Concetti fondamentali serie temporali

Di Davide Grimaldi

Di seguito verrano esposti i principali concetti alla base della comprensione e dello studio delle TS, fra cui quelli di statistica e le principali caratteristiche teoriche.

Rumore bianco

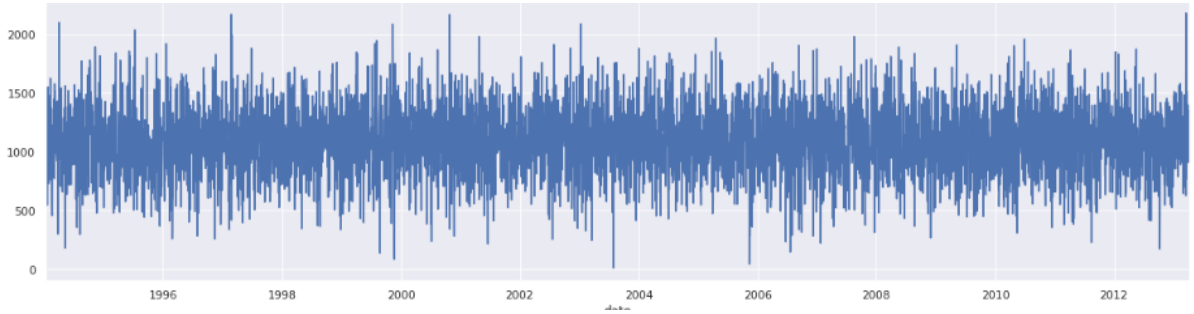
Il rumore bianco può essere visto come un tipo speciale di TS, dove non vale l'assunzione che gli andamenti nel passato continueranno nel futuro, i dati non seguono un particolare andamento.
Le caratteristiche del rumore bianco sono:

i. Valore atteso uguale a zero, $\mu = 0$
ii. Varianza costante, $\sigma^2 = const$
iii. Autocorrelazione nulla, $\rho = corr(x_t, x_{t-1}) = 0$

Questo ultimo punto rende il rumore bianco non predicibile.

È importante distinguere il rumore bianco da una TS, perchè se abbiamo un andamento a rumore bianco non possiamo fare nessuna previsione. Per distinguere basta osservare i grafici, creiamo una TS a rumore bianco ponendo la sua media uguale alla media e la sua varianza uguale a quella dell'indice S&P500.

```
wn =np.random.normal(loc= df.spx.mean(),scale=
df.spx.std(),size = len(df))
df['wn'] = wn
df.wn.plot(figsize = (20,5))
plt.show()
```



```
df.market_value.plot(figsize=(20,5))
plt.title("Prezzi S&P 500", size = 24)
plt.ylim(0,2300)
plt.show()
```

Come possiamo notare il rumore bianco ha cambi repentini e grandi variazioni fra un intervallo e l'altro, inoltre non presenta andamenti particolari: è evidente la differenza con una TS come quella di S&P500 disegnata per confronto.

Quando alleneremo il nostro modello per prevedere gli andamenti di una TS i suoi errori nel tempo dovrebbero essere il quanto più simili ad un rumore bianco, se così non fosse vorrà dire che al modello è sfuggita una caratteristica della TS.

Random walk

Anche il random walk può essere visto come un tipo speciale di TS, in esso i valori tendono a persistere nel tempo e le differenze fra i periodi sono semplice rumore bianco. In questo tipo di TS la stima di un valore dipende dal suo precedente e da un errore di tipo rumore bianco:

$$x_t = x_{t-1} + \epsilon_t$$

Dove $\epsilon_t = RumoreBianco\left(\mu, \sigma^2\right)$. Putroppo non è possibile fare previsioni accurate proprio per la natura di questo errore. Tramite il cocetto di random walk è possibile misurare la difficolta di previsione di una TS, più essa differisce da una random walk più facile è fare previsioni.

Mettiamo a confronto i grafici di una random walk con l'indice S&P500.
Per prima cosa carichiamo una random walk e settiamola in modo tale che abbia la stessa frequenza di S&P500

```
rw = pd.read_csv("RandWalk.csv")
rw.date = pd.to_datetime(rw.date, dayfirst = True)
rw.set_index("date", inplace = True)
rw = rw.asfreq('b')
df['rw'] = rw
```

Grafichiamo nella stessa figura i due andamenti

```
df.rw.plot(figsize = (20,5))
df.spx.plot()
plt.title("Random Walk vs S&P", size = 24)
plt.legend()
plt.show()
```

Come vediamo gli andamenti sono molto simili! Questo ci da una intuizione su quanto sia difficile prevedere gli indici di borsa!

Stazionarietà

Una TS si dice debolmente stazionaria se, differenti campioni sui dati, cioè insiemi di n intervalli consecutivi, hanno uguale covarianza senza considerare il punto di partenza. Più formalmente le proprietà sono:

i. media costante $\mu = const$
ii. varianza costante $\sigma^2 = const$
iii. $cov(x_n, x_{n+k}) = cov(x_m, x_{m+k})$

Questi principi sono sintetizzati nella seguente figura[Fig]

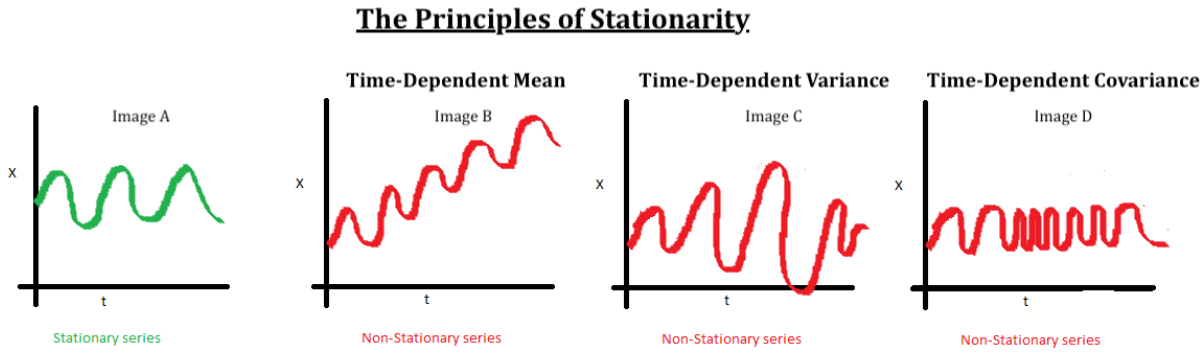


Figure 1: a) sono rispettati tutte le proprietà b) non rispettata la prima c) non rispettata la seconda d) non rispettata la terza

Un esempio di TS debolmente stazionaria è il rumore bianco, infatti esso oltre ad avere media e varianza costante ha correlazione nulla quindi

$$cov(x_m, x_{m+k}) = corr(x_m, x_{m+k}) \cdot \sigma_m \cdot \sigma_{m+k} = 0$$

sapendo che la covarianza è la correlazione moltiplicata per la deviazione standard. Per avere la stazionarietà stretta occorre anche che le distribuzioni di due campioni qualsiasi siano uguali

$$(x_m, x_{m+k}) \approx Dist(\mu, \sigma^2)$$
$$(x_{m+\tau}, x_{m+\tau+k}) \approx Dist(\mu, \sigma^2)$$
$$\forall \tau \in N$$

Questo tipo di stazionarietà si osserva raramente in natura, per questo quando si parla in generale di stazionarietà ci si riferisce a quella debole.

Dickey-Fuller test

Per testare se una TS sia stazionaria o meno si fa uso del Dickey-Fuller test. Nel test vengono consideratedue ipotesi:

1. l'*ipotesi nulla* H_o, la non stazionarietà

2. l'*ipotesi alternativa* H_1 la stazionarietà

Viene calcolata la propabilità che data per certo l'ipotesi nulla si ottenga una distribuzione dei dati come quella che abbiamo, questa viene chiamta *t-score* o *statistica t.* Se questo valore risulta inferiore a un valore di soglia si ritiene rigettata l'ipotesi nulla, quindi si ammette la stazionarietà.
Il test viene eseguito eseguendo l'apposita funzione del package *statsmodels.tsa.stattools.* Eseguiamo il test sui valori di S&P500 e analizziamone i risultati.

```
sts.adfuller(df.spx)
```

```
(-1.7369847452352405,
 0.41216456967706366,
 18,
 5002,
 {'1%': -3.431658008603046,
  '5%': -2.862117998412982,
  '10%': -2.567077669247375},
 39904.880607487445)
```

Capiamo il significato di ogni valore:

- $-1.736...$ è il valore della *statistica t*
- $0.4121...$ è il valore p associato alla statistica t, ovvero la probabilità con la quale falliamo nel rigettare l'ipotesi nulla
- 18 è il numero di intervalli che vengono considerati durante la regressione, ovvero il numero di intervalli precendeti a un singolo valore da analizzare che vengono considerati rilevanti nel calcolo della correlazione.
- 5002 è il numero di osservazioni utilizzate, sommando questo valore a quello precedente si ottiene il numero totale di osservazioni, infatti le prime diciotto osservazioni non hanno sufficienti intervalli precedenti
- I tre successivi valori sono tre valori di soglia (rispettivamente 1%, 5%, 10%)con cui confrontare il t-score
- $39904.880...$ è parametro il *criterio di massima informazione* un parametro che misura quanto il modello regressivo è accurato nelle previsioni in rapporto alla complicatezza del modello stesso, più è piccolo questo valore migliore è la capacità di previsione sulla data TS

Da questi valori vediamo che la statistica t è maggiore rispetto ai valori critici e quindi non è possibile rifiutare l'ipotesi nulla: la TS non è stazionaria.

Confrontiamo questi risultati con quelli del test eseguito sul rumore bianco e sul random walk.

```
sts.adfuller(df.wn) #sul rumore bianco
```

```
(-48.037079944642365,
 0.0,
 1,
 5019,
 {'1%': -3.4316535759402753,
  '5%': -2.8621160400844468,
  '10%': -2.567076626752987},
 70845.5091765096)
```

Possiamo nortare una statistica t molto bassa, inferiore di molto anche alla soglia critica dell'1%, quindi viene rigettata l'ipotesi nulla e la TS è stazionaria, cosa che ci aspettavamo essendo il rumore bianco, come detto in precedenza, stazionario. Il valore zero degli intervalli era anche prevedibile, visto che ogni valore in una TS a rumore bianco non dipende dal suo valore precende in quanto casuale. Il parametro del criterio di massima informazione è invece molto più alto, questo perchè il rumore bianco è impredicibile.

```
sts.adfuller(df.rw) #random walk
```

```
(-1.3286073927689712,
 0.6159849181617387,
 24,
 4996,
 {'1%': -3.4316595802782865,
  '5%': -2.8621186927706463,
  '10%': -2.567078038881065},
 46299.333497595144)
```

Una TS a random walk come ci si poteva aspettare non è stazionaria, anzi è meno stazionaria dell'indice S&P500 in quanto i cambiamenti da un intervallo all'altro sono puramente randomici.

Stagionalità

La stagionalità rappresenta gli andamenti ciclici all'interno di una TS, un esempio è la temperatura che dipende ciclicamente dall'orario della giornata o dal periodo dell'anno. Saper identificare delle componenti stagionali aiuta a costruire modelli più accurati per le previsioni, esistono diversi metodi per estrarre queste componenti, il più immediato è quello della *decomposizione.* Questa decompone la TS in tre parti:

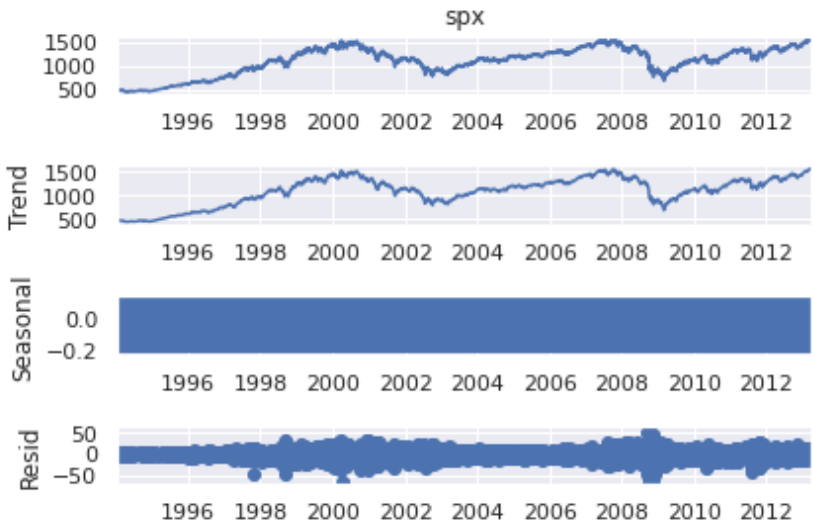
1. Trend, descrive i pattern su tutto l'intervallo dei dati
2. Stagionale, descirve gli effetti ciclici
3. Residuale, rappresenta le differenze fra le predizioni e i valori effettivi della TS

Si possono svolgere due tipi di decomposizioni, dette *naive,* quella additiva e quella moltiplicativa.

$$additiva: \quad x_t = Trend + Stagionale + Residuale$$
$$moltiplicativa: \quad x_t = Trend \cdot Stagionale \cdot Residuale$$

Applichiamo la decomposizione attraverso la funzione apposita del pacchetto *statsmodels.tsa.seasona*l

```
s_dec_multiplicative = seasonal_decompose(df.spx, model =
"additive")
s_dec_multiplicative.plot()
plt.show()
```



Possiamo vedere che la componente Trend è molto rappresentativa della TS, la parte stagionale ha la forma di un rettangolo perchè abbiamo una oscillazione velocissima del valore fra 0 e -0.2 per ogni intervallo, dato il basso valore e l'assenza di andamenti concludiamo che non abbiamo una particolare componente stagionale. I residui sono più marcati intorno al 2000 e 2008 in cui ci sono state le ultime crisi finanziarie, intervalli in cui il modello che ha eseguito la decomposizione ha commesso più errori.

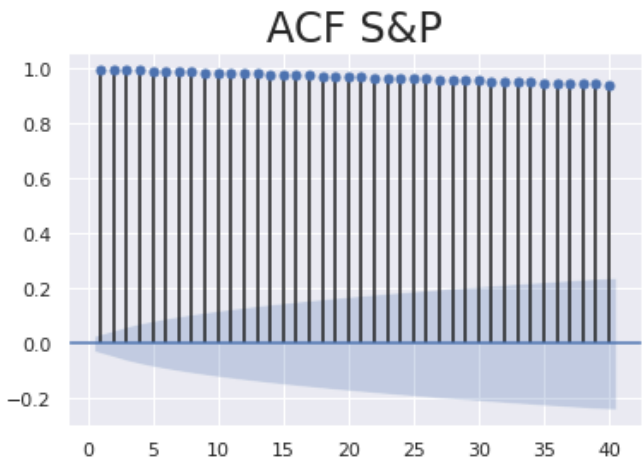
ACF e PACF

In una TS siamo interessati nell'influenza che i valori passati hanno sui valori presenti e che avranno sui valori futuri. Uno strumento utile per valutare questo è la *auto correlation function,* questa valuta per ogni osservazione il valore di autocorrelazione di essa con i valori valori fino a n intervalli precedenti:

$$\left.\begin{array}{c}\rho(x_t, x_{t-1}) \\ \rho(x_t, x_{t-2}) \\ \rho(x_t, x_{t-3}) \\ \dots \\ \rho(x_t, x_{t-n})\end{array}\right\} ACF$$

Valutiamo graficamente questa funzione applicandola sull'indice S&P500 utilizzando l'apposita funzione *plot_acf()* presente nel pacchetto *statsmodels.graphics.tsaplots* importato come sgt.

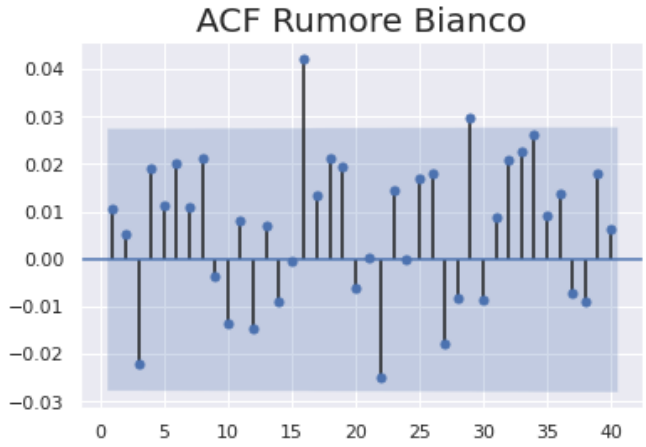
```
sgt.plot_acf(df.market_value, lags = 40, zero = False)
plt.title("ACF S&P", size = 24)
plt.show()
```



Il parametro "zero" impostato a false serve per escludere il valore stesso, inutile calcolare l'autocorrelazione con se stesso, il parametro "lags" indica il numero di intervalli (n) da considerare nella ACF, nell'analisi delle TS si utilizza di solito il valore 40, empiricamente ottimale.
Ogni linea del grafico corrisponde al valore di autocorrelazione all'ennesimo intervallo, i valori possono oscillare fra 1 e -1. L'area blu è detta *area di significanza,* tutti i valori che escono da questa area hanno una autocorrelazione significativa,

l'area aumenta man mano che aumentano gli intervalli perchè un valore più è lontano nel tempo maggiore deve essere il suo livello di autocorrelazione per influire sul valore presente. Notiamo che i valori di autocorrelazione partono da 1 e vanno man mano a scendere andando indietro nel tempo, capiamo quindi che tutti i valori passati incidono su quello corrente, in modo man mano descrescente, questo evidenzia una dipendenza temporale.
Se applichiamo la stessa funzione al rumore bianco otteniamo invece il seguente grafico:

```
sgt.plot_acf(df.wn, lags = 40, zero = False)
plt.title("ACF S&P", size = 20)
plt.show()
```



Come potevamo aspettarci la maggior parte dei valori non hanno autocorrelazione significativa, assunto rumore bianco, inoltre abbiamo valori positivi e negativi alternati dovuti proprio alla randomicità del fenomeno.

L'ACF misura l'autocorrelazione fra i valori di una TS e una sua versione ritardata (traslata nel tempo), in questo modo vengono presi in considerazione anche effetti indiretti. Prendiamo per esempio una TS formata da tre osservazioni [Fig3], il valore x_t può essere influenzato dal valore x_{t-2} direttamente o indirettamente tramite l'effetto che esso ha sul valore x_{t-1}.

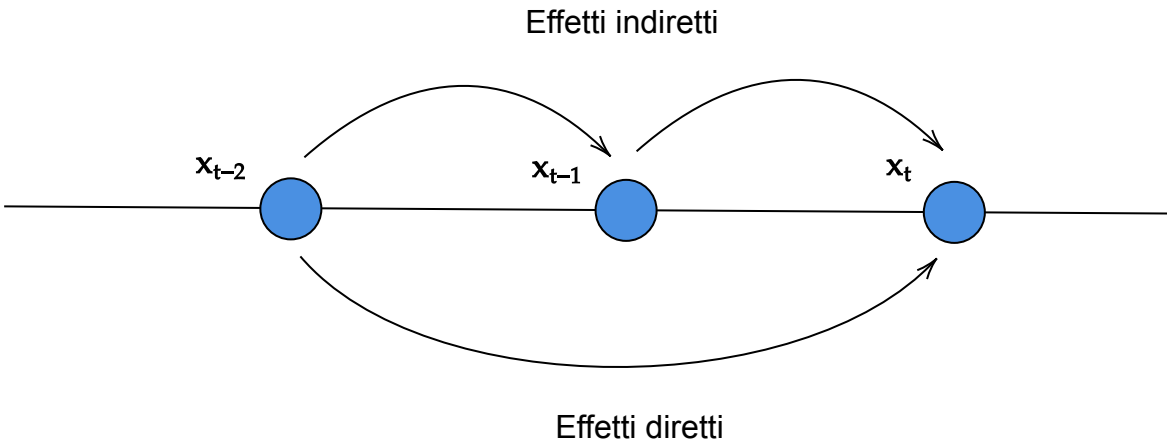
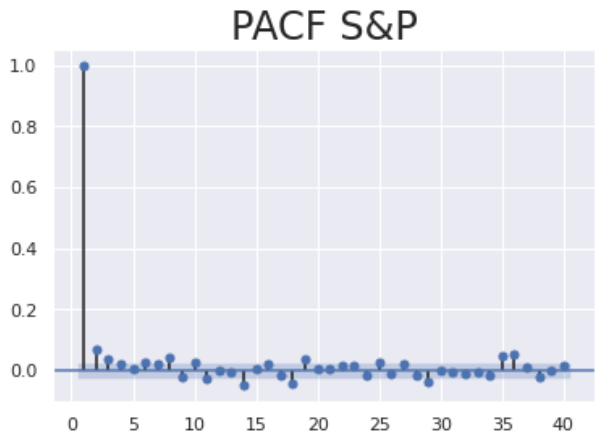


Fig 3: effetti diretti e indiretti

Quando costruiamo un modello è utile conoscere i singoli effetti diretti, in tal caso viene usato la *partial auto correlation function* PACF, questa calcola i valori dei *residui*, ovvero le componenti che rimangono quando vengono eliminati gli effetti indiretti degli intervalli fra i due valori considerati.

Grafichiamo in modo del tutto simile a prima la PACF dell'indice S&P500

```
sgt.plot_pacf(df.market_value, lags = 40, zero = False,
method = ('ols'))
plt.title("PACF S&P", size = 24)
plt.show()
```



Il parametro "method" specifica il metodo con cui viene calcolata la PACF.
Notiamo che solo i primi valori sono statisticamente significativi, situazione opposta

rispettoa alla ACF, questo è dovuto a come è costruita la PACF, infatti per esempio il valore x_{t-3} incide attraverso diversi *canali* per l'ACF.

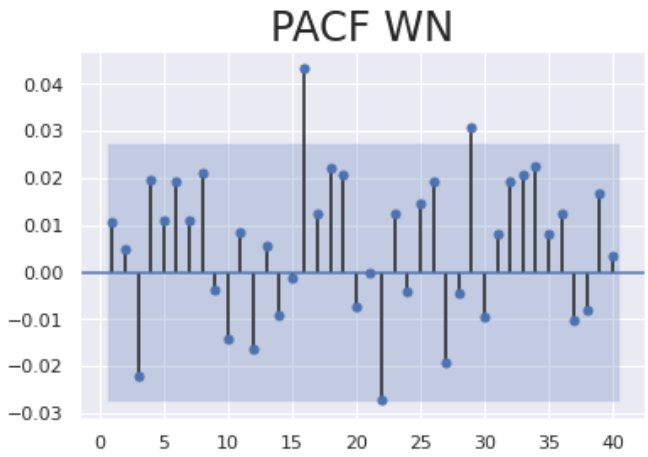
$$PACF: \; x_{t-3} \rightarrow x_t$$

$$ACF: \begin{cases} x_{t-3} \rightarrow x_{t-2} \rightarrow x_{t-1} \rightarrow x_t \\ x_{t-3} \rightarrow x_{t-1} \rightarrow x_t \\ x_{t-3} \rightarrow x_{t-2} \rightarrow x_t \\ x_{t-3} \rightarrow x_t \end{cases}$$

Questi effetti indiretti ci fanno capire le differenze fra i due grafici, l'unico valore che deve essere uguale è il primo perchè esso non ha canali alternativi.

Come nel caso precedente facendo la PACF del rumore bianco otteniamo dei valori randomici molto piccoli e non significativi.

```
sgt.plot_pacf(df.wn, lags = 40, zero = False, method =
('ols'))
plt.title("PACF WN", size = 24)
plt.show()
```



INDICE

Bibliografia

1. Pedamkar P. Introduction to Time series Analysis [Internet]. Available from: https://www.educba.com/time-series-analysis/