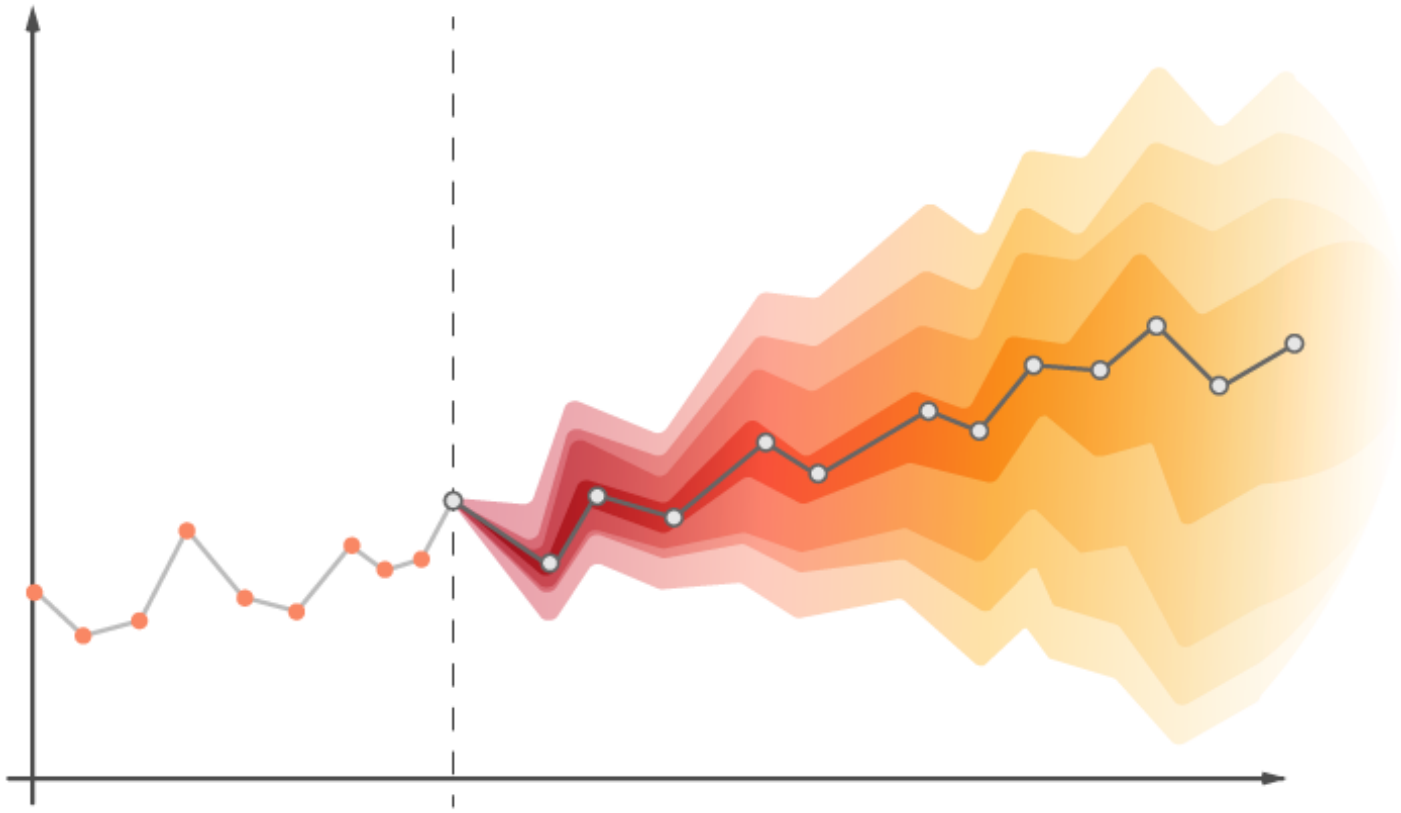
NOTES

Streaming Data Management and Time Series Analysis

2324-2-FDS01Q023



Vittorio Haardt
853268
vittoriohaardt@gmail.com

June 6, 2024

Abstract

The course illustrates methods and applications for managing, analyzing, and forecasting - possibly streaming - time series.

Besides data managing applications, our lessons cover linear (ARIMA, VAR, state-space/Kalman filter) and nonparametric (neural networks, support vector machine) methods.

The student who successfully follows this course will be able to manage streaming data and select, identify, and implement the time series model to fit the data and address the problem under analysis.

Contents

Chapter 1

Teoria della Previsione Statistica (Generale e Lineare)

1.1 Miglior Predittore

1.1.1 Introduzione

Prevedere un processo stocastico implica anticipare qualcosa che non può essere previsto con esattezza. Nella previsione di una variabile casuale continua, qualsiasi valore si cerchi di prevedere in questo processo, sarà errato con probabilità 1. L'unica certezza che abbiamo nella previsione (a meno che non si tratti di variabili categoriali) è che la probabilità che qualsiasi numero fornito come previsione corrisponda al valore reale che si otterrà è pari a 0, nel caso di variabili casuali continue.

Data una variabile casuale Y che è di interesse prevedere, e date le osservazioni $X_1, \ldots, X_n$. Si vuole prevedere Y date le osservazioni, ovvero vogliamo un modello che dia la distribuzione di Y date le osservazioni $X_1, \ldots, X_n$.

Il **predittore** è una funzione di $X_1, \ldots, X_n$, rappresentata da $\hat{p} = p(X_1, \ldots, X_n)$. Tale funzione p è progettata per minimizzare l'errore, ossia $\mathbb{E}(Y - p(X)) = 0$.

Qui, $p(X)$ denota il predittore. Va notato che non si tratta di uno stimatore in quanto Y non è ancora realizzato nella "storia", ma è una variabile casuale. Quando si prevede qualcosa di stocastico tramite un altro processo stocastico, è necessario stabilire il prezzo da pagare in caso di errore, ovvero definire una funzione di perdita o di costo, indicata come l.

La funzione $l(\cdot)$, chiamata **funzione di perdita** (loss function), mappa l'errore nel suo costo.

I costi possono differire dalle funzioni di previsione della perdita/costo; spesso queste ultime sono caratterizzate da funzioni asimmetriche. Pertanto, per realizzare una previsione ottimale è cruciale conoscere la funzione di costo, considerando anche l'asimmetria associata.

1.1.2 Miglior Predittore Generale

L'obiettivo è minimizzare la funzione p rispetto alla quale il rischio (il valore atteso dell'errore) sia ridotto al minimo. Il **predittore ottimale** $\hat{p}$ è la funzione che minimizza la previsione attesa della perdita:

$$\hat{p}(X_1, \ldots, X_n) = \arg\min_p \mathbb{E}(l(Y - p(X)))$$

Solitamente, se la funzione di perdita $l(\cdot)$ non viene specificata, si adotta la funzione di perdita quadratica:

$$l(x) = x^2$$

Se la funzione di perdita è quadratica e se entrambe le aspettative $\mathbb{E}(Y^2)$ e $\mathbb{E}(X_i)$ sono finiti, allora il predittore ottimale di Y basato su $X_1, \ldots, X_n$ è dato da:

$$\hat{p}(X_1, \ldots, X_n) = \mathbb{E}(Y|X_1, \ldots, X_n)$$

Tendenzialmente si minimizza il Mean Squared Error (MSE) perché ha buone proprietà matematiche e non vi è alcun motivo teorico per non farlo. Il MSE è simmetrico negli errori, trattando gli errori positivi e negativi nello stesso modo. Il MSE penalizza gli errori in modo quadratico all'aumentare della loro ampiezza: errori più grandi vengono penalizzati in modo significativamente maggiore. Un'alternativa al MSE potrebbe essere l'utilizzo del valore assoluto dell'errore, dove non vi è una penalizzazione quadratica all'aumentare dell'errore. Tuttavia, se il problema non specifica una funzione di perdita particolare, è del tutto ragionevole adottare il MSE. In breve, la scelta della funzione di perdita dipende dal contesto specifico e dalle preferenze dell'analista, ma in assenza di indicazioni specifiche, il MSE rimane una scelta comune e ragionevole.

L'errore di previsione è ortogonale (o incorrelato) con qualsiasi funzione $g(\cdot)$ applicata a X. Questo significa che l'aspettativa dell'errore di previsione moltiplicato per una qualsiasi funzione di X, $\mathbb{E}([Y - \mathbb{E}(Y|X)]g(X))$, è uguale a 0. In altre parole, l'errore di previsione è incorrelato con qualsiasi trasformazione della variabile X. Ciò implica che non è possibile migliorare le previsioni includendo qualsiasi funzione di X, poiché non si può trovare un modo per migliorare le previsioni.

Il miglior predittore per il valore assoluto è invece la mediana. Quando si utilizza una funzione di perdita asimmetrica, come ad esempio una funzione di perdita Pinball, la soluzione ottimale del problema è data dai percentili. In base alla scelta della funzione di perdita, il predittore ottimale può variare.

1.1.3 Miglior Predittore Lineare

Un predittore è definito come lineare se è costituito da funzioni lineari delle variabili in gioco, come ad esempio:

$$p(\cdot) \in \text{ funzioni lineari in } X_1, \ldots, X_n$$

Quando voglio trovare la funzione che minimizza l'errore nei miei dati X, posso limitare la ricerca solo alle funzioni lineari. Un esempio di predittore lineare è:

$$\mathbb{P}(X_1, \ldots, X_n) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

Se la funzione di perdita è quadratica e il predittore è espressamente lineare come sopra, allora il miglior predittore lineare sarà dato da:

$$\mathbb{P}(Y|X_1, \ldots, X_n) = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

Dove:

- μ_Y rappresenta il valore atteso di Y,
- Σ_{YX} è la matrice di covarianza tra Y e X,
- Σ_{XX}^{-1} è l'inverso della matrice di covarianza tra X e X,
- μ_X è il valore atteso di X.

Quando la distribuzione è gaussiana il miglior pedittore coincide con il miglior predittore lineare.

Per prevedere Y utilizzando una combinazione lineare delle variabili X, non è necessario conoscere l'intera distribuzione di Y dato X, ma solo i primi due momenti, ovvero la media e la covarianza. Questi possono essere stimati in modo coerente indipendentemente dalla distribuzione dei dati, utilizzando la media campionaria e la varianza/covarianza campionaria. Questo approccio rappresenta un metodo non parametrico di previsione, in quanto non è necessario stimare né un modello né la distribuzione, ma solo i primi due momenti con i momenti campionari.

Applicando il concetto ad una serie storica $Y_1, \ldots, Y_n$ si vuole prevedere il futuro Y_{n+1}. Anche in questo caso vale la stessa formula:

$$\mathbb{P}[Y_{n+1}|Y_1, \ldots, Y_n] = \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X)$$

Dove $X = Y_1, \ldots, Y_n$ e $Y = Y_{n+1}$. Dunque:

- $\mu_Y = \mathbb{E}[Y_t] = \mu_X$
- $\Sigma_{YX} = \mathbb{E}[Y_{n+1} - \mu] \begin{bmatrix} Y_n - \mu \\ Y_{n-1} - \mu \\ \vdots \\ Y_1 - \mu \end{bmatrix}^T = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_n \end{bmatrix}$
- $\Sigma_{XX} = \mathbb{E} \begin{bmatrix} Y_n - \mu \\ Y_{n-1} - \mu \\ \vdots \\ Y_1 - \mu \end{bmatrix} \begin{bmatrix} Y_n - \mu & \cdots & Y_1 - \mu \end{bmatrix} = \begin{bmatrix} \gamma_0 & \cdots & \gamma_{n-1} \\ \vdots & \ddots & \vdots \\ \gamma_{n-1} & \cdots & \gamma_0 \end{bmatrix}$

Chapter 2

Stazionarietà e Processi Integrati

2.1 Introduzione

2.1.1 Cosa è una Serie Storica

Una serie storica è una osservazione di una variabile o di un dato, tipicamente numerica, nel tempo (anche categoriali, ma più rare) indicato con X_t se discreto, oppure con $X(t)$ se continuo. Esistono serie storiche a tempo continuo (che vengono utilizzate in finanza), ma noi tratteremo quelle a tempo discreto.

Una serie storica è anche definibile come una mappatura tra il tempo discreto e i valori osservati in una variabile.

Il tempo nel nostro caso sarà **equispaziato**, ovvero la distanza tra t_i e t_{i+1} è costante per ogni i. Quindi tratteremo i dati con la stessa granularità, ovvero istanti di tempo omogenei, equispaziati nel tempo, per esempio con divisione trimestrale (4 osservazioni l'anno).

T è un insieme discreto, composto tipicamente da numeri interi (es. 1,2,3,4) ma anche da date (es. Gen, Feb, Mar).

I modelli a tempo continuo sono utili nel caso in cui si abbiano dati osservati in momenti non equispaziati nel tempo, come nel trading, dove si disctetizano questi modelli per usare i dati in modo concreto.

$$\{X_t\}_{t=1,\dots,n}$$

Se X_t è uno scalare la serie storica è detta **uni-variata**, se invece X_t è un vettore la serie storica è detta **multi-variata**. Esistono come detto anche serie storiche **discrete** o **con-**

tinue. Nel caso delle discrete ci sono diverse distribuzioni per approssimare id dati di questo tipo (es. Poisson). Nel caso i valori discreti siano molto grandi o per i valori continui esistono distribuzioni apposite. Esistono anche le serie storiche di dati **categoriali**, ma non verranno affrontate.

In un vettore di serie storiche ci sono informazioni comuni tra le varie serie storiche. Utilizzare il passato di un inseme di serie storiche per prevedere una serie a venire, fa si che tutte le informazioni possono risultare utili, quindi conviene prendere tutte le informazioni disponibili da un vettore di serie storiche. È anche possibile che ci siano dei movimenti comuni difficilmente osservabili, anche per questo motivo conviene utilizzare tutto il vettore di serie storiche.

2.1.2 Processo Stocastico

Un **processo stocastico** X_t è definito come una *sequenza di variabili casuali* indicizzate con t che rappresenta il tempo. Possono essere processi a tempo continuo o discreto.

Una serie storica $\{X_t\}_{t=1,\dots,n}$ *è la realizzazione finita di un processo stocastico.* È detta finita la realizzazione del processo è osservata su n tempi diversi.

A livello statistico la serie storica è un campione di dimensione 1. Per esempio per quanto riguarda il PIL italiano, non è possibile osservare tutti i possibili PIL che si sarebbero verificati se la storia fosse stata diversa. Questo porta ad un problema dal punto di vista statistico, che si risolve con una **supposizione di omogeneità**. Si possono fare previsioni solo se passato e futuro si comportano in maniera simile. Assumere forme di omogeneità ci permette di trattare il nostro unico campione di serie storica come se fosse uguale a campioni i-esimi.

2.2 Stazionarietà

2.2.1 Introduzione

La **stazionarietà** è l'assunzione di omogeneità nel tempo della serie storica. La stazionarietà può essere in senso debole o in senso forte.

Stazionarietà debole Un processo $X_{t=1,\dots,N}$ viene detto debolmente stazionario (a covarianza stazionaria) se:

- la sua media (valore atteso) è costante nel tempo.
  $\mathbb{E}(X_t) = \mu$
- la sua varianza esiste e non dipende dal tempo t.
  $Var(X_t) = \gamma_0 < \infty$
- la covarianza tra X_t e X_{t-1} dipende solamente da k (intero), e non da t. È una funzione (autocovarianza).
  $Cov(X_t, X_{t-k}) = \gamma_k$, con $k \in \mathbb{Z}$

γ_k è una **auto covariance function** (ACF), ovvero restituisce la covarianza di un'osservazione con la k-esima osservazione che la precede. Questa funzione ha le seguenti proprietà:

1. *Positivity of variance*: $\gamma(0) \geq 0$
2. *Cauchy-Schwartz inequality*: $|\gamma(h)| \leq \gamma(0)$
3. *Symmetry* $\gamma(h) = \gamma(-h)$
4. *Non-negative definiteness* $\sum_{i=1}^{m} \sum_{j=1}^{m} s_i \gamma(i-j) a_j \geq 0$

La più importante è la simmetria che ci dice $\gamma_k = \gamma_{-k}$ qjuesto è valido perché:

$$\gamma_k = Cov(X_t, X_{t-k}) = Cov(X_{t+k}, X_t) = Cov(X_t, X_{t+k}) = \gamma_{-k}$$

Se c'è relazione tra X_{t-k} e X_t allora c'è una relazione tra passato e presente. Se questa relazione resta costante nel tempo (non dipende da t) la si può stimare nei dati e quindi usarla per prevedere i dati futuri in modo lineare.

Il fatto che esista sempre la varianza non è da dare per scontato, qualora ci siano distribuzioni con valori estremi è pericoloso avere la varianza.

Stazionarietà forte Un processo stocastico è detto fortemente stazionario se qualunque k ha la stessa distribuzione di se stesso traslato nel tempo, ovvero è invariato nella traslazione temporale.

La stazionarietà in senso forte coinvolge tutta la distribuzione della serie storica, non solo i primi due momenti (distribuzione invariante a traslazione temporale in tutta la serie storica).

Un processo X_t è detto fortemente stazionario se per ogni $t_1, t_2, \ldots, t_n$ per ogni h e per ogni k.

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_k}) \stackrel{d}{=} (X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$$

Dove $\stackrel{d}{=}$ vuol dire che è equi-distribuito.

Esistono trasformazioni per modellare la serie storica che permettono di rendere stazionaria una serie storica che non lo è.

Modello Lineare

In un modello lineare il futuro dipende linearmente dal passato. Nel modello lineare prevediamo il futuro tramite funzioni del passato.

La formula di relazione tra futuro e passato necessaria per il modello lineare è la seguente:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{Cov(X_t, X_{t-k})}{\sqrt{Var(X_t)Var(X_{t-k})}}$$

La finalità dei modelli lineari di serie storiche è quella di modellare la funzione di autocovarianza/autocorrelazione facendolo in maniera efficiente, ovvero facendo si che dipenda da un numero "piccolo" di parametri.

Operatori Ritardo

Nel seguito si farà frequentemente uso dell'operatore ritardo B (dal l'Inglese backward, spesso anche indicato con la L di lag o la D di delay), che applicato ad una serie storica $\{X_t\}$ ha l'effetto, appunto, di ritardarla di un periodo:

$$B\{X_t\} = \{X_{t-1}\}$$

o più sinteticamente

$$BX_t = X_{t-1}$$

Si definisca prodotto dell'operatore B con se stesso, l'applicazione sequenziale dell'operatore

$$BBX_t = BX_{t-1} = X_{t-2}$$

e si usi la naturale simbologia B^k, con k intero positivo, per intendere n applicazioni sequenziali dell'operatore:

$$B^k X_t = B^{k-1}BX_t = \ B^{k-1}X_{t-1} = \ldots = X_{t-k}$$

Si ponga, inoltre, $B^0 = I$, dove I è l'operatore identità e si indichi con B^{-1} l'inverso dell'operatore B, cioè l'operatore anticipo:

$$BB^{-1}X_t = BX_{t+1} = X_t$$

da cui, per definizione,

$$BB^{-1} = I$$

Trasformazioni per la Stazionarietà

Esistono trasformazioni che permettono di rendere certe serie storiche stazionarie sia rispetto alla *varianza* che rispetto alla *media*. Una delle trasformazioni più comuni è quella **logaritmica**.

Il logaritmo, applicabile solo se le serie storiche sono positive, viene usato per risolvere la non stazionarietà della varianza quando essa è dipendente dalla media (cresce la variabilità quando cresce il livello della serie storica). Questo ha anche un vantaggio interpretativo, ovvero che tutti *gli incrementi assoluti diventano incrementi relativi*, si dice che il logaritmo relativizza i movimenti.

Gli operatori **differenza** e **differenza stagionale** possono essere definiti per mezzo degli operatori ritardo, rispettivamente come:

$$\Delta = (I - B)$$

$$\Delta_s = (I - B^s)$$

dove s è il numero di "stagioni" in un anno ($s = 12$ per serie mensili e $s = 4$ per serie trimestrali).

Quando applicata alla serie temporale, otteniamo:

$$\Delta X_t = X_t - X_{t-1}$$

che può essere derivata come:

$$\Delta X_t = (I - B)X_t = X_t - X_{t-1}$$

La differenza può anche essere di secondo ordine:

$$\Delta^2 = (I - B)^2 = (I - 2B + B^2)$$

dove vale che:

$$\Delta^2 X_t = \Delta\Delta X_t = \Delta(X_t - X_{t-1}) = X_t - X_{t-1} - X_{t-1} + X_{t-2} = X_t - 2X_{t-1} + X_{t-2}$$

Quando la serie non sembra neanche stazionaria in varianza, ma vi è una relazione funzionale tra media e varianza, allora la si può rendere stazionaria per mezzo di una trasformazione. Le trasformazioni più comuni sono quelle della famiglia di Box e Cox,

$$y_t(\lambda) = \begin{cases} \frac{x_t^{\lambda-1}}{\lambda} & \text{per} \quad \lambda \neq 0 \\ log(x_t) & \text{per} \quad \lambda = 0 \end{cases}$$

Se la serie X_t ha dei valori non positivi, prima di applicare la trasformazione si somma una costante arbitraria sufficientemente grade da rendere tutta la serie positiva. La trasformazione logaritmica è quella più comune in pratica perché rende stazionarie in varianza quelle serie per cui la deviazione standard cresce proporzionalmente alla media. Le trasformazioni di Box e Cox rendono stazionarie in varianza serie in cui esiste una relazione tra media e variabilità, se, come nelle serie stazionarie in media, la media non varia, l'effetto sulla varianza sarà nullo.

2.2.2 Stima ACF

Auto Correlation Function

Una volta applicate le trasfromazioni necessarie per rendere stazionaria la serie storica, tutta l'informazione viene fornita dall'ACF (**auto correlation function**).

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

Se non c'è correlazione con il passato non è possibile prevederla linearmente, ma è possibile provare con funzioni non lineari.

Plottano la stima di ogni ρ_k (i ritardi) viene costruito *l'autorcorrelogramma*, che fornisce informazioni su quanto il passato è in correlazione con il presente, e consente di prevedere il futuro utilizzano funzioni lineari (o combinazioni di funzioni lineari).

Ha senso calcolare ACF solamente dopo aver reso la serie storica stazionaria, se la serie storica è stazionaria prima o poi il suo autocorrelogramma deve tendere a 0. Un processo stazionario è un processo che prima o poi si deve scordare di quello che è successo nel passato, in altre parole non ci può essere nessuno shock passato che causi un cambiamento permanente permanentemente. Un processo stazionario torna sempre alla media, quindi ACF deve tendere a 0 per k crescente. La correlazione a ritardo 0 è uguale ad 1 per definizione.

Le bande di confidenza sono costruite sotto l'ipotesi che la serie storica sia incorrelata con il passato, ovvero una serie storica **white noise**. Se i valori sono sotto o sopra alle bande, si dice che sono significativamente diversi da zero.

In conclusione ACF sintetizza la memoria del presente con il passato che si può posso supporre essere continuo anche nel futuro.

Passaggi per la Stazionarietà

È di fondamentale importanza affrontare innanzitutto la non stazionarietà in varianza e solo successivamente quella in media mediante *differenziazioni semplici* o *stagionali*. Ciò perché

quando si lavora con incrementi logaritmici, tutti gli incrementi diventano relativi. Se si procede immediatamente con la differenziazione, si rischia di compromettere la relazione tra la media e la volatilità della serie storica.

Per l'utilizzo di modelli ARMA, è essenziale che la serie storica sia stazionaria sia in termini di varianza che di media. Gli ARMA modellano serie stazionarie, le quali presentano una memoria limitata nel tempo e possono essere interpretate come processi che trasformano rumore bianco casuale in una serie con una memoria di breve termine, poiché gli impatti degli shock vengono gradualmente dimenticati.

Gli shock devono avere una memoria limitata nei processi stazionari; in altre parole, la memoria della serie storica del suo passato deve decadere verso zero affinché il processo sia stazionario. Se ciò non accade, non si avrebbe un processo stazionario: ogni volta che uno shock entra in un processo stazionario, deve essere gradualmente dimenticato nel tempo.

I passaggi da fare sono i seguenti:

1. se non stazionaria in varianza $\rightarrow$ attuare una trasformazione (famiglia Box Cox)
2. se stagionale $\rightarrow$ applicare la differenza stagionale $\Delta_s = (1 - B^s)$

La stagionalità potrebbe essere giornaliera, settimanale, mensile, annuale, o a stagionalità multiple. Con serie giornaliere diventa più complicato.

È possibile usare dei test per controllare la stazionarietà, anche se dal punto di vista pratico sono poco utilizzati. Abbiamo il **test di Dickey-Fuller** dove l'ipotesi nulla è che il processo sia stazionario, e il **test KPSS** dove l'ipotesi nulla è invece che il processo non sia stazionario.

Qualora ci fosse ancora qualche forma di trend dopo aver applicato la differenza stagionale allora è necessario applicare la forma semplice del trend Δ.

Una volta resa stazionaria una serie storica si analizza auto-covarianza e/o auto-correlazione, perché analizzando modelli lineari interessa la correlazione.

Funzione di Auto-Covarianza

In un contesto di serie storica, come ad esempio quando si considera la stazionarietà debole, incontriamo le seguenti proprietà:

$$\mathbb{E}(X_t) = \mu$$

$$Var(X_t) = \gamma_0$$

$$Cov(X_t, X_{t-k}) = \gamma_k$$

Il nostro interesse è rivolto alla stima di queste grandezze, avendo già a disposizione un valore particolare per γ_0.

Nel caso di un processo stazionario, la covarianza non dovrebbe dipendere dal tempo t, ma solo dalla distanza k tra i due momenti. Pertanto, per stimare γ_k, non è necessario tener conto di t. La stima di γ_k avviene tramite la covarianza stazionaria:

$$\hat{\gamma_k} = \frac{1}{n} \sum_{t=k+1}^{n} (X_t - \overline{X}_n)(X_{t-k} - \overline{X}_n)$$

Dove $\overline{X}_n$ rappresenta la media campionaria:

$$\overline{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t$$

L'intervallo di somma parte da $k+1$ per garantire la considerazione della covarianza rispetto al passato.

Questa procedura si basa sulla fondamentale proprietà che la matrice di covarianza sia definita positiva. Se la matrice di varianza non è definita positiva, è possibile ottenere previsioni o combinazioni lineari con varianza negativa, il che non ha senso. La stima delle **autocorrelazioni** ρ è equivalente alla stima delle **autocovarianze** con ritardo k, o alla stima della varianza. Questo concetto esprime la dipendenza che un processo ha dal suo passato, indicando la possibilità di fare previsioni future basate su tale dipendenza.

In condizioni normali, la stima $\hat{\rho}$ si ottiene come:

$$\hat{\rho} = \frac{\hat{\gamma_k}}{\hat{\gamma_0}}$$

Dove:

$$\overline{X}_n \xrightarrow{d} \mu$$

$$\hat{\gamma_k} \xrightarrow{d} \gamma_k$$

Tutti i modelli di serie storica mirano a fornire una rappresentazione parsimoniosa della covarianza della popolazione utilizzando un numero limitato di parametri.

I grafici di autocorrelazione sono utilizzati per identificare il possibile processo ARMA che ha generato l'autocorrelogramma. Tuttavia, è importante notare che il numero di ritardi considerati non dovrebbe superare un terzo della lunghezza della serie storica, poiché altrimenti l'analisi risulterebbe poco affidabile.

Di solito, si rappresenta graficamente l'autocorrelazione rispetto al ritardo (lag), il che fornisce informazioni sulla quantità di memoria lineare o dipendenza dal passato presente nella serie temporale. Se l'autocorrelazione è completamente nulla, il processo in questione è un processo di white noise (WN).

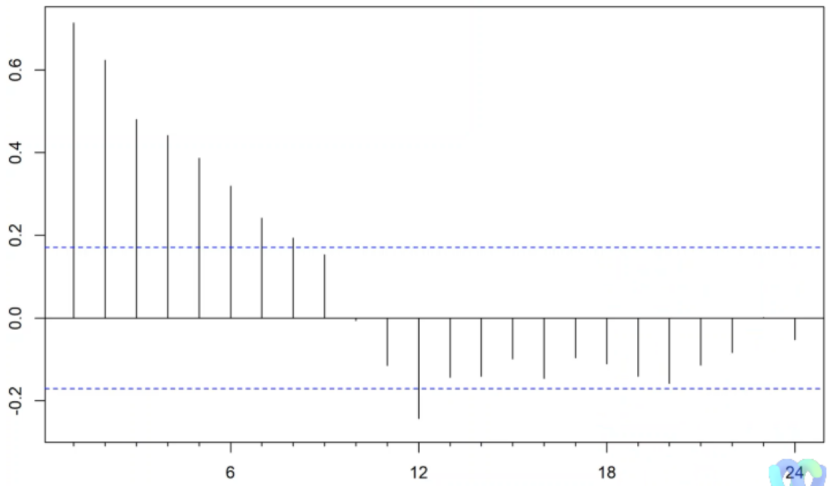


Figure 2.1: In questo caso il correlogramma ci dice che la serie storica ha buona memoria di se stessa soprattutto nel breve periodo.

White Noise

Processo stocastico stazionario a media zero, varianza costante e autocovarianza sempre nulla (non ha memoria lineare), imprevedibile linearmente, è detto **white noise** ed è indicato con ε_t. È importante perché si usa per costruire processi ARMA.

- $\mathbb{E}(\varepsilon_t) = 0$
- $Var(\varepsilon_t) = \sigma^2$
- $Cov(\varepsilon_t, \varepsilon_{t=k}) = 0$

2.3 Processi Integrati

2.3.1 Definizione

Quando si tratta di serie storiche, gli **integrated process** (processi integrati) giocano un ruolo cruciale nella comprensione e nell'analisi dei dati. Un processo X_t è considerato stazionario, indicato con $X_t \sim I(0)$, se non mostra alcuna tendenza sistematica nel tempo. D'altra parte, se le variazioni di X_t sono stazionarie solo dopo la differenziazione una volta, ossia $\Delta X_t \sim I(0)$, allora diciamo che il processo è **integrato di ordine 1**, $X_t \sim I(1)$. Se il processo richiede più di una differenziazione per diventare stazionario, ad esempio se $\Delta^{d-1} X_t$ è non stazionario ma $\Delta^d X_t$ è stazionario, allora diciamo che $X_t \sim I(d)$.

L'integrazione di un processo X_t può essere rappresentata come segue:

$$Y_t = Y_{t-1} + X_t$$

Se X_t è integrato all'ordine d, $X_t \sim I(d)$, allora $Y_t \sim I(d+1)$. Chiaramente, invertendo questo processo, si ottiene il processo originale.

Una volta che il processo è stato trasformato in una serie stazionaria tramite integrazione, possiamo utilizzare modelli ARIMA (AutoRegressive Integrated Moving Average) per l'analisi e la previsione dei dati storici. Questi modelli sono approssimativi ma potenti e sono ampiamente utilizzati nell'analisi delle serie storiche.

2.3.2 Random Walk

Un processo integrato può essere visualizzato come una random walk, in cui il valore al tempo corrente è uguale al valore al tempo precedente più un errore casuale. Matematicamente, possiamo rappresentare questo come:

$$X_t = X_{t-1} + \varepsilon_t$$

dove $\varepsilon_t \sim WN(\sigma^2)$, rappresentando il white noise. È importante notare che un processo di questo tipo è non stazionario e non ha una media definita. Tuttavia, se si parte da un valore iniziale x_0, la media del processo sarà x_0.

Si può scrivere come una somma parziale, assumendo che ci sia un punto di partenza x_0 che è un numero, allora si ottiene:

$$X_t = x_0 \sum_{j=1}^{t} \varepsilon_j$$

La media del numero è il numero stesso (x_0).

$$\mathbb{E}X_t = X_0 + 0 = x_0$$

La media del random walk non viola la condizione sulla media definita nella definizione di stazionarietà, quindi bisogna calcolare la varianza

La varianza di questo processo aumenta linearmente nel tempo, come mostrato da:

$$Var(X_t) = t \cdot \sigma^2$$

Mentre la covarianza tra due punti nel processo, assumendo $x_0 = 0$, è data da:

$$Cov(X_t, X_s) = \mathbb{E}(X_t \cdot X_s) = \gamma(t,s) = \mathbb{E}(\sum_{j=1}^{t} \varepsilon_J, \sum_{j=1}^{s} \varepsilon_j) = \min(t,s)\sigma^2$$

Vediamo, dato che la varianza dipende dal tempo, che X_t è non stazionario e richiede integrazione, $\Delta X_t = X_t - X_{t-1} = \varepsilon_t \sim I(0)$ quindi $X_t \sim I(1)$.

Random walk è un tipo di processo integrato del white noise (WN). Nei contesti reali, i processi di integrazione non superano di solito un ordine di 2.

Per quanto riguarda le previsioni condizionate sul passato, la migliore previsione per la prossima osservazione è data dalla media condizionata:

$$\hat{X}_{n+1|n} = \mathbb{E}[X_{n+1}|X_n, X_{n-1}, \ldots, X_1]$$

Applicando la definizione, otteniamo:

$$= \mathbb{E}[X_{n+1} + \varepsilon_{n+1}|X_n, X_{n-1}, \ldots, X_1]$$

Assumendo che ε_{n+1} sia indipendente e identicamente distribuito, rappresenta un nuovo shock non conosciuto:

$$= X_n$$

Quindi, la migliore previsione per il valore futuro in una random walk è semplicemente il valore attuale. Quindi, vale:

$$\hat{X}_{n+k|n} = X_n$$

Tuttavia, possiamo anche prevedere la varianza condizionata:

$$Var(X_{n+}|X_n) = K\sigma^2$$

Se ad esempio assumiamo una random walk gaussiana, la distribuzione condizionata sarà anch'essa normale:

$$X_{n+k}|X_n \sim N(X_n, K\sigma^2)$$

Possiamo quindi calcolare gli intervalli di confidenza per le previsioni, ad esempio al 95 percento:

$$(X_n - 1.96\sqrt{K\sigma^2}, (X_n - 1.96\sqrt{K\sigma^2})$$

Riassumendo, per fare previsioni, spesso ci interessa stimare il prossimo valore condizionato sull'andamento passato. In una random walk, la migliore previsione per il valore futuro è semplicemente il valore corrente. La varianza condizionata può essere calcolata anche in base alla varianza dei shock casuali. In particolare, se assumiamo una distribuzione normale per la random walk, possiamo calcolare gli intervalli di confidenza per le previsioni future.

Chapter 3

Modelli ARIMA

3.1 Autoregressive Processes (AR)

3.1.1 AR(p)

Introduzione

Un processo autoregressivo, noto come $AR(p)$, è un modello nel quale il valore attuale della serie temporale dipende linearmente da p suoi valori precedenti, oltre a un termine di errore. Matematicamente, possiamo esprimerlo come:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

dove Y_t è il valore corrente della serie temporale, $\phi_1, \phi_2, \ldots, \phi_p$ sono i coefficienti autoregressivi, $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ sono i valori precedenti della serie, e ε_t è un termine di errore che segue $\varepsilon_t \sim WN(\sigma^2)$.

Come è possibile intuire una Random Walk altro non è che un processo autoregressivo di ordine 1, $RW = AR(1)$.

Applicando l'operatore di ritardo B, possiamo riscrivere l'equazione come:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

dove $\phi_p(B)$ è il polinomio autoregressivo in B. Pertanto, l'equazione diventa:

$$\phi_p(B) Y_t = \varepsilon_t$$

che può essere scritta come:

$$(1 - \phi_1 B - \ldots - \phi_p B^p) Y_t = \varepsilon_t$$

Da qui, possiamo ottenere nuovamente la forma originale del processo autoregressivo $AR(p)$:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

È importante notare che i processi AR potrebbero non essere stazionari. La condizione di non stazionarietà è associata al polinomio $(1 - \phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p) Y_t$.

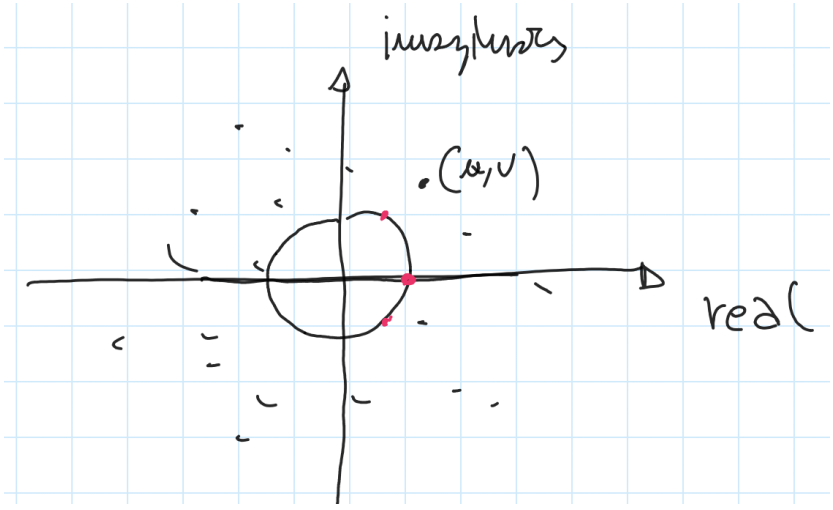
Inoltre, un processo $AR(p)$ ha soluzioni casuali e stazionarie se i valori degli zeri delle equazioni caratteristiche soddisfano $|z_i| > 1$:

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

Questo implica che le radici devono trovarsi al di fuori del cerchio unitario sulla piana complessa. In altre parole, se z_i è una soluzione, allora $|z_i| > 1$ per $i = 1, \ldots, p$. Vale la seguente relazione:

$$|z_i| > 1 \Longleftrightarrow AR(p) \sim I(0)$$

L'equazione di un processo autoregressivo $AR(p)$ definisce Y_t come dipendente da se stesso ritardato, oltre a un termine di rumore bianco. È legittimo chiedersi se esista un processo stocastico causale stazionario, in cui il presente è dipendente dal passato, rispettando questa equazione. Le soluzioni causali escludono scenari in cui il presente dipende dal futuro. Il processo in questione esiste se e solo se tutte le sue radici hanno modulo maggiore di 1. Le radici, essendo immaginarie, possono essere rappresentate come $z = u + iv$, dove la circonferenza unitaria (raggio 1) rappresenta la frontiera, con le soluzioni al di fuori di essa.



Se nel processo $AR(p)$ prendiamo i coefficienti, risolviamo l'equazione di grado p e troviamo che una radice è uguale a 1 (radice unitaria), allora il processo non è stazionario, ma integrato. Tuttavia, se differenziamo, possiamo renderlo stazionario. Se ci sono 2 radici/soluzioni con

valore 1, allora il processo è integrato di ordine 2. Le radici unitarie, in particolare, indicano che almeno una soluzione è uguale a 1. Il test di **Dickey-Fuller**, comunemente utilizzato, verifica la presenza di radici unitarie, determinando se esista una radice uguale a 1.

Inoltre, possono esistere radici stagionali distribuite lungo la circonferenza unitaria, che possono essere rilevanti soprattutto in contesti di serie temporali stagionali.

Verificare Radici Unitarie

Un processo autoregressivo di ordine uno $AR(1)$ è definito semplicemente da:

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

Dove ε_t è il termine di errore e ϕ è il coefficiente autoregressivo. Sappiamo che per l'equazione caratteristica, otteniamo:

$$1 - \phi z = 0$$

Da cui risolvendo per z, otteniamo:

$$z = \frac{1}{\phi}$$

E poiché dobbiamo assicurarci che $|z| > 1$, abbiamo:

$$|\frac{1}{\phi}| > 1$$

Se $z = 1$ è una soluzione dell'equazione caratteristica questo significa che $1 - \phi = 0 \rightarrow \phi = 0$ quindi una random walk, che è non stazionaria ma può essere resa stazionaria.

Questo implica che se $|\phi| < 1$, la soluzione è reale, quindi il processo è stazionario. In un processo $AR(1)$:

- $|\phi| < 1 \iff$ stazionario
- $\phi = 1$ possiede una radice unitaria, ovvero è integrato
- $|\phi| > 1$ esplosivo

Questa regola non va generalizzata, vale solo per processi $AR(1)$.

Per il processo $AR(2)$ definito da:

$$Y_t = 1.5Y_{t-1} - 0.7Y_{t-2} + \varepsilon_t$$

l'equazione caratteristica associata è:

$$1 - 1.5z + 0.7z^2 = 0$$

Risolvendo questa equazione, otteniamo le soluzioni complesse:

$$z_{1/2} = 1.07143 \pm 0.529728i$$

Calcolando il modulo di queste soluzioni, otteniamo $|z| = 1.19$. Poiché il modulo è maggiore di 1, possiamo affermare che il processo è stazionario.

Un modo semplice per verificare se c'è una radice unitaria è *sommare tutti i coefficienti ϕ, se la somma è uguale a 1, allora c'è una radice unitaria presente*, e quindi il processo è non stazionario.

Se $z = 1$ è una soluzione dell'equazione caratteristica:

$$1 - \phi_1 z - \ldots - \phi_p z^p = 0$$

allora significa che la somma dei coefficienti autoregressivi è uguale a 1:

$$1 = \phi_1 + \phi_2 + \ldots + \phi_p$$

Questo indica che il processo è integrato e che possiamo applicare l'operatore di differenziazione per trasformarlo in un processo stazionario. La presenza di una radice unitaria suggerisce che il processo è non stazionario e richiede quindi la differenziazione per ottenere la stazionarietà.

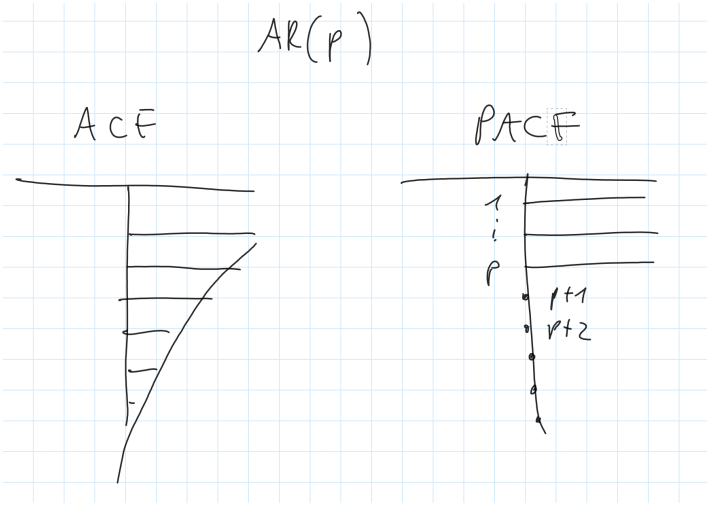


Figure 3.1: ACF e PACF di $AR(p)$

AR con costante

Se consideriamo un processo con costante c definito come:

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

Se questo processo è stazionario, la media del processo può essere calcolata come:

$$\mathbb{E}(Y_t) = \frac{c}{1 - \phi_1 - \ldots - \phi_p}$$

Per dimostrarlo, possiamo utilizzare la definizione di media e l'assunzione che il processo sia stazionario:

$$\mu = \mathbb{E}(Y_t) = c + \phi_1 \mathbb{E}(Y_{t-1}) + \ldots + \phi_p \mathbb{E}(Y_{t-p}) + \mathbb{E}(\varepsilon_t)$$

Dal momento che per definizione un processo stazionario ha una media costante, possiamo scrivere:

$$\mathbb{E}(Y_t) = \mathbb{E}(Y_{t-1})$$

Da cui otteniamo:

$$\mu = c + \phi_1 \mu + \ldots + \phi_p \mu$$

Poiché $\mathbb{E}(\varepsilon_t) = 0$, possiamo semplificare ulteriormente:

$$\mu(1 - \phi_1 - \ldots - \phi_p) = c$$

E quindi otteniamo la formula iniziale.

Se la costante è zero, allora la media del processo sarà zero. Questo implica che la media di un processo AR senza costante è zero.

3.1.2 Partial Auto-Correlation Function

PACF

La funzione di autocorrelazione parziale (PACF) è definita come la correlazione tra Y_t e Y_{t-k} una volta eliminato l'effetto delle osservazioni intermedie. In un processo stazionario, questa è semplicemente la correlazione tra Y_t e Y_{t-k}.

$$\rho_k = \mathrm{Cor}(Y_t, Y_{t-k})$$

La PACF è definita anche per processi non stazionari, ma in quel caso ha due indici poiché dipende anche dal tempo. Tuttavia, per ora concentriamoci sulla PACF del processo stazionario, poiché è possibile stimarla con accuratezza, a differenza di quella dei processi non stazionari.

La partial auto-correlation function rappresenta la correlazione tra due variabili una volta eliminato l'effetto delle variabili intermedie. Se consideriamo un processo stazionario, possiamo notare che se Y_{t-2} influisce su Y_{t-1} nello stesso modo in cui Y_{t-1} influenza Y_t, allora la PACF diventa zero dopo aver eliminato l'effetto delle variabili intermedie.

In un processo AR(1), ad esempio, dove ogni osservazione dipende solo dalla precedente più uno shock, eliminando l'effetto di Y_{t-1} (tenendo il suo valore fisso), non rimane correlazione tra Y_t e Y_{t-2} poiché non c'è più un effetto diretto.

La prima PACF, α_1 (tra Y_{t-2} e Y_{t-1}), è per definizione uguale alla correlazione, poiché non ci sono variabili intermedie. Ma in questo caso particolare, α_2 (tra Y_{t-1} e Y_t) è zero poiché la correlazione viene interrotta quando si tengono fissi alcuni valori.

Definiamo ora la PACF come:

$$\alpha_k = \frac{\mathbb{E}\left[(Y_t - \mathbb{P}[Y_t|Y_{t-1}, \ldots, Y_{t-k+1}]) \cdot (Y_{t-k} - \mathbb{P}[Y_{t-k}|Y_{t-1}, \ldots, Y_{t-k+1}])\right]}{\gamma_0}$$

Quindi, si prende la miglior predizione possibile di Y_t dato il passato fino a Y_{t-k+1} e si fa lo stesso con Y_{t-k}, anche se non è strettamente necessario. Si divide per la varianza e si ottiene la PACF, che ci indica se c'è un effetto diretto tra Y_t e Y_{t-k}.

ACF e PACF

L'autocorrelation function (ACF) e la partial autocorrelation function (PACF) hanno scopi leggermente diversi e forniscono informazioni complementari.

L'ACF misura la correlazione lineare tra un'osservazione e le sue precedenti a diversi ritardi temporali. Serve per aiutare ad identificare la struttura di dipendenza temporale nei dati. Ad esempio, può indicare se vi è una stagionalità o una tendenza nei dati. L'ACF viene calcolata senza considerare l'effetto di altre osservazioni intermedie, quindi riflette la correlazione diretta tra le osservazioni a diversi ritardi. È utile per determinare il numero di ritardi da includere in un modello autoregressivo (AR) o un modello di media mobile (MA).

La PACF misura la correlazione tra due osservazioni una volta eliminato l'effetto delle altre osservazioni intermedie. Questo fornisce informazioni sulla correlazione diretta tra due osservazioni a un determinato ritardo, controllando l'effetto di tutte le altre osservazioni intermedie. Serve ad aiutare ad identificare i ritardi significativi da includere in un modello AR. Quando la PACF di un certo ritardo diventa non significativa, suggerisce che tutti i ritardi precedenti sono stati già considerati nel modello. La PACF viene comunemente utilizzata per identificare l'ordine di un modello AR.

La PAFC serve per trovare l'ordine p dell'AR(p). Quando ci sono soluzioni complesse osserviamo una qualche forma di oscillazione. Per un AR di ordine p l'ACF rientra a velocità geometrica verso zero con pattern, la PAFC esce fino al ritardo p-esimo poi va a 0.

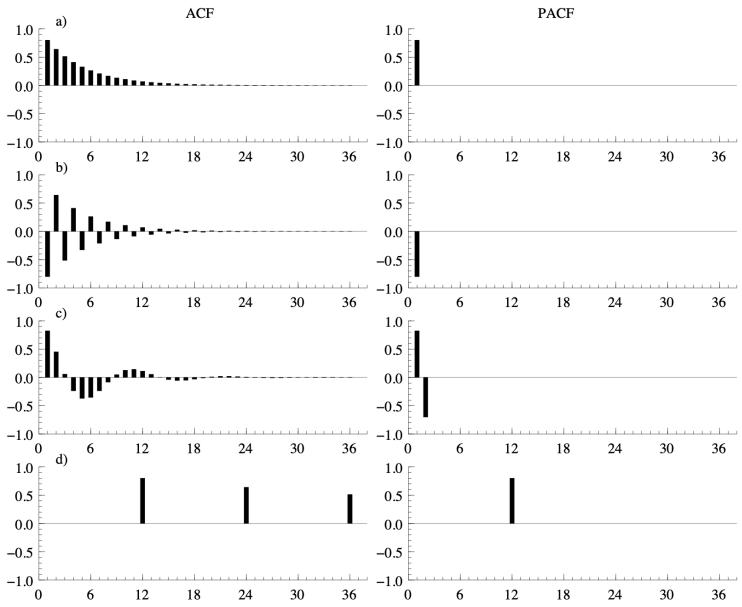


Figure 3.2: Funzioni di autocorrelazione (ACF) e autocorrelazione parziale (PACF) di quattro processi autoregressivi: (a) $AR(1)$ con $\phi > 0$, (b) $AR(1)$ con $\phi < 0$, (c) $AR(2)$ con radici complesse, (d) $SAR(1)_{12}$ con $\Phi > 0$.

3.1.3 Dickey-Fuller Test

DF Test

Il test di Dickey-Fuller è utilizzato per testare la presenza di radici unitarie in un processo, in altre parole, verifica se il processo è stazionario o meno. Consideriamo un processo generato da $Y_t = C + \phi Y_{t-1} + \varepsilon_t$, dove C è una costante. Il test Dickey-Fuller è formulato come:

- $H_0 : \phi = 1$
- $H_1 : \phi < 1$

Se $\phi = 1$, il processo è una random walk (RW), mentre se $\phi < 1$, il processo è stazionario, con ϕ che rappresenta l'effetto di ritorno al valore medio. Per stimare il parametro $\hat{\phi}$, possiamo eseguire una regressione OLS e quindi calcolare la statistica test:

$$\frac{\hat{\phi} - 1}{se(\hat{\phi})}$$

Questa statistica segue una distribuzione specifica chiamata distribuzione DF.

Per questo test l'ipotesi nulla ci dice che il processo è o una RW o una RW con drift. In generale vale che se è presente una costante e il processo è stazionario, il processo ha comunque una media ma la media non è 0.

Augmented DF Test

Il test di Dickey-Fuller è stato generalizzato per i processi AR, diventando il test aumentato di Dickey-Fuller (ADF). L'ADF test si applica a processi del tipo:

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

L'idea è riscrivere il processo in modo da avere un unico parametro, direttamente dalla stima. Di solito, l'equazione di Dickey-Fuller è riscritta come:

$$\Delta Y_t = c + \rho Y_{t-1} + \gamma_{1t-1} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

Dove ρ non è l'autocorrelazione, ma rappresenta un termine di ritorno al valore medio.

Raggruppando le varie differenze, otteniamo:

$$Y_t - Y_{t-1} = \rho Y_{t-1} + \gamma_{1t-1} - \gamma_{1t-2} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1-\gamma_p \Delta Y_{t-p} + \varepsilon_t}$$

Quindi, l'ADF test si basa su:

$$Y_t = (1 + \rho + \gamma_1) Y_{t-1} + (-\gamma_1 + \gamma_2) Y_{t-2} + \ldots$$

Ponendo questa equazione uguale alla prima e risolvendo per ρ, si ottiene:

$$\rho = \phi_1 + \ldots + \phi_p - 1$$

Quindi, l'ADF test si basa su:

- $H_0 : \rho = 0$ (radice unitaria)
- $H_1 : \rho < 0$ (stazionarietà)

La statistica del test è calcolata come:

$$\frac{\hat{\rho}}{se(\hat{\rho})}$$

Anche in questo caso, questa statistica segue una distribuzione DF. Se $\hat{\rho}$ è minore di zero, possiamo concludere che il processo è stazionario.

Per visualizzare questo in modo più semplice, pensiamo ad un processo $AR(3)$ come il seguente.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t$$

Si ha quindi:

$$\Delta Y_t = \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \varepsilon_t$$

Che viene sviluppato come:

$$Y_t = Y_{t-1} + \rho Y_{t-1} + \gamma_1 Y_{t-1} - \gamma_1 Y_{t-2} + \gamma_2 Y_{t-2} - \gamma_2 Y_{t-3} + \varepsilon_t$$

Raggruppando si ottiene:

$$Y_t = (1 + \rho + \gamma_1) Y_{t-1} + (-\gamma_1 + \gamma_2) Y_{t-2} - \gamma_2 Y_{t-3} + \varepsilon_t$$

Uguagliando il processo di partenza con il raggruppamento come di seguito:

$$\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \varepsilon_t = (1 + \rho + \gamma_1) Y_{t-1} + (-\gamma_1 + \gamma_2) Y_{t-2} - \gamma_2 Y_{t-3} + \varepsilon_t$$

Si arriva ad ottenere:

$$\phi_3 = -\gamma_3$$
$$\phi_2 = -\gamma_1 + -\gamma_2 = -\gamma_1 - \phi_3$$
$$\phi_1 = 1 + \rho + \gamma_1$$

Risolvendo poi i passaggi per ρ:

$$\rho = \phi_1 + \phi_2 + \phi_3 - 1$$

Vediamo quindi che ρ raggruppa la somma degli coefficienti auto regressivi meno 1, quindi se $\rho = 1$ allora abbiamo una radice unitaria. Se è minore di 1 abbiamo stazionarietà.

3.1.4 Seasonal AR

Se desideriamo modellare una serie storica in cui la memoria è influenzata principalmente da intervalli stagionali, possiamo utilizzare un modello autoregressivo stagionale, chiamato $SAR(P)$, dove s è la componente stagionale. Questo modello è definito come:

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \ldots + \Phi_P Y_{t-Ps} + \varepsilon_t$$

Funziona in modo simile ai processi $AR(p)$, ma opera su intervalli multipli di s. Matematicamente, possiamo esprimerlo come:

$$(1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}) Y_t = \varepsilon_t$$

Il correlogramma di un processo $SAR(P)$ è simile a quello di un $AR(p)$, ma mostra correlazioni solo nei multipli di s. Ad esempio, se $s = 4$, l'ACF tornerà a zero a multipli di 4. Inoltre, la PACF tornerà a zero alla componente stagionale.

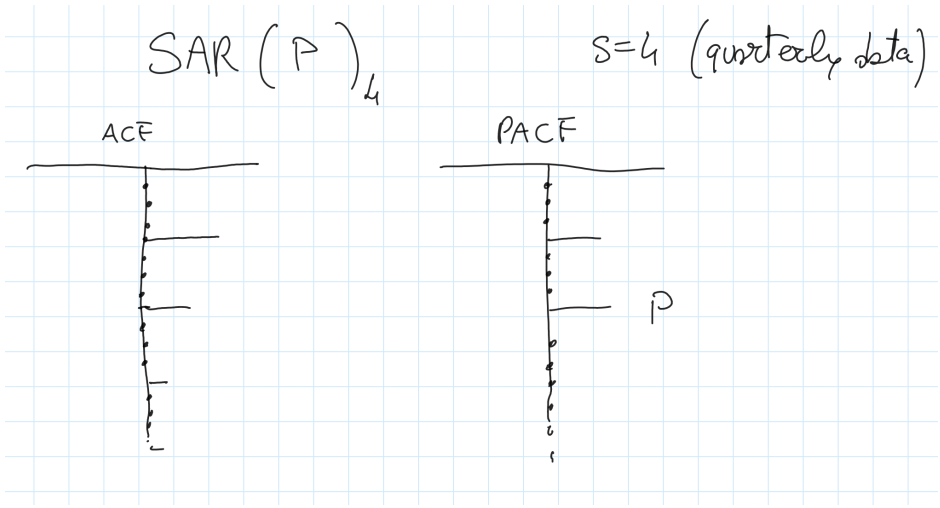


Figure 3.3: ACF e PACF di $SAR(P)$.

Questo tipo di modello è particolarmente utile per catturare le variazioni stagionali nei dati, come ad esempio nei dati economici che mostrano regolarità stagionali, come le vendite al dettaglio durante le festività o i dati meteorologici con variazioni stagionali nelle temperature.

3.2 Moving Average Processes (MA)

3.2.1 MA(q)

Un processo di **media mobile** $MA(q)$ è definito unicamente da una combinazione lineare dei white noise precedenti:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Questo tipo di processo è *sempre stazionario*. Senza l'introduzione di una costante, la media del processo sarà zero. Se una costante è inclusa, allora la media sarà uguale a quella costante:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \mu$$

Quindi:

$$\mathbb{E}(Y_t) = \mu$$

La varianza del processo è semplice da calcolare, in quanto i white noise non sono correlati:

$$Var(Y_t) = \sigma^2 + \theta_1^2 \sigma^2 + \ldots + \theta_q^2 \sigma^2 = \sigma^2 (1 + \theta_1^2 + \ldots + \theta_q^2)$$

La varianza risulta costante nel tempo.

Un processo $MA(q)$ rappresenta una combinazione lineare mobile dei white noise e dei loro ritardi. Questo genera una dipendenza nel processo, ma la dipendenza non si estende molto nel tempo. Mentre in un processo AR, la memoria si attenua lentamente nel tempo fino a zero, nel processo MA si ricorda solo degli ultimi q shock.

Ad esempio, in un processo $MA(1)$, abbiamo:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Di conseguenza:

- $\mathbb{E}(Y_t) = 0$
- $Var(Y_t) = (1 + \theta_1^2)\sigma^2$
- $Cov(Y_t, Y_{t-1}) = \mathbb{E}(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-1} + \theta_1 \varepsilon_{t-2}) = \mathbb{E}(\theta_1 \varepsilon_{t-1}^2) = \theta_1 \sigma^2$

Se non ci sono white noise comuni tra i due fattori, la covarianza sarà nulla. Ad esempio, la covarianza tra Y_t e Y_{t-2} sarà:

$$Cov(Y_t, Y_{t-2}) = \mathbb{E}(\varepsilon_t + \theta_1 \varepsilon_{t-1})(\varepsilon_{t-2} + \theta_1 \varepsilon_{t-3})$$

Nel processo $MA(q)$, l'ACF decresce gradualmente fino al ritardo q-esimo e poi va a zero dal $q+1$-esimo ritardo. Ciò indica una correlazione significativa solo per i primi q ritardi e poi una mancanza di correlazione. D'altra parte, la PACF nel processo $MA(q)$ va a zero gradualmente. Ciò significa che ogni ritardo successivo contribuisce alla correlazione tra le osservazioni, anche se in modo meno significativo. Questo è in contrasto con il processo $AR(p)$, dove l'ACF va a zero gradualmente e la PACF va a zero dal ritardo $p+1$.

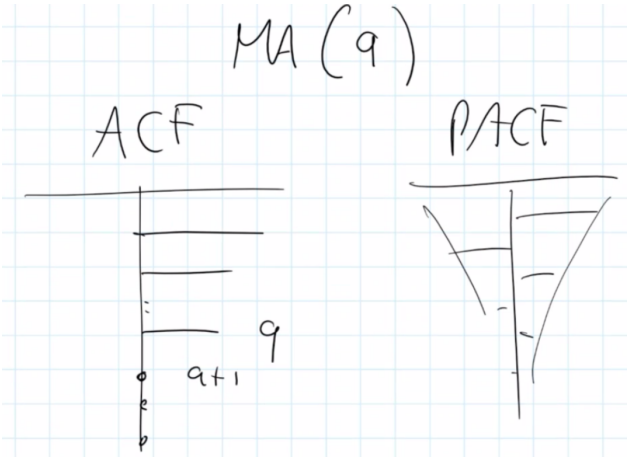


Figure 3.4: ACF e PACF di $MA(q)$.

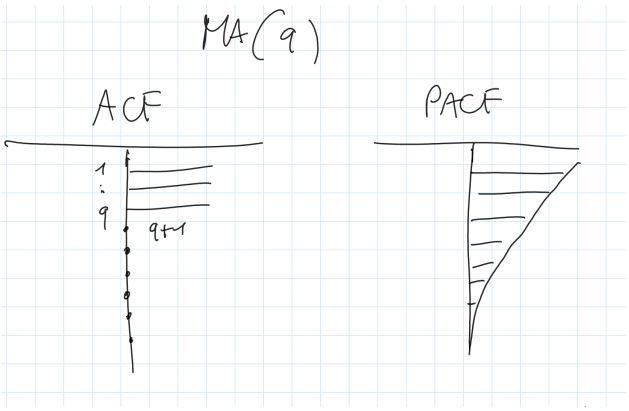


Figure 3.5: ACF e PACF di $MA(q)$.

Il processo $MA(q)$ può essere espresso in una forma più concisa utilizzando gli operatori di ritardo, noti come forma operatoriale.

Partendo dall'equazione del processo $MA(q)$:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Utilizzando l'operatore di ritardo, l'equazione del processo $MA(q)$ diventa:

$$Y_t = (1 + \theta_1 B + \ldots + \theta_q B^q)\varepsilon_t$$

Questa rappresentazione può essere abbreviata come:

$$Y_t = \theta_q(B)\varepsilon_t$$

3.2.2 Seasonal MA

La versione stagionale di un processo di media mobile di ordine q, indicata come **Seasonal** $MA(Q)$ o $SMA(Q)$, opera sui ritardi multipli di un periodo stagionale s. Matematicamente, può essere rappresentata come:

$$Y_t = \mu + (1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs})\varepsilon_t$$

La caratteristica principale di un processo MA è la sua memoria breve, il che significa che le osservazioni più recenti hanno un impatto maggiore sulle previsioni rispetto alle osservazioni più remote.

In questo contesto, l'operatore Θ rappresenta la componente stagionale del modello di media mobile. Essenzialmente, cattura la variazione stagionale nei dati e consente di incorporare tale variazione nelle previsioni.

Tuttavia, è importante notare che, poiché la memoria del processo MA è breve, una delle migliori previsioni possibili è la semplice media del processo.

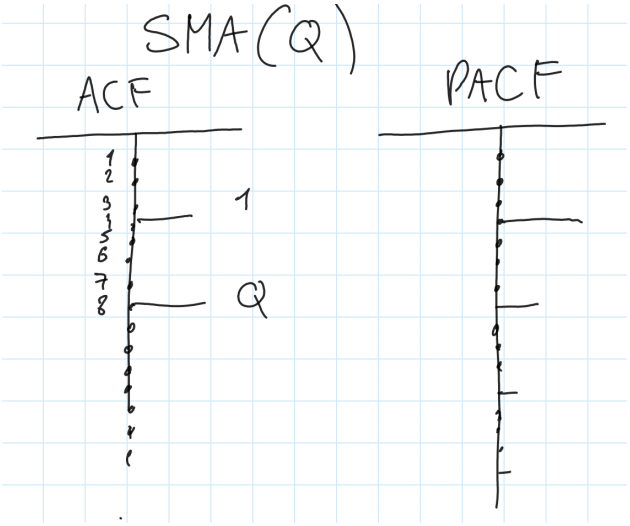


Figure 3.6: ACF e PACF di $SMA(Q)$.

3.3 ARMA

3.3.1 ARMA(p,q)

Unendo i modelli AR e MA, otteniamo il modello $ARMA(p, q)$ rappresentato dall'equazione:

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

È dimostrato che approssima efficacemente qualsiasi processo stazionario. Anche se con valori grandi di p e q si può approssimare qualsiasi processo, l'interesse principale risiede nel fatto che con valori ridotti di p e q è possibile ottenere una buona approssimazione di un ampio spettro di processi stazionari. Questo si traduce in una stima parsimoniosa dei parametri, il che rende il modello ARMA un metodo efficiente per parametrizzare l'ACF necessaria per le previsioni lineari delle serie temporali.

La stazionarietà del processo dipende principalmente dai coefficienti p, dalla parte AR, poiché la parte MA è sempre stazionaria. Il modello ARMA è stazionario solo se le radici dell'equazione caratteristica della parte AR sono unitarie.

Il vantaggio di combinare i modelli AR e MA è evidente: mentre entrambi potrebbero essere buoni approssimatori di un processo stazionario, unire le loro capacità permette di mantenere

il numero di parametri basso, rispetto a quello che sarebbe necessario se fossero utilizzati separatamente.

Il processo può essere riscritto in una forma più compatta come:

$$(1 - \phi_1 B - \ldots - \phi_p B^p) Y_t = c + (1 + \theta_1 B + \ldots + \theta_q B^q) \varepsilon_t$$

Che in notazione più semplice diventa:

$$\phi_p(B) Y_t = c + \theta_q(B) \varepsilon_t$$

Da cui otteniamo:

$$Y_t = \frac{\theta_q(B)}{\phi_p(B)} \varepsilon_t + \mu$$

Dove la media μ è data da:

$$\mu = \frac{c}{1 - \phi_1 - \cdots - \phi_p}$$

3.3.2 ARMA Definito ricorsivamente

Prendendo in considerazione un modello $AR(1)$, si desidera predire ciò che accadrà al tempo $t + 1$ basandosi su quanto è avvenuto fino al tempo t. Il modello segue un processo markoviano di ordine 1, dove tutta l'informazione necessaria per il tempo $t + 1$ è racchiusa nell'informazione al tempo t, senza la necessità di considerare il tempo $t - 1$. Pertanto, se si desidera conoscere l'informazione al tempo $t+2$, si può eliminare t poiché tutta l'informazione è contenuta nel tempo $t + 1$. Il tempo $t + 1$ dipende solo dal tempo t.

La previsione a un passo in avanti si ottiene considerando l'aspettativa condizionata di Y al tempo $t + 1$ data l'informazione fino al tempo t:

$$\hat{Y}_{t+1|t} = \phi Y_t$$

Per la previsione a due passi in avanti, si considera l'aspettativa condizionata di Y al tempo $t + 1$ data l'informazione fino al tempo t:

$$\hat{Y}_{t+2|t} = \phi \mathbb{E}(Y_{t+1} | Y_t)$$

Nel caso dei modelli lineari, la previsione ricorsiva è ottimale. Tuttavia, per i modelli non lineari, la situazione è diversa.

Nel modello lineare, si può controllare il comportamento in base ai parametri della parte autoregressiva. Se le radici sono al di fuori del cerchio unitario, la previsione converge alla

media del processo nel lungo termine. Nel caso di una sola radice unitaria, la convergenza avviene più lentamente, mentre se ci sono radici multiple unitarie, si verifica l'esplosione.

Per i modelli non lineari, la situazione è più complessa e dipende dall'esponente di Lyapunov. Costruire modelli multi-step diventa difficile.

Nei modelli ARMA invertibili, si può rappresentare il processo sia in termini di AR che di MA:

$$Y_t = k + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \ldots + \varepsilon_t$$

$$Y_t = \mu + \varepsilon_t + \psi\varepsilon_{t-1} + \psi\varepsilon_{t-2} + \ldots$$

Se il processo ARMA ha tutte le radici sia nella parte AR che nella parte MA al di fuori del cerchio unitario, può essere scritto in entrambi i modi. I coefficienti (π e ψ) convergono a zero all'aumentare del loro valore poiché il processo è stazionario e quindi ha varianza finita.

In sintesi, nei modelli lineari, conoscere il passato di Y e il passato di ε è approssimativamente equivalente, poiché entrambi i passati sono lineari.

3.4 ARIMA

3.4.1 Residui

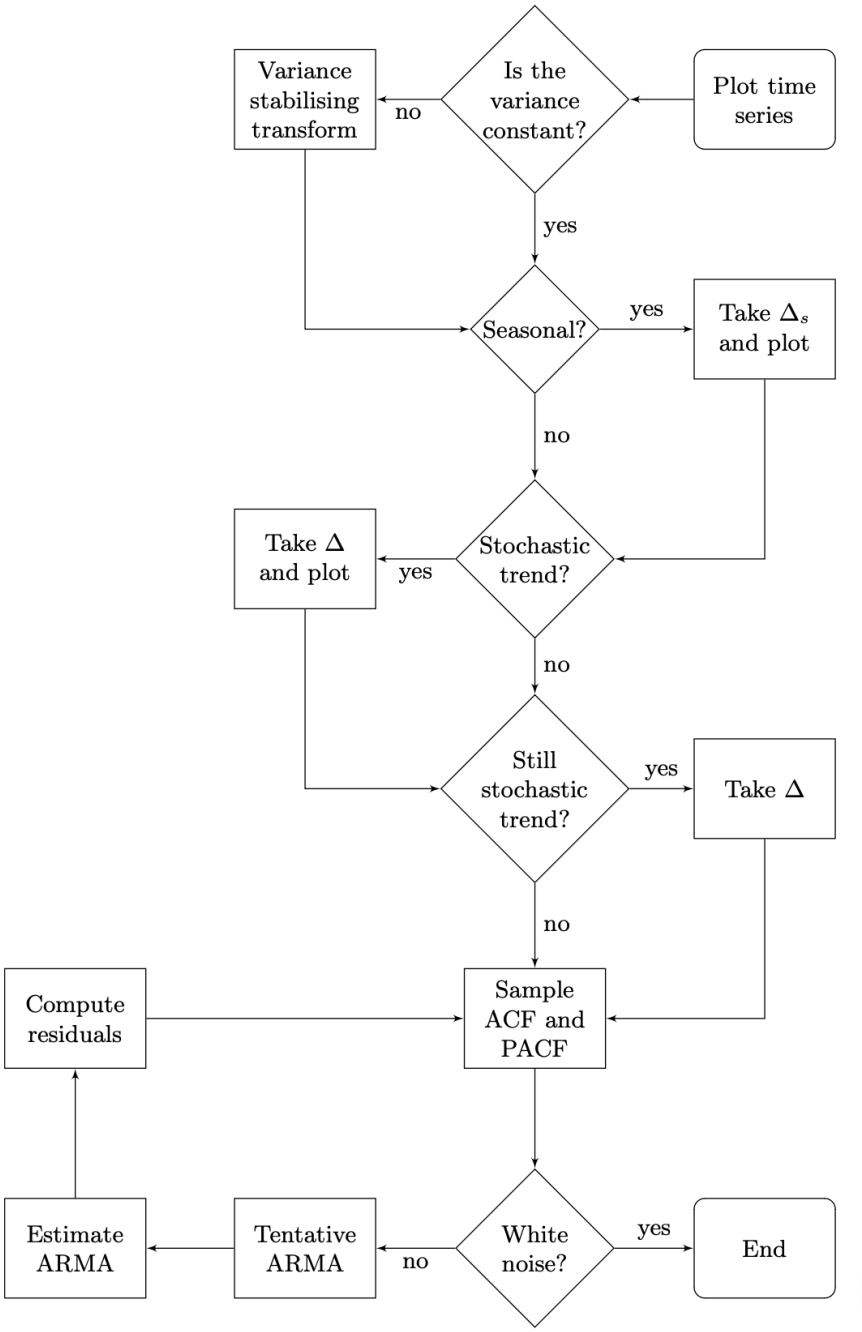
Iniziamo definendo cosa sono i **residui** in relazione alle serie temporali, anche conosciuti come innovazioni. Nei contesti delle serie temporali, i residui rappresentano la differenza tra l'osservazione effettiva e la previsione al passo successivo. Formalmente, possiamo esprimerli come:

$$\text{Residui} = Y_t - \hat{Y}_{t|n-1}$$

Dove il forecast, o previsione, al passo $n-1$ è definito come la proiezione lineare condizionata sull'intero storico delle osservazioni:

$$\hat{Y}_{t|n-1} = \mathbb{P}[Y_t | Y_{t-1}, \ldots, Y_1]$$

Quando stimiamo un modello ARIMA, il nostro obiettivo è minimizzare i residui. Pertanto, calcoliamo i residui, anche chiamati **innovazioni**, poiché rappresentano la parte di Y che non è spiegata o prevista dal modello. Desideriamo che le innovazioni siano dei rumori bianchi.



3.4.2 Modello ARIMA

Un modello ARIMA è comunemente denotato come:

$$ARIMA(p, d, q)(P, D, Q)_s$$

Dove (P, D, Q) rappresenta la parte stagionale e (p, d, q) la parte non stagionale. Il parametro d indica le differenze non stagionali (Δ), mentre D indica le differenze stagionali (Δ_s). p è l'ordine della parte AR, q è l'ordine della parte MA, e lo stesso vale per la parte stagionale P e Q.

Il processo che qui viene definito ARIMA, sarebbe in realtà SARIMA, ovvero Seasonal ARIMA, dato che la parte stagionale non è nulla; per semplicità lo chiameremo ARIMA comunque.

Ad esempio, se prendiamo una differenza stagionale di ordine 1 e vogliamo stimare un modello $AR(2)$ non stagionale dopo aver applicato questa differenza stagionale, scriviamo il modello come $ARIMA(2,0,0)(0,1,0)_{12}$. Questo indica che abbiamo una differenza stagionale di ordine 1 ($D = 1$), mentre il modello non stagionale è un $AR(2)$ ($p = 2$). Il parametro $s = 12$ indica la stagionalità con un periodo di 12, tipico ad esempio per dati mensili che presentano una stagionalità annuale.

Data la solita notazione:

- $\phi_p(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$
- $\theta_p(B) = (1 - \theta_1 B - \cdots - \theta_q B^q)$
- $\Phi_p(B) = (1 - \Phi_1 B - \cdots - \Phi_P B^P)$
- $\Theta_Q(B) = (1 - \Theta_1 B - \cdots - \Theta_Q B^Q)$

Dove le differenze sono:

- $\Delta^d = (1 - B)^d$
- $\Delta_s^D = (1 - B^s)^D$

Utilizzando questa notazione, possiamo scrivere il modello ARIMA in questione come:

$$\Delta^d \Delta_s^D Y_t = \frac{\theta_p(B)\Theta_Q(B)}{\phi_p(B)\Phi_p(B)}\varepsilon_t + \mu$$

Riprendendo quindi l'esempio precedente $ARIMA(1,0,0)(0,1,0)_{12}$, avremmo:

$$Y_t = \frac{1}{(1 - \phi_1 B)(1 - \Phi_1 B^{12})}\varepsilon_t + \mu$$

Moltiplicando a destra e a sinistra per $(1 - \phi_1 B)(1 - \Phi_1 B^1 2)$ si ottiene:

$$(1 - \phi_1 B)(1 - \Phi_1 B^1 2)Y_t = \varepsilon_t + (1 - \phi_1 B)(1 - \Phi_1 B^1 2)\mu$$

Da cui si ricava:

$$Y_t - \phi_1 Y_{t-1} - \Phi_1 Y_{t-12} + \phi_1 \Phi_1 Y_{t-13} = \varepsilon_t + \mu(1 - \phi_1 - \Phi_1 - \phi_1 \Phi_1)$$

Raggruppando $\mu(1 - \phi_1 - \Phi_1 - \phi_1 \Phi_1)$ come una costante c, otteniamo il processo:

$$Y_t = \phi_1 Y_{t-1} + \Phi_1 Y_{t-12} - \phi_1 \Phi_1 Y_{t-13} + c + \varepsilon_t$$

3.4.3 SARIMAX

Un modello SARIMAX, è un modello ARIMA che prende in considerazione più variabili esogene, può essere ottenuto sostituendo la variabile dipendente Y_t con $Y_t - \beta X_t^T$ nell'equazione:

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Quindi, l'equazione diventa:

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta^D(Y_t - \beta X_t^T) = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Dove X_t^T rappresenta un vettore di variabili esogene al tempo t e β è un vettore di coefficienti che rappresenta l'effetto di queste variabili esogene sul processo. In questo modo, il modello tiene conto delle relazioni tra le variabili esogene e la serie temporale Y_t, consentendo di incorporare informazioni aggiuntive nell'analisi e nelle previsioni.

In un modello ARIMAX, rappresentato dall'equazione:

$$\Delta^d\Delta_s^D(Y_t - \mu - X_t^T\beta) = ARMA(p, q)$$

Prendendo la regressione e calcolando le differenze dei residui, si osserva che l'errore della regressione non segue una distribuzione di rumore bianco, come di solito accade nelle regressioni lineari, ma segue un processo ARMA, che ha memoria. Quindi, il modello ARIMAX può essere interpretato come il risultato di prendere le differenze sia sulla serie temporale Y_t che sulle variabili esogene X_t^T, effettuare una regressione e modellare l'errore di regressione come un processo ARMA. In altre parole, anche quando si applica una regressione nelle serie temporali, si deve tenere conto della presenza di memoria negli errori di regressione. Se non si considera questa memoria, si otterranno non solo stime errate dei parametri di regressione, ma soprattutto non si terrà conto di tutta la memoria che può influenzare le previsioni future, compromettendo quindi la capacità del modello di fare previsioni accurate nel lungo termine.

Chapter 4

Modelli a Componenti Non Osservabili

4.1 Introduzione

Nei modelli a componenti non osservabili (**UCM**, da unobserved components models), anche detti modelli strutturali per serie storiche (STSM, da structural time series models), una serie storica è pensata come la somma di alcune componenti, appunto, non direttamente osservabili. Nella sua versione più tipica ed estesa un UCM è dato dalla somma di **trend**, **ciclo**, **stagionalità** e **rumore bianco** (a volte attribuibile all'errore con cui si misurano le serie storiche):

$$Y_t = \mu_t + \psi_t + \gamma_t + \varepsilon_t$$

Ciò permette di costruire previsori ottimi per la serie e per ciascuna componente, unite alle classiche misure di incertezza dell'inferenza statistica (errori standard, intervalli di confidenza, distribuzione dell'errore di previsione).

Come vedremo in questo capitolo, le singole componenti verranno derivate da funzioni deterministiche del tempo come la retta e le sinusoidi, rese stocastiche per mezzo dell'aggiunta di opportuni shocks casuali. Pertanto gli UCM possono essere visti come modelli di regressione su funzioni deterministiche del tempo in cui i parametri evolvono nel tempo.

Nei prossimi paragrafi analizzeremo i processi che definiscono le singole componenti, lasciando al capitolo successivo la spiegazione degli strumenti inferenziali che consentono la stima dei parametri ignoti del modello e delle componenti non osservabili.

4.2 La Componente Trend

La componente trend è responsabile delle variazioni della media del processo nel lungo periodo. Da un punto di vista economico, possiamo pensare al trend come a un movimento legato alla struttura di una economia, cioè come a quella componente che domina la crescita nel lungo periodo.

La funzione più semplice con cui molti tendono modellare il trend di una serie storica è sicuramente la retta:

$$\mu_t = \alpha + \beta t$$

Tuttavia, come si è visto nel capitolo precedente, un trend deterministico sembra essere un po' troppo rigido per adattarsi a molte serie storiche economiche osservate per un periodo di diversi anni o decenni. Per supplire a questa mancanza si sono introdotti i processi integrati, che contengono trend di natura stocastica.

La retta può essere scritta in forma incrementale come

$$\mu_0 = \alpha$$

$$\mu_t = \mu_{t-1} + \beta$$

Il lettore può verificare l'identità delle due rappresentazioni provando a fare i conti per $t = 0, 1, 2, \ldots$.

Il primo modo per rendere stocastica l'evoluzione della retta è quella di aggiungere un rumore bianco nell'equazione:

$$\mu_t = \mu_{t-1} + \beta + \eta$$

con $\eta_t \sim WN(0, \sigma^2_\eta)$. In questo modo μ_t è una passeggiata aleatoria con deriva (RWD). Tale processo corrisponde a una retta di coefficiente angolare β sommata ad un RW, che può essere visto come una intercetta stocastica della retta. È facile mostrare che la previsione (valore atteso condizionato) di un tale processo è data da

$$\hat{\mu}_{t+k|t} = \mu_t + \beta \cdot k$$

In tale processo, quindi, il coefficiente angolare non cambia mai e le previsioni si incrementano sempre della stessa quantità β qualunque sia la storia della serie.

Un modello di trend più flessibile si ottiene facendo evolvere il coefficiente angolare β come un RW. Facendo ciò si ottiene il **trend lineare locale** (LLT, da local linear trend)

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t$$

$$\beta_t = \beta_{t-1} + \zeta_t$$

con $\zeta_t \sim WN(0, \sigma_\zeta^2)$

Restringendo i parametri del LLT si ottengono alcuni casi interessanti per le applicazioni:

- passeggiata aleatoria quando $\sigma_\zeta^2 = 0$ e $\beta_0 = 0$
- passeggiata aleatoria con deriva quando $\sigma_\zeta^2 = 0$ e $\beta_0 \neq 0$
- passeggiata aleatoria integrata quando $\sigma_\eta^2 = 0$

Se il processo β_t fosse osservabile la previsione del LLT sarebbe

$$\mathbb{E}(\mu_{t+k}|\mu_t, \beta_t) = \mu_t + \beta_t \cdot k$$

che è ancora lineare, ma il coefficiente angolare β_t non è più costante per ogni t. Ovviamente, in pratica, noi non potremo mai osservare β_t direttamente e quindi la previsione si dovrà fare diversamente, ma della soluzione al problema delle previsioni parleremo più avanti.

4.3 La Componente Ciclo

La funzione "naturale" da cui partire per modellare un fenomeno ciclico, come un ciclo economico è una sinusoide

$$f(t) = R\cos(\varphi + \lambda t)$$

dove R è detta ampiezza della sinusoide, la quale oscilla infatti tra i valori R e $-R$, λ è al frequenza e rappresenta il numero di sinusoidi complete per unità di tempo e φ è la fase che, appunto, sfasa a sinistra o a destra il coseno. Infatti, quando la fase è nulla il valore del coseno per $t = 0$ è al suo valore massimo, mentre quando la fase è φ, il coseno raggiunge il suo massimo per $t = -\varphi/\lambda$. Dato che il coseno è funzione periodica di periodo 2π, cioè $\cos(x) = \cos(x + 2\pi)$, la funzione è periodica di periodo $2\pi/\lambda$, cioè $f(t) = f(t + 2\pi/\lambda)$ che può essere interpretato come il numero di unità temporali t necessarie affinché la sinusoide si ripeta.

Uno modo alternativo, spesso più conveniente, di scrivere la funzione è

$$f(t) = A\cos(\lambda t) + B\sin(\lambda t)$$

dove $A = R\cos(\varphi)$ e $B = -R\sin(\varphi)$.

Infatti, ricordando l'identità trigonometrica

$$\cos(a + b) = \cos a \cos b = \sin a \sin b$$

l'uguaglianza della è immediata. Per ricavare ampiezza e fase partendo da A e B è sufficiente risolvere il sistema rispetto a tali variabili:

$$R = \sqrt{A^2 + B^2}$$

$$\varphi = \arctan(-B/A)$$

per ottenere R è sufficiente prendere il quadrato di entrambe le righe e sommarle membro a membro, ricordando che per il teorema di Pitagora $\cos^2(x) + \sin^2(x) = 1$; per ottenere φ si divide la seconda equazione per la prima cambiata di segno e si inverta la funzione $\tan(\varphi) = \sin(\varphi)/\cos(\varphi) = -B/A$.

Quando, come nel nostro caso t è una variabile a valori in Z (numeri interi), possiamo riscrivere la funzione $f(t)$, che indicheremo con f_t, in forma incrementale:

$$\begin{bmatrix} f_0 \\ f_0^* \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix}$$

$$\begin{bmatrix} f_t \\ f_t^* \end{bmatrix} = \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-1}^* \end{bmatrix}$$

Si noti che f_t^* è solo una funzione di comodo utilizzata nella costruzione di f_t.

Per rendere stocastica la sinusoide f_t si può aggiungere una coppia di processi rumore bianco:

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}$$

con

$$\begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \sim WN(0, \sigma_\kappa^2 \boldsymbol{I}_2)$$

Si può mostrare che un tale processo è non stazionario, avendo nella sua parte autoregressiva due radici unitarie complesse coniugate. Spesso, tuttavia, si ritiene che un ciclo economico debba essere generato da un processo stazionario, rappresentando esso una reazione oscillatoria e transitoria dell'economia ad uno shock. La seguente modifica rende stazionario il ciclo ψ_t

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}$$

con $0 \leq \rho \leq 1$.

4.4 La Componenete Stagionale

Se la stagionalità fosse deterministica, cioè si ripetesse identicamente ogni anno, aggiungendosi alle altre componenti, potremmo definirla come funzione periodica di periodo s a somma nulla, dove s è il numero di osservazioni in un anno.

$$\sum_{i=0}^{s-1} \gamma_{t-i} = 0$$

o equivalente

$$\gamma_t = -\sum_{i=1}^{s-1} \gamma_{t-i}$$

Il fatto che tale somma sia nulla è necessario affinché la stagionalità sia una componente assente sulla serie annuale, cioè ottenuta come somma o media annua della serie originale.

Il modo più semplice per rendere stocastica, cioè evolutiva, la stagionalità, è quello di supporre che la somma della stagionalità sull'anno sia pari a un processo stazionario a media nulla, ω_t :

$$\gamma_t = -\sum_{i=1}^{s-1} \gamma_{t-i} + \omega_t$$

Se si sceglie ω_t rumore bianco, allora si ottiene la stagionalità a variabili di comodo stocastiche (stochastic dummies seasonality), che ha la seguente rappresentazione markoviana, che ci tornerà utile più avanti:

$$\gamma_t = \begin{bmatrix} -\mathbf{1} & -\mathbf{1}'_{s-1} \\ \boldsymbol{I}_{s-1} & \mathbf{0}_{s-1} \end{bmatrix} \gamma_{t-1} + \omega_t$$

dove 1_m e O_m sono vettori colonna di, rispettivamente, m uni e zeri, e $\boldsymbol{I}_m$ è la matrice identità $(m \times m)$.

La stagionalità deterministica trigonometrica rappresenta un approccio alternativo per gestire la stagionalità nelle serie storiche, particolarmente utile quando si lavora con dati a frequenza diversa (ad esempio, dati orari), utilizzando funzioni sinusoidali come seni e coseni.

La stagionalità può essere descritta come una funzione periodica che può essere analizzata tramite l'approccio di Fourier. Questo metodo sfrutta le proprietà delle funzioni periodiche, che si ripetono identicamente a intervalli regolari. In particolare, una funzione periodica di periodo s si ripete identicamente ogni s periodi, e la stagionalità dovrebbe esaurirsi entro un anno.

Un modo alternativo di modellare la stagionalità parte appunto dalla rappresentazione di una funzione periodica, $f : \mathbb{Z} \to \mathbb{R}$, a somma nulla, come somma di *sinusoidi* a frequenze di Fourier,

$$f(t) = f(t+s) = \sum_{j=1}^{\lfloor s/2 \rfloor} a_j \cos(\omega_j t) + b_j \sin(\lambda_j t)$$

dove $\lambda_j = 2\pi j/s$ e $\lfloor s/2 \rfloor$ è la parte intera di $s/2$. Per rendere stocastica questa rappresentazione possiamo utilizzare $\lfloor s/2 \rfloor$ cicli stocastici a frequenze di Fourier:

$$\gamma_t = \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_t^{(j)}$$

$$\begin{bmatrix} \gamma_t^{(j)} \\ \gamma_t^{(j)*} \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix} \begin{bmatrix} \gamma_{t-1}^{(j)} \\ \gamma_{t-1}^{(j)*} \end{bmatrix} + \begin{bmatrix} \omega_t^{(j)} \\ \omega_t^{(j)*} \end{bmatrix}$$

con

$$\begin{bmatrix} \omega_t^{(j)} \\ \omega_t^{(j)*} \end{bmatrix} \sim i.i.d. N_2(0, \sigma_\omega^2 \boldsymbol{I}_2)$$

Si noti che quando s è pari, la frequenza di Fourier più alta è

$$\frac{2\pi}{s} \cdot \frac{s}{2} = \pi$$

ed essendo $\sin\pi = 0$, l'equazione per $\gamma_t^{(s/2)*}$ diventa inutile, essendo azzerato il suo effetto su $\gamma_t^{(s/2)}$ tramite la moltipricazione per $\sin\pi$.

Anche la stagionalità così definita rispetta, ovviamente, la definizione base, ma ora ω_t si dimostra essere un processo $MA(s-1)$ a media nulla.

Dal punto di vista applicativo, la stagionalità con sinusoidi stocastiche tende a evolvere più gradualmente rispetto alla stagionalità con variabili di comodo stocastiche, infatti il processo MA che la fa evolvere è più liscia del processo rumore bianco che perturba la rappresentazione concorrente.

Chapter 5

Modelli in forma State-Space

5.1 Introduzione

Tutti i modelli visti fino ad ora (ARIMA e UCM) possono essere espressi in una forma comune molto flessibile che consente di condurre inferenza sia sui parametri ignoti, sia sulle componenti non osservabili. Tale forma, detta nello spazio degli stati, o più sinteticamente **state-space**, è molto generale e permette la rappresentazione di modelli anche non stazionari, non solo per via di radici unitarie, ma anche attraverso funzioni di autocovarianza che evolvono nel tempo. Sebbene in questo corso si esaminino solamente modelli univariati e uniequazionali, la forma state-space e la relativa strumentazione inferenziale verranno esposti per serie storiche multivariate, cioè per rappresentare una sequenza di vettori casuali $\{\boldsymbol{y}_t\}_{t=1}^n$, infatti questo non comporta alcuna complicazione, sia $t = 1$ nella definizione, sia nella derivazione degli stimatori.

L'utilizzo della forma state space e del **filtro di Kalman** (più avanti), insieme alle loro varianti, offre una soluzione completa per affrontare tre questioni cruciali nell'analisi di serie storiche:

1. *Modellazione congiunta di serie storiche*: La forma state space consente di modellare contemporaneamente più serie storiche, consentendo di trattare la variabile dipendente come un vettore. Questo permette di catturare le interazioni tra le diverse serie storiche e di incorporare tali informazioni nei modelli predittivi.
2. *Gestione di valori mancanti o anomali*: Entrambi, la forma state space e il filtro di Kalman, sono in grado di gestire serie storiche con valori mancanti o anomali. All'interno del filtro di Kalman, è possibile includere procedure per trattare in modo appropriato i valori mancanti, come l'imputazione automatica di dati mancanti o la

rimozione di valori anomali. Questo consente di mantenere la coerenza dei dati e di evitare distorsioni nei risultati analitici.

3. *Impatto dell'imputazione sui risultati di machine learning/deep learning*: In molte situazioni, è necessario affrontare i valori mancanti prima di applicare algoritmi di machine learning o deep learning. L'interpolazione o l'imputazione dei valori mancanti può influenzare significativamente i risultati ottenuti tramite tali algoritmi. Utilizzando il filtro di Kalman per l'imputazione dei dati mancanti, si garantisce che questa procedura non influenzi in modo distorto i risultati successivi, poiché il filtro tiene conto della struttura temporale dei dati e stima in modo coerente i valori mancanti.

In sintesi, l'approccio basato sulla forma state space e sul filtro di Kalman fornisce una metodologia robusta e completa per la gestione e l'analisi di serie storiche, consentendo di affrontare in modo efficace i problemi di dati mancanti, valori anomali e modellazione congiunta di più serie storiche, senza introdurre distorsioni nei risultati analitici successivi.

5.2 La forma State-Space

Sia $\boldsymbol{y}_t$ una serie storica di vettori casuali, di cui è osservabile una traiettoria finita. Nella forma state space $\boldsymbol{y}_t$ dipende linearmente da un vettore casuale $\boldsymbol{\alpha}_t$, detto vettore di stato, non osservabile (o parzialmente osservabile), che evolve secondo uno schema markoviano.

In formule, si ha la seguente coppia di sistemi di equazioni:

Equazione di misurazione (o di osservazione)

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{d}_t + \boldsymbol{\varepsilon}_t$$

Equazione di transizione (o di stato)

$$\boldsymbol{\alpha}_t = \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t$$

per $t = 1, \ldots, n$, con le seguenti proprietà

- $\boldsymbol{y}_t$ $(k \times 1)$ vettore di variabili osservabili,
- $\boldsymbol{\alpha}_t$ $(m \times 1)$ vettore di variabili (in genere) non osservabili,
- $\boldsymbol{Z}_t$ $(k \times m)$ matrice di (iper-)parametri,
- $\boldsymbol{d}_t$ $(k \times 1)$ vettore usato soprattutto per cambiare il valore medio di $\boldsymbol{y}_t$ per esempio per mezzo di regressori, $d_t = \boldsymbol{\beta}'\boldsymbol{x}_t$, (si può avere una rappresentazione equivalente del sistema anche senza $\boldsymbol{d}_t$),
- $\boldsymbol{\varepsilon}_t$ $(k \times 1)$ vettore di v.c. normali serialmente incorrelate con media nulla, $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$, e matrice di covarianza $\mathbb{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}') = \boldsymbol{H}_t$,
- $\boldsymbol{T}_t$ $(m \times m)$ matrice di transizione, contenente (iper-)parametri,

- $\boldsymbol{c}_t$ $(m \times 1)$ vettore di costanti con funzione analoga a $\boldsymbol{d}_t$,
- $\boldsymbol{\eta}_t$ $(g \times 1)$ vettore di v.c. normali serialmente incorrelate con con $\mathbb{E}(\boldsymbol{\eta}_t) = 0$, $\mathbb{E}(\boldsymbol{\eta}_t \boldsymbol{\eta}_t') = \boldsymbol{Q}_t$,
- $\boldsymbol{R}_t$ $(m \times g)$, matrice di (iper-)parametri usata per poter sempre definire la distribuzione del vettore casuale ηt propriamente (ovvero in modo che $\boldsymbol{Q}_t$ sia definita positiva).

Il sistema viene completato con le seguenti ipotesi sulla distribuzione del vettore di stato al tempo $t = 0$:

- $\boldsymbol{\alpha}_0 = N_m(\boldsymbol{\alpha}_0, \boldsymbol{P}_0)$,
- $\mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\eta}_s') = \boldsymbol{0}$, per ogni $s, t = 1, \ldots, n$ (questa ipotesi può essere rilassata, ma non lo faremo perché è valida per la grandissima parte dei modelli che si usano nelle applicazioni comuni),
- $\mathbb{E}[\varepsilon_t(\boldsymbol{\alpha}_0 - \boldsymbol{a}_0)'] = \boldsymbol{0}$, $\mathbb{E}[\eta_t(\boldsymbol{\alpha}_0 - \boldsymbol{a}_0)'] = \boldsymbol{0}$, per ogni $t = 1, \ldots, n$.

Il sistema è lineare, cioè $\boldsymbol{y}_t$ può essere espresso come combinazione lineare di valori presenti e passati di $\boldsymbol{\varepsilon}_t$, $\boldsymbol{\eta}_t$ e $\boldsymbol{\alpha}_0$. Ma poiché questi ultimi sono normali, anche i vettori $\boldsymbol{y}_t$ saranno distribuiti normalmente.

Se le matrici di sistema $\boldsymbol{Z}_t$, $\boldsymbol{d}_t$, $\boldsymbol{H}_t$, $\boldsymbol{T}_t$, $\boldsymbol{c}_t$, $\boldsymbol{R}_t$ e $\boldsymbol{Q}_t$ sono tutte costanti (cioè si possono omettere i pedici t), il sistema è detto time-invariant o time-homogeneous. I modelli stazionari sono un caso particolare dei sistemi time-invariant, cioè la condizione di omogeneità temporale è necessaria ma non sufficiente per la stazionarietà del sistema.

La forma state space va però completata aggiungendo la condizione iniziale.

$$\boldsymbol{\alpha}_1 \sim (\boldsymbol{a}_{1|0}, \boldsymbol{P}_{1|0})$$

Dove $\boldsymbol{a}$ indica la media e $\boldsymbol{P}$ indica la varianza. Se non si hanno informazioni e la componente è non stazionaria è possibile mettere media arbitraria (solitamente 0) e varianza infinita (che rappresenta la totale ignoranza su dove sia la componente al tempo $t = 1$). Se la componete invece è stazionaria è possibile mettere media marginale e varianza marginale (anche in mancanza di informazioni).

Per chiarire i concetti, possiamo iniziare esaminando un esempio di regressione lineare in forma state space (ssf), che è una forma più strutturata per rappresentare i modelli statistici. Iniziamo con un esempio semplice, dove la regressione lineare è espressa come:

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

dove ε segue una distribuzione normale con media zero e varianza σ^2. Riscrivendo questa equazione in forma state space, otteniamo:

$$y_t = Z_t \alpha_t + \varepsilon_t$$

dove Z_t è un vettore dipendente dal tempo che rappresenta i regressori x_t^T. Il rumore bianco (WN) rimane invariato. Ora, avendo specificato l'equazione di osservazione, dobbiamo definire l'equazione di transizione. Il parametro α, che qui corrisponde semplicemente a β dell'equazione, è costante nel tempo:

$$\alpha_{t+1} = \boldsymbol{I}\alpha_t$$

dove I è una matrice identità, poiché in questo caso non c'è un cambiamento nel tempo. Tuttavia, per completare il modello, è necessario specificare le condizioni iniziali, che in questo caso non sono solo il valore iniziale, ma rappresentano tutti i possibili valori di α poiché non cambiano nel tempo.

Possiamo anche considerare l'aggiunta di errori distribuiti come WN e far evolvere i coefficienti di regressione nel tempo:

$$\alpha_{t+1} = \boldsymbol{I}\alpha_t + \boldsymbol{I}\eta_t$$

Qui, i coefficienti della regressione evolvono come una random walk.

5.3 Modelli ARIMA in forma State-Space

Ora vediamo come è possibile adattare i modelli ARIMA alla forma state-space. Iniziamo con casi particolarmente semplici per poi dare la forma generale.

5.3.1 SSF per processi AR

AR(1)

Consideriamo il processo autoregressivo di primo ordine, $AR(1)$:

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

dove $\sigma^2 = Var(\varepsilon_t)$. La sua rappresentazione in forma state-space è data da:

$$\begin{cases} Y_t = 1\alpha_t \\ \alpha_{t+1} = \phi\alpha_t + \eta_t \end{cases}$$

dove $Z = 1$, $T = \phi$ e $\eta_t = \varepsilon_{t+1}$.

Se $|\phi| < 1$, allora α_t è stazionario, con:

- $\mathbb{E}(\alpha_1) = 0$
- $Var(\alpha_1) = \frac{\sigma^2}{1-\phi^2} = \gamma_0$

AR(1) con intercetta

Per un processo $AR(1)$ con intercetta:

$$Y_t = k + \phi Y_{t-1} + \varepsilon_t$$

definiamo l'equazione di transizione considerando le due parti di α_{t+1}:

$$\begin{bmatrix} k_{t+1} \\ \alpha_{t+1} \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} k_t \\ \alpha_t \end{bmatrix} + \begin{bmatrix} \ \end{bmatrix} + \eta_t$$

Per mantenere k costante nel tempo, otteniamo:

$$\begin{bmatrix} k_{t+1} \\ \alpha_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \end{bmatrix} \begin{bmatrix} k_t \\ \alpha_t \end{bmatrix} + \begin{bmatrix} 0 \end{bmatrix} + \eta_t$$

e per α:

$$\begin{bmatrix} k_{t+1} \\ \alpha_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & \phi \end{bmatrix} \begin{bmatrix} k_t \\ \alpha_t \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \eta_t$$

Otteniamo quindi un $AR(1)$ con intercetta:

$$\begin{cases} k_{t+1} = k_t \\ \alpha_{t+1} = k_t + \phi \alpha_t + \eta_t \end{cases}$$

dove $Z = 1$, $T = \phi$ e $\eta_t = \varepsilon_{t+1}$. Per quanto riguarda i valori iniziali, trattiamo le due componenti in modo diverso poiché una è una costante di cui non si conosce nulla. Pertanto, la media è zero e la varianza è infinita poiché non si conosce nulla di k.

Anche se è stazionario, non si sa nulla dell'intercetta iniziale:

$$\begin{bmatrix} k_1 \\ \alpha_1 \end{bmatrix} \sim dist(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \infty & 0 \\ 0 & \infty \end{bmatrix})$$

Ciò è dovuto alla mancanza di conoscenza del valore iniziale della costante.

AR(2)

Mettiamo il processo $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$ in forma state-space.

- Equazione di transizione

$$\begin{bmatrix} \alpha_t^{(1)} \\ \alpha_t^{(2)} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1}^{(1)} \\ \alpha_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t$$

- Equazione di osservazione

$$Y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_t^{(1)} \\ \alpha_t^{(2)} \end{bmatrix}$$

Dato che la seconda riga dell'equazione di transizione è l'identità $\alpha_t^{(2)} = \alpha_{t-1}^{(1)}$, sostituendo nella prima riga otteniamo

$$\alpha_t^{(1)} = \phi_1 \alpha_{t-1}^{(1)} + \phi_2 \alpha_{t-1}^{(2)} + \eta_t = \phi_1 \alpha_{t-1}^{(1)} + \phi_2 \alpha_{t-1}^{(1)} + \eta_t$$

che è un processo $AR(2)$. L'equazione di misurazione si riduce all'identità $Y_t = \alpha_t^{(1)}$.

Se $AR(2)$ è stazionario i valori iniziali sono

$$\begin{bmatrix} \alpha_1^{(1)} \\ \alpha_1^{(2)} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_0 \end{bmatrix} \sim dist(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma 1 & \gamma_0 \end{bmatrix})$$

AR(p)

È possibile generalizzare gli esempi appena visti per ogni processo $AR(p)$:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

- Equazione di transizione

$$\begin{bmatrix} \alpha_{t+1}^{(1)} \\ \vdots \\ \alpha_{t+1}^{(p)} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_p \\ & & & 0 \\ & \boldsymbol{I}_{p-1} & & \vdots \\ & & & 0 \end{bmatrix} \begin{bmatrix} \alpha_t^{(1)} \\ \vdots \\ \alpha_t^{(p)} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_t$$

- Equazione di osservazione

$$Y_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \alpha_t^{(1)} \\ \vdots \\ \alpha_t^{(p)} \end{bmatrix}$$

Anche in questo caso se il processo dovesse essere stazionario è possibile definire la distribuzione iniziale:

$$\begin{bmatrix} \alpha_1^{(1)} \\ \vdots \\ \alpha_1^{(p)} \end{bmatrix} \sim dist(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \ddots & \vdots \\ \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix})$$

5.3.2 SSF per processi MA

MA(1)

Mettiamo il processo $Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$ in forma state space.

- Equazione di transizione

$$\begin{bmatrix} \alpha_t^{(1)} \\ \alpha_t^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{t-1}^{(1)} \\ \alpha_{t-1}^{(2)} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t$$

- Equazione di osservazione

$$Y_t = \begin{bmatrix} 1 & \theta_1 \end{bmatrix} \begin{bmatrix} \alpha_t^{(1)} \\ \alpha_t^{(2)} \end{bmatrix}$$

Dato che la seconda riga dell'equazione di transizione è l'identità $\alpha_t^{(2)} = \alpha_{t-1}^{(1)}$, sostituendo nella prima riga otteniamo

$$\alpha_t^{(1)} = \varepsilon_t$$
$$\alpha_t^{(2)} = \alpha_{t-1}^{(1)} = \varepsilon_{t-1}$$

che sono, rispettivamente, un processo white noise e il medesimo processo ritardato di un periodo. L'equazione di misurazione assegna i coefficienti MA ai ritardi del white noise

$$Y_t = \alpha_t^{(1)} + \theta_1 \alpha_t^{(2)} = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

MA(q)

È possibile generalizzare gli esempi appena visti per ogni processo $MA(p)$:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- Equazione di transizione

$$\begin{bmatrix} \alpha_{t+1}^{(0)} \\ \vdots \\ \alpha_{t+1}^{(q)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ & & & 0 \\ & \boldsymbol{I}_q & & \vdots \\ & & & 0 \end{bmatrix} \begin{bmatrix} \alpha_t^{(0)} \\ \vdots \\ \alpha_t^{(q)} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_t$$

- Equazione di osservazione

$$Y_t = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_q \end{bmatrix} \begin{bmatrix} \alpha_{t+1}^{(0)} \\ \vdots \\ \alpha_{t+1}^{(q)} \end{bmatrix}$$

Per completare le condizioni iniziali, essendo un white noise la media sarà sicuramente zero e nella matrice varianza e covarianza avremo un whithe noise (con varianza σ^2 e il suo ritardo, ovvero la matrice identità con sulla diagmonale le varianze.

$$\begin{bmatrix} \alpha_1^{(0)} \\ \vdots \\ \alpha_1^{(q)} \end{bmatrix} \sim dist(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \boldsymbol{I}_{q+1}\sigma^2)$$

5.3.3 SSF per processi ARMA

ARMA(1,1)

Prendendo un processo MA e passandolo attraverso un filtro AR, si ottiene un processo ARMA. Viceversa, lo stesso vale se si prende un processo AR e si applica un filtro MA, ottenendo sempre un processo ARMA.

Ad esempio, considerando un processo $MA(1)$:

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

si può costruire un filtro AR (non un processo) e applicarlo a X invece di Y, generando una nuova serie temporale:

$$Y_t = \phi Y_{t-1} + X_t$$

dove X_t è il processo $MA(1)$. Tuttavia, questo processo non sarà un AR di ordine 1 poiché non è il rumore bianco a guidare il processo, ma il $MA(1)$. Invece, il processo risultante sarà un $ARMA(1,1)$, come si può vedere sostituendo il processo:

$$Y_t = \phi Y_{t-1} + \varepsilon + \theta\varepsilon_{t-1}$$

Questo principio si applica in generale a qualsiasi combinazione di processi e filtri $AR(p)$ e $MA(q)$.

Per dimostrare il contrario, consideriamo un processo $AR(1)$:

$$X_t = \phi X_{t-1} + \varepsilon_t$$

Possiamo costruire un filtro MA e applicarlo come segue:

$$Y_t = X_t + \theta X_{t-1}$$

Il primo processo può essere riscritto come:

$$(1 - \phi B)X_t = \varepsilon_t$$

da cui otteniamo:

$$X_t = (1 - \phi B)^{-1}\varepsilon_t$$

Sostituendo nella serie temporale, otteniamo:

$$Y_t = (1 - \phi B)^{-1}\varepsilon_t + (1 - \phi B)^{-1}\theta\varepsilon_{t-1}$$

Sviluppando, otteniamo:

$$(1 - \phi B)^{-1}Y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

e infine lo stesso processo $ARMA(1,1)$:

$$Y_t = \phi Y_{t-1} + \varepsilon + \theta\varepsilon_{t-1}$$

Il ragionamento che permette di trasformare un processo ARMA in forma state-space è fondamentale per semplificarne l'analisi. Consideriamo un processo $ARMA(1,1)$ definito come:

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$

La dimensione del vettore di stato, indicata con r, è determinata come il massimo tra p e $q+1$, dove p è l'ordine dell'AR e q è l'ordine del MA. Nel nostro caso, essendo presente un $AR(p=1)$ e un $MA(q=1)$, $r = 2$. Scriviamo ora l'equazione di transizione, concentrandoci sull'AR di ordine 1:

$$\begin{matrix} AR(1) \\ lagAR(1) \end{matrix} \begin{bmatrix} \alpha_{t+1}^{(0)} \\ \alpha_{t+1}^{(1)} \end{bmatrix} = \begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_t^{(0)} \\ \alpha_t^{(1)} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_t$$

Successivamente incorporiamo il filtro MA all'interno dell'equazione di stato:

$$Y_t = \begin{bmatrix} 1 & \theta \end{bmatrix} \begin{bmatrix} \alpha_t^{(0)} \\ \alpha_t^{(1)} \end{bmatrix} \begin{matrix} \rightarrow AR(1) \\ \rightarrow lagAR(1) \end{matrix}$$

Inizialmente, nel processo, le condizioni sono:

$$\boldsymbol{a}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{P}_1 = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 \end{bmatrix}$$

Per questo specifico caso, conosciamo che:

$$\gamma_0 = \frac{\sigma^2}{1 - \phi}$$

$$\gamma_1 = \frac{\sigma^2}{1 - \phi^2}\phi$$

ARMA(p,q)

Ora vediamo la rappresentazione di un generico modello $ARMA(p,q)$ (anche con radici unitarie) in forma state space.

Siano $r = \max(p, q+1)$, $\phi_j = 0$ per $j > p$ e σ_j per $j > q$.

- *Equazione di misurazione*

$$Y_t = \begin{bmatrix} 1 & \theta_1 & \cdots & \theta_{r-1} \end{bmatrix} \boldsymbol{\alpha}_t$$

  dove il vettore di stato $\boldsymbol{\alpha}$ è $(r \times 1)$.
- *Equazione di transizione*

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{r-1} & \phi_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \eta_t$$

L'equazione di stato definisce un processo $AR(r)$, infatti le righe dalla seconda all'ultima rappresentano le identità

$$\begin{aligned} \alpha_t^{(2)} &= \alpha_{t-1}^{(1)} \\ \alpha_t^{(3)} &= \alpha_{t-1}^{(2)} = \alpha_{t-2}^{(1)} \\ \ldots &= \ldots \\ \alpha_t^{(r)} &= \alpha_{t-1}^{(r-1)} = \ldots = \alpha_{r-1}^{(1)} \end{aligned}$$

e sostituendo nella prima riga

$$\alpha_t^{(1)} = \phi_1 \alpha_{t-1}^{(1)} + \ldots + \phi_r \alpha_{t-r}^{(1)} + \varepsilon_t$$

che può essere scritto in forma compatta

$$\alpha_t^{(1)} = \phi_r(B)^{-1} \varepsilon_t$$

Allo stesso modo, contenendo $\boldsymbol{\alpha}_t$ solamente ritardi dell'elemento in prima posizione, anche l'equazione di osservazione può essere riscritta in forma compatta come

$$Y_t = \theta_r(B) \alpha_t^{(1)}$$

e sostituendo il penultimo risultato nell'ultimo otteniamo

$$Y_t = \theta_r(B) \phi_r(B)^{-1} \varepsilon_t$$

che è un processo $ARMA(p,q)$.

5.4 Modelli UCM in forma State-Space

5.4.1 SSF per UCM

Vediamo come costruire il (sotto-)vettore di stato di ciascuna componente, per poi unirli tutti nel vettore di stato completo.

Trend lineare locale Usando la medesima notazione usata nel capitolo precedente, l'equazione di transizione di un LLT è data da

$$\begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} \eta_t \\ \zeta_t \end{bmatrix}$$

La matrice di covarianza dei disturbi è

$$Var \begin{bmatrix} \eta_t \\ \zeta_t \end{bmatrix} = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}$$

Solamente il trend μ_t dovrà entrare nell'equazione di osservazione.

Ciclo stocastico Il ciclo è già stato espresso in forma markoviana:

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}$$

Anche in questo caso solamente il primo elemento del vettore di stato dovrà entrare nell'equazione di osservazione.

Come si è visto, due sono i processi utilizzati più frequentemente per modellare la stagionalità. Vediamo la rappresentazione markoviana di entrambi.

Sinusoidi stocastiche stagionali La componente stagionale può essere ottenuta dalla somma di $v = \lfloor s/2 \rfloor$ sinusoidi stocastiche (non stazionarie).

$$\begin{bmatrix} \gamma_t^{(1)} \\ \gamma_t^{(1)*} \\ \vdots \\ \gamma_t^{(v)} \\ \gamma_t^{(v)*} \end{bmatrix} = \begin{bmatrix} \cos\lambda_1 & \sin\lambda_1 & \cdots & 0 & 0 \\ -\sin\lambda_1 & \cos\lambda_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos\lambda_v & \sin\lambda_v \\ 0 & 0 & \cdots & -\sin\lambda_v & \cos\lambda_v \end{bmatrix} \begin{bmatrix} \gamma_{t-1}^{(1)} \\ \gamma_{t-1}^{(1)*} \\ \vdots \\ \gamma_{t-1}^{(v)} \\ \gamma_{t-1}^{(v)*} \end{bmatrix} + \begin{bmatrix} \omega_t^{(1)} \\ \omega_t^{(1)*} \\ \vdots \\ \omega_t^{(v)} \\ \omega_t^{(v)*} \end{bmatrix}$$

Si noti che quando s è pari l'ultima riga può essere rimossa, in quanto irrilevante per la sinusoide $\gamma_t^{(1)}$ (viene moltiplicata per $\sin\pi = 0$). Solitamente la matrice di covarianza dei disturbi viene posta uguale a $\sigma_\omega^2 I$.

Nell'equazione di osservazione deve entrare la somma delle righe dispari del vettore di stato.

Dummy stocastiche stagionali La relazione $\gamma_t = -\gamma_{t-1} - \ldots - \gamma_{t-s+1} + \boldsymbol{\omega}_t$ può essere scritta in forma markoviana come

$$\begin{bmatrix} \gamma_t^{(1)} \\ \vdots \\ \gamma_t^{(s-1)} \end{bmatrix} = \begin{bmatrix} -\mathbf{1}'_{s-2} & -\mathbf{1} \\ \boldsymbol{I}_{s-2} & \mathbf{0}_{s-2} \end{bmatrix} \begin{bmatrix} \gamma_{t-1}^{(1)} \\ \vdots \\ \gamma_{t-1}^{(s-1)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\omega}_t \\ \mathbf{0}_{s-2} \end{bmatrix}$$

dove $\mathbf{1}_r$ e $\mathbf{0}_r$ sono vettori colonna di r elementi tutti pari, rispettivamente, a uno e a zero.

Nell'equazione di osservazione entra solamente il primo elemento del vettore di stato. La matrice di covarianza dei disturbi è nulla ovunque tranne il primo elemento che è pari a σ_ω^2.

Il vettore di stato completo, $\boldsymbol{\alpha}_t$, si ottiene concatenando verticalmente i singoli vettori di stato, mentre la matrice di transizione, $\boldsymbol{T}$, completa è data dalla concatenazione diagonale delle matrici di transizione delle singole componenti. Analogamente, il vettore dei disturbi, $\boldsymbol{\eta}_t$, si ottiene concatenando verticalmente i singoli vettori degli shock di ciascuna componente e la relativa matrice di covarianza, $\boldsymbol{Q}$, è data dalla concatenazione diagonale delle singole matrici di covarianza.

Per quanto riguarda l'equazione di osservazione, la matrice $\boldsymbol{Z}$ si riduce ad un vettore riga contenente 1 in corrispondenza agli elementi di $\boldsymbol{\alpha}_t$ che entrano direttamente nella definizione di Y_t e 0 altrove.

5.4.2 Sviluppo della SSF per varianti di UCM

I modelli UCM si collocano in un contesto noto come "*mondo modulare*", dove è possibile aggiungere o rimuovere componenti semplicemente manipolando le variabili di stato.

Per esempio, è fattibile creare un modello UCM che incorpori trend e stagionalità aggiungendo il vettore di stato della stagionalità al vettore di stato del trend e successivamente collegando diagonalmente le matrici $\boldsymbol{T}$ e $\boldsymbol{Q}$.

A differenza dei modelli ARIMA che approssimano qualsiasi serie storica attraverso filtri, i modelli UCM sfruttano informazioni preesistenti, come trend e stagionalità, e le integrano in modo modulare. Le varie componenti vengono unite sequenzialmente come variabili di stato, permettendo l'aggiunta agevole di ulteriori variabili senza complicazioni significative. Le matrici $\boldsymbol{T}$ e $\boldsymbol{Q}$ (covarianza/varianza) vengono concatenate diagonalmente, consentendo così una flessibilità nella gestione e nell'aggiunta delle variabili di stato.

Le matrici $\boldsymbol{Z}$, $\boldsymbol{T}$, $\boldsymbol{R}$, $\boldsymbol{Q}$, e $\boldsymbol{H}$ sono essenziali nella definizione delle condizioni iniziali per il sistema. È importante notare che, a livello software, non è necessario costruire manualmente queste matrici per poi passarle al filtro di Kalman. Esistono funzioni dedicate che

consentono di definire e gestire queste matrici in modo più efficiente e pratico, semplificando l'implementazione del filtro di Kalman.

LLT Per ottenere la forma state space di un sistema con rumore, come nel caso del modello LLT, possiamo definire le matrici $\boldsymbol{T}$ e $\boldsymbol{R}$ come segue:

$$\boldsymbol{T} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

La matrice $\boldsymbol{T}$ rappresenta la dinamica del sistema, dove μ_{t+1} è definito come la somma di se stesso ritardato e β_t.

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La matrice $\boldsymbol{R}$ è una matrice identità che associa il primo rumore η_t alla prima equazione e il secondo rumore ζ_t alla seconda equazione.

Quindi, la forma state space del sistema diventa:

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \end{bmatrix} = \underset{T}{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}} \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} + \underset{R}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} \begin{bmatrix} \eta_t \\ \zeta_t \end{bmatrix}$$

Dove μ_t e β_t sono le variabili di stato del sistema al tempo t, e η_t e ζ_t sono i rumori bianchi associati a ciascuna delle equazioni del sistema.

La matrice $\boldsymbol{Q}$ rappresenta la varianza-covarianza dei rumori η_t e ζ_t, che sono incorrelati tra loro per definizione:

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}$$

Dove σ_η^2 e σ_ζ^2 sono le varianze dei rumori associati a η_t e ζ_t rispettivamente.

La matrice $\boldsymbol{H}$ rappresenta la varianza del rumore ε, poiché ci troviamo in una situazione scalare:

$$\boldsymbol{H} = \sigma_\varepsilon^2$$

Se consideriamo solamente questa componente (eventualmente contaminata da rumore), la matrice $\boldsymbol{Z}$ dovrebbe permettere di selezionare solo μ_t e annullare lo slope β_t. In altre parole, la matrice $\boldsymbol{Z}$ è progettata per estrarre le componenti desiderate dalla variabile di stato.

$$Y_t = \underset{Z}{\begin{bmatrix} 1 & 0 \end{bmatrix}} \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} + \varepsilon_t$$

+ stagionalità con dummy Per costruire una componente stagionale per dati trimestrali e applicare le dummy stocastiche, possiamo definire la forma state space come segue:

$$Y_t = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{Z} \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} + \varepsilon_t$$

Considerando che ci sono 4 osservazioni all'anno e il vettore di stato è piccolo, composto da 3 elementi $(s-1)$, possiamo procedere con la definizione della forma state space.

Per quanto riguarda la condizione iniziale, assumiamo che μ e β non siano stazionari, pertanto impostiamo le seguenti condizioni iniziali:

$$\boldsymbol{a}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{P}_1 = \begin{bmatrix} \infty & 0 \\ 0 & \infty \end{bmatrix}$$

In questa configurazione, a_1 rappresenta la media iniziale delle variabili di stato e P_1 è la matrice di covarianza delle variabili di stato. Poiché le componenti non sono stazionarie, abbiamo impostato la varianza delle variabili di stato a infinito.

La forma finale del modello LLT con rumore è quindi definita dalle equazioni sopra, con le condizioni iniziali specificate. Adesso si può procedere con l'assemblaggio del modello UCM combinando le matrici $\boldsymbol{T}$, $\boldsymbol{Q}$, e $\boldsymbol{R}$.

Per creare una stagionalità utilizzando dummy stocastiche con $s = 3$, dobbiamo definire il meccanismo di generazione della stagionalità e dei suoi ritardi. Supponiamo di avere una stagionalità γ_t nel trimestre t, il ritardo della stagionalità nel trimestre precedente $\gamma_t^{(1)}$, e il ritardo della stagionalità nel secondo trimestre precedente $\gamma_t^{(2)}$.

Il meccanismo di aggiornamento della stagionalità può essere rappresentato come segue:

$$\begin{bmatrix} \gamma_{t+1} \\ \gamma_{t+1}^{(1)} \\ \gamma_{t+1}^{(2)} \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{T} \begin{bmatrix} \gamma_t \\ \gamma_t^{(1)} \\ \gamma_t^{(2)} \end{bmatrix} + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{R} \omega_t$$

Dove ω_t rappresenta il rumore associato alla stagionalità.

La matrice $\boldsymbol{Q}$ rappresenta la varianza del rumore ω_t, e poiché si tratta di un rumore scalare, $\boldsymbol{Q}$ sarà uno scalare:

$$\boldsymbol{Q} = \left[\sigma_\omega^2\right]$$

Se si considera la stagionalità osservata con rumore, si può definire $\boldsymbol{H} = \sigma_\varepsilon^2$ come la varianza del rumore ε_t.

Per estrarre dalla variabile di stato solo la componente corrispondente alla stagionalità (e mettere a 0 le altre due), si definisce la matrice $\boldsymbol{Z}$ come segue:

$$Y_t = \underset{Z}{\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}} \begin{bmatrix} \gamma_t \\ \gamma_t^{(1)} \\ \gamma_t^{(2)} \end{bmatrix} + \varepsilon_t$$

Dove Y_t rappresenta l'osservazione della serie temporale e ε_t è il rumore associato all'osservazione. In questa configurazione, la matrice $\boldsymbol{Z}$ selezionerà solo la prima variabile (stagionalità) dalla variabile di stato, azzerando le altre due (ritardi della stagionalità).

Per combinare un processo LLT con una stagionalità basata su dummy, è necessario creare un vettore di stato che contenga tutte le variabili di stato coinvolte nei due processi. Ecco come si può procedere:

1. Si inizia con le variabili di stato del modello LLT: μ, β, e la stagionalità γ, insieme ai suoi ritardi $\gamma^{(1)}$ e $\gamma^{(2)}$.
2. Si aggiunge il vettore di stato ritardato, che rappresenta lo stato al tempo $t-1$.
3. Infine, si include un vettore degli errori che contiene i rumori associati alle componenti del modello: η per il livello, ζ per lo slope e ω per la stagionalità.

Si inizia con l'equazione di aggiornamento delle variabili di stato:

$$\underset{1}{\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma_{t+1}^{(1)} \\ \gamma_{t+1}^{(2)} \end{bmatrix}} = \left[\begin{array}{cc|ccc} 1 & 1 & & & \\ 0 & 1 & & & \\ \hline & & & & \\ & & & & \\ & & & & \end{array}\right] \underset{2}{\begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma_t^{(1)} \\ \gamma_t^{(2)} \end{bmatrix}} + \begin{bmatrix} & & \\ & & \\ & & \\ & & \\ & & \end{bmatrix} \underset{3}{\begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}}$$

Nella matrice di transizione $\boldsymbol{T}$, i primi due blocchi riguardano il trend (μ_t e β_t), mentre gli ultimi tre blocchi riguardano la stagionalità (γ_t, $\gamma_t^{(1)}$, e $\gamma_t^{(2)}$), dove vengono introdotti i ritardi.

μ e β non dipendono dai γ, quindi tre zeri, ovvero γ non deve essere presente né nell'equazione di μ né in quella di β.

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma^{(1)}_{t+1} \\ \gamma^{(2)}_{t+1} \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline & & & & \\ & & & & \\ & & & & \end{array}\right] \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

dopo di che viene definita la stagionalità in tempo t, che è data dalla stagionalità dei trimestri precedenti $t-1$, $t-2$, $t-3$.

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma^{(1)}_{t+1} \\ \gamma^{(2)}_{t+1} \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline & & -1 & -1 & -1 \\ & & & & \\ & & & & \end{array}\right] \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

successivamente aggiungo il meccanismo di ritardo

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma^{(1)}_{t+1} \\ \gamma^{(2)}_{t+1} \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline & & -1 & -1 & -1 \\ & & 1 & 0 & 0 \\ & & 0 & 1 & 0 \end{array}\right] \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

Siccome μ non entra nell'equazione dei γ ci saranno tutti 0 nell'ultima parte della matrice.

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma^{(1)}_{t+1} \\ \gamma^{(2)}_{t+1} \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right] \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

Quindi si combinano diagonalmente le $\boldsymbol{T}$ delle componenti e si aggiunge 0 al di fuori delle diagonali. Il funzionamento è a blocchi, dove vengono riempiti i blocchi che definiscono le componenti e fuori da questi si pone tutto a 0.

Anche la $\boldsymbol{R}$ è definita allo stesso modo perché per la prima equazione va preso il primo errore, per la seconda equazione il secondo errore, per la terza (che sarebbe anche la prima della stagionalità) va preso il terzo errore, poi tutti 0 (le altre sono solo identità).

$$\begin{bmatrix} \mu_{t+1} \\ \beta_{t+1} \\ \gamma_{t+1} \\ \gamma^{(1)}_{t+1} \\ \gamma^{(2)}_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}}_{T} \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \end{bmatrix}$$

Anche la $\boldsymbol{R}$ si combina diagonalmente. Dove gli 1 indicano i tre errori, il primo blocco è il blocco della LLT, mentre il quarto è quello della stagionalità.

Vale lo stesso anche per la matrice $\boldsymbol{Q}$, ovvero la matrice di varianza dei tre errori, ovvero la combinazione a blocchi della matrice varianza-covarianza degli errori.

$$\boldsymbol{Q} = \left[\begin{array}{cc|c} \sigma^2_\eta & 0 & 0 \\ 0 & \sigma^2_\zeta & 0 \\ \hline 0 & 0 & \sigma^2_\omega \end{array}\right]$$

La matrice $\boldsymbol{Z}$ si combina orizzontalmente, il μ e azzera β, poi prende la prima componente che è la stagionalità e azzera i ritardi, più il solito ε.

$\boldsymbol{H}$ rimane sempre uguale a se stesso perché è solo un rumore che aggiunto.

$$Y_t = \left[\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \end{array}\right] \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma^{(1)}_t \\ \gamma^{(2)}_t \end{bmatrix} + \varepsilon_t$$

$$\boldsymbol{H} = \sigma^2_\varepsilon$$

È anche possibile a questo punto definire la condizione iniziale. Nel vettore delle medie del LLT in questo caso mettiamo delle medie arbitrarie (sempre 0) perché tanto metteremo varianza infinita. Anche nella parte del vettore riguardante la stagionalità (che è non stazionaria) verranno messi 0 e avrà varianza infinita.

$$\boldsymbol{a}_1 = \left[\begin{array}{c} 0 \\ 0 \\ \hline 0 \\ 0 \\ 0 \end{array}\right]$$

$$\boldsymbol{P}_1 = \left[\begin{array}{cc|ccc} \infty & 0 & 0 & 0 & 0 \\ 0 & \infty & 0 & 0 & 0 \\ \hline 0 & 0 & \infty & 0 & 0 \\ 0 & 0 & 0 & \infty & 0 \\ 0 & 0 & 0 & 0 & \infty \end{array}\right]$$

Tutte componenti non stazionarie, sia LLT che stagionalità; se si aggiunge un ciclo stazionario, viene aggiunta una componente con degli elementi finiti (non varianze infinite).

+ cliclo stocastico Complichiamo ulteriormente il modello con un ciclo stocastico. Ovvero si avrà un LLT (2 componenti), con stagionalità trimestrale (3 componenti), con ciclo stocastico (2 componenti). Le matrici saranno di dimensione 7×7 (tranne la $\boldsymbol{R}$ che dipende dal numero degli errori).

Innanzitutto avremo

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \psi_t \\ \psi_t^* \\ \gamma_t^{(1)} \\ \gamma_t^{(2)} \\ \gamma_t^{(3)} \end{bmatrix}$$

I blocchi della $\boldsymbol{T}$ sono distribuiti nel seguente ordine:

- le due componenti del LLT (μ e β)
- le due componenti del ciclo stocastico stazionario
- le tre componenti della stagionalità a dummy stocastiche

Sono combinate diagonalmente a blocchi le matrici del LLT, del cicli stocastico stazionario e della stagionalità a dummy stocastiche.

$$\boldsymbol{T} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho\cos(\lambda) & \rho\sin(\lambda) & 0 & 0 & 0 \\ 0 & 0 & -\rho\sin(\lambda) & \rho\cos(\lambda) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

La matrice $\boldsymbol{Q}$ viene utilizzata senza la matrice $\boldsymbol{R}$, la quale rappresenta i ritardi della stagionalità. Di conseguenza, all'interno della matrice di varianza-covarianza, sono presenti degli

zeri. Questo accade perché, se si combinassero le matrici $\boldsymbol{R}$ e $\boldsymbol{Q}$, non si avrebbero zeri sulle varianze della matrice varianza-covarianza. La matrice $\boldsymbol{R}$ è utilizzata per generare una $\boldsymbol{Q}$ con varianze nulle, quindi senza ulteriori variabili casuali.

Sulla diagonale della matrice $\boldsymbol{Q}$ ci sono diversi termini: σ_η^2 e σ_ζ^2, che rappresentano gli errori della parte del modello LLT; due volte σ_κ^2, poiché ci sono due varianze uguali nel ciclo stocastico; σ_ω^2, la varianza della stagionalità; ed infine due zeri, poiché la seconda e la terza componente della stagionalità a dummy stocastiche sono delle identità.

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_\eta^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\kappa^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\kappa^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\omega^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Se ci fosse una matrice $\boldsymbol{R}$, le ultime due righe e colonne della matrice $\boldsymbol{Q}$ non sarebbero presenti. La matrice $\boldsymbol{R}$ assocerebbe ciascun white noise a ciascuna componente del modello. Le prime cinque righe e colonne di $\boldsymbol{R}$ sarebbero una matrice identità, mentre le ultime due righe sarebbero nulle, in quanto rappresentano gli ultimi due ritardi della stagionalità, i quali non sono associati a rumore.

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I}_5 & \boldsymbol{0}_{5\times 2} \end{pmatrix}$$

$$\boldsymbol{Q} = \begin{bmatrix} \sigma_\eta^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\kappa^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\kappa^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\omega^2 \end{bmatrix}$$

Ovviamente $\boldsymbol{\eta}_t$ sarà:

$$\boldsymbol{\eta}_t = \begin{bmatrix} \eta_t \\ \zeta_t \\ \kappa_t \\ \kappa_t^* \\ \omega_t \end{bmatrix}$$

La matrice $\boldsymbol{Z}$ è:

- $\begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$ perché si seleziona il μ e il β

- $\begin{bmatrix}\cdot & \cdot & 1 & 0 & \cdot & \cdot & \cdot\end{bmatrix}$ perché si seleziona il primo ψ (ovvero la prima componente del ciclo stocastico) e si mette la seconda a 0 perché ne va preso solo una (prendere la prima o la seconda non cambierebbe, è arbitrario, però solitamente si sceglie di prendere la prima)
- $\begin{bmatrix}\cdot & \cdot & \cdot & \cdot & 1 & 0 & 0\end{bmatrix}$ perché c'è la stagionalità a dummy stocastiche, si mette a 1 solo il primo elemento della stagionalità, mentre gli altri sono posti a 0 perché non sono necessari per l'estrazione della componente desiderata.

$$\boldsymbol{Z} = \begin{bmatrix}1 & 0 & 1 & 0 & 1 & 0 & 0\end{bmatrix}$$

$\boldsymbol{H}$ rimane sempre una varianza, c'è sempre una componente noise che si somma.

$$\boldsymbol{H} = \begin{bmatrix}\sigma_\varepsilon^2\end{bmatrix}$$

Per quanto riguarda le condizioni iniziali, si possono delineare come segue:

- Tutti i valori di a_1 sono nulli poiché la componente stazionaria (il ciclo economico) ha una media pari a zero. Le altre componenti sono tutte arbitrariamente determinate: le componenti non stazionarie hanno varianza infinita, rendendo quindi qualsiasi numero associato alla media insignificante.

$$\boldsymbol{a}_1 = \begin{bmatrix}0\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

- La matrice P_1 rappresenta la varianza-covarianza iniziale. Inizialmente contiene le due componenti del LLT, entrambe stazionarie con varianza infinita. Successivamente, include le due componenti del ciclo stazionario, per il quale la varianza marginale esiste e si calcola come varianza del rumore bianco diviso per $1 - (\text{fattore di smorzamento})^2$. Qui, ρ rappresenta il fattore di smorzamento che determina la persistenza del processo. Infine, sono presenti le componenti della stagionalità e i suoi ritardi, che sono non stazionarie e quindi associate a varianza infinita.

$$\boldsymbol{P}_1 = \begin{bmatrix}\infty & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & \infty & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \sigma_\kappa^2/(1-\rho^2) & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \sigma_\kappa^2/(1-\rho^2) & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & \infty & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & \infty & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & \infty\end{bmatrix}$$

+ stagionalità trigonometrica Anche se si cambia da dummy a rappresentazione trigonometrica, il metodo di costruzione rimane lo stesso, ma i valori cambiano.

Utilizzando una frequenza di $\frac{2\pi}{4}$ per la prima coppia (ovvero avendo $s = 4$), non abbiamo più una coppia come ultimo elemento. Quando abbiamo s pari, possiamo eliminare la seconda componente della coppia delle sinusoidi stocastiche usate nella stagionalità, poiché otterremmo $2\pi \times \frac{2}{4}$, quindi π nella parte seno diventa 0.

Importante ricordare che $\lambda_j = \frac{2\pi}{s} j$

Tuttavia, la rappresentazione stagionale trigonometrica stocastica continua ad avere tre variabili di stato e non diventa più esosa rispetto alla rappresentazione con variabili stocastiche.

$$\boldsymbol{T} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho\cos(\lambda) & \rho\sin(\lambda) & 0 & 0 & 0 \\ 0 & 0 & -\rho\sin(\lambda) & \rho\cos(\lambda) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos(2\pi\frac{1}{4}) & \sin(2\pi\frac{1}{4}) & 0 \\ 0 & 0 & 0 & 0 & -\sin(2\pi\frac{1}{4}) & \cos(2\pi\frac{1}{4}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cos(2\pi\frac{2}{4}) \end{bmatrix}$$

Nelle varianze, mentre in precedenza solo la prima varianza della componente stagionale era diversa da zero, con la rappresentazione trigonometrica stocastica imponiamo la stessa varianza e quindi la stessa velocità di evoluzione per tutte le componenti sinusoidali (tre varianze σ^2_ω).

$$\boldsymbol{Q} = \begin{bmatrix} \sigma^2_\eta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_\zeta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_\kappa & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_\kappa & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_\omega & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_\omega & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_\omega \end{bmatrix}$$

La matrice $\boldsymbol{Z}$ in questo caso dovrebbe prendere la prima sinusoide e la seconda sinusoide e sommarle tra loro per creare la stagionalità sinusoidale. Quindi, per creare una stagionalità trimestrale, avremmo un vettore $\boldsymbol{Z}$ del tipo $\begin{bmatrix} \cdots & 1 & 0 & 1 \end{bmatrix}$, poiché sono necessarie due sinusoidi per formare una stagionalità trimestrale.

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

5.5 Inferenza per modelli in forma State-Space

5.5.1 Introduzione

Il **filtro di Kalman** è lo strumento principale per condurre inferenza sulle componenti non osservabili in $\boldsymbol{\alpha}$ e per la costruzione della funzione di verosimiglianza di un modello in forma state-space. Sia $Y_t = \{y_1, \ldots, y_t\}$ la collezione dei dati dalla prima fino alla t-esima osservazione. L'inferenza su $\boldsymbol{\alpha}_t$ sarà necessariamente basata sui dati osservabili e sulla struttura evolutiva definita dall'equazione di transizione. Come noto, la migliore stima di $\boldsymbol{\alpha}_t$, nel senso del minimo errore quadratico medio, basata sui dati Y_s è il valore atteso condizionato $\mathbb{E}(\boldsymbol{\alpha}_t|Y_s)$. A seconda del valore di s essa assume un diverso nome:

- **smoother**: $\boldsymbol{a}_{t|s} = \mathbb{E}(\boldsymbol{\alpha}_t|Y_s)$, con $s > t$, spesso $s = n$
- **filtro**: $\boldsymbol{a}_t = \boldsymbol{a}_{t|t} = \mathbb{E}(\boldsymbol{\alpha}_t|Y_t)$
- **previsore**: $\boldsymbol{a}_{t|s} = \mathbb{E}(\boldsymbol{\alpha}_t|Y_s)$, con $s < t$

A ciascuna stima è associata una matrice di covarianza dell'errore di stima:

$$\boldsymbol{P}_{t|s} = \mathbb{E}[(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|s})(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|s})^T]$$

anche in questo caso si abbrevierà la scrittura quando $s = t$, con $P_t = P_{t|t}$.

Quindi, $\boldsymbol{\alpha}$ rappresenta il vettore delle variabili di stato, contenente tutte le componenti non osservabili. $\boldsymbol{a}_{t|s}$ indica la proiezione di questo vettore, insieme a tutte le sue componenti, basandosi su tutte le informazioni disponibili dai dati dalla prima osservazione fino alla s-esima. In generale, l'obiettivo è prevedere le componenti nel futuro, quindi t deve essere successivo a s. Ad esempio, se si desidera prevedere il trend di una serie storica, che è la prima componente di $\boldsymbol{\alpha}$, è possibile proiettare il trend a 10 mesi avanti, sfruttando tutte le informazioni disponibili fino ad oggi, che è 10 mesi prima della previsione.

Tuttavia, è importante considerare che queste previsioni sono soggette a errori. Non è possibile prevedere il segnale stocastico in modo esatto. Quindi il vettore delle previsioni, $\boldsymbol{a}_{t|s}$, tiene conto di questo errore e può includere previsioni sia per il futuro che per il passato.

Questa operazione non può essere eseguita senza errore, poiché si sta cercando di inferire le componenti osservando la serie storica, ma queste sono solo stime e non si può mai conoscere esattamente il vero livello del trend, ad esempio.

La matrice $\boldsymbol{P}_{t|s}$ rappresenta l'errore di previsione, calcolato come la differenza tra il vero valore $\boldsymbol{\alpha}_t$ e la sua previsione ($\boldsymbol{a}_{t|s}$), trasposta moltiplicata per se stessa. Sia il filtro di Kalman che lo smoother calcolano questa coppia di valori.

Il filtro di Kalman opera in tempo reale, quindi considera situazioni in cui t e s sono uguali o dove t è al massimo un passo avanti rispetto a s. Al contrario, lo smoother lavora quando s è nel futuro rispetto a t.

Entrambi, filtro di **Kalman** e **smoother**, sono algoritmi per calcolare queste proiezioni e i relativi errori.

5.5.2 Il filtro di Kalman

Il filtro di Kalman è un algoritmo ricorsivo che permette di calcolare $\boldsymbol{a}_t$ e P_t partendo da $\boldsymbol{a}_{t|t-1}$ e $\boldsymbol{P}_{t|t-1}$ e viceversa. Come si è detto durante la trattazione della forma state-space, il modello è completato dalla definizione della media e dalla matrice di covarianza di $\boldsymbol{\alpha}_0$, della cui determinazione si parlerà più avanti. Questi due valori serviranno a inizializzare l'iterazione dell'algoritmo. Vediamo ora le diverse fasi del calcolo del filtro.

Equazione di previsione

$$\boldsymbol{a}_{t|t-1} = \mathbb{E}(\boldsymbol{\alpha}_t|Y_{t-1}) = \mathbb{E}(\boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t|Y_{t-1}) = \boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t$$

$$\begin{aligned}\boldsymbol{P}_{t|t-1} &= \mathbb{E}[(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|s})(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|s})^T] \\ &= \mathbb{E}[(\boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t - \boldsymbol{c}_t)(\boldsymbol{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{c}_t + \boldsymbol{R}_t\boldsymbol{\eta}_t - \boldsymbol{c}_t)^T] \\ &= \boldsymbol{T}_t\mathbb{E}[(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_{t-1})(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_{t-1})^T]\boldsymbol{T}_t^T + \boldsymbol{R}_t\mathbb{E}[\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T]\boldsymbol{R}_t^T \\ &= \boldsymbol{T}_t\boldsymbol{P}_{t-1}\boldsymbol{T}_t + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t^T\end{aligned}$$

Innovazioni La previsione dell'osservazione y_t basata sui dati Y_{t-1} è data da

$$\hat{y}_{t|t-1} = \mathbb{E}[y_t|Y_{t-1}] = \mathbb{E}[\boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{d}_t + \varepsilon_t|Y_{t-1}] = \boldsymbol{Z}_t\boldsymbol{a}_{t|t-1} + \boldsymbol{d}_t$$

e l'errore di previsione, o **innovazione**, è quindi

$$\boldsymbol{v}_t = y_t - \hat{y}_{t|t-1} = \boldsymbol{Z}_t(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}) + \varepsilon_t$$

con varianza

$$\boldsymbol{F}_t = \mathbb{E}\{[\boldsymbol{Z}_t(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}) + \varepsilon_t][\boldsymbol{Z}_t(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}) + \varepsilon_t]^T\} = \boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T + \boldsymbol{H}_t$$

Equazioni di aggiornamento Per ricavare le equazioni per $\boldsymbol{a}_t$ e $\boldsymbol{P}_t$ si noti che si è supposto che il sistema sia **normale** e che, pertanto,

$$\begin{bmatrix}\boldsymbol{\alpha}_t \\ y_t\end{bmatrix}|Y_{t-1} \sim N\left(\begin{bmatrix}\boldsymbol{a}_{t|t-1} \\ \hat{y}_{t-1}\end{bmatrix}, \begin{bmatrix}\boldsymbol{P}_{t|t-1} & \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T \\ \boldsymbol{Z}\boldsymbol{P}_{t|t-1} & \boldsymbol{F}_t\end{bmatrix}\right)$$

dove gli elementi sulla diagonale secondaria della matrice di covarianza sono stati ottenuti da

$$Cov(\boldsymbol{a}_t, y_t|Y_{t-1}) = \mathbb{E}\{[\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}][\boldsymbol{Z}_t(\boldsymbol{\alpha}_t - \alpha_{t|t-1}) + \varepsilon_t]^T\} = \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T$$

e dalla sua trasposta. A questo punto la distribuzione di $\boldsymbol{\alpha}_t|Y_t$ si ottiene condizionando $\boldsymbol{\alpha}_t|Y_{t-1}$ a $y_t|Y_{t-1}$, cioè applicando la Proprietà 1. $\boldsymbol{\alpha}_t|Y_t$ è pertanto, distribuito normalmente con vettore di medie e matrice di covarianza, rispettivamente,

$$\begin{aligned}\boldsymbol{a}_t &= \boldsymbol{a}_{t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T\boldsymbol{F}_t^{-1}\boldsymbol{v}_t \\ \boldsymbol{P}_t &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{Z}_t^T\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\boldsymbol{P}_{t|t-1}\end{aligned}$$

Smoother Per calcolare lo smoother è necessario "passare" il filtro di Kalman e immagazzinare $\boldsymbol{\alpha}_t$ e $\boldsymbol{P}_t$ per $t = 1, \ldots, n$. Lo smoother è dato dalle iterazioni

$$\begin{aligned}\boldsymbol{a}_{t|n} &= \boldsymbol{a}_t + \boldsymbol{O}_t(\boldsymbol{a}_{t+1|n} - \boldsymbol{T}_{t+1}\boldsymbol{a}_t - \boldsymbol{c}_{t+1}) \\ \boldsymbol{P}_{t|n} &= \boldsymbol{P}_t + \boldsymbol{O}_t(\boldsymbol{P}_{t+1|n} - \boldsymbol{P}_{t+1|t})\boldsymbol{O}_t^T\end{aligned}$$

per $t = n, n-1, \ldots, 0$ con $\boldsymbol{O}_t = \boldsymbol{P}_t\boldsymbol{T}_{t+1}^T\boldsymbol{P}_{t+1}^{-1}$. La procedura per ricavare lo smoother è più complicata rispetto al calcolo del filtro e pertanto la si omette.

Si noti che, quando il sistema non è gaussiano, il filtro di Kalman fornisce le migliori stime lineari, ma non fornendo più valori attesi, ma solo proiezioni lineari, esistono stimatori non lineari nei dati, che possono essere più precisi.

5.5.3 Inizializzazione del Filtro

La definizione di un sistema state-space è completata dal vettore di medie $\boldsymbol{a}_0$ e dalla matrice di covarianza $\boldsymbol{P}_0$ del vettore di stato iniziale $\boldsymbol{\alpha}_0$.

In genere è possibile utilizzare l'informazione a priori nota sulle serie da analizzare per potere definire medie e varianza delle componenti in $\boldsymbol{\alpha}_0$. Quando si è molto incerti sul valore iniziale atteso di un vettore di stato è sempre possibile porre la relativa varianza su valori molto alti, dichiarando quindi una sostanziale ignoranza.

Quando non si hanno informazioni a priori si possono inizializzare le componenti non stazionarie (per esempio trend e stagionalità in un UCM) per mezzo di distribuzioni diffuse (a varianza infinita), mentre le componenti stazionarie possono essere inizializzate per mezzo della distribuzione marginale. Si supponga che $\boldsymbol{\alpha}_t$ contenga solo componenti stazionarie, allora la distribuzione marginale di $\boldsymbol{\alpha}_t$ è normale con media

$$\mathbb{E}(\boldsymbol{\alpha}_t) = \boldsymbol{T}\mathbb{E}(\boldsymbol{\alpha}_{t-1}) + \boldsymbol{c} + \boldsymbol{R}\mathbb{E}(\boldsymbol{\eta}_t)$$

$$\boldsymbol{\alpha}_0 = \boldsymbol{T}\boldsymbol{\alpha}_0 + \boldsymbol{c}$$

$$\boldsymbol{\alpha}_0 = (\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{c}$$

e matrice di covarianza data dalla soluzione di

$$\mathbb{E}[(\boldsymbol{\alpha}_t - \boldsymbol{a}_0)(\boldsymbol{\alpha}_t - \boldsymbol{a}_0)^T] = \mathbb{E}\{[\boldsymbol{T}(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_0) + \boldsymbol{R}\boldsymbol{\eta}_t][\boldsymbol{T}(\boldsymbol{\alpha}_{t-1} - \boldsymbol{a}_0) + \boldsymbol{R}\boldsymbol{\eta}_t]^T\}$$

$$\boldsymbol{P}_0 = \boldsymbol{T}\boldsymbol{P}_0\boldsymbol{T}^T + \boldsymbol{R}\boldsymbol{Q}\boldsymbol{R}^T$$

che può essere ottenuta per mezzo di

$$vec(\boldsymbol{P}_0) = (\boldsymbol{I} - \boldsymbol{T} \otimes \boldsymbol{T})^{-1} vec(\boldsymbol{R}\boldsymbol{Q}\boldsymbol{R}^T)$$

dove *vec* è l'operatore vettorizzazione che incolonna i vettori colonna di una matrice, e $\otimes$ è il prodotto di Kronecker. Il risultato appena mostrato è dovuto alla proprietà

$$vec(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = (\boldsymbol{B}^T \otimes \boldsymbol{A}) vec(\boldsymbol{X})$$

Quando vi fossero variabili sia stazionarie sia non-stazionarie nel vettore di stato è possibile utilizzare la regola appena vista per il sotto vettore di variabili stazionarie e distribuzioni diffuse per le altre.

5.5.4 Stime di Massima Verosimiglianza di un Modello in Forma State-Space

Data la gaussianità del sistema e la linearità delle equazioni che definiscono il filtro di Kalman, il vettore delle innovazioni $\boldsymbol{v}_t$ è normale con media zero e matrice di covarianza $\boldsymbol{F}_t$. Pertanto, la funzione di **log-verosimiglianza** è data da

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\left\{kn \log 2\pi - \sum_{t=1}^{n} \log\det(\boldsymbol{F}_t) - \sum_{t=1}^{n} \boldsymbol{v}_t^T \boldsymbol{F}_t^{-1} \boldsymbol{v}_t\right\}$$

dove $\boldsymbol{\theta}$ è un vettore contenente tutti i parametri ignoti del modello. La log-verosimiglianza può essere massimizzata per mezzo di metodi numerici.

5.5.5 Funzionamento Pratico del Filtro di Kalman

Filtro di Kalman

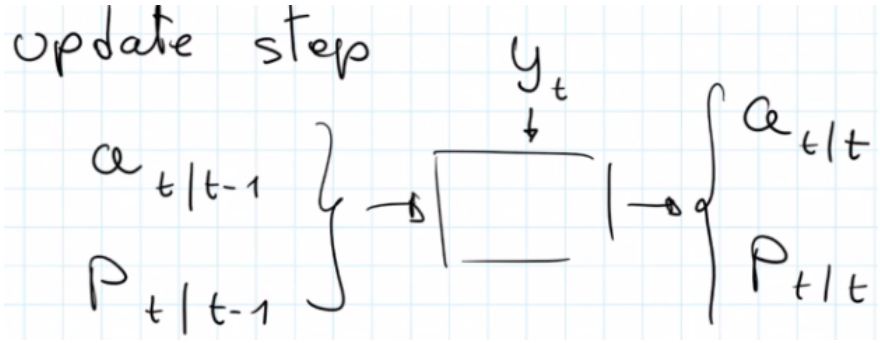
Vediamo ora come funzionamento pratico del filtro di Kalman. Come prima cosa viene definito l'**inizializzazione**. Viene inizializzato il vettore di stato con media sconosciuta (supposta gaussianità) e si predice $\boldsymbol{\alpha}_1$:

$$\boldsymbol{\alpha}_1 \sim (\boldsymbol{a}_{1|0}, \boldsymbol{P}_{1|0})$$

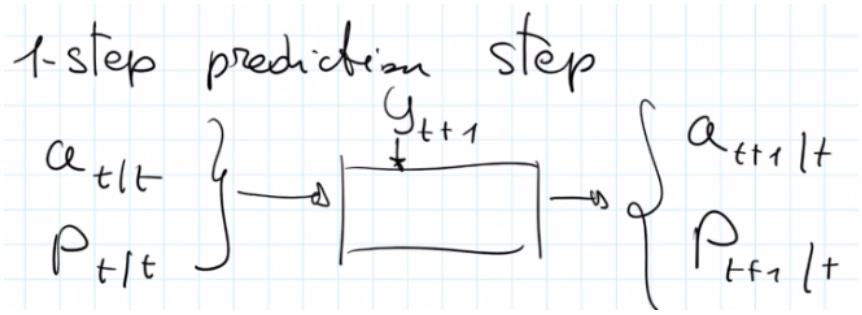
$\boldsymbol{a}_{1|0}$ è il valore atteso di $\boldsymbol{\alpha}_1$ non avendo osservato nessun dato e la varianza dell'errore prevedendo $\boldsymbol{\alpha}_1$ tramite $\boldsymbol{a}_{1|0}$. Semplicemente si specifica il valore atteso e la varianza iniziale.

Il filtro di Kalman è composto da due step:

1. **Update step**: prende $\boldsymbol{a}_{t|t-1}$ e $\boldsymbol{P}_{t|t-1}$ e il nuovo dato y_t e restituisce $\boldsymbol{a}_{t|t}$ e $\boldsymbol{P}_{t|t}$. Ovvero prende le previsioni un passo in avanti, osserva un nuovo dato e le trasforma nel filtro.



   Si tratta quindi di un algoritmo che date le previsioni un passo in avanti e i relativi errori, osserva un dato e aggiorna queste previsioni usando il dato al tempo t per avere una migliore stima di quello che sta dentro ad $\boldsymbol{\alpha}_t$.
2. **Prediction step**: operazione inversa rispetto all'update step. Dato $\boldsymbol{a}_{t|t}$ e $\boldsymbol{P}_{t|t}$, osserva il dato y_{t+1} e restituisce $\boldsymbol{a}_{t+1|t}$ e $\boldsymbol{P}_{t+1|t}$. Ovvero prende il filtro, osserva un nuovo dato e crea una nuova previsione un passo in avanti.



In pratica prima viene inizializzato l'algoritmo (imponendo le condizioni iniziali si inizializza l'algoritmo), poi si fa l'update step, poi il prediction step, poi si aggiorna il valore di t, e poi si ritorna all'update step e così via.

Il filtro di Kalman crea e restituisce una sequenza di previsioni un passo in avanti (sempre con il loro errore e la varianza del loro errore) e ne fornisce anche il filtro. L'algoritmo lavora in tempo reale, ci dà sempre la stima migliore di quello che succede in $\boldsymbol{\alpha}_t$ conoscendo tutte le osservazioni fino al tempo t e anche la previsione di quello che succederà un passo in avanti.

$$(\boldsymbol{a}_{t|t-1}, \boldsymbol{P}_{t|t-1})_{t=1,\dots,n}$$

$$(\boldsymbol{a}_{t|t}, \boldsymbol{P}_{t|t})_{t=1,\dots,n}$$

L'algoritmo è molto efficiente computazionalmente perché memorizza in due oggetti $(\boldsymbol{a}_t, \boldsymbol{P}_t)$ tutte le informazioni utili della serie storica. Questo si chiama **streaming**, nel filtro è già presente tutta la storia del dato di interesse (di tutto il resto, come il passato di y non serve più); quando arriverà il nuovo dato, si aggiorna e si tengono solo le nuove stime.

Il filtro di Kalman può essere utilizzato per costruire la funzione di verosimiglianza gaussiana. Nell'update step si è sempre parlato di previsioni un passo in avanti di $\boldsymbol{a}_t$ e $\boldsymbol{P}_t$, ma è possibile calcolare anche le previsioni un passo in avanti per y.

$$\hat{y}_{t|t-1} = \boldsymbol{Z}_t \boldsymbol{a}_{t|t-1}$$

$\hat{y}_{t|t-1}$ sono le previsioni 1-step della y al tempo t basate sul passato della serie storica fino al tempo $t-1$.

Le previsioni un passo in avanti dei miei dati non sono altro che la $\boldsymbol{Z}$ (serve per prendere il vettore di stato e distribuire le sue componenti, sommarle o moltiplicarle). Le previsioni un passo in avanti della $\boldsymbol{Z}$ non sono altro che le previsioni un passo in avanti della t pre-moltiplicate per la matrice $\boldsymbol{Z}$. Si può far vedere che l'errore di previsione $\boldsymbol{F}$ che è dato dalla vera y meno la sua previsione, moltiplicata per sé stessa trasposta.

$$\boldsymbol{F}_t = \mathbb{E}[(y_t - \hat{y}_{t|t-1})(y_t - \hat{y}_{t|t-1})^T]$$

Tramite il filtro di Kalman si calcolano gratis le previsioni un passo in avanti e la varianza dell'errore delle previsioni un passo in avanti.

Se si suppone che i dati sono gaussiani, allora anche la y è gaussiana, quindi $y_{t|t-1}$ (la distribuzione di y condizionata al passato della serie storica) è una normale con media $\hat{y}_{t|t-1}$ e varianza $\boldsymbol{F}_t$.

$$y_{t|t-1} \sim N(\hat{y}_{t|t-1}, \boldsymbol{F}_t)$$

Ovvero, supponendo gaussianità la previsione un passo in avanti e la varianza sono media e varianza di una normale.

Questo consente di scrivere la funzione di **log-verosimiglianza**. Funzione di verosimiglianza è densità di probabilità dei nostri dati vista come funzione dei parametri ignoti $\boldsymbol{\theta}$. Stima di massima verosimiglianza cerca di massimizzare questa funzione rispetto a $\boldsymbol{\theta}$, è come se si mettesse nella situazione di aver osservato i dati più probabili possibili rispetto al modello.

$$l(\boldsymbol{\theta}) = f_\theta(y_1, \dots, y_n)$$

La funzione massima verosimiglianza va interpretata nel seguente modo: si osservano dei dati in ogni modello, nel modello non si conoscono i parametri θ_i e si cerca il valore dei vari θ_i che rende più probabili possibile le osservazioni dei dati che sono state osservate.

Quando si fa statistica per campioni bernoulliani estratti indipendentemente dalla stessa distribuzione, la funzione di log-verosimiglianza può essere fattorizzata come il prodotto delle marginali. Tuttavia questo non può essere fatto per le serie storiche, perché *le osservazioni non sono indipendenti*. La dipendenza delle ossecrazioni di fatto è quello che permette di fare le previsioni delle serie storiche. Quindi la congiunta non è più data dal prodotto delle marginali, se i dati non sono indipendenti.

È comunque possibile fattorizzare le serie storiche. La congiunta è data dal prodotto delle condizionate, dove solo la prima è marginale.

$$l(\boldsymbol{\theta}) = f_{1,\theta}(y_1) \cdot f_{2,\theta}(y_2|y_1) \cdot f_{3,\theta}(y_3|y_1, y_2) \cdot \ldots \cdot f_{n,\theta}(y_n|y_1, y_2, \ldots, y_{n-1})$$

Dal risultato del filtro di Kalman $y_{t|t-1} \sim N(\hat{y}_{t|t-1}, \boldsymbol{F}_t)$, è possibile costruire la funzione di verosimiglianza gaussiana.

$$f_{t,\theta}(y_t|Y_{t-1}) = N(\hat{y}_{t|t-1}, \boldsymbol{F}_t)$$

dove come nel capitolo precedente $Y_{t-1} = y_1, \ldots, y_{n-1}$. Il filtro di Kalman permette di calcolare il generico elemento di questa produttoria.

Il filtro di Kalman permette di fare inferenze in contemporanea sulle componenti restituendo la stima in tempo reale di $\boldsymbol{\alpha}_t$ e la varianza dell'errore, permette inoltre di costruire tramite questi due oggetti (cioè le previsioni un passo in avanti e la varianza dell'errore di previsione) la funzione di log-verosimiglianza. Il filtro di Kalman in sostanza calcola $\boldsymbol{a}_{t|t-1}$, $\boldsymbol{P}_{t|t-1}$, $\boldsymbol{a}_t$ e $\boldsymbol{P}_t$ abbinati a $\hat{y}_{t|t-1}$ e $\boldsymbol{F}_{t|t-1}$. Alcune implementazioni del filtro di Kalman uniscono i due passiaggi.

Smoother

Lo smoother è un algoritmo per calcolare $\boldsymbol{a}_{t|n}$ e $\boldsymbol{P}_{t|n}$. In generale lo smoother è la situazione in cui s è nel futuro rispetto a t, in pratica l'unico smoother che di interesse è quello in cui s è l'ultima osservazione disponibile e quindi equivale ad n.

La proiezione delle componenti non osservabili del vettore di stato su tutti i dati restituisce una miglior stima della componente non osservabile di interesse. Ponendo $t = n$ si ottiene t, ovvero: all'ultimo dato lo smoother e il filtro di Kalman coincidono poiché non ho osservazioni future.

Da un punto di vista pratico per calcolare lo smoother si necessita di tutto il filtro di Kalman. Lo smoother è un algoritmo che lavora dall'ultima osservazione (il filtro) e torna indietro. Il filtro viene passato in avanti, lo smoother, al contrario, viene passato dall'ultimo dato al primo. Ovviamente sono presenti anche tutti i predittori, che sono poi previsioni.

$$\text{smoother} \begin{cases} \boldsymbol{a}_{t|n} & a_{1|1} \cdots a_{n|n} \\ \boldsymbol{P}_{t|n} & a_{1|n} \cdots \end{cases} \quad \text{previsioni} \begin{cases} \boldsymbol{a}_{n+k|n} & \hat{y}_{n+k|n} \\ \boldsymbol{P}_{n+k|n} & \hat{\boldsymbol{F}}_{n+k|n} \cdots \end{cases}$$

Il filtro restituisce le stime in tempo reale o un passo in avanti, lo smoother restituisce le stime ritardate perché basate anche su dati futuri e le previsioni restituiscono le stime del futuro delle componenti e del loro valore.

Dati Mancanti

Nel caso di **valori mancanti**, ovvero se manca y_t nell'update step (manca tra i vari input), si mette un uguale e rimane coincidente con le previsioni un passo in avanti, cosi facendo tutte le formule restano uguali (ovviamente questo aumenta l'incertezza e le previsioni al tempo $t+1$ saranno in realtà quelle al tempo $t+2$).

Lo smoother fa il lavoro da indietro in avanti, questo migliora la previsione del dato mancante usando anche l'altro verso, ovvero non solo usando dati passati ma anche quelli futuri. L'imputazione del dato mancante (y_s) diventa stima dello smoother dove viene preso $\boldsymbol{Z}$, e $\boldsymbol{Z}_s$ e lo va a sostituire con lo smoother trasformato linearmente.

$$\hat{y}_s = \boldsymbol{Z}_s \boldsymbol{a}_{s|n} = \mathbb{E}(y_s|Y_n)$$

Si tratta di una proiezione di y_t basata su tutti i dati tranne s.

Chapter 6

Distrubance Smoother

6.1 Distrubance Smoother per gli Outliers

6.1.1 Introduzione

Il **distrubance smoother** assume che una serie temporale $\{Y_1, \cdots, Y_n\}$ sia generata da un modello in forma state space e che il filtro di Kalman sia stato eseguito.

Il miglior modo di identificare gli *outlier* di diversa natura sono le quantità calcolate attraverso il distrubance smoother. Attraverso l'utilizzo dei così datti *residui ausiliari*, che altro non sono che particolari t-statistiche, è possibile verificare la presenza di diversi problemi.

Si può verificare la presenza di un "additive outlier" nella serie storica, ovvero valori anomali della y. Oppure valori anomali per le singole componenti. In particolare:

- level μ_t: l'outlier è un **level shift**
- slope β_t l'outlier è un **slope shift**
- stagionalità γ_t: l'outlier è un cambio nel pattern di stagionalità e potrebbe influenzare la sere storica per più di un periodi di tempo
- ciclo ψ_t: l'outlier può essere interpretato come uno shock eccezionale che influenza il ciclo

In ognuno di questi casi il metodo più semplice per adattare il modello UCM a trovare gli outlier è introducendo variabili dummy moltiplicate per il coefficiente da stimare nell'equazione dove sono stati rivelati gli outlier. Di seguito si pone particolare attenzione sugli addative outlier, level shift e slope shift.

Consideriamo il modello $y_t = \mu_t + \varepsilon_t$, dove μ_t rappresenta il trend e ε_t il rumore. Questo modello può essere espresso tramite un Local Linear Trend (LLT) più il rumore, definito come:

$$\underbrace{\begin{bmatrix}\mu_{t+1}\\ \beta_{t+1}\end{bmatrix}}_{\boldsymbol{\alpha}_{t+1}} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mu_t\\ \beta_t\end{bmatrix} + \underbrace{\begin{bmatrix}\eta_t & \zeta_t\end{bmatrix}}_{\boldsymbol{\eta}_t}$$

Dove η_t e ζ_t rappresentano il rumore.

Il processo di smoothing fornisce la migliore previsione lineare per α_{t+1} utilizzando tutte le osservazioni disponibili:

$$\boldsymbol{a}_{t|n} = \mathbb{P}(\boldsymbol{\alpha}_t | y_1, \ldots, y_n)$$

Analogamente, si può calcolare la matrice di covarianza associata:

$$\boldsymbol{P}_{t|n} = \mathbb{E}\{[\boldsymbol{\alpha}_t - \mathbb{P}(\boldsymbol{\alpha}_t|y_1, \ldots, y_n)][\boldsymbol{\alpha}_t - \mathbb{P}(\boldsymbol{\alpha}_t|y_1, \ldots, y_n)]^T\}$$

Ora, si vuole stimare i termini di rumore ε_t, η_t, e ζ_t per fare inferenza. Questo è utile quando si osserva un valore estremo (outlier) in y_t, che potrebbe indicare un'anomalia. Ad esempio se si monitorano dei dati sul il volume di acqua in un fiume, e si nota un anno in cui il valore misurato è estremo rispetto agli altri (basso o alto). Nel caso si verifichi uno shock estremo sul valore di η, questo implicherebbe un improvviso aumento o diminuzione nel livello μ in base al segno di η. Ovvero, un η molto grande implicherebbe che il livello sale, mentre un η molto basso (negativo) implicherebbe che il livello scende. Per quanto riguarda invece un valore estremo per ζ, questo indica un cambio nella slope β. Un balzo nella slope significa che magari la serie storica stava salendo e poi immediatamente scende, o magari punta ancora più su.

Ovviamente μ e β continuano a cambiare ogni osservazione, ma se η e ζ sono estremi, questo provoca cambi estremi da un osservazione all'altra.

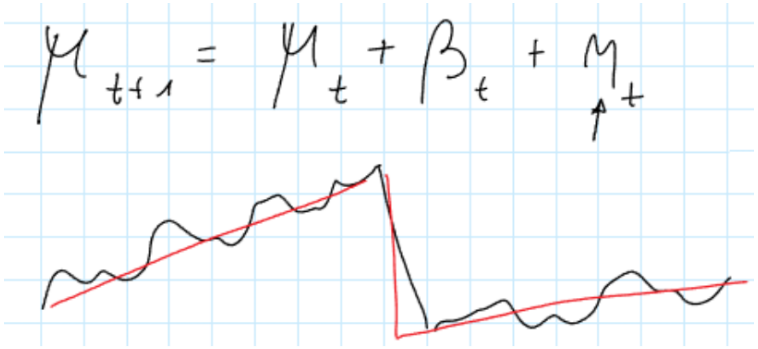


Figure 6.1: Level Shift.

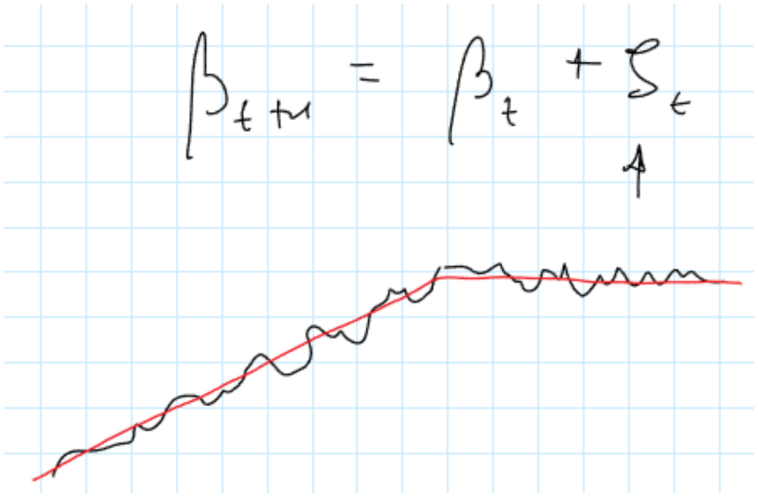


Figure 6.2: Slope Shift

6.1.2 Residui Ausiliari

Questo tipo di analisi, noto come **disturbance smoother**, consente di identificare e interpretare tali cambiamenti, viene indicato come:

$$\hat{\varepsilon}_{t|n} = \mathbb{P}(\varepsilon|y_1, \ldots, y_n)$$

e rappresenta la proiezione di ε su tutti i dati.

Si può anche considerare la proiezione di tutto $\boldsymbol{\eta}_t$ su tutte le osservazioni:

$$\hat{\boldsymbol{\eta}}_{t|n} = \mathbb{P}(\boldsymbol{\eta}|y_1, \ldots, y_n)$$

In particolare, $\hat{\varepsilon}_{t|n}$ rappresenta la proiezione di ε utilizzando tutte le osservazioni disponibili. Può essere usato per individuare eventuali **additive outliers** (AO). Allo stesso modo, $\hat{\boldsymbol{\eta}}_{t|n}$ rappresenta la proiezione di $\boldsymbol{\eta}_t$ su tutte le osservazioni. Può essere usato per individuare altri cambiamenti strutturali come level shift (LS) o slope shift (SS).

$\hat{\boldsymbol{\eta}}_{t|n}$ può essere visto come uno stimatore della vera (non sconosciuta) $\boldsymbol{\eta}_t$, ovvero del vero rumore. In questo caso si può calcolare il MSE e $\hat{\boldsymbol{\eta}_{t|n}}$ come stimatore di $\boldsymbol{\eta}_t$.

$$V_\eta = \mathbb{E}[(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{t|n})(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{t|n})^T]$$

Si può anche vedere $\hat{\boldsymbol{\eta}}_{t|n}$ come una variabile casuale con media zero, da cui

$$Var(\hat{\boldsymbol{\eta}}_{t|n}) = \mathbb{E}(\hat{\boldsymbol{\eta}}^2_{t|n})$$

Questo è di particolare interesse perché se si vuole comparare la stima fatta con la gaussiana, per poi identificare un osservazione estrema. L'utilizzo di questo disturbance smoother identifica solo movimenti estremi che risultano outlier o LS o SS e cosi via.

Per trovare se un valore è estremo serve un benchmark di comparazione. Come detto precedentemente, se il sistema è gaussiano anche lo smother lo è, quindi solitamente si procede costruendo quello che viene chiamato **auxiliary residuals** (residui ausiliari).

$$AR_\eta = \frac{\hat{\boldsymbol{\eta}}_{t|n}}{sd(\hat{\boldsymbol{\eta}}_{t|n})} \sim N(0,1)$$

I residui ausiliari essendo già normalizzati possono essere confrontati con i percentili della gaussiana e quindi è possibile capire quando un valore è particolarmente estremo rispetto a un altro. Se si identifica un valore fuori dalle linee di confidenza (95%), se è in η allora identifica un salto nella level component, se è la slope allora una salto nella slope component, se y allora è solo un oulier.

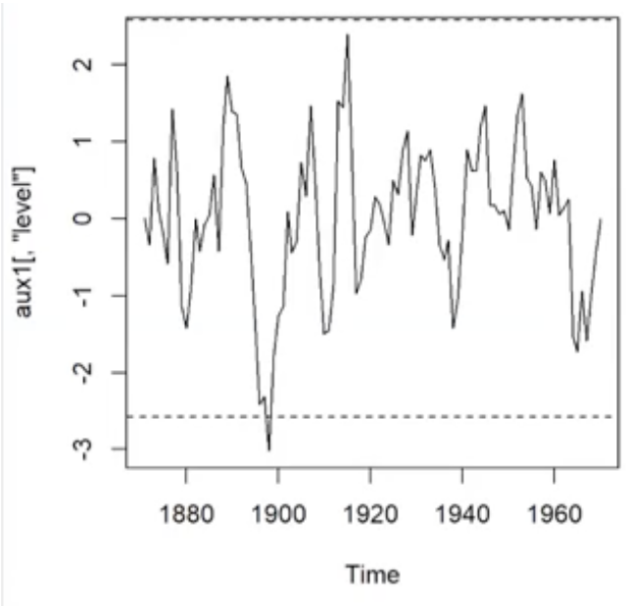


Figure 6.3: Utilizzo di residui ausiliari per outlier.

6.1.3 Variabili Dummy

Si è osservato che il "disturbance smother" rappresenta l'errore di osservazione o il disturbo dell'equazione di stato proiettato su tutte le osservazioni. Ciò equivale a inferire sugli shock che influenzano una serie temporale. Pertanto, oltre a trarre inferenze sulle variabili di stato,

si può fare altrettanto sugli shock. Questo può essere utile per identificare shock di notevole ampiezza che influenzano la serie temporale. Quando si individuano movimenti estremi in questa quantità, si verificano gli "additive outlier" (outlier additivi), ossia valori anomali che non alterano strutturalmente l'intera serie storica, ma solo un singolo valore.

Se lo shock riguarda l'errore di osservazione ε_t (identificato tramite i residui ausiliari $\frac{\hat{\varepsilon}_{t|n}}{se(\hat{\varepsilon}_{t|n})}$), ciò implica la presenza di outlier additivi (AO). Per individuare tali AO, è possibile utilizzare variabili dummy. In questo contesto, si utilizza una variabile *pulse*, che è sempre uguale a 0 tranne al tempo τ (in cui si è registrato un valore particolarmente estremo), assumendo il valore 1 e quindi ritornando a 0. Ciò rappresenta un cambiamento momentaneo e temporaneo di livello.

$$D_t^\tau = \begin{array}{ccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array}$$

$$y_t = \cdots + \delta D_t^\tau + \varepsilon_t$$

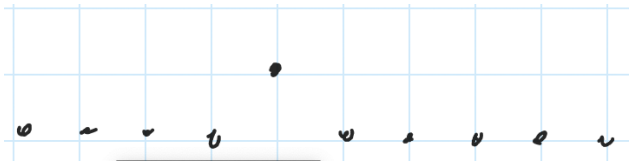


Figure 6.4: Pulse.

Si può incorporare il regressore nella forma state space, quindi introdurre un coefficiente δ che moltiplica la variabile dummy D_t^τ per il tempo τ, così da catturare il valore dell'outlier additivo. In alternativa, si potrebbe escludere l'osservazione, ponendo il valore di quell'osservazione a zero, senza che la forma state space ne risenta. Scegliere una delle due opzioni è del tutto equivalente. Aggiungendo una dummy e utilizzando il coefficiente δ, si ottiene una stima di quanto sia significativo quell'outlier (l'introduzione di una dummy di questo tipo consente di valutare l'impatto dell'evento; il logaritmo del valore della dummy indica la variazione percentuale relativa dell'effetto dell'evento dovuto all'evento stesso).

Se lo shock estremo coinvolgesse il disturbo dell'equazione di livello (identificato tramite i residui ausiliari $\frac{\hat{\eta}_{t|n}}{se(\hat{\eta}_{t|n})}$), ciò implica la presenza di LS. Si potrebbe utilizzare una variabile *step* che rimane costantemente a 0 e diventa 1 al tempo τ.

$$S_t^\tau = \begin{array}{ccccccc} 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array}$$

$$y_t = \cdots + \delta S_t^\tau + \varepsilon_t$$

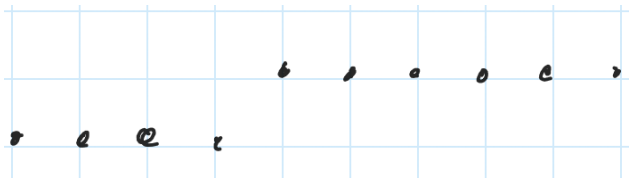


Figure 6.5: Step.

Per quanto riguarda gli shock estremi sulla perturbazione della pendenza, ciò implica la presenza di SS. In questo contesto, sarebbe necessario utilizzare una variabile *ramp*, che inizia da 0 e diventa una bisettrice a partire dal momento τ.

$$R_t^\tau = \begin{cases} 0 & t < \tau - 1 \\ t - \tau & t \geq \tau - 1 \end{cases}$$

$$y_t = \cdots + \delta R_t^\tau + \varepsilon_t$$

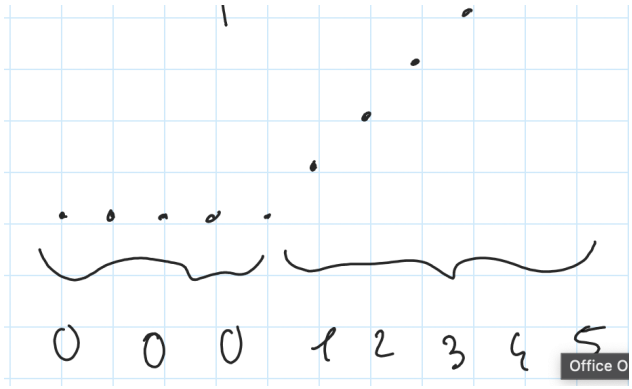


Figure 6.6: Ramp.

Immaginiamo di avere un LS rappresentato dall'equazione:

$$\mu_{t+1} = \mu_t + \beta_t + \eta_t$$

Quello che succede è che la varianza di η cambia in base al tempo:

$$Var(\eta) = \begin{cases} \sigma_\eta^2 & t \neq \tau \\ \sigma_\tau^2 & t = \tau \end{cases}$$

Questo significa che al momento dello shock, η_t avrà una varianza diversa. Questo rappresenta un modo alternativo per ottenere lo stesso effetto semplicemente modificando la varianza anziché la pendenza.

6.2 Esempio Campagna Pubblicitaria

Si consideri una campagna pubblicitaria che incide sulle vendite, con dati raccolti su base settimanale. Poiché i dati sono settimanali, non è possibile determinare esattamente quale giorno della settimana ha avuto maggiore esposizione pubblicitaria, né se questa sia stata uniformemente distribuita durante la settimana. Inoltre, c'è un ritardo nel processo pubblicitario poiché l'acquisto avviene probabilmente in un momento successivo (giorni o settimane) rispetto all'esposizione alla pubblicità. La relazione tra le vendite e la pubblicità può essere descritta da un modello come:

$$sales_t = \omega_0 adv_t + \omega_1 adv_{t-1} + \omega_2 adv_{t-2} + \dots$$

dove adv_t rappresenta l'effetto della pubblicità al tempo t e $\omega_0, \omega_1, \omega_2, \dots$ sono coefficienti che rappresentano l'effetto della pubblicità nei tempi precedenti. Tuttavia, poiché l'effetto della pubblicità probabilmente non è istantaneo ma dura nel tempo, ciò porta ad un gran numero di coefficienti, il che potrebbe essere problematico dal punto di vista statistico. Per affrontare questo problema, si usa un modo più parsimonioso per modellare l'effetto della pubblicità sulle vendite.

Questa serie sembra essere un filtro $MA(q)$ applicato alla variabile di pubblicità:

$$(\omega_0 + \omega_1 B + \omega_2 B^2 + \dots + \omega_q B^q) adv_t$$

In questo caso, anche ω_0 deve essere stimato insieme agli altri coefficienti. La memoria del filtro MA è breve e arriva solo fino a q ritardi.

D'altra parte, se consideriamo un filtro $AR(1)$:

$$sales_t = \omega_0 adv_t + \delta sales_{t-1}$$

dove $\omega_0 adv_t$ rappresenta lo shock e anziché includere molti ritardi della variabile di vendita, viene incluso solo un ritardo della variabile di vendita. Questo implica una memoria molto più lunga. Se $|\delta| < 1$, allora il filtro è stabile.

Rimpiazzando il lato destro dell'equazione con il lato sinistro in modo ricorsivo:

$$sales_t = \omega_0 \sum_{i=1}^{\infty} \delta^i adv_{t-i} = \omega_0 adv_t + \omega_0 \delta adv_{t-1} + \omega_0 \delta^2 adv_{t-2} + \dots$$

Questo assomiglia a un filtro $MA(\infty)$.

Nella prima formulazione, sarebbe stato necessario stimare molti parametri, il che avrebbe potuto generare rumore e overfitting. La nuova formulazione riduce il numero di parametri

da stimare a soli 2. Un altro effetto della nuova formulazione è che l'effetto della pubblicità diminuisce progressivamente nel tempo fino a essere dimenticato. Questo concetto è conosciuto come "impulse response function", che indica che l'effetto iniziale sulle vendite è di ω_0, ma dopo una settimana l'effetto è ancora positivo ma più piccolo, pari a $\omega_0\delta$, dove δ, per esempio, potrebbe essere 0.9.

Quello descritto finora è solo un filtro di tipo $AR(1)$, ma è possibile avere filtri più complessi. Ad esempio, se si desidera una risposta oscillante alla pubblicità, potremmo utilizzare un filtro $AR(2)$:

$$sales_t = \omega_0 adv_t + \delta_1 sales_{t-1} + \delta_2 sales_{t-2}$$

L'equazione caratteristica associata sarebbe:

$$(1 - \delta_1 B - \delta_2 B^2) sales = \omega_0 adv_t$$

Chapter 7

Splines

7.1 Spline, spline cubico e smoothing spline

Gli **splines** sono una tecnica versatile che può essere applicata a una vasta gamma di modelli, comprese le serie temporali. In una formulazione generale, consideriamo una funzione in cui la relazione tra la variabile indipendente x e la variabile dipendente y non è lineare e non è conosciuta a priori. Il nostro obiettivo è stimare la variabile dipendente in funzione dei regressori:

$$y_i = f(x_i) + \varepsilon_i$$

Le basi spline consentono di trasformare la variabile indipendente x in modo da approssimare qualsiasi relazione su y con una funzione lineare. L'approccio fondamentale prevede la suddivisione del dominio di x in sotto-intervalli e la posizione di nodi all'interno di tali intervalli. Oltre ai nodi interni, vi sono anche i nodi di confine, corrispondenti al minimo e al massimo del dominio.

Uno spline è una funzione che è polinomiale all'interno di ciascun intervallo ma continua attraverso i nodi.

L'idea è quella di dividere il dominio di x in sotto-intervalli e collocare nodi (ad esempio, due nodi interni). Vi sono anche nodi di confine, corrispondenti ai limiti del dominio.

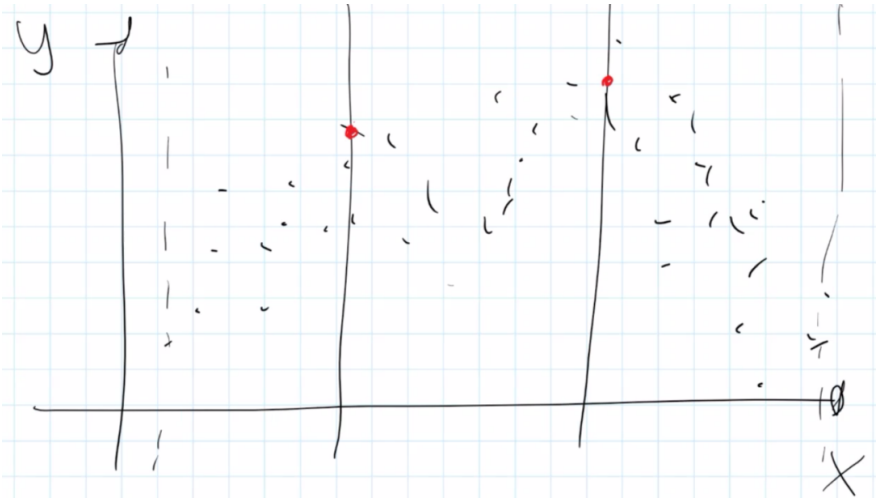


Figure 7.1: Spline con i nodi.

Un'importante variante è lo spline cubico, che approssima la relazione con una funzione cubica in ogni sotto-intervallo, mantenendo la continuità della funzione e delle sue derivate ai nodi. Questo comporta che la funzione cubica in ogni intervallo sia continua con le funzioni cubiche degli intervalli adiacenti.

Ogni spline curve può essere rappresentata come una combinazione lineare di funzioni note:

$$f(x) = \sum_{i=1}^{m} \alpha_i f_i(x)$$

dove m è il numero di nodi e $f_i(x)$ sono funzioni di base conosciute.

Trasformando la variabile indipendente x, si creano altre variabili che si occupano di ciascun sotto-intervallo. Questa tecnica è altamente flessibile: tramite una regressione si crea una relazione non lineare tra x e y utilizzando la regressione lineare. Gli splines permettono di approssimare con una combinazione lineare di queste trasformazioni qualsiasi funzione di x, con il controllo del numero di nodi o gradi di libertà.

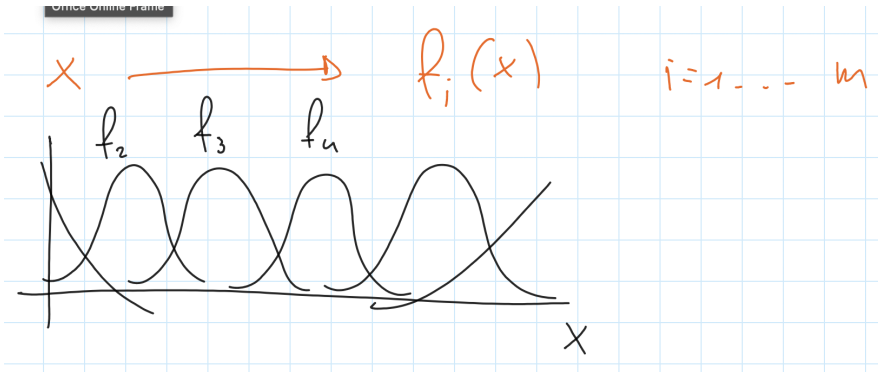


Figure 7.2: Trasformando la x, si creano altre variabili che si occupano ognuna di un sotto-intervallo.

Gli splines possono essere utilizzati per modellare relazioni non lineari, come ad esempio tra temperatura e consumo di energia elettrica, o per catturare relazioni stagionali, ad esempio mediante l'uso di splines periodici, che potrebbero essere utili in casi come la modellazione della stagionalità annuale.

Le spline possono essere implementate come *smoothing splines*, caratterizzate da un nodo per ogni osservazione. L'obiettivo è minimizzare la seguente funzione di costo, che include un parametro di regolarizzazione per prevenire l'overfitting:

$$\sum_{i=1}^{n}(y_i - \alpha_0 - \sum_{j=1}^{n}\alpha_j f_j(x_i))^2 + \lambda \sum_{i=1}^{n}\left|\frac{\partial^2}{\partial x_i}\left(\sum_{j=1}^{n}\alpha_j f_j(x_i)\right)^2\right|$$

Dove y_i sono i valori osservati, α_0 e α_j sono i coefficienti da stimare, $f_j(x_i)$ sono le funzioni di base spline, λ è il parametro di regolarizzazione. Il termine di regolarizzazione, rappresentato dalla somma delle derivate seconde, è progettato per penalizzare la flessibilità eccessiva del modello, riducendo così l'overfitting.

Chapter 8

Machine Learning

8.1 Introduzione

I metodi di machine learning (ML) sono principalmente utilizzati per effettuare previsioni future delle serie temporali. Una delle poche applicazioni del ML sulle serie temporali che non riguarda il forecasting è la ricostruzione di un segnale affetto da rumore. Tuttavia, per fare ciò è necessario aver osservato il segnale in precedenza. Pertanto, l'uso principale del ML nelle serie temporali è il forecasting.

La maggior parte degli algoritmi di ML non è in grado di fare previsioni al di fuori del range dei dati osservati durante il training. Questo è vero soprattutto per gli algoritmi basati sugli alberi, per KNN e per alcune reti neurali. Un aspetto positivo è che questi algoritmi sono in grado di apprendere bene la stagionalità, quindi non è necessario trattare esplicitamente questo problema. Tuttavia, il problema del range rimane, e può essere risolto rendendo la serie temporale additiva e la varianza stazionaria.

Si suppone $\mathbb{E}(y|x) = f(x)$, dove l'obiettivo è stimare la funzione $f(x)$. Per fare ciò, si minimizza solitamente il mean square error o altre funzioni di perdita. I modelli di ML sono non parametrici, il che significa che non dipendono dalla forma funzionale che lega il valore atteso condizionato con le x, né dalla distribuzione degli errori di questa funzione.

Le covariate spesso includono regressori creati ad hoc (come trend, dummy stagionali e regressori per cambi di livello) e spesso contengono la variabile y ritardata. Pertanto, applicando ML alle serie storiche, y diventa y_t e x diventa il ritardo a p periodi, sia della serie storica che dei regressori x_t (noti in precedenza).

È sempre importante considerare l'orizzonte predittivo. Ad esempio, un modello che predice un passo avanti avrà:

$$y \rightarrow y_{t+1}$$
$$x \rightarrow y_t, y_{t-1}, ..., y_{t-p}, x_t$$

8.2 Cross Validation per Serie Storiche

La **cross-validation** per le serie storiche è un processo cruciale per valutare l'efficacia dei modelli di previsione temporale. Tuttavia, poiché siamo interessati principalmente alla capacità predittiva sui dati futuri, è importante adottare un approccio che tenga conto dell'aspetto temporale dei dati.

Un metodo comune di cross-validation per serie storiche è lo **schema ricorsivo**. Questo approccio consente di mantenere l'ordine temporale dei dati. Ecco come funziona:

1. Si divide il campione iniziale in due parti: una per il training e una per la validation.
2. Si addestra il modello sul primo set di dati e usa il modello addestrato per prevedere i prossimi n passi in avanti.
3. Si aggiungono nuovi dati reali (non previsti) al campione di allenamento e si ripete il processo di training e previsione.
4. Si continua ad aggiungere dati reali e a fare previsioni finché non viene utilizzata l'intera serie storica.
5. Si calcolano gli errori di previsione per ogni passo in avanti.
6. Si valuta la bontà del modello considerando gli errori a uno, due, fino a n passi in avanti, sia rispetto al tempo che rispetto all'orizzonte temporale.

Una volta calcolati gli errori di previsione per ogni passo in avanti, è possibile valutare la performance del modello utilizzando metriche come l'errore quadratico medio (MSE) o altre metriche appropriate.

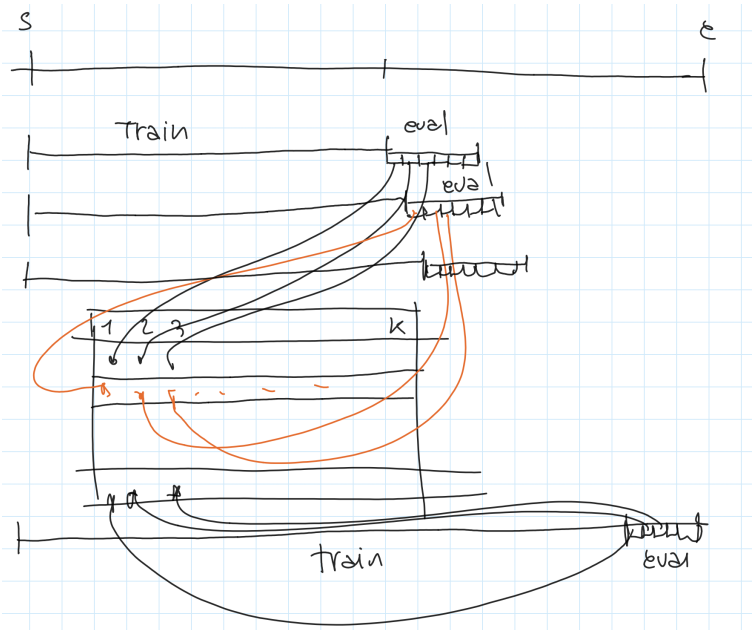


Figure 8.1: Cross validation ricorsiva.

È importante notare che, poiché i dati delle serie temporali possono provenire da processi complessi e non lineari, è possibile che alcuni modelli preformino meglio nel breve termine mentre altri nel lungo termine. Pertanto, la valutazione dei modelli dovrebbe tener conto sia della previsione a breve termine che di quella a lungo termine, poiché entrambi gli aspetti sono cruciali per comprendere la capacità predittiva complessiva del modello.

A volte nei modelli di ML si preferisce adottare una **moving window** anziché una previsione ricorsiva. In questa modalità, la stima avanza, utilizzando una finestra temporale di addestramento che si sposta senza prolungare il periodo di previsione. Questo approccio può essere preferibile in alcuni modelli poiché tende a non considerare i periodi più datati, che potrebbero essere stati influenzati da eventi anomali che hanno alterato la serie.

8.3 Multiple-Step Ahead Prediction

8.3.1 Metodo Ricorsivo

Abbiamo visto che per i modelli lineari la predizione è semplice e ottimale per qualsiasi orizzonte del modello. Questo perché la predizione risulta **ricorsiva**, permettendo di sostituire i valori futuri non disponibili con le loro predizioni.

Con il metodo ricorsivo, si sostituiscono appunto le osservazioni future mancanti con le predizioni ottenute:

1. Il modello viene costruito per l'osservazione successiva:

$$\hat{y}_{t+1|t} = \hat{f}(y_t, y_{t-1}, ..., y_{t-p+1})$$

2. Per predire due passi successivi, si utilizza la predizione fatta nello step precedente:

$$\hat{y}_{t+2|t} = \hat{f}(\hat{y}_{t+1|t}, y_t, y_{t-1}, ..., y_{t-p+1})$$

3. E così via:

$$\hat{y}_{t+3|t} = \hat{f}(\hat{y}_{t+2|t}, \hat{y}_{t+1|t}, y_t, y_{t-1}, ..., y_{t-p+1})$$

Questa predizione ricorsiva è ottimale per i modelli lineari, poiché se la predizione è una funzione lineare del passato, sostituire la miglior predizione del passato per lo step successivo risulta ottimale. La linearità garantisce che questa combinazione sia efficace.

Tuttavia, per i modelli non lineari, come quelli di ML, questo non è più valido. Minimizzando il mean square error, le predizioni sono i valori attesi condizionati del futuro dato il passato. Per la disuguaglianza di Jensen, i valori attesi condizionati di funzioni non lineari differiscono dalle funzioni non lineari dei valori attesi condizionati, indicando che stiamo facendo un'approssimazione.

Inoltre, nei modelli lineari è facile controllare la stabilità del modello controllando le radici del polinomio caratteristico della parte AR nei modelli $ARIMA$. Nei modelli non lineari, la funzione usata per le predizioni può essere molto instabile, rischiando che le predizioni esplodano nel futuro. Il controllo della stabilità per questi modelli è solo empirico.

Un ulteriore problema è che un errore nella prima predizione viene iterato in tutte le predizioni successive, amplificandosi.

Sebbene il metodo ricorsivo sia comune per fare predizioni in ML, non è l'unico e non è sempre la soluzione migliore, soprattutto se si desidera fare predizioni su un lungo orizzonte.

8.3.2 Metodo Diretto

Un alternativa al metodo appena visto è il metodo **diretto**. In questo metodo si costruiscono diversi modelli, uno per ogni orizzonte previsivo h. Ad esempio una predizione per un singolo mese h sarà uguale ad 1, mentre per 12 mesi h andrà da 1 a 12.

I modelli sono specificati come segue:

$$
\begin{aligned}
y_{t+1} &= f_1(y_t, ..., y_{t-p+1}) \\
y_{t+2} &= f_2(y_t, ..., y_{t-p+1}) \\
&\vdots \\
y_{t+h} &= f_h(y_t, ..., y_{t-p+1})
\end{aligned}
$$

Questo metodo varia molto la preparazione dei dati, che sarà:

$$
\begin{aligned}
y &\rightarrow y_{t+h} \\
x &\rightarrow y_t, y_{t-1}, ..., y_{t-p+1}, x_{t(+h)}
\end{aligned}
$$

Il metodo diretto è semplicemente un altro modo di gestire questo problema.

8.3.3 Metodo MIMO

Per molti modelli di ML il metodo diretto e il metodo ricorsivo sono gli unici due metodi, dato che molti accettano solo una variabile output. Se il modello di ML utilizzato permette l'utilizzo di target multipli, questo ad esempio vale per KNN e anche per le reti neurali, allora si usa il metodo **MIMO** (multiple input and multiple output).

In questo metodo la funzione è in grado di fare più previsioni insieme:

$$
\begin{bmatrix}
\hat{y}_{t+1|t} \\
\hat{y}_{t+2|t} \\
\vdots \\
\hat{y}_{t+h|t}
\end{bmatrix}
= f(y_t, ..., y_{t-p+1})
$$

Questo metodo è preferibile al metodo diretto in cui si costruiscono modelli diversi perché molto spesso l'informazione è comune per ogni orizzonte predittivo, come se vi fosse un segnale da estrarre dai dati per fare previsione. Se si riesce in un unico modello a costruite le funzioni del passato, che prevedono h periodi in avanti, si riesce meglio a distinguere il segnale e il rumore all'interno dei dati.

8.4 Soluzioni al Problema del Range

A seconda del modello che si utilizza, come detto nell'introduzione, è necessario fare considerazioni sul pre-processing dei dati. Ad esempio utilizzando modelli di ML basati su alberi decisionali si avrà il problema del range, ovvero che le previsioni non possono uscire dal

range osservato nel training del modello, dato che gli alberi si basano su suddivisioni del range iniziale dei dati in intervalli. Dunque i modelli ML tendono a fare previsioni simili a quello che hanno osservato in fase di training, questo fa si che questi modelli non sono in grado di prevedere un trend di crescita o di decrescita (che porterebbe ad una variazione nel range di valori possibili).

Consideriamo le trasformazioni applicate ai modelli $ARIMA$, come la differenziazione semplice e la differenziazione stagionale:

- *Differenziazione semplice*: chiediamo al modello di ML di apprendere la stagionalità. Il problema non è che il pattern stagionale possa evolversi nel tempo, ma che la stagionalità stessa è un pattern che può essere appreso e riproposto.
- *Differenziazione stagionale*: chiediamo al modello di apprendere l'incremento da un anno all'altro, superando la stagionalità. In questo modo, imponiamo al modello di ML una stagionalità che include una componente stocastica. Ad esempio, se il periodo è annuale, il modello apprende la stagionalità di un anno e somma la previsione del modello.

Per quanto riguarda la *non stazionarietà in varianza*, questa può rappresentare un problema. Se si applica la differenziazione semplice, si può comunque incorrere in problemi di range. Con la differenziazione stagionale, invece, tali problemi non si presentano. La differenziazione semplice può causare un aumento del range (i modelli di ML non possono prevedere nulla al di fuori del range osservato). In questi casi, potrebbe essere utile *linearizzare* i dati.

La soluzione potrebbe essere quella di applicare una *trasformazione logaritmica prima di utilizzare una differenziazione*, sia essa semplice o stagionale. Questo approccio consente di mitigare i problemi di varianza e di range, migliorando la capacità del modello di fare previsioni accurate.

Con la differenziazione stagionale, si elimina la stagionalità, ma può rimanere un certo trend. Se non si utilizza un modello $ARIMA$, l'inversione della trasformazione (da logaritmica a serie storica normale) deve essere effettuata manualmente, poiché non avviene automaticamente come accade con $ARIMA$.

Ecco il processo in dettaglio:

$$
\begin{aligned}
y_t &= \log(x_t) - \log(x_{t-h}) \\
\log(x_t) &= y_t + \log(x_{t-h}) \\
x_t &= x_{t-h} e^{y_t} \\
x_{t+h} &= x_t e^{y_{t+h}}
\end{aligned}
$$

Questo è ciò che il forecast di un modello $ARIMA$ fa automaticamente. In sostanza, se abbiamo le previsioni di y e abbiamo applicato trasformazioni sia logaritmiche che di dif-

ferenziazione stagionale, le nostre previsioni per y saranno l'esponenziale delle previsioni di y moltiplicate per la serie storica al tempo $t - h$.

Se avessimo utilizzato la differenziazione semplice, avremmo avuto $t - 1$ invece di $t - h$. Quando non utilizziamo $ARIMA$, dobbiamo eseguire manualmente questa operazione e costruire la trasformazione inversa.

8.5 Modelli ML

8.5.1 Alberi Decisionali

Quando si utilizzano metodi basati su alberi decisionali, come XGBoost o Random Forest, per gestire variabili continue o numeriche, queste vengono partizionate in base a soglie specifiche, ad esempio, "minore di" e "maggiore di" un certo valore.

Consideriamo come passare l'informazione della stagionalità che si ripete ogni 365 giorni.

Passare informazioni stagionali tramite funzioni sinusoidali funziona bene con modelli che gestiscono trasformazioni continue, ma con gli alberi decisionali, che effettuano trasformazioni non continue, queste sinusoidi vengono trasformate in gradini. Gli alberi decisionali individuano punti specifici delle sinusoidi e decidono soglie dove il valore è maggiore o minore in quei punti.

Un approccio efficace è passare direttamente il contatore dei giorni dall'inizio dell'anno. In questo caso l'algoritmo noterà che i valori a fine dicembre e inizio gennaio stabiliscono una soglia e partizionano i dati intorno al giorno 350. Un'altra partizione avverrà per i valori minori del sesto giorno dell'anno, e la stessa logica si applicherà ad altri periodi come agosto. Partizionando continuamente le variabili, passare il conteggio dei giorni dall'inizio dell'anno è una soluzione adeguata anche per stagionalità settimanali.

Sul singolo albero, questo approccio porta a una rappresentazione a gradini. Tuttavia, utilizzando tecniche di ensemble come XGBoost o Random Forest, che costruiscono un numero elevato di alberi, ciascuno con diverse partizioni e soglie, si ottiene una sorta di "smoothing" tra i risultati degli alberi. Questo processo elimina, in parte, la discontinuità.

Con XGBoost o Random Forest, ogni albero stabilisce soglie in giorni diversi e l'aggregazione dei risultati crea una previsione più fluida.

8.5.2 Support Vector Machines

Con Support Vector Machines (SVM), che includono regolarizzazione e utilizzano una funzione obiettivo diversa dalla minimizzazione del MSE, è possibile ottenere risultati simili in termini di gestione della stagionalità. L'SVM applica una funzione di perdita L2 che può gestire meglio la regolarizzazione e la continuità nei dati.

Le SVM partono da una funzione lineare e utilizzano il "kernel trick" per trasformare lo spazio delle feature, ampliandolo e creando interazioni tra le feature. I kernel, generalmente, calcolano interazioni tra coppie di variabili, permettendo di modellare le non linearità. Questo equivale ad avere avere una regressione lineare con una funzione obiettivo particolare, a cui si aggiunge la capacità di modellare interazioni e non linearità grazie al kernel. In questo contesto, le sinusoidi sono particolarmente utili.

Se si fornisce non linearità intelligenti (dato che il risultato delle SVM è lineare in output), l'algoritmo ne trarrà maggiore beneficio rispetto al semplice conteggio dei giorni dell'anno. Perciò, è utile distinguere tra modelli che producono funzioni continue, dove ha senso utilizzare funzioni lisce come le sinusoidi, e metodi che generano funzioni discontinue, come gli alberi decisionali, dove le funzioni continue verrebbero semplicemente frammentate.

Al contrario di altri metodi, le SVM possono apprendere la presenza di un trend, rendendole più adatte a catturare le dinamiche sottostanti nei dati temporali.

8.5.3 K-Nearest Neighbors

Supponiamo di voler prevedere una serie storica univariata. Consideriamo i seguenti parametri:

- h: orizzonte previsivo
- p: look back, ovvero il numero di osservazioni passate da utilizzare per prevedere il futuro
- k: numero di esempi più simili su cui basare le previsioni

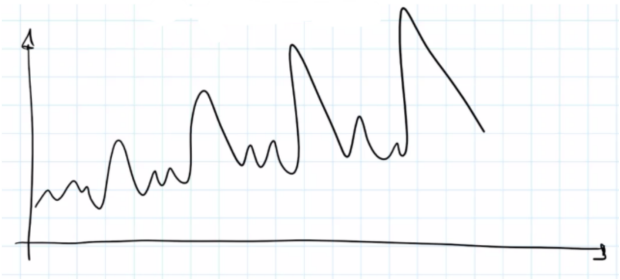


Figure 8.2: Serie storica univariata.

Un algoritmo K-NN (k-nearest neighbors) cerca nel passato le situazioni più simili a quelle attuali. Se si fissa $p = 12$ (un anno di dati) e si vuole prevedere h periodi in avanti, il K-NN cerca le k sequenze nella serie storica più simili a quella osservata nel presente. Successivamente, guarda cosa è successo in quelle k sottosequenze nei h passi successivi e sintetizza ciò che è avvenuto (ad esempio, facendo la media) per fare la previsione.

In pratica, osserviamo l'ultimo anno di dati e cerchiamo sequenze simili spostandoci nel passato di intervalli di 12 unità temporali. Tuttavia, ci sono problemi tipici delle serie storiche:

1. Senza trasformazioni sui dati, la previsione sarà basata sull'ultimo anno osservato, senza crescita e con ampiezza minore rispetto a quella reale.
2. Le stime dei minimi quadrati tendono a sottostimare le ampiezze se la serie ha un trend crescente o stagionalità amplificante.

Nel K-NN è strutturalmente impossibile generare trend o stagionalità che aumentano di ampiezza se queste non sono presenti nel passato. Pertanto, è necessario applicare trasformazioni ai dati che rendano la serie meno moltiplicativa.

Le trasformazioni possono essere:

1. Trasformazione logaritmica: Le serie storiche moltiplicative (dove il trend è moltiplicato per un fattore stagionale) diventano additive tramite la trasformazione logaritmica.
2. Differenziazione: Applicare la differenziazione semplice o stagionale ai dati logaritmici.

La trasformazione logaritmica e la differenziazione aiutano a stabilizzare la varianza e a rendere la serie additiva, facilitando l'identificazione delle similarità stagionali. Quando la differenza è priva di trend (senza ampliamento nel tempo), il K-NN può sfruttare la stagionalità per identificare le similarità.

In realtà, nella pratica viene utilizzato un metodo diverso di de-trendizzare il modello da confrontare con le sottosequenze. Quando si confrontano i due modelli (si calcola la metrica, tipicamente la distanza euclidea), dipende dalla scelta:

- Modello additivo: toglie la media a entrambe le sequenze e confronta le due sotto-sequenze private della media.
- Modello moltiplicativo: viene divisa la sotto-sequenza per la media, quindi le riscala.

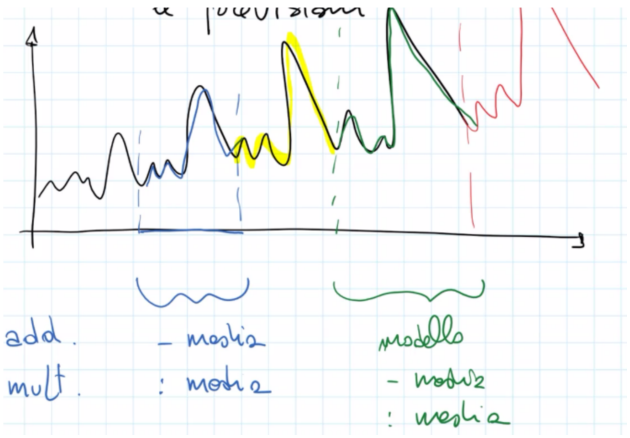


Figure 8.3: De-trenizzazione nella pratica.

Nel momento in cui deve andare a prevedere in avanti, supponiamo che abbia selezionato la sottosequenza blu come la più vicina (neutralizzata rispetto al livello della serie storica) a quella che ci interessa. Andiamo a vedere cosa è successo nei 12 passi successivi nella sottosequenza.

Per applicare l'anti-trasformazione a questo futuro:

- Modello additivo: dal futuro giallo viene tolta la media della sottosequenza blu.
- Modello moltiplicativo: la sottosequenza gialla viene divisa per la media della blu, poi viene presa la media del verde e moltiplicata per quella sottosequenza.

In seguito, si ricostruisce la differenza tra blu e giallo e si procede a ricostruirla dal verde al rosso. Questo è il modo in cui si riproduce il trend.

I metodi pricipali sono:

1. *Metodo MIMO*: Sostanzialmente, copiamo tutta la sequenza successiva alla sottosequenza che ci interessa.
2. *Metodo Ricorsivo*: Possiamo adottare un approccio simile alle random forest; possiamo fare previsioni un passo in avanti e poi sostituire la serie storica e rifare la previsione un passo in avanti, e così via.
3. *Metodo Diretto*: Costruiamo un modello diverso per ogni orizzonte previsivo (un modello per un passo in avanti, uno per due passi in avanti, ..., uno per k passi in avanti).

La metrica di default è quella euclidea. Di solito, nel mondo KNN, è necessario normalizzare i dati per evitare scale troppo diverse tra le feature. Tuttavia, quando si usa KNN su una singola serie storica, non si normalizza nulla, in quanto il modello gestisce la standardizzazione internamente.

Se si vogliono aggiungere altre features oltre alla serie storica, ci sono sfide nella standardizzazione delle variabili, in quanto potrebbe essere difficile dosare l'importanza di ciascuna feature nell'identificare la similarità. Mentre le tecniche di standardizzazione sono spesso utilizzate per fare previsioni basate solo sul passato della serie storica, quando si vogliono aggiungere altre features, potrebbero non funzionare altrettanto bene.

Una volta che abbiamo k esempi, dobbiamo scegliere come sintetizzare i k futuri che emergono da questi esempi. Ci sono vari metodi, come la media, la mediana e la trimmed mean (media che estrae le osservazioni outlier/estreme). Tuttavia, tutti i metodi basati su gaussiane possono funzionare male su funzioni con code pesanti, e la media è particolarmente sensibile agli outlier.