

Stima dei parametri di una catena di Markov dalle osservazioni

Stima dei parametri

Consideriamo il problema di stimare i *parametri* di un processo di Markov omogeneo, sulla base di *osservazioni di una traiettoria*, ossia la matrice

- ▶ delle probabilità di transizione Q per una catena di Markov $(X_n)_n$

Stima dei parametri

Consideriamo il problema di stimare i *parametri* di un processo di Markov omogeneo, sulla base di osservazioni di una traiettoria, ossia la matrice

- ▶ delle probabilità di transizione Q per una catena di Markov $(X_n)_n$
- ▶ o delle intensità di salto L per un processo di Markov a salti $(X_t)_t$,

+ eventualmente la distribuzione marginale
a un dato tempo

Due approcci

Si osserva che la catena X segue un cammino

$\gamma = (x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n)$, ossia $X_0 = x_0, X_1 = x_1, \dots, X_n = x_n$
(scriviamo $X = \gamma$). Consideriamo i due approcci:

- Massima verosimiglianza: massimizzare

$$Q \mapsto L(Q = Q; X = \gamma) = P(X = \gamma | Q = Q) = P(X_0 = x_0) Q_\gamma$$

Due approcci

Si osserva che la catena X segue un cammino

$\gamma = (x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n)$, ossia $X_0 = x_0, X_1 = x_1, \dots, X_n = x_n$ (scriviamo $X = \gamma$). Consideriamo i due approcci:

- Massima verosimiglianza: massimizzare

$$Q \mapsto L(Q = Q; X = \gamma) = P(X = \gamma | Q = Q) = P(X_0 = x_0) Q_\gamma$$

- Bayesiano: si introduce una densità a priori per Q (vista come variabile aleatoria) e si stima tramite Bayes la densità a posteriori, noto $X = \gamma$,

$$p(Q = Q | X = \gamma) \propto p(Q = Q) L(Q = Q; X = \gamma),$$

Q è v.d. a valori in $\mathbb{R}^{E \times E}$

Stima di massima verosimiglianza

Supponiamo per semplificare che sia $P(X_0 = x_0) = 1$, così

$$L(Q; X = \gamma) = Q_\gamma = \prod_{k=1}^{n-1} Q_{x_{k-1} \rightarrow x_k}.$$

- Raccogliendo i fattori ripetuti,

$$Q_\gamma = \prod_{k=1}^n Q_{x_{k-1} \rightarrow x_k} = \prod_{i,j \in E} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}},$$

osservato (pointing to $\gamma_{i \rightarrow j}$)
parametri (pointing to $Q_{i \rightarrow j}$)

dove $\gamma_{i \rightarrow j}$ è il numero di transizioni dallo stato $i \in E$ a $j \in E$ che avvengono in γ .

Stima di massima verosimiglianza

Supponiamo per semplificare che sia $P(X_0 = x_0) = 1$, così

$$L(Q; X = \gamma) = Q_\gamma = \prod_{k=1}^{n-1} Q_{x_{k-1} \rightarrow x_k}.$$

- Raccogliendo i fattori ripetuti,

$$Q_\gamma = \prod_{k=1}^n Q_{x_{k-1} \rightarrow x_k} = \prod_{i,j \in E} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}},$$

dove $\gamma_{i \rightarrow j}$ è il numero di transizioni dallo stato $i \in E$ a $j \in E$ che avvengono in γ .

- In particolare,

$$n = \sum_{i,j \in E} \gamma_{i \rightarrow j}.$$

Un massimo vincolato

Per calcolare il punto di massimo, di

$$Q \mapsto \prod_{i,j \in E} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}}$$

dobbiamo tenere conto del vincolo che la somma delle righe della matrice Q sia 1.

- Per determinare massimi o minimi di funzioni vincolate in generale si usano i moltiplicatori di Lagrange: nei punti critici il gradiente della funzione sia ortogonale al vincolo.

Una sostituzione

Nel nostro caso evitiamo i moltiplicatori esprimendo la diagonale di \mathcal{Q} in termini delle altre entrate sulla riga:

$$Q_{i \rightarrow i} = 1 - \sum_{j \neq i} Q_{i \rightarrow j} \quad \text{per ogni } i \in E.$$

► Riscriviamo la verosimiglianza

$$L(\mathcal{Q} = Q; X = \gamma) = \prod_{i \in E} \left(1 - \sum_{j \neq i} Q_{i \rightarrow j} \right)^{\gamma_{i \rightarrow i}} \prod_{j \neq i} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}}.$$

$Q_{i \rightarrow i}$

usando

$Q_{i \rightarrow j}$

con i fisso

Una sostituzione

Nel nostro caso evitiamo i moltiplicatori esprimendo la diagonale di \mathcal{Q} in termini delle altre entrate sulla riga:

$$Q_{i \rightarrow i} = 1 - \sum_{j \neq i} Q_{i \rightarrow j} \quad \text{per ogni } i \in E.$$

- Riscriviamo la verosimiglianza

$$L(\mathcal{Q} = Q; X = \gamma) = \prod_{i \in E} (1 - \sum_{j \neq i} Q_{i \rightarrow j})^{\gamma_{i \rightarrow i}} \prod_{j \neq i} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}}.$$

- Possiamo ragionare separatamente per ciascuna riga i , e massimizzare

$$(Q_{i \rightarrow j})_{j \neq i} \mapsto (1 - \sum_{j \neq i} Q_{i \rightarrow j})^{\gamma_{i \rightarrow i}} \prod_{j \neq i} Q_{i \rightarrow j}^{\gamma_{i \rightarrow j}},$$

Simile $(1-p)^{u_B} p^{u_R}$

$$\log\left((1 - \sum_{j \neq i} a_{ij})^{\gamma_{ii}} \prod_{j \neq i} a_{ij}^{\gamma_{ij}} \right) \text{ e derivo rispetto a } \underline{a_{ij}}$$

$$\frac{\partial}{\partial a_{ij}} \log(\quad) = \frac{\partial}{\partial a_{ij}} \left[\gamma_{ii} \log(1 - \sum_{j \neq i} a_{ij}) + \sum_{j \neq i} \gamma_{ij} \log a_{ij} \right]$$

$$= \gamma_{ii} \cdot \frac{1}{1 - \sum_{j \neq i} a_{ij}} \cdot (-1) + \gamma_{ij} \frac{1}{a_{ij}} = 0$$

$$\frac{\gamma_{ij}}{a_{ij}} = \frac{\gamma_{ii}}{1 - \sum_{j \neq i} a_{ij}} \Rightarrow \frac{a_{ij}}{\gamma_{ij}} = \frac{1 - \sum_{j \neq i} a_{ij}}{\gamma_{ii}} \Rightarrow a_{ij} = \gamma_{ij} \cdot \frac{a_{ii}}{1 - \sum_{j \neq i} a_{ij}}$$

Conclusione

Q_{ij} e γ_{ij} (anche se $j=i$)

$\Rightarrow Q$ è la versione "stocastica"

della matrice $\Gamma = (\gamma_{i \rightarrow j})_{i,j \in E}$

Esempio \hookrightarrow $(ON) \rightleftharpoons (OFF) \hookrightarrow \gamma = ON \rightarrow OFF \rightarrow OFF \rightarrow ON \rightarrow ON$
 $\rightarrow OFF \rightarrow ON$

$$\Gamma = \begin{matrix} & \begin{matrix} ON & OFF \end{matrix} \\ \begin{matrix} ON \\ OFF \end{matrix} & \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \end{matrix}$$

$$\leadsto Q_{MLE} = \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix}$$

Il metodo bayesiano

Se la densità a priori per Q è della forma (Dirichlet sulle righe)

$$p(Q = Q) \propto \prod_{i,j \in E} Q_{i \rightarrow j}^{\alpha_{ij}},$$

per opportuni parametri $\alpha_{ij} \geq 0$, la moda della densità sopra è data dalla matrice

$$Q_{i \rightarrow j} = \frac{\alpha_{ij}}{\sum_{k \in E} \alpha_{ik}} \propto \alpha_{ij}.$$

► La formula di Bayes darebbe quindi come densità a posteriori

$$p(Q = Q | X = \gamma) \propto \prod_{i,j \in E} Q_{i \rightarrow j}^{\alpha_{ij} + \gamma_{i \rightarrow j}},$$

con stima di massima ~~verosimiglianza~~ \rightarrow *posteriori*

$$Q_{i \rightarrow j} = \frac{\alpha_{ij} + \gamma_{i \rightarrow j}}{\sum_{k \in E} \alpha_{ik} + \gamma_{i \rightarrow k}} \propto \alpha_{ij} + \gamma_{ij}.$$

regolarizzazione di Laplace

Processi di Markov a salti

Nel caso di processi di Markov a salti, l'argomento è analogo ma si basa sulla formula per la “densità” di probabilità di un cammino.

- Presentiamo solo la stima di massima verosimiglianza. Si consideri un cammino $\gamma = (x_0 \rightarrow x_1 \dots x_n)$ che rimane per un tempo t_1 nello stato x_0 , t_2 nello stato x_1 ecc., e si supponga di osservare $X = \gamma$, ossia tutta la traiettoria da $X_0 = x_0$ fino a $X_{t_1+\dots+t_n} = x_n$.

Processi di Markov a salti

Nel caso di processi di Markov a salti, l'argomento è analogo ma si basa sulla formula per la "densità" di probabilità di un cammino.

- Presentiamo solo la stima di massima verosimiglianza. Si consideri un cammino $\gamma = (x_0 \rightarrow x_1 \dots x_n)$ che rimane per un tempo t_1 nello stato x_0 , t_2 nello stato x_1 ecc., e si supponga di osservare $X = \gamma$, ossia tutta la traiettoria da $X_0 = x_0$ fino a $X_{t_1+\dots+t_n} = x_n$.
- La verosimiglianza per L è

$$\mathcal{L}(L; \gamma) = \prod_{k=1}^n \exp(t_k L_{x_{k-1} \rightarrow x_{k-1}}) L_{x_{k-1} \rightarrow x_k} = \prod_{i \in E} \exp(\gamma_{i \rightarrow i} L_{i \rightarrow i}) \prod_{i \neq j \in E} L_{i \rightarrow j}^{\gamma_{i \rightarrow j}}$$

dove $\gamma_{i \rightarrow j}$ per $i \neq j$ è come prima ma $\gamma_{i \rightarrow i}$ è il tempo totale trascorso dal cammino nello stato $i \in E$.

Massimizzazione

$$L \text{ è tale che } \forall i \quad \sum_j L_{ij} = 0 \Rightarrow L_{ii} = -\sum_{j \neq i} L_{ij}$$

Eliminiamo il vincolo che la somma sulle righe di L è nulla:

$$\exp(\gamma_{i \rightarrow i} L_{i \rightarrow i}) = \exp\left(-\gamma_{i \rightarrow i} \sum_{j \neq i} L_{i \rightarrow j}\right).$$

- Passando ai logaritmi e derivando si ottiene che \mathcal{L}_{MLE} è data dall'espressione, per $i \neq j$,

$$L_{i \rightarrow j} = \frac{\gamma_{i \rightarrow j}}{\gamma_{i \rightarrow i}}.$$

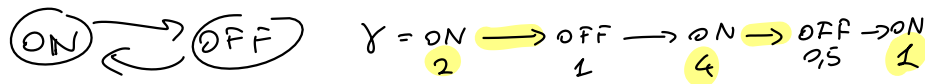
$$\left(\prod_i \prod_{j \neq i} \exp(-\gamma_{i \rightarrow i} L_{ij}) L_{ij}^{\gamma_{ij}} \right)$$

Modelli nascosti

Abbiamo supposto di osservare completamente la catena X in un intervallo (discreto o continuo) di tempi.

- Cosa accade se mancano le osservazioni delle variabili X_k in alcuni tempi?

Esempio di stima di L partendo da Γ



$$\Gamma = \begin{matrix} & \text{ON} & \text{OFF} \\ \text{ON} & 2 & 1 \\ \text{OFF} & 4 & 0.5 \end{matrix} \Rightarrow L = \begin{pmatrix} -\frac{2}{7} & \frac{2}{7} \\ \frac{4}{3} & -\frac{4}{3} \end{pmatrix}$$

Modelli nascosti

Abbiamo supposto di osservare completamente la catena X in un intervallo (discreto o continuo) di tempi.

- ▶ Cosa accade se mancano le osservazioni delle variabili X_k in alcuni tempi?
- ▶ oppure se si osserva solamente una funzione $g(X_k)$ della catena invece, di X_k , o più in generale una funzione $g(X_k, Z_k)$ dove Z è un processo indipendente da X ?

Modelli nascosti

Abbiamo supposto di osservare completamente la catena X in un intervallo (discreto o continuo) di tempi.

- ▶ Cosa accade se mancano le osservazioni delle variabili X_k in alcuni tempi?
- ▶ oppure se si osserva solamente una funzione $g(X_k)$ della catena invece, di X_k , o più in generale una funzione $g(X_k, Z_k)$ dove Z è un processo indipendente da X ?
- ▶ In questa situazione si parla di modelli di Markov nascosti (in inglese *Hidden Markov Models*, HMM) e la ricostruzione di X_k è il problema del filtraggio.

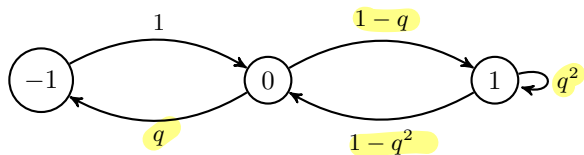
Modelli nascosti

Abbiamo supposto di osservare completamente la catena X in un intervallo (discreto o continuo) di tempi.

- ▶ Cosa accade se mancano le osservazioni delle variabili X_k in alcuni tempi?
- ▶ oppure se si osserva solamente una funzione $g(X_k)$ della catena invece, di X_k , o più in generale una funzione $g(X_k, Z_k)$ dove Z è un processo indipendente da X ?
- ▶ In questa situazione si parla di modelli di Markov nascosti (in inglese *Hidden Markov Models*, HMM) e la ricostruzione di X_k è il problema del filtraggio.
- ▶ Opportuni algoritmi (EM) permettono di stimare i parametri di un HMM.

Un esempio/esercizio

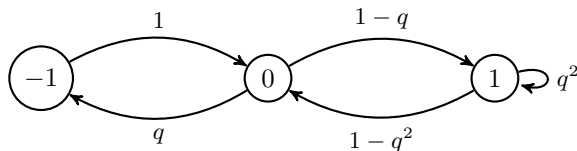
Si consideri una catena di Markov con probabilità di transizione rappresentate in figura, dove $q \in (0, 1)$ è un parametro (non aleatorio).



1. Supponendo che X sia stazionaria, dire al variare di $q \in (0, 1)$ se è più probabile che sia $X_0 = 0$ oppure $X_0 \neq 0$.

Un esempio/esercizio

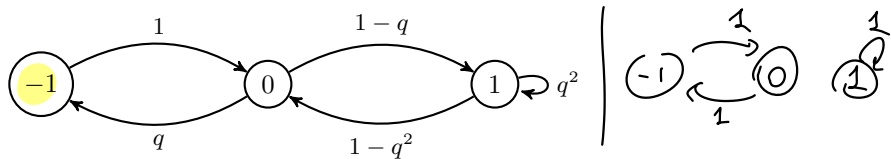
Si consideri una catena di Markov con probabilità di transizione rappresentate in figura, dove $q \in (0, 1)$ è un parametro (non aleatorio).



1. Supponendo che X sia stazionaria, dire al variare di $q \in (0, 1)$ se è più probabile che sia $X_0 = 0$ oppure $X_0 \neq 0$.
2. Si supponga noto a priori che $X_0 = 0$. Si osserva il cammino $0 \rightarrow -1 \rightarrow 0 \rightarrow -1 \rightarrow 0 \rightarrow 1 \rightarrow 1$. Determinare la stima di massima verosimiglianza q_{MLE} .

Un esempio/esercizio

Si consideri una catena di Markov con probabilità di transizione rappresentate in figura, dove $q \in (0, 1)$ è un parametro (non aleatorio).



1. Supponendo che X sia stazionaria, dire al variare di $q \in (0, 1)$ se è più probabile che sia $X_0 = 0$ oppure $X_0 \neq 0$.
2. Si supponga noto a priori che $X_0 = 0$. Si osserva il cammino $0 \xrightarrow{1} -1 \xrightarrow{1} 0 \xrightarrow{1} -1 \xrightarrow{1} 0 \xrightarrow{1-q} 1 \xrightarrow{q^2} 1$. Determinare la stima di massima verosimiglianza q_{MLE} .
3. Si supponga noto a priori che X sia stazionaria. Si osserva lo stesso cammino della domanda di prima. Determinare q_{MLE} .

Bilancio di flusso

$$\begin{cases} \mu_{-1} \cdot 1 = \mu_0 \cdot q \\ \mu_0 \cdot 1 = \mu_{-1} \cdot 1 + \mu_1 (1 - q^2) \end{cases}$$

1.

$$\mu_0 = \mu_0 \cdot q + \mu_1 (1 - q^2)$$

Se $q=1$

$$\mu_{-1} = \mu_0$$

\Downarrow

$$\mu = \left(\frac{d}{2}, \frac{d}{2}, 1-d \right)$$

$$d \in [0,1]$$

$$\leftarrow \begin{matrix} \mu_0(1-q) = \mu_1(1-q^2) \\ \text{Se } q \neq 1 \end{matrix}$$

$$\mu_0 = \mu_1(1+q)$$

$$\left(q(1+q)\mu_1, \mu_1(1+q), \mu_1 \right) \downarrow \text{normalizza}$$

$$\left(q(1+q), 1+q, 1 \right) \frac{1}{1+(1+q)^2}$$

$$2x = q \cdot q \cdot (1-q) \cdot q^2 = q^4(1-q)$$

$$\underline{P(X_0 = 0) = 1}$$

2. $L(q; X=x) = 1 \cdot 2x = q^4(1-q)$ ← Densität
Beob.

$$\frac{\partial}{\partial q} \log(L) = \frac{\partial}{\partial q} [4 \log q + \log(1-q)] =$$

$$= \frac{4}{q} - \frac{1}{1-q} = 0 \Rightarrow 4 - 4q = q$$

$$\Rightarrow \boxed{q = \frac{4}{5}}$$

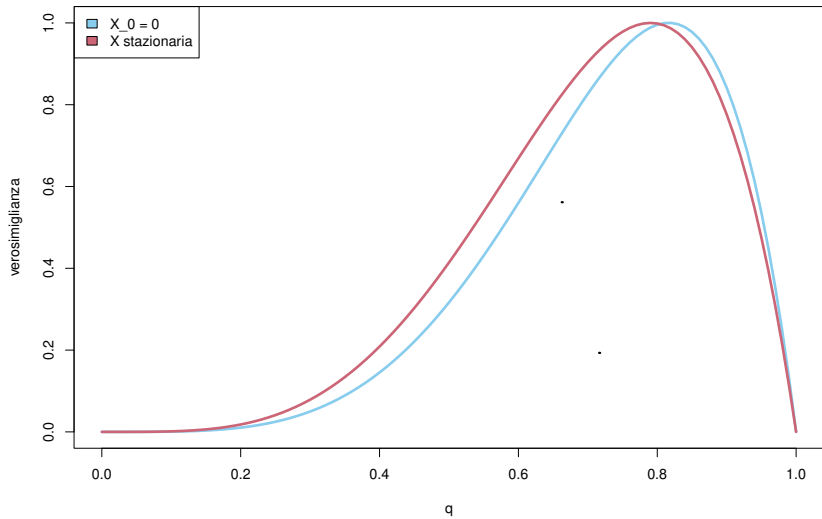
3. Stavolta

$$L(q; X = \gamma) = P(X_0 = 0 | X \text{ staz}, q) Q_\gamma = \frac{1+q}{2+2q+q^2} q^2 (1-q) q^2.$$
$$= P(X=\gamma | q, X \text{ stazionaria}) = \mu_0' \alpha_\gamma = \frac{1+q}{1+(4q)} q^4 (1-q)$$

Stimiamo numericamente q_{MLE} .

```
likelihood_stazionaria = function(q){  
  -(1-q^2)*q^4/(2+2*q+q^2)  
}  
  
x_mle = nlm(likelihood_stazionaria, 1/2)  
  
x_mle$estimate
```

```
## [1] 0.7902904
```



Cenni alla teoria delle code

Introduzione ai modelli di code

La *teoria delle code* studia le linee d'attesa che si possono formare in tante situazioni, ad esempio:

- ▶ persone che vogliono accedere ad un servizio (entrare in un negozio, o pagare alla cassa)

Introduzione ai modelli di code

La *teoria delle code* studia le linee d'attesa che si possono formare in tante situazioni, ad esempio:

- ▶ persone che vogliono accedere ad un servizio (entrare in un negozio, o pagare alla cassa)
- ▶ veicoli che si presentano ad un casello autostradale

Introduzione ai modelli di code

La *teoria delle code* studia le linee d'attesa che si possono formare in tante situazioni, ad esempio:

- ▶ persone che vogliono accedere ad un servizio (entrare in un negozio, o pagare alla cassa)
- ▶ veicoli che si presentano ad un casello autostradale
- ▶ istanze di calcolo che devono essere eseguite da una o più processori in un computer. . .

Introduzione ai modelli di code

La *teoria delle code* studia le linee d'attesa che si possono formare in tante situazioni, ad esempio:

- ▶ persone che vogliono accedere ad un servizio (entrare in un negozio, o pagare alla cassa)
- ▶ veicoli che si presentano ad un casello autostradale
- ▶ istanze di calcolo che devono essere eseguite da una o più processori in un computer. . .
- ▶ L'obiettivo è individuare strategie per migliorare l'esperienza di chi è in attesa (ridurre i tempi) rendendone più efficiente il servizio.

Introduzione ai modelli di code

La *teoria delle code* studia le linee d'attesa che si possono formare in tante situazioni, ad esempio:

- ▶ persone che vogliono accedere ad un servizio (entrare in un negozio, o pagare alla cassa)
- ▶ veicoli che si presentano ad un casello autostradale
- ▶ istanze di calcolo che devono essere eseguite da una o più processori in un computer. . .
- ▶ L'obiettivo è individuare strategie per migliorare l'esperienza di chi è in attesa (ridurre i tempi) rendendone più efficiente il servizio.
- ▶ La teoria delle code è un campo molto esteso, presentiamo i modelli più semplici come esempi di processi di Markov a salti.

Clienti e serventi

Per indicare le persone, le auto, le istanze ecc. da servire usiamo il termine **clienti** (in inglese si usa a volte *jobs*)

Usiamo il termine **serventi** (in inglese *servers*) per chi eroga il servizio richiesto dei clienti.

Aspetti da modellizzare:

- ▶ l'ingresso di uno o più clienti,

Clienti e serventi

Per indicare le persone, le auto, le istanze ecc. da servire usiamo il termine **clienti** (in inglese si usa a volte *jobs*)

Usiamo il termine **serventi** (in inglese *servers*) per chi eroga il servizio richiesto dei clienti.

Aspetti da modellizzare:

- ▶ l'ingresso di uno o più clienti,
- ▶ il tempo d'attesa che un servente prenda in carico il compito richiesto,

Clienti e serventi

Per indicare le persone, le auto, le istanze ecc. da servire usiamo il termine **clienti** (in inglese si usa a volte *jobs*)

Usiamo il termine **serventi** (in inglese *servers*) per chi eroga il servizio richiesto dei clienti.

Aspetti da modellizzare:

- ▶ l'ingresso di uno o più clienti,
- ▶ il tempo d'attesa che un servente prenda in carico il compito richiesto, + *Tempo di Servizio*
- ▶ e infine l'uscita dalla coda quando il compito è svolto

Clienti e serventi

Per indicare le persone, le auto, le istanze ecc. da servire usiamo il termine **clienti** (in inglese si usa a volte *jobs*)

Usiamo il termine **serventi** (in inglese *servers*) per chi eroga il servizio richiesto dei clienti.

Aspetti da modellizzare:

- ▶ l'ingresso di uno o più clienti,
- ▶ il tempo d'attesa che un servente prenda in carico il compito richiesto,
- ▶ e infine l'uscita dalla coda quando il compito è svolto
- ▶ Una volta introdotto un modello, è di interesse calcolare il tempo medio di attesa, il numero medio di clienti in coda e stimare i parametri di un modello sulla base di quantità osservate nella realtà.

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

- ▶ A indica un “processo” di arrivo dei clienti,

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

- ▶ A indica un “processo” di arrivo dei clienti,
- ▶ S la legge del tempo di servizio per ciascun cliente,

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

- ▶ A indica un “processo” di arrivo dei clienti,
- ▶ S la legge del tempo di servizio per ciascun cliente,
- ▶ c il numero dei serventi.

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

- ▶ A indica un “processo” di arrivo dei clienti,
- ▶ S la legge del tempo di servizio per ciascun cliente,
- ▶ c il numero dei serventi.
- ▶ Noi studiamo solo i modelli $M/M/c$: arrivi e tempi di servizio sono Markoviani (a tempi continui).

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

- ▶ A indica un “processo” di arrivo dei clienti,
- ▶ S la legge del tempo di servizio per ciascun cliente,
- ▶ c il numero dei serventi.
- ▶ Noi studiamo solo i modelli $M/M/c$: arrivi e tempi di servizio sono Markoviani (a tempi continui).
- ▶ Formalmente i modelli sono processi di Markov a salti negli stati $E = \mathbb{N}$.

Notazione di Kendall

Kendall propose una notazione abbreviata $A/S/c$:

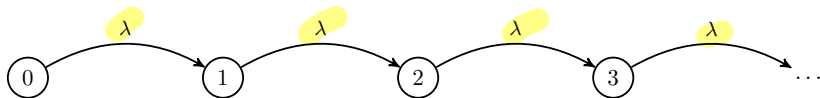
- ▶ A indica un “processo” di arrivo dei clienti,
- ▶ S la legge del tempo di servizio per ciascun cliente,
- ▶ c il numero dei serventi.
- ▶ Noi studiamo solo i modelli $M/M/c$: arrivi e tempi di servizio sono Markoviani (a tempi continui).
- ▶ Formalmente i modelli sono processi di Markov a salti negli stati $E = \mathbb{N}$.
- ▶ Lo stato n indica il numero di clienti in attesa o in corso di servizio.

Caso $M/M/0$: il processo di Poisson

Il modello più semplice è il caso in cui non vi siano serventi (oppure si è interessati solo al processo di arrivo dei clienti): il processo è detto *processo di Poisson* di intensità $\lambda > 0$.

Le intensità di salto sono

$$L_{n \rightarrow n+1} = \lambda, \quad L_{n \rightarrow n} = -\lambda \quad \text{e} \quad L_{n \rightarrow k} = 0 \quad k \neq n, k \neq n+1.$$



► Ogni stato è transitorio, non vi sono distribuzioni invarianti.

$$\begin{matrix} 0 \\ 1 \\ 2 \\ 3 \end{matrix} \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & 0 & -\lambda & \lambda & \dots \end{pmatrix}$$

Legame con la densità Poisson

Se $X_0 = 0$, allora la densità marginale di X_t è Poisson di intensità λt , ossia

$$P(X_t = n) = \mu_n^t \propto \frac{(t\lambda)^n}{n!} = \boxed{\frac{(t\lambda)^n}{n!} \exp(-t\lambda)}.$$

- Basta verificare che valga la *master equation*, per ogni $n \in \mathbb{N}$, $t \geq 0$,

$$\frac{d}{dt} \mu_n^t = (\mu^t L)_n = \begin{cases} -\lambda \mu_0^t & \text{se } n = 0, \\ \lambda(\mu_{n-1}^t - \mu_n^t) & \text{se } n \geq 1. \end{cases}$$

$$\frac{d}{dt} \mu^t = \mu^t L \quad \left[\leadsto \mu^t = \mu^0 \exp(tL) \right]$$

$$\frac{d}{dt} \mu_n^t = \lambda \left[\underbrace{t^{n-1} \frac{\lambda^{n-1}}{(n-1)!} \exp(-t\lambda)}_{\mu_{n-1}^t} \right] - \lambda \left[\underbrace{t^n \frac{\lambda^n}{n!} \exp(-t\lambda)}_{\mu_n^t} \right]$$

Legame con la densità Poisson

Se $X_0 = 0$, allora la densità marginale di X_t è Poisson di intensità λt , ossia

$$\mu_n^t \propto \frac{(t\lambda)^n}{n!} = \frac{(t\lambda)^n}{n!} \exp(-t\lambda).$$

- Basta verificare che valga la *master equation*, per ogni $n \in \mathbb{N}$, $t \geq 0$,

$$\frac{d}{dt} \mu_n^t = (\mu^t L)_n = \begin{cases} -\lambda \mu_0^t & \text{se } n = 0, \\ \lambda(\mu_{n-1}^t - \mu_n^t) & \text{se } n \geq 1. \end{cases}$$

- Calcoliamo quindi

$$\frac{d}{dt} \exp(-t\lambda) \frac{(t\lambda)^n}{n!} = \begin{cases} -\lambda \exp(-t\lambda) & \text{se } n = 0, \\ \frac{nt^{n-1}\lambda^n}{n!} \exp(-t\lambda) - \lambda \frac{(t\lambda)^n}{n!} \exp(-t\lambda) & \text{se } n \geq 1. \end{cases}$$

Legame con la densità Poisson

Se $X_0 = 0$, allora la densità marginale di X_t è Poisson di intensità λt , ossia

$$\mu_n^t \propto \frac{(t\lambda)^n}{n!} = \frac{(t\lambda)^n}{n!} \exp(-t\lambda).$$

- Basta verificare che valga la *master equation*, per ogni $n \in \mathbb{N}$, $t \geq 0$,

$$\frac{d}{dt} \mu_n^t = (\mu^t L)_n = \begin{cases} -\lambda \mu_0^t & \text{se } n = 0, \\ \lambda(\mu_{n-1}^t - \mu_n^t) & \text{se } n \geq 1. \end{cases}$$

- Calcoliamo quindi

$$\frac{d}{dt} \exp(-t\lambda) \frac{(t\lambda)^n}{n!} = \begin{cases} -\lambda \exp(-t\lambda) & \text{se } n = 0, \\ \frac{nt^{n-1}\lambda^n}{n!} \exp(-t\lambda) - \lambda \frac{(t\lambda)^n}{n!} \exp(-t\lambda) & \text{se } n \geq 1. \end{cases}$$

- Per concludere nel caso $n \geq 1$ basta notare che

$$\frac{nt^{n-1}\lambda^n}{n!} \exp(-t\lambda) = \lambda \frac{(t\lambda)^{n-1}}{(n-1)!} \exp(-t\lambda) = \lambda \mu_{n-1}^t.$$

Stima del parametro λ dalle osservazioni

Supponiamo di osservare un cammino $\gamma = (x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n)$ con tempi di permanenza t_1 (nello stato x_0), t_2 (in x_1), \dots , t_n .

- Poiché i salti avvengono solo tra n e $n + 1$, deve essere

$$x_1 = x_0 + 1, x_2 = x_0 + 2, \text{ ecc.}$$

Stima del parametro λ dalle osservazioni

Supponiamo di osservare un cammino $\gamma = (x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n)$ con tempi di permanenza t_1 (nello stato x_0), t_2 (in x_1), \dots , t_n .

- Poiché i salti avvengono solo tra n e $n+1$, deve essere $x_1 = x_0 + 1$, $x_2 = x_0 + 2$, ecc.
- La verosimiglianza è

$$L(\Lambda = \lambda; X = \gamma) = \prod_{k=1}^n \exp(-\lambda t_k) \lambda = \lambda^n \exp(-\lambda T).$$

dove supponiamo noto a priori che $X_0 = x_0$ e $T = \sum_{k=1}^n t_k$.

$$\hookrightarrow \frac{d}{d\lambda} (\lambda^n \exp(-\lambda T)) = 0$$

$$\Rightarrow n \lambda^{n-1} \exp(-\lambda T) - T \lambda^n \exp(-\lambda T) = 0 \rightarrow \lambda = \frac{n}{T}$$

Stima del parametro λ dalle osservazioni

Supponiamo di osservare un cammino $\gamma = (x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n)$ con tempi di permanenza t_1 (nello stato x_0), t_2 (in x_1), \dots , t_n .

- Poiché i salti avvengono solo tra n e $n + 1$, deve essere $x_1 = x_0 + 1$, $x_2 = x_0 + 2$, ecc.
- La verosimiglianza è

$$L(\Lambda = \lambda; X = \gamma) = \prod_{k=1}^n \exp(-\lambda t_k) \lambda = \lambda^n \exp(-\lambda T).$$

dove supponiamo noto a priori che $X_0 = x_0$ e $T = \sum_{k=1}^n t_k$.

- La stima di massima verosimiglianza si trova annullando la derivata rispetto a λ e vale

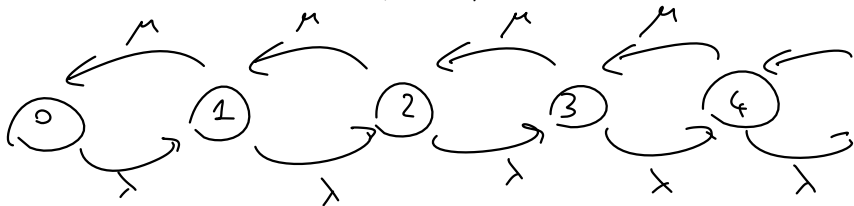
$$\frac{n}{\lambda_{MLE}} - T = 0 \quad \text{quindi} \quad \lambda_{MLE} = \frac{n}{T}.$$

Code $M/M/1$

Supponiamo vi sia un solo servente e che il tempo di servizio per ciascun cliente sia una variabile esponenziale di parametro μ (ogni cliente sia indipendente dagli altri).

- Per arrivare al modello come un processo di Markov a salti, supponiamo non vi siano arrivi: si salta solo da n verso $n - 1$ (se $n \geq 1$) con dei tempi di permanenza esponenziali di parametro μ . Pertanto, se $n \geq 1$,

$$L_{n \rightarrow n-1} = \mu.$$



Code M/M/1

Supponiamo vi sia un solo servente e che il tempo di servizio per ciascun cliente sia una variabile esponenziale di parametro μ (ogni cliente sia indipendente dagli altri).

- ▶ Per arrivare al modello come un processo di Markov a salti, supponiamo non vi siano arrivi: si salta solo da n verso $n - 1$ (se $n \geq 1$) con dei tempi di permanenza esponenziali di parametro μ . Pertanto, se $n \geq 1$,

$$L_{n \rightarrow n-1} = \mu.$$

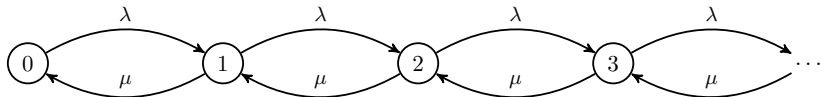
- ▶ Aggiungiamo gli arrivi come un processo di Poisson di intensità λ : per $n \geq 0$,

$$L_{n \rightarrow n+1} = \lambda,$$

e di conseguenza

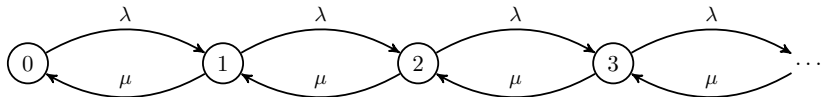
$$L_{n \rightarrow n} = \begin{cases} -\lambda & \text{se } n = 0 \\ -(\lambda + \mu) & \text{se } n \geq 1, \end{cases}$$

avendo posto $L_{k \rightarrow k} = 0$ se $k \notin \{n-1, n, n+1\}$

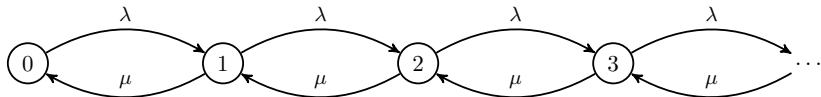


- Ogni stato è ricorrente, ma essendo infiniti stati non è ovvio che esista una distribuzione invariante.

$$L = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \dots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \dots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \dots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \dots \end{pmatrix}$$



- ▶ Ogni stato è ricorrente, ma essendo infiniti stati non è ovvio che esista una distribuzione invariante.
- ▶ Vi è una competizione tra il tasso di arrivo λ e di uscita μ .



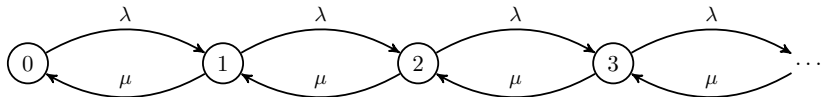
- ▶ Ogni stato è ricorrente, ma essendo infiniti stati non è ovvio che esista una distribuzione invariante.
- ▶ Vi è una competizione tra il tasso di arrivo λ e di uscita μ .
- ▶ Risolvendo l'equazione $\overline{\mu}L = 0$ (o imponendo il bilancio di flusso) si trova

$$\pi_n \propto \left(\frac{\lambda}{\mu}\right)^n$$

perché $\sum_{n=0}^{+\infty} \left(\frac{\lambda}{\mu}\right)^n < +\infty$

$$\pi_0 \cdot \lambda = \pi_1 \cdot \mu \rightarrow \pi_1 = \pi_0 \frac{\lambda}{\mu}$$

$$\pi_1 \cdot \lambda + \cancel{\pi_1 \cdot \mu} = \cancel{\pi_0 \cdot \lambda} + \pi_2 \cdot \mu \Rightarrow \pi_2 = \pi_1 \cdot \frac{\lambda}{\mu} = \pi_0 \left(\frac{\lambda}{\mu}\right)^2$$



- ▶ Ogni stato è ricorrente, ma essendo infiniti stati non è ovvio che esista una distribuzione invariante.
- ▶ Vi è una competizione tra il tasso di arrivo λ e di uscita μ .
- ▶ Risolvendo l'equazione $\mu L = 0$ (o imponendo il bilancio di flusso) si trova

$$\mu_n \propto \left(\frac{\lambda}{\mu} \right)^n.$$

- ▶ Per garantire che μ sia una densità di probabilità, bisogna che

$$\sum_n \left(\frac{\lambda}{\mu} \right)^n < \infty,$$

ossia che $\lambda < \mu$.

- La distribuzione invariante è *geometrica* di parametro $1 - \lambda/\mu$, con valor medio

$$\mathbb{E}[N] = \sum_n n \mu_n = \frac{\lambda}{\mu - \lambda},$$

$$\underline{\underline{\mu > \lambda}}$$

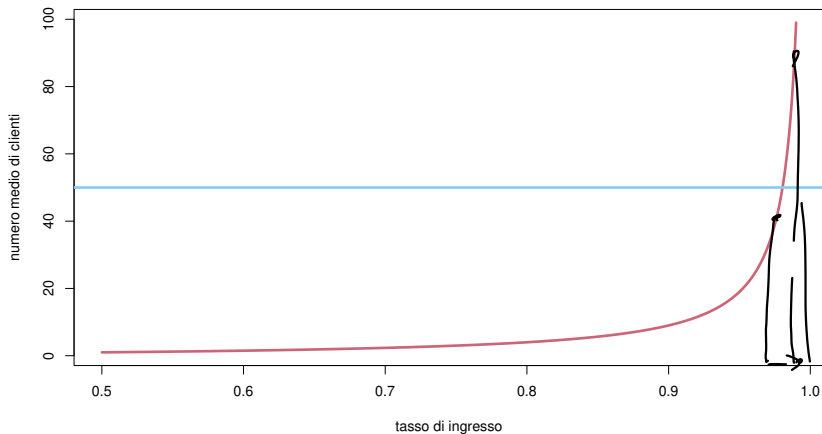


Figure 1: grafico di $\mathbb{E}[N]$ per $\mu = 1$ in funzione di λ (in rosso) e una soglia massima di possibili persone in coda (in azzurro)

Stima dei parametri dalle osservazioni

Stimiamo i parametri (λ, μ) avendo osservato un cammino

$\gamma = (n_0 \rightarrow n_1 \rightarrow \dots \rightarrow n_{\ell-1})$ con i tempi di permanenza $t_1, t_2, \dots, t_{\ell}$.

- Poniamo $T = \sum_{k=1}^{\ell} t_i$ e supponiamo inoltre che il cammino osservato non passi mai per lo stato 0.

Stima dei parametri dalle osservazioni

Stimiamo i parametri (λ, μ) avendo osservato un cammino $\gamma = (n_0 \rightarrow n_1 \rightarrow \dots \rightarrow n_{\ell-1})$ con i tempi di permanenza $t_1, t_2, \dots, t_{\ell}$.

- ▶ Poniamo $T = \sum_{k=1}^{\ell} t_k$ e supponiamo inoltre che il cammino osservato non passi mai per lo stato 0.
- ▶ La verosimiglianza è

$$L(\lambda, \mu; X = \gamma) = \exp(-(\lambda + \mu)T) \lambda^{\gamma_+} \mu^{\gamma_-},$$

dove γ_+ indica il numero di arrivi osservati in γ (ossia transizioni da uno stato n a $n+1$), mentre γ_- il numero di uscite.

$$\left| \lambda_{MLE} = \frac{\gamma_+}{T} \quad \mu_{MLE} = \frac{\gamma_-}{T} \right|$$

Stima dei parametri dalle osservazioni

Stimiamo i parametri (λ, μ) avendo osservato un cammino $\gamma = (n_0 \rightarrow n_1 \rightarrow \dots \rightarrow n_{\ell-1})$ con i tempi di permanenza $t_1, t_2, \dots, t_{\ell}$.

- ▶ Poniamo $T = \sum_{k=1}^{\ell} t_i$ e supponiamo inoltre che il cammino osservato non passi mai per lo stato 0.
- ▶ La verosimiglianza è

$$L(\lambda, \mu; X = \gamma) = \exp(-(\lambda + \mu)T) \lambda^{\gamma_+} \mu^{\gamma_-},$$

dove γ_+ indica il numero di arrivi osservati in γ (ossia transizioni da uno stato n a $n+1$), mentre γ_- il numero di uscite.

- ▶ La stima di massima verosimiglianza è

$$\lambda_{MLE} = \frac{\gamma_+}{T}, \quad \mu_{MLE} = \frac{\gamma_-}{T}.$$

Se il cammino osservato trascorre un tempo T_0 nello stato 0, l'espressione per la verosimiglianza cambia: al posto di $-(\lambda + \mu)T$ si trova $-\lambda T - \mu(T - T_0)$

► λ_{MLE} non cambia, invece

$$\mu_{MLE} = \frac{\gamma_-}{T - T_0}.$$

Se il cammino osservato trascorre un tempo T_0 nello stato 0, l'espressione per la verosimiglianza cambia: al posto di $-(\lambda + \mu)T$ si trova $-\lambda T - \mu(T - T_0)$

- ▶ λ_{MLE} non cambia, invece

$$\mu_{MLE} = \frac{\gamma_-}{T - T_0}.$$

- ▶ Interpretazione: il tempo in cui la coda è vuota non si può usare per stimare il tasso di uscita.

Se il cammino osservato trascorre un tempo T_0 nello stato 0, l'espressione per la verosimiglianza cambia: al posto di $-(\lambda + \mu)T$ si trova $-\lambda T - \mu(T - T_0)$

- ▶ λ_{MLE} non cambia, invece

$$\mu_{MLE} = \frac{\gamma_-}{T - T_0}.$$

- ▶ Interpretazione: il tempo in cui la coda è vuota non si può usare per stimare il tasso di uscita.
- ▶ *Esempio:* In un intervallo di 10 minuti si osservano 5 persone arrivare alla cassa di un supermercato e 3 persone uscirne. Supponendo che la cassa non sia mai senza lavoro si stimano i parametri $\lambda = 1/2$ persone al minuto, $\mu = 3/10$ persone al minuto. Se invece la cassa è rimasta priva di persone in coda per 4 minuti, si stima $\mu = 3/6 = 1/2$ persone al minuto.

Consideriamo la situazione con un numero arbitrariamente grande, idealmente infinito, di serventi ($M/M/\infty$).

- ▶ Il tempo di servizio per ciascun cliente sia una variabile esponenziale di parametro μ (e ogni cliente sia indipendente dagli altri).

Consideriamo la situazione con un numero arbitrariamente grande, idealmente infinito, di serventi ($M/M/\infty$).

- ▶ Il tempo di servizio per ciascun cliente sia una variabile esponenziale di parametro μ (e ogni cliente sia indipendente dagli altri).
- ▶ Per arrivare al modello, consideriamo il caso in cui non vi siano arrivi: si osservano salti da n a $n - 1$ con dei tempi di permanenza dati dal minimo di n variabili aleatorie esponenziali indipendenti tra loro (la transizione avviene appena il cliente che impegna meno tempo tra gli n in servizio lascia la coda).

$$\begin{aligned} \text{Tempo di permanenza in stato } n &= \min\{T_1, T_2, T_3, \dots, T_n\} \\ &\Rightarrow \text{Exp}(n\mu) \end{aligned}$$

Esercizio: il minimo di n variabili esponenziali indipendenti, tutte di parametro μ , ha densità esponenziale di parametro $n\mu$.

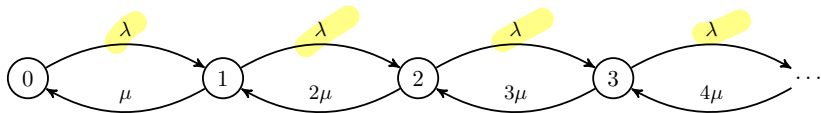
Pertanto, si avrà (se $n \geq 1$)

$$L_{n \rightarrow n-1} = n\mu.$$

Nel caso in cui vi siano arrivi con un processo di Poisson di intensità λ , poniamo, per $n \geq 0$,

$$L_{n \rightarrow n+1} = \lambda,$$

e di conseguenza



$\lambda < \mu$ ok

$\mu \leq \lambda < 2\mu$ pure ok ----

Distribuzione invariante

Come nel caso $M/M/1$, ogni stato è ricorrente.

- La competizione tra il tasso di arrivo λ e di uscita μ , è “smorzata” dal fatto che per n abbastanza grande si ha comunque $\lambda < n\mu$.

Bilancio di Flusso

$$\pi_0 \cdot \lambda = \pi_1 \cdot \mu \quad \Rightarrow \quad \pi_1 = \pi_0 \frac{\lambda}{\mu}$$

$$\pi_n \cdot \lambda + \pi_n \cdot n\mu = \pi_{n-1} \cdot \lambda + \pi_{n+1} \cdot (n+1)\mu$$

Supponendo $\pi_n \propto \left(\frac{\lambda}{\mu}\right)^n \cdot \frac{1}{n!}$

$$\left(\frac{\lambda}{\mu}\right)^n \cdot \frac{1}{n!} \cdot \lambda + \left(\frac{\lambda}{\mu}\right)^n \cdot \frac{n \cdot \mu}{n!}$$

$\stackrel{?}{=}$

$$\lambda \left(\frac{\lambda}{\mu}\right)^{n-1} \cdot \frac{1}{(n-1)!} + \left(\frac{\lambda}{\mu}\right)^{n+1} \cdot \frac{(n+1)}{(n+1)!} \mu$$

Distribuzione invariante

Come nel caso $M/M/1$, ogni stato è ricorrente.

- ▶ La competizione tra il tasso di arrivo λ e di uscita μ , è “smorzata” dal fatto che per n abbastanza grande si ha comunque $\lambda < n\mu$.
- ▶ Infatti una distribuzione invariante esiste sempre: resolvendo il sistema $\pi L = 0$ si trova una densità di Poisson di parametro λ/μ :

$$\pi_n \propto \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n.$$

$$\mathbb{E}[N] = \frac{\lambda}{\mu}$$

Stima dei parametri dalle osservazioni

Come nel caso $M/M/1$, per stimare (λ, μ) sulla base dell'osservazione di un cammino $\gamma = (n_0 \rightarrow n_1 \rightarrow \dots \rightarrow n_\ell)$ con i tempi di permanenza t_1, t_2, \dots, t_ℓ scriviamo la verosimiglianza:

$$L(\lambda, \mu; X = \gamma) \propto \exp(-\lambda T - \mu T_\gamma) \lambda^{\gamma_+} \mu^{\gamma_-},$$

dove $T = \sum_{k=1}^{\ell} t_k$, γ_+ e γ_- sono come nel caso $M/M/1$.

Il termine nuovo è

$$T_\gamma = \sum_{k=1}^{\ell} t_k n_k,$$

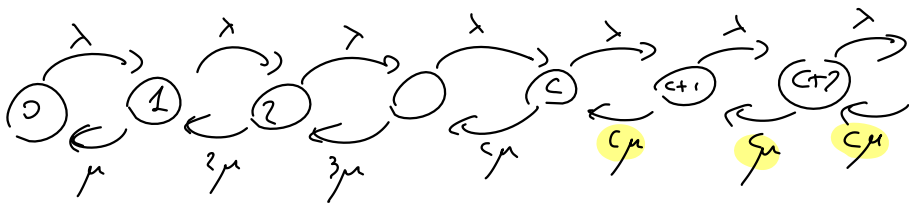
(il tempo totale trascorso da tutti i clienti osservati nella coda). La stima di massima verosimiglianza è

$$\lambda_{MLE} = \frac{\gamma_+}{T}, \quad \mu_{MLE} = \frac{\gamma_-}{T_\gamma}.$$

Il caso $M/M/c$

Il caso $M/M/c$ con $2 \leq c < \infty$ è intermedio tra i gli estremi che abbiamo considerato.

- Una distribuzione invariante esiste se e solo se $\lambda < c\mu$, ma le formule sono meno eleganti.



Capitolo Gaussiane

PCA

Analisi delle componenti principali (PCA)

Problema: *ridurre la dimensionalità* di una variabile $Y \in \mathbb{R}^d$ (o di un campione di taglia n), con $d \gg 1$, definendo una variabile $X \in \mathbb{R}^k$, con $k \ll d$.

- ▶ Questo può essere utile per rappresentare graficamente Y (ad esempio se $k = 2$) ma soprattutto anche per velocizzare l'esecuzione di algoritmi che in dimensione alta possono risultare lenti. o imprecisi

Analisi delle componenti principali (PCA)

Problema: *ridurre la dimensionalità* di una variabile $Y \in \mathbb{R}^d$ (o di un campione di taglia n), con $d \gg 1$, definendo una variabile $X \in \mathbb{R}^k$, con $k \ll d$.

- ▶ Questo può essere utile per rappresentare graficamente Y (ad esempio se $k = 2$) ma soprattutto anche per velocizzare l'esecuzione di algoritmi che in dimensione alta possono risultare lenti.
- ▶ Problema estremamente comune e molteplici tecniche per affrontarlo. L'**analisi delle componenti principali** (in inglese *principal component analysis*, abbreviato PCA) è forse la più semplice, ma spesso efficace.

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- ▶ Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- ▶ Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

- ▶ Posta $Y' = U_Y Y$, la matrice delle covarianze di Y' è diagonale.

Punto di vista teorico

La PCA si spiega a partire dalla *standardizzazione* di un vettore aleatorio.

- ▶ Data $Y \in \mathbb{R}^d$, la matrice delle covarianze Σ_Y può essere diagonalizzata ossia esiste $U_Y \in \mathbb{R}^{d \times d}$ ortogonale ($U_Y^T = U_Y^{-1}$) tale che

$$U_Y \Sigma_Y U_Y^T = D_Y$$

è diagonale (e contiene gli autovalori di Σ_Y).

- ▶ Posta $Y' = U_Y Y$, la matrice delle covarianze di Y' è diagonale.
- ▶ Si definisce $X \in \mathbb{R}^k$ come la variabile congiunta delle k coordinate di Y' che hanno varianza maggiore. Vale quindi

$$X = \Pi_Y Y.$$

dove Π_Y è la proiezione ortogonale sul sottospazio k -dimensionale di “maggior variabilità”.

Tre problemi

1. Le direzioni di maggior variabilità potrebbero essere dovute al fatto che i *valori* sono grandi, non che ci sia effettiva variabilità:
⇒ passare dalla matrice delle **covarianze** a quella di **correlazione** (ossia riscalare/centrare componente per componente).

Tre problemi

1. Le direzioni di maggior variabilità potrebbero essere dovute al fatto che i *valori* sono grandi, non che ci sia effettiva variabilità:
 \Rightarrow passare dalla matrice delle covarianze a quella di **correlazione** (ossia riscalare/centrare componente per componente).
2. Si dispone solo di n di osservazioni (y_1, \dots, y_n) associate a variabili aleatorie (Y_1, \dots, Y_n) , tutte indipendenti tra loro e con la stessa legge di Y (un campione). Come stimare Π_Y ?

Tre problemi

1. Le direzioni di maggior variabilità potrebbero essere dovute al fatto che i *valori* sono grandi, non che ci sia effettiva variabilità:
 \Rightarrow passare dalla matrice delle covarianze a quella di **correlazione** (ossia riscalare/centrare componente per componente).
2. Si dispone solo di n di osservazioni (y_1, \dots, y_n) associate a variabili aleatorie (Y_1, \dots, Y_n) , tutte indipendenti tra loro e con la stessa legge di Y (un campione). Come stimare Π_Y ?
3. la PCA è una stima che si può giustificare mediante MLE/MAP?

PCA empirica

- La matrice delle covarianze teorica Σ_Y non è nota \rightarrow

$$\Sigma_y = \frac{1}{n} \sum_{i=1} (y_i - \bar{y})(y_i - \bar{y})^T.$$

matrice campionaria delle covarianze

- ▶ La matrice delle covarianze teorica Σ_Y non è nota \rightarrow

$$\Sigma_y = \frac{1}{n} \sum_{i=1} (y_i - \bar{y})(y_i - \bar{y})^T.$$

- ▶ Il teorema spettrale applicato Σ_y determina una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

PCA empirica

- ▶ La matrice delle covarianze teorica Σ_Y non è nota \rightarrow

$$\Sigma_y = \frac{1}{n} \sum_{i=1} (y_i - \bar{y})(y_i - \bar{y})^T.$$

- ▶ Il teorema spettrale applicato Σ_y determina una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

- ▶ Definiamo $\Pi_y \in \mathbb{R}^{k \times d}$ come la proiezione nel sottospazio k -dim degli autovettori con autovalori il più grande possibile.

PCA empirica

- ▶ La matrice delle covarianze teorica Σ_Y non è nota \rightarrow

$$\Sigma_y = \frac{1}{n} \sum_{i=1} (y_i - \bar{y})(y_i - \bar{y})^T.$$

- ▶ Il teorema spettrale applicato Σ_y determina una matrice ortogonale $U_y \in \mathbb{R}^{d \times d}$ e una matrice diagonale $D_y \in \mathbb{R}^{d \times d}$ (contenente gli autovalori di Σ_y) tali che

$$U_y \Sigma_y U_y^T = D_y.$$

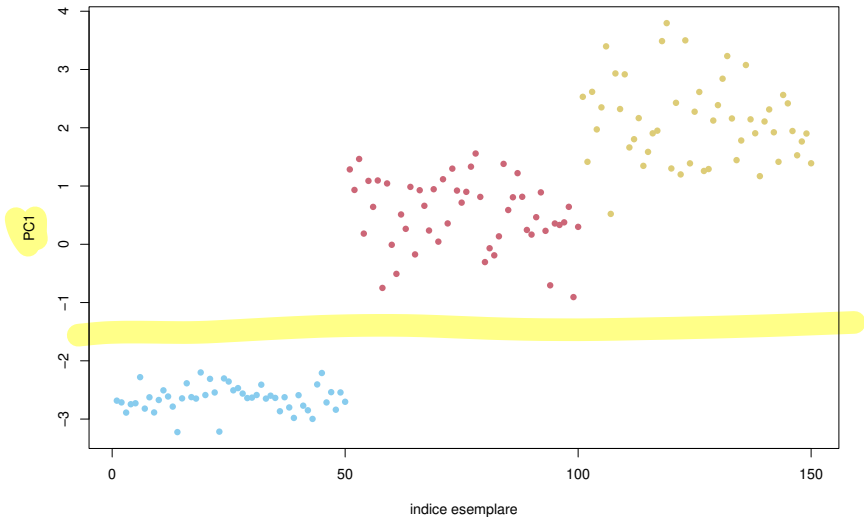
- ▶ Definiamo $\Pi_y \in \mathbb{R}^{k \times d}$ come la proiezione nel sottospazio k -dim degli autovettori con autovalori il più grande possibile.
- ▶ Il “riassunto” (loadings) è il vettore delle osservazioni proiettate $x_i = \Pi_y y_i$.

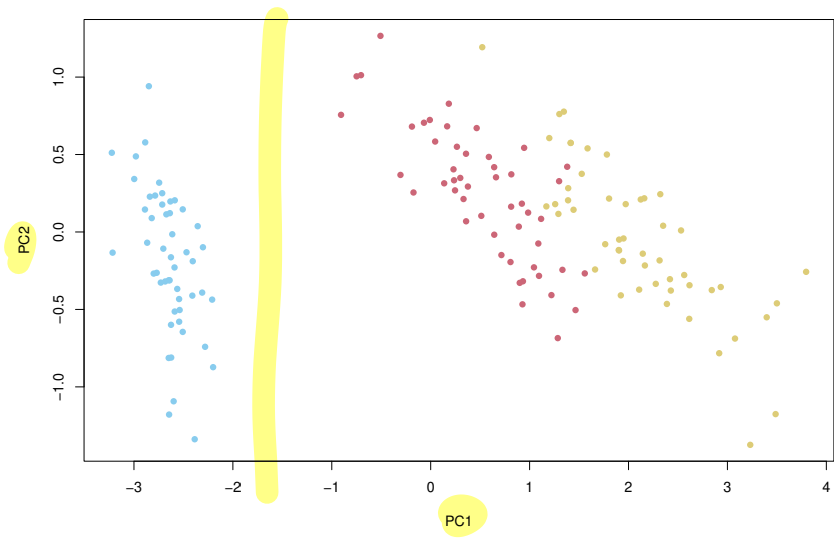
Applichiamo la PCA in R usando la funzione specifica `prcomp()` (oppure `princomp()`). Vediamo ad esempio sul dataset Iris:

L'oggetto contiene diverse informazioni sulla PCA, la base di vettori U (una matrice 4×4) e le deviazioni standard delle varie componenti.

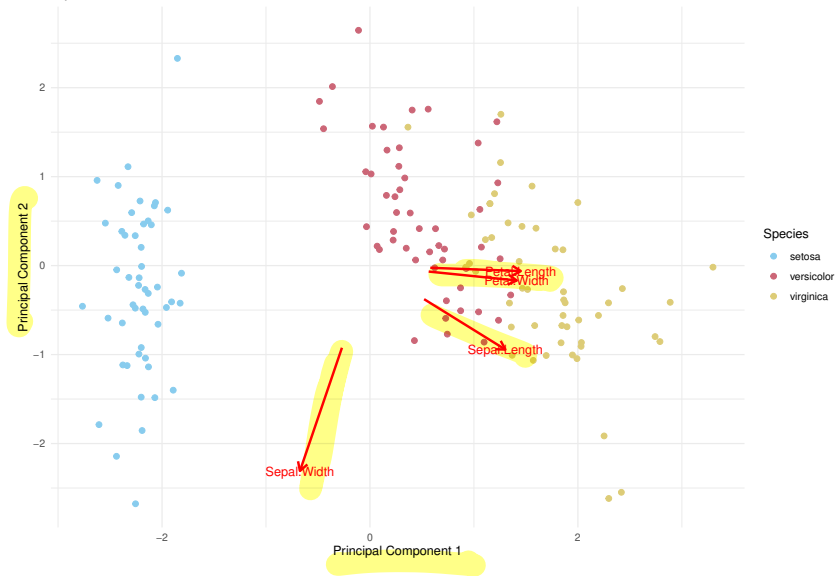
```
##              PC1              PC2              PC3
## Sepal.Length  0.36138659 -0.65658877  0.58202985  0.3154
## Sepal.Width   -0.08452251 -0.73016143 -0.59791083 -0.3197
## Petal.Length  0.85667061  0.17337266 -0.07623608 -0.4798
## Petal.Width   0.35828920  0.07548102 -0.54583143  0.7536
## [1] 2.0562689 0.4926162 0.2796596 0.1543862
```

La PCA aiuta a separare tre specie (almeno la prima dalle altre due) come evidenziamo con la diversa colorazione.



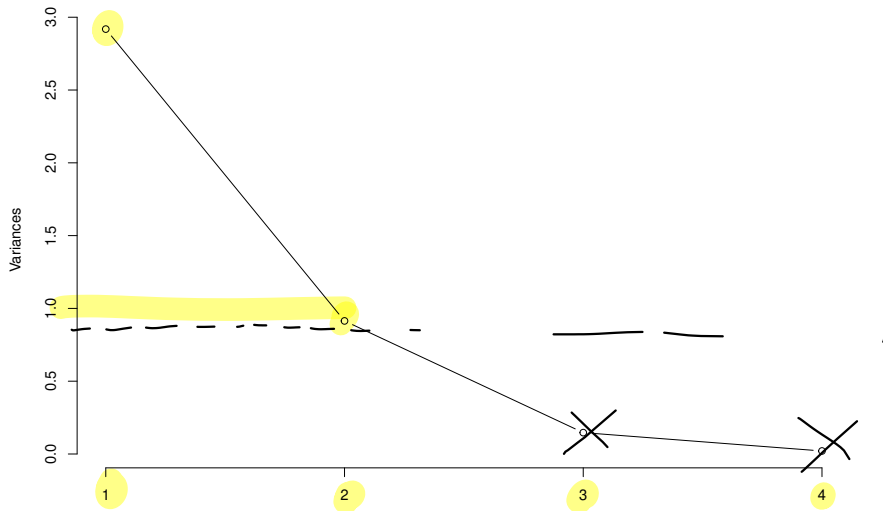


Biplot of PCA on Iris Dataset



Scree plot

Varianze delle componenti principali, dataset Iris



Una giustificazione tramite MLE della PCA

È possibile giustificare il metodo di PCA in termini di una stima di massima verosimiglianza per un opportuno modello gaussiano.

- ▶ **Idea:** con la PCA stiamo recuperando un “segnale” (X) osservandone una versione “rumorosa” e disposta su un sottospazio non noto.



- fissata la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- ▶ fissata la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- ▶ Il “segnale” da ricostruire è quindi AZ (quello che nella PCA abbiamo chiamato X) e W è una variabile che rappresenta il “rumore” aggiunto.

- ▶ fissata la dimensione k , introduciamo una variabile standardizzata $Z \in \mathbb{R}^k$ e imponiamo che valga

$$Y = AZ + W,$$

dove $A \in \mathbb{R}^{d \times k}$ è una matrice non nota (rispetto all'informazione priori).

- ▶ Il “segnale” da ricostruire è quindi AZ (quello che nella PCA abbiamo chiamato X) e W è una variabile che rappresenta il “rumore” aggiunto.
- ▶ Supponiamo che Z , W siano indipendenti con densità gaussiane centrate e, oltre a $\Sigma_Z = Id$, supponiamo che $\Sigma_W = \sigma_0^2 Id$, per una costante opportuna (nota a priori e sufficientemente piccola).

Supponendo nota la matrice A , allora la densità di Y , è anch'essa una gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$, ch     una funzione di A . Pertanto la verosimiglianza di A associata ad Y si scrive

$$L(A; y) = p(Y = y|A) \propto \exp \left(-\frac{1}{2} \left(y^T \sigma_Y^{-1} y + \log(\det(\Sigma_Y)) \right) \right).$$

- se supponiamo di avere n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – in particolare con la stessa matrice A , la verosimiglianza si ottiene come prodotto della funzione sopra (cambiando i valori osservati)

$$L(A; y_1, \dots, y_n) \propto \exp \left(-\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y)) \right) \right).$$

Supponendo nota la matrice A , allora la densità di Y , è anch'essa una gaussiana centrata, con covarianza $\Sigma_Y = AA^T + \sigma_0^2 Id$, ch     una funzione di A . Pertanto la verosimiglianza di A associata ad Y si scrive

$$L(A; y) = p(Y = y|A) \propto \exp \left(-\frac{1}{2} \left(y^T \Sigma_Y^{-1} y + \log(\det(\Sigma_Y)) \right) \right).$$

- se supponiamo di avere n osservazioni indipendenti $Y_i = y_i$, tutte gaussiane con gli stessi parametri – in particolare con la stessa matrice A , la verosimiglianza si ottiene come prodotto della funzione sopra (cambiando i valori osservati)

$$L(A; y_1, \dots, y_n) \propto \exp \left(-\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y)) \right) \right).$$

- la massima verosimiglianza si ottiene minimizzando

$$A \mapsto \frac{1}{n} \sum_{i=1}^n y_i^T \Sigma_Y^{-1} y_i + \log(\det(\Sigma_Y))$$

- La stima di massima verosimiglianza per A è data da

$$A_{\text{MLE}} = U_{y|k}(D_{y|k} - \sigma_0^2 Id)^{1/2},$$

dove $U_{y|k} \in \mathbb{R}^{d \times k}$ indica la matrice corrispondente ai k autovettori della covarianza campionaria $\Sigma_y = \sum_{i=1}^n y_i y_i^T$ con autovalori più grandi, e $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- La stima di massima verosimiglianza per A è data da

$$A_{\text{MLE}} = U_{y|k}(D_{y|k} - \sigma_0^2 Id)^{1/2},$$

dove $U_{y|k} \in \mathbb{R}^{d \times k}$ indica la matrice corrispondente ai k autovettori della covarianza campionaria $\Sigma_y = \sum_{i=1}^n y_i y_i^T$ con autovalori più grandi, e $D_{y|k} \in \mathbb{R}^{k \times k}$ indica la matrice diagonale contenente tali autovalori nell'ordine corrispondente.

- Tutto questo purché σ_0^2 sia sufficientemente piccolo. Nel limite $\sigma_0 \rightarrow 0$ si ottiene che $A_{\text{MLE}} = U_{y|k} D_{y|k}^{1/2}$ e la variabile $A_{\text{MLE}} X$ si identifica con $\Pi_y Y$.