

FAMIGLIA CONIUGATA NATURALE

Famiglia di distribuzioni il cui nucleo ha la stessa forma funzionale della verosimilità. In tal caso, applicando il teorema di Bayes, si giunge alla stessa forma funzionale, ma con parametri aggiornati.

STATISTICA SUFFICIENTE

Stabiliamo $T = t(x_1, \dots, x_n)$ funzione dei x_1, \dots, x_n .

È una informazione dell'oggetto x_1, \dots, x_n .

$$\text{Sufficiente} \Leftrightarrow G(\theta | H_n(x_1, \dots, x_n)) = G(\theta | H(T(x_1, \dots, x_n)))$$

$$\Leftrightarrow F(x_1, \dots, x_n | t(x_1, \dots, x_n, \theta)) \text{ non dipende da } \theta$$

$$\Leftrightarrow l(\theta | x_1, \dots, x_n) = Ml(\theta | x_1) = h(t(x_1, \dots, x_n) | \theta) k(x_1, \dots, x_n)$$

FAMIGLIA ESPONENZIALE

derutti o probabilità:

$$\frac{f_x(x | \theta)}{P(x_0 | \theta)} = B(\theta) h(x) \exp \left\{ \sum_{j=1}^n \alpha_j(\theta) u_j(x) \right\}$$

\Rightarrow stabilità delle α_j e diverse costanti
rispetto a n

PROCESSI STOCASTICI

Richiamo su processi di Poisson:

Sono anche detti processi di arrivi o di Puro Ingresso.

Consideriamo uno Spazio di Probabilità (Ω, \mathcal{F}, P) .

Su di esso siano definiti i seguenti numeri aleatori a valori in \mathbb{R} :

• $T(j, \omega)$: tempi di attesa tra arrivi successivi dell'evento di interesse;

• $S(n, \omega)$: tempi di arrivo dell'evento di interesse;

$$\text{Risulta che: } S(n, \omega) = \sum_{j=1}^n T(j, \omega) \quad \begin{array}{c} T_1 \quad T_2 \quad T_3 \\ \downarrow \quad \downarrow \quad \downarrow \\ S_1 \quad S_2 \quad S_3 \dots \end{array}$$

• $N(t, \omega)$: numero aleatorio di arrivi nell'intervallo $[0, t]$.

Poiché questi numeri aleatori sono definiti sul medesimo spazio di probabilità, essi sono collegati tra loro, infatti risulta che:

gli eventi $\{N(t, \omega) = n\} \sim \{S_n \leq t < S_{n+1}\}$

$$\text{equivalente} \quad \begin{array}{c} I_n \quad I_{n+1} \\ \downarrow \quad \downarrow \\ \sum_{j=1}^n T(j, \omega) \quad \sum_{j=1}^{n+1} T(j, \omega) \end{array}$$

Osserviamo che i numeri aleatori sono definiti mediante due argomenti:

• Il primo è l'indice che identifica il singolo numero aleatorio all'interno della famiglia costituita dal processo stocastico, ed è detto PARAMETRO OPERATIVO.

Esso può essere: - Discreto: si parla di Processo Stocastico a parametro discreto. Ne sono un esempio i processi dei $T(j, \omega)$ e degli $S(n, \omega)$, in cui j e n sono numeri naturali.

- Continuo: si parla di Processo Stocastico a parametro continuo. Un esempio è costituito da $N(t, \omega)$, dove t è un numero reale positivo.

Nel caso in cui il Parametro Operativo assuma un significato temporale, si parlerà allora rispettivamente di Processo Stocastico a Tempo Discreto e a Tempo Continuo. $N(t, \omega)$ è un esempio di Processo a Tempo Continuo.

• Il secondo argomento, ω , è il caso elementare della partizione.

Per conoscere i processi a tempo discreto dal punto di vista probabilistico basta conoscere la famiglia di tutte le distribuzioni congiunte finite-dimensionali; mentre ciò non basta per quelli a tempo continuo.

Un processo si dice OSSERVABILE se vi è la possibilità di tenere conto di tutte le sue possibili traiettorie.

Poiché un'osservazione nel tempo fornisce una sequenza di informazioni che, si suppone, arricchisce lo stato di informazione, aggiungeremo allo Spazio di Probabilità una famiglia di sotto- σ -algebre crescenti con t che tenga conto di tale evoluzione informativa, ossia $\{\mathcal{F}_t\}$ tale che $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, che viene chiamata FILTRAZIONE.

- Sia $T(j, \omega)$ un processo osservabile, ossia siano osservabili i primi numeri aleatori T_1, T_2, \dots, T_m . Essi sono Indipendenti e Identicamente Distribuiti e in un Processo di Poisson essi hanno comune distribuzione Esponenziale Negativa di intensità λ .

Essi sono dunque dotati di densità $f_T(t) = \lambda e^{-\lambda t}$.

Nella realtà, però, non tutto è noto: l'intensità λ è ignota, dunque aleatoria, e perciò viene indicata con il simbolo θ . La comune densità delle variabili osservabili sarà allora $f_T(t, \theta) = \theta e^{-\theta t}$.

Supponiamo di osservare l'evento $K = \bigwedge_{j=1}^m (T_j = t_j)$.

Su questa base informativa cerchiamo di costruire una stima (valore approssimato del numero aleatorio) di θ , ossia $\hat{\theta}_m$, in maniera tale da ridurre al minimo l'errore.

Seguendo il procedimento di stima di Massima Verosimiglianza otteriamo:

$$L(\theta, K) = \prod_{j=1}^m f(t_j, \theta) = \theta^m e^{-\theta \sum_{j=1}^m t_j} \rightarrow \text{Funzione di Verosimiglianza (funzione di } \theta \text{)}$$

$$\text{Sia } \tau = \sum_{j=1}^m t_j : L'(\theta, K) = m \theta^{m-1} e^{-\theta \tau} - \theta^m e^{-\theta \tau} = 0 \Leftrightarrow \underbrace{\theta e^{-\theta \tau}}_{\neq 0} (m - \tau \theta) = 0 \Leftrightarrow \hat{\theta}_m = \frac{m}{\tau}$$

Il numero m delle osservazioni può essere diverso ogni volta a causa di diversi fattori (anche esterni, come i costi), perciò introduciamo un processo stocastico, che tiene inoltre conto del cambiamento di informatività.

Problema: non è detto che nel momento in cui λ noto viene sostituito con θ ignoto sia mantenuta l'ipotesi di Indipendenza e Identica Distribuzione?

L'evento K può essere rappresentato sulla retta reale come una sequenza di punti: $t_1, t_2, t_3, \dots, t_n$

Chiamiamo questa sequenza di punti TRAIETTORIA o REALIZZAZIONE del processo, riferita ad un singolo evento elementare $\omega \in \Omega$: se cambia ω , cambia la sequenza di punti.

Un processo stocastico può essere dunque interpretato in due modi complementari:

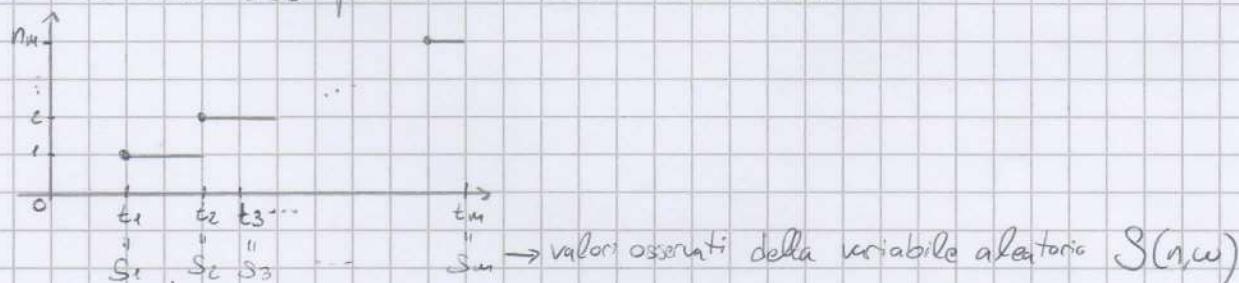
$$\{T(j, \omega), j \geq 1, \omega \in \Omega\}$$

- 1) Come famiglia di numeri aleatori: se fissiamo $\omega = \omega'$ abbiamo una Realizzazione o traiettoria del processo;
- 2) Come insieme di tutte le traiettorie possibili: se fissiamo $j=j'$ otteniamo un singolo numero aleatorio $T(j', \omega)$ dipendente da ω .

- Sia $N(t, \omega)$ un processo osservabile.

Supponiamo di osservare l'evento $K = \bigwedge_{j=1}^m [N(t_j, \omega) = n_j]$

La traiettoria del processo è descrivibile come:



Supporre osservabili un numeri aleatori fa sì che K rappresenti un INCREMENTO DI INFORMAZIONE rispetto allo stato di informazione precedente. Questa è una sostanziale differenza con la statistica classica, che invece considera come unico stato di informazione quello derivante dall'osservazione.

Poiché $N(t, \omega)$ è un Processo di Poisson risulta che i numeri aleatori

$N(s, \omega) \sim \text{Po}(s \lambda)$ IID, supponendo λ nota. $\Rightarrow \Pr(N(s, \omega) = n) = e^{-s\lambda} \frac{(s\lambda)^n}{n!}$

Se l'intensità λ non è nota, allora $\lambda = 0$. In tal caso

$\Pr(N(s, \omega) = n; 0) = e^{-s0} \frac{(s0)^n}{n!}$ e l'ipotesi di indipendenza e di identica distribuzione cade.

Supponiamo che passando da un'intensità nota ad una non nota rimanga valida questa valutazione di probabilità.

[Vedremo che in realtà l'espressione $\Pr(N(s_i, \omega) = n_i; \theta)$ andrebbe sostituita da una subordinazione $\Pr(N(s_i, \omega) = n_i | \theta)$].

Possiamo allora scrivere la funzione di Verosimiglianza: \rightarrow sequenza degli n_i non decrescente

$$L(\theta | K) = \prod_{i=1}^m \left[e^{-\theta(s_i - s_{i-1})} \frac{[\theta(s_i - s_{i-1})]^{n_i - n_{i-1}}}{(n_i - n_{i-1})!} \right] \quad \begin{matrix} n_1 & n_2 & n_{i-1} & n_i & n_m \\ s_1 & s_2 & s_{i-1} & s_i & s_m \end{matrix}$$

Trascurando le costanti moltiplicative si ottiene che la derivata prima è proporzionale a:

$$e^{-\theta s_m} \theta^{n_m-1} (-\theta s_m + n_m) = 0 \Rightarrow \hat{\theta}_{ml} = \frac{n_m}{s_m} \rightarrow \begin{matrix} \text{numero totale di eventi osservati} \\ \text{tempo totale di osservazione} \end{matrix}$$

Questo risultato è equivalente a quello ottenuto con $T(j, \omega)$, in quanto l'informazione acquisita è la stessa.

La scrittura $\Pr(N(s, \omega) = n; \theta) = e^{-\theta s} \frac{(\theta s)^n}{n!}$ è giustificata dal θ , condizionamento sui possibili valori del parametro non noto.

In generale, dato uno Spazio di Probabilità (Ω, \mathcal{F}, P) ,

sia $X(t, \omega)$ un processo stocastico tale che:

$\forall \omega \in \Omega$

$t \in \mathbb{R}_+$: parametro operativo di tempo da sì che lo studio verta sulla dinamica del calcolo della probabilità. In tal senso sostituiamo all'insieme delle probabilità coerenti una valutazione sulle loro evoluzioni nel tempo al variare dello stato di informazione.

$$(\Omega, \mathcal{F}, P) \xrightarrow{X(t, \omega)} (\mathbb{R}, \mathcal{B}, P_{X_t}^{-1}(-\infty, x))$$

↑ funzione di probabilità indotta da X

Il processo stocastico può essere interpretato in due modi complementari:

- Come insieme di Traiettorie: in funzione di t e fissando $\omega = \omega' \in \Omega$ si ottiene una Traiettoria o Realizzazione (o Sample Function) $X(t, \omega')$.
- Come famiglia di numeri aleatori: in funzione di ω e fissando $t = t' > 0$ si ottiene un numero aleatorio $X(t', \omega)$.

Notazione:

- X_n : processo a tempo discreto con $n \geq 0$ intero;
- $X(t)$: processo a tempo continuo con $t \geq 0$ reale.

Ciascun processo stocastico, come ciascun numero aleatorio, può essere considerato a diversi (infiniti) LIVELLI DI SPECIFICAZIONE

o DI CONOSCENZA.

Nel caso di un numero aleatorio X , per conoscerlo completamente possiamo avvalerci della funzione di ripartizione $F_X(x)$ (o della densità, se ne è dotato). Se il range di X è limitato la conoscenza di F_X è equivalente alla conoscenza di tutti i suoi momenti $E(X^n)$, $n \geq 1$.

Non sempre è possibile conoscere un numero aleatorio nella sua intera distribuzione, perciò molto diffuso è il più basso livello di specificazione su $E(X)$ e $\text{Var}(X)$.

Nel caso di un processo stocastico i principali livelli di conoscenza sul fenomeno considerati sono:

- LIVELLO MASSIMALE: Viene specificata la famiglia di tutte le distribuzioni congiunte finito-dimensionali della LEGGE TEMPORALE DEL PROCESSO ("Temporale" poiché il parametro operativo ha un'accelerazione temporale; nel caso generale è detta Legge di Probabilità del Processo).

$$(X_{n_1}, X_{n_2}, \dots, X_{n_m}) \Rightarrow \{F_{n_1, n_2, \dots, n_m}(x_1, \dots, x_m)\} \quad \forall m \rightarrow \text{Distribuzioni coerenti tra loro}$$

Consideriamo un processo stocastico a parametro discreto, allora possiamo sostituire alla famiglia $\{F_{n_1, \dots, n_m}(x_1, \dots, x_m)\}$ la famiglia $\{F_1(x), F_{1,2}(x,y), F_{1,2,3}(x,y,z) \dots\}$

Teorema di Kolmogorov: Le distribuzioni di probabilità congiunta

$\{F_1(x), F_{1,2}(x,y), F_{1,2,3}(x,y,z)\}$ sono COERENTI tra loro se e solo se ogni elemento della famiglia è distribuzione marginale per tutte le distribuzioni seguenti e ammette come sue distribuzioni marginali tutti gli elementi precedenti della famiglia.

Esempi: $\lim_{z \rightarrow +\infty} F_{1,2,3}(x,y,z) = F_{1,2}(x,y)$; $\lim_{y \rightarrow +\infty} F_{1,2}(x,y) = F_1(x)$

[Un analogo di questo teorema esiste per processi a tempo continuo].

Il livello di conoscenza massimale permette di determinare tutti i momenti.

- APPROCCIO DEL II ORDINE: Vengono specificati solo i momenti del I e del II ordine, ossia valori medi, varianze e covarianze.

Questo livello di conoscenza è molto diffuso nella pratica, soprattutto laddove l'informazione risulta essere limitata.

Questo approccio si riferisce unicamente alle due funzioni:

- $\varphi(n) = E(X_n)$ \Rightarrow FUNZIONE VALOR MEDIO : esprime tutti i momenti primi.

- $\psi(m,n) = \text{Cov}(X_m, X_n)$ \Rightarrow FUNZIONE DI COVARIANZA : esprime tutti i momenti secondi.
(se $m=n$ è la Varianza)

Esiste la Funzione di Covarianza solo se tutti i momenti secondi sono finiti.

Proprietà della Funzione di Covarianza:

$$1) \Psi(u, u) = V(X_u) \geq 0$$

$$2) \Psi(u, n) = \Psi(n, u)$$

Infatti, per definizione: $\Psi(u, n) = \text{Cov}(X_u, X_n) = E[(X_u - E(X_u))(X_n - E(X_n))]$

$$3) |\Psi(u, n)| \leq \sqrt{\Psi(u, u) \Psi(n, n)}$$

L'scarti quadratici medi

$$4) \sum_{i=1}^n \sum_{j=1}^n a_i a_j \Psi(i, j) \geq 0 \quad \forall a_i, a_j \in \mathbb{R}, \forall n \Rightarrow \text{E' una funzione semi-definita positiva}$$

Infatti:

$$0 \leq E\left[\left(\sum_{j=1}^n a_j X_j\right)^2\right] = E\left[\sum_{i=1}^n \sum_{j=1}^n a_i a_j X_i X_j\right] = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \overbrace{E(X_i X_j)}$$

Supponiamo, senza perdita di generalità,
che $E(X_i) = 0 \quad \forall i$ (numer. aleatori equi)

Se ordiniamo le covarianze in una matrice otteniamo la Matrice di Varianza e covarianza Ψ semi-definita positiva (ossia $\underline{Y}^T \Psi \underline{Y} \geq 0 \quad \forall \underline{Y} \in \mathbb{R}^n$):

$$\Psi = \text{Cov}(\underline{X}) = \begin{bmatrix} \Psi(1,1) & \Psi(1,2) & \dots & \Psi(1,n) \\ \Psi(2,1) & \Psi(2,2) & \dots & \Psi(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ \Psi(n,1) & \Psi(n,2) & \dots & \Psi(n,n) \end{bmatrix} = E\left[\left(\underline{X} - E(\underline{X})\right)\left(\underline{X} - E(\underline{X})\right)^T\right]$$

Qualsiasi funzione di due argomenti che soddisfa tali proprietà è una Funzione di Covarianza.

Cenni storico: I Processi Stocastici sono nati agli albori del secolo scorso.

Prima il calcolo delle probabilità era considerato solo una sorta di curiosità di carattere aritmetico, in quanto non esisteva una teoria matematica soddisfacente in merito. Si dice che tale disciplina sia stata formalizzata negli anni '30 con una monografia di Kolmogorov e con Bruno de Finetti (docente a Trieste); a Kolmogorov viene attribuito il merito di aver costruito formalmente il calcolo delle probabilità (1933), mentre de Finetti fondò la nozione di probabilità soggettiva (1935).

Mentre per Kolmogorov la probabilità è un caso particolare della Teoria della Misura, per de Finetti non esiste una probabilità oggettiva, bensì soggettiva, legata all'individuo e al suo stato di informazione. La teoria sui processi stocastici nasce: a Mosca con Kolmogorov, a Parigi con Lévy e a Copenaghen con Kramer. In Italia i massimi esponenti furono Cantelli e de Finetti.

A: L'ini dell'inferenza bayesiana considereremo i tre principali schemi di dipendenza stocastica presenti nei seguenti processi stocastici:

① PROCESSI DI MARKOV

Markov fu un matematico russo che nel 1906 formulò una struttura di dipendenza stocastica tra numeri aleatori. Egli si occupava di linguistica e voleva valutare le probabilità che all'ultimo posto in una parola comparisse una consonante in funzione delle lettere precedenti. La sua idea fu quella che l'influenza prevalente dipende dal carattere immediatamente precedente.

Tale studio di dipendenza fu una novità, in quanto in precedenza (secoli XXVII, XXVIII, XXIX) si considerava più comunemente l'indipendenza. Ci si accorse però che nelle applicazioni raramente i numeri aleatori possono essere assunti indipendenti tra loro (separata assurzere facilità i calcoli), dunque si iniziarono ad ipotizzare particolari modelli di dipendenza stocastica.]

Consideriamo il processo a tempo discreto costituito dai numeri aleatori $X_{n_1}, X_{n_2}, \dots, X_{n_m}, \dots$

Si parla di DIPENDENZA MARKOVIANA quando:

$$F(X_{n_m} | X_{n_1}, \dots, X_{n_{m-1}}) = F(X_{n_m} | X_{n_{m-1}})$$

Funzione di ripartizione di X_{n_m} condizionata ad una particolare ipotesi sull'ultimo punto sui valori precedenti

Supponiamo quindi che X_{n_m} sia un numero aleatorio non ancora osservabile mentre $X_{n_{m-1}}$ sia osservabile oggi. I numeri aleatori che precedono $X_{n_{m-1}}$ (ovvero che rappresentano il passato) non interverranno nella valutazione di X_{n_m} .

Dunque nell'ipotesi di dipendenza markoviana futuro e passato sono stocasticamente indipendenti, noto il presente; del passato conta solo, dunque, l'ultima osservazione disponibile. Perciò tali processi sono anche detti Processi A MEMORIA CORTA.

(2) PROCESSI STAZIONARI

Nascono negli anni '30 e rappresentano la svolta della statistica bayesiana, in quanto sostituiscono la nozione di Campione casuale semplice.

Il processo $X_{n_1}, X_{n_2}, \dots, X_{n_m}, \dots$ è detto PROCESSO STAZIONARIO se:

$$F(X_{n_1}, X_{n_2}, \dots, X_{n_m}) = F(X_{n+h}, X_{n+h}, \dots, X_{n+h}) \quad \forall n, n_1, \dots, n_m, \forall h \text{ intero arbitrario}$$

con $n+h, \dots, n+m \geq 0$ (anche negativo)

Si parla in questo caso di INVARIANZA PER TRASLAZIONI RIGIDE del parametro operativo intendendo la trasformazione di n, \dots, n_m in $n+h, \dots, n+m$.

In questo caso tutto il passato influenza la valutazione sul futuro.

Esempi:

- $m=1 \Rightarrow F(X_{n_1}) = F(X_{n+h}) \quad \forall h$

Cioè implica che tutti i numeri aleatori del processo hanno la medesima distribuzione univariata.

- $m=2 \Rightarrow F(X_{n_1}, X_{n_2}) = F(X_{n+h}, X_{n+h})$

Cioè non implica che tutte le coppie di numeri aleatori possibili abbiano la medesima distribuzione congiunta (infatti $F(X_1, X_3) = F(X_5, X_7)$, ma non è detto che $F(X_1, X_3) = F(X_1, X_2)$). Essa è presente solo nelle coppie di numeri aleatori per le quali è costante la differenza tra i parametri operativi.

(3) PROCESSI MARTINGALA

Nascono negli anni '30 ad opera dei Francesi Ville, Lévy, Borel.

Il processo $X_{n_1}, X_{n_2}, \dots, X_{n_m}, \dots$ è detto MARTINGALA se:

$$E[X_{n_m} | X_{n_1}, \dots, X_{n_{m-1}}] = X_{n_{m-1}}$$

\downarrow
Funzione di Regressione

Il numero aleatorio X_{n_m} nella sua valutazione non viene influenzato da tutti i numeri del passato, ma solo dal più recente.

Il riferimento è alla funzione di Regressione (Conditional Expectation) $\Psi(X_{n_1}, \dots, X_{n_{m-1}})$.

Il termine inglese "Conditional Expectation" è sovraintende, in quanto non distingue tra:

- $E(Y|X) = \Psi(X)$: numero aleatorio dipendente da $X \rightarrow$ Funzione di Regressione
- $E(Y|X=x)$: numero certo \rightarrow Valore medio condizionato

PROCESSI DI MARKOV

Il lavoro di Markov iniziato nel 1906 fu sviluppato da Kolmogorov (1931) nei suoi Metodi Analitici del Calcolo delle Probabilità e da William Feller (1936), che parla di un approccio analitico sulle probabilità di transizione. Tale approccio fu in seguito considerato nelle traiettorie dal probabilista giapponese K. Ito (1944).

Consideriamo le variabili aleatorie $X_{n_1}, X_{n_2}, \dots, X_{n_m}, \dots$, ciascuna con N stati possibili: $x_1, x_2, \dots, x_n, \dots, x_N$.

Tra esse c'è DIPENDENZA MARKOVIANA se: $F(X_{n_m} | X_{n_1}, \dots, X_{n_{m-1}}) = F(X_{n_m} | X_{n_{m-1}})$

n_1, n_2, \dots, n_m . Se la funzione di ripartizione è derivabile possono fornire la corrispondente condizione per la funzione di densità: $f(X_{n_m} | X_{n_1}, \dots, X_{n_{m-1}}) = f(X_{n_m} | X_{n_{m-1}})$

Considerazione la fattorizzazione:

$$f(X_{n_1}, \dots, X_{n_m}) = f(X_{n_m} | X_{n_1}, \dots, X_{n_{m-1}}) f(X_{n_1}, \dots, X_{n_{m-1}}) = f(X_{n_m} | X_{n_{m-1}}) \left[\underbrace{f(X_{n_{m-1}} | X_{n_1}, \dots, X_{n_{m-2}})}_{f(X_{n_{m-1}} | X_{n_{m-2}})} f(X_{n_1}, \dots, X_{n_{m-2}}) \right] \\ = \dots = f(X_{n_1}) \cdot \prod_{j=2}^m f(X_{n_j} | X_{n_{j-1}})$$

Conseguenze:

- La dipendenza markoviana implica che la funzione di ripartizione congiunta possa essere scritta usando solo le univariate e le condizionate univariate.
- Siano $l < m < n$ Teorema delle probabilità composte

$$f(X_n | X_l) = \int_R f(X_n, X_m | X_l) dX_m = \int_R \underbrace{f(X_n | X_m, X_l)}_{f(X_n | X_m)} f(X_m | X_l) dX_m$$

Questa relazione rappresenta una CONDIZIONE DI COERENZA tra densità condizionate.
(CONDIZIONE DI CHAPMAN-KOLMOGOROV)

Distinguiamo i Processi di Markov per:

- Tempo Discreto
continuo

- Numero di Stati Discreto
Continuo

Definizione: Un processo di Markov i cui numeri aleatori hanno tutti i medesimi N (discreto) stati possibili è detto CATENA MARKOVIANA.

CASO 1: CATENE DI MARKOV A TEMPO DISCRETO

Consideriamo la Catena Markoviana costituita dai numeri aleatori:

$X_0, X_1, X_2, \dots, X_n, \dots$ ciascuno con i medesimi N (discreto) stati possibili.

Sia $\Pr(X_{n+1}=j | X_0=a, X_1=b, \dots, X_n=i) \stackrel{\Delta}{=} P_{ij}(n)$ una

Probabilità condizionata di TRANSIZIONE (andrebbe indicata con il simbolo P_{ij} , ma per consuetudine si preferisce P_{ij}).

Definizione: Una Catena di Markov si dice OMOGENEA se $P_{ij}(n)=P_{ij} \forall i, j$, ovvero se le probabilità di Transizione condizionate non dipendono da n .

Sia la Catena Markoviana Omoogenea, ossia $P_{ij}(n)=P_{ij}$.

Sia \underline{a}_0 la distribuzione di probabilità di X_0 : gli elementi di a_0 .

Sono non negativi e di somma 1.

\underline{a}_0 è detto CONDIZIONE INIZIALE.

Sia P la matrice delle probabilità condizionate di Transizione.

Risulta allora che:

- $\underline{a}_1 = \underline{a}_0 P \rightarrow$ Distribuzione di probabilità di X_1

$$\text{Infatti: } \Pr(X_1=j) = \sum_{i=1}^N a_{0i} P_{ij}$$

- $\underline{a}_n = \underline{a}_{n-1} P = \underline{a}_0 P^{n-m} = \underline{a}_0 P^n \rightarrow$ Distribuzione di probabilità di X_n

Indichiamo con $P_{ij}^{(n)}$ il generico elemento della matrice P^n , ossia la probabilità di passare dallo stato i allo stato j in n passi.

Condizione di Coerenza di CHAPMAN-KOLMOGOROV: $P_{ij}^{(n)} = \sum_{h=1}^N P_{ih}^{(m)} P_{hj}^{(n-m)}$

\Rightarrow Si deve poter passare da i a j in n passi, passando da i a h in m passi e da h a j in $n-m$ passi, facendo variare h in tutti i modi possibili.

Dal punto di vista matriciale deve risultare che: $P^n = P^m \cdot P^{n-m}$ ($m < n$).

Definizione: Una Catena si dice REGOLARE quando esiste un intero n^* tale che $\forall n > n^*$ tutti gli elementi della matrice P^n sono positivi, ossia è positiva (definitivamente) la probabilità di passare da un qualsiasi stato i a un qualsiasi stato j in un numero finito di passi.

TEOREMA DI MARKOV: In una Catena Regolare la matrice P^n converge ad una matrice U con righe tutte uguali ed elementi tutti positivi, ossia:

$$\lim_{n \rightarrow \infty} P^n = U$$

(Ciò implica che la distribuzione asintotica è stazionaria).

Di conseguenza la distribuzione di probabilità da un certo punto in poi resta la stessa, infatti, sia \underline{u} la generica riga della matrice U , risulta che $\underline{u} = \underline{u}P$.

Possiamo concludere quindi che, in questo caso, la distribuzione del processo dipende solo dalla Condizione Iniziale $\underline{\alpha}_0$ e dalla matrice P .

CASO 2: CATENE DI MARKOV A TEMPO CONTINUO

Consideriamo la Catena di Markov costituita dai numeri aleatori: $X(t)$, $t \geq 0$, ciascuno con i medesimi N (discreto) stati possibili.

Poiché il parametro operativo è continuo, dovremo considerare:

- a_0 : distribuzione iniziale

- $\{\theta(t), t \geq 0\}$ = un'infinità continua di matrici $\theta(t)$ tali che:

$$\begin{aligned} - p_{ij}(t) &= \sum_{h=1}^N p_{ih}(s) \theta_{hj}(t-s) \quad \forall i, j, t \rightarrow \text{Condizione di Coerenza di} \\ &\quad \xrightarrow{\text{generico elemento di } \theta(t)} \text{Chapman-Kolmogorov. In termini} \\ &\quad \text{matriciali: } \theta(t) = \theta(s)\theta(t-s) \end{aligned}$$

$$\left. \begin{aligned} - p_{ij}(t) &\geq 0 \quad \forall i, j, t \\ - \sum_{j=1}^N p_{ij}(t) &= 1 \quad \forall i, t \end{aligned} \right\} \text{condizioni che definiscono } \theta(t) \text{ come una} \\ \xrightarrow{\theta(t) \cdot \underline{1} = \underline{1}} \text{MATRICE STOCASTICA}$$

La cardinalità infinita dell'insieme $\{\theta(t), t \geq 0\}$ ne rende complesso l'utilizzo e il rispetto delle condizioni di coerenza, per cui ricerchiamo una forma più semplice.
 \uparrow condizione molto forte

TEOREMA DI MARKOV (caso continuo): Se tutti gli elementi di tutte le matrici $\theta(t)$ sono positivi $\forall t$, allora $\theta(t)$ converge ad una matrice U

stazionaria, ossia tale che:

- $\underline{u} = \underline{u}\theta(t) \quad \forall t$
- $\underline{u} = \underline{u}U$

Teorema:

- $\theta(0) = I_N = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}_{N \times N}$
 Al tempo 0 non avvengono transizioni: se il sistema si trova in uno stato, esso rimane tale
- $\lim_{t \rightarrow 0} \theta(t) = \theta(0) = I_N$
 $\xrightarrow{\text{continuità nell'origine}}$
- $\theta(\frac{t}{2}) = \theta(s)\theta(t-s)$
 (Chapman-Kolmogorov)

Esiste una matrice unica G ($N \times N$)

\Rightarrow i cui elementi g_{ij} sono tali che:

$$\begin{cases} g_{ij} \geq 0 \quad \forall i, j \\ g_{ij} = -\sum_{j \neq i} g_{ij} \end{cases}$$

I g_{ij} sono detti INTENSITÀ DI TRANSIZIONE

La matrice G è tale che $\forall t > 0$ risulta:

$$P(t) = e^{tG} = \sum_{n=0}^{+\infty} \frac{(tG)^n}{n!}$$

ESPOENZIALE
DI MATRICE

Grazie a questo risultato tutte le matrici $P(t)$ possono essere determinate a partire da G , che è perciò detta GENERATORE DELLA CATENA A TEMPO CONTINUO.

Proprietà di G :

- G non è una matrice stocastica, in quanto la somma degli elementi di ogni riga è 0.
- Gli elementi g_{ij} sono detti Intensità in quanto, per t piccolo, risulta che $P_{ij}(t) = g_{ij}t + o(t)$
 $\xrightarrow{L \rightarrow \text{infinitesimo}}$

In questo caso, quindi, gli elementi per determinare la legge del processo sono a. e il generatore G .

Teorema:

- $\exists (0, t]$ di lunghezza finita esiste un numero finito di transizioni
- Nella transizione da i a j ci si ferma per un certo tempo T_j nello stato j e questi numeri aleatori T_j sono indipendenti con distribuzione esponenziale negativa, ossia: $f_{T_j}(t) = -g_{jj}e^{-tg_{jj}}$
(con $g_{jj} = -\sum_{h \neq j} g_{jh}$)
- $\forall i \sum_{j=1}^N P_{ij}(t) \cdot g_{jj} < +\infty$

⇒

$\bullet P'(t) = P(t) \cdot G$
EQUAZIONE DIFFERENZIALE PROSPETTIVA o FORWARD EQUATION

$\bullet P'(t) = G \cdot P(t)$
EQUAZIONE DIFFERENZIALE RETROSPETTIVA o BACKWARD EQUATION

(Equazioni Differenziali DETERMINISTICHE: le incognite sono delle funzioni)

Problema: Calcolare $\beta(t)$ sapendo che $G = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$, $\lambda, \mu > 0$.

Ci sono tre modi:

1) Metodo dell'approssimazione:

$$\beta(t) = e^{tG} = \sum_{n=0}^{\infty} \frac{(tG)^n}{n!} = I_N + tG + \frac{1}{2!} t^2 G^2 + \dots$$

Dunque, approssimando la somma,

$$\begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} + \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} + \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} = \begin{bmatrix} \lambda(\lambda+\mu) & -\lambda(\lambda+\mu) \\ -\mu(\lambda+\mu) & \mu(\lambda+\mu) \end{bmatrix} = (\lambda+\mu) \begin{bmatrix} \lambda & -\lambda \\ -\mu & \mu \end{bmatrix}$$

$$\text{Ottieniamo per esempio: } P_{11}(t) = 1 + t \underbrace{G_{11}}_{-\lambda} + \frac{1}{2!} t^2 (\lambda+\mu) \lambda + \dots$$

2) Approccio Analitico di Kolmogorov-Feller:

Questo metodo permette di ottenere direttamente gli elementi di $\beta(t)$ risolvendo le equazioni differenziali del teorema precedente:

- $\dot{\beta}(t) = \beta(t) \cdot G \rightarrow \text{PROSPETTIVA}$
- $\dot{\beta}^*(t) = G \cdot \beta(t) \rightarrow \text{RETROSPETTIVA}$

3) Approccio di Itô:

In questo metodo l'obiettivo non sono le probabilità condizionate di transizione, bensì le traiettorie del processo come soluzioni di EQUAZIONI DIFFERENZIALI STOCASTICHE (in cui la funzione incognita è un processo stocastico).

Esempio di equazione differenziale stocastica: $dX(t) = -dX(t) + dW(t)$

↓ processo stocastico noto

Processo stocastico incognito

Sia h un intervallo di tempo opportunamente piccolo.

Siano $P_{ij}(h)$ le probabilità di transizione dallo stato i allo stato j nel tempo h .

$= \Pr(X_{t+h} = j | X_t = i) \rightarrow$ non dipende da t perché la catena è OMogenea

Se h è sufficientemente piccolo possiamo considerare le seguenti approssimazioni:

$$\begin{cases} P_{ij}(h) = g_{ij} h + o(h) & \rightarrow \text{infinitesimo di ordine superiore ad } h \rightarrow \text{tiene conto del fatto che} \\ & \text{il sistema possa passare da } i \\ P_{ii}(h) = 1 + g_{ii} h + o(h) & \text{a } j \text{ in più di una transizione} \end{cases}$$

(con h piccolo la probabilità è praticamente nulla)

Alla base vi è l'ipotesi che in un tempo scalo h non avvenga in generale più di una transizione.

Omettendo gli infinitesimali possiamo scrivere: $\begin{cases} P_{ij}(h) \approx g_{ij} h \\ P_{ii}(h) \approx 1 + g_{ii} h \end{cases}$

Proposizione: $\lim_{h \downarrow 0} \frac{1}{h} [\underbrace{\mathbf{P}(h) - \mathbf{I}_N}_{\mathbf{P}'(0)}] = \mathbf{G}$ → \mathbf{G} rappresenta la derivata prima rispetto al tempo della matrice $\mathbf{P}(t)$ per $t=0$

Dove $\mathbf{P}'(0) = [p_{ij}'(0)] = \frac{d}{dt} [p_{ij}(t)] \Big|_{t=0}$

La derivata della matrice è la matrice delle derivate

(a rigore sarebbe una derivata parziale su t lasciando inalterati i e j)

Per quanto riguarda il generico elemento p_{ij} (con $i \neq j$): $\lim_{h \downarrow 0} \frac{p_{ij}(h) - 0}{h} = \frac{d}{dh} p_{ij}(h)$

Dimostrazione

- Ottieniamo \mathbf{G} da $\mathbf{P}(t)$: per il risultato precedente

Supponiamo che $\forall i \quad 1 = \sum_{j=1}^N p_{ij}(h) \cong 1 + h \sum_{j=1}^N g_{ij}$
 $\mathbf{P}(t)$ matrice stocistica

$$\text{Questo accade poiché: } = (1+g_{ii}h) + \left(\sum_{j \neq i} g_{ij}h\right)$$

$$= 1 + \underbrace{\left(g_{ii} + \sum_{j \neq i} g_{ij}\right)h}_{\text{per verificare l'approssimazione}} \quad \begin{cases} g_{ii} \geq 0 \text{ per } i \\ g_{ii} = -\sum_{j \neq i} g_{ij} \end{cases}$$

ossia se g_{ij} sono elementi della matrice \mathbf{G}

- Ottieniamo $\mathbf{P}(t)$ da \mathbf{G} (supposta nota):

Applichiamo le condizioni di coerenza di Chapman-Kolmogorov:

$$p_{ij}(t+h) = \sum_{k=1}^N p_{ik}(t)p_{kj}(h) \cong p_{ij}(t)(1+g_{ij}h) + \sum_{k \neq j} p_{ik}(t) \cdot (g_{kj}h) =$$

h molto piccolo, quindi $t+h \approx t$ per le ipotesi precedenti

$$= p_{ij}(t) + h \sum_{k=1}^N p_{ik}(t) g_{kj} \Leftrightarrow p_{ij}(t+h) - p_{ij}(t) \cong h \sum_{k=1}^N p_{ik}(t) g_{kj} / h$$

$$\Leftrightarrow \frac{p_{ij}(t+h) - p_{ij}(t)}{h} \cong \sum_{k=1}^N p_{ik}(t) g_{kj} \Rightarrow h \downarrow 0 \Rightarrow P'_i(t) = \sum_{k=1}^N p_{ik}(t) g_{kj}$$

Ottieniamo un sistema lineare di equazioni differenziali nelle funzioni

incognite $p_{ij}(t)$, che può essere scritto sinteticamente come:

$$\mathbf{P}'(t) = \mathbf{P}(t) \cdot \mathbf{G} \Rightarrow \text{Equazione Differenziale Prospettiva}$$

Analogamente si può giungere al sistema di:

$$\mathbf{P}'(t) = \mathbf{G} \cdot \mathbf{P}(t) \Rightarrow \text{Equazione Differenziale Retrospettiva.}$$

Utilizziamo dunque l'approccio analitico di Kolmogorov-Feller cercando una soluzione a queste equazioni differenziali.

CENNO SU EQUAZIONI DIFFERENZIALI

Le Equazioni Differenziali appartengono all'ambito delle Equazioni Funzionali (vasto ambito che comprende differenziali, integrali, a differenze finite,...), in cui le incognite non sono numeri, bensì funzioni.

In un'equazione differenziale compaiono una funzione incognita e le sue derivate.

Esempi:

$$\bullet \quad y'(t) = y(t) \Rightarrow y(t) = K e^t \quad : \text{Equazione differenziale del I ordine lineare}$$

↓
costante corrispondente al valore
della funzione in un punto assegnato

Compare solo la derivata prima

Per risolvere l'equazione bisogna impostare la cosiddetta Condizione Iniziale:

$$\left. \begin{array}{l} y'(t) = y(t) \\ y(0) = K \end{array} \right\} \rightarrow \text{Condizione Iniziale}$$

(Questa particolare equazione è detta
"Problema di Cauchy")

$$\left. \begin{array}{l} y'(t) = y(t) \\ y(0) = K \end{array} \right\} \downarrow \quad \text{istante iniziale}$$

$$\bullet \quad y''(t) + y(t) = 0 \Rightarrow \left. \begin{array}{l} y(t) = \sin(t) \text{ se } y(0)=0 \\ y(t) = \cos(t) \text{ se } y(0)=1 \end{array} \right\} \quad : \text{Equazione differenziale lineare}$$

del II ordine
(compare la derivata seconda)

Queste equazioni sono dette ORDINARIE, ossia intervengono funzioni di una sola variabile. Possono anche prendere in considerazione funzioni di più variabili; in tal caso interverrà anche la derivata parziale.

Esempio: Equazione di Laplace: $\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0$

Nei nostri problemi affronteremo prevalentemente equazioni lineari ordinarie del I ordine:

$$y'(t) = a y(t) + b$$

Casi:

a) $a=1, b=0$: $y'(t) = y(t) \Rightarrow y(t) = K e^t$

b) $b=0$: $y'(t) = a y(t) \Rightarrow y(t) = K e^{at}$

c) $a \neq 0, b \neq 0$ costanti (o funzioni note del tempo):

$$y'(t) = a y(t) + b \Rightarrow y(t) = K e^{at} + b \int_0^t e^{a(t-s)} ds$$

Adottiamo l'approccio analitico di Kolmogorov-Feller:

$$P'(t) = \begin{cases} \beta(t) \cdot G \\ G \cdot \beta(t) \end{cases} \Rightarrow \begin{bmatrix} P_{11}'(t) & P_{12}'(t) \\ P_{21}'(t) & P_{22}'(t) \end{bmatrix} = \begin{bmatrix} p_{11}(t) & p_{12}(t) \\ p_{21}(t) & p_{22}(t) \end{bmatrix} \cdot \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

Sistema lineare di equazioni differenziali del I° ordine

$$\Rightarrow \begin{cases} P_{11}'(t) = p_{11}(t)g_{11} + p_{12}(t)g_{21} \\ P_{12}'(t) = p_{11}(t)g_{12} + p_{22}(t)g_{22} \\ P_{21}'(t) = p_{21}(t)g_{11} + p_{22}(t)g_{21} \\ P_{22}'(t) = p_{21}(t)g_{12} + p_{22}(t)g_{22} \end{cases}$$

Consideriamo la prima equazione: $P_{11}'(t) = p_{11}(t)g_{11} + p_{12}(t)g_{21}$

Sappiamo che $p_{12}(t) = 1 - p_{11}(t) \Rightarrow P_{11}'(t) = p_{11}(t)g_{11} + (1 - p_{11}(t))g_{21}$

Ottieniamo un'equazione nella sola funzione incognita $p_{11}(t)$:

$$P_{11}'(t) = \underbrace{(g_{11} - g_{21})}_{<0 \text{ poiché } g_{11} \leq 0 \text{ e } g_{21} \geq 0} p_{11}(t) + g_{21}$$

Se questa equazione ammette soluzioni, queste saranno infinite. Per ottenere un'unica soluzione dobbiamo attribuire alla funzione incognita un qualche valore dipendente dalle nostre informazioni in un certo istante, ovvero dobbiamo fissare la CONDIZIONE INIZIALE.

In base all'informazione iniziale fissiamo $p_{11}(0) = 1$.

Sia $G = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$, allora il sistema lineare è: $\begin{cases} P_{11}'(t) = -(\lambda + \mu)p_{11}(t) + \\ p_{11}(0) = 1 \end{cases}$

Ottieniamo come soluzione: $p_{11}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t}$

Per ottenere $p_{21}(t)$ basta calcolare $p_{21}(t) = 1 - p_{11}(t)$

Analogamente si ottiene $p_{22}(t)$ come soluzione dell'equazione e in seguito $p_{22}^{(t)} = 1 - p_{11}(t)$

Per $t \rightarrow +\infty$ ottieniamo: $p_{11}(t) \rightarrow \frac{\mu}{\lambda + \mu}$ e $p_{22}(t) \rightarrow \frac{\lambda}{\lambda + \mu}$

Questi due valori costituiscono la DISTRIBUZIONE STAZIONARIA della catena

asintotica: $\underline{u} = \left(\frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu} \right)$ tale che $\underline{u} = \underline{u} \cdot \beta(t) \quad \forall t$; $\underline{u} \cdot G = \underline{0}$.

Cioè accade solo se tutti gli elementi di $\beta(t)$ sono definitivamente positivi.

Un esempio di Catenza Markoviana a tempo continuo è il Processo di Poisson, poiché il parametro operativo è continuo e il numero di Stati possibili è numerabile.

L'intensità del processo è un elemento della matrice G , che in questo caso è una matrice infinita:

$$G = \begin{matrix} \text{stato: } 0 & 1 & 2 & \dots \\ \left[\begin{array}{cccccc} 0 & -\lambda & \lambda & 0 & 0 & 0 & \dots \\ 1 & 0 & -\lambda & \lambda & 0 & 0 & \dots \\ 2 & 0 & 0 & -\lambda & \lambda & 0 & 0 & \dots \\ \vdots & & & & & \ddots & \end{array} \right] \end{matrix}$$

↳ diagonale principale di $-\lambda$

Per passare dallo stato 0 allo stato 1 l'intensità è λ . Non si può passare ad un altro stato:
 o si rimane in 0,
 o si passa in 1.

Proprietà del Processo di Poisson:

- 1) È una Catenza di Markov
- 2) È ad incrementi indipendenti su intervalli di tempo disgiunti (o eventi un unico punto in catena): $N(t_1)-N(t_0), N(t_2)-N(t_1), N(t_3)-N(t_2) \rightarrow$ indipendenza
 $t_1 \quad t_2 \quad t_3 \quad t_4$ (può anche essere $t_2 \equiv t_3$)
- 3) È ad incrementi omogenei (o stazionari), ossia: $N(t)-N(s) \stackrel{d}{=} N(t+h)-N(s+h) \quad \forall h$.
 Dunque incrementi relativi ad intervalli di medesima lunghezza sono identicamente distribuiti.

Un processo di Markov che abbia incrementi indipendenti e omogenei è detto PROCESSO DI LÉVY.

CASO 3: PROCESSI CON UN CONTINUO DI STATI A TEMPO DISCRETO

Sia $K(X, \mathcal{B})$ una FUNZIONE DI TRANSIZIONE in un periodo unitario dallo stato X all' insieme boreiano di stati $\mathcal{B}(\mathbb{R})$.

Tale funzione è anche chiamata STOCHASTIC KERNEL. NUCLEO STOCASTICO

[A rigore non possiamo definire tale funzione come probabilità di transizione, ovvero $K(X, \mathcal{B}) = \Pr(X_n \in \mathcal{B} | X_{n-1} = x)$, in quanto l'ipotesi condizionante ($X_{n-1} = x$), poiché siamo in un continuo di stati, ha probabilità 0.]

Una funzione $K(X, \mathcal{B})$ è un Nucleo Stocastico se:

- 1) $\forall \mathcal{B}$ $K(\cdot, \mathcal{B})$ è misurabile rispetto alla σ -algebra di (Ω, \mathcal{F}, P) , ossia $\mathcal{B} \in \mathcal{B}(\mathbb{R})$;
- 2) $\forall x \in \mathbb{R}$ $K(X, \cdot)$ è una misura di probabilità sulla σ -algebra dei boreiani di \mathbb{R} .

In particolare, se $\mathcal{B} = \mathbb{R}$ allora $K(X, \mathbb{R}) = 1$.

Consideriamo transizioni in più periodi, indicate con il simbolo $K^{(n)}(X, \mathcal{B})$.

Vi applichiamo le condizioni di coerenza di Chapman - Kolmogorov:

$$K^{(n)}(X, \mathcal{B}) = \int_{\mathbb{R}} K^{(n)}(X, dy) K^{(n-m)}(y, \mathcal{B}) \quad \rightarrow \text{nei primi } m \text{ passi si passa da } X \text{ a un intorno di } y, \text{ mentre nei successivi } n-m \text{ si arriva a } \mathcal{B}.$$

Se accade che $K(X, \mathcal{B}) = \int_{\mathbb{R}} K(x, y) dy$ allora esiste la DENSITÀ

DI TRANSIZIONE $K(x, y)$ a cui possiamo applicare la condizione

di Coerenza: $K(x, y) = \int_{\mathbb{R}} K^{(m)}(x, z) K^{(n-m)}(z, y) dz$.

Esempio: $X_n = aX_{n-1} + \varepsilon_n \Rightarrow$ MODELLO AUTOREGRESSIVO DI ORDINE 1

[Questo modello è tipico delle serie storiche approssimativamente stazionarie; in generale è detto ARMA(p, q) con $p=1$ e $q=0$ (Auto-Regressive Moving Average)].

ε_n è un processo stocastico, perciò in questa equazione funzionale si parla di EQUAZIONE ALLE DIFFERENZE FINITE STOCASTICA con soluzione costituita da una successione (poiché X_n è a tempo discreto).

L'equazione è inoltre lineare e a coefficienti a costanti.

Poiché ammette infinite soluzioni, per ottenerne una unica si fissano CONDIZIONE INIZIALE: $X_0 = k$

Comunemente si assume che $E_n \sim NWN(0, \sigma^2_\varepsilon)$, detto PROCESSO NORMAL WHITE NOISE: un processo stocastico a tempo discreto con variabili eque, non correlate tra loro e aventi la medesima varianza σ^2_ε .

[Questo processo è molto diffuso, in quanto rappresenta un "rumore", o errore, non prevedibile, impossibile da eliminare in quanto dovuto ad una miscela di cause minori (errore accidentale, distinto dall'errore sistematico, eliminabile di taratura).]

In generale si ha un processo WHITE NOISE se:

- $\text{Cov}(E_m, E_n) = \begin{cases} \sigma^2_\varepsilon & \text{se } m=n \\ 0 & \text{se } m \neq n \end{cases}$
- $E(E_n) \equiv 0 \forall n$

In particolare il processo si dice NORMAL WHITE NOISE se tutte le distribuzioni, sia univariate che multivariate, sono gaussiane. In tal caso la non correlazione implica la stocastica indipendenza.

Risolviamo l'equazione $X_n = aX_{n-1} + E_n$ con un procedimento iterativo ponendo $X_0 = K$:

- $X_1 = aX_0 + E_1 = aK + E_1 \Rightarrow X_1 \sim N(aK, \sigma^2_\varepsilon) \Rightarrow$ Il risultato sarà un processo stocastico di cui conosciamo la distribuzione
- $X_2 = aX_1 + E_2 = a^2K + aE_1 + E_2$
- \dots
- $X_n = a^nK + (E_n + aE_{n-1} + a^2E_{n-2} + \dots + a^{n-1}E_1) = a^nK + \sum_{h=0}^{n-1} a^h E_{n-h} \Rightarrow$ Soluzione dell'Equazione alle differenze finite stocastica

Il processo è dunque Gaussiano, in quanto la somma di numeri indipendenti è normale, perciò esso può essere completamente determinato attraverso i primi due momenti:

- Funzione Valor Medio: $\Psi(n) = E(X_n) = E\left[a^nK + \sum_{h=0}^{n-1} a^h E_{n-h}\right] = a^nK + \sum_{h=0}^{n-1} a^h E(E_{n-h}) = a^nK \Rightarrow$ costante
- Funzione di Covarianza: $\Psi(m, n) = \text{Cov}(X_m, X_n) = E[(X_m - a^mK)(X_n - a^nK)] =$
 $= E\left[\left(\sum_{k=0}^{m-1} a^k E_{m+k}\right) \cdot \left(\sum_{h=0}^{n-1} a^h E_{n-h}\right)\right] = E[(E_m + aE_{m-1} + a^2E_{m-2} + \dots + a^{m-1}E_1)(E_n + aE_{n-1} + \dots + a^{n-1}E_1)]$

Poiché $E[E_h E_k] = \begin{cases} \sigma^2_\varepsilon & \text{se } h=k \\ 0 & \text{se } h \neq k \end{cases}$ per incorrelazione, otteniamo contributi non nulli solo quando $h=k$, ossia da noi in poi:
 $= E[a^m E_m^2] + E[a^m \cdot a \cdot E_{m-1}^2] + \dots + E[a^m E_1^2] = a^m \sigma^2_\varepsilon + a^{m-1} \sigma^2_\varepsilon + \dots + a^2 \sigma^2_\varepsilon = a^m \sigma^2_\varepsilon (1 + a + a^2 + \dots + a^{m-1})$
 $= a^{n-m} \sigma^2_\varepsilon \cdot \frac{1-a^n}{1-a}$ \Rightarrow In particolare, se $m=n$ otteniamo la Funzione di Varianza: $\Psi(n, n) = \sigma^2_\varepsilon \frac{(1-a)^n}{1-a}$ progressione geometrica

Tramite queste informazioni possiamo inoltre calcolare la Densità di Transizione:

$$K(x, y) = \frac{f(x, y)}{f(x)} \quad (\text{poiché è una densità condizionata}), \text{ essendo } f(x, y) \text{ la densità di una } N^{(2)}$$

CASO 4: PROCESSI CON UN CONTINUO DI STATI A TEMPO CONTINUO

Indichiamo con $Q_t(x, B)$ la Funzione di Transizione (o Nucleo Stocastico) che in questo caso dipende anche dalla variabile continua t .

In quanto Nucleo Stocastico sussistono le seguenti condizioni:

- 1) Condizione di Coerenza: $Q_{t+s}(x, B) = \int_B Q_t(x, dy) Q_s(y, B)$
- 2) $Q_t(x, B) \geq 0$ (condizione di semigruppo)
- 3) $Q_t(x, \mathbb{R}) = 1$

Esempio:

• PROCESSO DI POISSON COMPOSTO

Sia $Y(t) = \sum_{j=1}^{N(t)} Y_j$ tale che:

- 1) $N(t)$ è un processo di arrivi di Poisson caratterizzato da un'intensità λ .
- 2) Y_j è un processo a tempo discreto di numeri aleatori IID detto Processo degli EFFETTI: rappresenta infatti gli effetti portati da ogni arrivo aleatorio.

In questo processo, dunque sono aleatori sia gli addendi che il numero di addendi.

Esempio: Operazioni di incassi/pagamenti di clienti ad uno sportello di banca: un numero aleatorio $N(t)$ di clienti, ciascuno dei quali incrementa o decremente il conto di una quantità Y_j .

Il processo di Poisson Composto inteso come somme successive di numeri aleatori

IID è un processo di cosiddetta PASSEGGIATA ALEATORIA.

■ Cenni su processi di PASSEGGIATA ALEATORIA o RANDOM WALK

Sono processi definiti da $S_n = \sum_{k=1}^n X_k$, $n \geq 1$, con X_k numeri aleatori IID, tali che $S_0 = 0$.

Equivalentemente il processo S_n può essere definito da: $S_0 = 0$

Si tratta di CATENE MARKOVIANE.

Sono X_k tali che $E(X_k) = \mu$ e $\text{Var}(X_k) = \sigma^2$ finiti (in particolare essi possono essere anche indicatori di eventi indipendenti ed equiprobabili), allora risulta:

- Funzione Valor Medio: $\psi(n) = E(S_n) = n\mu \rightarrow$ crescente con n
- Funzione di Covarianza: $\psi(\mu, n) = \text{Cov}(S_m, S_n) = \min(\mu\sigma^2, n\sigma^2)$
- Legge dei Grandi Numeri: $\Pr\left(\lim_{n \rightarrow \infty} \frac{S_n}{n} = \mu\right) = 1$
- Teorema Centrale del Limite: $\frac{S_n - n\mu}{\sqrt{n}\sigma^2} \xrightarrow{d} N(0, 1)$

• PROCESSO PSEUDO-Poisson

Sia $X(t) = \sum_{n=1}^{N(t)} X_n$ tale che:

1) $N(t)$ è un processo di arrivi di Poisson di intensità λ .

2) X_n è una Catena di Markov a tempo discreto

In tal caso il Nucleo Stocastico è:

$$Q_t(x, B) = \sum_{n=1}^{+\infty} \underbrace{\left[e^{-\lambda t} \frac{(\lambda t)^n}{n!} \right]}_{\substack{\text{Probabilità che tra } 0 \text{ et } \\ \text{si verifichino } n \text{ arrivi}}} K^{(n)}(x, B)$$

Funzione di Transizione della Catenza di Markov
tale che $K^{(n)}(x, B) = \int_R K(x, dy) K^{(n-1)}(y, B)$

Il generatore G di questo processo agisce su funzioni continue e limitate

$f(x) \in \mathcal{C}^{(\infty)}$, quindi $G: \mathcal{C}^{(\infty)} \rightarrow \mathcal{C}^{(\infty)}$:

$$\underbrace{G(f(x))}_{g(x)} = \lambda \int_R [f(y) - f(x)] K(x, dy)$$

In particolare risulta che $Q_t(x, B) = e^{tG}$

Dunque ogni processo di Markov, sotto certe condizioni, ammette un proprio generatore. Si può dimostrare inoltre che ogni processo di Markov può essere approssimato da processi Pseudo-Poissonianiani.

Cenni su FUNZIONE CARATTERISTICA:

Sia X un numero aleatorio, si definisce Funzione Caratteristica di X :

$$\Phi_X(u) = E[e^{iux}] = \int_R e^{iux} F_X(dx), \text{ dove } i \text{ è l'unità immaginaria.}$$

Proprietà:

- $\Phi_X(u)$ continua su \mathbb{R} ;
- $\Phi_X(0) = 1$, $|\Phi_X(u)| \leq 1 \forall u$
- $\Phi_{aX+b}(u) = e^{iub} \Phi_X(au) \quad \forall a, b \in \mathbb{R}$ costanti
- $\Phi'_X(0) = iE(X)$, $\Phi''_X(0) = -E(X^2)$

Funzione a valori complessi: $= \underbrace{\cos(ux)}_{\text{parte reale}} + i \underbrace{\sin(ux)}_{\text{parte immaginaria}}$

La Funzione Caratteristica generalizza la Funzione Generatrice dei Momenti (che non esiste sempre) in quanto esiste per tutte le distribuzioni.

Si può dimostrare che ogni Funzione caratteristica può essere approssimata da distribuzioni Poisson-Composte.

PROCESSI DI LÉVY

Rappresentano un sottoinsieme dei Processi di Markov a tempo continuo e appartengono ai processi cosiddetti **INFINITAMENTE DIVISIBILI**, strettamente connessi ad essi, ma senza una corrispondenza biunivoca.

Una prima caratterizzazione dei processi infinitamente divisibili fu data negli anni '30 da Bruno de Finetti come: dato $X_n F$, il numero aleatorio X è infinitamente divisibile se $\forall n \quad X \stackrel{d}{=} \sum_{i=1}^n X_i$ con X_i IID o, equivalentemente in termini di distribuzioni, se $\forall n \quad F_n | F_n^{*(n)} = F$

In pratica X è infinitamente divisibile se può essere scomposto come somma di numeri aleatori indipendenti e identicamente distribuiti.

De Finetti arrivò ad un risultato parziale implementato in seguito da Kolmogorov e infine da Paul Lévy, che riuscì a darne una caratterizzazione rigorosa in termini di Funzione Caratteristica come: X è infinitamente divisibile se $\Phi_X(n) = [\Phi_X(u)]^n$, ovvero se la Funzione Caratteristica di X può essere espressa in termini di Funzioni caratteristiche di altri n numeri aleatori IID. Risulta in particolare che l'insieme delle distribuzioni infinitamente divisibili rappresenta la Chiusura dell'insieme delle Poisson Composte (ovvero può essere rappresentato come loro limite).

Dare la rappresentazione generale per i processi infinitamente divisibili equivale a dare la descrizione matematica per i processi di Lévy: si può provare che ogni processo di Lévy ha legge temporale costituita da distribuzioni infinitamente divisibili e che fissata una distribuzione infinitamente divisibile esiste un processo di Lévy la cui legge temporale contiene quella distribuzione.

$X(t)$ è un PROCESSO DI LÉVY se:

- 1) $X(0) = 0$ q.c.
- 2) $X(t)$ è a INCREMENTI INDIPENDENTI: incrementi su intervalli disgiunti sono indipendenti.
- 3) $X(t)$ è a INCREMENTI OMogenei (o STAZIONARI) incrementi su intervalli della stessa lunghezza sono ugualmente distribuiti;
- 4) Vale la CONTINUITÀ STOCHASTICA: $\lim_{\substack{s \rightarrow t \\ (\text{continua in probabilità})}} \Pr(|X(t) - X(s)| > \varepsilon) = 0$

Dimostriamo che i processi di Lévy sono un sottinsieme dei processi di Markov:

Consideriamo $\Pr(X(t)=x | X(t_1), \dots, X(t_n))$ con $t_1 < t_2 < \dots < t_n < t$

$$= \Pr \left[\left(\underbrace{X(t) - X(t_n)}_{\text{...}} \right) + X(t_n) \leq x \mid X(t_1), \dots, X(t_n) \right] =$$

↳ Questo incremento è stocasticamente indipendente dagli: $X(t_0) - X(0)$, $X(t_1) - X(t_0)$, ..., $X(t_n) - X(t_{n-1})$, dunque possiamo trascurare tutti i numeri allestori precedenti a $X(t_0)$

$$= \Pr \left[(X(t) - X(t_n)) + X(t_n) \leq x \mid X(t_n) \right] = \Pr \left[X(t) \leq x \mid X(t_n) \right] \Rightarrow \text{D. tutti i numeri aleatori passati influiscono solo il più recente } X(t)$$

↳ Dipenderza Markoviana

Proposizione: Le distribuzioni che costituiscono la legge di probabilità dei processi di Lévy sono infinitamente divisibili.

Dimostrazione

(\Rightarrow) Sia $X(E)$ un processo di Lévy con $t > 0$ fissato.

Dobbiamo dimostrare che $X(t) = \sum_{j=1}^n Y_j$ con Y_j IID.

Suddividiamo l'intervallo $[0, t]$ in n sottointervalli della stessa lunghezza

$$X(t_n \cdot j) \xrightarrow{\text{IID}} X(t_n \cdot (j+1)) \quad (j=1 \dots n) \quad \text{sono IID} \quad \Rightarrow \text{Per il processo di Lévy gli incrementi su questi intervalli}$$

$$\underbrace{X(t_n \cdot (j+1)) - X(t_n \cdot j)}_{Y_j} \quad \text{e} \quad X(t) = \sum_{j=1}^n Y_j$$

\Leftarrow) Fissata una distribuzione infinitamente divisibile arbitraria, esiste un processo di Levy (non necessariamente unico) la cui legge temporale ha distribuzione dello stesso tipo di quella fissata in partenza. ✓ C.v.s.

I processi di Lévy sono tra i più semplici processi di Markov e sono

Semi-Martingale.

Proposizione

(a) 1) $\Psi(t) = E(X(t)) = a \cdot t \rightarrow$ La Funzione Valor Medio è lineare in t

Dimostrazione

Per l'omogeneità risulta che $X(t) - X(s) \stackrel{d}{=} X(t-s) - \underbrace{X(0)}_{=0 \text{ q.c.}}$

$$\Rightarrow E(X(t) - X(s)) = E(X(t-s)) \Leftrightarrow E(X(t)) - E(X(s)) = E(X(t-s))$$

Sia $E[X(\cdot)] = u(\cdot)$ $\Rightarrow u(t) = u(s) + u(t-s)$: Equazione funzionale di Cauchy

L'equazione ha infinite soluzioni a meno che non si aggiunga una condizione.

Imponiamo la condizione che $u(t)$ sia continua.

L'unica soluzione continua di tale equazione è $E[X(t)] = u(t) = kt$ ~~c.v.d.~~

Esempi: • Processo di Poisson \rightarrow intensità λt
• Processo di Wiener $\rightarrow K=0$

2) $\text{Var}(X(t)) = a^* \cdot t$

Dimostrazione

Risulta che $\text{Var}(X(t) - X(s)) = \text{Var}(X(t-s)) \Leftrightarrow \text{Var}(X(t)) + \text{Var}(X(s)) - 2\text{Cov}(X(t), X(s))$

$= \text{Var}(X(t-s))$ Assumiamo, senza perdita di generalità, che i numeri aleatori siano equi.

Poiché siamo in un processo di Lévy, per $s < t$, risulta che: $\text{Cov}(X(t), X(s)) = \text{Var}(X(s))$

$$\Rightarrow \text{Var}(X(t)) - \text{Var}(X(s)) = \text{Var}(X(t-s))$$

Sia $V(\cdot) = \text{Var}(X(\cdot)) \Rightarrow V(t) - V(s) = V(t-s) \Rightarrow$ Imponiamo la continuità di $V(t)$:

otteniamo l'unica soluzione $V(t) = a^* \cdot t$ ~~c.v.d.~~

Osservazione: Nel processo di Poisson, poiché la media è uguale alla varianza, $a = a^*$.

(b) $\Psi(s, t) = \text{Cov}(X(s), X(t)) = \text{Var}(X(s \wedge t))$
 $\hookrightarrow = \min\{s, t\}$

Dimostrazione

Assumiamo, senza perdita di generalità, che $\Psi(t)$ sia identicamente nulla (ossia $= 0 \forall t$).

$$\Psi(s, t) = E[X(s)X(t)] = E[\underbrace{X(s)(X(t) - X(s) + X(s))}_{\text{indipendente da } X(s)}] = E[X(s)(X(t) - X(s)) + X^2(s)] =$$

$$= E[X(s)(X(t) - X(s))] + E[X^2(s)] = \underbrace{E[X(s)]}_{0} E[X(t) - X(s)] + \underbrace{E[X^2(s)]}_{\text{Var}(X(s))} = \text{Var}(X(s))$$

$s = \min\{t, s\}$

(C) $\Phi_{X(t)}(u) = E[e^{iuX(t)}] = e^{\underline{t \cdot \eta(u)}}$ (o simbolo)

↳ ESPONENTE DI LÉVY : è una funzione a valori complessi che caratterizza il particolare processo di Lévy.

Dimostrazione

$$\begin{aligned} \Phi_{X(t+s)}(u) &= E[e^{iuX(t+s)}] = E[e^{iu[X(t+s)-X(s)+X(s)]}] = E[e^{iu[X(t+s)-X(s)]} e^{iuX(s)}] = \\ &= E[e^{iu[X(t+s)-X(s)]}] \cdot E[e^{iuX(s)}] = \Phi_{X(t)}(u) \Phi_{X(s)}(u) \\ &\stackrel{| X(t+s)-X(s) \perp\!\!\!\perp X(s) \quad | \stackrel{d}{=} X(t)}{\Rightarrow} \Phi_{X(t+s)}(u) = \Phi_{X(t)}(u) \Phi_{X(s)}(u) \Rightarrow \text{Imponiamo la condizione di continuità,} \\ &\text{ottenendo l'unica soluzione di } \Phi_{X(t)}(u) = e^{t\eta(u)}. \end{aligned}$$

(d) $\Phi_{X(t)}(u) = E[e^{iuX(t)}] = \exp \left\{ t \left[iub - \frac{\sigma^2 u^2}{2} + \int_R (e^{iuy} - 1 - iuy I_{[-1,1]}(y)) \nu(dy) \right] \right\}$

FORMULA DI LÉVY-KHINTCHINE

Dove:

- $\eta(u) = iub - \frac{\sigma^2 u^2}{2} + \int_R (e^{iuy} - 1 - iuy I_{[-1,1]}(y)) \nu(dy) \rightarrow$ ESPONENTE DI LÉVY
- $\nu(\cdot) \rightarrow$ MISURA DI LÉVY

- Essa è tale che:
- Se $\nu(R) = \infty$ abbiamo processi di Lévy più complessi, in quanto in qualsiasi intervallo finito di tempo possono verificarsi infiniti arrivi;
 - Se $\nu(R) < \infty$ allora in qualunque intervallo finito di tempo può verificarsi solo un numero finito di eventi, caratterizzando processi più semplici, come quelli di Poisson.

• $I_{[-1,1]}(y)$ è un indicatore che tiene conto di situazioni in cui in un intervallo finito di tempo arrivano infiniti eventi.

• La terna (b, σ^2, ν) caratterizza completamente il processo di Lévy.

Po esempio nel processo di Wiener: $b=0$, $\sigma^2 > 0$, ν = misura di Wiener.

Nell'espressione intervengono:

- Una componente gaussiana: $\frac{\sigma^2 u^2}{2}$
- Una componente Poisson-Composta: integrale di Lebesgue-Stieltjes deterministica

Esempi:

- Moto Browniano: $\Phi_{W(t)}(u) = e^{-\frac{\sigma^2 t}{2} u^2}$ → non interviene l'integrale, né l'addendo immaginario

- Poisson: $\Phi_{N(t)}(u) = \exp \{ \lambda t (e^{iu} - 1) \} \rightarrow e^{iu}$ è la Funzione Caratteristica di un numero aleatorio che vale 1 con probabilità 1, è dunque (Se i salti avessero ampiezza a , avremmo e^{iau}) la Funzione Caratteristica dei Salti (unitari) di un processo di Poisson.

- Poisson-Composto: $\Phi_{X(t)}(u) = \exp \{ \lambda t (\Phi_Y(u) - 1) \} \rightarrow \Phi_Y(u)$ è la Funzione Caratteristica delle discontinuità (incrementi) aleatorie di un processo Poisson-Composto, ossia gli effetti.

RAPPRESENTAZIONE DI LÉVY-ITO: Il processo di Lévy può essere visto come soluzione di un'equazione differenziale stocastica:

$$X(t) = \mu t + b \underbrace{\int_0^t}_{\substack{\text{costante} \\ b}} + \underbrace{\int_0^t}_\text{Processo di Wiener} x \bar{N}_w(t, dx) + \underbrace{\int_{|x| \geq 1} x N_w(t, dx)}_{\substack{\text{componente di} \\ \text{discontinuità ridotte del} \\ \text{processo} \\ (\text{intorno di } 0, \\ \text{il valore } b \text{ è} \\ \text{arbitrario})}} + \underbrace{\int_{|x| < 1} x \bar{N}_w(t, dx)}_{\substack{\text{componente di} \\ \text{discontinuità} \\ \text{ampie del processo} \\ \text{Integrati di Poisson} \\ \text{con } N_w \text{ misura aleatoria di Poisson} \\ \text{(stocastica, non deterministica)}}}$$

È opportuno distinguere le discontinuità piccole da quelle grandi in quanto gli incrementi piccoli causano problemi per l'intensità (poiché talvolta sono omissi), mentre gli incrementi grandi costituiscono incrementi di tipo Poisson-Composto.

Il primo integrale, quindi, considera le anomalie del processo di Lévy, tenendo conto delle infinite discontinuità in intervalli limitati.

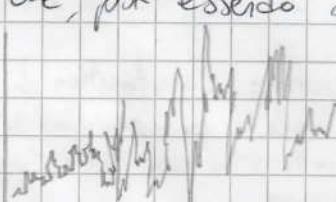
Esempi di Processi di Lévy:

- Processo di Poisson
- Processo di WIENER o di MOTO BROWNAVO :

Definito nel 1923 da Norbert Wiener, quando ancora non esisteva una descrizione rigorosa del calcolo delle probabilità. È anche definito il Moto Browniano in quanto associato dal botanico Brown alle traiettorie casuali degli atomi di una particella immersa in un liquido prodotte dai loro urti.

Tale osservazione provò l'esistenza degli atomi (assieme ad Einstein nel 1905).

Le traiettorie di un processo di Wiener hanno un'irregolarità così forte che, pur essendo funzioni continue, non sono derivabili in alcun punto.



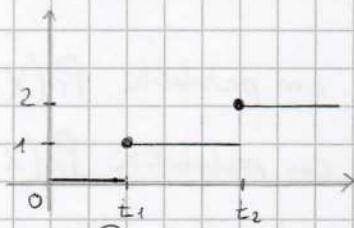
Il moderno calcolo delle probabilità è costruito sui due fondamentali processi di Poisson (per il caso discreto) e di Wiener (per il caso continuo).

Il processo di Wiener ha la particolarità di essere contemporaneamente un processo di Lévy (dunque markoviano), Gaussiano e Martingala.

• Processo di Poisson - Composto

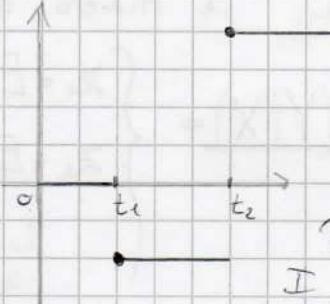
Essendo $X(t) = \sum_{j=1}^{N(t)} Y_j$, il processo contiene due parti d'incertezza: $N(t)$ e Y_j , tra esse indipendenti.

Confrontiamo la traiettoria generica di un Processo di Poisson con quella di un Poisson-Composto:



Poisson

Le discontinuità sono tutte di ampiezza unitaria, in quanto contano il numero di arrivi



Poisson- Composto

I salti possono essere di qualsiasi ampiezza finita e qualsiasi segno con traiettorie costanti a tratti

Consideriamo la funzione di ripartizione di $X(t)$:

$$\Pr(X(t) \leq x) = \Pr(\{X(t) \leq x\} \cap \{N(t) = n\}) = \sum_{n=0}^{+\infty} \Pr(X(t) \leq x | N(t) = n) \Pr(N(t) = n)$$

$$= \sum_{n=0}^{+\infty} F(x) e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

$\underbrace{\text{Convoltione } n\text{-esima degli } Y_j}_{\text{Convoluzione } n\text{-esima degli } Y_j}$

$\sum_{j=0}^{N(t)=n} Y_j \xrightarrow{\text{Numero d'eventi fissato}} e^{-\lambda t} \frac{(\lambda t)^n}{n!}$

Poiché non è agevole lavorare con convoluzioni, è preferibile utilizzare la Funzione Caratteristica d- Y_i nella sua potenza n -esima: $[\phi_{Y_i}(u)]^n$.

Si può dimostrare che:

$$\cdot E(X(t)) = \lambda t E(Y_1) \quad \cdot \text{Var}(X(t)) = \lambda t E(Y_1^2)$$

Osservazioni: - $X(t)$ ha un insieme continuo di stati solo se gli Y_j sono reali;

$$- F_{X(t)}(x) = \sum_{n=0}^{+\infty} P_{N(t)} F_Y^{*n}(x)$$

- $F_{X(t)}(x)$ ha una forma simile alla transizione di un processo Pseudo-Poisson $Q_t(x, \beta) = \sum_{n=0}^{+\infty} P_{N(t)} K^{(n)}(x, \beta)$, ma le due espressioni hanno significati diversi.

- Poiché i processi di Lévy sono omogenei nel tempo e nello spazio risulta che $Q_t(x, \beta) = Q_t(\beta - x)$, ossia la totalità degli stati di tipo $Y - X$ con $Y \in B$.

In generale per un processo d-Markov le funzioni di transizione possono essere scritte come $Q_{s-t}(x, \beta) \sim \Pr(X(t) \in B | X(s) = x)$ (per processi non omogenei).

Se il processo d-Markov è omogeneo nel tempo possiamo scrivere $Q_{t-s}(x, \beta)$.

Se il processo, oltre che essere omogeneo nel tempo, è omogeneo anche nello spazio degli stati, possiamo scriverlo come $Q_{t-s}(B - x)$: conta anche la distanza tra gli elementi dell'insieme B e lo stato d-partenza x , oltre che agli intervalli di tempo.

Ricchiamo sulla Funzione di REGRESSIONE:

[Nasce negli anni '30 ad opera di Kolmogorov]

Definizioni:

(1) Dati X, Y numeri aleatori tali che $E(Y) < +\infty$ e X discreto si dice

FUNZIONE DI REGRESSIONE il numero aleatorio:

$$Z(w) = E(Y|X) = \begin{cases} Z_1 = E[Y|X=x_1] & \text{con probabilità } P_r(X=x_1) \\ Z_2 = E[Y|X=x_2] & \text{con probabilità } P_r(X=x_2) \\ \vdots & \end{cases}$$

L'osservazione medì condizionata sono le possibili determinazioni del numero aleatorio Funzione di Regressione

Osservazione: Anche se i valori di probabilità sono gli stessi di X non si può dire che Z abbia la stessa distribuzione di probabilità di X , in quanto hanno determinazioni diverse.

(2) Definizione Assiomatica: Siano X, Y numeri aleatori tali che $E[Y] < +\infty$.

Si dice FUNZIONE DI REGRESSIONE il numero aleatorio $Z = E(Y|X)$ tale che:

$$E[(Y-Z)f(x)] = E[(Y-E(Y|X))f(x)] = 0$$

$f(x)$ per cui tale scrittura
abbia senso
(continua e l.a. tata)

Osservazioni:

- La condizione equivale a: $E[Yf(x)] = E[E(Y|X)f(x)]$
- Poiché risulta che $E[Y-E(Y|X)] = 0$, il valor medio $E[(Y-E(Y|X))f(x)]$ rappresenta la $\text{Cov}[Y-E(Y|X), f(x)]$, che è dunque nulla.

La Funzione di Regressione è dunque quel numero aleatorio in corrispondenza al quale le differenze $Y-E(Y|X)$ sono non correlate con $f(x)$, funzione continua e limitata. Essa può dunque essere vista come una proiezione ortogonale.

(3) Definizione Operativa: La Funzione di Regressione è il numero aleatorio $\hat{Y}(X) = E(Y|X)$ tale che:

$$E[(Y-\hat{Y}(X))^2] \geq E[\underbrace{Y - E(Y|X)}_{\hat{Y}(X) \rightarrow \text{dipendente da } X}]^2$$

Osservazione: La Funzione di Regressione può essere interpretata come approssimazione dei minimi quadrati per Y non osservabile utilizzando X osservabile. Essa è dunque il miglior approssimatore per Y nel senso dei minimi quadrati. \rightarrow Minimizza $E[(Y-\hat{Y}(X))^2]$

Dimostrazione: Minimizziamo la media degli scarti quadratici:

$$E[(Y-\hat{Y}(X))^2] = E[(Y-E(Y|X)+E(Y|X)-\hat{Y}(X))^2] = E[(Y-E(Y|X))^2] + 2E[(Y-E(Y|X))(E(Y|X)-\hat{Y}(X))] + E[(E(Y|X)-\hat{Y}(X))^2] = \infty$$

Se \hat{Y} è la Funzione di Regressione lo scarto $E[(E(Y|X)-\hat{Y}(X))^2]$ si annulla.
otteniamo $E[(Y-E(Y|X))\cdot Y(X)] = 0$ per la definizione (2) Cvd (2) c

►9.7. Properties of conditional expectation: a list

These properties are proved in Section 9.8. All X 's satisfy $E(|X|) < \infty$ in this list of properties. Of course, \mathcal{G} and \mathcal{H} denote sub- σ -algebras of \mathcal{F} . (The use of 'c' to denote 'conditional' in (cMON), etc., is obvious.)

- (a) If Y is any version of $E(X|\mathcal{G})$ then $E(Y) = E(X)$. (Very useful, this.)
- (b) If X is \mathcal{G} measurable, then $E(X|\mathcal{G}) = X$, a.s.
- (c) (Linearity) $E(a_1 X_1 + a_2 X_2|\mathcal{G}) = a_1 E(X_1|\mathcal{G}) + a_2 E(X_2|\mathcal{G})$, a.s.

Clarification: if Y_1 is a version of $E(X_1|\mathcal{G})$ and Y_2 is a version of $E(X_2|\mathcal{G})$, then $a_1 Y_1 + a_2 Y_2$ is a version of $E(a_1 X_1 + a_2 X_2|\mathcal{G})$.

- (d) (Positivity) If $X \geq 0$, then $E(X|\mathcal{G}) \geq 0$, a.s.
- (e) (cMON) If $0 \leq X_n \uparrow X$, then $E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G})$, a.s.
- (f) (cFATOU) If $X_n \geq 0$, then $E[\liminf X_n|\mathcal{G}] \leq \liminf E[X_n|\mathcal{G}]$, a.s.
- (g) (cDOM) If $|X_n(\omega)| \leq V(\omega)$, $\forall n$, $EV < \infty$, and $X_n \rightarrow X$, a.s., then

$$E(X_n|\mathcal{G}) \rightarrow E(X|\mathcal{G}), \quad \text{a.s.}$$

- (h) (cJENSEN) If $c : \mathbf{R} \rightarrow \mathbf{R}$ is convex, and $E|c(X)| < \infty$, then

$$E[c(X)|\mathcal{G}] \geq c(E[X|\mathcal{G}]), \quad \text{a.s.}$$

Important corollary: $\|E(X|\mathcal{G})\|_p \leq \|X\|_p$ for $p \geq 1$.

- (i) (Tower Property) If \mathcal{H} is a sub- σ -algebra of \mathcal{G} , then

$$E[E(X|\mathcal{G})|\mathcal{H}] = E[X|\mathcal{H}], \quad \text{a.s.}$$

Note. We shorthand LHS to $E[X|\mathcal{G}|\mathcal{H}]$ for tidiness.

- (j) ('Taking out what is known') If Z is \mathcal{G} -measurable and bounded, then

$$(*) \quad E[ZX|\mathcal{G}] = ZE[X|\mathcal{G}], \quad \text{a.s.}$$

If $p > 1$, $p^{-1} + q^{-1} = 1$, $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbf{P})$ and $Z \in \mathcal{L}^q(\Omega, \mathcal{G}, \mathbf{P})$, then (*) again holds. If $X \in (\mathbf{m}\mathcal{F})^+$, $Z \in (\mathbf{m}\mathcal{G})^+$, $E(X) < \infty$ and $E(ZX) < \infty$, then (*) holds.

- (k) (Rôle of independence) If \mathcal{H} is independent of $\sigma(\sigma(X), \mathcal{G})$, then

$$E[X|\sigma(\mathcal{G}, \mathcal{H})] = E(X|\mathcal{G}), \quad \text{a.s.}$$

In particular, if X is independent of \mathcal{H} , then $E(X|\mathcal{H}) = E(X)$, a.s.

$$(\ell) \quad \left. \begin{array}{l} \{ \subseteq \mathcal{G} \\ X \in \mathcal{G} \\ Y \perp\!\!\!\perp \mathcal{G} \end{array} \right\} \quad \begin{array}{l} E[f(x, Y)|\mathcal{G}] = G_f(x), \text{ where } G_f(x) = E[f(x, Y)] \\ \downarrow \\ E(X+Y|\mathcal{G}) = X, E(Y) \end{array} , \quad \forall x \in \mathbf{R}.$$

Nell'espressione $E(Y|X)$ condizionarsi al numero aleatorio X equivale a condizionarsi all'intera σ -algebra generata da X , dunque possiamo scrivere:

$$E(Y|X) = E(Y|\mathcal{G}(X))$$

Introduciamo infatti nel nostro Spazio di Probabilità (Ω, \mathcal{F}, P) una famiglia di σ -algebre $\{\mathcal{F}_n\}$, detta Filtrazione, tale che:

$$\{\mathcal{F}_n \subseteq \mathcal{F}\}$$

$$\mathcal{F}_n \subseteq \mathcal{F}_{n+1} \quad (\text{è un insieme crescente di sotto-}\sigma\text{-algebre di } \mathcal{F})$$

$$\text{In particolare } \mathcal{F}_\infty = \sigma\left(\bigcup_{i=1}^{\infty} \mathcal{F}_i\right) \subseteq \mathcal{F}.$$

Otteniamo così lo spazio $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ detto SPAZIO DI PROBABILITÀ FILTRATO o BASE STOCASTICA.

La presenza di una Filtrazione fa sì che lo stato di informazione individuale su un l'insieme $\omega \in \Omega$ sia crescente nel tempo.

Una Filtrazione può anche essere generata da un Processo Stocastico, che possiamo supporre osservabile: $\mathcal{F}_1 = \sigma(X_1), \mathcal{F}_2 = \sigma(X_1, X_2), \mathcal{F}_3 = \sigma(X_1, X_2, X_3), \dots$

In questo caso, se oltre a conoscere X_1 conosciuto anche X_2 , la nostra informazione su ω cresce. È dunque l'osservabilità del processo a generare la Filtrazione.

Si parla di Filtrazione Naturale per il processo Stocastico come σ -algebra generata proprio dal processo X .

Si dice che $\{Y_n\}$ è ADAPTEDA A $\{\mathcal{F}_n\}$ se $Y_n \in \mathcal{F}_n \quad \forall n$, ossia se Y_n è misurabile su \mathcal{F}_n .

Proprietà della Funzione di Regressione:

$$1) \text{ Linearità: } E[\alpha Y_1 + \beta Y_2 | X] = \alpha E(Y_1 | X) + \beta E(Y_2 | X)$$

$$2) \text{ Positività: } Y \geq 0 \Rightarrow E(Y | X) \geq 0$$

$$3) E[E(Y | X)] = E(Y)$$

$$4) E[Y f(x) | X] = f(x) E(Y | X) \rightarrow \text{Poiché } X \text{ è condizionante, ogni ipotesi su } X \text{ si riflette su } f(x)$$

$$5) Y \perp\!\!\!\perp X \Rightarrow E(Y | X) = E(Y)$$

$$6) \text{ Tower Property: } E[E(Y | X) | X, Z] = E[E(Y | X, Z) | X] = E(Y | X)$$

↓
Funzione di Regressione della Funzione di Regressione
(doppio condizionamento)

→ Prende il
condizionamento meno
restrietivo

$$7) E[(Y - E(Y|X))^2] \leq E[(Y - f(X))^2]$$

$$8) Y \perp\!\!\!\perp X \Rightarrow E[f(X, Y)|X] = E_y[f(X, Y)] \rightarrow E \text{ una funzione di } X \times Y(x)$$

Calcolato con la distribuzione di Y

Esempi:

- Nell'Analisi di Regressione per studiare l'influenza di X su Y vorremo studiare $F_{Y|X}(y|x)$, ma ci accontentiamo di osservare la relazione fra $E(Y)$ e X utilizzando uno strumento meno potente, ossia la Funzione di Regressione $E(Y|X)$.

Per costrurla si assume l'**IPOTESI DI LAVORO** che $E(Y|X)$ sia data in un certo modo, ad esempio $E(Y|X) = \alpha + \beta X$ (nell'analisi di regressione lineare). In tal caso l'obiettivo diventa la stima dei coefficienti α e β a partire da una sequenza di coppie di valori per X e Y , (x_i, y_i) $i=1 \dots n$.

- Siano N il numero uscito nel lancio di un dado equo e D l'evento "esce un numero dispari". Dunque $|D|=1$ se esce 1, 3, 5.

$ D $	1	2	3	4	5	6	\rightarrow Distribuzione di Probabilità condizionata
1	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	Sappiamo che $E(N)=3.5$
0	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	

Calcoliamo $E(N| |D|=1) = \begin{cases} E(N| |D|=1) = \frac{1+3+5}{3} = 3 \\ E(N| |D|=0) = \frac{2+4+6}{3} = 4 \end{cases} \Rightarrow E[E(N| |D|)] = \frac{3+4}{2} = 3.5 = E(N)$

$E(|D| | N)$ 6 valori possibili

PROCESSI MARTINGALA

Nascono in Francia negli anni '30 ad opera dei probabilisti J. Ville e P. Lévy. La teoria fu sviluppata negli anni '40 dal probabilista americano Doob.

Consideriamo la Base Stocastica $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ e il processo

$X_n(\omega) : \Omega \rightarrow \mathbb{R} \in \mathcal{F}_n$ (Adapted) $\forall n$ tale che $E(|X_n|) < +\infty$.

Il processo X_n si dice MARTINGALA se:

$$\underline{E[X_n | \mathcal{F}_{n-1}] = E[X_n | X_1, \dots, X_{n-1}] = X_{n-1}}$$

Tale definizione è nel caso particolare di Martingale a tempo discreto.

Essa può essere data anche in presenza di un processo stocastico Y_n non necessariamente collegato con X_n , ossia: $E[X_n | Y_1, \dots, Y_{n-1}] = X_{n-1}$.

Proprietà:

- Funzione Valor-Medio: $\Psi(X_n) = E(X_n) = K$ costante

Dimostrazione: $E[E(X_n | X_{n-1})] =$ Per la Martingalità: $E(X_{n-1})$ \rightarrow uguali $\forall n \Rightarrow$ costanti
Per la proprietà (3) della Funzione di Regressione: $E(X_n)$ \checkmark v.d.

- Funzione di Covarianza: $\Psi(m, n) = \text{Cov}(X_m, X_n) = \text{Var}(X_{m \wedge n})$, ($m \wedge n = \min\{m, n\}$)

Dimostrazione:

Assumiamo, senza perdita di generalità, che gli X_n siano identicamente equi, ossia: $E(X_n) \equiv 0$

$$\text{Cov}(X_m, X_n) = E(X_m, X_n) = E[(X_n - X_m + X_m)X_m] = E[\underbrace{(X_n - X_m)X_m}_{=0} + \underbrace{E(X_m^2)}_{\text{Var}(X_m)}]$$

$$E[(X_n - X_m)X_m] = E[E[(X_n - X_m)X_m | X_m]] = E[X_m E(X_n - X_m | X_m)] = 0$$

Definizione: I numeri aleatori di una successione Y_n sono detti ASSOLUTAMENTE EQUI se: $E(Y_1) = 0$, $E(Y_2 | Y_1) = 0$, $E(Y_3 | Y_1, Y_2) = 0$, $E(Y_4 | Y_1, Y_2, Y_3) = 0$, ...

Proposizioni:

- Assoluta equità \Rightarrow equità $E(X_2) = E[\underline{E(X_2 | X_1)}] = \dots$

- Assoluta equità \Rightarrow non correlazione

Dimostrazione

$$\text{Cov}(Y_i, Y_j) = \overbrace{E(Y_i Y_j)}^{i < j} = E[\overbrace{E(Y_i Y_j | Y_{j-1})}^{O \text{ per l'assoluta equità}}] = E[Y_i \cdot \overbrace{E(Y_j | Y_{j-1})}^{O \text{ per l'assoluta equità}}] = 0$$

\checkmark c.v.d.

\downarrow

$E(Y_i) = E(Y_j) = 0$ per equità

\downarrow

Poiché $Y_i \in Y_{j-1}$ (proprietà 4 funzione di regressione)

• Stocistica indipendenza ed equità \Rightarrow assoluta equità

Dimostrazione

Siano i numeri aleatori Z_j equi ed indipendenti, allora:

$$E(Z_j) = 0 \quad \forall j \text{ per equità}$$

$$\text{Per l'indipendenza: } E[Z_2 | Z_1] = E(Z_2) = 0$$

$$E[Z_3 | Z_2, Z_1] = E(Z_3) = 0$$

\Rightarrow assolutamente equi ~~c.v.d.~~

... -

Osserviamo che mentre l'assoluta equità implica equità e non correlazione, queste due non bastano per ottenere l'assoluta equità, bensì serve la stocastica indipendenza.

TEOREMA DI RAPPRESENTAZIONE PER MARTINGALE

X_n è una Martingala $\Leftrightarrow X_n = \sum_{j=1}^n Y_j + C$ con Y_j assolutamente equi e C costante

Dimostrazione

(\Rightarrow) Sia X_n una Martingala, definiamo:

$$Y_1 = X_1 - \underbrace{E(X_1)}_C$$

Creiamo una corrispondenza \rightarrow In forma vettoriale:

$$Y_2 = X_2 - X_1$$

\Rightarrow biunivoca tra X e Y

$$Y_3 = X_3 - X_2$$

con una trasformazione lineare

affine

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & \ddots & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} - \begin{bmatrix} C \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

:

Risulta allora che: $E(Y_1) = E(X_1 - C) = 0$

$$E(Y_2 | Y_1) = E[X_2 - X_1 | X_1] \stackrel{\substack{\text{Linearità} \\ X_1}}{=} E[X_2 | X_1] - E[X_1 | X_1] = 0$$

Qualunque sia l'ipotesi su Y_1 vi corrisponde
una e una sola ipotesi per X_1 , quindi possiamo
sostituire il condizionamento da Y_1 a X_1 \rightarrow per Martingalità

... - $\Rightarrow Y_j$ assolutamente equi

\checkmark

(\Leftarrow) Siano $X_n = \sum_{j=1}^n Y_j + C$ con Y_j assolutamente equi

$$E(X_n | X_1, \dots, X_{n-1}) = E\left[\sum_{j=1}^n Y_j + C | Y_1, \dots, Y_{n-1}\right] = E\left[\underbrace{\sum_{j=1}^{n-1} Y_j}_{=0} + C | Y_{n-1}\right] =$$

$$= \sum_{j=1}^{n-1} Y_j + C = X_{n-1} \Rightarrow \text{Martingala}$$

per l'assoluta equità

~~c.v.d.~~

Proposizione: $E[X_n | X_1, \dots, X_{n-1}] = X_{n-1} \Leftrightarrow E[X_{n+h} | X_1, \dots, X_{n-1}] = X_{n-1} \quad \forall h$

Dimostrazione

(\Leftarrow) Poiché l'uguaglianza vale $\forall h$ basta porre $h=0$ per ottenere la tesi.

$$\Rightarrow E\left[E(X_{n+h} | X_{n-1}) | X_{n-1}\right] = E(X_{n+h-1} | X_{n-1}) = E\left[E(X_{n+h-1} | X_{n-2}) | X_{n-1}\right] = \dots = X_{n-1}$$

↓
 Martingalità
 Tower Property della funzione
 ↓ Regressione

Reiterando il procedimento
 di sostituzione C.V.D.

Osserviamo che ogni sottosuccessione \downarrow una Martingala è ancora una Martingala,

ossia: $E[X_{n+h} | X_t, \dots, X_{t-1}] = X_{t-1}$.

Esempi:

(1) Siano X_n indipendenti ed equi (non necessariamente hanno la medesima distribuzione),

allora $S_n = \sum_{j=1}^n X_j$ è una Martingala, ossia: $E[S_n | S_1, \dots, S_{n-1}] = E[S_n | X_1, \dots, X_{n-1}] = S_{n-1}$

Dimostrazione

$$S_n = X_n + S_{n-1} \Rightarrow E[X_n + S_{n-1} | X_1, \dots, X_{n-1}] = E[X_n | X_1, \dots, X_{n-1}] + E[S_{n-1} | X_1, \dots, X_{n-1}] = S_{n-1}$$

Linearietà
 ↓
 = $E(X_n)$ per $= S_{n-1}$: è discutibile rispetto alla
 indipendenza, $\hookrightarrow = 0$ per egualità
 e algebrà generata da X_1, \dots, X_{n-1}

(2) Somme successive di numeri aleatori IID ~~non~~ equi sono una Martingala

(3) Martingala di Lévy: Sia $X_n = E(Z | Y_1, \dots, Y_n)$ una funzione di regressione tale che

$E(Z|Y_i) < \infty$ (senza nessun particolare legame tra Z e Y_i), allora X_n è una Martingala.

rispetto agli Y_i , ossia: $E(X_n | Y_1, \dots, Y_{n-1}) = X_{n-1}$.

Dimostrazione: $E(X_n | Y_1, \dots, Y_{n-1}) = E[E(Z | Y_n) | Y_{n-1}] = E(Z | Y_{n-1}) = X_{n-1}$ // C.V.D.

(4) Urna di Polya: b bianche e r rosse: ad ogni estrazione la pallina pescata viene reimessa aggiungendo C (≥ 1) palline dello stesso colore. Assumiamo $C=1$.

La proporzione X_n di palline bianche nell'urna dopo le prime n estrazioni è una Martingala.

Dimostrazione: Sia $h = \sum_{j=1}^n |E_j|$ la frequenza di bianche pescate

Dopo n estrazioni nell'urna ci saranno $b+r+n$ palline, quindi: $X_n = \frac{b+h}{b+r+n}$

Dunque X_{n+1} può assumere due valori: $\begin{cases} \frac{b+h+1}{b+r+n+1} & \text{se } E_{n+1}, \text{ con probabilità } \frac{b+h}{b+r+n} \\ \frac{b+h}{b+r+n+1} & \text{se } \bar{E}_{n+1}, \text{ con probabilità } 1 - \frac{b+h}{b+r+n} \end{cases}$

Ottieniamo che $E(X_{n+1}) = X_n$ $\begin{cases} \frac{b+h}{b+r+n+1} & \text{se } E_{n+1}, \text{ con probabilità } 1 - \frac{b+h}{b+r+n} \\ \frac{b+h+1}{b+r+n+1} & \text{se } \bar{E}_{n+1}, \text{ con probabilità } \frac{b+h}{b+r+n} \end{cases}$

Condizionatamente all'ipotesi h , ossia: $E[X_{n+1} | X_1, \dots, X_n] = X_n$ // C.V.D.

Osservazione: Il numero di palline bianche dopo n estrazioni è invece una Catena di Markov. 36

MARTINGALE A TEMPO CONTINUO

Il processo a tempo continuo $X(t)$ è una Martingala se: $E[X(t)|\mathcal{F}_s] = X(s) \quad \forall s < t$

Esempi:

(1) Processo di Wiener o di Moto Browniano: $W(t)$ è una Martingala

Dimostrazione: Sia $\mathcal{F}_s^W = \sigma(W(s), s \leq t)$ la σ -algebra generata da $W(s)$ per $s \leq t$ (filtrazione naturale)

$$E[W(t)|\mathcal{F}_s^W] = E[W(t) - W(s) + W(s)|\mathcal{F}_s^W] = E[W(t) - W(s)|\mathcal{F}_s^W] + E[W(s)|\mathcal{F}_s^W] = \underbrace{E[W(t) - W(s)]}_{=0 \text{ funzione } \varphi \text{ identicamente nulla}} + W(s) = W(s)$$

$\xrightarrow{\text{C.V.D}}$

poiché $W(t)$ è ad incrementi indipendenti su intervalli disgiunti,
 $W(t) - W(s) \perp\!\!\!\perp W(s)$ e dunque $W(t) - W(s) \perp\!\!\!\perp \mathcal{F}_s^W$

(2) Processo di Poisson: Sia $N(t)$ il numero d'arrivi dell'intervallo $[0, t]$ con intensità costante λ

Risulta che $\varphi(t) = E[N(t)] = \lambda t$ non è costante, dunque non può trattarsi di una

Martingala, che ha funzione Valor Medio costante. Se però consideriamo

$N^*(t) = N(t) - \lambda t$ otteniamo un processo stocastico con funzione Valor Medio identicamente nulla e, in particolare, una Martingala.

(3) Processo di Poisson-Calposto: Il processo $X(t) = \sum_{j=1}^{N(t)} Y_j$ non ha funzione Valor Medio costante, in quanto $E(X_t) = E(N(t)) \cdot E(Y_i) = \lambda t E(Y_i)$.

Se però consideriamo $X(t) - \lambda t E(Y_i)$ otteriamo una Martingala.

TEOREMA DI CONVERGENZA PER MARTINGALE: Sia X_n una Martingala tale che i momenti secondi siano uniformemente limitati, ossia $\exists M$ certo tale che $E(X_t^2) < M < \infty \quad \forall t$, allora: $\exists X$ tale che $X_n \xrightarrow{q.c.} X$.

PROCESSI STAZIONARI

Un processo $X_1, X_2, \dots, X_n, \dots$ si dice STAZIONARIO se:

$$F_{i_1, i_2, \dots, i_n}(x) = F_{i_1 + h, i_2 + h, \dots, i_n + h}(x) \quad \forall n, \forall i_1, \dots, i_n, \forall x \in \mathbb{R}^n, \forall h$$

↳ Condizione di invarianza della distribuzione congiunta
rispetto a traslazioni rigide

Implicazioni:

- I numeri aleatori del processo sono identicamente distribuiti (basta porre $n=1$);
- Le n -uple che hanno la medesima differenza h dei componenti sono identicamente distribuite (esempio: $F_{1,2}(x,y) = F_{100,101}(x,y)$ poiché $100-1=101-2$), mentre non è invece detto che tutte le n -uple siano in generale uguali in distribuzione (esempio: $F_{1,2}(x,y) \neq F_{1,3}(x,y)$);
- La definizione vale anche per processi a tempo continuo;
- Un processo di numeri aleatori IID è stazionario.

Proprietà:

- Funzione Valor Medio: $\Psi(n) = E(X_n) = \text{costante}$
 - Funzione di Covarianza: $\Psi(m,n) = \text{Cov}(X_m, X_n) = \Psi(|m-n|) \rightarrow$ Funzione dell'unico argomento $|m-n|$
- ⇒ La varianza è costante: $\text{Var}(X_n) = f(0)$

Distinguiamo due tipologie di stazionarietà:

- 1) Stazionarietà in senso stretto: $F_{i_1, i_2, \dots, i_n}(x) = F_{i_1 + h, i_2 + h, \dots, i_n + h}(x)$
- 2) Stazionarietà in senso lato o del II ordine: $\begin{cases} \Psi_X(n) = E(X_n) = \text{costante} \\ (\text{implicata dalla 1) se esistono finiti i primi due momenti}) \end{cases} \quad \begin{cases} \Psi_X(m,n) = \text{Cov}(X_m, X_n) = f(|m-n|) \end{cases}$

Nella pratica, infatti, può capitare di riuscire a specificare, per carenza di informazioni, soltanto i primi due momenti, ottenendo così soltanto la stazionarietà in senso lato. Se conoscessimo invece tutte le distribuzioni congiunte potremmo raggiungere il livello di specificazione massimale.

La scelta del livello di specificazione è consentita dalle informazioni disponibili. Quunque il livello di specificazione del II ordine possa sembrare "povero", esso consente comunque di compiere diverse operazioni.

PROBLEMA ERGODICO [Birkov, Von Neumann]

Sia $X_1, X_2, \dots, X_n, \dots$ un processo stazionario.

Sia $\frac{1}{n} \sum_{j=1}^n X_j^h$ il momento campionario h -esimo.

Sia $E(X_1^h)$ il comune momento SPAZIALE h -esimo (calcolato rispetto alla misura di probabilità dello spazio di probabilità (Ω, \mathcal{F}, P)).

Il processo si dice ERGODICO se: $\frac{1}{n} \sum_{j=1}^n X_j^h \rightarrow E(X_1^h) \forall h$

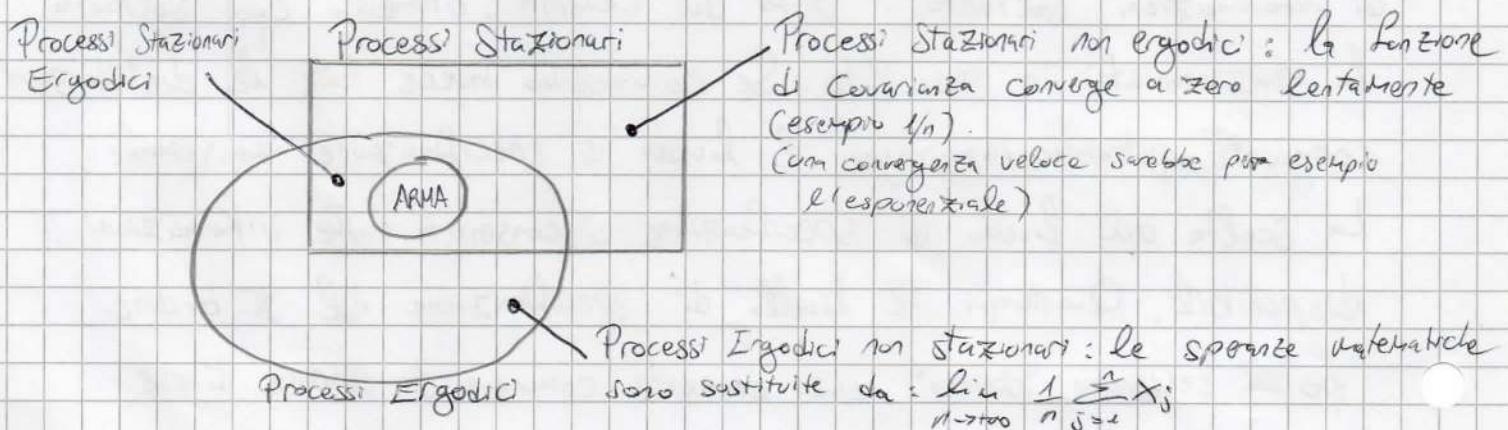
In tal caso i momenti spaziali possono essere approssimati, purché n sia sufficientemente grande, dai momenti campionari (o temporali).

Alla base vi è la Legge dei Grandi Numeri, ma noi assumiamo stazarietà, che è una condizione di dipendenza stocastica (e non di indipendenza come assunto da tale legge).

Considerando il livello di specificazione del II ordine, ci basta considerare la convergenza dei soli primi due momenti per l'ergodicità.

Si può provare che l'ergodicità dei momenti del I e II ordine è legata alla velocità di convergenza a 0 della funzione di covarianza al crescere di $|u-n|$. Se tale velocità è sufficientemente alta, allora vi è l'ergodicità.

Quasi tutta la statistica tradizionale si poggia sul problema ergodico, dando per scontato che la media campionaria stima la media della popolazione, assumendo però campioni IID. Ciò non è però così scontato in situazioni diverse, non IID, bensì staziarie o di autoreverse.



PROCESSI STAZIONARI DEL II ORDINE O IN COVARIAZIA

Siano:

$$\begin{cases} \Psi_x(n) = \text{costante} \\ \Psi_{x(m,n)} = \delta(m-n) \end{cases}$$

Esempi:

- Processo White Noise $WN(0, \sigma_E^2)$

→ [Caratterizzato da una DENSITÀ SPETTRALE COSTANTE, alla stregua delle componenti di frequenza della luce bianca tutte allo stesso livello]

Siano $E_1, E_2, \dots, E_n, \dots$ equi, con uguale varianza σ_E^2 e incorrelati.

Se non vi è alcuna assunzione sulla distribuzione di probabilità, il processo può essere specificato al livello del II ordine. Il processo è Stazionario in senso lato, in quanto:

- $\Psi_E(n) \equiv 0$
- $\Psi_E(m,n) = \begin{cases} \sigma_E^2 & \text{se } m=n \Leftrightarrow m-n=0 \\ 0 & \text{se } m \neq n \Leftrightarrow m-n \neq 0 \end{cases} \Rightarrow \text{Dipende solo da } |m-n|$

Questo processo stocastico normalmente è non prevedibile: qualunque sia l'informazione sui primi n valori del processo, nulla possiamo dire sull' $n+1$ -esimo.

Non c'è un processo analogo nel caso continuo.

- Sia $X_n = \sum_{j=1}^n E_j$ con $E_j \sim WN(0, \sigma_E^2)$

Poiché il processo E_n è stazionario, ci chiediamo se lo è anche il processo delle somme successive, specificando al livello del II ordine:

- $\Psi_X(n) \equiv 0$
- $\Psi_X(m,n) = E(X_m X_n)$

In particolare $\text{Var}(X_n) = n \sigma_E^2$ (poiché gli E_j sono un corretto) non è costante, dunque il processo X_n non è stazionario.

Osservazioni:

- Scrivendo il processo come $X_n = X_{n-1} + E_n$ otteniamo un processo di PASSAGGIATA ALÉATORIA;
- Il processo è un caso particolare del Modello AR1 quando $a=1$;
- C'è analogia tra il processo X_n e una Martingala con la differenza che in questo caso la varianza degli E_n è la stessa. La mancanza di tale vincolo in una Martingala permette di guadagnare un grado di libertà.

• MODELLO ARMA(p,q) (Auto-Regressive, Moving Average)

$$X_n - \underbrace{\sum_{j=1}^p a_j X_{n-j}}_{\text{Componente AUTO-REGRESSIVA}} = \underbrace{\varepsilon_n + \sum_{i=1}^q b_i \varepsilon_{n-i}}_{\text{Componente MEDIA MOBILE}}$$

con $\varepsilon_n \sim NWN(0, \sigma_\varepsilon^2)$

• Se $p=0 \Rightarrow$ Modello MA(q)

• Se $q=0 \Rightarrow$ Modello AR(p)

Alcuni casi:

- ARMA(1,0) \Rightarrow Modello AR(1) : $X_n = a X_{n-1} + \varepsilon_n$, $\varepsilon_n \sim NWN(0, \sigma_\varepsilon^2)$

Per la gaussianità otteniamo: $\Psi_x(n) = a^n K$

$$\Psi_x(m,n) = a^{(n-m)} \sigma_\varepsilon^2 \frac{1-a^m}{1-a^n}$$

Se imponiamo la condizione: $-1 < a < 1$, per $n \rightarrow \infty, m \rightarrow \infty$:

$$\Psi_x(n) \rightarrow 0$$

$a^n \rightarrow 0$ e a^{n-m} dipende solo dalla differenza $n-m$, dunque la funzione di

Covarianza tende ad una funzione che dipende solo da $n-m$.

Il processo stocastico, dunque, tende ad una situazione di Stazionarietà in senso lato, detta perciò QUASI-STAZIONARIETÀ o STAZIONARIETÀ ASINTOTICA.

In generale, dunque, il modello AR(1) non genera un processo Stazionario, ma se $-1 < a < 1$ e al divergere di n e m , possiamo parlare di Stazionarietà Asintotica, molto utilizzata nella pratica in quanto la vera Stazionarietà nella realtà è difficile da incontrare.

- ARMA(2,0) \Rightarrow Modello AR(2): $X_n - a_1 X_{n-1} - a_2 X_{n-2} = \varepsilon_n$, $\varepsilon_n \sim NWN(0, \sigma_\varepsilon^2)$

Una trasformazione lineare consente di ricondursi ad un AR(1) vettoriale:

$$\text{Sia } \underline{Z}_n = \begin{bmatrix} X_n \\ X_{n-1} \end{bmatrix} = \underbrace{\begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix}}_A \begin{bmatrix} X_{n-1} \\ X_{n-2} \end{bmatrix} \underbrace{+ \begin{bmatrix} \varepsilon_n \\ 0 \end{bmatrix}}_{\underline{\varepsilon}_n} \Leftrightarrow \underline{Z}_n = A \underline{Z}_{n-1} + \underline{\varepsilon}_n \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(Equazione alle differenze finite del secondo ordine lineare stocastica)

Le soluzioni di quest'equazione sono processi stocastici. Per ottenere un'unica soluzione dobbiamo impostare due condizioni iniziali (in generale in un'equazione di ordine p dobbiamo impostare p condizioni iniziali).

Questa equazione è detta **COMPLETA** in quanto prevede sia un input a livello stocastico che uno a livello deterministico. Il processo E_n è specificato completamente fino al II ordine, quindi lo consideriamo come un dato.

Si dice **equazione OMogenea associata**: $X_n - a_1 X_{n-1} - a_2 X_{n-2} = 0$.

Si dice **Soluzione Generale** l'insieme di tutte le possibili soluzioni.

La **Soluzione generale dell'equazione Completa** si ottiene cercando la **Soluzione generale dell'equazione omogenea associata** aggiungendovi la **Soluzione Particolare** (con condizioni e vincoli) di quella completa.

Risoluzione: Poniamo $X_n = \lambda^n$

$$\Rightarrow \lambda^n - a_1 \lambda^{n-1} - a_2 \lambda^{n-2} = 0 \Leftrightarrow \lambda^{n-2}(\lambda^2 - a_1 \lambda - a_2) = 0 \Leftrightarrow \begin{cases} \lambda_0 = 0 & \text{Non di interesse} \\ \lambda_1 & \rightarrow \text{Reali o} \\ \lambda_2 & \text{complessi} \end{cases}$$

LEQUAZIONE CARATTERISTICA associata
all'equazione omogenea

Aggiungiamo la condizione che tutte le radici siano in modulo < 1 : $\begin{cases} |\lambda_1| < 1 \\ |\lambda_2| < 1 \end{cases}$

In generale da un'equazione di ordine P otteniamo $P+1$ radici, una delle quali non è di interesse (come $\lambda_0 = 0$).

- ARMA(0,1) \Rightarrow Modello MA(1): $X_n = E_n + b E_{n-1}$, $E_n \sim N(0, \sigma_E^2)$

• $\Psi_X(n) \equiv 0$ (E è equi)

$$\bullet \Psi_X(m,n) = E(X_m X_n) = E[(E_m + b E_{m-1})(E_n + b E_{n-1})] = E[E_m E_n + b E_m E_{n-1} + b E_n E_{n-1} + b^2 E_{m-1} E_{n-1}] =$$

$$= \begin{cases} \sigma_E^2 (1+b^2) & \text{se } m=n \rightarrow \text{comune varianza degli } X_n \end{cases}$$

$$\underbrace{E(E_{n-1} E_n)}_0 + \underbrace{E(b E_{n-1} E_{n-2})}_0 + \underbrace{E(b E_n E_{n-2})}_0 + \underbrace{E(b^2 E_{n-2} E_{n-1})}_0 = b \sigma_E^2 \quad \text{se } m=n-1$$

$$0 \quad \text{se } m \leq n-2 \rightarrow \text{DEFINITIVAMENTE NULLA}$$

In generale in un modello d'ordine q la covarianza sarà nulla per $m \leq n-q$.

[Questi modelli sono molto usati in economia in quanto si presa che valori istantanei nel tempo non siano tra loro correlati (esempio: PIL di oggi con PIL di 5 anni fa).]

Osserviamo che la covarianza dipende solo dalla differenza fra m e n .

Teorema di Hergen-Wold [1938]: Sia X_n un processo

stazionario del II ordine tale che $E(X_n) = 0$, allora:

$$\exists \text{ coefficienti } C_j \text{ tali che } \sum_{j=1}^{+\infty} C_j^2 < +\infty \mid X_n = \sum_{j=1}^{+\infty} C_j E_{n-j} + Y_n.$$

Dove:

- Y_n è un PROCESSO STOCASTICO DETERMINISTICO, ossia un processo che è perfettamente noto dopo un numero finito di osservazioni statistiche.

$$Y_n \perp\!\!\!\perp E_n$$

- E_n è un Processo White Noise

Osservazioni:

- Tutti i processi del II ordine sono quindi delle trasformazioni lineari di un processo di rumore bianco.
- In questo modello intervengono infiniti coefficienti C_j che devono essere stimati, perciò questo risultato non può essere utilizzato in maniera operativa e dobbiamo accontentarci di un'approssimazione con n opportunamente grande: N .

Anche in tal caso per stimare N coefficienti sono necessarie svariate centinaia di dati.

$$X_n = \sum_{j=1}^N C_j E_{n-j} + Y_n$$

Sì può però dimostrare che modelli lineari del tipo ARMA (p, q) con $p, q > 0$ costituiscono una buona stima per questa approssimazione. La presenza della componente autoregressiva riduce il numero di coefficienti da stimare a $p+q+1$.

Esempio: $X_n = E_n + Y$, $E_n \sim WN(0, \sigma_E^2)$

Con Y numero aleatorio tale che $E(Y) = 0$, $\text{Var}(Y) = \sigma_Y^2$, $\text{Corr}(Y, E_n) = 0$.

$$\cdot \Psi_X(n) = 0$$

$$\cdot \Psi_X(m, n) = E(X_m X_n) = E[(E_m + Y)(E_n + Y)] = \begin{cases} E[(Y+E_m)^2] = E(Y^2) + 2E(Y E_m) + E(E_m^2) = \sigma_Y^2 + 0 + \sigma_E^2 & \text{se } m=n \\ \sigma_Y^2 & \text{se } m \neq n \rightarrow \text{costante} \end{cases}$$

(La covarianza non risente della lontananza tra X_m e X_n nel tempo)

Osserviamo che in questo modello numeri anche molto distanti tra loro hanno la medesima covarianza (a differenza del modello ARMA); diremo quindi che X_n è SCAMBIABILE in covarianza.

PROCESSI SCAMBIABILI

Sono particolari processi stazionari proposti nel 1928 da Bruno de Finetti, formulando un teorema di rappresentazione della scambiabilità e mostrando come applicarla all' inferenza statistica. In quegli anni vi erano ancora dubbi sulla qualità matematica del calcolo delle probabilità, dunque il suo lavoro passò inosservato per un certo periodo. Nel 1933 Kolmogorov diede la propria definizione matematica della probabilità, vista come misura numerabile additiva (de Finetti parlò invece di finita additività, non ancora sviluppata tutt'oggi). Nel 1935 de Finetti diede la propria teoria della probabilità soggettiva, mentre nel 1937, in una conferenza internazionale a Ginevra, propose una nozione di scambiabilità parziale, che permette di considerare anche situazioni in cui le variabili osservabili non siano omogenee tra loro.]

L'idea della scambiabilità costituisce la base della statistica Bayesiana: il classico campione casuale semplice viene sostituito con un campione scambiabile in cui le variabili osservabili sono mutuamente dipendenti in maniera tale da rendere possibile l'apprendimento derivante dall'esperienza maturata dall'osservazione statistica. Da un punto di vista logico l'assunzione di indipendenza stocastica ci vieta qualsiasi miglioramento della valutazione su un numero aleatorio non ancora osservabile. Vedremo che l'ipotesi di scambiabilità, dunque di dipendenza, equivale ad un'ipotesi di indipendenza ed identica distribuzione condizionate.

Definizione: Il processo X_n è SCAMBIABILE se: $F_{i_1, \dots, i_n}(x) = F_{1, 2, \dots, n}(x) \quad \forall n \quad (i_1, \dots, i_n), x$

Quest'uguaglianza può essere interpretata in due modi:

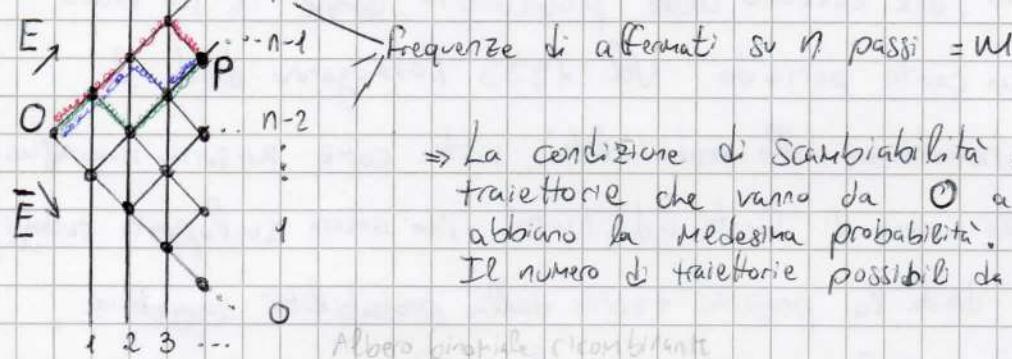
- 1) $\{i_1, \dots, i_n\}$ come permutazione arbitraria di $\{1, 2, \dots, n\}$: parliamo di INVARIANZA della distribuzione congiunta rispetto all'ordine, che non è quindi influente.
- 2) $\{i_1, \dots, i_n\}$ come qualsiasi n -upla di numeri interi distinti fra loro: una qualsiasi distribuzione n -variata dipende soltanto da n , dunque tutte le distribuzioni n -dimensionali hanno la medesima informatività, che può essere ricondotta alla prima n -upla $\{1, 2, \dots, n\}$.

PROCESSI SCAMBIABILI DI ALTERNATIVA

Sia $X_n = |E_n|$ un processo di INDICATORI DI EVENTO.

Sia X_n SCAMBIABILE, ossia: $\Pr(E_{i_1} = x_1 \wedge \dots \wedge E_{i_n} = x_n) = \Pr(E_1 = x_1 \wedge \dots \wedge E_n = x_n)$

Poiché $x_1, \dots, x_n \in \{0, 1\}$ possiamo tracciare le traiettorie possibili del processo su un reticolo in cui si sale se l'evento è affermato e si scende altrimenti:



⇒ La condizione di scambiabilità comporta che tutte le traiettorie che vanno da O a un generico punto P abbiano la medesima probabilità.
Il numero di traiettorie possibili da O a P è $\binom{n}{m}$ → numero totale di passi → numero di affermati

Per esempio: $\Pr(E_{i_1} \wedge \bar{E}_{i_2} \wedge E_{i_3}) = \Pr(E_1 \wedge \bar{E}_2 \wedge E_3) \rightarrow$ La probabilità di questo percorso coincide con quella di ogni altro percorso che abbia la stessa frequenza di successi.

Indichiamo questa probabilità con il simbolo $\underline{w}_m^{(n)}$ = probabilità che su n eventi m siano affermati.

Risulta quindi:

$$1) \underline{w}_j^{(n)} \geq 0 \quad \forall j \leq n$$

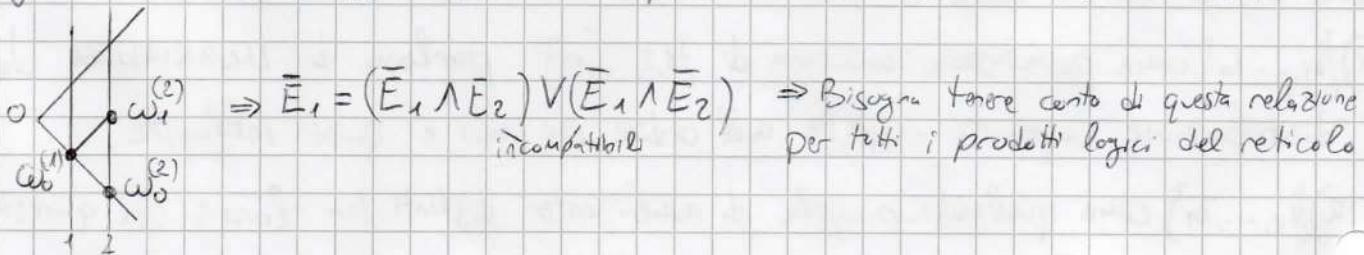
$$2) \sum_{j=0}^n \underline{w}_j^{(n)} = 1$$

$$3) \text{Coerenza: } \frac{\underline{w}_m^{(n)}}{\binom{n}{m}} = \frac{\underline{w}_m^{(n+1)}}{\binom{n+1}{m}} + \frac{\underline{w}_{m+1}^{(n+1)}}{\binom{n+1}{m+1}} \quad \forall n, \forall m \quad (\text{rappresenta infinite condizioni})$$

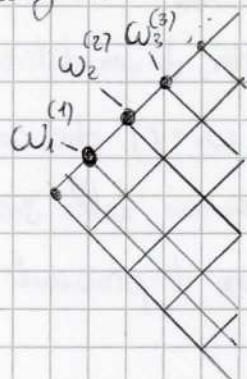
Poiché ci sono $\binom{n}{m}$ traiettorie, rappresenta la probabilità di una singola traiettoria

Supponendo infatti di assegnare una probabilità a tutti i vertici del reticolo $\{\underline{w}_m^{(n)}, m \leq n\}$ le prime due condizioni non bastano.

Bisogna infatti tenere conto che in un passo ci sono solo due alternative:



Si dimostra che per assegnare la struttura probabilistica del processo basta assegnare in modo coerente le probabilità sui nodi della scatola superiore:



Basta dunque assegnare le probabilità $w_n^{(n)}$, che indicheremo con w_n .

$$\Rightarrow \{w_n : n \geq 0, w_0 = 1\}$$

La condizione di coerenza che lega tra loro le probabilità w_n è detta COMPLETA MONOTONIA: $(-1)^r \Delta^r(w_n) \geq 0 \quad \forall r, n$

Data la successione $w_0, w_1, w_2, \dots, w_n, \dots$ consideriamo gli incrementi del processo $\Delta(w_n)$ degli incrementi di probabilità in un passo [utili per stabilizzare le serie stocastiche: dopo un certo numero di passi la serie diventa approssimativamente stazionaria]:

$$\text{- Ordine 0: } \Delta^0(w_n) = w_n$$

$$\text{- Ordine 1: } \Delta^1(w_n) = w_{n+1} - w_n$$

$$\text{- Ordine 2: } \Delta^2(w_n) = w_{n+2} - 2w_{n+1} + w_n = (w_{n+2} - w_{n+1}) - (w_{n+1} - w_n)$$

$$\text{- Ordine } r: \Delta^r(w_n) = \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} w_{n+j}$$

La condizione di completa monotonia diventa allora:

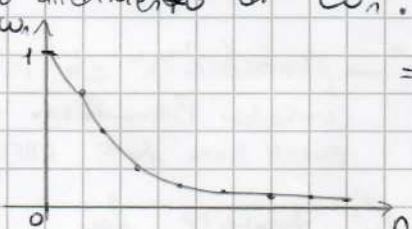
$$(-1)^r \Delta^r(w_n) = (-1)^r \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} w_{n+j} = \sum_{j=0}^r \binom{r}{j} (-1)^{2r-j} w_{n+j} = \sum_{j=0}^r \binom{r}{j} (-1)^j w_{n+j} \geq 0$$

Poiché $2r$ è pari, dal punto di vista del segno, questo è lo stesso con j .

$$\text{Se } j=1: -(w_{n+1} - w_n) \geq 0 \Leftrightarrow w_n - w_{n+1} \geq 0$$

$$\text{Se } j=2: -(w_{n+2} - w_{n+1}) + (w_{n+1} - w_n) \geq 0$$

Tenendo conto solo di queste due condizioni per ogni n possiamo tracciare l'andamento di w_n :



\Rightarrow La successione (considerando solo le differenze prime e seconde) è monotona non crescente a tassi decrescenti.

Teorema di Hausdorff: I momenti di una distribuzione concentrata sull'intervallo $[0, 1]$ soddisfano la Completa Monotonia.

Nel 1927 Hausdorff diede una dimostrazione non costruttiva, ma d'uso esistenza, di una distribuzione di probabilità composta da una successione numerica monotona normalizzata (con $w_0 = 1$) che costituisce i momenti di una distribuzione Completamente Monotona. A partire dai momenti della successione Completamente Monotona sarà poi de Finetti a definire la distribuzione di probabilità.

Dimostrazione

$$(-1)^r \Delta^r(w_n) = (-1)^r \Delta^r \left(\int_0^1 x^n F(dx) \right) = (-1)^r \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} \int_0^1 x^{n+j} F(dx) = \sum_{j=0}^r \binom{r}{j} (-1)^j \int_0^{n+j} F(dx) =$$

\hookrightarrow Per il Teorema di de Finetti: $w_n = \int_0^1 x^n F(dx)$

$$= \int_0^1 x^n \left[\sum_{j=0}^r \binom{r}{j} (-1)^j x^j \right] F(dx) = \int_0^1 \underbrace{x^n (1-x)^r}_{\substack{\geq 0 \\ \text{funzione su } [0, 1]}} F(dx) \geq 0 \Rightarrow \text{Completa Monotonia}$$

\checkmark c.v.s.

Binomio di Newton: $(1-x)^r$

Osservazione: De Finetti prese come successione completamente monotonica la successione delle probabilità w_n e provò che la relativa distribuzione di probabilità concentrata sull'intervallo $[0, 1]$ del teorema è: $\lim_{n \rightarrow +\infty} \frac{F_{S_n}(x)}{n} = F(x)$.

In sintesi ci sono tre modi per assegnare la distribuzione di probabilità di X_n :

i) $w_m^{(n)} = \Pr(S_n = m)$ con $w_m^{(n)} \geq 0$, $\sum_{j=0}^n w_j^{(n)} = 1$, $\frac{w_m^{(n)}}{\binom{n}{m}} = \frac{w_m}{\binom{n+m}{m}} + \frac{w_{m+1}}{\binom{n+m+1}{m+1}}$;

ii) $w_n^{(n)} = w_n = \Pr(S_n = n)$ con $(-1)^r \Delta^r(w_n) \geq 0$ (infinte condizioni non semplici da rispettare)

iii) $F(x)$ tale che $\int F(x) dx = 0$



$\Rightarrow F$ concentrata su $p \Rightarrow$ otteniamo un processo BERNOULLIANO,

Ciò implica il legare con: $w_n = \int_0^1 x^n F(dx) = p^n \rightarrow$ probabilità che in un

processo Bernoulliano n eventi siano tutti veri

Osserviamo che fissate le w_n possiamo ricavare le $w_m^{(n)}$ tramite la seguente relazione: $w_m^{(n)} = (-1)^{n-m} \Delta^{n-m}(w_n)$.

Esempio: $w_0^{(3)} = -\Delta^3(w_0) = 1 - (w_3 - w_2 + w_1)$; $w_1^{(3)} = \Delta^2(w_1) = w_3 - 2w_2 + w_1$; $w_2^{(3)} = -\Delta(w_2) = w_2 - w_3$;

$w_3^{(3)} = w_3 \Rightarrow$ sommando queste quantità otteniamo $= w_3 - w_2 + w_1$.

Teorema di Bruno de Finetti: Tutti e solo i processi scambiabili illimitati sono combinazioni lineari convesse di processi stocastici i cui numeri aleatori sono bernoulliani, dunque indipendenti e identicamente distribuiti.

In generale sia E un evento logicamente dipendente da un numero finito di indicatori scambiabili, allora:

$$\Pr(E) = \int_0^1 P_x(E) F(dx) \rightarrow \text{misura di probabilità}$$

\leftarrow Probabilità dell'evento E valutata nell'ipotesi che gli indicatori non siano scambiabili, bensì elementi di un processo bernoulliano (ossia IID) caratterizzato dalla comune probabilità X degli eventi
 $X \in [0,1]$
 in quanto comune probabilità degli eventi

In particolare:

$$\cdot \Pr(S_n = n) = \omega_n = \int_0^1 x^n F(dx) = \lim_{n \rightarrow +\infty} F_{S_n/n}(x)$$

$$\cdot \Pr(S_n = m) = \omega_m^{(m)} = \int_0^1 \binom{n}{m} x^m (1-x)^{n-m} F(dx)$$

Il teorema implica che i numeri aleatori osservabili del processo scambiabile, se condizionati ad un dato ente aleatorio, sono indipendenti ed identicamente distribuiti. Il condizionamento avviene rispetto alla σ -algebra degli EVENTI SIMMETRICI (eventi il cui valore logico vero o falso non varia rispetto a permutazioni arbitrarie degli argomenti di una distribuzione di probabilità congiunta).

Dimostrazione

La dimostrazione procede per passi:

(1) Si dimostra che per $n \rightarrow +\infty$ le frequenze relative di successo del processo scambiabile di indicatori convergono in probabilità ad un numero aleatorio bernoulliano, ossia:

$$\frac{S_n}{n} = \frac{1}{n} \sum_{j=1}^n |E_j| \xrightarrow[n \rightarrow +\infty]{P} X \text{ tale che } E(X) = \omega_i \xrightarrow{\text{comune probabilità}} \text{ degli eventi del processo scambiabile}$$

(2) Dalla convergenza delle frequenze relative discende la convergenza delle distribuzioni: $F_{\frac{S_n}{n}}(x) \xrightarrow{n \rightarrow \infty} F(x)$
dove $F(x)$ è la distribuzione di probabilità di X .

In questo punto del teorema de Finetti riesce a definire la distribuzione di probabilità F di cui Hausdorff riuscì solo a dimostrare l'esistenza.]

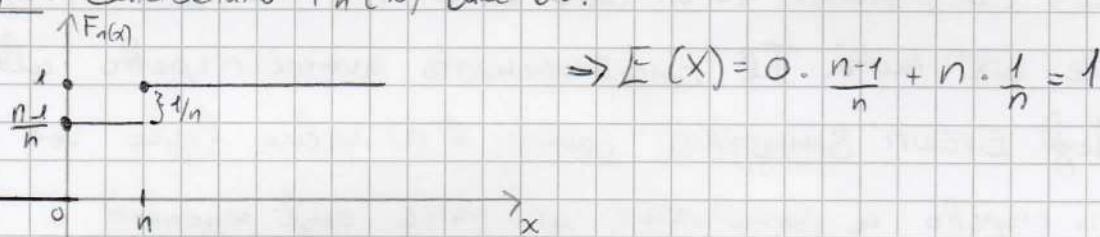
(3) Si prova che il generico momento delle frequenze relative di successo è tale che per $n \rightarrow \infty$: $E\left[\left(\frac{S_n}{n}\right)^h\right] \xrightarrow{n \rightarrow \infty} \int_0^1 x^h F(dx) \rightarrow$ per la limitatezza delle distribuzioni:

Per l'unicità del limite discende che: $c_{lh} = \int_0^1 x^h F(dx)$

↓
Coincide con il momento h -esimo della distribuzione limite

Osservazione: Non necessariamente la convergenza in distribuzione implica la convergenza dei momenti. Ciò avviene solo se la probabilità è concentrata su intervalli di grandezza finita, dunque limitati.

Esempio: Consideriamo $F_n(x)$ tale che:



Per $n \rightarrow \infty$ la probabilità associata a n decresce tendendo a 0, mentre la probabilità associata a 0 tende a 1: $\xrightarrow{n \rightarrow \infty} \Rightarrow E(X) = 0$.

La sequenza converge ad una distribuzione limite, ma la speranza matematica di queste distribuzioni non converge alla speranza matematica della distribuzione limite.

Esempio:

• Processo di BAYES-LAPLACE

È un processo scambiabile con funzione di ripartizione uniforme sull'intervallo $[0,1]$ e dunque densità costante identicamente uguale a 1 su $[0,1]$.



Per il teorema di de Finetti: $\omega_n = \int_0^1 x^n f(x) dx = \int_0^1 x^n dx = \frac{x^{n+1}}{n+1} \Big|_0^1 = \frac{1}{n+1}$

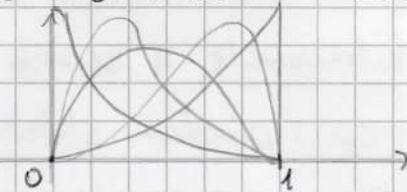
Si può inoltre verificare che $\omega_m^{(n)} = \frac{1}{n+1} \forall n, m$, dunque le determinazioni della frequenza relativa hanno la medesima probabilità.

• Processo di POLYA

Il processo è determinato da una distribuzione dotata di densità del tipo Beta.

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

$\alpha, \beta \in \mathbb{R}_+$



→ concentrata su $[0, 1]$

Per il teorema di de Finetti: $\omega_n = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \underbrace{x^{\alpha-1} (1-x)^{\beta-1}}_{\text{nucleo } \downarrow \text{ una Beta}(\alpha, \beta)} dx$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+n)\Gamma(\beta)}{\Gamma(\alpha+n)\Gamma(\beta+n)} = \frac{(\alpha)_n}{(\alpha+\beta)_n}$$

$$= \frac{\Gamma(\alpha+n)\Gamma(\beta)}{\Gamma(\alpha+\beta+n)}$$

[Se $\alpha \in \mathbb{R}$ e $n \in \mathbb{N}$ vale la PROPRIETÀ DI EULERO: $\Gamma(\alpha+n) = (\alpha)_n \cdot \Gamma(\alpha)$
 $\Rightarrow = \alpha(\alpha+1) \cdots (\alpha+n-1)$
(SIMBOLO DI POCHHAMMER)]

Osservazione: I processi scambiabili sono processi stazionari

non ergodici in quanto la funzione di covarianza non si annulla asintoticamente o, se lo fa, non tende a zero rapidamente.

Inoltre i momenti campionari di un certo ordine non convergono ai corrispondenti momenti spaziali.

Nonostante ciò la scambiabilità, grazie al teorema di de Finetti, riesce comunque a conservare molti vantaggi dello schema classico (come il principio di induzione).

PROCESSI SCAMBIABILI NON DI ALTERNATIVA, ILLIMITATI

Teorema di Bruno de Finetti (caso generale): Sia E un evento logicamente dipendente da un numero finito di numeri aleatori scambiabili, allora:

$$\Pr(E) = \sum_F P_F(E) \mu(dF)$$

Spazio
funzionale

Probabilità di E valutata nell'ipotesi che i numeri aleatori X_i non siano scambiabili, bensì IID con comune funzione di ripartizione F .

Otteniamo una mistura di probabilità $P(F(E))$ con pesi una misura di probabilità μ definita sull'insieme \mathcal{F} di tutte le possibili funzioni di ripartizione univariate.

A rigor: μ andrebbe fissata come F nel caso precedente, ossia in maniera generica, ma aggiriamo questo ostacolo fissando la forma funzionale della comune distribuzione delle variabili arbitrariamente a seconda del tipo di problema da affrontare (allo stesso modo con cui i problemi statistici suggeriscono forme funzionali campionarie come gaussiane, Poisson, etc.). Una volta fissata la forma funzionale la mistura sarà realizzata in base agli specifici parametri di quella forma, attribuendo ad essa una distribuzione di probabilità.

Esempi:

- Gaussiana $\Rightarrow F(x) = \int_{-\infty}^x (2\pi\sigma^2)^{-1/2} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$

In questo caso i parametri a cui attribuire una distribuzione sono (μ, σ^2) .

- Poisson $\Rightarrow \Pr(N_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$

Qui il parametro a cui attribuire una distribuzione è λ .

Dunque ogni processo stocastico scambiabile illimitato è rappresentabile come mistura di processi i cui numeri aleatori sono IID.

INFERENZA STATISTICA BAYESIANA

La Statistica Bayesiana si dice sia nata nel '600 ad opera del filosofo ecclesiastico Thomas Bayes. Tale disciplina è stata dimenticata per molto tempo in quanto mancava la nozione fondamentale di probabilità soggettiva.

Sia H il parametro "non completamente noto" della distribuzione di un processo stocastico osservabile X .

Nell'impostazione classica tale parametro da stimare è assunto incognito al punto da essere considerato come una costante. Nell'impostazione bayesiana, invece, non si afferma che il parametro sia completamente incognito (se non altro perché se ne possono conoscere i possibili valori), dunque non è del tutto corretto parlare di parametro "incognito" ed è preferibile definirlo come "parzialmente o malamente noto", in quanto si pensa di potervi attribuire una distribuzione di probabilità dipendente dalle conoscenze soggettive in avанto al fenomeno.

Alla base di quest'impostazione, infatti, vi è il cosiddetto "Dogma Bayesiano", secondo cui è sempre possibile, qualunque informazione si disponga, scegliere una distribuzione unica all'interno di qualsiasi famiglia.

Sia indicato con H lo stato di informazione iniziale dell'individuo I; H contiene non solo l'informazione disponibile sul fenomeno in analisi, ma anche tutta la conoscenza e l'insieme di opinioni dell'individuo. Ciascuna valutazione sarà dunque strettamente legata ad H .

È improbabile che un secondo individuo J abbia il medesimo stato H . Tale stato di informazione non è però statico: si assume infatti che esso possa crescere (e dunque fornire risultati più precisi) mediante l'esperienza, ossia l'osservazione del fenomeno. Se l'osservazione delle variabili porta informazioni sulle future variabili, allora cade la classica ipotesi di indipendenza dal punto di vista logico.

Ipotesi della Statistica Bayesiana PARAMETRICA:

(1) $X_i | \mathcal{H} = \theta$ IID FO, ossia X_i SCAMBIABILI

\leftarrow X_i condizionate alla generica ipotesi sul valore del parametro non noto

↳ Le variabili osservabili sono equivalenti dal punto di vista informativo, dunque l'incremento d'informazione non dipende dalla singola variabile che si osserva.

(2) $X_i | \mathcal{H} = \theta \sim f(x|\theta)$ → Data la distribuzione campionaria condizionata

(3) $\text{Oltre } g(\theta | \mathcal{H}) \stackrel{d}{=} g_{\mathcal{H}}(\theta)$ → Opinione soggettiva sulla distribuzione di probabilità del parametro θ nello stato di informazione iniziale H .

Osserviamo che queste tre ipotesi rappresentano tre scelte di lavoro soggettive e arbitrarie, sempre condizionate ad H che non rappresenta un vero evento, bensì un più ampio insieme di osservazioni precedenti, opinioni, suggestioni e molti altri fattori soggettivi complessi.

Queste tre ipotesi individuano la cosiddetta LEGGE TEMPORALE INIZIALE

• STRUTTURA PROBABILISTICA per il processo: $\Lambda(H) = \{f_{1,2,\dots,n}(x), n \geq 1\}$, ossia l'insieme di tutte le possibili probabilità congruenti per un numero finito arbitrario di numeri aleatori osservabili. Siamo dunque in grado di attribuire una valutazione al livello massimale, pur non avendo effettuato ancora alcuna osservazione.

Supponiamo di poter osservare le prime n variabili del processo, ottenendo l'INCREMENTO DI INFORMAZIONE $K_n = (\underline{X}_n - \underline{x})$.

Nell'impostazione classica K_n sarebbe l'unica informazione campionaria, senza alcuna informazione pregressa; qui invece è aggiuntiva alla già presente opinione iniziale.

Osserviamo che l'incremento d'informazione è assunto soltanto come insieme di valori osservati e da nessun altro elemento, quando in realtà la complessità d'un'osservazione sarebbe maggiore: in un'osservazione, infatti, concretamente otteniamo anche ulteriori informazioni (potremmo ad esempio notare che la sequenza è crescente) che potrebbero indurci a cambiare le ipotesi e dunque modificare il modello.

Essendo le variabili scambiabili, le prime n danno la stessa informazione di qualsiasi altra nupla.

L'arrivo di nuova informazione ci costringe, per coerenza, a modificare l'opinione iniziale, ottenendo una distribuzione a posteriori dalla quale potremmo, per esempio, ricavare una stima puntuale (anche se ciò costituisce una perdita di informazione).

L'evoluzione dello stato di informazione è assunta crescente nel tempo, perciò è formalizzata mediante l'introduzione di una **FILTRAZIONE**, che permette di passare in maniera coerente da: $P(E)$ a $P(E)$.

$$I, H_1 \quad I, H_2$$

Per sfruttare razionalmente l'incremento di informazione K_n vi sono due metodi:

(1) Approccio Inferenziale Bayesiano INDIRETTO: Utilizza la norma di coerenza del Teorema di Bayes per passare dalla distribuzione iniziale $g_H(\theta)$ a quella finale dopo l'osservazione $g(\theta | H \wedge K_n) = g_H(\theta | K_n) \rightarrow$ Poiché H è l'evento \wedge l'evento

Il procedimento si dice Indiretto perché il parametro non noto θ ha solo un rapporto indiretto con le variabili osservabili ed interviene solo nella distribuzione condizionata.

$$\Rightarrow \text{Teorema di Bayes per densità di probabilità: } g_H(\theta | K_n) = \frac{g_H(\theta) \cdot l(\theta | K_n)}{\int_R g_H(\theta) l(\theta | K_n) d\theta} \propto g_H(\theta) \cdot l(\theta | K_n)$$

Sarebbe $\int_R g_H(\theta) d\theta = 1$
costante di normalizzazione
tale che $\int_R g_H(\theta) d\theta = 1$

A questo punto, avendo osservato X_1, \dots, X_n ora costanti, e nota la distribuzione a posteriori, possiamo trovare la distribuzione congiunta delle variabili non ancora osservate, che indichiamo con $\xi_{n+1}, \dots, \xi_{n+m} \rightarrow$ PROCESSO OSSERVABILE RESIDUO

Tale distribuzione $f(\xi)$ è detta DISTRIBUZIONE PREVISIONALE:

$$f_{H\xi}(\xi_{n+1}, \dots, \xi_{n+m} | K_n) = \int_R f(\xi | \theta) g_H(\theta | K_n) d\theta \rightarrow \text{combinazione linea convessa}$$

(Teorema di de Finetti)

Osservazioni:

- Abbiamo assunto che valga formalmente l'egualianza: $f(x | \theta) = f(\theta | x)$ che sono in generale funzioni diverse (gli integrali assumono diversi valori), ma, se sia X che θ sono noti, esse assumono lo stesso valore da più essere vista come:
 - densità di X rispetto al valore θ del parametro;
 - Verosimiglianza di θ rispetto al particolare valore di X .
 Inoltre normalizzando la verosimiglianza ($\frac{f(\theta | x)}{\int_\theta f(\theta | x) d\theta}$) il rapporto si comporta come una densità (non negativa con integrale = 1).

• Il procedimento Indiretto, poiché sfrutta il Teorema di de Finetti, vale soltanto per processi SCAMBIABILI.

• Alla base vi è l'ipotesi che la scambiabilità non venga modificata dal processo di osservazione: il processo scambiabile resta tale anche dopo le osservazioni. Ciò non è scattato, in quanto in generale l'opinione sulla natura del processo potrebbe cambiare (per esempio da scambiabilità a markovianità).

• Supponiamo che lo stato iniziale di informazione H sia scarso, allora la distribuzione iniziale è poco attendibile. Se le osservazioni statistiche sono tante l'opinione finale sarà costituita da una combinazione tra il paio che sapevamo prima e il consistente incremento acquistato. In tal caso l'opinione iniziale acquista sempre meno importanza in favore della nuova opinione.

(2) Approccio Inferenziale Bayesiano DIRETTO: Utilizza la norma di coerenza del Teorema delle Probabilità Composte per passare dalla Legge Temporale Iniziale $\Lambda(H)$ alla Legge Temporale Finale $\Lambda(H \wedge K_n)$.

Il procedimento si dice Diretto in quanto non fa intervenire alcun parametro, basandosi esclusivamente sull'osservabilità dei numeri aleatori.

A differenza dell'approccio Indiretto, questo approccio non richiede la scambiabilità, dunque può essere applicato a qualsiasi processo.

Siano $\xi_{i1}, \xi_{i2}, \dots, \xi_{im}$ una generica m -pla osservabile residua (con $i_1, i_2, \dots, i_m > n$):

\Rightarrow Teorema delle Probabilità Composte : $f_H(\xi_{i1}, \dots, \xi_{im} | K_n) = \frac{f_H(x_1, \dots, x_n, \xi_{i1}, \dots, \xi_{im})}{f_H(x_1, \dots, x_n)} \in \Lambda(H \wedge K_n)$

Esempio: Sia $f_H(x_i, x_i) \in \Lambda(H)$ e $K_1 = (X_1 = 3)$: Il fatto di aver osservato che $X_1 = 3$ costringe a modificare la legge tenendone conto. Se prima era ammesso qualsiasi valore per x_1 , ora tutte le densità corrispondenti a cui $x_1 \neq 3$ sono nulle.

Proposizione: Nell'ipotesi di scambiabilità delle variabili osservabili i due procedimenti Diretto e Indiretto giungono allo stesso risultato.

Dimostrazione:

$$f_H(\xi | K_n) = \int_R f(\xi | \theta) g_H(\theta | K_n) d\theta = \int_R f(\xi | \theta) \cdot g_H(\theta) \frac{f(x_1, \dots, x_n | \theta)}{\int_R g_H(\theta) f(x | \theta) d\theta} d\theta$$

$$= \int_R f(\xi | \theta) \frac{f(x | \theta) g_H(\theta)}{\int_R g_H(\theta) f(x | \theta) d\theta} d\theta = \frac{f_H(x_1, \dots, x_n; \xi_{i1}, \dots, \xi_{im})}{f_H(x_1, \dots, x_n)}$$

può essere interpretata come
verosimiglianza o densità $I(\theta | K_n)$

IID congruenza di Markovitch

Le densità hanno lo stesso condizionante

c.v.d.

Tale equivalenza sussiste solo se è applicabile il teorema di de Finetti e se le variabili sono scambiabili. In caso contrario è applicabile soltanto il procedimento Diretto.

Cenni su PROBABILITÀ IMPRECISE:

Negli anni '70, per perorare il Dogma Bayesiano, de Finetti fece un esperimento: propose a diversi giocatori di schedine di attribuire una probabilità soggettiva alle partite. Inizialmente questi non erano in grado di farlo, dunque tendevano ad attribuire la stessa probabilità di $\frac{1}{3}$ ai risultati 1, x, 2. Col passare del tempo si sono abituati sempre meglio a formulare valutazioni numeriche. Poiché soggettivamente è difficile dare una valutazione puntuale, la valutazione era invece intervalle (per esempio, anziché dire 0.43 si diceva tra 0.4 e 0.45). Nacque così la Teoria delle Probabilità Imprecise.]

Se l'informazione iniziale è scarsa non è pensabile scegliere con convinzione dei valori numerici puntuali per i parametri della distribuzione, perciò si può pensare che questi appartengano ad un intervallo (per esempio nella Beta(α' , β') e $\beta''(\mu', \mu'')$). Assumendo che sia però sempre possibile fornire gli estremi degli intervalli. In tal caso sono scelte famiglie di infinite distribuzioni iniziali e potremo pensare di applicare il teorema di Bayes a ciascuna di queste infinite densità, ottenendo un'ultima legge finale costituita da infinite distribuzioni: $\mathcal{F}_1 \xrightarrow{K_0} \mathcal{F}_2$.

Ottieniamo quindi, per esempio, un intervallo di valori di probabilità:

$$\inf_{g \in \mathcal{F}_2} P(E) \leq P_{\pi}(E|K) \leq \sup_{g \in \mathcal{F}_2} P(E)$$

Esiste anche un altro approccio bayesiano detto Gerarchico, in cui intervengono distribuzioni su vari livelli. Si può pensare, per esempio, che (α, β) siano parametri di II livello, ossia aventi anch'essi una distribuzione (di II livello) che a sua volta avrà dei parametri (di III livello).

La teoria di de Finetti si sviluppò nel 1935, due anni dopo la monografia di Kolmogorov. Quest'ultimo diede un punto di vista formale delle probabilità, mentre de Finetti ne diede uno interpretativo.]

• SCHEMA POISSON - GAMMA

Consideriamo un processo di arrivi subordinatamente poissoniano di intensità non nota Θ . In corrispondenza alla generica ipotesi $\Theta = \Theta(\epsilon R^+)$ possiamo considerare le variabili:

- $N(T)$: numero aleatorio di arrivi nell'intervallo $[0, T]$. Il processo è a tempo continuo ma può essere ricordato a tempo discreto suddividendo il tempo in intervalli unitari $(N_1, N_2, \dots, N_i, \dots)$.
- T_j : tempo d'attesa tra l'arrivo $j-1$ e il j -esimo. Il processo è a tempo discreto. (INTER-ARRIVAL TIME)
- S_n : tempo d'attesa per l' n -esimo arrivo. $S_n = \sum_{j=1}^n T_j$ è a tempo discreto.

Se consideriamo Θ aleatorio i primi due processi sono scambiabili condizionatamente a $\Theta = \Theta$ (inoltre N_T e T_j IID), mentre ciò non vale per S_n , che diventa un processo di Passeggiate Aleatorie. ($S_2 \geq S_1$, quindi non sono indipendenti).

- Inferenza sui tempi d'attesa T_j

Ipotesi:

(1) $T_j | \Theta = \Theta$ IID $\forall j$, ossia T_j scambiabili

(2) $T_j | \Theta = \Theta \sim \text{Exp}(\Theta)$, ossia $f(t | \Theta = \Theta) = \Theta e^{-\Theta t}$

(3) $g_\Theta(\Theta) \sim \text{Gamma}(\alpha, \beta)$, ossia $g_\Theta(\Theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \Theta^{\alpha-1} e^{-\beta\Theta}$

Supponiamo di avere informazioni sufficienti in H per fissare α e β .

Sulla base delle ipotesi e nello stato H valutiamo la densità su una variabile: $f_{T_1}^{(H)}(t)$

Per il Teorema di de Finetti: $f_{T_1}^{(H)}(t) = \int_0^{+\infty} f(t | \Theta) g_\Theta(\Theta) d\Theta = \int_0^{+\infty} (\Theta e^{-\Theta t}) \cdot \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \Theta^{\alpha-1} e^{-\beta\Theta} \right) d\Theta = \frac{\beta^\alpha}{\Gamma(\alpha)} \underbrace{\int_0^{+\infty} \Theta^{\alpha-1} e^{-\Theta(\beta+t)} d\Theta}_{\text{integrale del nucleo di una Gamma } (\alpha+1, \beta+t)} = \frac{\alpha! \beta^\alpha}{(\beta+t)^\alpha} \rightarrow$ densità di probabilità di T_1 , uguale per tutte le altre

$$\text{integrale del nucleo di una Gamma } (\alpha+1, \beta+t) = \frac{\Gamma(\alpha+1)}{(\beta+t)^\alpha}$$

Per quanto riguarda invece la densità congiunta di n -uple diverse tra loro:

$$f_{T_1, T_n}^{(H)}(t) = \int_0^{+\infty} f(t | \Theta) g_\Theta(\Theta) d\Theta = \int_0^{+\infty} (\Theta e^{-\Theta t}) \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \Theta^{\alpha-1} e^{-\beta\Theta} \right) d\Theta = \frac{\beta^\alpha}{\Gamma(\alpha)} \underbrace{\int_0^{+\infty} \Theta^{\alpha-1} e^{-\Theta(\beta+ \sum_{j=1}^n t_j)} d\Theta}_{\text{nucleo di una Gamma } (\alpha+n, \beta + \sum_{j=1}^n t_j)}$$

I numeri condizionati per la scambiabilità sono IID,

dunque la densità congiunta è il prodotto delle marginali:

$$= \frac{(\alpha)_n \beta^\alpha}{(\beta + \sum_{j=1}^n t_j)^{\alpha+n}} \in \Lambda(H) \rightarrow \text{Legge Temporale del Processo}$$

Osserviamo che i valori osservati intervergono attraverso una funzione simmetrica che non dipende dal loro ordine, ossia $\sum_{j=1}^n t_j$. Ciò avviene a causa della scambiabilità. 57

Sia $K_n = \prod_{j=1}^n (T_j = t_j)$ l'incremento di informazione osservato.

La corrispondente funzione di Verosimiglianza è:

$$l(\theta | K_n) = f(t | \theta) = \prod_{j=1}^n \theta e^{-\theta t_j} = \theta^n e^{-\theta \sum_{j=1}^n t_j}$$

↓ Variabili osservabili
Scambieribili

Possiamo in questo modo ottenere la Stima di Massima Verosimiglianza per θ , data

$$\hat{\theta} = \frac{n}{\sum_{j=1}^n t_j}$$

↳ tempo complessivo di osservazione

(nell'esponentiale negativo il tempo medio d'attesa è infatti $\frac{1}{\theta}$)

Sfruttiamo l'informazione per trovare la distribuzione a posteriori mediante il procedimento indiretto: $g_H(\theta | K_n) \propto g_H(\theta) l(\theta | K_n) = (\theta^{\alpha-1} e^{-\beta\theta}) \cdot (\theta^n e^{-\theta \sum_{j=1}^n t_j})$

$$= \theta^{\alpha+n-1} e^{-\theta(\beta + \sum_{j=1}^n t_j)} \rightarrow \text{Nucleo di una Gamma, detta GAMMA FINALE.}$$

$$(\alpha+n, \beta + \sum_{j=1}^n t_j) \rightarrow \Lambda(H|K) = \frac{f_H(t|K)}{f_H(t|H)} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)} \left(\frac{\beta}{\beta + \sum_{j=1}^n t_j} \right)^{\alpha+n}$$

Dunque, partendo da una Gamma, abbiamo ottenuto una distribuzione finale che è ancora una Gamma, ma con parametri aggiornati rispetto all'osservazione. Ciò non è scritto che avverga: se avessimo scelto un'altra distribuzione iniziale non sarebbe successo.

Il valor medio della distribuzione aggiornata è: $E(\theta | H|K_n) = \frac{\alpha+n}{\beta + \sum_{j=1}^n t_j}$

L'espressione può essere riscritta come:

$$E(\theta | H|K_n) = \frac{\beta \cdot \frac{\alpha}{\beta}}{\beta + \sum_{j=1}^n t_j} + \frac{\sum_{j=1}^n t_j \cdot n}{\beta + \sum_{j=1}^n t_j} = \frac{\beta \cdot E(\theta | H)}{\beta + \sum_{j=1}^n t_j} + \frac{\sum_{j=1}^n t_j \hat{\theta}}{\beta + \sum_{j=1}^n t_j}$$

Ottieniamo una combinazione lineare convessa tra il valor medio della distribuzione

Gamma iniziale ($\frac{\alpha}{\beta}$) e la stima di Massima verosimiglianza derivante dalla nuova informazione

$(\hat{\theta} = \frac{n}{\sum_{j=1}^n t_j})$, con pesi tali che al crescere del numero delle osservazioni $\sum_{j=1}^n t_j$

aumenta, dando più pesa a $\hat{\theta}$ a scopo del valor medio iniziale

(al divergere di n la mistura converge stocasticamente a $\hat{\theta}_n$). Alla lunga

dunque l'opinione iniziale tende a perdere importanza e i due approcci Diretto e Indiretto tendono a coincidere.

È pur vero che sono rare le situazioni in cui ci è concesso ottenere un numero arbitrario

↓ osservazioni omogenee tra loro (per esempio in Medicina o nelle serie storiche).

Un modo per fissare i valori α, β : Possiamo pensare di attribuire dei valori ai primi due momenti nello stato H in un sistema che, se ammette una sola soluzione, allora può fornire (α, β) :

$$E_H(\theta) = \alpha/\beta = M \quad \rightarrow \text{È solo uno degli infiniti modi, per cui il risultato non va considerato per certo.}$$

$$(V_H(\theta)) = \alpha/\beta^2 = Z^2$$

- Inferenza sul numero di arrivi $N_M \rightarrow$ Modello Poisson-Gamma

Ipotesi:

(1) $N_M | \Theta = \alpha \sim \text{ID} \forall \alpha$, ossia N_M scambiabili

(2) $N_M | \Theta = \alpha \sim \text{Po}(\alpha)$, ossia $\Pr(N_M = n | \Theta = \alpha) = e^{-\alpha} \frac{\alpha^n}{n!}$

(3) $g_H(\alpha) \sim \text{Gamma}(\alpha, \beta)$, ossia $g_H(\alpha) = \frac{\alpha^M}{M(\alpha)} \alpha^{\alpha-1} e^{-\beta\alpha}$
Assunti

Legge Temporale Iniziale: $P_H \left(\prod_{h=1}^M (N_h = n_h) \right) = \frac{M(d + \sum_{h=1}^M n_h)}{M(d)} \frac{\beta^d}{(\beta + M)} d + \sum_{h=1}^M n_h \in \Delta(H)$
Probabilità attribuita a M osservazioni prima di effettuarle

Sia $K_M = \prod_{h=1}^M (N_h = n_h)$ l'incremento di informazione osservato.

La corrispondente funzione di Verosimiglianza è: $l(\alpha | K_M) = \Pr \left(\prod_{h=1}^M (N_h = n_h) | \alpha \right) = \prod_{h=1}^M \left[e^{-\alpha} \frac{\alpha^{n_h}}{n_h!} \right] =$
 $= e^{-M\alpha} \frac{\alpha^{\sum_{h=1}^M n_h}}{M!} \Rightarrow$ Stima di Massima Verosimiglianza: $\hat{\alpha} = \frac{\sum_{h=1}^M n_h}{M}$

Utilizziamo la nuova informazione seguendo i due approcci:

1) Indiretto: $g_H(\alpha | K_M) \propto g_H(\alpha) l(\alpha | K_M) = e^{-\alpha(M+n)} \alpha^{(M+\sum_{h=1}^M n_h)\alpha} \sim \text{Gamma}\left(\frac{M+\sum_{h=1}^M n_h}{\alpha}, M+n\right)$
Anche in questo caso siamo passati da una Gamma ad un'altra Gamma con parametri aggiornati.

Osservazione: Della informazione K_M emergono soltanto M e la somma dei valori osservati, che rappresentano una STATISTICA SUFFICIENTE (conoscere N_1, \dots, N_M equivale a conoscere M e $\sum_{h=1}^M n_h$).
o Riassunto Caso 1

Non è un caso, in quanto la famiglia esponenziale si contraddistingue per essere dotata di statistiche sufficienti finito-dimensionali (in questo caso bidimensionale $(M, \sum n_h)$). Se la statistica non avesse dimensione finita (esempio: statistiche d'ordine) tale corrispondenza non sarebbe avvenuta.

Dunque: $P_H \left(\prod_{k=1}^N N_{M+k} = y_k | K_M \right) = \int_0^{+\infty} \left(\frac{e^{-\alpha N}}{\prod_{k=1}^N y_k!} \right) \left(K \alpha^{\left(\alpha + \sum_{h=1}^M n_h \right) - 1} e^{-(\beta + \alpha) \alpha} \right) d\alpha$
 $\in \Delta(H|K)$

2) Diretto: $P_H \left(\prod_{k=1}^N N_{M+k} = y_k | K_M \right) = \frac{P_H \left(\prod_{h=1}^M (N_h = n_h) \prod_{k=1}^N (N_{M+h} = y_k) \right)}{P_H \left(\prod_{h=1}^M (N_h = n_h) \right)}$

Osservazione: $0 \xrightarrow{T_1} 1 \xrightarrow{T_2} 2 \xrightarrow{T_3} 3 \dots \xrightarrow{T_M} M \xrightarrow{T_N} h \rightarrow \hat{\alpha}_N = \frac{\sum_{h=1}^M n_h}{M} ; \hat{\alpha}_T = \frac{M}{\sum_{j=1}^h t_j}$

Questi due tipi di osservazioni non sono esattamente equivalenti: nel caso di T noi teniamo conto del numero di osservazioni e del tempo complessivo di osservazione. Il tempo di osservazione non coincide con i tempi di arrivo: potremmo avere n arrivi all'inizio e null'altro dopo. Nel secondo caso contiamo solo i tempi necessari agli arrivi, e qui il tempo totale corrisponde al vero tempo per n arrivi.

SCHEMA BERNoulli - BETA, Inferenza su PROPORZIONI

Consideriamo un processo di estrazione da un'urna di Polya, in cui la pallina estratta viene reimbuossolata. Sia Θ il parametro malamente noto interpretabile come percentuale di palline bianche nell'urna. Vi sono due possibilità:

1) Se conoscessimo il numero N di palline nell'urna allora i possibili valori

per Θ sarebbero finiti: $\left\{ \frac{h}{N}, h=1,2,\dots,N \right\}$.

2) Non conoscendo N , i possibili valori per Θ sarebbero tutti i numeri razionali nell'intervallo $[0,1]$.

Supponiamo di non conoscere N .

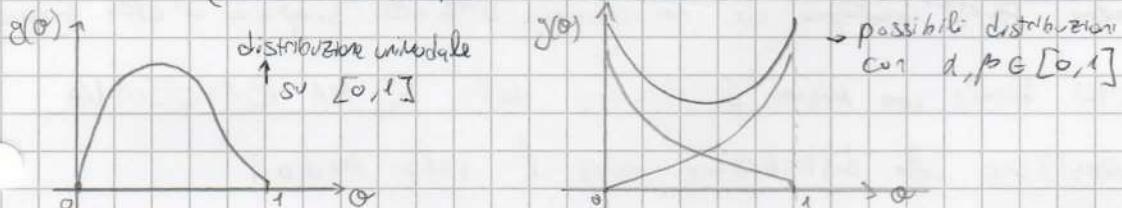
Ipotesi:

(1) $X_i = |E_i| \mid \Theta = \theta \text{ IID } \forall \theta \rightarrow$ I numeri aleatori osservabili e scambiabili sono gli indicatori dell'evento "esce pallina bianca" in ipotesi di estrazione regolare.

(2) $X_i \mid \Theta = \theta \sim Be(\theta)$, ossia $\Pr(|E_i|=x \mid \Theta = \theta) = \theta^x (1-\theta)^{1-x}, x \in \{0,1\}$.

(3) $g_{\Theta}(\theta) \sim Beta(\alpha, \beta)$, ossia: $g_{\Theta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \alpha, \beta \in \mathbb{R}^+$

Osserviamo che α e β , seppur non direttamente riconducibili agli indicatori di evento, sono fissati (scelta non semplice in quanto non intuitiva, ma esistono diversi metodi con diverse affidabilità).



Osservazioni:

- L'utilizzo di un modello statistico è complesso, perciò talvolta le ipotesi di lavoro sono solo provvisorie. Il procedimento è iterativo; se le ipotesi non risultano soddisfacenti, si torna indietro e si modificano.
- Con questo modello abbiamo trovato una terza procedura per individuare la legge temporale iniziale di un processo di alternativa: $W_m^{(n)}, W_n = w_n^{(n)}, f(\theta)$.
- Con la scambiabilità si giudica che qualunque sia il risultato dell'estrazione esso sia equivalente ad una qualsiasi altra estrazione, senza considerare la presenza di eventuali sottogruppi diversi. In tal caso si dovrebbe passare ad uno schema di SCAMBIABILITÀ PARZIALE, in cui le osservazioni sono divise in sottogruppi omogenei interamente.
- Ci chiediamo se sia corretto interpretare la Beta come possibile distribuzione per la proporzione incognita dell'urna, che può assumere solo valori razionali, mentre la Beta considera valori reali tra 0 e 1. Essa è dunque più da interpretare come strumento matematico che consente di esprimere un'opinione probabilistica sui valori razionali di Θ .

Sia $K_n = \left(\sum_{i=1}^n |E_i| = m \right) = (S_n = m)$ l'incremento di informazione osservato.

La corrispondente funzione di verosimiglianza è: $l(\theta | K) = \Pr(S_n = m) = \theta^m (1-\theta)^{n-m}$

Il fattore di proporzionalità (n) rientra se non conosciuto l'ordine di estrazione;

per semplificare la scrittura, assumiamo che l'esperimento sia avvenuto con ordine noto,

dunque: $l(\theta | K) = \Pr(S_n = m) = \theta^m (1-\theta)^{n-m}$

La stima di massima Verosimiglianza è: $\hat{\theta} = \frac{m}{n}$.

Osservazione: Sarebbe un errore sfruttare solo la stima di massima Verosimiglianza trascurando il contesto e lo stato di informazione, in quanto ciò ci renderebbe disposti a credere che la proporzione sia proprio $\frac{m}{n}$. Per esempio, se nel caso di un abile musicista che riconosce qualsiasi brano dalle prime battute (dunque $\theta = 1$) saremmo anche propensi a pensare che realmente $\theta = 1$, nel caso di un ubriaco che azzecca tutti i lanci di una moneta (quindi sempre $\frac{m}{n} = 1$), saremmo molto meno propensi a pensare che realmente $\theta = 1$.

Sfruttiamo la nuova informazione seguendo l'approccio Indiretto:

$$g_H(\theta | K_n) \propto g_H(\theta) l(\theta | K_n) \propto (\theta^{\alpha-1} (1-\theta)^{\beta-1}) \cdot (\theta^m (1-\theta)^{n-m}) = \underbrace{\theta^{\alpha+m-1} (1-\theta)^{\beta+n-m-1}}_{\text{Nucleo della Beta iniziale}} \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{Costante}} \underbrace{\theta^{\alpha+\beta-1}}_{\text{Nucleo di una Beta}} (\theta^{\alpha}, \theta^{\beta})$$

Osserviamo che anche in questo caso, partendo da un'Alpha, siamo giunti ad un'altra Beta con parametri aggiornati. La scelta della distribuzione iniziale non è casuale, bensì è voluta per ottenere quest'analogia, che noi avremmo ottenuto scegliendo un'altra famiglia.

Ricaviamo dalla densità finale una misura di sintesi, detta STIMA BAYESIANA, ossia un valore di posizione della distribuzione, come il valor medio:

$$E(\theta | H \wedge K_n) = \frac{\alpha'}{\alpha' + \beta'} = \frac{\alpha + m}{\alpha + \beta + n} = \frac{(\alpha + \beta) \frac{\alpha}{\alpha + \beta}}{\alpha + \beta + n} + \frac{n \frac{m}{n}}{\alpha + \beta + n}$$

Ottieniamo una mistura tra il valor medio della distribuzione iniziale $\frac{\alpha}{\alpha + \beta}$ e la stima di massima Verosimiglianza per l'estrazione $\frac{m}{n}$. Al crescere del numero delle estrazioni la stima assume un peso sempre maggiore a scapito dell'opinione iniziale, che non può però mai essere trascurata.

È pur vero che nella realtà la situazione in cui il numero di esperimenti è aumentabile a volontà è molto rara.

Osservazione: Per il teorema di Bayes risulta: $P_H(E|K_n) = \frac{P_H(E) P(K_n|E)}{P_H(E) P(K_n|E) + P_H(\bar{E}) P(\bar{K}_n|\bar{E})}$

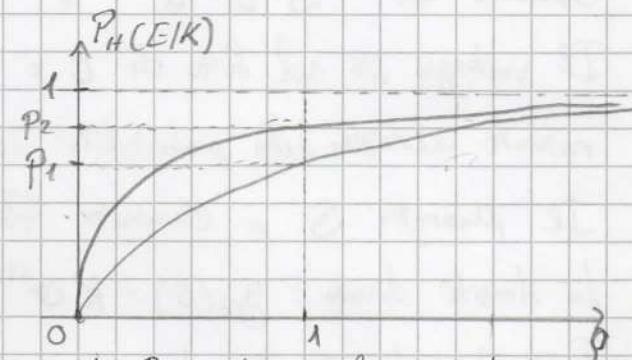
Dividendo per $P(K_n|E)$ otteniamo: $= \frac{P_H(E)}{P_H(E) + [1 - P_H(E)]} \frac{P(K_n|E)}{P(K_n|E)}$ Indichiamolo $= \frac{1}{\varphi}$

In questo modo $\varphi = \frac{P(K_n|E)}{P(K_n|\bar{E})}$ è il rapporto di verosimiglianza.

Sia inoltre $P_H(E) = p \Rightarrow \varphi = \frac{p}{p + (1-p) \cdot \frac{1}{\varphi}} = \psi_p(\varphi) \rightarrow$ Funzione di: - p : valutazione iniziale
- φ : rapporto di verosimiglianza

Vediamo che qualunque sia p attribuito inizialmente, $P_H(E|K)$ converge a 1, ma non uniformemente rispetto a p . Ciò vuol dire che se fissiamo un intorno di 1 esistono infiniti valori per p in corrispondenza ai quali lo scarto tra la probabilità iniziale e quella finale è $> \epsilon$. Quindi a rigore l'informazione iniziale non è mai trascurabile (dal punto di vista concettuale), anche quando l'evidenza campionaria è estremamente grande.

Dunque anche in presenza di un'informazione campionaria elevata esistono sempre opinioni iniziali p tali che la differenza tra la funzione e il limite 1 è sempre $> \epsilon$. $\rightarrow H \geq 0$ in $[p^*, \infty)$ risulta: $|P_H(E|p) - 1| > \epsilon$.



\rightarrow Curve diverse col variaz di p

Esempio: $K = E_1 \wedge E_2 \wedge \bar{E}_3 = (S_3=2)$

$$P_H(K|O) = \int_0^1 [\varphi^2(1-\varphi)] g_H(\varphi) d\varphi = \frac{\Gamma(\alpha+M)}{\Gamma(\alpha)\Gamma(M)} \int_0^1 \varphi^{(\alpha+2)-1} (1-\varphi)^{(M+1)-1} d\varphi \rightarrow$$

Nucleo di una Beta($\alpha+2, M+1$)

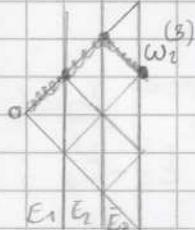
$$\Rightarrow \text{Appross Indiretto: } g_H(O|K) \propto g_H(O) P(K|O) \propto [\varphi^{(\alpha-1)}(1-\varphi)^{M-1}] \varphi^2(1-\varphi) = \varphi^{(\alpha+2)-1} (1-\varphi)^{(M+1)-1}$$

$$\Rightarrow \frac{\Gamma(\alpha+M)}{\Gamma(\alpha)\Gamma(M)} \int_0^1 \varphi^{(\alpha+2)-1} (1-\varphi)^{(M+1)-1} d\varphi = \frac{\Gamma(\alpha+M)}{\Gamma(\alpha)\Gamma(M)} \cdot \frac{\Gamma(\alpha+2)\Gamma(M+1)}{\Gamma(\alpha+2+\alpha+M+1)} = \frac{\Gamma(\alpha) \cdot \alpha(\alpha+1) \Gamma(M)}{\Gamma(\alpha+M)(\alpha+M+1)(\alpha+M+2)} =$$

$$= \frac{\alpha(\alpha+1) \beta}{(\alpha+M)(\alpha+M+1)(\alpha+M+2)} \in \Lambda(H)$$

Consideriamo il processo residuo non assento

Calcoliamo $P_H(K_n | E_1 \wedge E_2 \wedge \bar{E}_3)$



Dobbiamo modificare in maniera coerente le probabilità assegnate a tutti i nodi successivi, per esempio daremo all'evento $S_n=n$ probabilità $w_{n,n}=0$ (mentre prima era positiva).

Sia φ la probabilità di tutte le traiettorie che non partono con la traiettoria osservata.

$\underbrace{F'_4 \wedge F'_5 \wedge \dots \wedge F'_{n+3}}_{\text{ordine fissato (per semplicità)}} = (S_n=m) \rightarrow$ Valore logico fissato

$$F'_4 \wedge F'_5 \wedge \dots \wedge F'_{n+3} = (S_n=m)$$

\rightarrow ordine fissato (per semplicità)

$$(1) \text{ Indiretto: } P_H(K_n | E_1 \wedge E_2 \wedge \bar{E}_3) = \int_0^1 \varphi^{(\alpha-1)} (1-\varphi)^{M-1} g_H(\varphi | E_1 \wedge E_2 \wedge \bar{E}_3) d\varphi$$

$$(2) \text{ Diretto: } P_H(K_n | E_1 \wedge E_2 \wedge \bar{E}_3) = \frac{P_H(E_1 \wedge E_2 \wedge \bar{E}_3 \wedge F'_4 \wedge \dots \wedge F'_{n+3})}{P_H(E_1 \wedge E_2 \wedge \bar{E}_3)}$$

I due approcci sono equivalenti in quanto abbiano assunto che il processo è scambiabile e resta tale anche dopo l'osservazione.

cenno su analisi di robustezza

Una stima si dice Robusta se, cambiando il modello, non varia o varia di poco.

Poiché (α, β) sono solo indirettamente collegati con le variabili osservabili, potremmo pensare di modificarli per avvicinarli ad esse.

Consideriamo allora la seguente riparametrizzazione: $t = \frac{\alpha}{\alpha+\beta}$, $S = \alpha + \beta$

$$\text{Otteniamo che: } S \cdot t = \alpha \text{ e } S - \alpha = \beta \Leftrightarrow \beta = S - St = S(1-t)$$

Il vantaggio sta nel fatto che t è il valor medio della Beta e dunque rappresenta la naturale valutazione della probabilità dell'evento osservabile, infatti: $E_H(\theta) = \int_0^1 \theta g_H(\theta) d\theta = \frac{\alpha}{\alpha+\beta} = t$.

Il parametro S è chiamato PARAMETRO DI APPRENDIMENTO.

La densità d'una: $g_H(\theta) = K \theta^{S-1} (1-\theta)^{S(1-t)-1}$ con parametri scelti come: $-t \in (t_0', t_0'')$

In questo modo scegliere t sarà più semplice.

Sarebbe conveniente associare ad una distribuzione di probabilità un indice che suggerisca la quantità di informazione che ciascun punto alla valutazione (per esempio: affermiamo 0,5 al lancio di una moneta perché non abbiamo informazioni o potremo ne abbiamo un numero altissimo con probabilità 0,9) : la stessa probabilità può derivare da stati di informazione diversi.

La differenza $t''_0 - t'_0$ si riduce a 0 quando abbiamo così tanta informazione da poter scegliere un unico valore per t . Il massimo valore è 1.

Possiamo dunque interpretare $t''_0 - t'_0$ come un indice della quantità di informazione che porta alla scelta del modello.

Supponiamo di fare un esperimento su M eventi, dei quali r affermati. Dopo l'applicazione del teorema di Bayes si giunge ad una distribuzione finale $g_H(\theta|K_M)$ con i seguenti parametri: $t'_M = \frac{S_0 t_0' + r}{S_0 + M}$, $t''_M = \frac{S_0 t_0'' + r}{S_0 + M}$ perche' $\frac{S_0}{S_0+M} < 1$

L'indice di informazione si è in questo modo aggiornato in $t''_M - t'_M < t''_0 - t'_0$, il che implica che l'informazione è aumentata.

Da ciò si capisce perché S è detto Parametro di Apprendimento: Se fosse $S=0$ avremmo: $t''_M - t'_M = \frac{t''_0 - t'_0}{2}$; dunque esso indica il numero di osservazioni che permette di dimezzare l'indice di informazione iniziale. Concretamente se si sa molto poco, S dovrebbe essere molto elevato.]

• SCHEMA NORMALE - NORMALE

Consideriamo il problema di misurazione della temperatura di liquefazione dell'elio (prossima allo 0 assoluto, ossia -279°C). Ad ogni misurazione si può andare incontro a:

1) Errori SISTEMATICI: La grandezza di interesse Θ è sempre maggiorata o è sempre minorata (per esempio per uno strumento tarato male). Supponiamo che non ci siano.

2) Errori ACCIDENTALI: Sono dovuti ad una molteplicità di cause minori singolarmente trascurabili che però, se unite, possono dar luogo ad errori consistenti. Una buona rappresentazione di tali errori è una Normale con media 0 e varianza il reciproco della precisione dello strumento di misura (assunta nota). Nella teoria degli Errori questi sono assunti IID. In ogni misurazione sono presenti gli errori accidentali, che possono essere ridotti in alcuni momenti, ma sono ineliminabili.

Un possibile schema è il seguente: Siano Θ la temperatura di liquefazione e X_n le variabili osservabili corrispondenti alle misurazioni. In quanto tali, esse saranno la somma tra il vero valore e uno scarto accidentale:

$$X_n = \Theta + U_n, \text{ dove } U_n \sim N(0, \sigma_u^2) \text{ IID.}$$

Assumiamo inoltre che:

- $\Theta \perp\!\!\!\perp U_n \forall n$
- $\Theta | H \sim N(\mu_0, \sigma_0^2)$

Di conseguenza:

- $X_n | H \sim N(\mu_0, \sigma_u^2 + \sigma_0^2)$ in quanto somma di normali indipendenti

$$\begin{aligned} - \text{Cov}_n(X_n, X_m) &= E(X_n X_m) - E(X_n)E(X_m) = E[(\Theta + U_n)(\Theta + U_m)] - \mu_0^2 = E[\Theta^2 + \Theta(U_n + U_m) + U_n U_m] \\ &= \sigma_0^2 + \mu_0^2 + 0 + E(U_n U_m) + \mu_0^2 = \begin{cases} \sigma_0^2 & \text{se } m \neq n \\ \sigma_0^2 + \sigma_u^2 & \text{se } m = n \end{cases} \rightarrow \text{Covarianza dipende solo da } |m-n| \end{aligned}$$

\Rightarrow Il processo X_n è scambiatore.

Perciò: $f_H(x) \sim N^{(n)}(\mu_0, \sigma_0^2 + \sigma_u^2, \sigma_0^2) \in \Lambda(H)$

Con una misurazione $K_n = \lambda_{i=1}^n (X_i = x_i)$ possiamo ottenere la legge a posteriori per le variabili non ancora osservate ξ_1, \dots, ξ_m :

$$P_m(\xi_1, \dots, \xi_m | H \wedge K_n) = \frac{f_H(x_1, \dots, x_n, \xi_1, \dots, \xi_m)}{f_H(x_1, \dots, x_n)} \in \Lambda(H \wedge K_n)$$

Seguendo questo modello i calcoli per ottenere la legge a posteriori sono molto complessi, per cui esiste un analogo modello statistico equivalente che li semplifica:

Ipotesi:

- (1) $X_n | \Theta = \theta$ IID $\rightarrow X_n$ scambiabili e correlati fra loro (per l'apprendimento dall'esperienza)
- (2) $X_n | \Theta = \theta \sim N(\theta, \sigma_u^2)$, ossia $f(x|\theta) = (2\pi\sigma_u^2)^{-1/2} \exp\left\{-\frac{(x-\theta)^2}{2\sigma_u^2}\right\}$
- (3) $\theta | \Theta = \theta \sim N(\mu_0, \sigma_0^2)$, ossia $f_{\Theta}(\theta) = (2\pi\sigma_0^2)^{-1/2} \exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\}$

A questa terza di scelte corrisponde una legge temporale iniziale $\Lambda(H)$ a cui appartenere:

$$f_H(\xi_1, \dots, \xi_n) = \int_R f(\xi|\theta) g_{\Theta}(\theta) d\theta \sim N^{(n)} \left(\mu_0 \cdot 1_n, \underbrace{\begin{pmatrix} \sigma_0^2 + \sigma_u^2 & \sigma_0^2 & \dots & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma_u^2 & \dots & \sigma_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_0^2 & \sigma_0^2 & \dots & \sigma_0^2 + \sigma_u^2 \end{pmatrix}}_{\sum} \right)$$

Supponiamo di osservare $K_n = \hat{\Lambda}(X_i = x_i)$ con valori diversi a causa degli scatti accidentali (che non sono osservabili; se lo fossero conosciamo con certezza la grandezza incognita).

⇒ Procedimento Indiretto: $f_{\Theta}(O|K_n) \propto f_{\Theta}(O) f(O|K_n) \propto \exp\left\{-\frac{(O-\mu_0)^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\frac{n(O-\bar{x})^2}{2\sigma_u^2}\right\}$

$= \exp\left\{-\frac{(O-\mu_0)^2}{2\sigma_0^2} + \left[-n \frac{(O-\bar{x})^2}{2\sigma_u^2}\right]\right\}$ → All'esponente comparono due forme quadratiche in O . Sfruttando alcuni risultati sulle forme quadratiche ottenute alla fine:

$\sim N\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma_u^2}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma_u^2}}\right)$ → La distribuzione finale è ancora gaussiana con parametri aggiornati dalla statistica sufficiente (n, \bar{x}) , che contiene tutte le informazioni essenziali sull'estrazione.

Osserviamo che $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ è invariante per permutazioni.

- Osservazioni:
- Il valore medio finale è una combinazione lineare convessa tra il valore medio iniziale μ_0 e la media campionaria \bar{x} , con pesi tali che $\frac{1}{\sigma_0^2}$ varia con il numero delle misurazioni, mentre $\frac{n}{\sigma_u^2}$ cresce con n : più misurazioni sono effettuate e più costa σ_u^2 l'esperienza a posteriori.
 - La varianza della distribuzione finale è una funzione monotona decrescente con n , se assumiamo tale varianza come indice di informazione contenuta nella distribuzione, questa aumenta in modo monotono solo in base al numero di osservazioni, indipendentemente da quali valori assumono. (quando di solito la varianza dipende anche dai valori)

Sia $f(x, \Theta)$ la distribuzione congiunta di tipo normale. Il teorema delle probabilità composte permette di scrivere: $f(\Theta) f(x|\Theta) = f(x, \Theta) = f(x) f(\Theta|x)$, dove la densità marginale $f(x)$ rappresenta una generica distribuzione di $\Lambda(H)$, mentre $f(\Theta|x)$ corrisponde alla distribuzione finale del parametro, ossia $\in \Lambda(H|K_n)$.

Multivariate Normal Distributions

The nondegenerate r -variate normal p.d.f. is

$$f_N^r(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-r/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right]$$

where \mathbf{x} and $\boldsymbol{\mu}$ are $r \times 1$ vectors and Σ is an $r \times r$ symmetric positive definite matrix. The mean and variance are

$$E(\mathbf{x}) = \boldsymbol{\mu}$$

$$V(\mathbf{x}) = \Sigma.$$

Partition \mathbf{x} , $\boldsymbol{\mu}$, and Σ conformably:

$$\mathbf{x}' = (\mathbf{x}'_I \mathbf{x}'_J), \boldsymbol{\mu} = (\boldsymbol{\mu}'_I \boldsymbol{\mu}'_J) \quad \text{and}$$

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix}$$

Σ_{II}, Σ_{JJ} sono matrice di dispersione
 Σ_{IJ}, Σ_{JI} sono solo matrice di covarianza

with \mathbf{x}_I having q elements and \mathbf{x}_J having $r - q$ elements. The marginal and conditional distributions are

$$f(\mathbf{x}_I) = f_N^q(\mathbf{x}_I | \boldsymbol{\mu}_I, \Sigma_{II})$$

$$f(\mathbf{x}_I | \mathbf{x}_J) = f_N^q(\mathbf{x}_I | \boldsymbol{\mu}_I + \Sigma_{II}\Sigma_{JJ}^{-1}(\mathbf{x}_J - \boldsymbol{\mu}_J); \Sigma_{II} - \Sigma_{II}\Sigma_{JJ}^{-1}\Sigma_{JI}).$$

Notation. $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$ stands for “ \mathbf{x} is normally distributed with mean $\boldsymbol{\mu}$ and variance Σ .”

$$\underline{\Sigma} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\theta} \end{bmatrix} \quad \begin{array}{l} \text{Variabili osservabili} \\ \text{Parametri non noti} \end{array}$$

$$f(\underline{x} | \underline{\theta}) \sim N(F\underline{\theta}; V)$$

$$g(\underline{\theta}) \sim N(\underline{\mu}; \Sigma)$$

$$\downarrow$$

$$\underline{\Sigma} \sim N\left(\begin{bmatrix} F\underline{\mu} \\ \underline{\mu} \end{bmatrix}; \begin{bmatrix} F\Sigma F' + V & F\Sigma \\ \Sigma F' & \Sigma \end{bmatrix}\right)$$

$$\Downarrow$$

$$\begin{array}{l} \text{Distribuzione congiunta} \\ \text{Marginal } \mathbf{x} \in \Lambda(H) \end{array}$$

$$\mathbf{x} \sim N(F\underline{\mu}; F\Sigma F' + V)$$

$$\underline{\theta} | \mathbf{x} \sim N\left(\underline{\mu} + \Sigma F'(F\Sigma F' + V)^{-1}(\mathbf{x} - F\underline{\mu}); \Sigma - \Sigma F'(F\Sigma F' + V)^{-1}F\Sigma\right)$$

- Siano:
- F la matrice nota nel problema, definita come vettore unitario: $F = \underline{1}_n$
 - V la matrice di varianza e covarianza della distribuzione: $V = \delta_u^2 \cdot I_n$
- Di conseguenza risulta:
- $\mathcal{L}(\underline{x}|@) \sim N^{(n)}\left(\underline{1}_n \mu_0, \delta_u^2 I_n\right)$ condizionale
 - $g_H(@) \sim N(\mu_0, \delta_u^2)$

Da questa coppia otteniamo la densità congiunta per i due elementi

$$\begin{bmatrix} \underline{X} \\ @ \end{bmatrix} \sim N^{(n+1)}\left(\begin{bmatrix} \underline{1}_n \mu_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \underline{1}_n \delta_u^2 \underline{1}_n^T + \delta_u^2 I_n & \underline{1}_n \delta_u^2 \\ \delta_u^2 \underline{1}_n & \delta_u^2 \end{bmatrix}\right)$$

\underline{X} rappresenta il valore medio dello stato di informazione iniziale, ma anche il valore medio di tutte le variabili osservabili.

$$\text{Con } \underline{1}_n \delta_u^2 \underline{1}_n^T + \delta_u^2 I_n = \begin{bmatrix} \delta_u^2 & \delta_u^2 & \dots & \delta_u^2 \\ \delta_u^2 & \delta_u^2 & \dots & -\delta_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \delta_u^2 & \delta_u^2 & \dots & \delta_u^2 \end{bmatrix} + \begin{bmatrix} \delta_u^2 & 0 & \dots & 0 \\ 0 & \delta_u^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \delta_u^2 \end{bmatrix} = \begin{bmatrix} \delta_u^2 + \delta_u^2 & \delta_u^2 & \dots & \delta_u^2 \\ \delta_u^2 & \delta_u^2 + \delta_u^2 & \dots & -\delta_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \delta_u^2 & \delta_u^2 & \dots & \delta_u^2 + \delta_u^2 \end{bmatrix}$$

Cioè implica che le variabili sono scambiabili con tutte le varianze uguali e tutte le covarianze uguali.

Di conseguenza la marginale: $\underline{X} \sim N^{(n)}\left(\underline{1}_n \mu_0, \delta_u^2 (\underline{1}_n \underline{1}_n^T) + \delta_u^2 I_n\right)$

La distribuzione del parametro condizionata al nuovo stato di informazione è:

$$\textcircled{(1)} | \underline{X} \sim N\left(\underbrace{\mu_0 + \delta_u^2 \underline{1}_n^T (\underline{1}_n \delta_u^2 \underline{1}_n^T + \delta_u^2 I_n)^{-1} (\underline{X} - \underline{1}_n \mu_0)}_{E(\textcircled{(1)} | \underline{X} = \underline{x})}, \Sigma \sim \Sigma F (F \Sigma F^T + V)^{-1} F \Sigma\right)$$

Esempio: Se $n=2$: $E(\textcircled{(1)} | \underline{X} = \underline{x}) = \frac{\mu_0 \delta_u^2 + 2 \bar{x} \delta_u^2}{\delta_u^2 + 2 \delta_u^2} = \frac{\frac{\mu_0}{\delta_u^2} + \frac{2 \bar{x}}{\delta_u^2}}{\frac{1}{\delta_u^2} + \frac{2}{\delta_u^2}}$

\rightarrow Combinazione lineare convessa tra μ_0 e \bar{x} .

• SCHEMA POISSON-COMPOSTO

Il processo di Poisson-Composto è un particolare processo di Lévy (Markoviano con incrementi indipendenti e omogenei) in cui intervengono due processi stocastici:

- Processo degli arrivi : $N_j, j \geq 1$ (poissoniano) $\Rightarrow X(t) = \sum_{j=1}^{N(t)} Y_j$: Processo di Poisson-Composto
- Processo degli Effetti o Pesi : Y_j (con $Y_0 = 0$)

Il processo ha come funzione di ripartizione: $F_{X_t}(x) = \Pr(X_t \leq x) = \sum_{n=0}^{+\infty} \Pr(N_t = n) \Pr(Y_t \leq x | N_t = n)$

Inoltre: $\Phi_X(\theta) = E(X(\theta)) = \sum_{\lambda \in \mathbb{N}} \lambda^{\theta} E(N_{\theta}) E(Y_{\lambda})$

Sia Θ_1 l'intensità non completamente nota del processo di arrivi. Per semplificare discuteremo le osservazioni considerando una sequenza di periodi unitari, in ciascuno dei quali osserviamo il numero di eventi e la somma degli effetti, insieme anche al numero totale di intervalli unitari di osservazione.

Poiché siamo in presenza di due processi stocastici, con due distribuzioni diverse, consideriamo due parametri non noti: Θ_1 per $N(t)$ e Θ_2 per Y_j .

Assumiamo allora che:

- $N(t) | \Theta_1 = \theta_1 \sim N Po(\theta_1)$, ossia: $\Pr(N(t) = n_1 | \Theta_1 = \theta_1) = e^{-\theta_1} \frac{\theta_1^{n_1}}{n_1!}$ con
- $Y_j | \Theta_2 = \theta_2 \sim Esp(\theta_2)$, ossia $f(y) = \theta_2 e^{-\theta_2 y}$, IID condizionatamente a Θ_2

Sia K l'incremento di informazione costituito da:

- Numero M di periodi di osservazione unitari
- Sequenza n_1, \dots, n_M di numeri di arrivi negli M intervalli
- Sequenza y_1, \dots, y_M di somme degli effetti in ciascun intervallo ($y_1 = \sum_{j=1}^{n_1} Y_{1j}, \dots, y_M = \sum_{j=1}^{n_M} Y_{Mj}$)

Allora, se subordinatamente ad ogni ipotesi sui due parametri assumiamo l'indipendenza degli incrementi (scambiabilità), allora risulterà:

$$l(\theta_1, \theta_2 | K) = l_1(\theta_1 | K) l_2(\theta_2 | K) = \Pr(K | \Theta_1 = \theta_1) \Pr(K | \Theta_2 = \theta_2) \alpha(\theta_1, \theta_2) \exp \left[-\theta_1 \sum_{j=1}^M n_j - \theta_2 \sum_{j=1}^M y_{1j} \right]$$

In questo modo l'informazione usata è $K = (m, \sum_{j=1}^M n_j, \sum_{j=1}^M y_{1j})$ in cui gli effetti intervengono solo come somma.

Così procedendo, i parametri vengono stimati separatamente.

Indicando con $S_h = \sum_{j=1}^{n_h} Y_{1j}$ e con $S = \sum_{h=1}^m S_h$ e con $N = \sum_{h=1}^m n_h$, otteniamo per $K = \prod_{j=1}^m K_j$: $l_H(\theta_1, \theta_2 | K) = \prod_{h=1}^m l_h(\theta_1, \theta_2 | K_h) \alpha(\theta_1 | \theta_2) (\theta_1 e^{-\theta_1 N} / \theta_2^{S_h}) = (\theta_1 \theta_2)^{-S} e^{-\theta_1 N - \theta_2 S}$

da cui si ricavano le stime di Massima Verosimiglianza: $\hat{\theta}_1 = \frac{N}{m}$, $\hat{\theta}_2 = \frac{S}{\hat{\theta}_1}$ 68

Il problema è che concretamente fra il processo degli arrivi e quello degli effetti potrebbe esserci una correlazione (per esempio tra i danni denunciati ed il numero complessivo di denunce). Di conseguenza il prodotto $l_1(O_1/K) \cdot l_2(O_2/K)$ perde di attendibilità. Ci troviamo invece in presenza di due processi scambiabili tra loro correlati, ovvero di SCAMBIABILITÀ PARZIALE.

In presenza di una correlazione possiamo ricavare informazioni utili sul primo processo anche solo osservando il secondo processo.

La distribuzione iniziale scelta per i parametri, dunque, dovrà introdurre una correlazione, e dovrà risultare che $g_{14}(O_1, O_2) \neq g_{14}(O_1) g_{14}(O_2)$ (condizione sufficiente per la scambiabilità generale).

Se non abbiamo elementi per introdurre una qualche dipendenza stocastica fra i due parametri, allora possiamo supporre che: $g_{14}(O_1, O_2) = g_{14}(O_1) g_{14}(O_2)$.

In tal caso, assumendo che:

- $g_{14}(O_1) \sim \text{Gamma}(a, c)$
- $g_{14}(O_2) \sim \text{Gamma}(b, d)$ (parametri fissati)

risulterebbe che: $g_{14}(O_1, O_2) \propto (O_1^{a-1} e^{-cO_1})(O_2^{b-1} e^{-dO_2})$

Allora, sfruttando l'informazione K , otterremo:

$$g_{14}(O_1, O_2 | K) \propto (O_1^{a+n-1} e^{-c(O_1+m)}) (O_2^{b+n-1} e^{-d(O_2+S)})$$

con stime Bayesiane: $\hat{O}_1 = \frac{\hat{a}+n}{\hat{c}+m}$, $\hat{O}_2 = \frac{\hat{b}+n}{\hat{d}+S}$

che, al crescere dell'informazione disponibile (n, m) convergono stocasticamente rispettivamente a \hat{O}_1 e a \hat{O}_2 .

La scambiabilità nasce con de Finetti al congresso dei matematici a Bologna nel 1928, mentre la Scambiabilità Parziale viene presentata dallo stesso de Finetti in un convegno sulla probabilità a Ginevra nel 1937.]

Una Scambiabilità Parziale l'analogia, seppur non presente a livello globale, è presente ad un livello inferiore in sottosistemi differenti di variabili. Questo giustifica le correlazioni tra esse.

SCAMBIABILITÀ PARZIALE

SCAMBIABILITÀ

- Fa intervenire un unico processo stocastico
- Si assume quando si ritiene che tra le variabili osservabili ci sia una qualche analogia informativa, intesa come IID condizionata
- $g_H(O_1, O_2) = g_H(O_1) g_H(O_2)$

• Invarianza per permutazioni finite, ossia:

$$F_{1,2,\dots,n}(x) = F_{i_1, \dots, i_n}(x) = F_n(x)$$

dove i₁, ..., i_n sono diversi fra loro

Teorema di Rappresentazione

- Indicatori di evento:

$$X_i = |E_i| \Rightarrow w_n = \int_0^1 x^n F(dx)$$

$$P(E) = \int_0^1 P_x(E) F(dx)$$

↳ probabilità di E nell'ipotesi che gli indicatori siano bernulliani

- Numeri aleatori qualsiasi:

$$F_{1,\dots,n}(x) = \int_{\Omega} \left[\prod_{i=1}^n F(x_i) \right] G(dw)$$

Forma funzionale
fissa da cui dipendono le distribuzioni

Esempio: Assumendo G una Beta bivariata, ossia: $\text{Beta}(\nu_1, \nu_2, \nu_3) \rightarrow f(x, y) = \frac{\Gamma(\nu_1 + \nu_2 + \nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2)\Gamma(\nu_3)} x^{\nu_1-1} y^{\nu_2-1} z^{\nu_3-1}$

$$\text{otteniamo: } P\left(\bigcap_{i=1}^n E_i\right) = \int_0^1 \int_0^1 x^n y^n z^n \text{Beta}(dx, dy)$$

$$= \frac{(\nu_1)_{n_1} (\nu_2)_{n_2}}{(\nu_1 + \nu_2 + \nu_3)_{n_1+n_2}}$$

Analogia con la

cambiabilità semplice, in cui: $P\left(\bigcap_{i=1}^n E_i\right) = w_n = \int_0^1 x^{n_1} [k x^{\nu_1} (1-x)^{\nu_2}]^{n_2} dx = \frac{(n_1)_n}{(n_1 + \nu_2)_n}$

SCAMBIABILITÀ PARZIALE

• Fa intervenire 2 o più processi stocastici

• Parte dal presupposto che si riesca ad individuare tra le variabili osservabili dei sottosistemi diversi fra loro, un analoghi interamente

$$g_H(O_1, O_2) \neq g_H(O_1) g_H(O_2)$$

• Invarianza per permutazioni finite all'interno di ciascun sottosistema di variabili, ossia:

$$F_{1,\dots,n_1,\dots,n_k}(x, y) = F_{i_1,\dots,i_k,j_1,\dots,j_n}(x, y) = F_{n,k}(x, y)$$

per $x \in \mathbb{R}^k$ e $y \in \mathbb{R}^n$

(Il discorso può essere generalizzato con N processi stocastici)

Teorema di Rappresentazione

- Indicatori di evento

$$X_i = |E_i| \Rightarrow w_n = \int_0^1 \int_0^1 x^n y^n F(dx, dy)$$

N_1 eventi del I tipo affermati e introduce
 N_2 eventi del II tipo affermati ora dipendenti
tra X e Y

$$P(E) = \int_0^1 \int_0^1 P_{x,y}(E) F(dx, dy)$$

↳ Probabilità di E nell'ipotesi
che X e Y siano bernulliani indipendenti

\Rightarrow Indipendenza Stocastica Generalizzata: indipendenza
tra gli eventi del I tipo, tra quelli del II tipo e
tra quelli del I e II tipo.

- Numeri aleatori qualsiasi:

$$F_{1,\dots,n_1,\dots,n_k}(x, y) = \int_{\Omega} \left[\prod_{i=1}^k F(x_i) \right] \prod_{j=1}^n F(y_j) G(dw, dz)$$

comuni distribuzioni condizionate

Se G fosse il prodotto tra $G(dw)$ e $G(dz)$ il
processo sarebbe composto da due processi scambiabili
indipendenti (condizione sufficiente), ma in questo caso
sono correlati.

Le variabili sono condizionalmente indipendenti.

Marginali: $f_x(x) \propto \text{Beta}(\nu_1, \nu_2 + \nu_3)$ dipendenti
 $f_y(y) \propto \text{Beta}(\nu_2, \nu_1 + \nu_3)$

\rightarrow È nulla nell'estremo
del triangolo $(0, 1, 1)$, dunque
non permette densità positiva se
due valori prossimi a 1.

È utilizzato quando sono supposti valori
piccoli sia per x che per y .

MODELLO INFERENZIALE

Sia $\underline{\Omega} = (\underline{\theta}_1, \underline{\theta}_2)$. Ipotesi: (1) $X_i = |E_i| \mid \underline{\Omega} = \underline{\theta}$ IID $\forall i$ con $P(E_i \mid \underline{\theta}) \equiv \theta_1$, (2) $Y_j = |F_j| \mid \underline{\Omega} = \underline{\theta}$ IID $\forall j$ con $P(F_j \mid \underline{\theta}) \equiv \theta_2$, (3) $X_i \perp\!\!\!\perp Y_j$ $\forall i, j$ ipotesi $\underline{\Omega} = \underline{\theta}$ (4) $g_{\#}(\theta_1, \theta_2) \sim \text{Beta}(v_1, v_2, v_3)$

Siano: $K_1 = \prod_{i=1}^{n_1} E_i$ e $K_2 = \prod_{j=1}^{n_2} F_j$ gli incrementi di informazione rispettivamente fissati.
 → valori logici fissati e noti (po' evitare di inserire il coefficiente binomiale)

Supponiamo che: $K_1 \rightarrow S_{n_1} = s_1$, $K_2 \rightarrow S_{n_2} = s_2$

La corrispondente verosimiglianza è:

$$l(\theta_1, \theta_2 \mid K_1, K_2) \stackrel{D}{=} P(K_1 \mid \underline{\theta}) P(K_2 \mid \underline{\theta}) = \left[\theta_1^{s_1} (1-\theta_1)^{n_1-s_1} \right] \left[\theta_2^{s_2} (1-\theta_2)^{n_2-s_2} \right] \xrightarrow{\text{I.I.D sotto l'ipotesi } \underline{\theta} = \underline{\theta}} \hat{\theta}_1 = \frac{s_1}{n_1}, \hat{\theta}_2 = \frac{s_2}{n_2}$$

Osserviamo che non c'è ancora intervenuto alcun legame tra i due parametri incogniti.

Andiamo ora a calcolare la legge finale:

$$g_{\#}(\theta_1, \theta_2 \mid K_1, K_2) d\theta_1 d\theta_2 = \theta_1^{v_1-1} \theta_2^{v_2-1} (1-\theta_1-\theta_2)^{v_3-1} \left(\theta_1^{s_1} (1-\theta_1)^{n_1-s_1} \theta_2^{s_2} (1-\theta_2)^{n_2-s_2} \right) \xrightarrow{\text{Non c'è più una Beta bivariata}}$$

Se avessimo solo questa espressione la forma funzionale sarebbe analogia a quella iniziale con parametri aggiornati dalle frequenze di successo

Fattori di disturbo che scomparirebbero se le frequenze di successo dei due sottosampioni fossero pari alle numerosità (tutti affermati)

Strutturando la formula del Binomio di Newton possiamo riscrivere l'espressione:

$$= \sum_{i=0}^{n_1-s_1} \sum_{j=0}^{n_2-s_2} (-1)^{i+j} \binom{n_1-s_1}{i} \binom{n_2-s_2}{j} \theta_1^i \theta_2^j (1-\theta_1-\theta_2)^{v_3-1}$$

Otteniamo una combinazione lineare convessa di Beta bivariate.

Dunque la famiglia di distribuzioni che rimane invariata nell'applicazione del teorema di Bayes non è la famiglia delle Beta bivariate (Anite), bensì la famiglia di tutte le combinazioni lineari convesse di Beta bivariate (Anite): partendo da una beta bivariata (caso particolare di mistura di se stessa con peso unitario) si è giunti ad un'altra mistura di beta bivariate.

Esempio: Supponiamo di aver fatto un certo numero di misurazioni della temperatura di liquefazione dell'olio e che vi sia qualche differenza in presenza di un fattore esterno. Allora, anziché assumere un solo numero aleatorio per la temperatura, assumiamo che ce ne siano due: uno in presenza del fattore esterno, e uno in sua assenza.

Ipotesi: $\forall \underline{\omega} = \underline{\Omega}$: (1) X_i IID con $f_1(x_i | \underline{\Omega}) \sim N(\Omega_1, 1)$ \rightarrow senza fattore esterno

(2) Y_j IID con $f_2(y_j | \underline{\Omega}) \sim N(\Omega_2, 1)$ \rightarrow con fattore esterno

(3) $X_i \perp\!\!\!\perp Y_j$

(4) $g_{12}(\Omega_1, \Omega_2) \sim N^{(2)}(\underline{\mu}, \Sigma)$ con $\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \gamma & \gamma \\ \gamma & \gamma_2 \end{bmatrix}$ fissati

Effettuiamo l'esperimento: $K_1 = \sum_{i=1}^n (X_i - \bar{x}_i)$, $K_2 = \sum_{j=1}^m (Y_j - \bar{y}_j)$

La corrispondente funzione di verosimiglianza è:

$$l(\Omega_1, \Omega_2 | K_1, K_2) = f_1(\bar{x} | \Omega_1) f_2(\bar{y} | \Omega_2) = \prod_{i=1}^n f_1(x_i | \Omega_1) \prod_{j=1}^m f_2(y_j | \Omega_2) \propto e^{-\frac{n_1(\Omega_1 - \bar{x})^2}{2}} e^{-\frac{n_2(\Omega_2 - \bar{y})^2}{2}}$$

Scambiabilità

Da cui ottieniamo: $\hat{\Omega}_1 = \bar{x}$, $\hat{\Omega}_2 = \bar{y}$ \rightarrow medie aritmetiche, ciascuna invariante per permutazioni

Legge Temporale Iniziale: $f_{12}(x_i, y_j) = \int_R \int_R f(x_i, y_j | \underline{\Omega}) g_{12}(\Omega_1, \Omega_2) d\Omega_1 d\Omega_2 \sim N^{(2)}\left(\underline{\mu}, \begin{bmatrix} 1 + \gamma_1 & \gamma \\ \gamma & 1 + \gamma_2 \end{bmatrix}\right)$

Possiamo scrivere: $\begin{bmatrix} \bar{x} + \gamma \\ \bar{y} + \gamma \end{bmatrix} \in \Lambda(H)$

stessa covarianza
dei parametri

1 è la
varianza delle
distribuzioni iniziali

Il modello ha 7 parametri: $\mu_1, \mu_2, \gamma, \gamma_1, \gamma_2, \bar{x}, \bar{y}$ e le covarianze interne condizionate tra \underline{X} e \underline{Y} , che mancano per ora. Per trovarle consideriamo le generiche coppe:

$$\downarrow \begin{aligned} f_1^{(H)}(x_i, x_j) &= \int_R \int_R f_1(x_i, x_j | \underline{\Omega}) g_{12}(\Omega_1, \Omega_2) d\Omega_1 d\Omega_2 = \int_R \underbrace{f_1(x_i, x_j | \Omega_1)}_{\text{Intervale } \Omega_1} \underbrace{g_{12}(\Omega_1)}_{N^{(2)}(\mu_1, 1)} d\Omega_1 \sim N^{(2)}\left(\begin{bmatrix} \bar{x} + \gamma \\ \bar{x} + \gamma \end{bmatrix}, \begin{bmatrix} 1 + \gamma_1 & \gamma \\ \gamma & 1 + \gamma_2 \end{bmatrix}\right) \end{aligned}$$

Ottieniamo che nello stato H

$$\text{Cov}(X_i, X_j) = \gamma_1$$

Lo stesso procedimento con $f_2^{(H)}(y_i, y_j)$ porta a $\text{Cov}(Y_i, Y_j) = \gamma_2$.

Ottieniamo la Legge Temporale Finale:

(Per indipendenza se i misurati
ha varianza 1, allora na misuraz.
hanno varianza n)

$$\text{Riscriviamo la verosimiglianza: } l(\Omega_1, \Omega_2 | K_1, K_2) \propto e^{-\frac{n_1(\Omega_1 - \bar{x})^2}{2}} e^{-\frac{n_2(\Omega_2 - \bar{y})^2}{2}} = \exp \left\{ -\frac{1}{2} (\underline{\Omega} - \underline{\mu})^T D (\underline{\Omega} - \underline{\mu}) \right\}$$

$\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \leftarrow \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$

where D è l'inversa della matrice di dispersione $\begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix}$.

$$\Rightarrow g_{12}(\Omega_1, \Omega_2 | K_1, K_2) \propto g_{12}(\Omega_1, \Omega_2) l(\Omega_1, \Omega_2 | K_1, K_2) \propto \exp \left\{ -\frac{1}{2} (\underline{\Omega} - \underline{\mu}^*)^T (D^{-1} + D) (\underline{\Omega} - \underline{\mu}^*) \right\}$$

dove $\underline{\mu}^*$ è il vettore medio finale ottenuto

combinando il vettore medio iniziale $\underline{\mu}$ con il vettore delle Medie aritmetiche $\underline{\bar{x}}$:

$$\underline{\mu}^* = (D^{-1} + D)^{-1} (D^{-1} \underline{\mu} + D \underline{\bar{x}}) \quad \text{con} \quad D = (D^{-1} + D)^{-1} \text{ matrice di dispersione finale.}$$

PROCESSI DEL II ORDINE

Indichiamo con il simbolo $\mathcal{L}_2(\Omega, \mathbb{F}, P)$ l'insieme di tutte le variabili aleatorie definite sullo spazio (Ω, \mathbb{F}, P) con momento secondo finito.

Tratta di uno spazio lineare, in quanto chiuso per combinazioni lineari.

Sia $\mathcal{H} = \mathcal{L}_2(\Omega, \mathbb{F}, P)$ lo spazio detto SPAZIO DI HILBERT.

Sia \mathcal{H}_0 lo Spazio di Hilbert caratterizzato da variabili aventi valor medio nullo, ossia $\mathcal{H}_0 = \{X(\omega) \mid E(X) = 0, E(X^2) < +\infty\}$.

Si dice Spazio di Hilbert uno spazio lineare di dimensione infinita completo

Così tale che ogni successione di Cauchy di elementi dello spazio converge a un elemento dello stesso. → in media quadratica

Uno Spazio di Hilbert può essere caratterizzato dalla funzione PRODOTTO INTERNO (o SCALARE)

che associa ad ogni coppia un numero reale: $(X, Y) \rightarrow \langle X, Y \rangle \in \mathbb{R} = E(XY) = \text{Cov}(X, Y)$:

1) Lineare: $\langle \alpha X_1 + \beta X_2, Y \rangle = \alpha \langle X_1, Y \rangle + \beta \langle X_2, Y \rangle \quad \forall X_1, X_2, Y \in \mathcal{H}_0$

2) Semidefinita positiva: $\langle X, Y \rangle \geq 0 \quad \forall X, Y \in \mathcal{H}_0$

2.1) $\langle X, X \rangle = 0 \iff \Pr(X=0) = 1$

3) Simmetrica: $\langle X, Y \rangle = \langle Y, X \rangle$

→ L'introduzione di questa funzione rende lo spazio \mathcal{H}_0 Euclideo, ossia prendono significato nozioni come distanza, ortogonalità e angolo tra elementi

Attraverso questa funzione possiamo definire la NORMA o LUNGHEZZA di $X \in \mathcal{H}_0$:

$$\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{E(X^2)} = \sqrt{3_X} \quad \text{con le seguenti proprietà:}$$

1) $\|X\| \geq 0$

2) $\|\lambda X\| = |\lambda| \|X\|, \lambda \in \mathbb{R}$

3) Disegualanza triangolare: $\|X+Y\| \leq \|X\| + \|Y\|$

$$\Rightarrow \text{Distanza tra due elementi di } \mathcal{H}_0: d(X, Y) = \|X - Y\| = \sqrt{\text{Var}(X-Y)} = \sqrt{3_{X-Y}}$$

$$\Rightarrow \text{Angolo tra due elementi di } \mathcal{H}_0: \cos[\alpha(X, Y)] = \frac{\langle X, Y \rangle}{\|X\| \|Y\|} = \frac{\text{Cov}(X, Y)}{\sqrt{3_X} \sqrt{3_Y}} \rightarrow \begin{array}{l} \text{ORTOGONALITÀ} \\ \Leftrightarrow \langle X, Y \rangle = 0 \\ \text{INDICE DI CORRELAZIONE} \\ \Leftrightarrow \text{Correlazione lineare} \\ \Leftrightarrow \alpha = \frac{\pi}{2} \end{array}$$

L'introduzione della nozione di distanza permette di parlare di convergenza in \mathcal{H}_0 :

$X_1, X_2, \dots \rightarrow X \Leftrightarrow d(X_i, X) \xrightarrow{i \rightarrow \infty} 0$ (convergenza in norma o in distanza). Questa convergenza

implica la convergenza in media quadratica ($\sqrt{E(X_i - X)^2}$), utile per le valutazioni.

I processi del II ordine sono ampiamente utilizzati in quanto richiedono poco rispetto a modelli più complessi, che invece richiedono la specificazione di intere distribuzioni.

Naturalmente, tale semplificazione limita ciò che è possibile ricavare dal modello. 73

Sia $Y \in \mathcal{H}_0$ non osservabile e siano $X_1, \dots, X_n \in \mathcal{H}_0$ osservabili.

L'obiettivo è costruire un'approssimazione per Y in termini di numeri aleatori osservabili, detti in questo caso APPROXIMATORI \Rightarrow Problema di INDUZIONE STATISTICA

In particolare considereremo un'approssimazione lineare, stimando Y mediante un'opportuna combinazione lineare degli osservabili, costruita con il metodo dei Minimi Quadrati, che consiste nel minimizzare il valore medio del quadrato dell'errore di approssimazione (aleatorio).

• Caso $n=1$ \hookrightarrow Teoria del Filtraggio Ottimale [anni '50]

Consideriamo una sola osservazione $X \in \mathcal{H}_0$. Definiamo approssimatore dei minimi quadrati

\hat{Y} per Y la quantità tale che: $E[(Y - \hat{Y})^2] \leq E[(Y - kx)^2] \Rightarrow \min_K E[(Y - kx)^2]$

(Ipotizzano che errori positivi abbiano lo stesso peso di errori negativi, cosa non scontata).

$$\Rightarrow \Phi(K) = E(Y^2) - 2KE(XY) + K^2E(X^2) \Rightarrow \Phi(K) = -2E(XY) + 2K^2E(X^2) = 0 \Leftrightarrow K^* = \frac{E(XY)}{E(X^2)} = \frac{\text{cov}(XY)}{\text{Var}(X)}$$

Minimizzando questa quantità abbiamo cercato l'elemento dello spazio lineare KX

(una retta) a minima distanza dal vettore Y .

Abbiamo dunque un numero aleatorio che gode di due proprietà:

$$1) \hat{Y} \in L(X) \Leftrightarrow \exists h | \hat{Y} = hX$$

$$2) Y - \hat{Y} \perp L(X) : \text{L'errore è ortogonale a tutti gli elementi dello spazio lineare generato da } X$$

$$\Rightarrow \text{Cov}(Y - \hat{Y}, X) = 0 \Leftrightarrow 0 = E[(Y - \hat{Y})X] = E[(Y - hX)X] = E(XY) - hE(X^2) \Leftrightarrow h = \frac{E(XY)}{E(X^2)} = K^*$$

Queste due condizioni definiscono l'operatore PROIEttORE ORTOGONALE sullo spazio lineare generato da X . Esso soddisfa 3 condizioni fondamentali:

$$1) \text{È un operatore lineare: } P(\alpha Y_1 + \beta Y_2 | X) = \alpha P(Y_1 | X) + \beta P(Y_2 | X)$$

$$2) \text{Idempotente: } P(P(Y | X) | X) = P(Y | X) \quad \hookrightarrow \text{per brevità, al posto di } L(X)$$

$$3) \text{Autoaggiunto: } \langle P(X | Z), Y \rangle = \langle X, P(Y | Z) \rangle$$

L'interpretazione come proiezione ortogonale prende il nome di Principio di Ortogonalità: il punto del sottospazio lineare a minima distanza da Y coincide con la sua proiezione ortogonale sul sottospazio.

Esempi di approssimazioni lineari: • $Y_n = X + \epsilon_n$ con $\epsilon_n \sim WN(0, \sigma^2_\epsilon)$

• $Y_n = a Y_{n-1} + \epsilon_n$ con $\epsilon_n \sim WN(0, \sigma^2_\epsilon)$ ($X \perp \epsilon_n, E(X) = 0, V(X) = \sigma^2_X$)

• $Y_n = \epsilon_n + b \epsilon_{n-1} \sim NARMA(1, 1)$

Consideriamo ora un approssimatore non lineare, ossia una funzione $f(X)$ scelta all'interno dello spazio funzionale in cui $E[f(X)]$ è soluzione di: $\min_{f \in \mathcal{F}} E[(Y - f(X))^2]$. Risulta che la funzione che minimizza tali scarti è la FUNZIONE DI REGRESSIONE $E(Y|X) = \hat{Y} : E[(Y - E(Y|X))^2] \leq E[(Y - f(X))^2]$. Il problema è che, mentre per ottenere l'approssimatore lineare basta conoscere solo i primi due momenti della distribuzione di X , per quello non lineare serve conoscere tutta la distribuzione di X , conoscenza molto spesso non disponibile, per cui talvolta ci si accontenta del primo. I due stimatori coincidono quando la distribuzione condizionata è assunta gaussiana o di Student.

Osservazione: L'utilizzo di un'osservazione statistica ci ha permesso di ottenere una previsione migliore di quella che avremmo dato nello stato di informazione iniziale, ossia il vero valore medio di Y (assunto nullo). Se X e Y non fossero correlati, infatti, il prevedere sarebbe nullo, ovunque pari a $E(Y)$. Se il processo di osservazione è scambabile, ogni singola osservazione gioca il medesimo ruolo delle altre.

• Caso $n > 1$:

Sia $Y \in \mathbb{H}_0$ non osservabile e siano $X_1, \dots, X_n \in \mathbb{H}_0$ osservabili e correlati con Y .

Cerchiamo l'approssimatore lineare soluzione di: $\min_{\underline{\alpha} \in \mathbb{R}^n} E[(Y - \sum_{j=1}^n \alpha_j X_j)^2]$

Ponendo le derivate parziali uguali a 0 ottieniamo:

$$\text{Cov}(\underline{X}) \cdot \underline{\alpha} = E(Y \cdot \underline{X}) \Leftrightarrow \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \vdots & \ddots & & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \text{Var}(X_n) \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \text{Cov}(Y, X_1) \\ \vdots \\ \text{Cov}(Y, X_n) \end{bmatrix}$$

↳ Matrice di dispersione di \underline{X}
(semidefinita positiva)

Se la matrice di dispersione è invertibile, ossia se le X_i sono tra loro linearmente indipendenti ($\text{Cov}(X_i, X_j) = 0 \forall i \neq j$), allora ottieniamo:

$$\hat{\underline{\alpha}} = \text{Cov}(\underline{X})^{-1} E(Y \cdot \underline{X}), \text{ per cui il previsor lineare ottimale è:}$$

$$\hat{Y} = \sum_{j=1}^n \hat{\alpha}_j X_j = \underline{X}^T \cdot \text{Cov}(\underline{X})^{-1} E(Y \cdot \underline{X}) = E(Y \cdot \underline{X}^T) \text{Cov}(\underline{X})^{-1} \underline{X}$$

La soluzione esiste solo se $E(\underline{X} \cdot \underline{X}^T) \neq 0$, ossia se Y è correlato con \underline{X} ,

e se la matrice $\text{Cov}(\underline{X})$ è diagonale ($\text{Cov}(\underline{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$). L'inversa di una matrice diagonale è ancora diagonale con elementi reciproci: $\text{Cov}(\underline{X})^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{bmatrix}$

Alla base vi è l'assunzione che lo stato di informazione iniziale ci consente di valutare ogni varianza e ogni covarianza, aggiornate dall'incremento di informazione costituito da \underline{X} , seppur non esista un algoritmo che consenta di aggiornarle correntemente.

Anche in questo caso \hat{Y} può essere interpretato come proiezione ortogonale di Y sullo spazio lineare $L(\underline{X})$ generato dal vettore \underline{X} .

$$L(\underline{x}_1, \dots, \underline{x}_n) = \left\{ \sum_{j=1}^n \alpha_j \underline{x}_j \right\} \triangleq L_n$$

Possiamo anche in questo caso definire l'operatore PROETTORE ORTOGONALE

$P(Y|L_n) : H_0 \rightarrow L_n$ con le seguenti proprietà:

- Idempotenza: $P[P(Y|L_n)|L_n] = P(Y|L_n)$
- Linearità: $P(\alpha Y_1 + \beta Y_2 | L_n) = \alpha P(Y_1 | L_n) + \beta P(Y_2 | L_n)$
- Autoaggiunzione: $\langle P(Y_1 | L_n), Y_2 \rangle = \langle Y_1, P(Y_2 | L_n) \rangle$

Inoltre, essendo $L_{n-1} \subseteq L_n$ (spazio lineare generato dai primi $n-1$ numeri aleatori osservabili):

- $P(Y | L_{n-1}) = P[P(Y | L_{n-1}) | L_n] = P[P(Y | L_n) | L_{n-1}] \rightarrow$ prevale lo spazio di dimensione inferiore
- $P(Y | X_1, \dots, X_n) = \sum_{j=1}^n P(Y | X_j)$ se $X_i \perp X_j (i \neq j)$ (infatti: $\hat{Y} = \underline{X} \begin{bmatrix} E(Y|X_1) \\ \vdots \\ E(Y|X_n) \end{bmatrix} = \sum_{j=1}^n E(Y|X_j) X_j = \sum_{j=1}^n P(Y|X_j) X_j$)
- Proprietà ITERATIVA: $\hat{Y}_n = P(Y | L_n) = P(Y | L_{n-1}) + P[Y | X_n - P(X_n | L_{n-1})]$

→ Ottieniamo lo stesso risultato aggiornando $P(Y | L_{n-1})$ con la proiezione su $X_n - P(X_n | L_{n-1})$, che è detta INNOVAZIONE di X_n e corrisponde alla novità informativa che interviene con la possibilità d'osservare X_n . La componente $P(X_n | L_{n-1})$ è interpretabile come ciò che di X_n già si poteva dire sulla base dei primi $n-1$ numeri aleatori. L'informazione di X_n può infatti essere scomposta in una parte già acquisita con l'osservabilità di X_1, \dots, X_{n-1} e una parte tipica di X_n che non poteva essere prevista dalle precedenti osservazioni.

Tale proprietà è sfruttata nel procedimento di GRAHAM-SMITH, in cui X_1 e X_2 possono essere sostituiti con due vettori ortogonali fra loro X_1 e $\hat{X}_2 = \text{differenza fra } X_2 \text{ e la sua proiezione sullo spazio lineare generato da } X_1$ (ossia l'innovazione di X_2).

Se è presente un ulteriore vettore X_3 , questo sarà sostituito con la differenza fra X_3 e la sua proiezione sullo spazio generato da X_1 e X_2 .

Questa proprietà del proiettore ortogonale consente un notevole risparmio al livello computazionale: invece di invertire l'intera matrice di dimensione n basta aggiungere l'innovazione alle $n-1$.

Osservazione: Se interpretiamo l'insieme dei momenti secondi come stato di informazione sappiamo che questi non hanno ancora un algoritmo per essere aggiornati, nonostante i tentativi in passato, ad esempio:

$$\cdot \underline{X} = \begin{bmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_n \end{bmatrix} \sim N(\underline{\mu}, \Sigma) = N\left(\begin{bmatrix} \underline{\mu}_1 \\ \vdots \\ \underline{\mu}_n \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix}\right) \Rightarrow f(\underline{X}_1 | \underline{X}_2) \sim N\left(\underline{\mu}_1, \Sigma_{11}^{-1}\right)$$

Dove $f(\underline{X}_1 | \underline{X}_2)$ rappresenta l'aggiornamento dell'informazione su \underline{X}_1 avendo osservato solo \underline{X}_2 .

Prima di tale osservazione lo stato di informazione valeva \underline{X}_1 con $f(\underline{X}_1) \sim N(\underline{\mu}_1, \Sigma_{11})$.

Ma questo è potuto avvenire perché in precedenza avevamo già una conoscenza completa della distribuzione di \underline{X}_1 , e questo ci ha permesso di aggiornarne anche i momenti.

La gaussiana (come la t di Student) ha una proprietà che permette di interpretare i nuovi momenti come approssimazioni degli aggiornamenti.

Non è però possibile, dal punto di vista bayesiano, aggiornare i soli momenti secondi a meno che non d'apartiranno in ipotesi di gaussianità.

- Un altro tentativo di aggiornare solo i momenti secondi fu dato dall'autore Michael Goldstein negli anni '80 nella cosiddetta Statistica Bayesiana Lineare, in cui tentò di dare un adeguamento bayesiano solo per i momenti di primo e secondo ordine. L'impostazione non è esatta, ma mette in punto dei programmi operativi utili.

MODELLO LINEARE DINAMICO

Megli anni '60 R.E. Kalman mise a punto un procedimento di stima iterativo che permetteva un notevole salto di qualità nelle previsioni della posizione di un satellite date alcune posizioni occupate in alcuni istanti antecedenti. Tale procedimento è lineare. Al suo passo in tempo non lineare avvennero negli anni '90 con le equazioni differenziali stocastiche sul procedimento non lineare del Filtro di Kalman. Il modello prevede una componente dinamica che riguarda soltanto le osservazioni statistiche che possono crescere nel tempo, mentre la variabile di interesse resta sempre la stessa. Il problema diventa concretamente dinamico quando il dinamismo riguarda già l'acquisizione di nuove informazioni che la variabile di interesse, che viene assunta verosimilmente variabile nel tempo.

Il procedimento di stima di Kalman si basa su un algoritmo chiamato **FILTRA** di KALMAN, in quanto in quegli anni cominciarono ad apparire i primi operatori elettronici, che permettevano di svolgere agevolmente procedimenti iterativi.

Consideriamo un processo stocastico non osservabile composto da **VARIABILI DI STATO** Y_n di cui conosciamo soltanto la regola evolutiva nel tempo.

Ciò che invece possiamo osservare è il processo X_n , linearmente legato a Y_n , e in numero crescente nel tempo, per approssimare Y_n .

CASO 1: SPECIFICAZIONE DEL II ORDINE

→ AR(1)

Sia $Y_{n+1} = a Y_n + V_n$ con $V_n \sim WN(0, \sigma_v^2)$, a, σ_v^2 noti (ipotesi non necessaria in quanto si può imporre su essi) assumiamo che lo stato di informazione ci consenta di specificare soltanto i momenti primi e secondi del processo Y (soluzione dell'equazione alle differenze finite con coefficienti costanti). Di conseguenza non potremo ottenere di più dei momenti I e II.

Per risolvere l'equazione imponiamo la condizione iniziale $E[Y_0] = 0$, ottenendo:

$$Y_{n+1} = a^{n+1} Y_0 + \sum_{h=0}^n a^h V_{n+1-h} \rightarrow \text{Processo stocastico caratterizzato dai seguenti momenti:}$$

$$\cdot \Psi_Y(n) \equiv 0$$

$$\cdot \Psi_Y(n, n+h) = a^h \sigma_v^2 \frac{(1-a)^{2(n+h)}}{(1-a^2)}$$

Se $|a| < 1$ asintoticamente la funzione di covarianza dipende solo da h (differenza tra i deponenti) e ciò implica **STAZIONARITÀ ASINTOTICA**, ossia quasi-stazionarità per n sufficientemente grande.

Assumiamo che $Y_0 \perp V_n$.

Consideriamo il processo osservabile $X_n = bY_n + W_n$ con $W_n \sim WN(0, \sigma_n^2)$ e b, σ_n^2 noti.

Assumiamo che $W_n \perp V_n$. X rappresenta una trasformazione lineare affine stocastica (perde' subendo un processo di rumore) di $Y \Rightarrow X_n = b[a^n Y_0 + \sum_{h=0}^{n-1} a^h V_{n-h}] + W_n$

L'incremento di informazione è costituito dall'osservazione statistica di X correlato a Y .

Sia $K_{n+1} = \bigwedge_{j=1}^{n+1} (X_j = x_j)$ l'incremento di informazione che consente di ottenere $\hat{Y}_{n+1|n+1}$, ossia una previsione (stimatore) per Y_{n+1} ottenuta da $n+1$ osservazioni di X .

In generale, disponendo delle osservazioni X_1, \dots, X_{n+1} , lo stimatore potrà essere ottenuto come proiezione ortogonale di Y su \mathcal{L}_{n+1} , spazio lineare generato da X_1, \dots, X_{n+1} .

Possiamo allora prendere in considerazione:

- $\hat{Y}_{n+1|n+1} = \mathbb{P}(Y_{n+1}|d_{n+1}) = X_{n+1} \text{Cov}(X_{n+1})^{-1} E(Y_{n+1}|X_{n+1}) \rightarrow$ Problema di **FILTRAGGIO**
- $\hat{Y}_{n+1|n+1} = \mathbb{P}(Y_{n+1}|d_{n+1}) \rightarrow$ Problema di **PREVISIONE**: cerciamo di dire qualcosa sul futuro Y_{n+1} usando $n+1$ osservazioni
- $\hat{Y}_{n+1|n+1} = \mathbb{P}(Y_n|d_{n+1}) \rightarrow$ Problema di **PERCONEZIONE**. L'**ESITAMENTO**: si utilizza una informazione esubiente per stimare il passato Y_n ("post mortem")

FILTO DI KALMAN: È un algoritmo che, sfruttando la proprietà iterativa del proiettore ortogonale, consente di stimare Y . Per farlo si analisi delle seguenti equazioni:

$$(1) \hat{Y}_{n+1|n} = a \hat{Y}_{n|n}$$

$$(2) P_{n+1|n} = E[(Y_{n+1} - \hat{Y}_{n+1|n})^2] = a^2 P_{n|n} + \sigma_w^2$$

Dimostrazione \hookrightarrow valore medio del quadrato dell'errore di approssimazione, ossia la sua varianza (dato che è egual)

Sia $\mathcal{L}_{n+1} \subseteq \mathcal{L}_n \subseteq \mathcal{H}_0$ e sia $X_n^* = X_n - \mathbb{P}(X_n|d_{n+1})$ l'innovazione di X_n

Proprietà iterativa di \mathbb{P} : $\mathbb{P}(Y|X_1, \dots, X_n, X_n) = \mathbb{P}(Y|X_1, \dots, X_n, X_n^*) = \mathbb{P}(Y|d_{n+1}) + \mathbb{P}(Y|X_n^*)$

$$(1) \hat{Y}_{n+1|n} = \mathbb{P}(Y_{n+1}|d_n) = \mathbb{P}(aY_n + V_n|d_n) \stackrel{\text{lineari}}{=} a\mathbb{P}(Y_n|d_n) + \mathbb{P}(V_n|d_n) = a\hat{Y}_{n|n} \quad \begin{matrix} \text{P}_{n|n} \\ \text{P}_{n+1|n} \\ \sigma_w^2 \end{matrix}$$

$$(2) P_{n+1|n} = E[(Y_{n+1} - \hat{Y}_{n+1|n})^2] = E[(aV_n + V_n - a\hat{Y}_{n|n})^2] = E[(a(V_n - \hat{Y}_{n|n}) + V_n)^2] = a^2 E[(V_n - \hat{Y}_{n|n})^2] + E(V_n^2) \quad \begin{matrix} \text{E}(V_n - \hat{Y}_{n|n})V_n = 0 \end{matrix}$$

$$(3) \hat{Y}_{n+1|n+1} = \mathbb{P}(Y_{n+1}|d_{n+1}) = \hat{Y}_{n+1|n} + \mathbb{P}(Y_{n+1}|X_{n+1} - X_{n+1|n}) = \hat{Y}_{n+1|n} + \frac{\text{Cov}(Y_{n+1}, X_{n+1}^*)}{\text{Var}(X_{n+1}^*)} \cdot X_{n+1}^* \quad \begin{matrix} \text{Proiezione su} \\ \text{spazio} \\ \text{univariato} \end{matrix}$$

Essendo:

$$\bullet X_{n+1}^* = X_{n+1} - \mathbb{P}(X_{n+1}|d_n) = X_{n+1} - \mathbb{P}(bY_{n+1} + W_{n+1}|d_n) = bY_{n+1} + W_{n+1} - b\hat{Y}_{n+1|n} - \underbrace{\mathbb{P}(W_{n+1}|d_n)}_{=0} = b(Y_{n+1} - \hat{Y}_{n+1|n}) + W_{n+1}$$

$$\bullet \text{Var}(X_{n+1}^*) = b^2 \text{Var}(Y_{n+1} - \hat{Y}_{n+1|n}) + \text{Var}(W_{n+1}) = b^2 P_{n+1|n} + \sigma_w^2$$

$$\bullet \text{Cov}(Y_{n+1}, X_{n+1}^*) = E(Y_{n+1} X_{n+1}^*) = E(Y_{n+1} [b(Y_{n+1} - \hat{Y}_{n+1|n}) + W_{n+1}]) = b P_{n+1|n}$$

$$\hookrightarrow E[Y_{n+1} b(Y_{n+1} - \hat{Y}_{n+1|n})] = 0$$

$$\text{Per la proprietà della covarianza possiamo sostituire } Y_{n+1} \text{ con } Y_{n+1} - \hat{Y}_{n+1|n}$$

$$= E[(Y_{n+1} - \hat{Y}_{n+1|n})(Y_{n+1} - \hat{Y}_{n+1|n})b] = P_{n+1|n} \cdot b$$

Fasi del Filtro di Kalman: \rightarrow L'algoritmo parte da X_0 e Y_0 .

(1) Fase di PREVISIONE: Consente di passare dalla coppia $\hat{Y}_{n|n}, P_{n|n}$ a $\hat{Y}_{n+1|n}, P_{n+1|n}$.

Durante questa fase lo stato di informazione non è cambiato (X_0, \dots, X_n) e si sfruttano le equazioni (1) e (2) per prevedere il futuro Y_{n+1} . Per questo motivo esse sono anche dette Equazioni di Previsione.

(2) Fase di AGGIORNAMENTO o INDUTTIVA: Sfrutta le equazioni (3) e (4) per ottenere $\hat{Y}_{n+1|n+1}$ e $P_{n+1|n+1}$ a partire dalla nuova osservazione X_{n+1} .

Esse sono perciò dette Equazioni di Aggiornamento.

Osserviamo che le varianze non dipendono dai valori osservati, dunque la sequenza delle P può essere pre-calcolata prima di fare qualunque osservazione.

Questa è la fase in cui interviene la proprietà iterativa del proiettore ortogonale:

Intuitivamente è il prevedere per Y_{n+1} basato su tutte le nostre variabili osservabili disponibili e costruito sulla previsione data con le sole prime n ($\hat{Y}_{n|n|n}$) più un fattore correttivo e l'innovazione di X_{n+1} , ossia il contenuto informativo presente nell'osservazione $n+1$ -esima che non poteva essere previsto linearmente sulla base delle precedenti n osservazioni. (è un'informazione tipica di X_{n+1} che non hanno X_0, \dots, X_n).

Questa fase non è altro che il procedimento di Graham-Smith su numeri aleatori.

Osservazioni:

- Il modello può essere complicato a volontà senza che le conclusioni siano concettuali:
 - Il coefficiente a_{ij} anziché essere costante potrebbe essere variabile nel tempo
 - L'equazione potrebbe essere vettoriale, anziché scalare
 - Potrebbero esserci ulteriori variabili
- I processi di disturbo V e W sono non osservabili, dunque "inquinano" le nostre variabili, che vanno perciò "filtrate" dai disturbi o perturbazioni, che vanno neutralizzati.
- In questo modello ci sono due componenti di dinamicità nel tempo:
 - Attività delle variabili osservabili
 - Evoluzione dell'oggetto di studio
- Il modello appena analizzato è a tempo discreto. Per l'analisi a tempo continuo esiste un analogo algoritmo, detto Filtro di Kalman-Bucy [1961]

CASO 2: SPECIFICAZIONE COMPLETA: Modello Lineare Gaussiano

Assumiamo che lo stato di informazione iniziale consenta di attribuire ai due processi di rotore V_n e W la distribuzione gaussiana. L'ipotesi non è troppo forte, in quanto la Teoria degli Errori assume tale distribuzione per gli scarti accidentali.

Supponiamo che Y_0 sia specificato, per esempio $E(Y_0) = 0$ o $Y_0 = 0$.

Assumiamo di disporre, a partire dall'osservazione di $\underbrace{X_1, \dots, X_n}_{D_n}$, di $\hat{Y}_{n|n}$ e $P_{n|n}$.

Il tragitto del filtro di Kalman diventa in questo caso: $f(V_n | D_n) \rightarrow f(W_n | D_n) \rightarrow f(Y_n | D_n)$

$$\sim N(\hat{Y}_{n|n}, P_{n|n}) \quad \sim N(\hat{Y}_{n+1|n}, P_{n+1|n}) \quad \sim N(\hat{Y}_{n+1|n+1}, P_{n+1|n+1})$$

Si tratta dunque di calcolare la DISTRIBUZIONE PREVISIONALE $f(X_{n+1} | D_n)$: X_{n+1} non è infatti ancora osservato nello stato D_n .

$$X_{n+1} = b Y_{n+1} + W_{n+1} = b(a V_n + V_n) + W_{n+1} \text{ con } V \perp Y_n \text{ e } Y_{n+1} \perp W_{n+1}$$

$\hookrightarrow N(f(Y_n | D_n))$

Poiché la somma di Gaussiane indipendenti è ancora Gaussiana: $X_{n+1} \sim N(b\hat{Y}_{n|n}, b^2 P_{n|n} + \sigma_w^2) = f(X_{n+1} | D_n)$

$$\text{Inoltre } Y_{n+1} \sim N\left(\frac{\hat{Y}_{n|n}}{P_{n|n}}, \frac{\sigma_y^2}{P_{n|n}}\right)$$

Questo consente di determinare la distribuzione condizionata: $X_{n+1} | Y_{n+1}, D_n \sim N(b Y_{n+1}, \sigma_w^2)$

In questo modo possiamo trovare la distribuzione congiunta:

$$\begin{bmatrix} X_{n+1} \\ Y_{n+1} \end{bmatrix} | D_n \sim N^{(2)}\left(\begin{bmatrix} b \hat{Y}_{n|n} \\ \hat{Y}_{n|n} \end{bmatrix}, \begin{bmatrix} b^2 P_{n|n} + \sigma_w^2 & b P_{n|n} \\ b P_{n|n} & P_{n|n} \end{bmatrix}\right)$$

Ottengono infine la distribuzione finale dopo l'osservazione di X_{n+1} che va ad aggiungersi allo stato di informazione D_n :

$$f(Y_{n+1} | \underbrace{X_{n+1}, D_n}_{D_{n+1}}) \sim N\left(\frac{\hat{Y}_{n|n} + b P_{n|n} (b^2 P_{n|n} + \sigma_w^2)^{-1} (X_{n+1} - b \hat{Y}_{n|n})}{P_{n|n} + b^2 P_{n|n} (b^2 P_{n|n} + \sigma_w^2)^{-1}}\right)$$

MODELLO LINEARE DINAMICO GAUSSIANO

Consideriamo il seguente modello nello spazio degli stati:

$$(1) \quad Y_{n+1} = a.Y_n + V_n, \text{ ove } V_n \approx NWN(0; \sigma_V^2) \text{ e ove } V_n \perp Y_m, \forall m \leq n,$$

$$(2) \quad X_n = b.Y_n + W_n, \text{ ove } W_n \approx NWN(0; \sigma_W^2), W_n \perp V_m, \forall n, m \text{ e } W_n \perp Y_m, \forall n, m.$$

Assumeremo inoltre che il n.a. Y_0 abbia una distribuzione Gaussiana con $E(Y_0) = 0$ e $V(Y_0) = \sigma_0^2$ cosicché il processo non osservabile $\{Y_n\}$ e quello osservabile $\{X_n\}$ sono entrambi Gaussiani e le ipotesi di non correlazione diventano ora condizioni di indipendenza stocastica.

In questo modello, il filtro di Kalman consente non soltanto l'adeguamento degli stimatori dei minimi quadrati $\hat{Y}_{n/n}$ e $P_{n/n}$ alle nuove osservazioni, ma addirittura l'adeguamento delle distribuzioni di probabilità delle variabili di stato. Indicando con D_n l'informazione ricavabile dalla osservabilità dei n.a. X_1, X_2, \dots, X_n , ognuna delle fasi del filtro di Kalman realizza le seguenti trasformazioni:

$$f(y_n / D_n) \rightarrow f(y_{n+1} / D_n) \rightarrow f(y_{n+1} / D_{n+1}), \text{ ove}$$

$$f(y_n / D_n) \approx N(\hat{y}_{n/n}, P_{n/n}), \quad f(y_{n+1} / D_n) \approx N(\hat{y}_{n+1/n}, P_{n+1/n}) \text{ e } f(y_{n+1} / D_{n+1}) \approx N(\hat{y}_{n+1/n+1}, P_{n+1/n+1}).$$

Assumendo nota la distribuzione iniziale $f(y_n / D_n) \approx N(\hat{y}_{n/n}, P_{n/n})$ ed utilizzando l'equazione (1) e le ipotesi relative si determina senza problemi la successiva distribuzione previsionale $f(x_{n+1} / D_n) \approx N(\hat{y}_{n+1/n}, P_{n+1/n})$. Per la trasformazione successiva, che porta alla distribuzione finale $f(y_{n+1} / D_{n+1}) \approx N(\hat{y}_{n+1/n+1}, P_{n+1/n+1})$, si possono seguire varie strade; noi utilizzeremo proprietà note delle distribuzioni Gaussiane il che consente un notevole risparmio di elaborazione formale.

L'equazione (2) riferita al periodo $n+1$ consente la facile determinazione della distribuzione condizionale $f(x_{n+1} / y_{n+1}, D_n) \approx N(b.y_{n+1}, \sigma_W^2)$ che, assieme alla distribuzione previsionale

$f(y_{n+1} / D_n) \approx N(\hat{y}_{n+1/n}, P_{n+1/n})$, porta alla distribuzione congiunta

$$f(y_{n+1}, x_{n+1} / D_n) \approx N^{(2)} \left(\begin{bmatrix} \hat{y}_{n+1/n} \\ \hat{y}_{n+1/n} \end{bmatrix}; \begin{bmatrix} P_{n+1/n} & b.P_{n+1/n} \\ b.P_{n+1/n} & b^2.P_{n+1/n} + \sigma_W^2 \end{bmatrix} \right).$$

Per note proprietà generali riguardanti le distribuzioni congiunte Gaussiane, dalla precedente si ha:

$$f(y_{n+1} / x_{n+1}, D_n) = f(y_{n+1} / D_{n+1}) \approx N\left(\hat{y}_{n+1/n+1}, P_{n+1/n+1}\right),$$

con $\hat{y}_{n+1/n+1} = \hat{y}_{n+1/n} + b.P_{n+1/n} \cdot \left(b^2.P_{n+1/n} + \sigma_W^2\right)^{-1} \cdot \left(x_{n+1} - b.\hat{y}_{n+1/n}\right)$

e $P_{n+1/n+1} = P_{n+1/n} - b^2.P_{n+1/n}^2 \left(b^2.P_{n+1/n}^2 + \sigma_W^2\right)^{-1}.$

MODELLI A TEMPO CONTINUO

Con il Filtro di Kalman a tempo discreto si giunge alla seguente equazione alle differenze finite: $\hat{X}_{n+1/n} = (a - abK_n) \hat{Y}_{n/n} + K_n X_{n+1}$, con $K_n = \frac{b(a^2 P_{n/n} + b^2 v)}{b^2(a^2 P_{n/n} + b^2 v) + b^2 u}$

Lo stesso accade nel caso a tempo continuo con il Filtro KALMAN-BUCY, in cui si giunge ad equazioni alle differenze finite e differenziali negli stimatori dei minimi quadrati. In questo caso le equazioni sono differenziali stocastiche.

Se $|a| < 1$ allora la soluzione dell'equazione è asintoticamente un processo stazionario.

CENNO SU EQUAZIONI DIFFERENZIALI ORDINARIE (una sola variabile)

In un'equazione differenziale interviene una qualche derivata della funzione non nota $y(t)$.

Siano $a(t)$ e $f(t)$ (Funzioni di INPUT) funzioni note continue su un intervallo I :

$$y'(t) + a(t)y(t) = f(t) \quad (\text{Equazione Completa})$$

Casi:

- 1) $a(t) \equiv 0 \Rightarrow y'(t) = f(t) \Rightarrow$ soluzione: $y(t) = F(t) + K$
- 2) $a(t) = -1, f(t) \equiv 0 \Rightarrow y'(t) - y(t) = 0 \Rightarrow$ soluzione: $y(t) = K e^{-t}$ primitiva di f \rightarrow costante reale
- 3) $f(t) \equiv 0 \Rightarrow y'(t) + a(t)y(t) = 0 \rightarrow$ È la versione OMogenea dell'equazione di partenza

La sua soluzione è detta SOLUZIONE GENERALE: $y(t) = K e^{-\int a(t) dt}$ primitiva di $a(t)$
ed è composta da infinite funzioni a secondi di K .

Per risolvere un'equazione lineare:

- Si risolve l'equazione omogenea associata ottenendo la Soluzione Generale;
- Si sceglie un valore per K (a seconda dello stato di informazione) assegnando un'ulteriore condizione per l'equazione (per esempio $y(3) = -1$, se si fosse $y(0)$ si parla invece di Condizione Iniziale).

$$\Rightarrow y(t) = e^{-\int a(t) dt} \left[\int_0^t f(\tau) e^{\int_a^t a(\tau) d\tau} d\tau + K \right] = \underbrace{K e^{-\int a(t) dt}}_{\substack{\text{Soluzione Generale} \\ \text{dell'equazione} \\ \text{omogenea}}} + \underbrace{e^{-\int a(t) dt} \int_0^t f(\tau) e^{\int_a^t a(\tau) d\tau} d\tau}_{\substack{\text{Soluzione particolare} \\ \text{dell'equazione completa} \\ (\text{corrispondente a } K=0)}}$$

CENNO SU EQUAZIONI DIFFERENZIALI STOCASTICHE → intervengono processi stocastici

La prima equazione differenziale stocastica nasce nel 1906 con il fisico Paul Langevin per il Moto Browniano.

Prima definizione di Langevin: $\frac{d}{dt} X(t) = -\alpha X(t) + \xi(t), \alpha > 0$

Dove $\xi(t)$ introduce un processo di errore, assunto gaussiano, tale che $E(\xi(t))=0$
 \Rightarrow Processo Stocastico di INPUT

$\text{Var}(\xi(t)) = \sigma^2$

$\text{Cov}(\xi(s), \xi(t)) = 0, s \neq t$

Il problema è che non esiste un processo stocastico gaussiano con numeri indipendenti e varianza finita. In letteratura è detto Processo Stocastico GENERALIZZATO.

Chi risolse questo problema fu K. Ito nel 1940 dando vita al Calcolo Stocastico.

Ito interpretò $\int_0^t \xi(s) ds$ come un processo di Wiener $W(t)$.

(un altro modo per esprimere questa equivalenza è $dW(t) = \xi(t) dt$) \Rightarrow L'incremento del processo inesistente è un incremento del processo di Wiener (da non confondere erroneamente come derivata, in quanto un processo di Wiener non è derivabile).

L'equazione di Langevin viene oggi scritta come:

$dX(t) = -\alpha X(t) dt + \sigma dW(t) \rightarrow$ L'incremento infinitesimo del processo X è proporzionale allo stesso incremento più un errore

Questa non è altro che una versione semplificata della seguente equazione:

$X(t) = X(0) - \alpha \int_0^t X(s) ds + \sigma \int_0^t W(s) ds$

Nasce come velocità in
una direzione di una particella
in Moto Browniano immersa in
un fluido per i continui urti con lo stesso.

\hookrightarrow Esiste una versione più semplice di
calcolo stocastico, detta Calcolo in
Media Quadratica con convergenza, derivabilità,
integrità in media quadratica.
È un calcolo matematicamente povero, ma utile
dal punto di vista pratico.

$\Rightarrow X(t) = X(0) e^{-\alpha t} + \sigma \int_0^t e^{-\alpha(t-s)} dW(s)$

Integrale \int WIENER ITO \rightarrow La funzione integranda è deterministica
mentre l'integratore è un processo di Wiener.

Questo integrale può essere visto come
un numero aleatorio con valore medio nullo e
funzione di covarianza pari a $\langle W(t), s \rangle = e^{-\alpha(t-s)}$ $\left[\sqrt{X(0) + \frac{\sigma^2}{2\alpha} (e^{-2\alpha(t-s)} - 1)} \right]$

Se poniamo come condizione iniziale $X(0) \sim N(0, \frac{\sigma^2}{2\alpha})$ allora il
processo $X(t)$ diventa stazionario e viene chiamato Processo ORNSTEIN-UHLENBECK.

Problemi di stima per variabili latenti: modelli a tempo continuo.

\times non osservabili
 \vee osservabili

Considereremo ora un semplice modello lineare e stocastico costituito dalle equazioni differenziali seguenti:

$$16) \quad dX(t) = -a.X(t)dt + dW(t), \quad X(0) \approx N(0, \sigma_0^2), \quad W(t) \text{ } \underline{\text{non osservabile}} \text{ } ,$$

$$17) \quad dY(t) = X(t)dt + dV(t), \quad Y(0) = 0, \quad V(t) \text{ } \underline{\text{osservabile}} \text{ } ,$$

ove la prima equazione e' gia' nota al lettore (si veda a pag. 3); la seconda afferma che le osservazioni sono costituite dalle variabili di stato alle quali si aggiunge un rumore aleatorio rappresentato da un secondo processo standardizzato di Wiener $V(t)$, indipendente da $W(t)$.

Più precisamente, l'equazione 17) suggerisce l'equivalente equazione $Y(t) = X(t) + V(t)$ ove il secondo addendo e' un disturbo Gaussiano formalmente coincidente con la derivata prima del processo di Wiener $V(t)$. Più corretto e' interpretare l'equazione 17) come l'equazione integrale

$$Y(t) = Y(0) + \int_0^t X(s)ds + V(t) = \int_0^t X(s)ds + V(t)$$

ove l'integrale stocastico va inteso "in media quadratica".

Se $\sigma_0^2 = 1/2a$ si dimostra che le variabili di stato $X(t)$ formano un processo stazionario detto processo di Ornstein – Uhlenbeck e talvolta l'attuale modello con variabili latenti viene denominato in letteratura "osservazioni di un processo di Ornstein – Uhlenbeck affette da errori accidentali di misura".

Filtro di Kalman - Bucy

Un problema di stima analogo a quello posto per il modello precedente a tempo discreto consiste nella individuazione dello stimatore dei minimi quadrati per $X(t)$ basato sull'osservabilità del processo $Y(s)$ nell'intervallo $[0, t]$. Un famoso risultato del 1961, il teorema di Kalman – Bucy, mostra che lo stimatore ottimale $\hat{X}(t)$ e' dato dalla soluzione dell'equazione differenziale stocastica

$$18) \quad d\hat{X}(t) = -a\hat{X}(t)dt + P(t) \left[dY(t) - \hat{X}(t)dt \right], \quad \hat{X}(0) = 0,$$

ove la funzione deterministica $P(t) = E \left[X(t) - \hat{X}(t) \right]^2$, varianza dell'errore di stima, e' data dalla soluzione dell'equazione differenziale non lineare e deterministica, detta equazione di Riccati,

$$19) \quad \dot{P}(t) = 1 - P^2(t) - 2.a.P(t), \quad P(0) = \sigma_0^2 .$$

Per una semplice presentazione e dimostrazione del filtro di Kalman – Bucy suggeriamo al lettore il testo di M.H.A. Davis (paragrafo 4.4).