# Probablità, Statistica e Processi Stocastici

Franco Flandoli, Università di Pisa

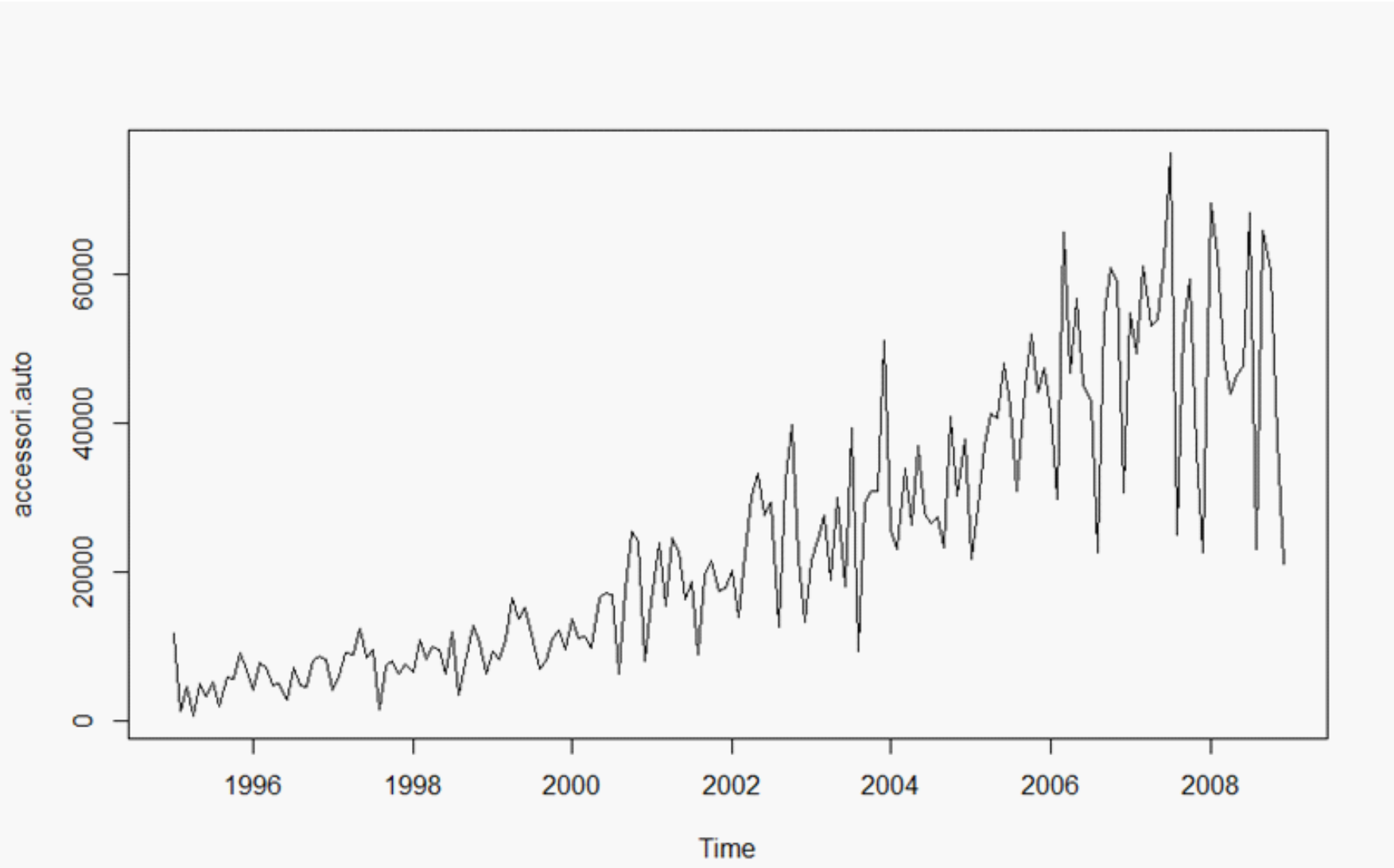
Corso per la Scuola di Dottorato in Ingegneria

# Serie storiche (verso fPCA)

- La tecnica chiamata fPCA (functional PCA) esamina serie storiche utilizzando paradigmi propri di PCA.
- E' utile premettere un po' di terminologia riguardante le serie storiche, per capire meglio.
- Quindi invertiamo leggermente l'ordine del programma, cioè anticipiamo alcune idee sulle serie storiche.
- Una serie storica è una sequenza di numeri $x_1, ..., x_n$ in cui l'indice $1, ..., n$ corrisponde al tempo, discetizzato come serve a seconda del problema (giorni, mesi anni ecc.).
- Ad esempio, il volume mensile di esportazioni italiane di automobili è una serie storica, reperibile sul sito Eurostat.

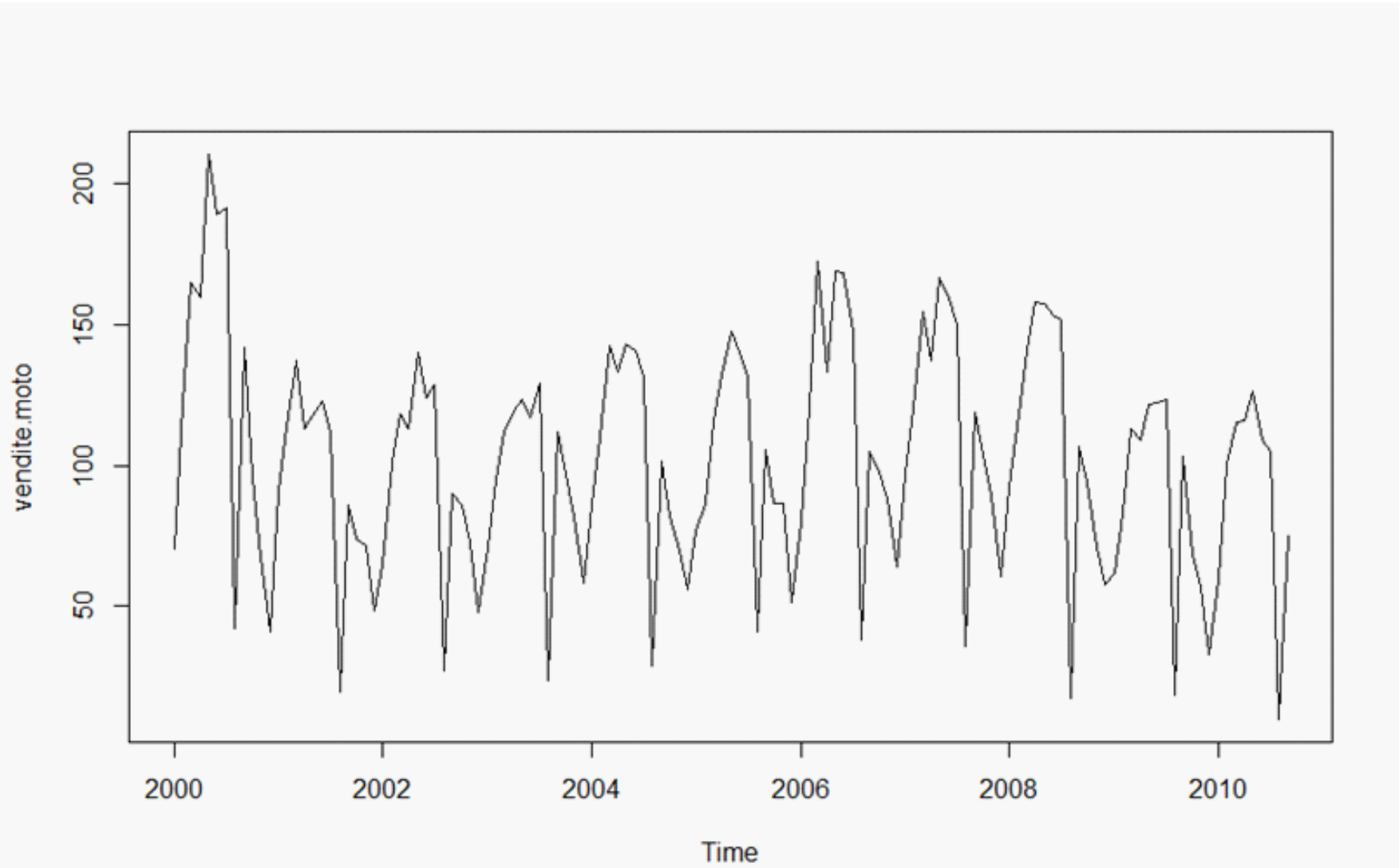
# Esempio di serie storica

Esportazioni italiane di pezzi di accessori auto (trend accentuato, poca stagionalità)



# Esempio di serie storica

Esportazioni italiane di motocicli (trend debole variabile, molta stagionalità)



# Analisi visive col software

1. funzione di autocorrelazione
   - quando c'è un forte trend, è tutta alta
   - altrimenti i picchi indicano le periodicità
2. decomposizione in trend, stagionalità e residui
   - si può fare in modo semplice ed uniforme con `decompose`
   - oppure in modo complesso e parametrizzato con `stl`.

# Processi stocastici

- Una serie storica è una successione finita concreta di valori, un po' come un campione sperimentale.
- Dietro un campione sperimentale immaginiamo spesso ci sia una variabile aleatoria.
- Dietro una serie storica, possiamo immaginare ci sia un *processo stocastico*.
- Questa immaginazione spesso è ancor più ardita che per le singole variabili aleatorie: la ragione è che spesso si possiede una sola serie storica di un problema specifico (es. il PIL italiano dal 2000 ad ora). Questo corrisponderebbe ad avere un solo valore sperimentale di una variabile aleatoria.
- Comunque, ragioniamo su questa immaginazione. Un processo stocastico a tempo discreto è una successione (finita o infinita) di v.a. $X_0, X_1, X_2, ..., X_n, ...$
- $X_n$ rappresenta il valore al tempo $n$ (o tempo $t_n$, secondo la convenzione che si preferisce).

# Funzioni medie associate ad un processo stocastico

Assumiamo tacitamente che tutti i valori medi scritti nel seguito siano ben definiti e finiti.
Chiamiamo *funzione valor medio* la funzione (a tempo discreto)

$$m(t) = E[X_t], \qquad t = 0, 1, 2, ..., n, ...$$

e *funzione di autocorrelazione* la funzione di due variabili

$$R(t, s) = E[X_t X_s], \qquad t, s = 0, 1, 2, ..., n, ...$$

Questo è il linguaggio preferito in ingegneria delle telecomunicazioni. Nel linguaggio invece più direttamente ispirato ai concetti base della probabilità, la *funzione di autocorrelazione* è:

$$\rho(t, s) = \frac{E[X_t X_s] - E[X_t] E[X_s]}{\sqrt{Var[X_t] Var[X_s]}}$$

(una versione standardizzata di $R(t, s)$).

# La funzione di autocorrelazione

- La funzione $\rho(t,s)$ (ma anche la $R(t,s)$) cattura degli aspetti dinamici, di legame statistico tra i valori ad un istante e quelli ad un istante successivo.
- Se ad esempio $\rho(t,0) \sim 0$ per $t$ grande, significa che il processo perde memoria al trascorrere del tempo.
- Se un processo è approssimativamente periodico, di periodo $P$, cioè $X_{t+P}$ è molto correlato a $X_t$, vale

$$\rho(P,0) \sim 1, \qquad \rho(2P,0) \sim 1 \qquad \text{ecc.}$$

  mentre gli altri valori $\rho(t,0)$ sono meno vicini ad 1. Quindi la funzione $\rho(t,0)$ può essere usata per indagare la periodicità.
- *acf* èuna versione empirica di $\rho(t,0)$.

# Processi stazionari in senso debole

Si chiamano così quando vale

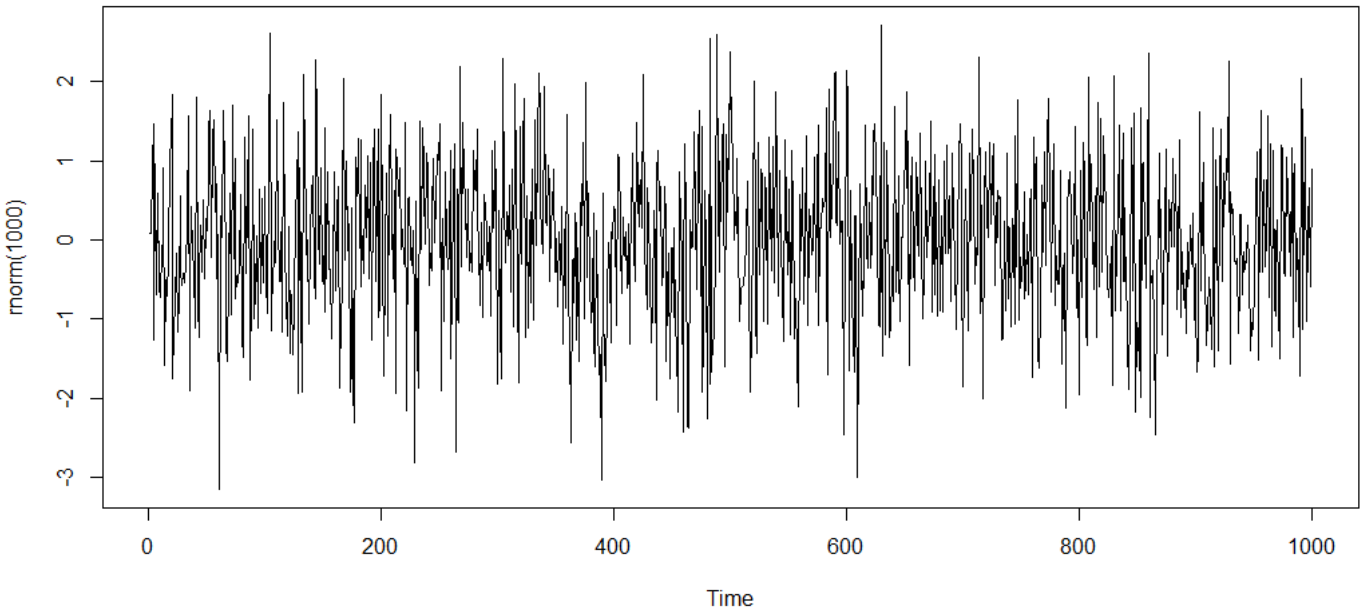
$$m(t) =: m \text{ costante}$$

$$\rho(t+h, s+h) = \rho(t,s) =: \rho(t-s)$$

per ogni $h > 0$. Come appena scritto, sono descritti dalla costante $m$ e dalla funzione di una sola variabile $\rho(t)$.
La teoria di questi processi è molto sviluppata. Tuttavia, entrambe le serie storiche dei nostri esempi sono ben lungi da essere stazionarie. Quindi non esauriscono di certo i processi di interesse pratico.

# White noise

Il rumore bianco a tempo discreto, di intensità $\sigma$, è una successione di v.a. $X_n$ indipendenti gaussiane di media zero e varianza $\sigma^2$ (a volte si danno definizioni più deboli). Si verifica che è stazionario.
La sua funzione di autocorrelazione $\rho(t)$ è una "delta di Dirac" discreta: vale 1 per $t = 0$, 0 per $t > 0$. Ecco una traittoria simulata, con ts.plot(rnorm(1000)):



# Decomposizione

L'usuale decomposione a cui si allude è della forma (additiva)

$$X_n = T_n + S_n + \epsilon_n$$

oppure (moltiplicativa)

$$X_n = T_n S_n \epsilon_n$$

(che però è del tipo additivo se appicata a $Y_n = \log X_n$).
Qui $T_n$ è un processo con forte trend, $S_n$ un processo fortemente periodico, $\epsilon_n$ un processo più vicino possibile ad un white noise.
Purtroppo questi concetti sono vaghi, indifinibili in modo esatto. Anche per questo non c'è un unico algoritmo di decomposizione, un unico trend ecc.

# Ricerca di un trend

Basilare per innescare algoritmi o in sé per motivi applicativi, è determinare un buon trend di una serie storica data.
Un algoritmo semplice è la *media mobile*:

$$t_n = \frac{x_n + x_{n-1} + x_{n-2} + x_{n-3} + x_{n-4}}{5}$$

che abbiamo illustrato con *finestra* pari a 5, per non scrivere formule astratte.
Il calcolo precedente può essere utilizzato o per calcolare un possibile trend al tempo $n$, oppure al tempo centrato $n-2$, ecc.; oppure per stimare il valore successivo (*previsione*), magari incognito, al tempo $n+1$.
Il comando `decompose` trova il trend usando la media mobile simmetrica (per questo perde un pezzo all'inizio ed uno alla fine).

# Smorzamento esponenziale

Simile alla media mobile è il metodo SE, di Smorzamento Esponenziale.
In esso si sommano i valori presente, $x_n$ e passati ($x_{n-1}, x_{n-2}$ ecc.) pesandoli in modo esponenziale, invece che in modo uniforme come nella media mobile.
Precisamente, si deve introdurre una serie storica ausiliaria $p_1, p_2, ...., p_n, ...$ che ha il seguente significato:

- $p_{n+1}$ è la previsione, relativa al valore del tempo $n+1$, effettuata al tempo $n$.

Si impone allora la formula iterativa

$$p_{n+1} = \alpha x_n + (1-\alpha)\, p_n.$$

La logica è: nel fare la previsione, usiamo una componente *innovativa* $\alpha x_n$ ed una *conservativa* $(1-\alpha)\, p_n$.

# Smorzamento esponenziale

Iterando, troviamo

$$p_{n+1} = \alpha x_n + (1-\alpha)\left(\alpha x_{n-1} + (1-\alpha)\, p_{n-1}\right) = \ldots$$

da cui

$$p_{n+1} = \alpha x_n + \alpha(1-\alpha)\, x_{n-1} + \alpha(1-\alpha)^2 x_{n-2} + +\alpha(1-\alpha)^3 x_{n-3} + \ldots$$

ovvero una media pesata esponenzialmente dei valori presente e passati.
L'iterazione va inizializzata, ponendo ad es. $p_1 = x_1$.
Il parametro $\alpha$ può essere assegnato a piacere o trovato col metodo dei minimi quadrati: posto $\epsilon_n = x_n - p_n$, si minimizza

$$\sum_n \epsilon_n^2.$$

Vedremo una serie di versioni più evolute di metodi, basati però inizialmente su queste semplici idee.