se non addirittura eliminarlo), mettendo un bel polinomio anche alla destra del segno di uguale nella (6.5): così facendo, arriviamo ai processi GARCH.

6.2.2 Processi GARCH

I processi GARCH sono una generalizzazione degli ARCH (GARCH sta per *Generalized ARCH*), nello stesso modo in cui i processi ARMA sono una generalizzazione degli AR: infatti, in questo tipo di processi la varianza condizionale dipende, oltre che dai valori passati di ϵ_t^2 (come negli ARCH), anche dai propri valori passati. Avremo perciò

$$h_t = c + A(L)\epsilon_{t-1}^2 + B(L)h_{t-1}. \tag{6.6}$$

Un processo la cui varianza condizionale segua la (6.6) si dice GARCH(p,q), dove i polinomi $A(\cdot)$ e $B(\cdot)$ sono di ordine $p-1$ e $q-1$ rispettivamente.

Con la stessa logica che abbiamo applicato nel sottoparagrafo precedente, possiamo mostrare che un GARCH(p,q) può anche essere scritto come un ARMA. Infatti, ricordando la definizione di η_t,

$$\eta_t = \epsilon_t^2 - h_t,$$

si ricava facilmente

$$\epsilon_t^2 = c + A(L)\epsilon_{t-1}^2 + B(L)\left[\epsilon_{t-1}^2 - \eta_{t-1}\right] + \eta_t$$

ossia

$$\{1 - [A(L) + B(L)]\,L\}\,\epsilon_t^2 = c + (1 - B(L)L)\eta_t \tag{6.7}$$

e quindi ϵ_t^2 è un'ARMA(max$(p,q),q$)[5].

Anche in questo caso, il parallelismo fra ARMA e GARCH ci rivela una cosa interessante: un GARCH è fondalmentalmente un ARCH(∞). Infatti, la (6.7) si potrebbe anche scrivere come

$$\epsilon_t^2 = \frac{c}{1 - [A(1) + B(1)]} + \frac{(1 - B(L)L)}{1 - [A(L) + B(L)]\,L}\eta_t = \mu + C(L)\eta_t,$$

che è un MA(∞), e quindi ϵ_t è un ARCH(∞).

Esempio 6.2.2 (GARCH(1,1)) *Facciamo un esempio con un GARCH(1,1). Se*

$$h_t = 0.5 + 0.3\epsilon_{t-1}^2 + 0.6h_{t-1}.$$

[5]Notate la finezza: $B(L)$ può essere uguale a 0, e avremmo un semplice ARCH, ma se $A(L)$ fosse uguale a 0 la rappresentazione ARMA della (6.7) conterrebbe un bel fattore comune, per cui $B(L)$ non sarebbe identificato. Vedi il sottoparagrafo 2.7.2.

Ripetiamo il ragionamento di prima e scriviamo

$$\epsilon_t^2 = 0.5 + 0.3\epsilon_{t-1}^2 + 0.6\left[\epsilon_{t-1}^2 - \eta_{t-1}\right] + \eta_t.$$

ossia

$$[1 - 0.9L]\,\epsilon_t^2 = 0.5 + (1 - 0.6L)\eta_t,$$

cioè un ARMA(1,1). Anche qui, la varianza non condizionale è costante:

$$E\left[\epsilon_{t-1}^2\right] = E(h_t) = V(\epsilon_t) = \frac{0.5}{1 - 0.9} = 5.$$

Anche qui, perché il modello abbia senso, dobbiamo imporre che i parametri α_i siano maggiori di 0 e le β_i siano non-negative[6].

6.2.3 Stima dei GARCH

Come si stima un GARCH? Uno potrebbe prendere l'equazione (6.7) e ragionare: se ϵ_t^2 è un ARMA, dov'è il problema? Calcolo i quadrati e via.

A far così, c'è qualche problema. Per cominciare: non è detto che osserviamo ϵ_t. In effetti, quasi sempre i modelli che stimiamo sono del tipo

$$y_t = \mu_t + \epsilon_t,$$

dove c'è una parte di media condizionale (che naturalmente può contenere un nucleo deterministico, variabili esogene, valori passati della y_t eccetera) i cui parametri non sono noti, e quindi ϵ_t non è osservabile direttamente. Si potrebbe pensare di stimare la parte di media condizionale con le solite tecniche (OLS o che so io) e poi lavorare sui quadrati dei residui. Vi sarete accorti da soli, però, che questa procedura è un po' rabberciata. Non che non funzioni mai. Anzi, in molti casi si può dimostrare che produce stimatori consistenti; anche in questi casi fortunati, però, gli stimatori non sono efficienti. La perdita di efficienza deriva in sostanza da due fatti: primo, stiamo stimando i parametri un po' per volta anziché tutti insieme. Secondo, l'errore di previsione a un passo della (6.7) η_t, è evidentemente non normale: ad esempio, come facevamo notare poco fa, il suo supporto è limitato. Inoltre, non è detto che la varianza di η_t esista (si può dimostrare, ma non è essenziale), ciò che compromette seriamente le proprietà asintotiche degli stimatori ARMA, che normalmente sono basati su una verosimiglianza gaussiana.

Meglio stimare tutto con la massima verosimiglianza: ancora una volta, ci viene in soccorso la fattorizzazione sequenziale. Infatti, la distribuzione condizionale di y_t, sotto le ipotesi che abbiamo fatto, è gaussiana:

$$y_t | F_{t-1} \sim N\left[\mu_t, h_t\right],$$

[6]Come altrove, ho semplificato. In realtà, il modello ha senso solo per quei valori dei parametri per cui la rappresentazione ARCH(∞) abbia solo coefficienti positivi. Nel caso GARCH(1,1), condizione necessaria e sufficiente è effettivamente che ambedue i coefficienti siano positivi e che la loro somma sia minore di 1. Per casi di ordini superiore, valgono condizioni più complicate, note come condizioni di Nelson-Cao, che però qui non vi racconto.

per cui, se prendiamo y_0 come fisso, la log-verosimiglianza si può scrivere

$$L(\theta) = \sum_{t=1}^{T} \ell_t, \tag{6.8}$$

dove

$$\ell_t = \log\left[\frac{1}{\sqrt{h_t(\theta)}}\phi\left(\frac{y_t - \mu_t(\theta)}{\sqrt{h_t(\theta)}}\right)\right] = -\frac{1}{2}\left[\log(h_t(\theta)) + \frac{(y_t - \mu_t(\theta))^2}{h_t(\theta)}\right]$$

e θ è, naturalmente, un vettore che contiene i parametri della media e della varianza condizionali (uso come di consueto la notazione $\phi(\cdot)$ per indicare la densità della normale standardizzata).

La massimizzazione di questa funzione avviene coi soliti metodi iterativi di cui ho già parlato al sottoparagrafo 2.7.1. In questi casi, però, la convergenza è spesso più difficile che nei modelli ARMA, per cui il modello "che non converge" è un'eventualità tristemente non rara.

Quasi sempre la colpa è dei vincoli da imporre sui parametri. Può succedere che il massimo della funzione di verosimiglianza si trovi su un punto esterno all'insieme dei valori ammissibili per i parametri (ad esempio, $\alpha + \beta < 1$ per un GARCH(1,1)). Questo può accadere quando il modello è mal specificato, oppure quando il campione è troppo piccolo[7].

Normalmente, le routine di stima contengono dei controlli per far sì che la log-verosimiglianza non venga valutata per valori "assurdi" dei parametri. Poiché i metodi iterativi a cui ho accennato nel sottoparagrafo 2.7.1 sono congegnati per andare sempre in salita, a volte si assiste allo spettacolo malinconico dell'algoritmo che fa come la mosca contro la finestra chiusa: si avventa sul vetro, sbatte sul vetro, rimbalza, si ri-avventa e così via *ad libitum*. In quei casi c'è poco da fare, se non cercare di specificare meglio il modello e affidarsi a una divinità di vostra scelta.

Si noti: il problema in questi casi viene dai dati, non dal *software* che utilizziamo per le stime. Spesso, però è a quest'ultimo che si dà la colpa. Per evitare questo problema, a volte i programmatori pescano nel torbido: ho visto coi miei occhi un noto pacchetto commerciale produrre una stima di un GARCH(1,1) in cui α era negativo. Lo stesso pacchetto ha la surrettizia abitudine di presentare le stime come se la convergenza fosse avvenuta anche quando in realtà l'algoritmo si ferma perché è stato raggiunto il numero massimo di iterazioni. Mah.

6.3 Un esempio

Prendiamo la serie mostrata in figura 6.2 e mettiamoci su un modello GARCH, ossia stimiamo i parametri contenuti nella coppia di equazioni

$$y_t = \mu + \varphi y_{t-1} + \epsilon_t \tag{6.9}$$
$$h_t = c + a\epsilon_{t-1}^2 + bh_{t-1} \tag{6.10}$$

[7] Si badi che in questo contesto 100 osservazioni sono ancora un campione *molto* piccolo. Per fortuna, nelle applicazioni di finanza la scarsità di dati non è quasi mai un problema.

Rispetto all'esposizione dei sottoparagrafi precedenti, qui supponiamo che il processo osservabile non sia più ϵ_t, bensì y_t. La parte eteroschedastica sta nel fatto che il processo GARCH ϵ_t di cui ci interessa stimare i parameteri è semplicemente la differenza fra y_t e il suo valore atteso condizionale (il "termine di disturbo").

In pratica, stiamo stimando un modello AR(1) con errori GARCH(1,1); il termine autoregressivo all'equazione della media (6.9) non dovrebbe servire, ma non si sa mai, male non fa (questo è solo un esempio, per un lavoro "vero" si ragiona con meno pressappochismo — almeno io ci provo).

Tabella 6.2: Stime GARCH (ML)

Coefficiente	Stima	Errore std.	Statistica t	p-value
μ	0.038	0.043	0.891	0.373
φ	-0.007	0.028	-0.240	0.810
c	0.009	0.009	1.034	0.301
a	0.060	0.014	4.468	<1e-05
b	0.938	0.014	68.741	<1e-05

Diamo un'occhiata alle stime: come previsto, di persistenza nella media non ce n'è (il parametro φ non risulta significativamente diverso da 0). Visto che anche la costante μ è praticamente zero, possiamo continuare dicendo che la funzione di media condizionale $E(y_t|\Im_{t-1})$ è, a tutti i fini pratici, zero. Di persistenza nella varianza, invece, ce n'è eccome: i parametri a e b sono ambedue ben diversi da 0, e indicano una consistente persistenza nella varianza. Anzi, dirò di più: dalle stime si deduce che la rappresentazione ARMA per ϵ_t^2 è data da

$$(1 - (a + b)L)\epsilon_t^2 = c + \eta_t - b\eta_{t-1},$$

ovvero

$$(1 - 0.998L)\epsilon_t^2 = 0.009 + \eta_t - 0.938\eta_{t-1}.$$

Si vede facilmente che ϵ_t^2 è un processo molto persistente se non proprio $I(1)$; non è un caso: moltissime serie finanziarie esibiscono lo stesso comportamento a frequenza giornaliera. Anzi, ci sono dei modelli (i cosiddetti IGARCH) che *impongono* la radice unitaria su ϵ_t^2; ne parleremo più avanti.

Da quanto sopra, si può dedurre che la varianza non condizionale di y_t è uguale a

$$V(y_t) = \frac{c}{1 - a - b} = 4.9963;$$

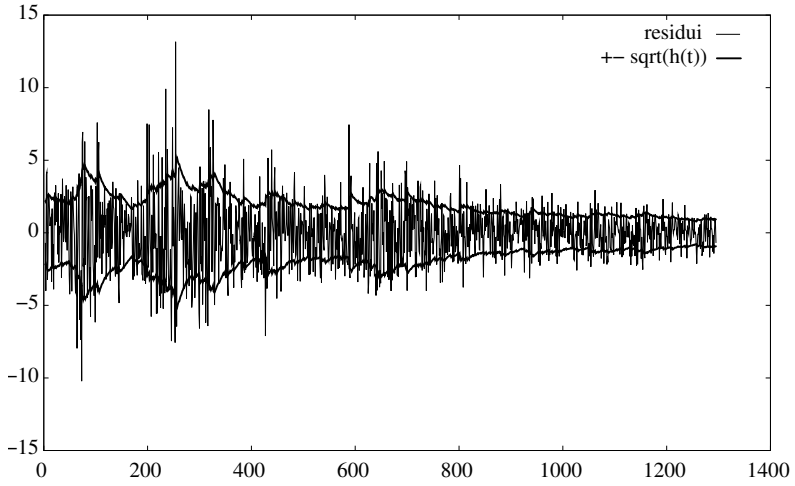
l'ordine di grandezza della varianza non condizionale avrebbe potuto anche essere stimato utilizzando la varianza campionaria di y_t, che infatti risulta 4.8022. Questo indicatore, tuttavia, non dà la misura del fatto che la varianza condizionale h_t è molto variabile. Quest'ultima può essere ricostruita *ex post* dagli errori di previsione e_t e dalle stime dei parametri come segue: fissiamo $\hat{h}_0$ ad un valore "plausibile" (una stima qualunque della varianza non

condizionale va benissimo), dopodiché possiamo calcolare $\hat{h}_t$ come

$$\hat{h}_t = \hat{c} + \hat{a}e_{t-1}^2 + \hat{b}\hat{h}_{t-1};$$

un modo abbastanza comune di presentare il tutto è in un grafico come quello in figura 6.5.

Figura 6.5: Rendimenti Nasdaq – residui e deviazione standard stimata



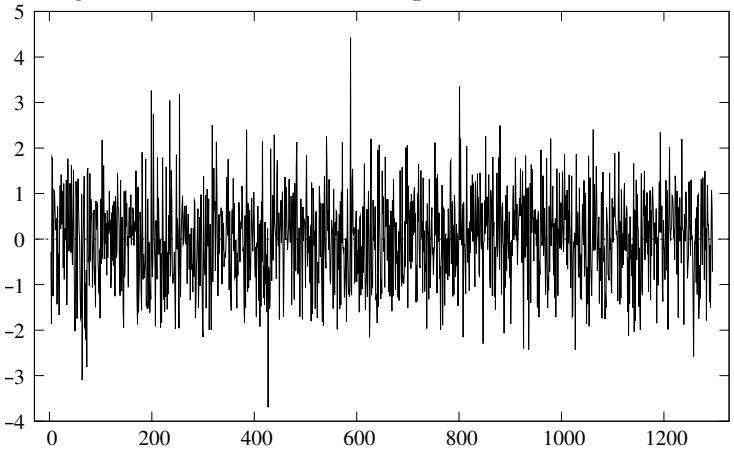
Le serie disegnate con la linea più spessa sono $\sqrt{\hat{h}_t}$, ossia la deviazione standard condizionale stimata e il suo negativo. Come uno si aspetterebbe, i residui, che più o meno coincidono con la serie stessa visto che la media condizionale è press'a poco zero, sono, per così dire, "contenuti" all'interno della deviazione standard un buon numero di volte. Si noti anche come, ogni volta che si verifica una variazione consistente nell'indice (e quindi un valore della serie molto lontano da 0), il valore di h_t stimato aumenti immediatamente dopo, per poi ridiscendere gradualmente fino al prossimo shock. Se andiamo a vedere alcune statistiche descrittive di $\hat{h}_t$, si notano alcuni particolari interessanti: la media è 4.8406 (di nuovo, una stima consistente della varianza non condizionale), mentre minimo e massimo sono pari, rispettivamente, a 0.59961 e 26.667 (0.77434 e 5.16401 in termini di deviazione standard), numeri che rendono molto bene l'idea di come la volatilità fluttui all'interno del campione considerato (che, ricordo, comprende sia l'11 settembre 2001 che tutto il periodo del *dot-com crash* che ha portato l'indice Nasdaq a perdere quasi l'80% del suo valore fra marzo 2000 e settembre 2002).

Dovrebbe essere lampante che, con le stime in mano, possiamo anche impostare un modello previsivo: non tanto per la media (ché non c'è persistenza sfruttabile), quanto piuttosto per la volatilità. Infatti, basta prendere la (6.10) e sostituire ai valori veri quelli stimati.

È anche interessante notare che il modello rende bene conto del fatto che i rendimenti abbiano una distribuzione marginale leptocurtica: se infatti calcoliamo i rendimenti "standardizzati", cioè definiamo una serie

$$u_t = \frac{y_t}{\sqrt{\hat{h}_t}}$$

Figura 6.6: Rendimenti Nasdaq – serie standardizzata



otteniamo la serie mostrata in figura 6.6. Si vede "a occhio" che l'eteroschedasticità è scomparsa. Più interessante è notare che la curtosi in eccesso della serie così trasformata si riduce a 0.15422, ed il test Jarque-Bera è pari a 2.747, con un *p-value* di 0.29; se ne deduce che si può accettare l'ipotesi che il modello con errori normali sia una buona rappresentazione dei dati, e quindi che la non-normalità della distribuzione marginale di y_t sia dovuta interamente all'effetto GARCH.

6.4 Estensioni

6.4.1 Distribuzioni non-normali

Nell'esempio precedente avevamo osservato come la persistenza in volatilità fornisse una spiegazione esauriente dell'eccesso di curtosi. Non sempre è così: a volte, anche normalizzando la serie dei residui per la deviazione standard stimata, permangono *outlier* tali che l'ipotesi di normalità risulta insostenibile. Per cui, invece della normale si adoperano a volte distribuzioni la cui curtosi possa prendere un valore diverso da 3; le due più usate sono la t di Student o la *Generalised Error Distribution*, o GED per brevità.

Queste due distribuzioni hanno una funzione di densità che dipende da un parametro aggiuntivo ν che deve essere anch'esso stimato, assieme a quelli della media e della varianza. Per la t il parametro ν è sempre positivo (può essere non intero) e si chiama "gradi di libertà"; per qualunque valore del parametro, la densità corrispondente è leptocurtica, ma tende alla normale per $\nu \to \infty$. Anche nella GED, ν è un reale positivo, ma la curtosi in eccesso è positiva per $\nu < 2$ e negativa per $\nu > 2$. Il valore 2 è speciale perché la GED con parametro 2 coincide con la normale (in altre parole, la normale è un caso particolare della GED).

Vi risparmio i dettagli di come sono fatte le funzioni di densità, ma il senso della scelta può essere apprezzato visivamente considerando la figura

6.7, che mostra la densità di alcuni esempi di t e di GED (normalizzate ad avere varianza 1) per diversi valori di ν.

Quindi, una possibile strategia empirica può essere quella di stimare un modello GARCH con una diversa ipotesi distribuzionale, ciò che ammonta a usare una definizione diversa di ℓ_t nella 6.8. Se si usa la GED, fra l'altro sia ha anche il vantaggio che si può testare l'ipotesi di normalità attraverso l'ipotesi $\nu = 2$. Il problema che si ha però in questi casi è che la massimizzazione numerica della log-verosimiglianza è considerevolmente più onerosa dal punto di vista computazionale: visto che, come ho già detto, la non convergenza di una stima GARCH è un evento raro ma non impossibile, questo ulteriore problema rischia di complicare ulteriormente le cose.

Una strada alternativa molto ingegnosa è quella di usare un metodo di stima conosciuto come quasi-massima verosimiglianza. L'idea è: sotto alcune ipotesi di regolarità (che non vi dico, ma non sono particolarmente stringenti), la stima dei parametri che si ottiene massimizzando una verosimiglianza gaussiana è consistente, ancorché non efficientee, anche se la densità "vera" è diversa. Ciò che *non* è consistente è lo stimatore della matrice varianze-covarianze dei parametri stessi. A questo, tuttavia, c'è rimedio. È sufficiente utilizzare uno stimatore robusto, e questo è spesso disponibile nei principali pacchetti econometrici. Quello che normalmente si usa in questo contesto è il cosiddetto stimatore di Bollerslev e Wooldridge.

La logica è la stessa che conduce, ad esempio, ad utilizzare lo stimatore di White in una regressione cross-section: in quel caso, continuiamo a usare la statistica OLS come stimatore, perché tanto è consistente e ci rassegniamo a una perdita di efficienza, che peraltro può essere trascurabile. Tuttavia, per poter fare inferenza in modo corretto, dobbiamo usare uno stimatore *ad hoc* per la matrice varianze-covarianze.

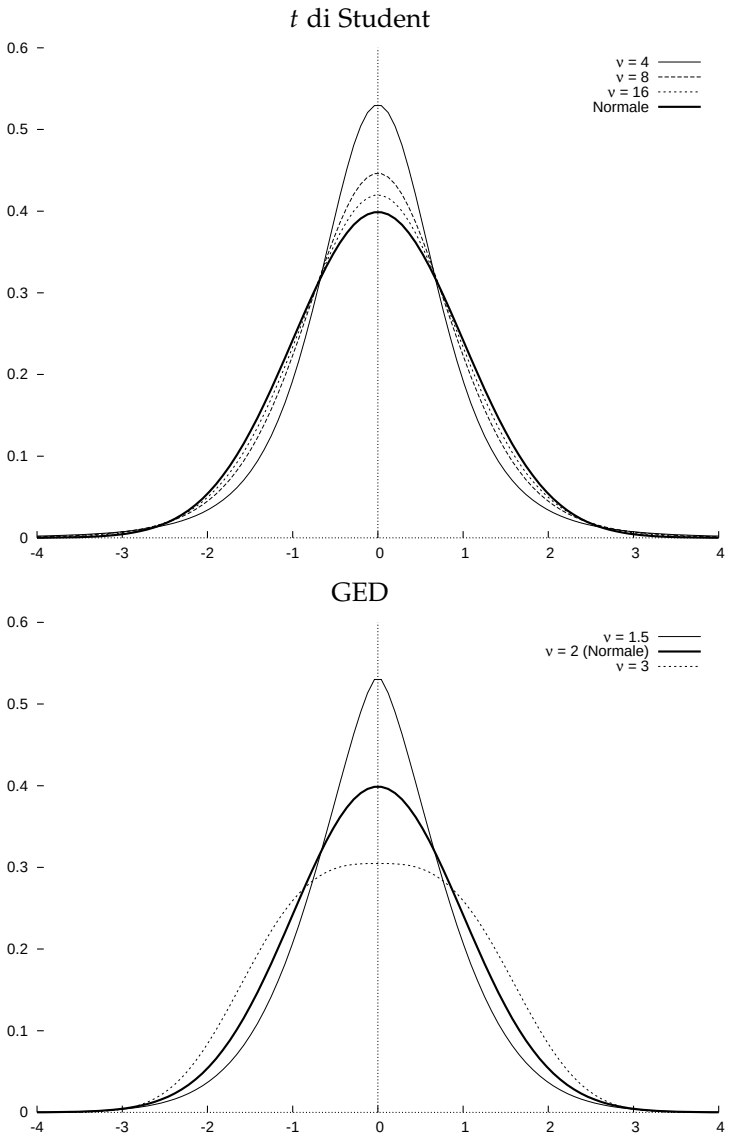
La similarità fra le due situazioni non è casuale: tutta la teoria che c'è dietro è stata prodotta da uno studio attento e complesso dei modelli scorrettamente specificati, il cui pioniere è, guarda un po', proprio Hal White.

La tavola 6.3 riporta le stesse stime dell'esempio riportato nella tavola 6.2, ma con lo stimatore robusto per la matrice varianze-covarianze. Come si vede, le stime dei coefficienti sono le stesse, ma cambiano quelle degli errori standard, anche se in questo caso non di tanto.

Tabella 6.3: Stime GARCH (QML)

Coefficiente	Stima	Errore std.	Statistica t	p-value
μ	0.038	0.045	0.842	0.400
ϕ	-0.007	0.026	-0.258	0.797
c	0.009	0.009	1.040	0.299
a	0.060	0.016	3.855	0.000
b	0.938	0.015	60.911	<1e-05

Figura 6.7: Distribuzioni alternative alla normale



6.4.2 Effetti asimmetrici

$$h_t = c + A(L)\left[\epsilon_{t-1} + \gamma|\epsilon_{t-1}|\right]^2 + B(L)h_{t-1}. \tag{6.11}$$

news impact curve

6.4.3 EGARCH

Questa è una formulazione meno generale di quella "vera", ma è quella che più spesso si adotta:

$$\ln h_t = c + A(L)\left[u_{t-1} + \gamma|u_{t-1}|\right]^2 + B(L)\ln h_{t-1}. \tag{6.12}$$

Vantaggi dell'EGARCH:

- Non c'è bisogno di imporre vincoli di non-negatività sui parametri. Ergo, non si verificano i problemi di convergenza che ogni tanto si hanno sui GARCH;
- include l'idea degli effetti asimmetrici in modo molto naturale.

Svantaggi:

- Scrivere le derivate analitiche è molto complicato e quasi nessun pacchetto che io conosca lo fa: per lo più, ci si affida alla differenziazione numerica con la conseguenza che se c'è qualche parametro in cui la prima cifra non zero viene 4-5 posti dopo la virgola ci possono essere problemi di convergenza e/o di precisione numerica.
- Se interessa una previsione di h_t, non è immediatissimo capire come calcolarla. Si ricordi che, di solito, si utilizza come previsore la media condizionale, ma l'EGARCH ci fornisce, tutt'al più, il valore atteso di $\ln h_t$. Come si sa,
$$E(e^X) \neq e^{E(X)}$$
e quindi la semplice esponenziazione di $\widehat{\ln h_t}$ ha delle proprietà, come previsiore, non molto chiare.

6.4.4 GARCH-in-mean

$$y_t = x_t'\beta + \varphi h_t + \epsilon_t \tag{6.13}$$
$$h_t = V(\epsilon_t|\Im_{t-1}) \tag{6.14}$$

In un contesto in cui le y_t siano rendimenti di un'attività finanziaria, il parametro φ è facile da interpretare come misura del premio al rischio: infatti, esso ci dice, di quanto deve variare il rendimento all'aumentare della volatilità dello stesso. In pratica, però, questi modelli stanno andando un po' in disuso perché spesso il termine φh_t va a cogliere, più che questo, altri effetti non compresi nel set informativo osservabile e la stima del parametro φ risulta incompatibile con valori ragionevoli.

6.4.5 IGARCH

6.4.6 Modelli multivariati

Problemi:

1. Troppi parametri
2. definitezza di Σ_t

- CCC/DCC
- BEKK

Capitolo 7

Per approfondimenti

7.1 In generale

In questa dispensa, non si presuppone nel lettore più che una conoscenza degli elementi di base di statistica e di econometria. Questo ha portato, in molti casi, a semplificazioni draconiane. Tanto per cominciare, il lettore già navigato avrà notato un disinteresse sovrano per tutti i punti più squisitamente tecnici di natura probabilistico-inferenziale, che sopratutto nell'analisi dei processi non stazionari possono essere decisamente impegnativi. Se proprio vi piace la teoria (del tipo che volete sapere *veramente* cos'è l'ergodicità), allora date un'occhiata a Davidson (1994), che è meraviglioso (ma tozzo assai). Più abbordabile McCabe e Tremayne (1993).

In realtà, quel che c'è da sapere per rincorrere le tante occasioni in cui dico di consultare la letteratura rilevante lo si trova in tutti i testi di econometria, a qualunque livello. Fra i più diffusi ci sono in inglese Greene (1997), che è diventato un po' lo standard oppure Davidson e McKinnon (1993), più avanzato ma assolutamente da consigliare per uno studio approfondito. Bello e recente, snello ma abbastanza completo è anche Verbeek (2000). In italiano, Peracchi (1995) è molto bello, forse addirittura troppo; ad un economista applicato consiglierei probabilmente in alternativa Favero (1994); una via di mezzo è Cappuccio e Orsi (1992).

Va detto inoltre che un argomento che brilla per la sua assenza in questa dispensa è l'analisi spettrale. Per chi volesse, in italiano si può consigliare Piccolo (1990) per un'introduzione, mentre per approfondire vanno benissimo Hamilton (1994) (che è stato anche tradotto in italiano) e Brockwell e Davis (1991), che però non è specificamente rivolto ad economisti.

7.2 Processi univariati

Sulle serie storiche univariate, un riferimento un po' datato, ma che vale sempre la pena di avere a portata di mano è Granger e Newbold (1986). Più recenti e piuttosto completi sono i già menzionati Hamilton (1994) e Brockwell e Davis (1991). Un testo di taglio molto personale, che può piacere molto

o non piacere affatto, è Harvey (1993)[1]. Di grande interesse è anche Sargent (1987), soprattutto perché mostra come certi concetti di analisi delle serie (che peraltro spiega molto bene) siano applicabili in un contesto economico-teorico piuttosto che statistico.

7.3 Processi VAR

Anche qui l'ubiquo Hamilton (1994). Volendo approfondire ancora, è quasi obbligatoria la lettura dell'articolo che ha dato il via a tutta la letteratura sui VAR, e cioè Sims (1980), e che contiene anche interessanti riflessioni sull'uso dell'analisi delle serie in economia; più in generale, consiglio Lütkepohl (1991) oppure (ma è un po' verboso) Ooms (1994). Un altro riferimento eccellente è Canova (1995).

Sui VAR strutturali una eccellente monografia è Amisano e Giannini (1997), che però è un pochino avanzata per il lettore medio.

7.4 Processi $I(1)$ e cointegrazione

Qui entriamo in un campo più recente, ed è difficile consigliare cose che non siano un po' avanzate. In italiano, è secondo me ottimo per iniziare il contributo di uno dei maggiori esperti italiani di cointegrazione, è cioè Rocco Mosconi, in Mosconi (1994).

In inglese, anche in questo caso Hamilton (1994) è un eccellente punto di partenza, anche se su certe cose lo considero un po' involuto. Maddala e Kim (1998) è, invece, un testo più recente e di impianto del tutto diverso: non ci sono grandi dimostrazioni di teoremi, ma una rassegna mostruosamente completa della letteratura rilevante, unita a giudizi spesso taglienti ma sempre pregnanti. Recente e per certi aspetti geniale è anche Davidson (2000).

Un testo interessantissimo, che a mio avviso bilancia molto bene teoria e prassi, è Banerjee *et al.* (1993). A livello più abbordabile, consiglio Enders (1995) e Cuthberson *et al.* (1992). A livello teorico, le caratteristiche delle serie integrate sono state studiate a fondo soprattutto da P.C.B. Phillips, il quale si è espresso al suo meglio in una serie di articoli, fra i quali consiglio in particolare Phillips (1986) e Phillips e Durlauf (1986). Molto interessante e più mirato all'economista applicato è anche Campbell e Perron (1991).

Per quanto riguarda i VAR non stazionari, un riferimento recente e molto gradevole è Mills (1998), che contiene una rassegna leggibile anche di argomenti di solito considerati piuttosto esoterici, come ad esempio i test di Granger-causalità in VAR cointegrati. Per un riferimento più esteso sull'interpretazione dei modelli cointegrati, e soprattutto sul loro legame coi modelli ECM, non si può non consigliare il papà dei modelli ECM, e cioè David Hendry: fra tutta la sua sterminata produzione, conviene segnalare Hendry (1995), oltre al già detto Banerjee *et al.* (1993).

[1] Per quel che conta, a me piace.

Per quanto riguarda invece la procedura di Johansen, la prima fonte è naturalmente Johansen stesso in Johansen (1995). Questo libro è un po' impegnativo, cosicché chi volesse semplicemente farsi un'idea può tranquillamente rifarsi a Hamilton (1994), che spiega anche abbastanza bene la tecnica Fully-Modified OLS, oppure a Johansen (2000), che è recente e sintetico; Boswijk e Doornik (2003) è un articolo ben fatto, che fra l'altro ha anche il pregio di poter essere scaricato da Internet (per adesso). Da segnalare anche Hargreaves (1994), che contiene una disamina molto distesa, seppure un po' datata, dei principali metodi di stima dei vettori di cointegrazione. Di testi sulla cointegrazione, comunque, ne sono usciti così tanti che si fa fatica anche solo a tenerne il conto; uno molto introduttivo, che considero particolarmente efficace dal punto di vista didattico è Harris (1995).

7.5 Processi ad eteroschedasticità condizionale

Siccome ormai ve lo sarete fotocopiato o scaricato (o meglio ancora, comprato), Hamilton (1994) contiene anche una discussione non superficiale dei modelli ARCH/GARCH, anche se mi corre l'obbligo di indicare a chi volesse approfondire Bollerslev *et al.* (1994), che è l'assoluta Bibbia sull'argomento, anche se ormai non più aggiornatissima. Aggiornatissimo, introduttivo e molto ben ragionato è anche Zivot (2008).

È però vero che, con l'esplosione delle applicazioni dei GARCH in finanza, i riferimenti bibliografici non si contano più. Anzi, ormai c'è proprio un genere letterario autonomo, che è quello che mischia abilmente finanza ed econometria. Su questo tenore, il riferimento classico è Campbell *et al.* (1997). Due ottime trattazioni in italiano, che consiglio caldamente anche perché ben tarate sui problemi pratici della finanza applicata sono Pastorello (2001) e Gallo e Pacini (2002).

Bibliografia

Amisano, G. e Giannini, C. (1997). *Topics in Structural VAR Econometrics*. Springer-Verlag, 2a ed.

Banerjee, A., Dolado, J., Galbraith, J. e Hendry, D. (1993). *Co-Integration, Error Correction and the Econometric Analysis of Non-Stationary Data*. Oxford University Press.

Bollerslev, T., Engle, R. F. e Wooldridge, J. (1994). *ARCH models*. In *Handbook of Econometrics* (curato da Engle, R. F. e McFadden, D. L.), vol. 4, pp. 2959–3031. Elsevier.

Boswijk, H. P. e Doornik, J. (2003). *Identifying, estimating and testing restricted cointegrated systems: An overview*. Rap. tecn., Economics Group, Nuffield College, University of Oxford.

Brockwell, P. J. e Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, 2a ed.

Campbell, J. Y., Lo, A. W. e Mckinley, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Campbell, J. Y. e Perron, P. (1991). *Pitfalls and opportunities: What macroeconomists should know about unit roots*. In *NBER Macroeconomics Annual 1991* (curato da Blanchard, O. J. e Fischer, S.), pp. 141–201. MIT Press.

Canova, F. (1995). *Vector autoregressive models: Specification, estimation, inference and forecasting*. In *Handbook of Applied Econometrics* (curato da Pesaran, H. e Wickens, M.), vol. I: Macroeconomics. Blackwell.

Cappuccio, N. e Orsi, R. (1992). *Econometria*. Il Mulino.

Cuthberson, K., Hall, S. G. e Taylor, M. P. (1992). *Applied Econometric Techniques*. Philip Allan.

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

— (2000). *Econometric Theory*. Blackwell.

Davidson, R. e McKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press.

Enders, W. (1995). *Applied Economic Time Series Analysis*. John Wiley & Sons.

Favero, C. A. (1994). *Econometria*. Nuova Italia Scientifica.

Gallo, G. M. e Pacini, B. (2002). *Metodi quantitativi per i mercati finanziari*. Carocci.

Granger, C. W. e Newbold, P. (1986). *Forecasting Economic Time Series*. Academic Press.

Greene, W. (1997). *Econometric Analysis*. Prentice Hall, 3a ed.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

Hargreaves, C. P. (1994). *A review of methods of estimating cointegrating relationships*. In *Nonstationary Time Series Analysis and Cointegration* (curato da Hargreaves, C. P.), pp. 87–131. Oxford University Press.

Harris, R. (1995). *Using Cointegration Analysis in Econometric Modelling*. Prentice-Hall.

Harvey, A. C. (1993). *Time Series Models*. Harvester Wheatsheaf, 2a ed.

Hendry, D. E. (1995). *Dynamic Econometrics*. Oxford University Press.

Johansen, S. (1995). *Maximum Likelihood Inference in Co-Integrated Vector Autoregressive Processes*. Oxford University Press.

— (2000). *Modelling of cointegration in the vector autoregressive model*. Economic Modelling, 17: 359–373.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Maddala, G. e Kim, I. (1998). *Unit Roots, Cointegration and Structural Change*. Cambridge University Press.

McCabe, B. e Tremayne, A. (1993). *Elements of modern asymptotic theory with statistical applications*. Manchester University Press.

Mills, T. C. (1998). *Recent developments in modelling nonstationary vector autoregressions*. Journal of Economic Surveys, 12(3): 279–312.

Mosconi, R. (1994). *Cointegrazione e modelli econometrici: teoria e applicazioni*. In *Ricerche quantitative per la politica economica 1993*, Contributi all'analisi economica, numero speciale. Banca d'Italia.

Ooms, M. (1994). *Empirical Vector Autoregressive Modeling*. Springer Verlag.

Pastorello, S. (2001). *Rischio e rendimento. Teoria finanziaria e applicazioni econometriche*. il Mulino.

Peracchi, F. (1995). *Econometria*. McGraw-Hill.

Phillips, P. C. (1986). *Understanding spurious regression in econometrics*. Journal of Econometrics, 33: 311–40.

Phillips, P. C. e Durlauf, S. (1986). *Multiple time series regression with integrated processes*. Review of Economic Studies, 53: 473–95.

Piccolo, D. (1990). *Introduzione all'analisi delle serie storiche*. NIS.

Sargent, T. J. (1987). *Macroeconomic Theory*. Academic Press, 2a ed.

Sims, C. A. (1980). *Macroeconomics and reality*. Econometrica, 48: 1–48.

Sims, C. A., Stock, J. e Watson, M. (1990). *Inference in linear time series models with some unit roots*. Econometrica, 58: 113–44.

Verbeek, M. (2000). *A guide to modern econometrics*. Wiley.

Zivot, E. (2008). *Practical issues in the analysis of univariate GARCH models*. `http://d.repec.org/n?u=RePEc:udb:wpaper:uwec-2008-03-fc&r=for`.

Indice analitico

Autocorrelazione, 5
    parziale, 42
Autocovarianza, 5

Beveridge-Nelson, scomposizione, 66
    in sistemi cointegrati, 130

Cholesky, scomposizione di, 100
Cointegrazione
    attrattore, 112
    definizione, 107
    vettori di, 108
Companion form, 84
Correlogramma, 7
Criteri di informazione, 42

DOLS, 130

ECM (Error Correction Mechanism), 110
Ergodicità, 4

Fattori comuni (COMFAC), 43
FM-OLS, 130
Funzione di risposta di impulso
    nei processi multivariati, 98
    nei processi univariati, 36

Granger
    causalità, 95
    teorema di rappresentazione, 116

Identificazione
    in senso Box-Jenkins, 42
    in senso econometrico, 42

Johansen, procedura di, 123

Nucleo deterministico
    in sistemi cointegrati, 119
    nei test di radice unitaria, 72

Operatore $\perp$, 117
Operatore ritardo, 13

Persistenza, 2
Previsore, 32
Processo stocastico
    ARMA stagionale, 31
    multivariato, 81
    TS (Trend-Stationary), 60
Processo stocastico, 2
    $I(1)$, 61
    AR (autoregressivo), 23
    ARCH, 138
    ARMA, 28
    ARMA moltiplicativo, 31
    DS (Difference-Stationary), 61
    EGARCH, 148
    GARCH, 140
    MA (a media mobile), 18
    VAR (autoregressivo vettoriale), 83

Radice unitaria, 24
    test
        KPSS, 73
Radice unitaria, 62
    test, 68
        Augmented Dickey-Fuller (ADF), 71
        Dickey-Fuller (DF), 70
        Phillips-Perron (PP), 72
Random Walk, 62
Regressione spuria, 76

Set informativo, 5
Shock strutturali, 99
Stazionarietà, 3
    dei processi AR, 28
    dei processi AR(1), 24
    dei processi ARMA, 29

dei processi VAR, 83

Test
LM-ARCH, 139
Trend
comuni, 118
deterministico, 60, 65
stocastico, 65

Verosimiglianza, 39
fattorizzazione sequenziale, 44

White noise, 16
multivariato, 82
Wold, teorema di rappresentazione di, 22